



Encyclopedia of Modern Optics

Second Edition

Editors in Chief

Bob Guenther
Duncan Steel

**ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION**

ENCYCLOPEDIA OF MODERN OPTICS SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 1

Fiber Optics ■ Coherence ■ Diffraction ■ History of Lasers
■ Geometrical Optics ■ Optical Modulation ■ Polarization
■ Photonic Crystals ■ Communications ■ Lasers and Amplifiers
■ TeraHertz ■ Quantum Cavity Physics ■ Coherent Control



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 1

Editorial Board	v
List of Contributors for Volume 1	ix
Contents of All Volumes	xi
Editors in Chief	xvii
Introduction	xix

VOLUME 1

Dispersion	<i>L. Thévenaz</i>	1
Nonlinear Optics	<i>K Thyagarajan and AK Ghatak</i>	10
Fabrication of Optical Fiber	<i>D Hewak</i>	27
Overview: Coherence	<i>A Sharma, AK Ghatak, and HC Kandpal</i>	34
Diffraction Gratings	<i>J Turunen and T Vallius</i>	51
Fraunhofer Diffraction	<i>BD Guenther</i>	62
Fresnel Diffraction	<i>BD Guenther</i>	81
Early History of Quantum Electronics	<i>CH Townes</i>	97
Lenses and Mirrors	<i>A Nussbaum</i>	101
Aberrations	<i>A Nussbaum</i>	114
Prisms	<i>A Nussbaum</i>	124
Acousto-Optics	<i>M Gottlieb and D Suhre</i>	130
Electro-Optics	<i>LR Dalton</i>	141
Polarization Introduction	<i>JM Bennett</i>	148
Matrix Analysis	<i>BD Guenther</i>	162
Electromagnetic Theory	<i>SG Johnson and JD Ioannopoulos</i>	169
Nonlinear Optics in Photonic Crystal Fibers	<i>JE Sharping and P Kumar</i>	177
Photonic Crystal Lasers, Cavities and Waveguides	<i>J O'Brien and W Kuang</i>	185
Microstructure Fibers	<i>RS Windeler</i>	195
Omnidirectional Surfaces and Fibers	<i>S Hart, G Benoit, and Y Fink</i>	207
Guided Wave Optics	<i>Alan Mickelson</i>	221
Optical Fiber Gratings	<i>Paul S Westbrook and Tristan Kremp</i>	229
Optical Amplifiers: SOAs	<i>Michael J Connelly</i>	242
Wavelength Division Multiplexing	<i>Klaus Grobe</i>	255
Coherent Lightwave Systems	<i>Michael J Connelly</i>	291
Measuring Fiber Characteristics	<i>A Girard</i>	308
Optical Fiber Cables	<i>G Galliano</i>	328

Passive Optical Components	<i>D Suino</i>	336
Nonlinear Effects (Basics)	<i>G Millot and P Tchofo-Dinda</i>	345
Basic Concepts of Optical Amplifiers	<i>MFS Ferreira</i>	351
Erbium Doped Fiber Amplifiers for Lightwave Systems	<i>P Bollond</i>	355
All-Optical Signal Regeneration	<i>O Leclerc</i>	364
Heterodyning	<i>T-C Poon</i>	373
Terahertz Lasers	<i>Benjamin S Williams and Qing Hu</i>	379
THz Molecular Spectroscopy	<i>Zbigniew Kisiel</i>	387
Broadband Terahertz Sources	<i>Kang Liu and Xi-Cheng Zhang</i>	403
Terahertz Detectors	<i>Antoni Rogalski</i>	418
Gravitational Wave Detection	<i>Marie-Anne Bizouard and Nelson Christensen</i>	432
Cavity QED Effects in Molecular Systems	<i>Stéphane Kéna-Cohen</i>	445
Cavity QED	<i>H Walther</i>	451
Cavity QED in Semiconductors	<i>M Kira, W Hoyer, SW Koch, G Khitrova, and HM Gibbs</i>	458
Silicon Qubits	<i>Thaddeus D Ladd and Malcolm S Carroll</i>	467
Entanglement and Quantum Information	<i>PG Kwiat and DFK James</i>	478
Applications in Semiconductors	<i>HM van Driel and JE Sipe</i>	486

LIST OF CONTRIBUTORS FOR VOLUME 1

JM Bennett
Michelson Laboratory, China Lake, CA, USA

G Benoit
Massachusetts Institute of Technology, Cambridge, MA, USA

Marie-Anne Bizouard
Paris-Saclay University, Orsay, France

P Bollond
IDS Uniphase Corporation, Ewing, NJ, USA

Malcolm S Carroll
Sandia National Laboratory, Albuquerque, NM, United States

Nelson Christensen
Carleton College, Northfield, MN, United States and University of Côte d'Azur, Nice, France

Michael J Connelly
University of Limerick, Limerick, Ireland

LR Dalton
University of Washington, Seattle, WA, USA

MFS Ferreira
University of Aveiro, Aveiro, Portugal

Y Fink
Massachusetts Institute of Technology, Cambridge, MA, USA

G Galliano
Telecom Italia Lab, Torino, Italy

AK Ghatak
Indian Institute of Technology, New Delhi, India

HM Gibbs
University of Arizona, Tucson, AZ, USA

A Girard
EXFO, Quebec, Canada

M Gottlieb
Carnegie Mellon University, Pittsburgh, PA, USA

Klaus Grobe
ADVA Optical Networking SE, Martinsried, Germany

BD Guenther
Duke University, Durham, NC, USA

S Hart
Massachusetts Institute of Technology, Cambridge, MA, USA

D Hewak
University of Southampton, Southampton, UK

W Hoyer
Philipps-University, Marburg, Germany

Qing Hu
MIT, Cambridge, MA, United States

DFV James
Los Alamos National Laboratory, Los Alamos, NM, USA

JD Joannopoulos
Massachusetts Institute of Technology, Cambridge, MA, USA

SG Johnson
Massachusetts Institute of Technology, Cambridge, MA, USA

Stéphane Kéna-Cohen
Polytechnique Montréal, Department of Engineering Physics, Montreal, QC, Canada

HC Kandpal
National Physical Laboratory, New Delhi, India

G Khitrova
University of Arizona, Tucson, AZ, USA

M Kira
Philipps-University, Marburg, Germany

Zbigniew Kisiel
Institute of Physics of the Polish Academy of Sciences, Warsaw, Poland

SW Koch
Philipps-University, Marburg, Germany

Tristan Kremp
OFS Fitel, LLC, Somerset, NJ, United States

W Kuang
University of Southern California, Los Angeles, CA, USA

P Kumar
Northwestern University, Evanston, IL, USA

PG Kwiat
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Thaddeus D Ladd
HRL Laboratories, LLC, Malibu, CA, United States

O Leclerc
Alcatel Research & Innovation, Marcoussis, France

Kang Liu
University of Rochester, Rochester, NY, United States

Alan Mickelson
University of Colorado at Boulder, Boulder, CO, United States

G Millot
Université de Bourgogne, Dijon, France

A Nussbaum
University of Minnesota, Minneapolis, MN, USA

J O'Brien
University of Southern California, Los Angeles, CA, USA

T-C Poon
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Antoni Rogalski
Military University of Technology, Warsaw, Poland

A Sharma
Indian Institute of Technology, New Delhi, India

JE Sharping
Cornell University, Ithaca, NY, USA

JE Sipe
University of Toronto, Toronto, ON, Canada

D Suhre
Carnegie Mellon University, Pittsburgh, PA, USA

D Suino
Telecom Italia Lab, Torino, Italy

P Tchofo-Dinda
Université de Bourgogne, Dijon, France

L Thévenaz
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

K Thyagarajan
Indian Institute of Technology, New Delhi, India

CH Townes
University of California, Berkeley, CA, USA

J Turunen
University of Joensuu, Joensuu, Finland

T Vallius
University of Joensuu, Joensuu, Finland

HM van Driel
University of Toronto, Toronto, ON, Canada

H Walther
University of Munich and Max-Planck-Institute for Quantum Optics, Garching, Germany

Paul S Westbrook
OFS Fitel, LLC, Somerset, NJ, United States

Benjamin S Williams
University of California, Los Angeles, CA, United States

RS Windeler
OFS Laboratories, Murray Hill, NJ, USA

Xi-Cheng Zhang
University of Rochester, Rochester, NY, United States; ITMO University, Saint-Petersburg, Russia; and Capital Normal University, Beijing, China

CONTENTS OF ALL VOLUMES

VOLUME 1

Dispersion	<i>L Thévenaz</i>	1
Nonlinear Optics	<i>K Thyagarajan and AK Ghatak</i>	10
Fabrication of Optical Fiber	<i>D Hewak</i>	27
Overview: Coherence	<i>A Sharma, AK Ghatak, and HC Kandpal</i>	34
Diffraction Gratings	<i>J Turunen and T Vallius</i>	51
Fraunhofer Diffraction	<i>BD Guenther</i>	62
Fresnel Diffraction	<i>BD Guenther</i>	81
Early History of Quantum Electronics	<i>CH Townes</i>	97
Lenses and Mirrors	<i>A Nussbaum</i>	101
Aberrations	<i>A Nussbaum</i>	114
Prisms	<i>A Nussbaum</i>	124
Acousto-Optics	<i>M Gottlieb and D Suhre</i>	130
Electro-Optics	<i>LR Dalton</i>	141
Polarization Introduction	<i>JM Bennett</i>	148
Matrix Analysis	<i>BD Guenther</i>	162
Electromagnetic Theory	<i>SG Johnson and JD Ioannopoulos</i>	169
Nonlinear Optics in Photonic Crystal Fibers	<i>JE Sharping and P Kumar</i>	177
Photonic Crystal Lasers, Cavities and Waveguides	<i>J O'Brien and W Kuang</i>	185
Microstructure Fibers	<i>RS Windeler</i>	195
Omnidirectional Surfaces and Fibers	<i>S Hart, G Benoit, and Y Fink</i>	207
Guided Wave Optics	<i>Alan Mickelson</i>	221
Optical Fiber Gratings	<i>Paul S Westbrook and Tristan Kremp</i>	229
Optical Amplifiers: SOAs	<i>Michael J Connelly</i>	242
Wavelength Division Multiplexing	<i>Klaus Grobe</i>	255
Coherent Lightwave Systems	<i>Michael J Connelly</i>	291
Measuring Fiber Characteristics	<i>A Girard</i>	308
Optical Fiber Cables	<i>G Galliano</i>	328
Passive Optical Components	<i>D Suino</i>	336
Nonlinear Effects (Basics)	<i>G Millot and P Tchofo-Dinda</i>	345
Basic Concepts of Optical Amplifiers	<i>MFS Ferreira</i>	351
Erbium Doped Fiber Amplifiers for Lightwave Systems	<i>P Bollond</i>	355
All-Optical Signal Regeneration	<i>O Leclerc</i>	364
Heterodyning	<i>T-C Poon</i>	373
Terahertz Lasers	<i>Benjamin S Williams and Qing Hu</i>	379

THz Molecular Spectroscopy	<i>Zbigniew Kisiel</i>	387
Broadband Terahertz Sources	<i>Kang Liu and Xi-Cheng Zhang</i>	403
Terahertz Detectors	<i>Antoni Rogalski</i>	418
Gravitational Wave Detection	<i>Marie-Anne Bizouard and Nelson Christensen</i>	432
Cavity QED Effects in Molecular Systems	<i>Stéphane Kéna-Cohen</i>	445
Cavity QED	<i>H Walther</i>	451
Cavity QED in Semiconductors	<i>M Kira, W Hoyer, SW Koch, G Khitrova, and HM Gibbs</i>	458
Silicon Qubits	<i>Thaddeus D Ladd and Malcolm S Carroll</i>	467
Entanglement and Quantum Information	<i>PG Kwiat and DFV James</i>	478
Applications in Semiconductors	<i>HM van Driel and JE Sipe</i>	486

VOLUME 2

Transient Holographic Grating Techniques in Chemical Dynamics	<i>E Vauthey</i>	1
Atomic Physics	<i>G Kurizki, AG Kofman, and D Petrosyan</i>	12
Terahertz Physics of Semiconductor Heterostructures	<i>Juraj Darmo and Karl Unterrainer</i>	19
Strong-Field Terahertz Excitations in Semiconductors	<i>Ulrich Huttner, Rupert Huber, Mackillo Kira, and Stephan W Koch</i>	33
Rydberg States in Semiconductors	<i>Manfred Bayer and Marc Assmann</i>	40
Two-Dimensional Coherent Spectroscopy of Transition Metal Dichalcogenides	<i>Galan Moody</i>	52
Excitons in Magnetic Fields	<i>Kankan Cong, G Timothy Noe II, and Junichiro Kono</i>	63
Ultrafast Studies of Semiconductors	<i>J Shah</i>	82
Band Structure and Optical Properties	<i>W Zawadzki</i>	87
Excitons	<i>I Galbraith</i>	93
Quantum Wells and GaAs-Based Structures	<i>P Blood</i>	98
Recombination Processes	<i>PT Landsberg</i>	110
Coherent Terahertz Sources	<i>L Wang</i>	118
Using Ultrafast Optical Spectroscopy to Unravel the Properties of Correlated Electron Materials	<i>Rohit P Prasankumar, Dmitry A Yarotski, and Antoinette J Taylor</i>	123
Tutorial on Multidimensional Coherent Spectroscopy	<i>Mark Siemens</i>	150
Two-Dimensional Infrared (2D IR) Spectroscopy	<i>Lauren E Buchanan and Wei Xiong</i>	164
Two-Dimensional Electronic Spectroscopy	<i>Yin Song, Xiaoqin Li, and Jennifer P Ogilvie</i>	184
Multidimensional Terahertz Spectroscopy	<i>Michael Woerner, Klaus Reimann, and Thomas Elsaesser</i>	197
Second Harmonic Generation Spectroscopy of Hidden Phases	<i>Liuyan Zhao, Darius Torchinsky, John Harter, Alberto de la Torre, and David Hsieh</i>	207
Saturated Absorption Spectroscopy for Diode Laser Locking	<i>Bachana Lomsadze</i>	227
Attosecond Spectroscopy	<i>Agnieszka Jaroń-Becker and Andreas Becker</i>	233
Nonlinear Spectroscopies	<i>SR Meech</i>	244
Alternative Plasmonic Materials	<i>Gururaj V Naik</i>	252
Raman Lasers	<i>Marco Santagiustina</i>	265

GaN Lasers	<i>Harumasa Yoshida</i>	271
Optically Pumped Semiconductor Lasers	<i>Jerome V Moloney and Alexandre Laurain</i>	280
Optical Parametric Amplifiers	<i>Giulio Cerullo, Sandro De Silvestri, and Cristian Manzoni</i>	290
Mode-Locked Lasers	<i>Ladan Arissian and Jean-Claude Diels</i>	302
Few-Cycle and Attosecond Lasers	<i>Francesca Calegari and Caterina Vozzi</i>	311
Chirped Pulse Amplification	<i>GA Mourou</i>	324
Carbon Dioxide Laser	<i>CR Chatwin</i>	325
Dye Lasers	<i>FJ Duarte and A Costela</i>	336
Edge Emitters	<i>JJ Coleman</i>	350
Excimer Lasers	<i>JJ Ewing</i>	358
Metal Vapor Lasers	<i>DW Coutts</i>	368
Noble Gas Ion Lasers	<i>WB Bridges</i>	376
Planar Waveguide Lasers	<i>S Bhandarkar</i>	384
Up-Conversion Lasers	<i>A Brenier</i>	394
Thin Disk Lasers	<i>Mikhail Larionov</i>	407
Microchip Lasers	<i>John J Zayhowski</i>	415
Supercontinuum Generation	<i>James R Taylor</i>	424
Infrared Transition Metal Solid-State Lasers	<i>Kenneth L Schepler</i>	435
UV Lasers	<i>Yushi Kaneda</i>	446
Single-Frequency Lasers	<i>Xiushan Zhu</i>	451
Semiconductor Lasers	<i>Stephan W Koch and Martin R Hofmann</i>	462

VOLUME 3

Displays – Introduction	<i>Han-Ping D Shieh</i>	1
Liquid Crystal Physics and Materials	<i>Huang-Ming Philip</i>	8
TFT Materials and Devices	<i>Po-Tsun Lin</i>	12
Color Formation of LCD – Spatial and Temporal	<i>Fang-Cheng Lin</i>	17
LCD Components	<i>Fang-Cheng Lin</i>	25
Projection Displays – LCD/MEMS/LCoS Based	<i>Fleming Chuang</i>	32
3D Displays, Stereoscopic/Autostereoscopic 3D	<i>Yi-Pai Huang and Chun-Ho Chen</i>	44
Wearable Displays	<i>Paul Yang</i>	51
View Angles of Liquid Crystal Displays	<i>Huang-Ming Philip Chen</i>	59
Organic Light-Emitting Diode (OLED)	<i>Chin H (Fred) Chen, Wen-Shi Wen, and Chin-Ti Chen</i>	64
Optical Characteristics of Display Devices	<i>Han-Ping D Shieh</i>	70
Reflective Display Technologies	<i>Han-Ping D Shieh</i>	79
In Situ Optical Tissue Diagnostics/Miniaturized Optoelectronic Sensors for Tissue Diagnostics	<i>Seung Yup Lee, Yooree Grace Chung, and Mary-Ann Mycek</i>	86

In Situ Optical Tissue Diagnostics/Laser Speckle-Based Spectroscopy Techniques in Biomedicine	<i>Karthik Vishwanath and Sara Zanfardino</i>	95
Photodynamic Therapy and Photobiomodulation: Can All Diseases be Treated with Light?	<i>Michael R Hamblin</i>	100
Resolution and Multiple Scattering in Imaging	<i>Edwin A Marengo, Zambrano-Nunez Maytee, and Edson S Galagarza</i>	136
Coherent Diffractive Imaging	<i>Lei Tian</i>	146
Nonuniqueness in Imaging	<i>Greg Gbur</i>	156
Phase Space and Imaging	<i>Markus Testorf</i>	164
Information Theory in Imaging	<i>FO Huck and CL Fales</i>	175
Inverse Problems and Computational Imaging	<i>M Bertero and P Boccacci</i>	187
Adaptive Optics	<i>C Pernechele</i>	197
Phase Conjugation and Image Correction	<i>EN Leith</i>	205
Hyperspectral Imaging	<i>MI Huebschman, RA Schultz, and HR Garner</i>	211
Imaging Through Scattering Media	<i>AC Boccara</i>	219
Infrared Imaging	<i>K Krapels and RG Driggers</i>	229
Interferometric Imaging	<i>DL Marks</i>	241
Multiplex Imaging	<i>A Lacourt</i>	247
Photon Density Wave Imaging	<i>V Toronov</i>	256
Three-Dimensional Field Transformations	<i>R Piestun</i>	262
Volume Holographic Imaging	<i>G Barbastathis</i>	267
Wavefront Sensors and Control (Imaging Through Turbulence)	<i>CL Matson</i>	272

VOLUME 4

Basic Concepts of Optical Communication Systems	<i>S Lee and AE Willner</i>	1
Historical Development of Optical Communication Systems	<i>G Keiser</i>	13
Dispersion Management	<i>AE Willner, Y-W Song, J McGeehan, Z Pan B Hoanca</i>	20
All-Optical Multiplexing/Demultiplexing	<i>Z Ghassemlooy and G Swift</i>	34
Pulse Characterization Techniques	<i>DJ Kane</i>	48
Optical Time Division Multiplexing	<i>LP Barry</i>	60
Lightwave Transmitters	<i>JG McInerney</i>	67
Broadband Passive Optical Access Networks	<i>Elaine Wong, Maluge P Imali Dias, Zhengxuan Li, and Lilin Yi</i>	73
Holographic Recording Media and Devices	<i>Pierre-Alexandre Blanche</i>	87
Colour Holography: Perception Versus Technical Reality	<i>Andrew Pepper</i>	102
High-Resolution Underwater Holographic Imaging	<i>John Watson</i>	106
Digital Holographic Display	<i>Daping Chu, Jia Jia, and Jhensi Chen</i>	113
Holography: Computer Generated Holograms	<i>WJ Dallas and AW Lohmann</i>	130

Module: Digital Holography	Wolfgang Osten	139
Overview: Holography	C. Shakher and AK Ghatak	151
The Fractional Order Fourier Transform and Fresnel Diffraction	Pierre Pellat-Finet and Yezid Torres Moreno	157
Ambiguity Function in Optics	JP Guigay	164
Phase Space Tomography in Optics	Tatiana Alieva, José A Rodrigo, and Antonio Picón	174
Coordinate Transformations and the Hough Transform	Filippus S Roux	182
Single-Pixel Imaging Using the Hadamard Transform	Fernando Soldevila, Pere Clemente, Enrique Tajahuerce, and Jesús Lancis	193
Linear Canonical Transforms	Kurt B Wolf	199
Phase-Space Representations of Freeform Optical Systems	Alois M Herkommer	205
Silicon Photonics; Ring Modulator Transmitters	M Ashkan Seyedi and Marco Fiorentino	216
Optical Switches	Dritan Celo, Dominic J Goodwill, and Eric Bernier	224
Indium Phosphide Photonic Integrated Circuits	Yuliya Akulova	242
CMOS Transceiver Circuits for Optical Interconnects	Samuel Palermo	254
Foundations of Coherent Transients in Semiconductors	Torsten Meier and Stephan W Koch	264
Nonlinear Optics in Disordered Media: Anderson Localization	Arash Mafi	278
Second-Harmonic Generation in Two-Dimensional Materials	Myrta Grüning	284
Parity-Time Symmetry in Optics	Mercedeh Khajavikhan, Mohammad-Ali Miri, Andrea Alu, and Demetrios N Christodoulides	291
Laser-Induced Damage in Optical Materials	Wolfgang Rudolph and Luke A Emmert	302
Four-Wave Mixing	L Canioni and L Sarger	310
Kramers–Krönig Relations in Nonlinear Optics	M Sheik-Bahae	317
Nonlinear Optical Phase Conjugation	BY Zeldovich	322
Photorefraction	M Cronin-Golomb and B Kippelen	327
Ultrafast and Intense-Field Nonlinear Optics	AL Gaeta and RW Boyd	335
Electromagnetically Induced Transparency	JP Marangos	339
Nonlinear Optics, Basics: Nomenclature and Units	MP Hasselbeck	347
Raman Spectroscopy	R Withnall	354
Spatial Heterodyne	Mark F Spencer	369
3D Metrics for Airborne Topographic Lidar	Shea T Hagstrom and Myron Z Brown	401
Multiple Input, Multiple Output, MIMO, Active Electro-Optical Sensing	Paul F McManamon and Jeffrey R Kraczek	407
InGaAs Linear-Mode Avalanche Photodiodes	Andrew S Huntington	415
Very High Range Resolution Lidars	Zeb W Barber	430
Multi-Dimensional Laser Radars	Vasyl Molebny	444
A Review of Laser Range Profiling for Target Recognition	Ove Steimvall	474
Micro-Lidars for Short Range Detection and Measurement	Vasyl V Molebny	496

VOLUME 5

RF Coherent Detection on Top of Direct Detection Lidar	<i>Barry I. Stann, John F Dammann, Mark M Giza, Brian C Redman, and William C Ruff</i>	1
Optical Masks Using Walsh Functions	<i>Lakshminarayan Hazra and Pubali Mukherjee</i>	14
Coherent Control: Theory	<i>H Rabitz</i>	30
Coherent Control: Experimental	<i>RJ Levis</i>	39
Optics of the Human Eye	<i>David A Atchison</i>	43
Measuring Retinal Blood Flow	<i>Alberto de Castro and Stephen A Burns</i>	64
Adaptive Optics Retinal Imaging Techniques and Clinical Applications	<i>Jessica IW Morgan</i>	72
Femtosecond Lasers in Retinal Imaging	<i>Christina Schwarz and Jennifer J Hunter</i>	85
Confocal and Multiphoton Imaging of Cornea	<i>Gopal S Jayabalan and Josef F Bille</i>	97
Refractive Error and Wavefront Sensing	<i>Larry N Thibos</i>	108
Methods of Vision Correction	<i>Len Zheleznyak, Ramkumar Sabesan, and Geunyoung Yoon</i>	116
Laser and Light in Ophthalmology	<i>Michael Mrochen, Nicole Lemanski, and Bojan Pajic</i>	130
Probing Ocular Mechanics With Light	<i>Susana Marcos, Giuliano Scarcelli, Antoine Ramier, and Seok H Yun</i>	140
Optical Coherence Tomography and Its Application to Imaging of Skin and Retina	<i>Michael Pircher</i>	155
TIRF Microscopy and Variants	<i>Daniel Axelrod</i>	168
Super-Resolution Depth Measurements: Variable Angle TIRF, Super-Critical Angle Fluorescence, MIET	<i>Alexey Chizhik</i>	175
Overview: Microscopy	<i>CJR Sheppard</i>	185
Confocal Microscopy	<i>T Wilson</i>	194
Atom Optics	<i>AD Cronin and DE Pritchard</i>	203
Organic Semiconductors	<i>Fang-Chung Chen</i>	220
OLED/Introduction	<i>Dae-Gyu Moon</i>	232
Harvesting Triplet Excitons in OLED	<i>Gufeng He</i>	240
Organic Light-Emitting Diodes/Light Extraction	<i>Jiun-Haw Lee</i>	247
Conjugated Polymer-Based Solar Cells	<i>Gang Li, Widhya Budiawan, Pen-Cheng Wang, and Chih Wei Chu</i>	256
Dye-Sensitized Solar Cells	<i>Lu-Yin Lin and Kuo-Chuan Ho</i>	270
Organic Inorganic Hybrid Perovskite Materials and Devices	<i>Yihua Chen, Ning Zhou, and Huanping Zhou</i>	282
Light Harvesting in Organic Solar cells	<i>Supriya Pillai, Matthew Wright, and Kah H Chan</i>	292
Organic Lasers	<i>Graham A Turnbull</i>	309
Organic Photodetectors	<i>Jaehoon Jeong, Hwajeong Kim, and Youngkyoo Kim</i>	317
Index		331

EDITORS IN CHIEF



Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.



Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Dispersion

L Thévenaz, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

D_λ	Chromatic dispersion [ps nm ⁻¹ km ⁻¹]	V	Normalized frequency
E	Electric field [V m ⁻¹]	ν	Optical frequency [s ⁻¹]
N	Group index	β	Propagation constant [m ⁻¹]
V_g	Group velocity [m s ⁻¹]	τ_D	Propagation delay [s]
D_v	Group Velocity Dispersion GVD [s ² m ⁻¹]	P	Polarization density field [A s m ⁻²]
α	Linear attenuation coefficient [m ⁻¹]	n	Refractive index
χ	Medium susceptibility	c_0	Vacuum light velocity [m s ⁻¹]
		λ	Wavelength [m]

Introduction

Dispersion designates the property of a medium to propagate the different spectral components of a wave with a different velocity. It may originate from the natural dependence of the refractive index of a dense material to the wavelength of light, and one of the most evident and magnificent manifestations of dispersion is the appearance of a rainbow in the sky. In this case dispersion gives rise to a different refraction angle for the different spectral components of the white light, resulting in an angular dispersion of the sunlight spectrum and explicating the etymology of the term dispersion. Despite this spectacular effect, dispersion is mostly seen as an impairment in many applications, in particular for optical signal transmission. In this case, dispersion causes a rearrangement of the signal spectral components in the time domain, resulting in temporal spreading of the information and eventually in severe distortion.

There is a trend to designate all phenomena resulting in a temporal spreading of information as a type of dispersion, namely the polarization dispersion in optical fibers that should be properly named as polarization mode delay (PMD). In this article chromatic dispersion will be solely addressed, that designates the dependence of the propagation velocity on wavelength and that corresponds to the strict etymology of the term.

Effect of Dispersion on an Optical Signal

An optical signal may always be considered as a sum of monochromatic waves through a normal Fourier expansion. Each of these Fourier components propagates with a different phase velocity if the optical medium is dispersive, since the refractive index n depends on the optical frequency ν . This phenomenon is linear and causal and gives rise to a signal distortion that may be properly described by a transfer function in the frequency domain.

Let $n(\nu)$ be the frequency-dependent refractive index of the propagation medium. When the optical wave propagates in a waveguiding structure the effective wavenumber is properly described by a propagation constant β , that reads:

$$\beta(\nu) = \frac{2\pi\nu}{c_0} n(\nu) \quad (1)$$

where c_0 is the vacuum light velocity. The propagation constant corresponds to an eigenvalue of the wave equation in the guiding structure and normally takes a different value for each solution or propagation mode. For free-space propagation this constant is simply equal to the wavenumber of the corresponding plane wave.

The complex electrical field $E(z, t)$ of a signal propagating in the z direction may be properly described by the following expression:

$$\begin{aligned} E(z, t) &= A(z, t) e^{i(2\pi\nu_0 t - \beta_0 z)} \\ \text{with } \nu_0 &: \text{the central optical frequency} \\ \text{and } \beta_0 &= \beta(\nu_0) \end{aligned} \quad (2)$$

where $A(z, t)$ represents the complex envelope of the signal, supposed to be slowly varying compared to the carrier term oscillating with the frequency ν_0 . Consequently, the signal will spread over a narrow band around the central frequency ν_0 and the propagation constant β can be conveniently approximated by a limited expansion to the second order:

$$\beta(\nu) = \beta_0 + \left. \frac{d\beta}{d\nu} \right|_{\nu=\nu_0} (\nu - \nu_0) + \left. \frac{d^2\beta}{d\nu^2} \right|_{\nu=\nu_0} (\nu - \nu_0)^2 \quad (3)$$

For a known signal $A(0, t)$ at the input of the propagation medium, the problem consists in determining the signal envelope $A(z, t)$ after propagation over a distance z . The linearity and the causality of the system make possible a description using a transfer function $H_z(v)$ such as:

$$\tilde{A}(z, v) = H_z(v) \tilde{A}(0, v) \quad (4)$$

where $\tilde{A}(z, v)$ is the Fourier transform of $A(z, t)$.

To make the transfer function $H_z(v)$ explicit, let us assume that the signal corresponds to an arbitrary harmonic function:

$$A(0, t) = A_0 e^{i2\pi f t} \quad (5)$$

Since this function is arbitrary and the signal may always be expanded as a sum of harmonic functions through a Fourier expansion, there is no loss of generality. The envelope identified as a harmonic function actually corresponds to a monochromatic wave of optical frequency $v = v_0 + f$ as defined by Eq. (2). Such a monochromatic wave will experience the following phase shift through propagation:

$$\begin{aligned} E(z, t) &= A_0 e^{i[2\pi(v_0+f)t - \beta(v_0+f)z]} \\ &= A_0 e^{i2\pi f t} e^{i[2\pi v_0 t - \beta(v_0+f)z]} \end{aligned} \quad (6)$$

On the other hand, an equivalent expression may be found using the linear system described by Eqs. (2) and (4):

$$\begin{aligned} E(z, t) &= A(z, t) e^{i(2\pi v_0 t - \beta_0 z)} \\ &= A_z e^{i2\pi f t} e^{i(2\pi v_0 t - \beta_0 z)} \end{aligned} \quad (7)$$

Since Eqs. (6) and (7) must represent the same quantities and using the definition in Eq. (4), a simple comparison shows that the transfer function must take the following form:

$$H_z(v) = e^{-i[\beta(v)z - \beta_0 z]} \quad (8)$$

The transfer function takes a more analytical form using the approximation in Eq. (3):

$$H_z(v) = e^{-i2\pi(v-v_0)\tau_D} e^{-i\pi D_v(v-v_0)^2 z} \quad (9)$$

In the transfer function interpretation the first term represents a delay term. It means that the signal is delayed after propagation by the quantity:

$$\tau_D = \frac{1}{2\pi} \frac{d\beta}{dv_0} z = \frac{z}{V_g} \quad (10)$$

where V_g represents the signal group velocity. This term therefore brings no distortion for the signal and thus states that the signal is replicated at the distance z with a delay τ_D .

The second term is the distortion term which is similar in form to a diffusion process and normally results in time spreading of the signal. In the case of a light pulse it will gradually broaden while propagating along the fiber, like a hot spot on a plate gradually spreading as a result of heat diffusion. The effect of this distortion is proportional to the distance z and to the coefficient D_v , named group velocity dispersion (GVD):

$$D_v = \frac{1}{2\pi} \frac{d^2\beta}{dv^2} = \frac{d}{dv} \left(\frac{1}{V_g} \right) \quad (11)$$

It is important to point out that the GVD may be either positive (normal) or negative (anomalous) and the distortion term in the transfer function in Eq. (9) may be exactly cancelled by propagating in a medium with D_v of opposite sign. It means that the distortion resulting from chromatic dispersion is reversible and this is widely used in optical links through the insertion of dispersion compensators. These are elements made of specially designed fibers or fiber Bragg gratings showing an enhanced GVD coefficient, with a sign opposite to the GVD in the fiber.

From the transfer function in Eq. (9) it is possible to calculate the impulse response of the dispersive medium:

$$h_z(t) = \frac{1}{\sqrt{i|D_v|z}} e^{i\pi \frac{(t-\tau_D)^2}{D_v z}} \quad (12)$$

so that the distortion of the signal may be calculated by a simple convolution of the impulse response with the signal envelope in the time domain.

The effect of dispersion on the signal can be more easily interpreted by evaluating the dispersive propagation of a Gaussian pulse. In this particular case the calculation of the resulting envelope can be carried out analytically. If the signal envelope takes the following Gaussian distribution at the origin:

$$A(0, t) = A_0 e^{-\frac{t^2}{\tau_0^2}} \quad (13)$$

with τ_0 the $1/e$ half-width of the pulse, the envelope at distance z is obtained by convoluting the initial envelope with the impulse response $h_z(t)$:

$$\begin{aligned} A(z, t) &= h_z(t) \otimes A(0, t) \\ &= A_0 \sqrt{\frac{iz_0}{z + iz_0}} e^{i\pi \frac{(t-\tau_D)^2}{D_v(z+iz_0)}} \end{aligned} \quad (14)$$

where

$$z_0 = -\frac{\pi\tau_0^2}{D_v} \quad (15)$$

represents the typical dispersion length, that is the distance necessary to make the dispersion effect noticeable.

The actual pulse spreading resulting from dispersion can be evaluated by calculating the intensity of the envelope at distance z :

$$|A(z, t)|^2 = A_0 \frac{\tau_0}{\tau(z)} e^{-\frac{2(t-\tau_D)^2}{\tau^2(z)}} \quad (16)$$

that is still a Gaussian distribution centered about the propagation delay time τ_D , with $1/e^2$ half-width:

$$\tau(z) = \tau_0 \sqrt{1 + (z/z_0)^2} \quad (17)$$

The variation of the pulse width $\tau(z)$ is presented in [Fig. 1](#) and clearly shows that the pulse spreading starts to be nonnegligible from the distance $z=z_0$. This gives a physical interpretation for the dispersion length z_0 . It must be pointed out that there is a direct formal similarity between the pulse broadening of a Gaussian pulse in a dispersive medium and the spreading of a free-space Gaussian beam as a result of diffraction. Asymptotically for distances $z \gg z_0$, the pulsewidth increases linearly:

$$\tau(z) \simeq |D_v| \frac{z}{\pi\tau_0} \quad (18)$$

It must be pointed out that the width increases proportionally to the dispersion D_v , but also inversely proportionally to the initial width τ_0 . This results from the larger spectral width corresponding to a narrower pulsewidth, giving rise to a stronger dispersive effect.

The chromatic dispersion does not modify the spectrum of the transmitted light, as any linear effect. This can be straightforwardly demonstrated by evaluating the intensity spectrum of the signal envelope at any distance z , using the [Eqs. \(4\)](#) and [\(8\)](#):

$$\begin{aligned} |\tilde{A}(z, \nu)|^2 &= |H_z(\nu) \tilde{A}(0, \nu)|^2 = |e^{-i[\beta(\nu)z - \beta_0 z]}|^2 |\tilde{A}(0, \nu)|^2 \\ &= |\tilde{A}(0, \nu)|^2 \end{aligned} \quad (19)$$

It means that the pulse characteristics in the time and frequency domains are no longer Fourier-transform limited, since after the broadening due to dispersion, the spectrum should normally spread over a narrower spectral width. This feature results from a rearrangement of the spectral components within the pulse. This can be highlighted by evaluating the distribution of instantaneous frequency through the pulse. The instantaneous frequency ω_i is defined as the time-derivative of the wave phase factor $\phi(t)$ and is uniformly equal to the optical carrier pulsation $\omega_i = 2\pi\nu_0$ for the initial pulse, as can be deduced from the phase factor at $z=0$ by combining [Eqs. \(2\)](#) and [\(13\)](#). This constant instantaneous frequency means that all spectral components are uniformly present within the pulse at the origin.

After propagation through the dispersive medium the phase factor $\phi(t)$ can be evaluated by combining [Eqs. \(2\)](#) and [\(14\)](#) and evaluating the argument of the resulting expression. The instantaneous frequency ω_i is obtained after a simple time derivative

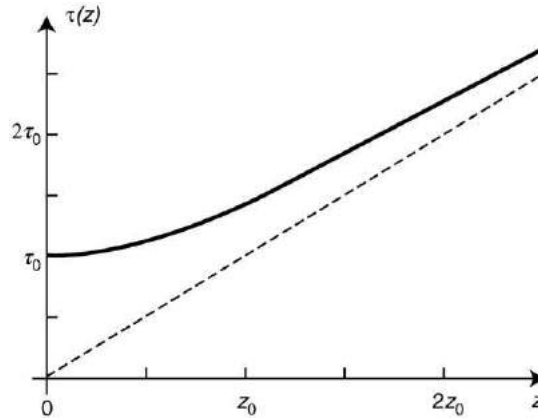


Fig. 1 Variation of the $1/e^2$ half-width of a Gaussian pulse, showing the pulse spreading effect of dispersion. The dashed line shows the asymptotic linear spreading for large propagation distance.

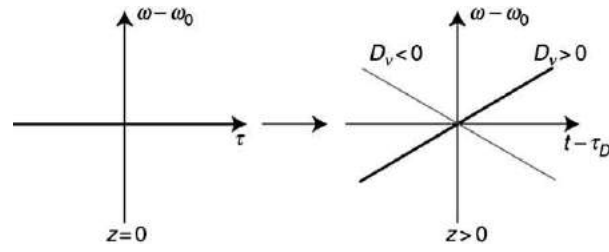


Fig. 2 Distribution of the instantaneous frequency through a Gaussian pulse. At the origin the distribution is uniform (left) and the dispersion induces a frequency chirp that depends on the sign of the GVD coefficient D_v .

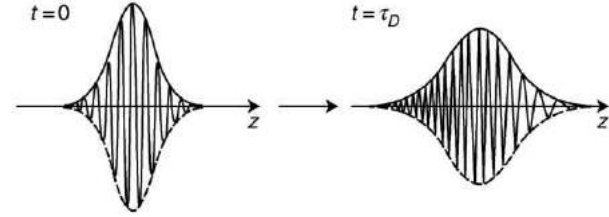


Fig. 3 The dispersion results in a pulse broadening together with a frequency chirp, here for a normal GVD, that can be seen like a frequency modulation.

of $\phi(t)$ and reads:

$$\omega_i(t) = \omega_0 + \frac{2\pi z}{D_v(z^2 + z_0^2)}(t - \tau_D) \quad (20)$$

For $z > 0$ the instantaneous frequency varies linearly over the pulse, giving rise to a frequency chirp. The frequency components are re-arranged in the pulse, so that the lower frequency components are in the leading edge of the pulse for a normal dispersion $D_v > 0$ and the higher frequencies in the trailing edge. For an anomalous dispersion $D_v < 0$, the arrangement is opposite, as can be seen in Fig. 2. The effect of this frequency chirp can be visualized in Fig. 3, showing that the effect of dispersion is equivalent to a frequency modulation over the pulse. The chirp is maximal at position z_0 and the pulsewidth takes its minimal value τ_0 when the chirp is zero. It is evident with this description that the pulse broadening may be entirely compensated through propagation in a medium of opposite group velocity dispersion; this is equivalent to reversing the time direction in Fig. 3. Moreover a pre-chirped pulse can be compressed to its Fourier-transform limited value after propagation in a medium with the proper dispersion sign. This feature is widely used for pulse compression after pre-chirping through propagation in a medium subject to optical Kerr effect.

The above description of the propagation of a Gaussian pulse implicitly states that the light source is perfectly coherent. In the early stages of optical communications it was not at all the case, since most sources were either light emitting diodes or multimode lasers. In this case the spectral extent of the signal was much larger than actually required for the strict need of modulation. Each spectral component may thus be considered as independently propagating the signal and the total optical wave can be identified to light merged from discrete sources emitting simultaneously the same signal at a different optical frequency. The group velocity dispersion will cause a delay $\delta\tau$ between different spectral components separated by a frequency interval $\delta\nu$ that can be simply evaluated by a first-order approximation:

$$\delta\tau = \frac{d\tau_D}{d\nu} \delta\nu = \frac{d}{d\nu} \left(\frac{z}{V_g} \right) \delta\nu = D_v z \delta\nu \quad (21)$$

where Eqs. (10) and (11) have been used. The delay $\delta\tau$ is thus proportional to the GVD coefficient D_v , to the propagation distance z and the frequency separation $\delta\nu$. This description can be extended to a continuous frequency distribution with a spectral width σ_ν , resulting to the following temporal broadening σ_τ :

$$\sigma_\tau = |D_v| \sigma_\nu z \quad (22)$$

Traditionally the spectral characteristics of a source are given in units of wavelength and the GVD coefficient is expressed in optical fibers accordingly. Following the same description as above, the temporal broadening σ_τ , for a spectral width σ_λ in wavelength units, reads:

$$\sigma_\tau = |D_\lambda| \sigma_\lambda z \quad (23)$$

Since equal spectral widths must give equal broadening the value of the GVD in units of wavelength can be deduced from the natural definition in frequency units:

$$D_\lambda = \frac{d}{d\lambda} \left(\frac{1}{V_g} \right) = \frac{d}{d\nu} \left(\frac{1}{V_g} \right) \frac{d\nu}{d\lambda} = -\frac{c_0}{\lambda^2} D_v \quad (24)$$

It must be pointed out that the coefficient D_λ takes a sign opposite to D_ν ; in other words, a normal dispersion corresponds to a negative GVD coefficient D_λ . It is usually expressed in units of picoseconds of temporal broadening, per nanometer of spectral width and per kilometer of propagating distance, or ps/nm km. For example, a pulse showing a spectral width of 1 nm propagating through a 100 km fiber having a dispersion D_λ of +10 ps/nm km, will experience, according to Eq. (23), a pulse broadening σ_τ of $10 \times 1 \times 100 = 1000$ ps or 1 ns.

Material Group Velocity Dispersion

Any dense material shows a variation of its index of refraction n as a function of the optical frequency ν . This natural property is called material dispersion and is the dominant contribution in weakly guiding structures such as standard optical fibers. This natural dependence results from the noninstantaneous response of the medium to the presence of the electric field of the optical wave. In other words, the polarization field $P(t)$ corresponding to the material response will vary with some delay or inertia to the change of the incident electric field $E(t)$. This delay between cause and effect generates a memory-type response of the medium that may be described using a time-dependent medium susceptibility $\chi(t)$. The relation between medium polarization at time t and incident field results from the weighted superposition of the effects of $E(t')$ at all previous times $t' < t$. This takes the form of the following convolution:

$$P(t) = \epsilon_0 \int_{-\infty}^{+\infty} \chi(t - t') E(t') dt' \quad (25)$$

Through application of a simple Fourier transform this relation reads in the frequency domain as:

$$P(\nu) = \epsilon_0 \chi(\nu) E(\nu) \quad (26)$$

showing clearly that the noninstantaneous response of the medium results in a frequency-dependent refractive index using the standard relationship with the susceptibility χ :

$$n(\nu) = \sqrt{1 + \chi(\nu)} \quad \text{where} \quad \chi(\nu) = \text{FT}\{\chi(t)\} \quad (27)$$

This means that a beautiful natural phenomenon such as a rainbow, originates on a microscopic scale from the sluggishness of the medium molecules to react to the presence of light. For signal propagation, it results in a distortion of the signal and in most cases in a pulse spreading, but the microscopic causes are in essence identical. This noninstantaneous response is tightly related to the molecules' vibrations that also give rise to light absorption. For this reason it is convenient to express the propagation in an absorptive medium by adding an imaginary part to the susceptibility $\chi(\nu)$:

$$\chi(\nu) = \chi'(\nu) + i\chi''(\nu) \quad (28)$$

so that the refractive index $n(\nu)$ and the absorption coefficient $\alpha(\nu)$ reads in a weakly absorbing medium:

$$\begin{aligned} n(\nu) &= \sqrt{1 + \chi'(\nu)} \\ \alpha(\nu) &= -\frac{2\pi\chi''(\nu)}{\lambda n(\nu)} \end{aligned} \quad (29)$$

Since the response of the medium, given by the time-dependent susceptibility $\chi(t)$ in Eq. (25), is real and causal, the real and imaginary part of the susceptibility in Eq. (28) are not entirely independent and are related by the famous Kramers–Kronig relations:

$$\begin{aligned} \chi'(\nu) &= \frac{2}{\pi} \int_0^\infty \frac{s\chi''(s)}{s^2 - \nu^2} ds \\ \chi''(\nu) &= \frac{2}{\pi} \int_0^\infty \frac{\nu\chi'(s)}{\nu^2 - s^2} ds \end{aligned} \quad (30)$$

Absorption and dispersion act in an interdependent way on the propagating optical wave and knowing either the absorption or the dispersion spectrum is theoretically sufficient to determine the entire optical response of the medium. The interdependence between absorption and dispersion is typically illustrated in Fig. 4. The natural tendency is to observe a growing refractive index for increasing frequencies in low absorption regions. In this case, the dispersion is called normal and this is the most observed situation in transparency regions of a material, which obviously offer the largest interest for optical propagation. In an absorption line the tendency is opposite, a diminishing index for increasing wavelength, and such a response is called anomalous dispersion. The dispersion considered here is the phase velocity dispersion, represented by the slope of $n(\nu)$, that must not be mistaken with the group velocity dispersion (GVD) that only matters for signal distortion. The difference between these two quantities is clarified below.

To demonstrate that a frequency-dependent refractive index $n(\nu)$ gives rise to a group velocity dispersion and thus a signal distortion we use Eqs. (1) and (10), so the group velocity V_g can be expressed:

$$V_g = \frac{c_0}{N} \quad \text{with} \quad N = n + \nu \frac{dn}{d\nu} = n - \lambda \frac{dn}{d\lambda} \quad (31)$$

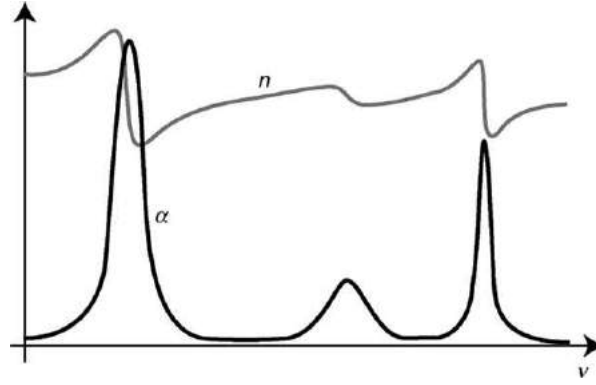


Fig. 4 Typical optical response of a transparent medium, showing the spectral interdependence between the absorption coefficient α and the index of refraction n .

N is called group velocity index and differs from the phase refractive index n only if n shows a spectral dependence. In a region of normal dispersion ($dn/d\nu > 0$), the group index is larger than the phase index and this is the situation observed in the great majority of transparent materials. From Eq. (31) and using Eq. (24) the GVD coefficient D_λ can be expressed as a function of the refractive index $n(\lambda)$:

$$D_\lambda = \frac{d}{d\lambda} \left(\frac{1}{V_g} \right) = \frac{d}{d\lambda} \left(\frac{N}{c_0} \right) = - \frac{\lambda}{c_0} \frac{d^2 n}{d\lambda^2} \quad (32)$$

The GVD coefficient is proportional to the second derivative of the refractive index n with respect to wavelength and is therefore minimal close to a point of inflexion of $n(\lambda)$. As can be seen in Fig. 4 such a point of inflexion is always present where absorption is minimal, at the largest spectral distance from two absorption lines. It means that the conditions of low-group velocity dispersion and high transparency are normally fulfilled in the same spectral region of an optical dielectric material. In this region the phase velocity dispersion is normal at any wavelength, but the group velocity is first normal for shorter wavelengths, then is zero at a definite wavelength corresponding to the point of inflexion of $n(\lambda)$, and finally becomes anomalous for longer wavelengths. The situation in which normal phase dispersion and anomalous group dispersion are observed simultaneously is in no way exceptional.

In pure silica the zero GVD wavelength is at 1273 nm, but is subject to be moderately shifted to larger wavelengths in optical fibers, as a result of the presence of doping species to raise the index in the fiber guiding core. This shift normally never exceeds 10 nm using standard dopings; larger shifts are observed resulting from waveguide dispersion and this aspect will be addressed in the next section. The zero GVD wavelength does not strictly correspond to the minimum attenuation in silica fibers, because the dominant source of loss is Rayleigh scattering in this spectral region and not molecular absorption. This scattering results from fluctuations of the medium density as observed in any amorphous materials such as vitreous silica and is therefore a collective effect of many molecules that does not impinge on the microscopic susceptibility $\chi(t)$. It has therefore no influence on the material dispersion characteristics and this explains the reason for the minimal attenuation wavelength at 1550 nm, mismatching and quite distant from the zero material GVD at 1273 nm.

Material GVD in amorphous SiO_2 can be accurately described using a three-term Sellmeier expansion of the refractive index:

$$n(\lambda) = \sqrt{1 + \sum_{j=1}^3 \frac{C_j \lambda^2}{\lambda^2 - \lambda_j^2}} \quad (33)$$

and performing twice the wavelength derivative. The coefficients C_j and λ_j are found in most reference handbooks and result in the GVD spectrum shown in Fig. 5. Such a spectrum explains the absence of interest for propagation in the visible region through optical fibers, the dispersion being very important in this spectral region. It also explains the large development of optical fibers in the 1300 nm region as a consequence of the minimal material dispersion there. It must be pointed out that it is quite easy to set up propagation in an anomalous GVD regime in optical fibers, since this regime is observed in the lowest attenuation spectral region. Anomalous dispersion makes possible interesting propagation features when combined with a third-order nonlinearity as the optical Kerr effect, namely soliton propagation and efficient spectral broadening through modulation instability.

Waveguide Group Velocity Dispersion

Solutions of the wave equation in an optical dielectric waveguide such as an optical fiber are discrete and limited. These solutions, called modes, are characterized by an unchanged field distribution along the waveguides and by a uniform propagation constant β over the wavefront. This last feature is particularly important if one recall that the field extends over regions presenting different

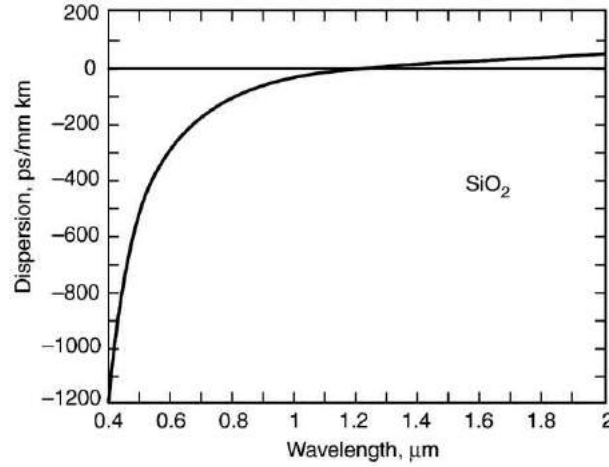


Fig. 5 Material dispersion of pure silica. The visible region (0.4–0.7 μm) shows a strong normal GVD that decreases when moving into the infrared and eventually vanishes at 1273 nm. In the minimum attenuation window (1550 nm) the material GVD is anomalous.

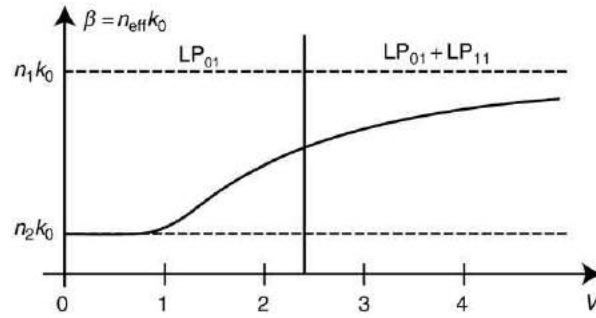


Fig. 6 Propagation constant β as a function of the normalized frequency V in a step-index optical fiber. The single mode region is in the range $0 < V < 2.405$.

refractive indices in a dielectric waveguide. For a given mode, the propagation constant β defines an effective refractive index n_{eff} for the propagation by similarity to [Eq. \(1\)](#):

$$\beta = \frac{2\pi\nu}{c_0} n_{\text{eff}} \quad (34)$$

The value of this effective refractive index n_{eff} is always bound by the value of the core index n_1 and of the cladding index n_2 , so that $n_2 < n_{\text{eff}} < n_1$. For a given mode, the propagation constant β , and so the effective refractive index n_{eff} , only depend on a quantity called normalized frequency V that essentially scales the light frequency to the waveguide optical parameters:

$$V = \frac{2\pi}{\lambda} a \sqrt{n_1^2 - n_2^2} \quad (35)$$

where a is the core radius. [Fig. 6](#) shows the dependence of the propagation constant β of the fundamental mode LP_{01} on the normalized frequency V . The variation is nonnegligible in the single-mode region and gives rise to a chromatic dispersion, since V depends on the wavelength λ , even in the fictitious case of dispersion-free refractive indices in the core and the cladding materials. This type of chromatic dispersion is called waveguide dispersion.

To find an expression for the waveguide dispersion in function of the guiding properties, let us define another normalized parameter, the normalized phase constant b , such as:

$$\beta = \frac{2\pi}{\lambda} \sqrt{n_2^2 + b(n_1^2 - n_2^2)} \quad (36)$$

The parameter b takes values in the interval $0 < b < 1$, is equal to 0 when $n_{\text{eff}} = n_2$ at the mode cutoff, and is equal to 1 when $n_{\text{eff}} = n_1$. This latter situation is never observed and is only asymptotic for very large normalized frequencies V . Solving the wave equation provides the dispersion relation $b(V)$ and it is important to point out that this relation between normalized quantities depends only on the shape of the refractive index profile. Step-index, triangular or multiple-clad index profiles will result in different $b(V)$ relations, independently of the actual values of the refractive indices n_1 and n_2 and of the core radius a .

From the definitions in Eqs. (10) and (35) and in the fictitious case of an absence of material dispersion, the propagation delay per unit length reads:

$$\frac{1}{V_g} = \frac{1}{2\pi} \frac{d\beta}{dv} = \frac{1}{2\pi} \frac{d\beta}{dV} \frac{dV}{dv} = \frac{\lambda}{2\pi} V \frac{d\beta}{dV} \quad (37)$$

so that the waveguide group velocity dispersion can be expressed using the relation in Eq. (24):

$$\begin{aligned} D_\lambda^w &= \frac{d}{d\lambda} \left(\frac{1}{V_g} \right) = \frac{d}{dv} \left(\frac{1}{V_g} \right) \frac{dv}{d\lambda} \\ &= -\frac{v}{\lambda} \frac{d}{dv} \left(\frac{1}{V_g} \right) = -\frac{1}{2\pi c_0} V^2 \frac{d^2 \beta}{dV^2} \end{aligned} \quad (38)$$

The calculation of the combined effect of material and waveguide dispersions results in very long expressions in which it is difficult to highlight the relative effect of each contribution. Nevertheless, by making the assumption of weak guidance:

$$\frac{n_1 - n_2}{n_2} \simeq \frac{N_1 - N_2}{N_2} \ll 1 \quad (39)$$

where N_1 and N_2 are the group indices in the core and the cladding, respectively, defined in Eq. (31), the complete expression can be drastically simplified to obtain for the delay per unit length:

$$\frac{1}{V_g} = \frac{1}{c_0} \left[N_2 + (N_1 - N_2) \frac{d(bV)}{dV} \right] \quad (40)$$

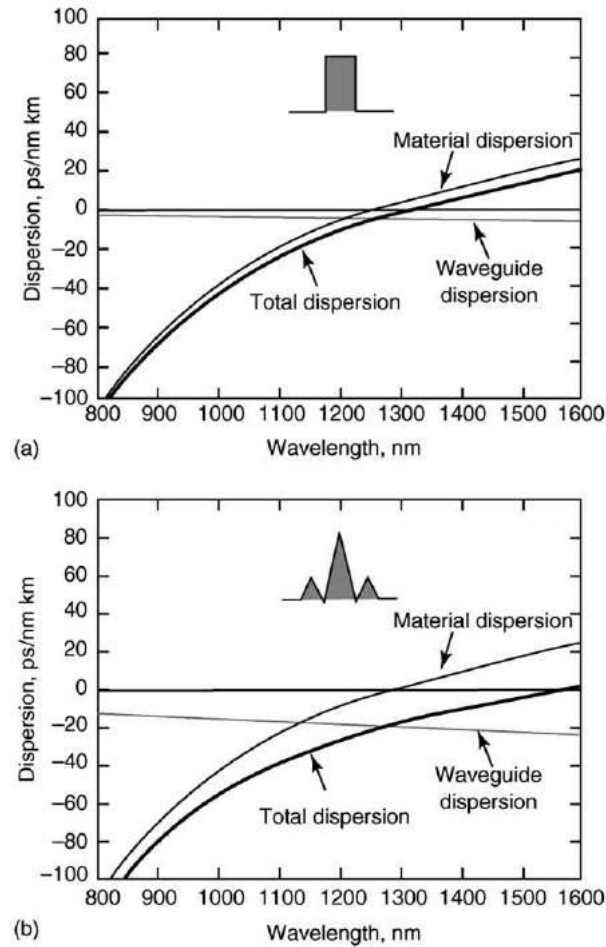


Fig. 7 Material, waveguide and total group velocity dispersions for: (a) step-index fiber; (b) triangular profile fiber. The waveguide dispersion can be significantly enhanced by changing the shape of the index profile, making possible a shift of the zero GVD to the minimum attenuation window.

and for the total dispersion:

$$D_{\lambda} = D_2 + (D_1 - D_2) \frac{d(bV)}{dV} - (N_1 - N_2) \frac{N_2}{n_1} \frac{1}{\lambda c_0} V \frac{d^2(bV)}{dV^2} \quad (41)$$

where

$$D_1 = -\frac{\lambda}{c_0} \frac{d^2 n_1}{d\lambda^2} \quad \text{and} \quad D_2 = -\frac{\lambda}{c_0} \frac{d^2 n_2}{d\lambda^2} \quad (42)$$

are the material GVD in the core and the cladding, respectively.

The first two terms in Eq. (41) represent the contribution of material dispersion weighted by the relative importance of core and cladding materials for the propagating mode. In optical fibers, the difference between D_1 and D_2 is small, so that this contribution can be often well approximated by D_2 , independently of any guiding effects.

The last term represents the waveguide dispersion and is scaled by 2 factors:

- The core-cladding index difference $n_1 - n_2 \simeq N_1 - N_2$. The waveguide dispersion will be significantly enhanced by increasing the index difference between core and cladding.
- The shape factor $V(d^2(bV)/dV^2)$. This factor uniquely depends on the shape of the refractive index profile and may substantially modify the spectral dependence of the waveguide dispersion, making possible a great variety of dispersion characteristics.

Due to this degree of freedom brought by waveguide dispersion it is possible to shift the zero GVD wavelength to the region of minimal attenuation at 1550 nm in silica optical fibers. Fig. 7(a) shows the total group velocity dispersion of a step-index core fiber, together with the separate contributions of material and waveguide GVD. In this case, the material GVD is clearly the dominating contribution, while the small waveguide GVD results in a shift of the zero GVD wavelength from 1273 nm to 1310 nm. A larger shift could be obtained by increasing the core-cladding index difference, but this also gives rise to an increased attenuation from the doping and no real benefit can be expected from such a modification. In Fig. 7(b), the core shows a triangular index profile to enhance the shape factor in Eq. (41), so that the contribution of waveguide GVD is significantly increased with no impairing attenuation due to excessive doping. This makes it possible to realize the ideal situation of an optical fiber showing a zero GVD at the wavelength of minimum attenuation. These dispersion-shifted fibers (DSF) have now become successful in modern telecommunication networks.

Nevertheless, the absence of dispersion favors the efficiency of nonlinear effects and several classes of fibers are now proposed, showing a small but nonzero GVD at 1550 nm, with positive or negative sign, nonzero DSF (NZDSF). By interleaving fibers with positive and negative GVDs it is possible to propagate along the optical link in a dispersive medium and thus minimize the impact of nonlinearities, while maintaining the overall GVD of the link close to zero and canceling any pulse spreading accordingly.

See also: Dispersion Management. Guided Wave Optics

Further Reading

- Ainslie, B.J., Day, C.R., 1986. A review of single-mode fibers with modified dispersion characteristics. *IEEE Journal of Lightwave Technology* LT-4 (8), 967–979.
- Bass M (ed.) (1994) *Handbook of Optics*, sponsored by the Optical Society of America, 2nd edn, vol. II, chap. 10 & 33. New York: McGraw-Hill.
- Buck JA (1995) In: Goodman JW (ed.) *Fundamentals of Optical Fibers*, chap. 1, 5, 6. New York: Wiley Series in Pure and Applied Optics.
- Gloge, D., 1971. Dispersion in weakly guiding fibers. *Applied Optics* 10, 2442–2445.
- Jeunhomme, L.B., 1989. *Single Mode Fiber Optics: Principles and Applications*, Optical Engineering Series, No. 23., New York: Marcel Dekker.
- Marcuse, D., 1974. *Theory of Dielectric Optical Waveguides*, Series on Quantum Electronics – Principles and Applications., New York: Academic Press, chap. 2.
- Marcuse, D., 1980. Pulse distortion in single-mode fibers. *Applied Optics* 19, 1653–1660.
- Marcuse, D., 1981. Pulse distortion in single-mode fibers. *Applied Optics* 20, 2962–2974.
- Marcuse, D., 1981. Pulse distortion in single-mode fibers. *Applied Optics* 20, 3573–3579.
- Mazurin, O.V., Streltsina, M.V., Shvaiko-Shvaikovskaya, T.P., 1983. *Silica Glass and Binary Silicate Glasses (Handbook of Glass Data; Part A)*, Series on Physical Sciences Data, vol. 15. Amsterdam: Elsevier.
- Murata, H., 1988. *Handbook of Optical Fibers and Cables*, Optical Engineering Series, No. 15., New York: Marcel Dekker, chap. 2.
- Saleh BEA and Teich MC (1991) In: Goodman JW (ed.) *Fundamentals of Photonics*, chap. 5, 8 & 22. New York: Wiley Series in Pure and Applied Optics.

Nonlinear Optics

K Thyagarajan and AK Ghatak, Indian Institute of Technology, New Delhi, India

© 2018 Elsevier Inc. All rights reserved.

Glossary

Attenuation The decrease in optical power of a propagating mode usually expressed in dB/km

Birefringent medium A medium characterized by two different modes of propagation characterized by two different polarization states of a plane light wave along any direction

Cross-phase modulation (XPM) The phase modulation of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by another propagating light wave

Effective length (L_{eff}) The length over which most of the nonlinear effect is accumulated in a propagating light beam

Four-wave mixing (FWM) The mixing of four propagating light waves due to intensity dependent refractive index of the medium. This mixing leads to the generation of new frequencies from a set of two or three input light beams

Mode effective area (A_{eff}) The cross sectional area of the mode which determines the nonlinear effects on the propagating mode. This is different from the core area

Nonlinear polarization (P_{NL}) When the response of the medium to an applied electric field is not proportional to the applied electric field, then this results in a nonlinear polarization

Optical waveguides Devices which are capable of guiding optical beams by overcoming diffraction effects using the phenomenon of total internal reflection

Phase coherence length The minimum crystal length upto which the second harmonic power generated by the nonlinear interaction increases

Phase matching Condition under which the nonlinear polarization generating an electromagnetic wave propagates at the same phase velocity as the generated electromagnetic wave

Pulse dispersion The broadening of a light pulse as it propagates in an optical fiber

Quasi-phase matching (QPM) Periodic modulation of the nonlinear coefficient to achieve efficient nonlinear interaction

Second-harmonic generation (SHG) The generation of a wave of frequency 2ω from an incident wave of frequency ω through nonlinear interaction in a medium

Self-phase modulation (SPM) The modulation of the phase of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by itself

Supercontinuum generation (SC) The phenomenon in which a nearly continuous spectrally broadband output is produced through nonlinear effects on high peak power picosecond and sub-picosecond pulses

Walk off length (L_{wo}) Length of the fiber required for two interacting pulses at different wavelengths to walk off relative to each other

Introduction

The invention of the laser provided us with a light source capable of generating extremely large optical power densities (several MW/m^2). At such large power densities, matter behaves in a nonlinear fashion and we come across new optical phenomena such as second-harmonic generation (SHG), sum and difference frequency generation, intensity dependent refractive index, mixing of various frequencies, etc. In SHG, an incident light beam at frequency ω interacts with the medium and generates a new light wave at frequency 2ω . In sum and difference frequency generation, two incident beams at frequencies ω_1 and ω_2 mix with each other producing sum ($\omega_1 + \omega_2$) and difference ($\omega_1 - \omega_2$) frequencies at the output. Higher-order nonlinear effects, such as self-phase modulation, four-wave mixing, etc. can also be routinely observed today. The field of nonlinear optics dealing with such nonlinear interactions is gaining importance due to numerous demonstrated applications in many diverse areas such as optical fiber communications, all-optical signal processing, realization of novel sources of optical radiation, etc.

Nonlinear optical interactions become prominent when the optical power densities are high and interaction takes place over long lengths. The usual method to increase optical intensity is to focus the light beam using a lens system. For a given optical power, the tighter the focusing, the larger will be the intensity for a given optical power; however, greater will be the divergence of the beam. Thus tighter focusing produces larger intensities, but over shorter interaction lengths (see Fig. 1(a)). Optical waveguides, in which the light beam is confined to a small cross-sectional area, are currently being explored for realizing efficient nonlinear devices. In contrast to bulk media, in waveguides, diffraction effects are balanced by waveguiding and the beam can have small cross-sectional areas over much longer interaction lengths (see Fig. 1(b)). A simple optical waveguide consists of a high index dielectric medium surrounded by a lower index dielectric medium so that light waves can be trapped in the high index region by the phenomenon of total internal reflection. Fig. 2 shows a planar waveguide, a channel waveguide and an optical fiber. In the planar waveguide a film of refractive index n_f is deposited/diffused on a substrate of refractive index n_s and has a cover of refractive index n_c (with $n_s, n_c < n_f$). The waveguide has typical cross-sectional dimensions of a few micrometers. Unlike planar waveguides,

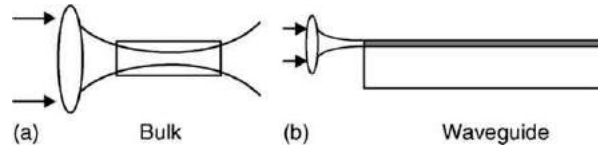


Fig. 1 (a) In bulk media, tighter focusing produces larger intensities, but over shorter interaction lengths. (b) In optical waveguides diffraction effects are balanced by waveguiding and the interaction lengths are much larger.

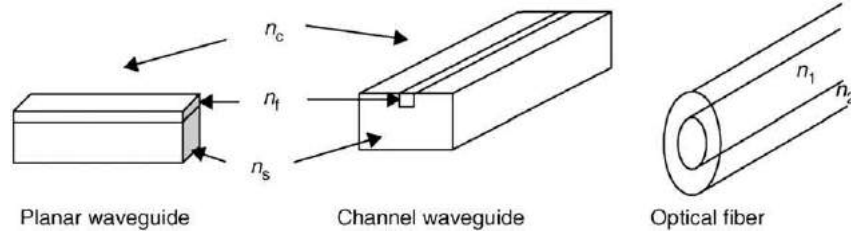


Fig. 2 A planar waveguide, a channel waveguide and an optical fiber.

in which guidance takes place only in one dimension, in channel waveguides the diffused region in a substrate is surrounded on all sides by lower index media. Optical fibers are structures with cylindrical symmetry and have a central cylindrical core of doped silica surrounded by a concentric cylindrical cladding of pure silica which has a slightly lower refractive index. Light guidance in all these waveguides takes place through the phenomenon of total internal reflection.

In contrast to bulk interactions requiring beam focusing, in the case of optical waveguides, the beam can be made to have extremely small cross-sectional areas ($\sim 25 \mu\text{m}^2$ over very long interaction lengths ($\sim 20 \text{ mm}$ in integrated optical waveguides and tens of thousands of kilometers in the case of optical fibers). This leads to very much increased nonlinear interaction efficiencies even at moderate powers (approximately a few tens of mW).

In the following sections, we will discuss some of the important nonlinear interactions that are being studied with potential applications to various branches of science and engineering. Obviously it is impossible to cover all aspects of nonlinear optics – several books have been written in this area (see Further Reading section at the end of this article) – what we will do is to discuss the physics of some of the important nonlinear effects.

Apart from intensity and length of interaction, one of the most important requirements for many efficient nonlinear interactions is the requirement of phase matching. The concept of phase matching can be easily understood from the point of view of SHG. In SHG, the incident wave at frequency ω generates a nonlinear polarization at frequency 2ω and this nonlinear polarization is responsible for the generation of the wave at 2ω . Now, the phase velocity of the nonlinear polarization wave at 2ω is the same as the phase velocity of the electromagnetic wave at frequency ω , which is usually different from the phase velocity of the electromagnetic wave at frequency 2ω ; this happens due to wavelength dispersion in the medium. When the two phase velocities are unequal, the polarization wave at 2ω (which is the source) and the electromagnetic wave at 2ω pass in and out of phase with each other as they propagate through the medium. Due to this, the energy flowing in from ω to 2ω cannot add constructively and the efficiency of second-harmonic generation is limited. If the phase velocities of the waves at ω and 2ω are matched then the polarization wave and the wave at 2ω remain in phase leading to drastically increased efficiencies. This condition is referred to as phase matching and plays a very important role in most nonlinear interactions.

Nonlinear Polarization

In a linear medium, the electric polarization P is assumed to be a linear function of the electric field E :

$$P = \epsilon_0 \chi E \quad (1)$$

where, for simplicity, a scalar relation has been written. The quantity χ is termed as linear dielectric susceptibility. At high optical intensities (which corresponds to high electric fields), all media behave in a nonlinear fashion. Thus Eq. (1) is modified to

$$P = \epsilon_0 (\chi E + \chi^{(2)} E^2 + \chi^{(3)} E^3 + \dots) \quad (2)$$

where $\chi^{(2)}, \chi^{(3)}, \dots$ are higher order susceptibilities giving rise to the nonlinear terms. The second term on the right-hand side is responsible for SHG, sum and difference frequency generation, parametric interactions, etc. while the third term is responsible for third-harmonic generation, intensity dependent refractive index, self-phase modulation, four-wave mixing, etc. For media possessing an inversion symmetry, $\chi^{(2)}$ is zero and there is no second-order nonlinear effect. Thus silica optical fibers, which form the heart of today's communication networks, do not possess the second-order nonlinearity.

We will first discuss second-harmonic generation and parametric amplification which arises due to the second-order nonlinear term and then go on to effects due to third-order nonlinear interaction.

Second-Harmonic Generation in Crystals

The first demonstration of SHG was made in 1961 by focusing a 3 kW ruby laser pulse ($\lambda_0 = 6943 \text{ \AA}$) on a quartz crystal. An incident beam from a ruby laser (red color) after passing through a crystal of KDP, gets partly converted into blue light which is the second harmonic. Ever since, SHG has been one of the most widely studied nonlinear interactions.

We consider a plane wave of frequency ω propagating along the z -direction through a medium and consider the generation of the second-harmonic frequency 2ω as the beam propagates through the medium. Now, the field at ω generates a polarization at 2ω , which acts as a source for the generation of an electromagnetic wave at 2ω . Corresponding to the frequencies ω and 2ω , the electric fields are assumed to be given by

$$E^{(\omega)} = \frac{1}{2} (E_1(z) e^{i(\omega t - k_1 z)} + \text{c.c.}) \quad (3)$$

and

$$E^{(2\omega)} = \frac{1}{2} (E_2(z) e^{i(2\omega t - k_2 z)} + \text{c.c.}) \quad (4)$$

respectively; c.c. stands for the complex conjugate of the preceding quantities. The quantities:

$$k_1 = \omega \sqrt{(\epsilon_1 \mu_0)} = (\omega/c) n^\omega \quad (5)$$

and

$$k_2 = 2\omega \sqrt{(\epsilon_2 \mu_0)} = (2\omega/c) n^{2\omega} \quad (6)$$

represent the propagation vectors at ω and 2ω , respectively; ϵ_1 and ϵ_2 represent the dielectric permittivities at ω and 2ω , and n^ω and $n^{2\omega}$ represent the corresponding wave refractive indices. It should be noted that the amplitudes E_1 and E_2 are assumed to be z dependent – this is because at $z=0$ (where the beam is incident on the medium) the amplitude E_2 is zero and this would develop as the beam propagates through the medium. We will now develop an approximate theory for the generation of the second harmonic.

We start with the wave equation:

$$\nabla^2 E - \epsilon \mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P_{\text{NL}}}{\partial t^2} \quad (7)$$

where P_{NL} is the nonlinear polarization. An incident wave at frequency ω generates a polarization at 2ω which acts as a source for the generation of an electromagnetic wave at 2ω .

In order to consider SHG, we write the wave equation corresponding to 2ω with the nonlinear polarization at 2ω given by

$$P_{\text{NL}}^{(2\omega)} = \frac{1}{2} (\tilde{P}_{\text{NL}}^{(2\omega)} e^{2i(\omega t - k_1 z)} + \text{c.c.}) \quad (8)$$

where

$$\tilde{P}_{\text{NL}}^{(2\omega)} = d E_1 E_1 \quad (9)$$

and

$$d = \frac{\epsilon_0 \chi^{(2)}}{2} \quad (10)$$

represents the effective nonlinear coefficient and depends on the nonlinear material, the polarization states of the fundamental and the second harmonic and also on the propagation direction. Simple manipulations give us, for the rate of change of amplitude of the second harmonic wave:

$$\frac{dE_2}{dz} = - \frac{i\mu_0 d c \omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \quad (11)$$

where

$$\Delta k = k_2 - 2k_1 = (2\omega/c)(n^{2\omega} - n^\omega) \quad (12)$$

and we have assumed:

$$d^2 E_2 / dz^2 \ll k_2 (dE_2 / dz) \quad (13)$$

In order to solve Eq. (11) we neglect depletion of the fundamental field, i.e., $E_1(z)$ is almost a constant and the quantity E_1^2 on the right-hand side can be assumed to be independent of z . If we now integrate Eq. (11), we obtain:

$$E_2(z) = - \frac{i\mu_0 d c \omega}{n^{2\omega}} z E_1^2 e^{i\sigma} \frac{\sin \sigma}{\sigma} \quad (14)$$

where

$$\sigma = \frac{1}{2}(\Delta k)z = (\omega/c)(n^{2\omega} - n^\omega)z \quad (15)$$

Now, the powers associated with the beams corresponding to ω and 2ω are given by:

$$P_1 = \frac{n^\omega}{2c\mu_0} S |E_1|^2, \quad P_2 = \frac{n^{2\omega}}{2c\mu_0} S |E_2|^2 \quad (16)$$

where S represents the cross-sectional area of the beams. Substituting for $|E_2|^2$ from Eq. (16), we get after some elementary simplifications:

$$\eta = \frac{P_2}{P_1} = \frac{2c^2\mu_0^3 d^2 \omega^2}{(n^\omega)^2 n^{2\omega}} z^2 \frac{P_1}{S} \left(\frac{\sin \sigma}{\sigma} \right)^2 \quad (17)$$

when η represents the SHG efficiency. Note that the efficiency of SHG increases if P_1/S the intensity of the fundamental wave increases. Also η increases if the nonlinear coefficient d increases (obviously) and if the frequency increases. However, for a given power P_1 , the conversion efficiency increases if the area of the beam decreases – thus a focused beam will have a greater SHG efficiency. The most important factor is $(\sin \sigma / \sigma)^2$ which is a sharply peaked function around $\sigma = 0$, attaining a maximum value of unity at $\sigma = 0$. Thus for maximum SHG efficiency:

$$\sigma = 0 \Rightarrow n^{2\omega} = n^\omega \quad (18)$$

i.e., the refractive index at 2ω must be equal to the refractive index at ω – this is known as the phase matching condition.

The phase matching condition can be pictorially represented by a vector diagram (see Fig. 3). In Fig. 3, $k_1\omega$ and $k_2\omega$ represent the wave vectors of the fundamental and the second harmonic respectively. To achieve reasonably effective SHG, it is very important to satisfy the phase-matching condition. Efficiencies under nonphase matched operation can be orders of magnitude lower than under phase-matched operation.

We see from Eq. (17) that the smallest z for which η is maximum is:

$$z = L_c = \frac{\pi}{\Delta k} = \frac{\pi c}{2\omega(n^{2\omega} - n^\omega)} \quad (19)$$

where we have used Eq. (15) for Δk . The length L_c is called the phase coherence length and represents the maximum crystal length up to which the second-harmonic power increases. Thus, if the length of the crystal is less than L_c , the second-harmonic power increases almost quadratically with z . For $z > L_c$, the second-harmonic power begins to reduce again.

In general, because of dispersion, it is very difficult to satisfy the phase-matching condition. However, in a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the o -wave for ω equals the refractive index of the e -wave for 2ω (see Fig. 4). For media with $n_e > n_o$, the direction would correspond to that along which the refractive index of the e -wave for ω equals the refractive index for the o -wave for 2ω . This can readily be understood by

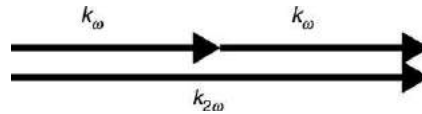


Fig. 3 Vector diagram representing the phase matching condition for SHG.

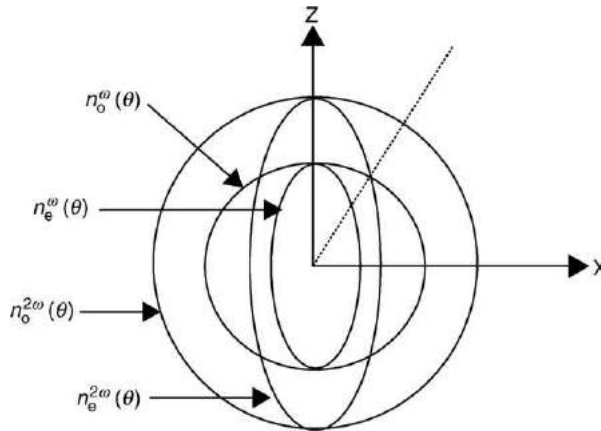


Fig. 4 In a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the o -wave for ω equals the refractive index of the e -wave for 2ω .

considering a specific example. We consider the SHG in KDP corresponding to the incident ruby laser wavelength ($\lambda_0 = 0.6943 \mu\text{m}$, $\omega = 2.7150 \times 10^{15} \text{ Hz}$). For this frequency:

$$\begin{cases} n_o^\omega = 1.50502, & n_e^\omega = 1.46532 \\ n_o^{2\omega} = 1.53269, & n_e^{2\omega} = 1.48711 \end{cases} \quad (20)$$

The refractive index variation for the e -wave is given by:

$$n_e(\theta) = \left(\frac{\sin^2 \theta}{n_e^2} + \frac{\cos^2 \theta}{n_o^2} \right)^{-\frac{1}{2}} \quad (21)$$

where θ is the angle that the wave propagation direction makes with the optic axis. Now, as can be seen from the above equations:

$$n_e < n_e(\theta) < n_o \quad (22)$$

Since n_o at $0.6943 \mu\text{m}$ lies between n_e and n_o at $0.34715 \mu\text{m}$, there will always exist an angle θ_m along which $n_e^{2\omega}(\theta_m) = n_o^\omega$. We can solve for θ_m and obtain:

$$\cos \theta_m = \left[\frac{(n_o^\omega)^2 - (n_e^{2\omega})^2}{(n_o^{2\omega})^2 - (n_e^{2\omega})^2} \right]^{1/2} \frac{n_o^{2\omega}}{n_o^\omega} \quad (23)$$

For the values given by Eq. (20), we find $\theta_m = 50.5^\circ$.

Quasi Phase Matching (QPM)

As mentioned earlier, phase matching is extremely important for any efficient nonlinear interaction. Recently the technique of quasi phase matching (QPM) has become a convenient way to achieve phase matching at any desired wavelength in any material. QPM relies on the fact that the phase mismatch in the two interacting beams can be compensated by periodically readjusting the phase of interaction through periodically modulating the nonlinear characteristics of the medium at a spatial frequency equal to the wavevector mismatch of the two interacting waves. Thus in SHG, when the nonlinear polarization at 2ω and the electromagnetic wave at 2ω have an accumulated phase difference of π , then the sign of the nonlinear coefficient is reversed so that the energy flowing from the polarization to the wave can add constructively with the existing energy (see Fig. 5). Thus, by properly choosing the period of the spatial modulation of the nonlinear coefficient, one could achieve phase matching. This scheme, QPM, is being very widely studied for application to nonlinear interactions in bulk and in waveguides.

In a ferroelectric material such as lithium niobate, the signs of the nonlinear coefficients are linked to the direction of the spontaneous polarization. Thus a periodic reversal of the domains of the crystal can be used for achieving QPM (see Fig. 6). This is the currently used technique to obtain high-efficiency SHG and other nonlinear interactions in LiNbO_3 , LiTaO_3 , and KTP . The most popular technique today, to achieve periodic domain reversal in LiNbO_3 , is the technique of electric field poling. In this method, a high electric field pulse is applied to properly oriented lithium niobate crystal using lithographically defined electrode patterns to produce a permanent periodic domain reversed pattern. Such a periodically domain reversed LiNbO_3 crystal with the periodically reversed domains going through the entire depth of the crystal is also referred to as PPLN (periodically poled lithium niobate, pronounced 'piplin') and is now commercially available.

In order to analyze SHG in a periodically poled material, let us assume that the nonlinear coefficient d varies sinusoidally with a period Λ . In such a case we have:

$$d = d_0 \sin(Kz) \quad (24)$$

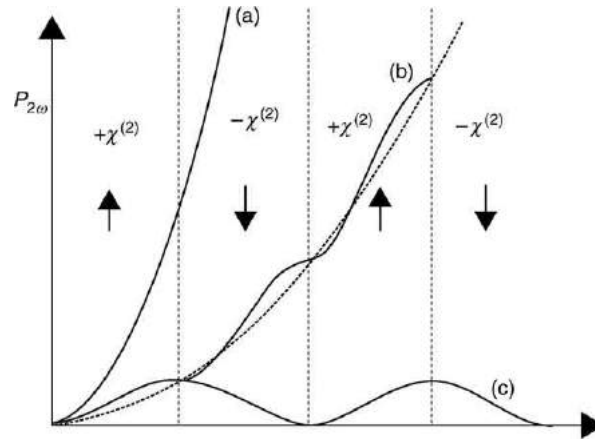


Fig. 5 By reversing the sign of the nonlinear coefficient after every coherence length, the energy in the second harmonic can be made to grow. (a) Perfect phase matching; (b) Quasi phase matching; (c) Nonphase matched.

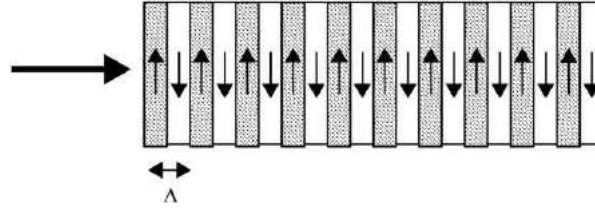


Fig. 6 QPM can be achieved by a periodic reversal of the domains of a ferroelectric material. Arrows represent the direction of the spontaneous polarization of the crystal.

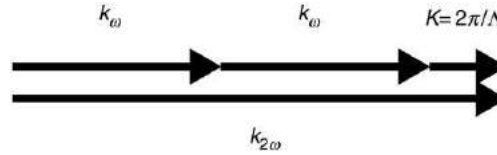


Fig. 7 Vector diagram corresponding to QPM-SHG.

where d_0 is the amplitude of modulation of the nonlinear coefficient and $K(= 2\pi/\Lambda)$ represents the spatial frequency of the periodic modulation. For easier understanding we are assuming the modulation to be sinusoidal; in general, the modulation will be periodic but not sinusoidal. Any periodic modulation can be written as a superposition of sinusoidal and cosinusoidal variations. Thus our discussion is valid for one of the Fourier components of the variation.

By using Eq. (24), Eq. (11) for the variation of the amplitude of the second harmonic becomes:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 d_0 c \omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \sin(Kz) \quad (25)$$

which can be written as:

$$\frac{dE_2}{dz} = -\frac{\mu_0 d_0 c \omega}{2n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} [e^{iKz} - e^{-iKz}] \quad (26)$$

Using similar arguments as earlier, it can be shown that if $\Delta k - K = 0$, then only the second term within the brackets in Eq. (26) contributes to the growth of the second harmonic and similarly if $\Delta k + K = 0$, then only the first term within the brackets contributes to the growth of the second harmonic. The first condition implies that:

$$2 \frac{2\pi}{\lambda_0} (n^{2\omega} - n^\omega) - \frac{2\pi}{\Lambda} = 0 \quad (27)$$

where $\Lambda = 2\pi/K$ represents the spatial period of the modulation of the nonlinear coefficient, and λ_0 is the wavelength of the fundamental. Thus the modulation period Λ required for QPM SHG is:

$$\Lambda = \frac{\lambda_0}{2(n^{2\omega} - n^\omega)} = 2L_c \quad (28)$$

Fig. 7 shows the vector diagram corresponding to QPM SHG. The phase mismatch between the fundamental and second harmonic is compensated by the spatial frequency vector of the periodic variation of d .

In the case of waveguides, n^ω and $n^{2\omega}$ would represent the effective indices of the modes at the fundamental and the second-harmonic frequency. It may be noted that the index difference between the fundamental and the second harmonic is typically 0.1 and thus the required spatial periods for a fundamental wavelength of 800 nm is $\sim 4 \mu\text{m}$.

The main advantage of QPM is that it can be used at any wavelength within the transparency window of the material; only the period needs to be correctly chosen for a specific fundamental wavelength. One can also choose appropriate polarization to make use of the largest nonlinear optical coefficients. Another advantage is the possibility of varying the domain reversal period (chirp) to achieve specific interaction characteristics such as increased bandwidth, etc.

In general, the spatial variation of the nonlinear grating is not sinusoidal. In this case the efficiency of interaction would be determined by the Fourier component of the spatial variation at the spatial frequency corresponding to the period given by Eq. (28). Also, since the required periods are very small, it is possible to use a higher spatial period of modulation and use one of the Fourier components for the nonlinear interaction process. Thus, in the case of periodic reversal of the nonlinear coefficient with a spatial period given by:

$$\Lambda_g = m \frac{\lambda_0}{2(n^{2\omega} - n^\omega)}; \quad m = 1, 3, 5, \dots \quad (29)$$

which happens to be the m^{th} harmonic of the fundamental spatial frequency required for QPM, the corresponding nonlinear coefficient that would be responsible for SHG would be the Fourier amplitude at that spatial frequency. This can be taken into

account by defining an effective nonlinear coefficient:

$$d_{\text{QPM}} = \frac{2d_0}{m\pi} \quad (30)$$

Of course the largest effective nonlinear coefficient is achieved by using the fundamental frequency with $m=1$. Higher spatial periods are easier to fabricate but would lead to reduced nonlinear efficiencies.

As compared to bulk devices, in the case of waveguides, the interacting waves are propagating in the form of modes having specific intensity distributions in the transverse plane. Because of this, the nonlinear interaction also depends on the overlap between the periodically inverted nonlinear medium, the fields of the fundamental and second-harmonic waves. Thus if $E_\omega(x, y)$ and $E_{2\omega}(x, y)$ represent the electric field distributions of the waveguide modes at the fundamental and second-harmonic frequency, then the efficiency of SHG depends on the following overlap integral:

$$I_{\text{ovl}} = \int \int d_0(x, y) E_\omega^2(x, y) E_{2\omega}(x, y) dx dy \quad (31)$$

where $d_0(x, y)$ represents the transverse variation of the nonlinear coefficient of the waveguide. Thus optimization of SHG efficiency in the case of waveguide interactions has to take account of the overlap integral.

Since QPM relies on periodic phase matching, it is highly wavelength dependent. Thus the required period Λ is different for different fundamental frequencies. Any deviation in the frequency of the fundamental would lead to a reduction in the efficiency due to deviation from QPM condition. Thus the pump laser needs to be highly stabilized at a frequency corresponding to the fabricated period.

Apart from SHG, QPM has been used for other three-wave interaction processes such as difference frequency generation, parametric amplification, etc. Among the many materials that have been studied for SHG using QPM, the important ones are lithium niobate, lithium tantalite, and potassium titanyl phosphate (KTP). Many techniques have been developed for periodic domain reversal to achieve a periodic variation in the nonlinear coefficient. This includes electric field poling, electron bombardment, thermal diffusion treatment, etc.

Third-Order Nonlinear Effects

In the earlier sections, we discussed effects arising out of second-order nonlinearity, i.e., the term proportional to E^2 in the nonlinear polarization. This nonlinear term is found only in media not possessing an inversion symmetry. Thus amorphous materials or crystals possessing inversion symmetry do not exhibit second-order effects. The lowest-order nonlinear effects in such a medium is of an order three wherein the nonlinear polarization is proportional to E^3 .

Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) represent some of the very important consequences of third-order nonlinearity. These effects have become all the more important as they play a significant and important role in wavelength division multiplexed optical fiber communication systems.

Self-Phase Modulation (SPM)

Consider the propagation of a plane light wave at frequency ω through a medium having $\chi^{(3)}$ nonlinearity. The polarization generated in the medium is given by

$$P = \epsilon_0 \chi E + \epsilon_0 \chi^{(3)} E^3 \quad (32)$$

If we consider a single frequency wave with an electric field given by

$$E = E_0 \cos(\omega t - kz) \quad (33)$$

then

$$P = \epsilon_0 \chi E_0 \cos(\omega t - kz) + \epsilon_0 \chi^{(3)} E_0^3 \cos^3(\omega t - kz) \quad (34)$$

Expanding $\cos^3 \theta$ in terms of $\cos \theta$ and $\cos 3\theta$, we obtain the following expression for the polarization at frequency ω :

$$P = \epsilon_0 \left(\chi + \frac{3}{4} \chi^{(3)} E_0^2 \right) E_0 \cos(\omega t - kz) \quad (35)$$

For a plane wave given by Eq. (33), the intensity is

$$I = \frac{1}{2} c \epsilon_0 n_0 E_0^2 \quad (36)$$

where n_0 is the refractive index of the medium at low intensity. Then

$$P = \epsilon_0 \left(\chi + \frac{3}{2} \frac{\chi^{(3)}}{c \epsilon_0 n_0} I \right) E \quad (37)$$

The polarization P and electric field are related through the following equation:

$$P = \epsilon_0 (n^2 - 1) E \quad (38)$$

where n is the refractive index of the medium. Comparing Eqs. (37) and (38), we get

$$n^2 = n_0^2 + \frac{3}{2} \frac{\chi^{(3)}}{c\epsilon_0 n_0} I \quad (39)$$

where

$$n_0^2 = 1 + \chi \quad (40)$$

Since the last term in the Eq. (39) is usually very small, we get

$$n = n_0 + n_2 I \quad (41)$$

where

$$n_2 = \frac{3}{4} \frac{\chi^{(3)}}{c\epsilon_0 n_0^3} \quad (42)$$

is the nonlinear coefficient.

For fused silica $n_0 \approx 1.47$, $n_2 \approx 3.2 \times 10^{-20} \text{ m}^2/\text{W}$ and if we consider power of 100 mW having a cross-sectional area of $100 \mu\text{m}^2$, the resultant intensity is 10^9 W/m^2 and the corresponding change in refractive index is

$$\Delta n \approx n_2 I \approx 3.2 \times 10^{-11} \quad (43)$$

Although this is very small, when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant.

In the case of an optical fiber, the light beam propagates as a mode having a specific transverse electric field distribution and thus the intensity is not constant across the cross-section. In such a case, it is convenient to express the nonlinear effect in terms of the power carried by the mode (rather than in terms of intensity). If the linear propagation constant of the mode is represented by β , then in the presence of nonlinearity, the effective propagation constant is given by

$$\beta_{\text{NL}} = \beta + \frac{k_0 n_2}{A_{\text{eff}}} P \quad (44)$$

where $k_0 = 2\pi/\lambda_0$, P is the power carried by the mode. The quantity A_{eff} represents the effective transverse cross-sectional area of the mode and is defined by

$$A_{\text{eff}} = \frac{\left[\int_0^\infty \int_0^{2\pi} \psi^2(r) r \, dr \, d\phi \right]^2}{\int_0^\infty \int_0^{2\pi} \psi^4(r) r \, dr \, d\phi} \quad (45)$$

where $\psi(r)$ represents the transverse mode field distribution of the mode. For example, under the Gaussian approximation:

$$\psi(r) = \psi_0 e^{-r^2/w_0^2} \quad (46)$$

where ψ_0 is a constant and $2w_0$ represents the mode field diameter (MFD), we get:

$$A_{\text{eff}} = \pi w_0^2 \quad (47)$$

It is usual to express the nonlinear characteristic of an optical fiber by the coefficient given by

$$\gamma = \frac{k_0 n_2}{A_{\text{eff}}} \quad (48)$$

Thus, for the same input power and same wavelength, smaller values of A_{eff} lead to greater nonlinear effects in the fiber. Typically:

$$\begin{aligned} A_{\text{eff}} &\sim 50 - 80 \mu\text{m}^2 \quad \text{and} \\ \gamma &\approx 2.4 \text{W}^{-1} \text{km}^{-1} \quad \text{to} \quad 1.5 \text{W}^{-1} \text{km}^{-1} \end{aligned} \quad (49)$$

When a light beam propagates through an optical fiber, the power decreases because of attenuation. Thus, the corresponding nonlinear effects also reduce. Indeed, the phase change suffered by a beam in propagating from 0 to L is given by

$$\phi = \int_0^L \beta_{\text{NL}} \, dz = \beta L + \gamma \int_0^L P \, dz \quad (50)$$

If α represents the attenuation coefficient, then

$$P(z) = P_0 e^{-\alpha z} \quad (51)$$

and we get

$$\phi = \beta L + \gamma P_0 L_{\text{eff}} \quad (52)$$

where

$$L_{\text{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \quad (53)$$

is referred to as the effective length. For $\alpha L \gg 1$, $L_{\text{eff}} \approx 1/\alpha$ and for $\alpha L \ll 1$, $L_{\text{eff}} \approx L$.

The effective length represents the length of the fiber over which most of the nonlinear effects has accumulated. For a loss coefficient of 0.20 dB/km, $L_{\text{eff}} \approx 21$ km.

If we consider a fiber length much longer than L_{eff} , then to have reduced impact of SPM, we must have

$$\gamma P_0 L_{\text{eff}} \ll 1 \quad (54)$$

or

$$P_0 \ll \frac{1}{\gamma L_{\text{eff}}} \approx \frac{\alpha}{\gamma} \quad (55)$$

For $\alpha = 4.6 \times 10^{-2} \text{ km}^{-1}$ (which corresponds to an attenuation of 0.2 dB/km) and $\gamma = 2.4 \text{ W}^{-1} \text{ km}^{-1}$, we get

$$P_0 \ll 19 \text{ mW} \quad (56)$$

Propagation of a Pulse

When an optical pulse propagates through a medium, it suffers from the following effects:

- (i) attenuation;
- (ii) dispersion; and
- (iii) nonlinearity.

Attenuation refers to the reduction in the pulse energy due to various mechanisms, such as scattering, absorption, etc. Dispersion is caused by the fact that a light pulse consists of various frequency components and each frequency component travels at a different group velocity. Dispersion causes the temporal width of the pulse to change; in most cases it results in an increase in pulse width, however, in some cases the temporal width could also decrease. Dispersion is accompanied by chirping, the variation of the instantaneous frequency of the pulse within the pulse duration. Since both attenuation and dispersion cause a change in the temporal variation of the optical power, they closely interact with nonlinearity in deciding the pulse evolution as it propagates through the medium.

Let $E(x, y, z, t)$ represent the electric field variation of an optical pulse. It is usual to express E in the following way:

$$E(x, y, z, t) = \frac{1}{2} [A(z, t) \psi(x, y) e^{i(\omega_0 t - \beta_0 z)} + \text{c.c.}] \quad (57)$$

where $A(z, t)$ represents the slowly varying complex envelope of the pulse, $\psi(x, y)$ represents the transverse electric field distribution, ω_0 represents center frequency, and β_0 represents the propagation constant at β_0 .

In the presence of attenuation, second-order dispersion and third-order nonlinearity, the complex envelope $A(z, t)$ can be shown to satisfy the following equation:

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2} A - \beta_1 \frac{\partial A}{\partial t} + i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial t^2} - i \gamma |A|^2 A \quad (58)$$

Here

$$\beta_1 = \left. \frac{d\beta}{d\omega} \right|_{\omega=\omega_0} = \frac{1}{v_g} \quad (59)$$

represents the inverse of the group velocity of the pulse, and

$$\beta_2 = \left. \frac{d^2 \beta}{d\omega^2} \right|_{\omega=\omega_0} = -\frac{\lambda_0^2}{2\pi c} D \quad (60)$$

where D represents the group velocity dispersion (measured in ps/km nm).

The various terms on the RHS of the Eq. (58) represent the following:

- I term:** attenuation
- II term:** group velocity term
- III term:** second-order dispersion
- IV term:** nonlinear term

If we change to a moving frame defined by coordinates $T = t - \beta_1 z$, Eq. (58) becomes

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2} A + i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2} - i \gamma |A|^2 A \quad (61)$$

If we neglect the attenuation term, we obtain the following equation which is also referred to as the nonlinear Schrödinger equation:

$$\frac{\partial A}{\partial z} = i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2} - i \gamma |A|^2 A \quad (62)$$

The above equation has a solution given by

$$A(z, t) = A_0 \text{sech} \sigma T e^{-igz} \quad (63)$$

with

$$A_0^2 = -\frac{\beta_2}{\gamma} \sigma^2, \quad g = -\frac{\sigma^2}{2} \beta_2 \quad (64)$$

Eq. (63) represents an envelope soliton and has the property that it propagates undispersed through the medium. The full width at half maximum (FWHM) of the pulse envelope will be given by $\tau_f = 2\tau_0$, where

$$\text{sech}^2 \sigma \tau_0 = \frac{1}{2} \quad (65)$$

which gives the FWHM τ_f :

$$\tau_f = 2\tau_0 = \frac{2}{\sigma} \ln(1 + \sqrt{2}) \approx \frac{1.7627}{\sigma} \quad (66)$$

The peak power of the pulse is:

$$P_0 = |A_0|^2 = \frac{|\beta_2|}{\gamma} \sigma^2 \quad (67)$$

Replacing σ by τ_f , we obtain

$$P_0 \tau_f^2 \approx \frac{\lambda_0^2}{2c\gamma} D \quad (68)$$

where we have used Eq. (60). The above equation gives the required peak for a given σ by τ_f for the formation of a soliton pulse.

As an example, we have σ by $\tau_f = 10$ ps, $\gamma = 2.4 \text{ W}^{-1} \text{ km}^{-1}$, $\lambda_0 = 1.55 \text{ } \mu\text{m}$, $D = 2 \text{ ps/km nm}$ and the required peak power will be $P_0 = 33 \text{ mW}$.

Soliton pulses are being extensively studied for application to long distance optical fiber communication. In actual systems, the pulses have to be optically amplified at regular intervals to compensate for the loss suffered by the pulses. The amplification could be carried out using erbium doped fiber amplifiers (EDFAs) or fiber Raman amplifiers.

Spectral Broadening due to SPM

In the presence of only nonlinearity, Eq. (61) becomes

$$\frac{dA}{dz} = -i\gamma |A|^2 A \quad (69)$$

whose solution is given by

$$A(z, t) = A(z=0, t) e^{-i\gamma P z} \quad (70)$$

where $P = |A|^2$ is the power in the pulse. If P is a function of time, then time dependent phase term at $z=L$ becomes

$$e^{i\varphi(t)} = e^{i[\omega_0 t - \gamma P(t)L]} \quad (71)$$

We can define an instantaneous frequency as

$$\omega(t) = \frac{d\varphi}{dt} = \omega_0 - \gamma L \frac{dP}{dt} \quad (72)$$

for a Gaussian pulse:

$$P = P_0 e^{-2T^2/\tau_0^2} \quad (73)$$

giving

$$\omega(t) = \omega_0 + \frac{4\gamma L T P_0 e^{-2T^2/\tau_0^2}}{\tau_0^2} \quad (74)$$

Thus the instantaneous frequency within the pulse changes with time, leading to chirping of the pulse (see Fig. 8). Note that since the pulsewidth has not changed, but the pulse is chirped, the frequency spectrum of the pulse has increased. Thus SPM leads to the generation of new frequencies. By Fourier transform theory, an increased spectral width implies that the pulse can now be compressed in the temporal domain by passing it through a medium with the proper sign of dispersion. This is indeed one of the standard techniques to realize ultrashort femtosecond optical pulses.

Cross Phase Modulation (XPM)

Like SPM, cross-phase modulation also arises due to the intensity dependence of refractive index, leading to spectral broadening. Unlike SPM, in the case of XPM, intensity variations of a light beam at a particular frequency modulate the phase of light beam at another frequency. If the signals at both frequencies are pulses, then due to difference in group velocities of the pulses, there is a

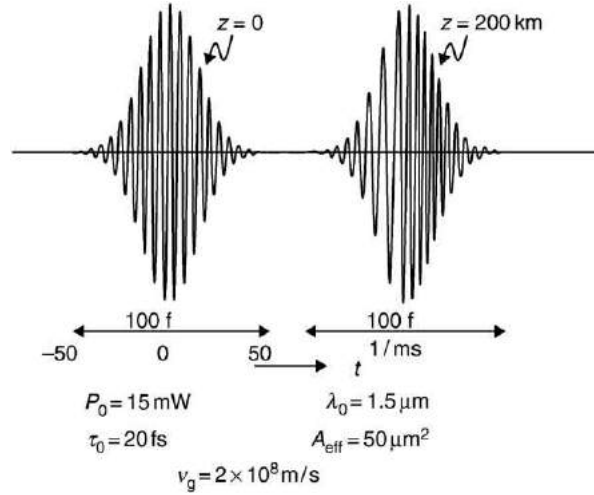


Fig. 8 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse.

walk off between the two pulses, i.e., if they start together, they will separate as they propagate through the medium. Nonlinear interaction takes place as long as they physically overlap in the medium. Smaller the dispersion, smaller will be the difference in group velocities (assuming closely spaced wavelengths) and the longer they will overlap. This would lead to stronger XPM effects. At the same time, if two pulses pass through each other, then since one pulse will interact with both the leading and the trailing edge of the other pulse, XPM effects will be nil provided there is no attenuation. In the presence of attenuation in the medium, the pulse will still get modified due to XPM. A similar situation can occur when the two interacting pulses are passing through an optical amplifier.

To study XPM, we assume simultaneous propagation of two waves at two different frequencies through the medium. If ω_1 and ω_2 represent the two frequencies, then one obtains for the variation of the amplitude A_1 of the frequency ω_1 :

$$\frac{dA_1}{dz} = -i\gamma(\tilde{P}_1 + 2\tilde{P}_2)A_1 \quad (75)$$

where $\tilde{P}_1$ and $\tilde{P}_2$ represent the powers at frequencies ω_1 and ω_2 , respectively. The first term in Eq. (75) represents SPM while the second term corresponds to XPM. If the powers are assumed to attenuate at the same rate, i.e.:

$$\tilde{P}_1 = P_1 e^{-\alpha z}, \quad \tilde{P}_2 = P_2 e^{-\alpha z} \quad (76)$$

then the solution of Eq. (75) is

$$A_1(L) = A_1(0) e^{-i\gamma(P_1 + 2P_2)L_{\text{eff}}} \quad (77)$$

where, as before, L_{eff} represents the effective length of the medium. When we are studying the effect of power at ω_2 on the light beam at frequency ω_1 we will refer to the wave at frequency ω_2 as pump, and the wave at frequency ω_1 as probe or signal. From Eq. (77) it is apparent that the phase of signal at frequency ω_1 is modified by the power at another frequency. This is referred to as XPM. Note also that XPM is twice as effective as SPM.

Similar to the case of SPM, we can now write for the instantaneous frequency in the presence of XPM as Eq. (72):

$$\omega(t) = \omega_0 - 2\gamma L_{\text{eff}} \frac{dP_2}{dt} \quad (78)$$

Hence the part of the signal that is influenced by the leading edge of the pump will be down-shifted in frequency (since in the leading edge $dP_2/dt > 0$) and the part overlapping with the trailing edge will be up-shifted in frequency (since $dP_2/dt < 0$). This leads to a frequency chirping of the signal pulse just as in the case of SPM.

If the probe and pump beams are pulses, then XPM can lead to induced frequency shifts depending on whether the probe pulse interacts only with the leading edge or trailing edge or both, as they both propagate through the medium. Let us consider a case when the group velocity of pump pulse is greater than that of the probe pulse. Thus, if both pulses enter the medium together, then since the pump pulse travels faster, the probe pulse will interact only with the trailing edge of the pump. Since in this case dP_2/dt is negative, the probe pulse suffers a blue-induced frequency shift. Similarly if the pulses enter at different instants but completely overlap at the end of the medium, then $dP_2/dt > 0$ and the probe pulse would suffer a red-induced frequency shift. Indeed, if the two pulses start separately and walk through each other, then there is no induced shift due to cancellation of shifts induced by leading and trailing edges of the pump. Fig. 9 shows measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber.

When pulses of light at two different wavelengths propagate through an optical fiber, due to different group velocities of the pulses, they pass through each other, resulting in what could be termed as a collision. In the linear case, the pulses will pass through without affecting each other, but when intensity levels are high, XPM induces phase shifts in both pulses. We can define a

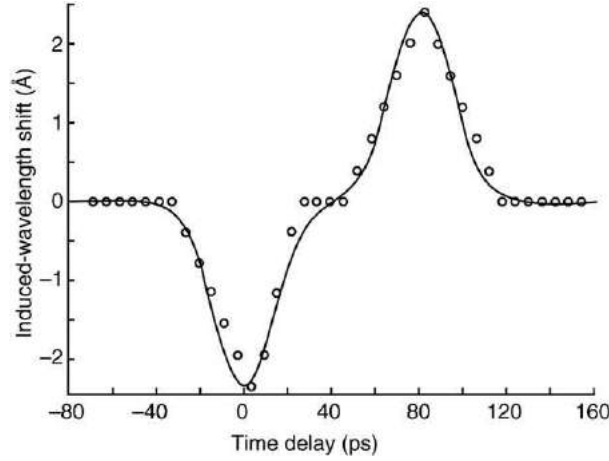


Fig. 9 Measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber. Reproduced from Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941 with permission from the American Institute of Physics.

parameter termed walk off length L_{wo} which is the length of the fiber required for the interacting pulses to walk off relative to each other. The walk off length is given by

$$L_{wo} = \frac{\Delta\tau}{D\Delta\lambda} \quad (79)$$

where D represents the dispersion coefficient, and $\Delta\lambda$ represents the wavelength separation between the interacting pulses. For return to zero (RZ) pulses, $\Delta\tau$ represents the pulse duration while for nonreturn to zero (NRZ) pulses, $\Delta\tau$ represents the rise term or fall time of the pulse. Closely spaced channels will thus interact over longer fiber lengths, thus leading to greater XPM effects. Larger dispersion coefficients will reduce L_{wo} and thus the effects of XPM. Since the medium is attenuating, the power carried by the pulses decreases as they propagate and thus leading to reduced XPM effect. The characteristic length for attenuation is the effective length L_{eff} , defined by Eq. (53). If $L_{wo} \ll L_{eff}$, then over the length of interaction of the pulses, the intensity levels do not change appreciably and the magnitude of the XPM-induced effects will be proportional to the wavelength spacing $\Delta\lambda$. For small $\Delta\lambda$'s, $L_{wo} \gg L_{eff}$ and the interaction length is now determined by the fiber losses (rather than by walk off) and the XPM-induced effects become almost independent of $\Delta\lambda$. Indeed, if we consider XPM effects between a continuous wave (cw) probe beam and a sinusoidally intensity modulated pump beam, then the amplitude of the XPM-induced phase shift ($\Delta\Phi_{pr}$) in the probe beam is given by:

$$\begin{aligned} \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{eff} & \text{for } L_{wo} \gg L_{eff} \\ \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{wo} & \text{for } L_{wo} \ll L_{eff} \end{aligned} \quad (80)$$

Here P_{2m} is the amplitude of the sinusoidal power modulation of the pump beam.

XPM-induced intensity interference can be studied by simultaneously propagating an intensity modulated pump signal and a cw probe signal at a different wavelength. The intensity modulated signal will induce phase modulation on the cw probe signal and the dispersion of the medium will convert the phase modulation to intensity modulation of the probe. Thus, the magnitude of the intensity fluctuation of the probe signal serves as an estimate of the XPM induced interference. Fig. 10 shows the intensity fluctuations on a probe signal at 1550 nm, induced by a modulated pump for a channel separation of 0.6 nm. Fig. 11 shows the variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. The experiment has been performed over four amplified spans of 80 km of standard single mode fiber (SMF) and nonzero dispersion shifted fiber (NZDSF). The large dispersion in SMF has been compensated using dispersion compensating chirped gratings. The probe modulation, in the case of SMF, decreases approximately linearly with $1/\Delta\lambda$ for all $\Delta\lambda$'s; the modulation is independent of $\Delta\lambda$. In contrast the NZDSF shows a constant modulation for $\Delta\lambda \leq 1$ nm. This is consistent with the earlier discussion in terms of L_{wo} and L_{eff} .

Four-Wave Mixing (FWM)

Four-wave mixing (FWM) is a nonlinear interaction that occurs in the presence of multiple wavelengths in a medium, leading to the generation of new frequencies. Thus, if light waves at three different frequencies $\omega_2, \omega_3, \omega_4$ are launched simultaneously into a medium, the same nonlinear polarization that led to intensity dependent refractive index, leads to nonlinear polarization component at a frequency:

$$\omega_1 = \omega_3 + \omega_4 - \omega_2 \quad (81)$$

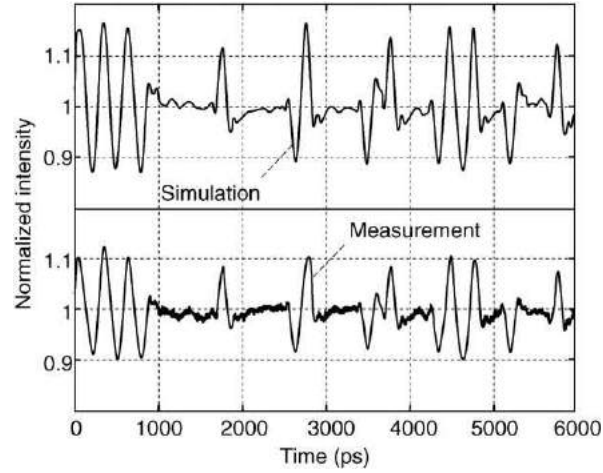


Fig. 10 Intensity fluctuations induced by cross phase modulation on a probe signal at 1550 nm by a modulated pump for a channel separation of 0.6 nm. Reproduced with permission from Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595, © 2004 IEEE.

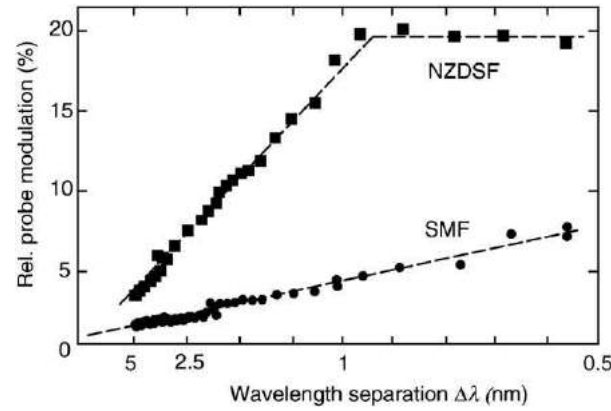


Fig. 11 Variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. Reproduced with permission from Shtai M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90, © 2004 IEEE.

This nonlinear polarization, under certain conditions, leads to the generation of electromagnetic waves at ω_1 . This process is referred to as four-wave mixing due to the interaction between four different frequencies. Degenerate four-wave mixing corresponds to the case when two of the input waves have the same frequency. Thus, if $\omega_3 = \omega_4$, then inputting waves at ω_2 and ω_3 leads to the generation of waves at the frequency

$$\omega_1 = 2\omega_3 - \omega_2 \quad (82)$$

During FWM process, there are four different frequencies present at any point in the medium. If we write the electric field of the waves as

$$E_i = \frac{1}{2} [A_i(z) \psi_i(x, y) e^{i(\omega_i t - \beta_i z)} + \text{c.c.}], \quad i = 1, 2, 3, 4 \quad (83)$$

where, as before, $A_i(z)$ represents the amplitude of the wave, $\psi_i(x, y)$ the transverse field distribution, and β_i the propagation constant of the wave. The total electric field is given by

$$E = E_1 + E_2 + E_3 + E_4 \quad (84)$$

Substituting for the total electric field in the equation for nonlinear polarization, the term with frequency ω_1 comes out as

$$P_{\text{NL}}^{(\omega_1)} = \frac{1}{2} [P_{\text{NL}}^{(\omega_1)} e^{i(\omega_1 t - \beta_1 z)} + \text{c.c.}] \quad (85)$$

where

$$P_{NL}^{(\omega_1)} = \frac{3\epsilon_0}{2} \chi^{(3)} A_2^* A_3 A_4 \psi_2 \psi_3 \psi_4 e^{-i\Delta\beta z} \quad (86)$$

and

$$\Delta\beta = \beta_3 + \beta_4 - \beta_2 - \beta_1 \quad (87)$$

In writing Eq. (86), we have only considered the FWM term, neglecting the SPM and XPM terms.

Substituting the expression for $P_{NL}^{(\omega_1)}$ in the wave equation for ω_1 and making the slowly varying approximation (in a manner similar to that employed in the case of SPM and XPM), we obtain the following equation for $A_1(z)$:

$$\frac{dA_1}{dz} = -2i\gamma A_2^* A_3 A_4 e^{-i\Delta\beta z} \quad (88)$$

where γ is defined by Eq. (48) with $k_0 = \omega/c$, ω representing the average frequency of the four interacting waves, and A_{eff} is the average effective area of the modes.

Assuming all waves to have the same attenuation coefficient α and neglecting depletion of waves at frequencies ω_2 , ω_3 and ω_4 , due to nonlinear conversion we obtain for the power in the frequency ω_1 as

$$P_1(L) = 4\gamma^2 P_2 P_3 P_4 L_{eff}^2 \eta e^{-\alpha L} \quad (89)$$

where

$$\eta = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2} \left[1 + \frac{4e^{-\alpha L} \sin^2 \frac{\Delta\beta L}{2}}{(1 - e^{-\alpha L})^2} \right] \quad (90)$$

and L_{eff} is the effective length (see Eq. (53)). Maximum four-wave mixing takes place when $\Delta\beta = 0$, since in such a case $\eta = 1$. Now:

$$\Delta\beta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_2) - \beta(\omega_1) \quad (91)$$

Since the frequencies are usually close to each other, we can make a Taylor series expansion about any frequency, say ω_2 . In such a case, we obtain

$$\Delta\beta = (\omega_3 - \omega_2)(\omega_3 - \omega_1) \left. \frac{d^2\beta}{d\omega^2} \right|_{\omega=\omega_2} \quad (92)$$

In optical fiber communication systems, the channels are usually equally spaced. Thus, we assume the frequencies to be given by

$$\begin{aligned} \omega_4 &= \omega_2 + \Delta\omega, & \omega_3 &= \omega_2 - 2\Delta\omega & \text{and} \\ \omega_1 &= \omega_2 - \Delta\omega \end{aligned} \quad (93)$$

Using these frequencies and Eq. (60), Eq. (92) gives us:

$$\Delta\beta = -\frac{4\pi D\lambda^2}{c} (\Delta\nu)^2 \quad (94)$$

where $\Delta\omega = 2\pi\Delta\nu$. Thus maximum FWM takes place when $D=0$. This is the main problem in using wavelength division multiplexing (WDM) in dispersion shifted fibers which are characterized by zero dispersion at the operating wavelength of 1550 nm, as FWM will then lead to crosstalk among various channels. FWM efficiency can be reduced by using fiber with nonzero dispersion. This has led to the development of nonzero dispersion shifted fiber (NZDSF) which have a finite nonzero dispersion of about ± 2 ps/km nm at the operating wavelength.

From Eq. (94), we notice that for a given dispersion coefficient D , FWM efficiency will reduce as $\Delta\nu$ increases.

In order to get a numerical appreciation, we consider a case with $D=0$, i.e., $\Delta\beta=0$. For such a case $\eta=1$. If all channels were launched with equal power P_{in} then:

$$P_1(L) = 4\lambda^2 P_{in}^3 L_{eff}^2 e^{-\alpha L} \quad (95)$$

Thus the ratio of power generated at ω_1 , due to FWM and that existing at the same frequency, is

$$\frac{P_g}{P_{out}} = \frac{P_1(L)}{P_{in} e^{-\alpha L}} = 4\gamma^2 P_{in}^2 L_{eff}^2 \quad (96)$$

Typical values are $L_{eff}=20$ km, $\gamma=2.4$ W⁻¹ km⁻¹. Thus:

$$\frac{P_g}{P_{out}} \approx 0.01 P_{in}^2 (\text{mW}^2) \quad (97)$$

Fig. 12 shows the output spectrum measured at the output of a 25 km-long dispersion shifted fiber ($D = -0.2$ ps/km nm) when three 3 mW wavelengths are launched simultaneously. Notice the generation of many new frequencies by FWM. Fig. 13 shows the ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. It can be seen that by choosing a nonzero value of dispersion, the four-wave mixing efficiency can be reduced. Larger the dispersion coefficient, smaller can be the channel spacing for the same crosstalk.

Since dispersion leads to increased bit error rates in fiber optic communication systems, it is important to have low dispersion. On the other hand, lower dispersion leads to crosstalk due to FWM. This problem can be resolved by noting that FWM depends on

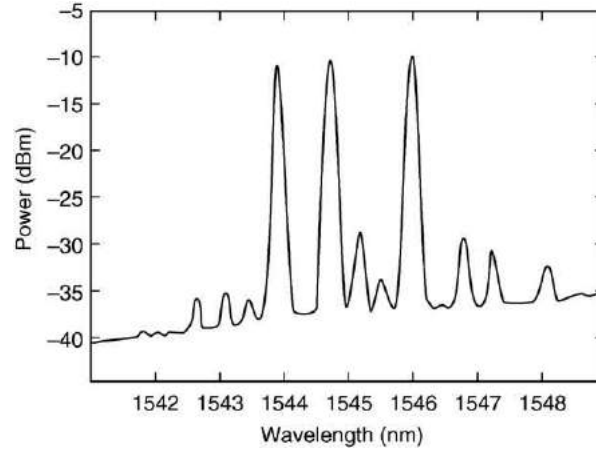


Fig. 12 Generation of new frequencies because of FWM when waves at three frequencies are incident in the fiber. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.

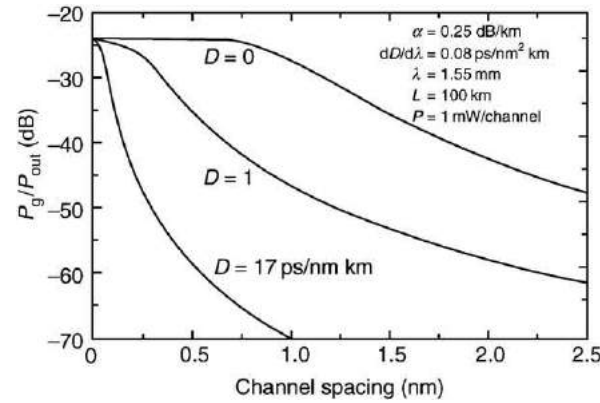


Fig. 13 Ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.

the local dispersion value in the fiber, while the pulse spreading at the end of a link depends on the overall dispersion in the fiber link. If one chooses a link made up of positive and negative dispersion coefficients, then by an appropriate choice of the lengths of the positive and negative dispersion fibers, it would be possible to achieve a zero total link dispersion while at the same time maintaining a large local dispersion. This is referred to as dispersion management in fiber optic systems.

Although FWM leads to crosstalk among different wavelength channels in an optical fiber communication system, it can be used for various optical processing functions such as wavelength conversion, high-speed time division multiplexing, pulse compression, etc. For such applications, there is a concerted worldwide effort to develop highly nonlinear fibers with much smaller mode areas and higher nonlinear coefficients. Some of the very novel fibers that have been developed recently include holey fibers, photonic bandgap fibers, or photonic crystal fibers which are very interesting since they possess extremely small mode effective areas ($\sim 2.5 \mu\text{m}^2$ at 1550 nm) and can be designed to have zero dispersion even in the visible region of the spectrum. This is expected to revolutionize nonlinear fiber optics by providing new geometries to achieve highly efficient nonlinear optical processing at lower powers.

Supercontinuum Generation

Supercontinuum (SC) generation is the phenomenon in which a nearly continuous spectrally broadened output is produced through nonlinear effects on high peak power picosecond and subpicosecond pulses. Such broadened spectra find applications in spectroscopy, wavelength characterization of optical components such as a broadband source from which many wavelength channels can be obtained by slicing the spectrum.

Supercontinuum generation in an optical fiber is a very convenient technique since it provides a very broad bandwidth ($> 200 \text{ nm}$), the intensity levels can be maintained high over long interaction lengths by choosing small mode areas, and the dispersion profile of the fiber can be appropriately designed by varying the transverse refractive index profile of the fiber.

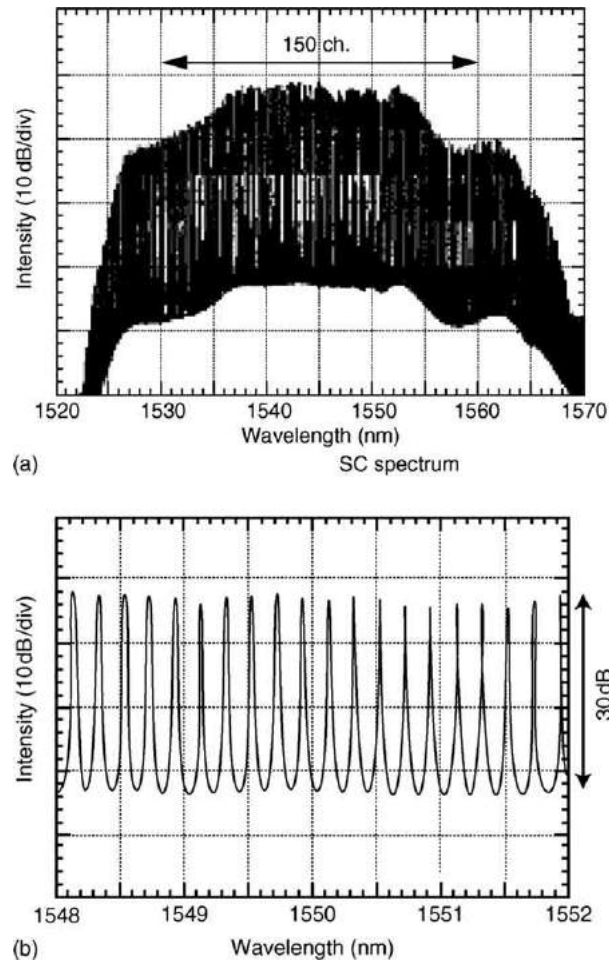


Fig. 14 Super Continuum spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. Reproduced from Yamada E, Takara H, Ohara T, Satc K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306 with permission from IEEE.

The spectral broadening that takes place in the fiber is attributed to a combination of various third-order effects such as SPM, XPM, FWM, and Raman scattering. Since dispersion plays a significant role in the temporal evolution of the pulse, different dispersion profiles have been used in the literature to achieve broadband SC.

Some studies have used dispersion decreasing fibers, dispersion flattened fibers, while others have used a constant anomalous dispersion fiber followed by a normal dispersion fiber.

Fig. 14 shows the SC spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. The fiber used for SC generation is a polarization maintaining dispersion flattened fiber. The output is a broad spectrum containing more than 150 spectral components at a spacing of 25 GHz with a flat top spectrum over 18 nm. The output optical powers range from -9 dBm to $+3$ dBm. Such sources are being investigated as attractive solutions for dense WDM (DWDM) optical fiber communication systems.

Conclusions

The advent of lasers provided us with an intense source of light and led to the birth of nonlinear optics. Nonlinear optical phenomena exhibit a rich variety of effects and coupled with optical waveguides and fibers, such effects can be observed even at moderate optical powers. With the realization of compact femtosecond lasers, highly nonlinear holey optical fibers, quasi phase matching techniques, etc, the field of nonlinear optics is expected to grow further and lead to commercial exploitation such as in compact blue lasers, soliton laser sources and multiwavelength sources and for fiber optic communication.

See also: Guided Wave Optics. Nonlinear Effects (Basics)

Further Reading

- Agrawal, G.P., 1989. *Nonlinear Fiber Optics*. Boston, MA: Academic Press.
- Armstrong, J.A., Bloembergen, N., Ducuing, J., Pershan, P.S., 1962. Interactions between light waves in a nonlinear dielectric. *Physics Review* 127, 1918.
- Baldeck, P.L., Alfano, R.R., Agrawal, G.P., 1998. Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52, 1939–1941.
- Baldi P (1992) *Generation de fluorescence parametrique guidée sur niobate et tantalite de lithium polarizes périodiquement. Etudes préliminaire d'un oscillateur paramétrique optique intégrée*. Doctoral Thesis of the University of Nice, France.
- Chiang, T.K., Kagi, N., Marhic, M.E., Kazovsky, L.G., 1996. Cross phase modulation in fiber links with multiple optical amplifiers and dispersion compensators. *Journal of Lightwave Technology* 14, 249–259.
- Franken, P.A., Hill, A.E., Peter, C.W., Weinreich, G., 1961. Generation of optical harmonics. *Physics Review Letters* 7, 118.
- Ghatak, A.K., Thyagarajan, K., 1987. *Optical Electronics*. Cambridge, UK: Cambridge University Press.
- Ghatak, A.K., Thyagarajan, K., 1998. *Introduction to Fiber Optics*. Cambridge: Cambridge University Press.
- Lim, E.J., Fejer, M.M., Byer, R.L., 1989. Second-harmonic generation green light in periodically poled planar lithium niobate waveguide. *Electronic Letters* 25, 174.
- Lim, E.J., Fejer, M.M., Byer, R.L., Kozovsky, W.J., 1989. *Electronic Letters* 25, 731–732.
- Myers, L.E., Bosenberg, W.R., 1997. Periodically poled lithium niobate and quasi phase matcher parametric oscillators. *IEEE Journal of Quantum Electronics* 33, 1663–1672.
- Newell, A.C., Moloney, J.V., 1992. *Nonlinear Optics*. Redwood City, CA: Addison Wesley Publishing Co.
- Rapp, L., 1997. Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9, 1592–1595.
- Shen, Y.R., 1989. *Principles of Nonlinear Optics*. New York: Wiley.
- Shtaif, M., Eiselt, M., Garret, L.D., 2000. Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12, 88–90.
- Tkach, R.W., Chraplyvy, A.R., Forghieri, F., Gnanck, A.H., Derosier, R.M., 1995. Four-photon mixing and high speed WDM systems. *Journal of Lightwave Technology* 13, 841–849.
- Wang, Q.Z., Lim, Q.D., Lim, D.H., Alfano, R., 1994. High resolution spectra of self phase modulation in optical fibers. *Journal of Optical Society of America B-11*, 1084.
- Webjorn, J., Siala, S., Nan, D.W., Waarts, R.G., Lang, R.J., 1997. Visible laser sources based frequency doubling in nonlinear waveguides. *IEEE Journal of Quantum Electronics* 33, 1673–1686.
- Yamada, M., Nada, N., Saitoh, M., Watanabe, K., 1993. *Applied Physics Letters* 62, 435–436.
- Yamada, E., Takara, H., Ohara, T., Sato, K., Morioka, T., Jinguji, K., Itoh, M., Ishi, M., 2001. A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37, 304–306.
- Yariv, A., 1997. *Optical Electronics in Modern Communications*, 5th ed. New York: Oxford University Press.

Fabrication of Optical Fiber

D Hewak, University of Southampton, Southampton, UK

© 2018 Elsevier Inc. All rights reserved.

Glossary

Dopants Elements or compounds added, usually in small amounts, to a glass composition to modify its properties.

Fiber drawing The process of heating and thus softening an optical fiber preform and then drawing out a thin thread of glass.

Fiber loss The transmission loss of light as it propagates through a fiber, usually measured in dB of loss per unit length of fiber. Loss can occur through the absorption of light in the core or scattering of light out of the core.

Glass An amorphous solid formed by cooling from the liquid state to a rigid solid with no long range structure.

Modified chemical vapor deposition (MCVD) A process for the fabrication of an optical preform where gases flow into the inside of a rotating tube, are heated and react to form particles of glass which are deposited onto the wall of

the glass tube. After deposition, the glass particles are consolidated into a solid preform.

Outside vapor deposition (OVD) A process for the fabrication of an optical preform where glass soot particles are formed in an oxy-hydrogen flame and deposited on a rotating rod. After deposition, the glass particles are consolidated into a solid preform.

Preform A fiber blank, a bulk glass rod consisting of a core and cladding glass composite which is drawn into fiber.

Refractive index A characteristic property of glass, which is defined by the speed of light in the material relative to the speed of light in a vacuum.

Silica A transparent glass formed from silicon dioxide.

Vapor-axial deposition A process similar to OVD, where the core and cladding layers are deposited simultaneously.

Introduction

The drawing of optical fibers from silica preforms has, over a short period of time, progressed from the laboratory to become a manufacturing process capable of producing millions of kilometers of telecommunications fiber a year. Modern optical fiber fabrication processes produce low-cost fiber of excellent quality, with transmission losses close to their intrinsic loss limit. Today, fiber with transmission losses of 0.2 dB per kilometer of fiber are routinely drawn through a two-stage process that has been refined since the 1970s.

Although fibers of glass have been fabricated and used for hundreds of years, it was not until 1966 that serious interest in the use of optical fibers for communication emerged. At this time, it was estimated that the optical transmission loss in bulk glass could be as low as 20 dB km⁻¹ if impurities were sufficiently reduced, a level at which practical applications were possible. At this time, no adequate fabrication techniques were available to synthesize glass of high purity, and fiber-drawing methods were crude.

Over the next five years, efforts worldwide addressed the fabrication of low-loss fiber. In 1970, a fiber with a loss of 20 dB km⁻¹ was achieved. The fiber consisted of a titania doped core and pure silica cladding. This result generated much excitement and a number of laboratories worldwide actively began researching optical fiber. New fabrication techniques were introduced, and by 1986, fiber loss had been reduced close to the theoretical limit.

All telecommunications fiber that is fabricated today is made of silica glass, the most suitable material for low-loss fibers. Early fiber research studied multicomponent glasses, which are perhaps more familiar optical materials; however, low-loss fiber, could not be realized, partly due to the lack of a suitable fabrication method. Today other glasses, in particular the fluorides and sulfides, continue to be developed for speciality fiber applications, but silica fiber dominates in most applications.

Silica is a glass of simple chemical structure containing only two elements, silicon and oxygen. It has a softening temperature of about 2,000°C at which it can be stretched, i.e. drawn into fiber. An optical fiber consists of a high purity silica glass core, doped with suitable oxide materials to raise its refractive index (Fig. 1). This core, typically on the order of 2–10 microns in diameter, is surrounded by silica glass of lower refractive index. This cladding layer extends the diameter to typically 125 microns. Finally a protective coating covers the entire structure. It is the phenomenon of total internal reflection at the core cladding interface that confines light to the core and allows it to be guided. The basic requirements of an optical fiber are as follows:

1. The material used to form the core of the fiber must have a higher refractive index than the cladding material, to ensure the fiber is a guiding structure.
2. The materials used must be low loss, providing transmission with no absorption or scattering of light.
3. The materials used must have suitable thermal and mechanical properties to allow them to be drawn down in diameter into a fiber.

Silica (SiO₂) can be made into a glass relatively easily. It does not easily crystallize, which means that scattering from unwanted crystalline centers within the glass is negligible. This has been a key factor in the achievement of a low-loss fiber. Silica glass has

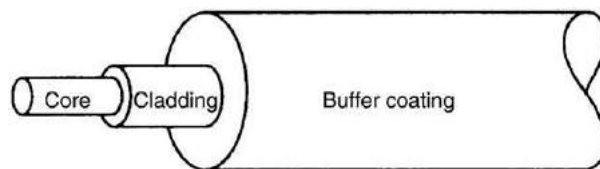


Fig. 1 Structure of an optical fiber.

high transparency in the visible and near-infrared wavelength regions and its refractive index can be easily modified. It is stable and inert, providing excellent chemical and mechanical durability. Moreover, the purification of the raw materials used to synthesize silica glass is quite straightforward.

In the first stage of achieving an optical fiber, silica glass is synthesized by one of three main chemical vapor processes. All use silicon tetrachloride (SiCl_4) as the main precursor, with various dopants to modify the properties of the glass. The precursors are reacted with oxygen to form the desired oxides. The end result is a high purity solid glass rod with the internal core and cladding structure of the desired fiber.

In the second stage, the rod, or preform as it is known, is heated to its softening temperature and stretched to diameters of the order of 125 microns. Tens to hundreds of kilometers of fiber are produced from a single preform, which is drawn continuously, with minimal diameter fluctuations. During the drawing process one or more protective coatings are applied, yielding long lengths of strong, low-loss fiber, ready for immediate application.

Preform Fabrication

The key step in preparing a low-loss optical fiber is to develop a technique for completely eliminating transition metal and OH ion contamination during the synthesis of the silica glass. The three methods most commonly used to fabricate a glass optical fiber preform are: the modified chemical vapor deposition process (MCVD); the outside vapor deposition process (OVD); and the vapor-axial deposition process (VAD).

In a typical vapor phase reaction, halide precursors undergo a high temperature oxidation or hydrolysis to form the desired oxides. The completed chemical reaction for the formation of silica glass is, for oxidation:



For hydrolysis, which occurs when the deposition occurs in a hydrogen-containing flame, the reaction is:



These processes produce fine glass particles, spherical in shape with a size of the order of 1 nm. These glass particles, known as soot, are then deposited and subsequently sintered into a bulk transparent glass. The key to low-loss fiber is the difference in vapor pressures of the desired halides and the transition metal halides that cause significant absorption loss at the wavelengths of interest. The carrier gas picks up a pure vapor of, for example, SiCl_4 , and any impurities are left behind.

Part of the process requires the formation of the desired core/cladding structure in the glass. In all cases, silica-based glass is produced with variations in refractive index produced by the incorporation of dopants. Typical dopants used are germania (GeO_2), titania (TiO_2), alumina (Al_2O_3), and phosphorous pentoxide (P_2O_5) for increasing the refractive index, and boron oxide (B_2O_3) and fluorine (F) for decreasing it. These dopants also allow other properties to be controlled, such as the thermal expansion of the glass and its softening temperatures. In addition, other materials, such as the rare earth elements, have also been used to fabricate active fibers that are used to produce optical fiber amplifiers and lasers.

MCVD Process

Optical fibers were first produced by the MCVD method in 1974, a breakthrough that completely solved the technical problems of low-loss fiber fabrication (**Fig. 2**).

As shown schematically in **Fig. 3**, the halide precursors are carried in the vapor phase by oxygen carrier gas into a pure silica substrate tube. An oxy-hydrogen burner traverses the length of the tube, which it heats externally. The tube is heated to temperatures of about $1,400^\circ\text{C}$ which then oxidizes the halide vapor materials. The deposition temperature is sufficiently high to form a soot made of glassy particles which are deposited on the inside wall of the substrate tube but low enough to prevent the softened silica substrate tube from collapsing. The process usually takes place on a horizontal glass-working lathe.

During the deposition process, the repeated traversing of the burner forms multiple layers of soot. Changes to the precursors entering the tube and thus the resulting glass composition are introduced for layers which will form the cladding and then the core. The MCVD method allows germania to be doped into the silica glass and the precise control of the refractive index profile of the preform. When deposition is complete, the burner temperature is increased and the hollow, multilayered structure is collapsed to a solid rod. A characteristic of fiber formed by this process is a refractive index dip in the center of the core of the fiber. In addition, the deposition takes place in a closed system, which dramatically reduces contamination by OH^- ions and maintains



Fig. 2 Fabrication of an optical fiber preform by the MCVD method.

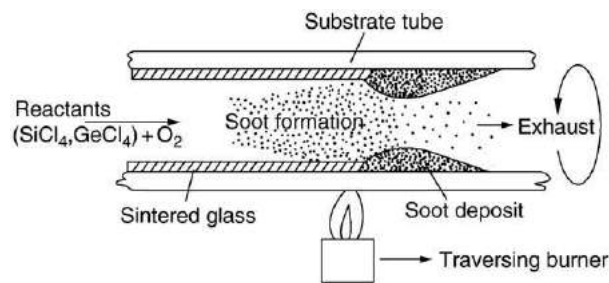


Fig. 3 Schematic of preform fabrication by the MCVD method.

low levels of other impurities. The high temperature allows high deposition rates compared to traditional chemical vapor deposition (CVD) and large preforms, built up from hundreds of layers can be produced.

The MCVD method is still widely used today though it has some limitations, particularly on the preform size that can be achieved and thus the manufacturing cost. Diameter of the final preform is determined by the size of the initial silica substrate tube, which, due to the high purity required, accounts for a significant portion of the cost of the preform. A typical preform fabricated by MCVD yields about 5 km of fiber. Its success has spurred improvements to the process, in particular to address the fiber yield from a single preform.

OVD Process

The outside vapor deposition (OVD) methods, also known as outside vapor phase oxidation (OVPO), synthesize the fine glass particles within a burner flame. The precursors, oxygen, and fuel for the burner, are introduced directly into the flame. The soot is deposited onto a target rod that rotates and traverses in front of the burner. As in MCVD, the preform is built up layer by layer, though now initially by depositing the core glass and then building up the cladding layers over this. After the deposition process is complete, the preform is removed from the target rod and is collapsed and sintered into a transparent glass preform. The center hole remains, but disappears during the fiber drawing process. This technique has advantages in both size of preform which can be obtained and the fact that a high-quality silica substrate tube is no longer required. These two advantages combine to make a more economical process. From a single preform, several hundred kilometers of fiber can be produced.

VAD Process

The most recent refinement to the fabrication process was developed again to aid the mass production of high-quality fibers. In the VAD process, both core and cladding glasses are deposited simultaneously. Like OVD, the soot is synthesized and deposited by flame hydrolysis, as shown in **Fig. 4**, the precursors are blown from a burner, oxidized, and deposited onto a silica target rod.

Burners for the VAD process consist of a series of concentric nozzles. The first delivers an inert carrier and the main precursors SiCl_4 , the second delivers an inert carrier glass and the dopants, the third delivers hydrogen fuel, and the fourth delivers oxygen. Gas flows are up to a liter per minute and deposition rates can be very high. With the VAD process, both core and cladding glasses can be deposited simultaneously. The main advantage is that this is a continuous process, as the soot which forms the core and cladding glasses are deposited axially onto the end of the rotating silica rod, it is slowly drawn upwards into a furnace which sinters

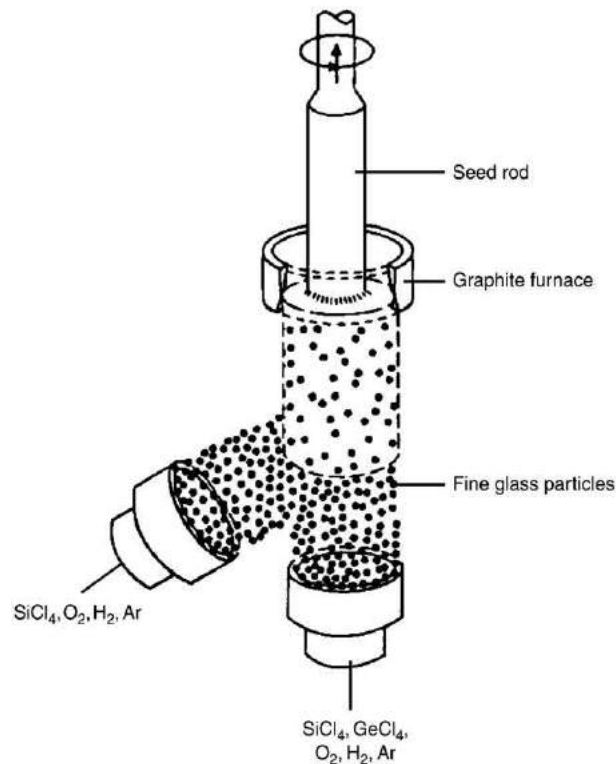


Fig. 4 Schematic of preform fabrication by the VAD method.

and consolidates the soot into a transparent glass. The upward motion is such that the end at which deposition is occurring remains in a fixed position and essentially the preform is grown from this base.

The advantages of the VAD process are a preform without a central dip and, most importantly, the mass production associated with a continuous process. The glass quality produced is uniform and the resulting fibers have excellent reproducibility and low loss.

Other Methods of Preform Fabrication

There are other methods of preform fabrication, though these are not generally used for today's silica glass based fiber. Some methods, such as fabrication of silica through sol-gel chemistry could not provide the large low-loss preforms obtained by chemical vapor deposition techniques. Other methods, such as direct casting of molten glass into rods, or formation of rods and tubes by extrusion, are better suited to and find application in other glass chemistries and speciality fibers.

Fiber Drawing

Once a preform has been made, the second step is to draw the preform down in diameter on a fiber drawing tower. The basic components and configuration of the drawing tower have remained unchanged for many years, although furnace design has become more sophisticated and processes like coating have become automated. In addition, fiber drawing towers have increased in height to allow faster pulling speeds (**Fig. 5**).

Essentially, the fiber drawing process takes place as follows. The preform is held in a chuck which is mounted on a precision feed assembly that lowers the preform into the drawing furnace at a speed which matches the volume of preform entering the furnace to the volume of fiber leaving the bottom of the furnace. Within the furnace, the fiber is drawn from the molten zone of glass down the tower to a capstan, which controls the fiber diameter, and then onto a drum. During the drawing process, immediately below the furnace, is a noncontact device that measures the diameter of the fiber. This information is fed back to the capstan which speeds up to reduce the diameter or slows down to increase the fiber diameter. In this way, a constant diameter fiber is produced. One or more coatings are also applied in-line to maintain the pristine surface quality as the fiber leaves the furnace and thus to maximize the fiber strength.

The schematic drawing in **Fig. 6** shows a fiber drawing tower and its key components. Draw towers are commercially available, ranging in height from 3 m to greater than 20 m, with the tower height increasing as the draw speed increases.

Typical drawing speeds are on the order of 1 m s^{-1} . The increased height is needed to allow the fiber to cool sufficiently before entering the coating applicator, although forced air-cooling of the fiber is commonplace in a manufacturing environment.



Fig. 5 A commercial scale optical fiber drawing tower.

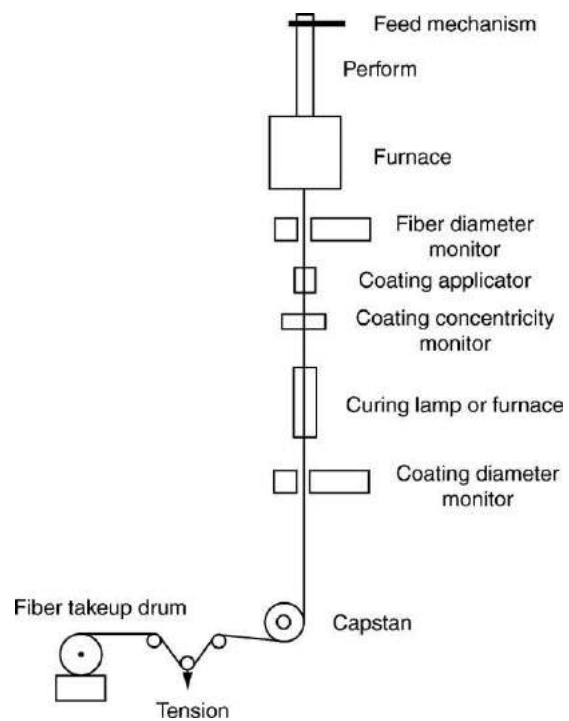


Fig. 6 Schematic diagram of an optical fiber drawing tower and its components.

Furnaces

A resistance furnace is perhaps the most common and economical heating source used in a fiber drawing tower. A cylindrical furnace, with graphite or tungsten elements, provides heating to the preform by blackbody radiation. These elements must be

surrounded by an inert gas, such as nitrogen or argon, to prevent oxidation. Gas flow and control is critical to prevent variations in the diameter of the fiber. In addition, the high temperature of the elements may lead to contamination of the preform surface that can then weaken the fiber.

An induction furnace provides precise clean heating through radio frequency (RF) energy that is inductively coupled to a zirconia susceptor ring. This type of furnace is cleaner than the graphite resistance type and is capable of continuous running for several months. It also has the advantage that it does not require a protective inert atmosphere and consequently the preform is drawn in a turbulence-free environment. These advantages result in high-strength fibers with good diameter control.

Diameter Measurement

A number of noncontact methods of diameter measurement can be applied to measure fiber diameter. These include laser scanning, interferometry, and light scattering techniques. To obtain precise control of the fiber diameter, the deviation between the desired diameter and the measured diameter is fed into the diameter control system. In order to cope with high drawing speeds, sampling rates as high as 1,000 times per second are used. The diameter control is strongly affected by the gas flow in the drawing furnace and is less affected by the furnace temperature variation. The furnace gas flow can be used to achieve suppression of fast diameter fluctuations. This is used in combination with drawing speed control to achieve suppression of both fast and slow diameter fluctuations. Current manufacturing processes are capable of producing several hundreds of kilometers of fiber with diameter variations of ± 1 micron.

There are two major sources of fiber diameter fluctuations: short-term fluctuations caused by temperature fluctuations in the furnace; and long-term fluctuations caused by variations in the outer diameter of the preform. Careful control of the furnace temperature, the length of the hot zone, and the flow of gas minimize the short-term fluctuations. Through optimization of these parameters, diameter errors of less than ± 0.5 microns can be realized. The long-term fluctuations in diameter are controlled by the feedback mechanism between the diameter measurement and the capstan.

Fiber Coating

A fiber coating is primarily used to preserve the strength of a newly drawn fiber and therefore must be applied immediately after the fiber leaves the furnace. The fiber coating apparatus is typically located below the diameter measurement gauge at a distance determined by the speed of fiber drawing, the tower height, and whether there is external cooling of the fiber. Coating is usually

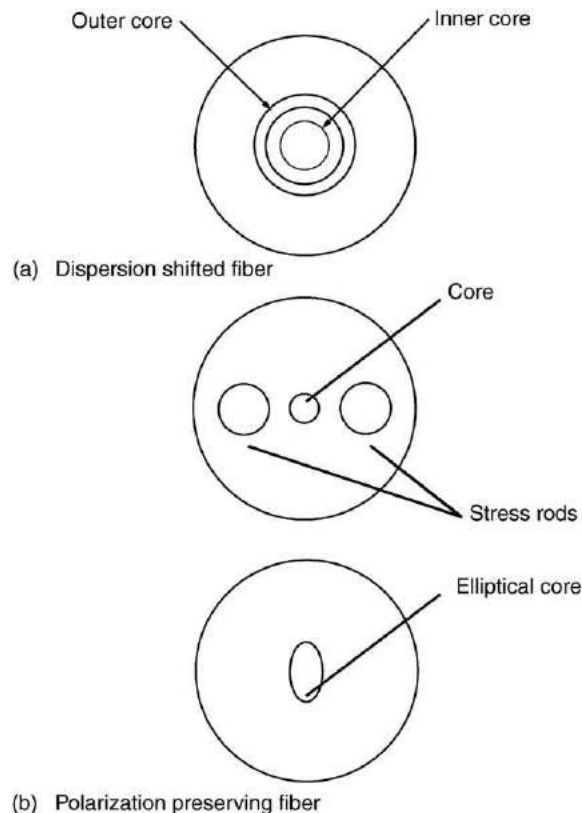


Fig. 7 Structure of (a) dispersion shifted fiber and (b) two methods of achieving polarization preserving fiber.

one of the limiting factors in the speed of fiber drawing. To minimize any damage or contamination of the pristine fiber leaving the furnace, this portion of the tower, from furnace to the application of the first coating, is enclosed in a clean, filtered-air chamber.

A wide variety of coating materials has been applied; however, conventional, commercial fiber generally relies on a UV curable polymer as the primary coating. Coating thickness is typically 50–100 microns. Subsequent coatings can then be applied for specific purposes. A dual coating is often used with an inner primary coating that is soft and an outer, secondary coating which is hard. This ratio of low to high elastic modulus can minimize stress on the fiber and reduce bending loss.

Alternative coatings include hermetic coatings of a low melting temperature metal, ceramics, or amorphous carbon. These can be applied in-line before the polymeric coating. Metallic coatings are applied by passing the fiber through a molten metal while ceramic or amorphous coatings utilize an in-line chemical vapor deposition reactor.

Types of Fiber

The majority of silica fiber drawn today is single mode. This structure consists of a core whose diameter is chosen such that, with a given refractive index difference between the core and cladding, only a single guide mode propagates at the wavelength of interest. With a discrete index difference between core and cladding, it is often referred to as a step index fiber. In typical telecommunications fiber, single mode operation is obtained with core diameters of 2–10 microns with a standard outer diameter of 125 microns.

Multimode fiber has core diameters considerably larger, typically 50, 62.5, 85, and 110 microns, again with a cladded diameter of 125 microns. Multimode fibers are often graded index, that is the refractive index is a maximum in the center of the fiber and smoothly decreases radially until the lower cladding index is reached. Multimode fibers find use in non-telecommunication applications, for example optical fiber sensing and medicine.

Single mode fibers, which are capable of maintaining a linear polarization input to the fiber, are known as polarization preserving fibers. The structure of these fibers provides a birefringence that removes the degeneracy of the two possible polarization modes. This birefringence is a small difference in the effective refractive index of the two polarization modes that can be guided and it is achieved in one of two ways. Common methods for the realization of this birefringence are an elliptical core in the fiber, or through stress rods, which modify the refractive index in one orientation. These fiber structures are shown in [Fig. 7](#).

While the loss minimum of silica-based fiber is near 1.55 microns, step index single-mode fiber offers zero dispersion close to 1.3 micron wavelengths and dispersion at the loss minimum is considerable. A modification of the structure of the fiber, and in particular a segmented refractive index profile in the core, can be used to shift this dispersion minimum to 1.55 microns. This fiber, illustrated in [Fig. 7](#) is known as dispersion shifted fiber. Similarly fibers, with a relatively low dispersion over a wide wavelength range, known as dispersion flattened fibers, can be obtained by the use of multiple cladding layers.

See also: Guided Wave Optics. Optical Fiber Cables

Further Reading

- Agrawal, G.P., 1992. *Fiber-Optic Communication Systems*. New York: John Wiley & Sons.
- Bass, M., Van Stryland, E.V. (Eds.), 2001. *Fiber Optics Handbook: Fiber, Devices, and Systems for Optical Communications*. New York: McGraw-Hill Professional Publishing.
- Fujiura, K., Kanamori, T., Sudo, S., 1997. Fiber materials and fabrications. In: Sudo, S. (Ed.), *Optical Fiber Amplifiers*. Boston, MA: Artech House, pp. 193–404. ch. 4.
- Goff, D.R., Hansen, K.S., 2002. *Fiber Optic Reference Guide: A Practical Guide to Communications Technology*, 3rd edn Oxford, UK: Butterworth-Heinemann UK.
- Hewak, D. (Ed.), 1998. *Glass and Rare Earth Doped Glasses for Optical Fibres*. EMIS Datareview Series, INSPEC. London: The Institution of the Electrical Engineers.
- Keck, D.B., 1981. Optical fibre waveguides. In: Barnoski, M.K. (Ed.), *Fundamentals of Optical Fiber Communications*. New York: Academic Press.
- Li, T. (Ed.), 1985. *Optical Fiber Communications: Fiber Fabrication*. New York: Academic Press.
- Personick, S.D., 1985. *Fiber Optics: Technology and Applications*. New York: Plenum Press.
- Schultz, P.C., 1979. In: Bendow, B., Mitra, S.S. (Eds.), *Fiber Optics*. New York: Plenum Press.

Overview: Coherence

A Sharma and AK Ghatak, Indian Institute of Technology, New Delhi, India
HC Kandpal, National Physical Laboratory, New Delhi, India

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Coherence refers to the characteristic of a wave that indicates whether different parts of the wave oscillate in tandem, or in a definite phase relationship. In other words, it refers to the degree of confidence by which one can predict the amplitude and phase of a wave at a point, from the knowledge of these at another point at the same or a different instant of time. Emission from a thermal source, such as a light bulb, is a highly disordered process and the emitted light is incoherent. A well-stabilized laser source, on the other hand, generates light in a highly ordered manner and the emitted light is highly coherent. Incoherent and coherent light represent two extreme cases. While describing the phenomena of physical optics and diffraction theory, light is assumed to be perfectly coherent in both spatial as well as temporal senses, whereas in radiometry it is generally assumed to be incoherent. However, practical light sources and the fields generated by them are in between the two extremes and are termed as partially coherent sources and fields. The degree of order that exists in an optical field produced by a source of any kind may be described in terms of various correlation functions. These correlation functions are the basic theoretical tools for the analyses of statistical properties of partially coherent light fields.

Light fields generated from real physical sources fluctuate randomly to some extent. On a microscopic level quantum mechanical fluctuations produce randomness and on macroscopic level the randomness occurs as a consequence of these microscopic fluctuations, even in free space. In real physical sources, spontaneous emission causes random fluctuations and even in the case of lasers, spontaneous emission cannot be suppressed completely. In addition to spontaneous emission, there are many other processes that give rise to random fluctuations of light fields. Optical coherence theory was developed to describe the random nature of light and it deals with the statistical similarity between light fluctuations at two (or more) space–time points.

As mentioned earlier, in developing the theory of interference or diffraction, light is assumed to be perfectly coherent, or, in other words, it is taken to be monochromatic and sinusoidal for all times. This is, however, an idealization since the wave is obviously generated at some point of time by an atomic or molecular transition. Furthermore, a wavetrain generated by such a transition is of a finite duration, which is related to the finite lifetime of the atomic or molecular levels involved in the transition. Thus, any wave emanating from a source is an ensemble of a large number of such wavetrains of finite duration, say τ_c . A simplified visualization of such an ensemble is shown in Fig. 1 where a wave is shown as a series of wavetrains of duration τ_c . It is evident from the figure that the fields at time t and $t + \Delta t$ will have a definite phase relationship if $\Delta t \ll \tau_c$ and will have no or negligible phase relationship when $\Delta t \gg \tau_c$. The time τ_c is known as the coherence time of the radiation and the field is said to remain coherent for time $\sim \tau_c$. This property of waves is referred to as the time coherence or the temporal coherence and is related to the spectral purity of the radiation. If one obtains the spectrum of the wave shown in Fig. 1 by taking the Fourier transform of the time variation, it would have a width of $\Delta\nu$ around ν_0 which is the frequency of the sinusoidal variation of individual wavetrains. The spectral width $\Delta\nu$ is related to the coherence time as

$$\Delta\nu \sim 1/\tau_c \quad (1)$$

For thermal sources such as a sodium lamp, $\tau_c \sim 100$ ps, whereas for a laser beam it could be as large as a few milliseconds. A related quantity is the coherence length l_c which is the distance covered by the wave in time τ_c

$$l_c = c\tau_c \sim \frac{c}{\Delta\nu} = \frac{\lambda_0^2}{\Delta\lambda} \quad (2)$$

where λ_0 is the central wavelength ($\lambda_0 = c/\nu_0$) and $\Delta\lambda$ is the wavelength spread corresponding to the spectral width $\Delta\nu$. In a two-beam interference experiment (e.g., Michelson interferometer, Fig. 2), the interfering beam derived from the same source will produce good interference fringes if the path difference between the two interfering beams is less than the coherence length of the radiation given out by the source.

It must be added here that for the real fields, generated by innumerable atoms or molecules, the individual wavetrains have different lengths around an average value τ_c . Furthermore, several wavetrains in general are propagating simultaneously,

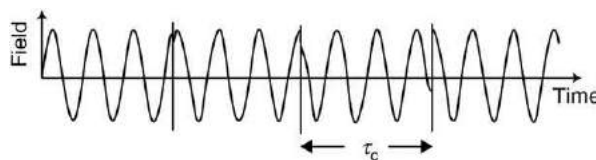


Fig. 1 Typical variation of the radiation field with time. The coherence time $\sim \tau_c$.

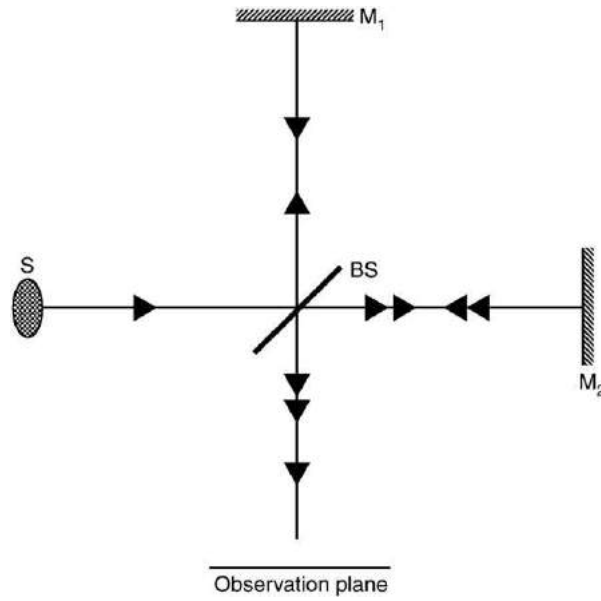


Fig. 2 Michelson interferometer to study the temporal coherence of radiation from source S; M_1 and M_2 are mirrors and BS is a 50%–50% beamsplitter.

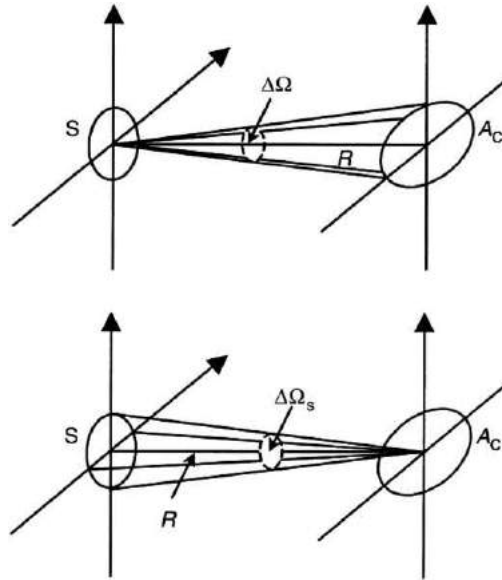


Fig. 3 Spatial coherence and coherence area.

overlapping in space and time, to produce an ensemble whose properties are best understood in a statistical sense as we shall see in later sections.

Another type of coherence associated with the fields is the space coherence or the spatial coherence, which is related to the size of the source of radiation. It is evident that when the source is an ideal point source, the field at any two points (within the coherence length) would have definite phase relationship. However, the field from a thermal source of finite area S can be thought as resultant of the fields from each point on the source. Since each point source would usually be independent of the other, the phase relationship between the fields at any two points would depend on their position and size of the source. It can be shown that two points will have a strong phase relationship if they lie within the solid angle $\Delta\Omega$ from the source (**Fig. 3**) such that

$$\Delta\Omega \sim \frac{\lambda_0^2}{S} \quad (3)$$

Thus, on a plane R distance away from the source, one can define an area $A_c = R^2 \Delta\Omega$ over which the field remains spatially coherent. This area A_c is called the coherence area of the radiation and its square root is sometimes referred to as the transverse

coherence length. It is trivial to show that

$$A_c \sim \frac{\lambda_0^2}{\Delta\Omega_s} \quad (4)$$

where $\Delta\Omega_s$ is the solid angle subtended by the source on the plane at which the coherence area is defined. In Young's double-hole experiment, if the two pinholes lie within the coherence area of the radiation from the primary source, good contrast in fringes would be observed.

Combining the concepts of coherence length and the coherence area, one can define a coherence volume as $V_c = A_c L_c$. For the wavefield from a thermal source, this volume represents that portion of the space in which the field is coherent and any interference produced using the radiation from points within this volume will produce fringes of good contrast.

Mathematical Description of Coherence

Analytical Field Representation

Coherence properties associated with fields are best analyzed in terms of complex representation for optical fields. Let the real function $V^{(r)}(\mathbf{r}, t)$ represent the scalar field, which could be one of the transverse Cartesian components of the electric field associated with the electromagnetic wave. One can then define a complex analytical signal $V(\mathbf{r}, t)$ such that

$$V^{(r)}(\mathbf{r}, t) = \text{Re}[V(\mathbf{r}, t)], \quad (5)$$

$$V(\mathbf{r}, t) = \int_0^\infty \tilde{V}^{(r)}(\mathbf{r}, \nu) e^{-2\pi i \nu t} d\nu$$

where the spectrum $\tilde{V}^{(r)}(\mathbf{r}, \nu)$ is the Fourier transform of the scalar field $V^{(r)}(\mathbf{r}, t)$ and the spectrum for negative frequencies has been suppressed as it does not give any new information since $\tilde{V}^{(r)*}(\mathbf{r}, \nu) = \tilde{V}^{(r)}(\mathbf{r}, -\nu)$.

In general, the radiation from a quasi-monochromatic thermal source fluctuates randomly as it is made of a large number of mutually independent contributions from individual atoms or molecules in the source. The field from such a source can be regarded as an ensemble of a large number of randomly different analytical signals such as $V(\mathbf{r}, t)$. In other words, $V(\mathbf{r}, t)$ is a typical member of an ensemble, which is the result of a random process representing the radiation from a quasi-monochromatic source. This process is assumed to be stationary and ergodic so that the ensemble average is equal to the time average of a typical member of the ensemble and that the origin of time is unimportant. Thus, the quantities of interest in the theory of coherence are defined as time averages:

$$\langle f(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(t) dt \quad (6)$$

Mutual Coherence

In order to define the mutual coherence, the key concept in the theory of coherence, we consider Young's interference experiment, as shown in Fig. 4, where the radiation from a broad source of size S is illuminating a screen with two pinholes P_1 and P_2 . The light emerging from the two pinholes produces interference, which is observed on another screen at a distance R from the first screen. The field at point P , due to the pinholes, would be $K_1 V(\mathbf{r}_1, t - t_1)$ and $K_2 V(\mathbf{r}_2, t - t_2)$ respectively, where $\mathbf{r}_1$ and $\mathbf{r}_2$ define the positions of P_1 and P_2 , t_1 and t_2 are the times taken by the light to travel to P from P_1 and P_2 , and K_1 and K_2 are imaginary constants that depend on the geometry and size of the respective pinhole and its distance from the point P . Thus, the resultant field at point P would be given by

$$V(\mathbf{r}, t) = K_1 V(\mathbf{r}_1, t - t_1) + K_2 V(\mathbf{r}_2, t - t_2) \quad (7)$$

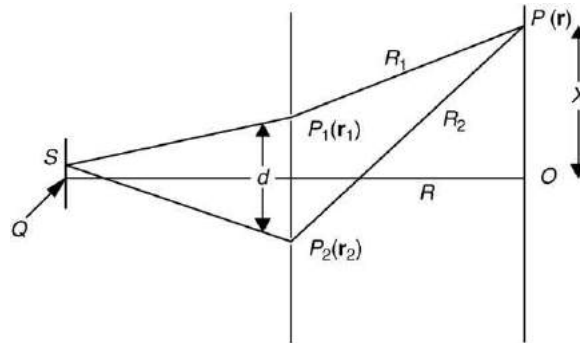


Fig. 4 Schematic of Young's double-hole experiment.

Since the optical periods are extremely small as compared to the response time of a detector, the detector placed at point P will record only the time-averaged intensity:

$$I(P) = \langle V^*(\mathbf{r}, t) V(\mathbf{r}, t) \rangle \quad (8)$$

where some constants have been ignored. With Eq. (7), the intensity at point P would be given by

$$I(P) = I_1 + I_2 + 2 \operatorname{Re} \{ K_1 K_2^* \Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2) \} \quad (9)$$

where I_1 and I_2 are the intensities at point P , respectively, due to radiations from pinholes P_1 and P_2 independently (defined as $I_i = \langle |K_i V(\mathbf{r}_i, t)|^2 \rangle$) and

$$\begin{aligned} \Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2) &\equiv \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) \\ &= \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \rangle \end{aligned} \quad (10)$$

is the mutual coherence function of the fields at P_1 and P_2 , and it depends on the time difference $\tau = t_2 - t_1$, since the random process is assumed to be stationary. This function is also sometimes denoted by $\Gamma_{12}(\tau)$. The function $\Gamma_{ii}(\tau) \equiv \Gamma(\mathbf{r}_i, \mathbf{r}_i, \tau)$ defines the self-coherence of the light from the pinhole at P_i , and $|K_i|^2 \Gamma_{ii}(0)$ defines the intensity I_i at point P , due to the light from pinhole P_i .

Degree of Coherence and the Visibility of Interference Fringes

One can define a normalized form of the mutual coherence function, namely:

$$\begin{aligned} \gamma_{12}(\tau) &\equiv \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{12}(0)} \sqrt{\Gamma_{22}(0)}} \\ &= \frac{\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \rangle}{[\langle |V(\mathbf{r}_1, t)|^2 \rangle \langle |V(\mathbf{r}_2, t)|^2 \rangle]^{1/2}} \end{aligned} \quad (11)$$

which is called the complex degree of coherence. It can be shown that $0 \leq |\gamma_{12}(\tau)| \leq 1$. The intensity at point P , given by Eq. (9), can now be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2} \operatorname{Re}[\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)] \quad (12)$$

Expressing $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ as

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \exp[i\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi i \nu_0 \tau] \quad (13)$$

where $\alpha(\vec{r}_1, \vec{r}_2, \tau) = \arg[\gamma(\vec{r}_1, \vec{r}_2, \tau)] + 2\pi \nu_0 \tau$ and ν_0 is the mean frequency of the light, the intensity in Eq. (12) can be written as

$$\begin{aligned} I(P) &= I_1 + I_2 + 2\sqrt{I_1 I_2} |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \\ &\quad \times \cos[\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi \nu_0 \tau] \end{aligned} \quad (14)$$

Now, if we assume that the source is quasi-monochromatic, i.e., its spectral width $\Delta \nu \ll \nu_0$, the quantities $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ and $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$ vary slowly on the observation screen, and the interference fringes are mainly obtained due to the cosine term. Thus, defining the visibility of fringes as $v = (I_{\max} - I_{\min}) / (I_{\max} + I_{\min})$, we obtain

$$v = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \quad (15)$$

which shows that for maximum visibility, the two interfering fields must be completely coherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$). On the other hand, if the fields are completely incoherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 0$), no fringes are observed ($I_{\max} = I_{\min}$). The fields are said to be partially coherent when $0 < |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| < 1$. When $I_1 = I_2$, the visibility is the same as $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|$. The relation in Eq. (15) shows that in an interference experiment, one can obtain the modulus of the complex degree of coherence by measuring I_1 , I_2 , and the visibility. Similarly, Eq. (14) shows that from the measurement the positions of maxima, one can obtain the phase of the complex degree of coherence, $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$.

Temporal and Spatial Coherence

If the source illuminating the pinholes is a point source of finite spectral width situated at point Q , the fields at point P_1 and P_2 (Fig. 4) at any given instant are the same and the mutual coherence function becomes

$$\begin{aligned} \Gamma_{11}(\tau) &= \Gamma(\mathbf{r}_1, \mathbf{r}_1, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_1, t + \tau) \rangle \\ &= \langle V^*(\mathbf{r}_2, t) V(\mathbf{r}_2, t + \tau) \rangle = \Gamma_{22}(\tau) \end{aligned} \quad (16)$$

The self-coherence, $\Gamma_{11}(\tau)$, of the light from pinhole P_1 , is a direct measure of the temporal coherence of the source. On the other hand, if the source is of finite size and we observe the interference at point O which corresponds to $\tau = 0$, the mutual coherence function would be

$$\Gamma_{12}(0) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t) \rangle \equiv I_{12} \quad (17)$$

which is called the mutual intensity and is a direct measure of the spatial coherence of the source. In general, however, the function $\Gamma_{12}(\tau)$ measures, for a source of finite size and spectral width, a combination of the temporal and spatial coherence, and in only some limiting cases, the two types of coherence can be separated.

Spectral Representation of Mutual Coherence

One can also analyze the correlation between two fields in the spectral domain. In particular, one can define the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ which defines the correlation between the amplitudes of the spectral components of frequency ν of the light at the points P_1 and P_2 . Thus:

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu)\delta(\nu - \nu') = \langle V^*(\mathbf{r}_1, \nu)V(\mathbf{r}_2, \nu') \rangle \quad (18)$$

Using the generalized Wiener–Khinchine theorem, the cross-spectral density function can be shown to be the Fourier transform of the mutual coherence function:

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu) e^{-2\pi i \nu \tau} d\nu \quad (19)$$

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^\infty \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) e^{2\pi i \nu \tau} d\tau \quad (20)$$

If the two points P_1 and P_2 coincide (i.e., there is only one pinhole), the cross-spectral density function reduces to the spectral density function of the light, which we denote by $S(\mathbf{r}, \nu) [\equiv W(\mathbf{r}, \mathbf{r}, \nu)]$. Thus, it follows from Eq. (20) that the spectral density of light is the inverse Fourier transform of the self-coherence function. This leads to the Fourier transform spectroscopy, as we shall see later. One can also define spectral degree of coherence at frequency ν as

$$\begin{aligned} \mu(\mathbf{r}_1, \mathbf{r}_2, \nu) &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{W(\mathbf{r}_1, \mathbf{r}_1, \nu)W(\mathbf{r}_2, \mathbf{r}_2, \nu)}} \\ &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \nu)S(\mathbf{r}_2, \nu)}} \end{aligned} \quad (21)$$

It is easy to see that $0 \leq |\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)| \leq 1$. It is also sometimes referred to as the complex degree of coherence at frequency ν . It may be noted here that in the literature the notation $G^{(1,1)}(\nu) \equiv G(\mathbf{r}_1, \mathbf{r}_2, \nu)$ has also been used for $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$.

Propagation of Coherence

The van Cittert–Zernike Theorem

Perfectly coherent waves propagate through diffraction formulae, which have been discussed in this volume elsewhere. However, incoherent and partially coherent waves would evidently propagate somewhat differently. The earliest treatment of propagation noncoherent light is due to van Cittert, which was later generalized by Zernike to obtain what is now the van Cittert–Zernike theorem. The theorem deals with the correlations developed between fields at two points, which have been generated by a quasi-monochromatic and spatially incoherent planar source. Thus, as shown in Fig. 5, we consider a quasi-monochromatic ($\Delta\nu \ll \nu_0$) planar source σ , which has an intensity distribution $I(\mathbf{r}')$ on its plane and is spatially incoherent, i.e., there is no correlation between the fields at any two points on the source. The field, due to this source, would develop finite correlations after propagation, and the theorem states that

$$\begin{aligned} \gamma(\mathbf{r}_1, \mathbf{r}_2, 0) &= \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)\Gamma_{22}(0)}} \\ &= \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}} \int \int_\sigma I(\mathbf{r}') \frac{e^{2\pi i \nu_0 (R_2 - R_1)/c}}{R_1 R_2} d^2 r' \end{aligned} \quad (22)$$

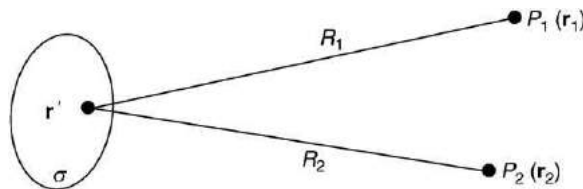


Fig. 5 Geometry for the van Cittert–Zernike theorem.

where $R_i = |\mathbf{r}_i - \mathbf{r}'|$ and $I(\mathbf{r}_i)$ is the intensity at point P_i . This relation is similar to the diffraction pattern produced at point P_1 due to a wave, with a spherical wavefront converging towards P_2 and with an amplitude distribution $I(\mathbf{r}')$ when it is diffracted by an aperture of the same shape, size, and the intensity distribution as those of the incoherent source σ . Thus, the theorem shows that a radiation, which was incoherent at the source, becomes partially coherent as it propagates.

Generalized Propagation

The expression $e^{-2\pi i \nu_0 R_1/c}/R_1$ can be interpreted as the field obtained at P_1 due to a point source located at the point $\mathbf{r}'$ on the planar source. Thus, this expression is simply the point spread function of the homogeneous space between the source and the observation plane containing the points P_1 and P_2 . Hopkins generalized this to include any linear optical system characterized by a point spread function $h(\mathbf{r}_j - \mathbf{r}')$ and obtained the formula for the complex degree of coherence of a radiation emerging from an incoherent, quasi-monochromatic planar source after it has propagated through such a linear optical system:

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}} \int \int_{\sigma} I(\mathbf{r}') h(\mathbf{r}_1 - \mathbf{r}') h^*(\mathbf{r}_2 - \mathbf{r}') d^2 r' \quad (23)$$

It would thus seem that the correlations propagate in much the same way, as does the field. Indeed, Wolf showed that the mutual correlation function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ satisfies the wave equations:

$$\nabla_1^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) \quad (24)$$

$$\nabla_2^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$$

where ∇_j^2 is the Laplacian with respect to the point $\mathbf{r}_j$. Here the first of Eqs. (24), for instance, gives the variation of the mutual coherence function with respect to $\mathbf{r}_1$ and τ for fixed $\mathbf{r}_2$. Further, the variable τ is the time difference defined through the path difference, since $\tau = (R_2 - R_1)/c$, and the actual time does not affect the mutual coherence function (as the fields are assumed to be stationary).

Using the relation in Eq. (20), one can also obtain from Eq. (24), the propagation equations for the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$:

$$\begin{aligned} \nabla_1^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) + k^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) &= 0 \\ \nabla_2^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) + k^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) &= 0 \end{aligned} \quad (25)$$

where $k = 2\pi\nu/c$.

Thompson and Wolf Experiment

One of the most elegant experiments for studying various aspects of coherence theory was carried out by Thompson and Wolf by making slight modifications in the Young's double-hole experiment. The experimental set-up shown in Fig. 6 consists of a quasi-monochromatic broad incoherent source S of diameter $2a$. This was obtained by focusing filtered narrow band light from a mercury lamp (not shown in the figure) on to a hole of size $2a$ in an opaque screen A . A mask consisting of two pinholes, each of diameter $2b$ with their axes separated by a distance d , was placed symmetrically about the optical axis of the experimental setup at plane B between two lenses L_1 and L_2 , each having focal length f . The source was at the front focal plane of the lens L_1 and the observations were made at the plane C at the back focus of the lens L_2 . The separation d was varied to study the spatial coherence function on the mask plane by measuring the visibility and the phase of the fringes formed in plane C .

Using the van Cittert–Zernike theorem, the complex degree of coherence at the pinholes P_1 and P_2 on the mask after the lens L_1 is

$$\gamma_{12} = |\gamma_{12}| e^{i\beta_{12}} = \frac{2J_1(\nu)}{\nu} \text{ with } \nu = \left(\frac{2\pi\nu a d}{cf} \right) \quad (26)$$

for symmetrically placed pinholes about the optical axis. Here β_{12} is the phase of the complex degree of coherence and in this special case where the two holes are equidistant from the axis, β_{12} is either zero or π , respectively, for positive or negative values of $2J_1(\nu)/\nu$. Let us assume that the intensities at two pinholes P_1 and P_2 are equal. The interference pattern observed at the back focal

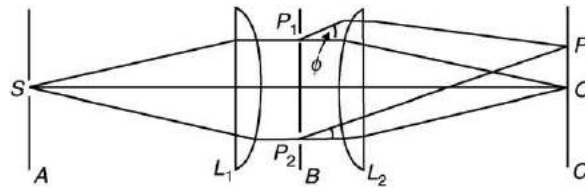


Fig. 6 Schematic of the Thompson and Wolf experiment.

plane of lens L_2 is due to the superposition of the light diffracted from the pinholes. The beams are partially coherent with degree of coherence given by Eq. (26). Since the pinholes are symmetrically placed, the intensity due to either of the pinholes at a point P at the back focal plane of the lens L_2 is the same and is given by the Fraunhofer formula for diffraction from circular apertures, i.e.:

$$I_1(P) = I_2(P) = \left| \frac{2J_1(u)}{u} \right|^2, \text{ with } u = \frac{2\pi v}{c} b \sin \phi \quad (27)$$

and ϕ is the angle that the diffracted beam makes from normal to the plane of the pinholes. The intensity of the interference pattern produced at the back focal plane of the lens L_2 is:

$$I(P) = 2I_1(P) \left[1 + \left| \frac{2J_1(v)}{v} \right|^2 \cos(\delta + \beta_{12}) \right] \quad (28)$$

where $\delta = d \sin \phi$ is the phase difference between the two beams reaching P from P_1 and P_2 . For the on-axis point O , the quantity δ is zero.

The maximum and minimum values of $I(P)$ are given by

$$I_{\max}(P) = 2I_1(P) [1 + |2J_1(v)/v|^2] \quad (29(a))$$

$$I_{\min}(P) = 2I_1(P) [1 - |2J_1(v)/v|^2] \quad (29(b))$$

Fig. 7 shows an example of the observed fringe patterns obtained by Thompson (1958) in a subsequent experiment.

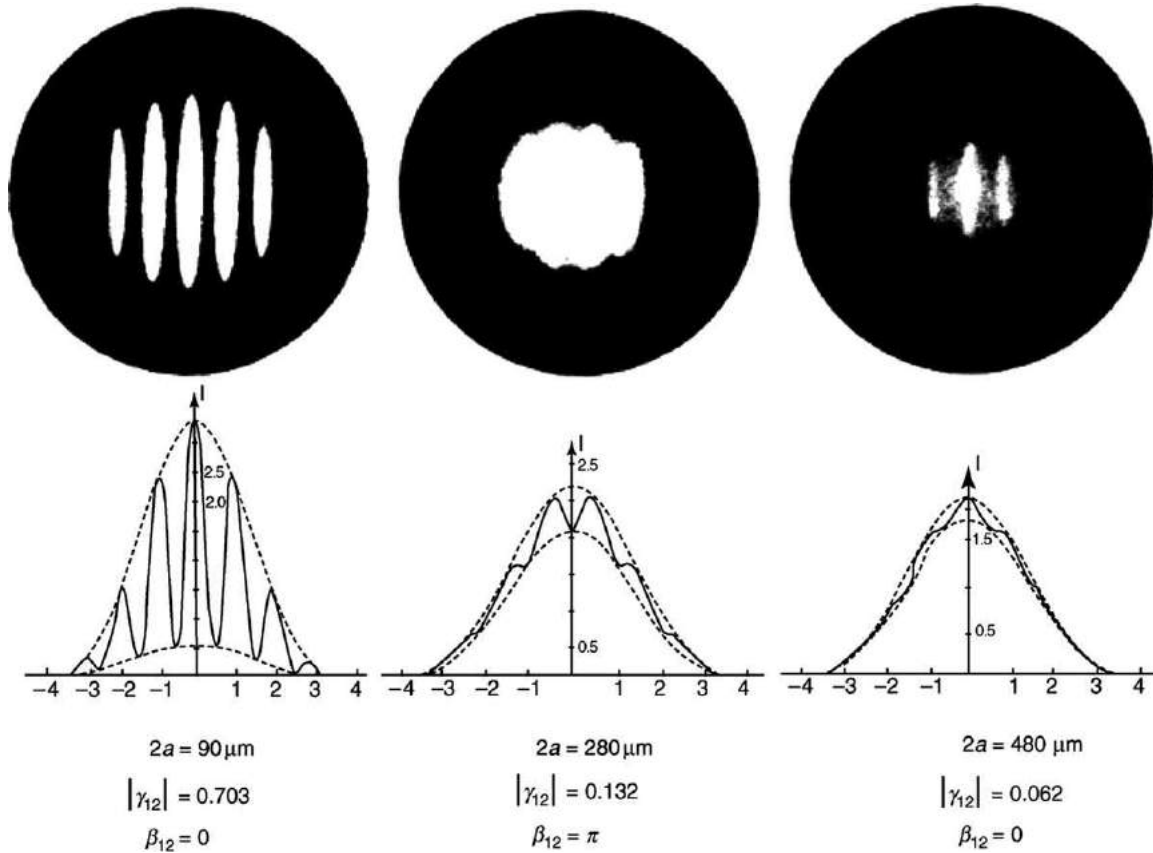


Fig. 7 Two beam interference patterns obtained by using partially coherent light. The wavelength of light used was $0.579 \mu\text{m}$ and the separation d of the pinholes in the screen at B was 0.5 cm . The figure shows the observed fringe pattern and the calculated intensity variation for three sizes of the secondary source. The dashed lines show the maximum and minimum intensity. The values of the diameter, $2a$, of the source and the corresponding values of the magnitude, $|\gamma_{12}|$, and the phase, β_{12} , are also shown in each case. Reproduced with permission from Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

Types of Fields

As mentioned above, $\gamma_{12}(\tau)$ is a measure of the correlation of the complex field at any two points P_1 and P_2 at specific time delay τ . Extending this definition, an optical field may be coherent or incoherent, if $|\gamma_{12}(\tau)|=1$ or $|\gamma_{12}(\tau)|=0$, respectively, for all pairs of points in the field and for all time delays. In the following, we consider some specific cases of fields and their properties.

Perfectly Coherent Fields

A field would be termed as perfectly coherent or self-coherent at a fixed point, if it has the property that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)|=1$ at some specific point $\mathbf{R}$ in the domain of the field for all values of time delay τ . It can also be shown that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)|$ is periodic in time, i.e.:

$$\gamma(\mathbf{R}, \mathbf{R}, \tau) = e^{-2\pi i \nu_0 \tau} \quad \text{for } \nu_0 > 0 \quad (30)$$

For any other point $\mathbf{r}$ in the domain of the field, $\gamma(\mathbf{R}, \mathbf{r}, \tau)$ and $\gamma(\mathbf{r}, \mathbf{R}, \tau)$ are also unimodular and also periodic in time.

A perfectly coherent optical field at two fixed points has the property that $|\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)|=1$ for any two fixed points $\mathbf{R}_1$ and $\mathbf{R}_2$ in the domain of the field for all values of time delay τ . For such a field, it can be shown that $\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \exp[i(\beta - 2\pi\nu_0\tau)]$ with ν_0 (> 0) and β are real constants. Further, as a corollary, $|\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau)|=1$ and $|\gamma(\mathbf{R}_2, \mathbf{R}_2, \tau)|=1$ for all τ , i.e., the field is self-coherent at each of the field points $\mathbf{R}_1$ and $\mathbf{R}_2$, and

$$\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau) = \gamma(\mathbf{R}_2, \mathbf{R}_2, \tau) = \exp(-2\pi i \nu_0 \tau) \text{ for } \nu_0 > 0$$

It can be shown that for any other point $\mathbf{r}'$ within the field

$$\gamma(\mathbf{r}, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0) \gamma(\mathbf{r}, \mathbf{R}_1, \tau) = \gamma(\mathbf{r}, \mathbf{R}_1, \tau) e^{i\beta} \quad (31)$$

$$\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0) \exp[-2\pi i \nu_0 \tau]$$

A perfectly coherent optical field for all points in the domain of the field has the property that $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|=1$ for all pairs of points $\mathbf{r}_1$ and $\mathbf{r}_2$ in the domain of the field, for all values of time delay τ . It can then be shown that $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \exp\{i[\alpha(\mathbf{r}_1) - \alpha(\mathbf{r}_2)] - 2\pi i \nu_0 \tau\}$, where $\alpha(\mathbf{r}')$ is real function of a single position $\mathbf{r}'$ in the domain of the field. Further, the mutual coherence function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ for such a field has a factorized periodic form as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = U(\mathbf{r}_1) U^*(\mathbf{r}_2) \exp(-2\pi i \nu_0 \tau) \quad (32)$$

which means the field is monochromatic with frequency ν_0 and $U(\mathbf{r})$ is any solution of the Helmholtz equation:

$$\nabla^2 U(\mathbf{r}) = k_0^2 U(\mathbf{r}) = 0, \quad k_0 = 2\pi\nu_0/c \quad (33)$$

The spectral and cross-spectral densities of such fields are represented by Dirac δ -functions with their singularities at some positive frequency ν_0 . Such fields, however, can never occur in nature.

Quasi-Monochromatic Fields

Optical fields are found in practice for which spectral spread $\Delta\nu$ is much smaller than the mean frequency $\bar{\nu}$. These are termed as quasi-monochromatic fields. The cross-spectral density $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ of the quasi-monochromatic fields attains an appreciable value only in the small region $\Delta\nu$, and falls to zero outside this region.

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = 0, \quad |\nu - \bar{\nu}| > \Delta\nu \quad \text{and} \quad \Delta\nu \ll \bar{\nu} \quad (34)$$

The mutual coherence function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ can be written as the Fourier transform of cross-spectral density function as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = e^{-2\pi i \bar{\nu} \tau} \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu) e^{-2\pi i (\nu - \bar{\nu}) \tau} d\nu \quad (35)$$

If we limit our attention to small τ such that $\Delta\nu|\tau| \ll 1$ holds, the exponential factor inside the integral is approximately equal to unity and Eq. (35) reduces to

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \exp(-2\pi i \bar{\nu} \tau) \int_0^\infty [W(\mathbf{r}_1, \mathbf{r}_2, \nu) d\nu] \quad (36)$$

which gives

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) \exp(-2\pi i \bar{\nu} \tau) \quad (37)$$

Eq. (37) describes the behavior of $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ for a limited range of τ values for quasi-monochromatic fields and in this range it behaves as a monochromatic field of frequency $\bar{\nu}$. However, due to the factor $\Gamma(\mathbf{r}_1, \mathbf{r}_2, 0)$ the quasi-monochromatic field may be coherent, partially coherent, or even incoherent.

Cross-Spectrally Pure Fields

The complex degree of coherence, if it can be factored into a product of a component dependent on spatial coordinates and a component dependent on time delay, is called reducible. In the case of perfectly coherent light, the complex degree of coherence is reducible, as we have seen above, e.g., in Eq. (32), and in the case of quasi-monochromatic fields, this is reducible approximately as

shown in Eq. (37). There also exists a very special case of a cross-spectrally pure field for which the complex degree of coherence is reducible. A field is called a cross-spectrally pure field if the normalized spectrum of the superimposed light is equal to the normalized spectrum of the component beams, a concept introduced by Mandel. In the space-frequency domain, the intensity interference law is expressed as the so-called spectral interference law:

$$S(\mathbf{r}, \nu) = S^{(1)}(\mathbf{r}, \nu) + S^{(2)}(\mathbf{r}, \nu) + 2[\sqrt{S^{(1)}(\mathbf{r}, \nu)}\sqrt{S^{(2)}(\mathbf{r}, \nu)}]\text{Re}[\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)e^{-2\pi i\nu\tau}] \quad (38)$$

where $\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)$ is the spectral degree of coherence, defined in Eq. (21) and τ is the relative time delay that is needed by the light from the pinholes to reach any point on the screen; $S^{(1)}(\mathbf{r}, \nu)$ and $S^{(2)}(\mathbf{r}, \nu)$ are the spectral densities of the light reaching P from the pinholes P_1 and P_2 , respectively (see Fig. 4) and it is assumed that the spectral densities of the field at pinholes P_1 and P_2 are the same [$S(\mathbf{r}_1, \nu) = CS(\mathbf{r}_2, \nu)$]. Now, if we consider a point for which the time delay is τ_0 , then it can be seen that the last term in Eq. (38) would be independent of frequency, provided that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)\exp(-2\pi i\nu\tau_0) = f(\mathbf{r}_1, \mathbf{r}_2, \tau_0) \quad (39)$$

where $f(\mathbf{r}_1, \mathbf{r}_2, \tau_0)$ is a function of $\mathbf{r}_1$, $\mathbf{r}_2$ and τ_0 only and the light at this point would have the same spectral density as that at the pinholes. If a region exists around the specified point on the observation plane, such that the spectral distribution of the light in this region is of the same form as the spectral distribution of the light at the pinholes, the light at the pinholes is cross-spectrally pure light.

In terms of the spectral distribution of the light $S(\mathbf{r}_1, \nu)$ at pinhole P_1 and $S(\mathbf{r}_2, \nu) = CS(\mathbf{r}_1, \nu)$ at pinhole P_2 , the mutual coherence function at the pinholes can be written as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \sqrt{C} \int S(\mathbf{r}_1, \nu)\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)\exp(-2\pi i\nu\tau)d\nu \quad (40)$$

and using Eq. (39), we get the very important condition for the field to be cross-spectrally pure, i.e.:

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\gamma(\mathbf{r}_1, \mathbf{r}_1, \tau - \tau_0) \quad (41)$$

The complex degree of coherence $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is thus expressible as the product of two factors: one factor characterizes the spatial coherence at the two pinholes at time delay τ_0 and the other characterizes the temporal coherence at one of the pinholes. Eq. (41) is known as the reduction formula for cross-spectrally pure light. It can further be shown that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\exp(2\pi i\nu\tau_0) \quad (42)$$

Thus, the absolute value of the spectral degree of coherence is the same for all frequencies and is equal to the absolute value of the degree of coherence for the point specified by τ_0 . It has been shown that cross-spectrally pure light can be generated, for example, by linear filtering of light that emerges from the two pinholes in Young's interference experiments.

Types of Sources

Primary Sources

A primary source is a set of radiating atoms or molecules. In a primary source the randomness comes from true source fluctuations, i.e., from the spatial distributions of fluctuating charges and currents. Such a source gives rise to a fluctuating field. Let $Q(\mathbf{r}, t)$ represent the fluctuating source variable at any point $\mathbf{r}$ at time t , then the field generated by the source is represented by fluctuating field variable $V(\mathbf{r}, t)$. The source is assumed to be localized in some finite domain such that $Q(\mathbf{r}, t) = 0$ at any time $t > 0$ outside the domain. Assuming that field variable $V(\mathbf{r}, t)$ and the source variable $Q(\mathbf{r}, t)$ are scalar quantities, they are related by an inhomogeneous equation as

$$\nabla^2 V(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} V(\mathbf{r}, t) = -4\pi Q(\mathbf{r}, t) \quad (43)$$

The mutual coherence functions of the source $\Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle Q^*(\mathbf{r}_1, t)Q(\mathbf{r}_2, t + \tau) \rangle$ and of the field $\Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle V^*(\mathbf{r}_1, t)V(\mathbf{r}_2, t + \tau) \rangle$ characterize the statistical similarity of the fluctuating quantities at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. Following Eq. (20), one can define, respectively, the cross-spectral density function of the source and the field as

$$W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau)e^{-2\pi i\nu\tau}d\tau \quad (44)$$

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau)e^{-2\pi i\nu\tau}d\tau$$

The cross-spectral density functions of the source and of the field are related as

$$(\nabla_2^2 + k^2)(\nabla_1^2 + k^2)W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = 4\pi^2 W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) \quad (45)$$

The solution of Eq. (45) is represented as

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) \frac{e^{ik(R_2 - R_1)}}{R_1 R_2} d^3 r'_1 d^3 r'_2 \quad (46)$$

where $R_1 = |\mathbf{r}_1 - \mathbf{r}'_1|$ and $R_2 = |\mathbf{r}_2 - \mathbf{r}'_2|$ (see Fig. 8). Using Eq. (46), one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u}$) in the far field ($r \gg r'_1, r'_2$) of a source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{1}{r^2} \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) e^{-iku \cdot (\mathbf{r}'_1 - \mathbf{r}'_2)} d^3 r'_1 d^3 r'_2 \quad (47)$$

where $\mathbf{u}$ is the unit vector along $\mathbf{r}$. The integral in Eq. (47), i.e., the quantity defined by $r^2 S^\infty(r\mathbf{u}, \nu)$, is also defined as the radiant intensity, which represents the rate of energy radiated at the frequency ν from the source per unit solid angle in the direction of $\mathbf{u}$.

Secondary Sources

Sources used in a laboratory are usually secondary planar sources. A secondary source is a field, which arises from the primary source in the region outside the domain of the primary sources. This kind of source is an aperture on an opaque screen illuminated either directly or via an optical system by primary sources. Let $V(\rho, t)$ represent the fluctuating field in a secondary source plane σ at $z=0$ (Fig. 9) and $W_0(\rho_1, \rho_2, \nu)$ represent its cross-spectral density (the subscript 0 refers to $z=0$). One can then solve Eq. (25) to obtain the propagation of the cross-spectral density from this planar source. For two points P_1 and P_2 located at distances which are large compared to wavelength, the cross-spectral density is then given by

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \left(\frac{k}{2\pi} \right)^2 \int_\sigma \int_\sigma W_0(\rho_1, \rho_2, \nu) \frac{e^{ik(R_2 - R_1)}}{R_1 R_2} \cos\theta'_1 \cos\theta'_2 d^2 \rho_1 d^2 \rho_2 \quad (48)$$

where $R_j = |\mathbf{r}_j - \rho_j|$, ($j=1,2$), θ'_1 and θ'_2 are the angles that R_1 and R_2 directions make with the z -axis (Fig. 9). Using Eq. (48), one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u}$) in the far field ($r \gg \rho_1, \rho_2$) of a planar source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2 \cos^2 \theta}{2\pi^2 r^2} \int_\sigma \int_\sigma W_0(\rho_1, \rho_2, \nu) e^{-iku_\perp \cdot (\rho_1 - \rho_2)} d^2 \rho_1 d^2 \rho_2 \quad (49)$$

where $\mathbf{u}_\perp$ is the projection of the unit vector $\mathbf{u}$ on the plane σ of the source.

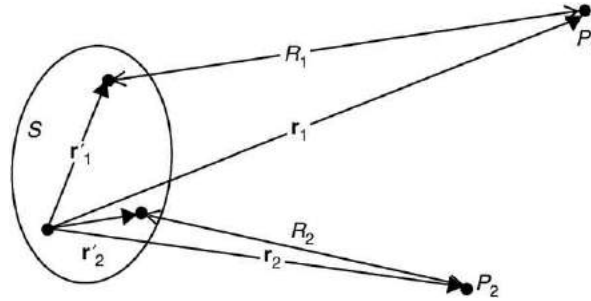


Fig. 8 Geometry of a 3-dimensional primary source S and the radiation from it.

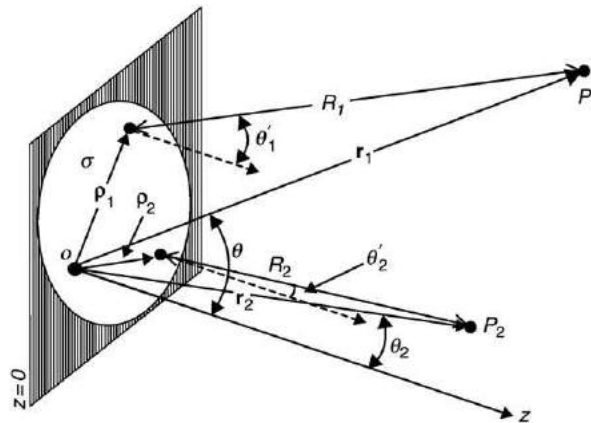


Fig. 9 Geometry of a planar source σ and the radiation from it.

Schell-Model Sources

In the framework of coherence theory in space-time domain, two-dimensional planar model sources of this kind were first discussed by Schell. Later, the model was adopted for formulation of coherence theory in space-frequency domain. Schell-model sources are the sources whose degree of spectral coherence $\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu)$ (for either primary or secondary source) is stationary in space. It means that $\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu)$ depends on $\mathbf{r}_1$ and $\mathbf{r}_2$ only through the difference $\mathbf{r}_2 - \mathbf{r}_1$, i.e., of the form:

$$\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu) \equiv \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) \quad (50)$$

for each frequency ν present in the source spectrum. Here A stands for field variables V , in the case of a Schell-model secondary source and for source variables Q , in the case of a Schell-model primary source. The cross-spectral density function of a Schell-model source is of the form:

$$W_A(\mathbf{r}_1, \mathbf{r}_2, \nu) = [S_A(\mathbf{r}_1, \nu)]^{1/2} [S_A(\mathbf{r}_2, \nu)]^{1/2} \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) \quad (51)$$

where $S_A(\mathbf{r}, \nu)$ is the spectral density of the light at a typical point in the primary source or on the plane of a secondary source. Schell model sources do not assume low coherence, and therefore, can be applied to spatially stationary light fields of any state of coherence. The Schell-model of the form given in Eq. (51) has been used to represent both three-dimensional primary sources and two-dimensional secondary sources.

Quasi-Homogeneous Sources

Useful models of partially coherent sources that are frequently encountered in nature or in the laboratory are the so-called quasi-homogeneous sources. These are an important sub-class of Schell-model sources. A Schell-model source is called quasi-homogeneous if the intensity of a Schell model source is essentially constant over any coherence area. Under these approximations, the cross-spectral density function for a quasi-homogeneous source is given by

$$\begin{aligned} W_A(\mathbf{r}_1, \mathbf{r}_2, \nu) &= S_A\left[\frac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2), \nu\right] \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) \\ &= S_A[\mathbf{r}, \nu] \mu_A(\mathbf{r}', \nu) \end{aligned} \quad (52)$$

where $\mathbf{r} = (\mathbf{r}_1 + \mathbf{r}_2)/2$, and $\mathbf{r}' = \mathbf{r}_2 - \mathbf{r}_1$. The subscript A stands for either V or Q for the field variable or a source variable, respectively. It is clear that for a quasi-homogeneous source the spectral density $S_A(\mathbf{r}, \nu)$ varies so slowly with position that it is approximately constant over distances across the sources that are of the order of the correlation length Δ , which is a measure of the effective width of $|\mu_A(\mathbf{r}', \nu)|$. Therefore, $S_A(\mathbf{r}, \nu)$ is a slowly varying function of $\mathbf{r}$ (Fig. 10(b)) and $|\mu_A(\mathbf{r}', \nu)|$ is a fast varying function of $\mathbf{r}'$ (Fig. 10(a)). In addition, the linear dimensions of the source are large compared with the wavelength of light and with the correlation length Δ (Fig. 10(c)).

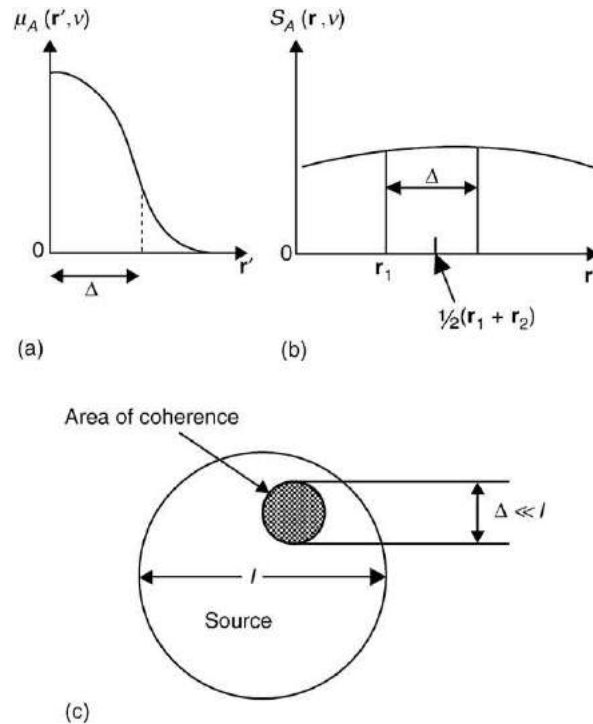


Fig. 10 Concept of quasi-homogeneous sources.

Quasi-homogeneous sources are always spatially incoherent in the 'global' sense, because their linear dimensions are large compared with the correlation length. This model is very good for representing two-dimensional secondary sources with sufficiently low coherence such that the intensity does not vary over the coherence area on the input plane. It has also been applied to three-dimensional primary sources, three-dimensional scattering potentials, and two-dimensional primary and secondary sources.

Equivalence Theorems

The study of partially coherent sources led to the formulations of a number of equivalence theorems, which show that sources of any state of coherence can produce the same distribution of the radiant intensity as a fully spatially coherent laser source. These theorems provide conditions under which sources of different spatial distribution of spectral density and of different state of coherence will generate fields, which have the same radiant intensity. It has been shown, by taking examples of Gaussian-Schell model sources, that sources of completely different coherence properties and different spectral distributions across the source generate identical distribution of radiant intensity. Experimental verifications of the results of these theorems have also been carried out. For further details on this subject, the reader is referred to Mandel and Wolf (1995).

Correlation-Induced Spectral Changes

It was assumed that spectrum is an intrinsic property of light that does not change as the radiation propagates in free-space, until studies on partially coherent sources and radiations from them in 1980s, revealed that this was true only for specific type of sources. It was discovered on general grounds that the spectrum of light, which originates in an extended source, either a primary source or a secondary source, depends not only on the source spectrum but also on the spatial coherence properties of the source. It was also predicted theoretically by Wolf that the spectrum of light would, in general, be different from the spectrum of the source, and be different at different points in space on propagation in free space.

For a quasi-homogeneous planar secondary source defined by Eq. (52), whose normalized spectrum is the same at each source point, one can write the spectral density as

$$S_0(\mathbf{p}, \nu) = I_0(\mathbf{p})g_0(\nu) \quad \text{with} \quad \int_0^\infty g_0(\nu)d\nu = 1 \quad (53)$$

where $I_0(\mathbf{p})$ is the intensity of light at point $\mathbf{p}$ on the plane of the source, $g_0(\nu)$ is the normalized spectrum of the source and the subscript 0 refers to the quantities of the source plane. Using Eq. (49), one can obtain an expression for the far-field spectrum due to this source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2 \cos^2 \theta}{2\pi^2 r^2} \int_\sigma \int_\sigma I_0(\mathbf{p})g_0(\nu)\mu_0(\mathbf{p}', \nu)e^{-ik\mathbf{u}_\perp \cdot (\mathbf{p}_1 - \mathbf{p}_2)} d^2 \rho_1 d^2 \rho_2 \quad (54)$$

Noting that $\mathbf{p} = (\mathbf{p}_1 + \mathbf{p}_2)/2$ and $\mathbf{p}' = \mathbf{p}_2 - \mathbf{p}_1$, one can transform the variables of the integration and obtain after some manipulation:

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2 \cos^2 \theta}{(2\pi r)^2} \tilde{I}_0 \tilde{\mu}_0(k\mathbf{u}_\perp, \nu)g_0(\nu) \quad (55)$$

where

$$\tilde{\mu}_0(k\mathbf{u}_\perp, \nu) = \int_\sigma \mu_0(\mathbf{p}', \nu)e^{-ik\mathbf{u}_\perp \cdot \mathbf{p}'} d^2 \rho' \quad (56)$$

and

$$\tilde{I}_0 = \int_\sigma I_0(\mathbf{p})d^2 \rho \quad (57)$$

Eq. (55) shows that the spectrum of the field in the far-zone depends on the coherence properties of the source through its spectral degree of coherence $\mu_0(\mathbf{p}', \nu)$ and on the normalized source spectrum $g_0(\nu)$.

Scaling Law

The reason why coherence-induced spectral changes were not observed until recently is that the usual thermal sources employed in laboratories or commonly encountered in nature have special coherence properties and the spectral degree of coherence has the function form:

$$\mu^{(0)}(\mathbf{p}_2 - \mathbf{p}_1, \nu) = f[k(\mathbf{p}_2 - \mathbf{p}_1)] \quad \text{with} \quad k = \frac{2\pi\nu}{c} \quad (58)$$

which shows that the spectral degree of coherence depends only on the product of the frequency and space coordinates. This formula expresses the so-called scaling law, which was enunciated by Wolf. Commonly used sources satisfy this property.

For example, the spectral degree of coherence of Lambertian sources and black-body sources can be shown to be

$$\mu_0(\mathbf{p}_2 - \mathbf{p}_1, \nu) = \frac{\sin(k|\mathbf{p}_2 - \mathbf{p}_1|)}{k|\mathbf{p}_2 - \mathbf{p}_1|} \quad (59)$$

This expression evidently satisfies the scaling law. If the spectral degree of coherence does not satisfy the scaling law, the normalized far-field spectrum will, in general, vary in different directions in the far-zone and will differ from the source spectrum.

Spectral Changes in Young's Interference Experiment

Spectral changes in Young's interference experiment with broadband light are not as well understood as in experiments with quasi-monochromatic, probably because in such experiments no interference fringes are formed. However, if one were to analyze the spectrum of the light in the region of superposition, one would observe changes in the spectrum of light in the region of superposition in the form of a shift in the spectrum for narrowband spectrum and spectral modulation for broadband light. One can readily derive an expression for the spectrum of light in the region of superposition. Let $S^{(1)}(P, \nu)$ be the spectral density of the light at P which would be obtained if the small aperture at P_1 alone was open; $S^{(2)}(P, \nu)$ has a similar meaning if only the aperture at P_2 was open (Fig. 4).

Let us assume, as is usually the case, that $S^{(2)}(P, \nu) \approx S^{(1)}(P, \nu)$ and let d be the distance between the two pinholes. Consider the spectral density at the point P , at distance x from the axis of symmetry in an observation plane located at distance of R from the plane containing the pinholes. Assuming that $x/R \ll 1$, one can make the approximation $R_2 - R_1 \approx xd/R$. The spectral interference law (Eq. (38)) can then be written as

$$S(P, \nu) \approx 2S^{(1)}(P, \nu) \{1 + |\mu(P_1, P_2, \nu)| \cos[\beta(P_1, P_2, \nu) + 2\pi\nu xd/cR]\} \quad (60)$$

where $\beta(P_1, P_2, \nu)$ denotes the phase of the spectral degree of coherence. Eq. (60) implies the two results:

- (i) at any fixed frequency ν , the spectral density varies sinusoidally with the distance x of the point from the axis, with the amplitude and the phase of the variation depending on the (generally complex) spectral degree of coherence $\mu(P_1, P_2, \nu)$; and
- (ii) at any fixed point P in the observation plane the spectrum $S(P, \nu)$ will, in general, differ from the spectrum $S^{(1)}(P, \nu)$, the change also depending on the spectral degree of coherence $\mu(P_1, P_2, \nu)$ of the light at the two pinholes.

Experimental Confirmations

Experimental tests of the theoretical prediction of spectral invariance and noninvariance due to correlation of fluctuations across the source were performed just after the theoretical predictions. Fig. 11 shows results of one such experiment in which spectrum changes in the Young's experiment were studied. Several other experiments also reported confirmation of the source correlation-dependent spectral changes. One of the important applications of these observations has been to explain the discrepancies in the maintenance of the spectroradiometric scales by national laboratories in different countries. These studies also have potential application in determining experimentally the spectral degree of coherence of partially coherent fields. The knowledge of spectral degree of coherence is often important in remote sensing, e.g., for determining angular diameters of stars.

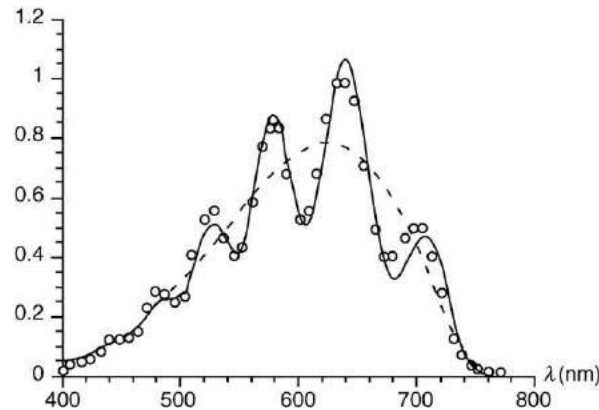


Fig. 11 Correlation-induced changes in spectrum in Young's interference. Dashed line represents the spectrum when only one of the slits is open, the continuous curve shows the spectrum when both the slits are open and the circles are the measured values in the latter case. Reproduced with permission from Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

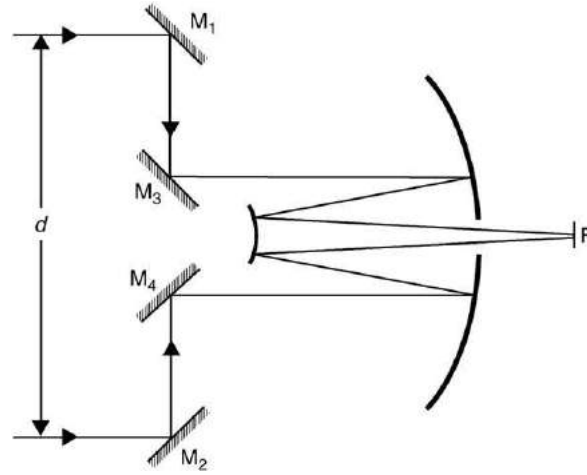


Fig. 12 Schematic of the Michelson stellar interferometer.

Applications of Optical Coherence

Stellar Interferometry

The Michelson stellar interferometer, named after Albert Michelson, was used to determine the angular diameters of stars and also the intensity distribution across the star. The method was devised by Michelson without using any concept of coherence, although subsequently the full theory of the method was developed on the basis of propagation of correlations. A schematic of the experiment is shown in **Fig. 12**. The interferometer is mounted in front of a telescope, a reflecting telescope in this case. The light from a star is reflected from mirrors M_1 and M_2 and is directed towards the primary mirror (or the objective lens) of the telescope. The two beams thus collected superpose in the focal plane F of telescope where an image crossed with fringes is formed. The outer mirrors M_1 and M_2 can be moved along the axis defined as $M_1M_3M_4M_2$ while the inner mirrors M_3 and M_4 remain fixed. The fringe spacing depends on the position of mirrors M_3 and M_4 and hence is fixed, while the visibility of the fringes depends on the separation of the mirrors M_1 and M_2 and hence, can be varied. Michelson showed that from the measurement of the variation of the visibility with the separation of the two mirrors, one could obtain information about the intensity distribution of the stars, which are rotationally symmetric. He also showed that if the stellar disk is circular and uniform, the visibility curve as a function of the separation d of the mirrors M_1 and M_2 will have zeros for certain values of d , and that the smallest of these d values for which zero occurs is given by $d_0 = 0.61 \lambda_a / \alpha$, where α is the semi-angular diameter of the star and λ_a is the mean wavelength of the filtered quasi-monochromatic light from the star. Angular diameters of several stars down to 0.02 second of an arc were determined.

From the standpoint of second-order coherence theory the principles of the method can be readily understood. The star is considered an incoherent source and according to the van Cittert–Zernike theorem, the light reaching the outer mirrors M_1 and M_2 of the interferometer will be partially coherent. This coherence would depend on the size of and the intensity distribution across the star. Let (x_1, y_1) and (x_2, y_2) be the coordinates of the positions of the mirrors M_1 and M_2 , respectively, and (ξ, η) the coordinates of a point on the surface plane of the star which is assumed to be at a very large (astronomical) distance R from the mirrors. The complex degree of coherence at the mirrors would then be given by **Eq. (22)** which can now be written as

$$\gamma(\Delta x, \Delta y, 0) = \frac{\int_{\sigma} I(u, v) e^{-ik_a(u\Delta x + v\Delta y)} du dv}{\int_{\sigma} I(u, v) du dv} \quad (61)$$

where $I(u, v)$ is the intensity distribution across the star disk σ as a function of the angular coordinates $u = \xi/R$, $v = \eta/R$, $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, and $k_a = 2\pi/\lambda_a$, λ_a being the mean wavelength of the light from the star. **Eq. (61)** shows that the equal-time ($\tau=0$) complex degree of coherence of the light incident at the outer mirrors of the interferometer is the normalized Fourier-transform of the intensity distribution across the stellar disk. Further, **Eq. (15)** shows that the visibility of the interference fringes is the absolute value of γ , if the intensity of the two interfering beams is equal, as in the present case. The phase of γ can be determined by the position of the intensity maxima of the fringe pattern (**Eq. (14)**). If one is interested in determining only the angular size of the star and the star is assumed to be a circularly symmetric disk of angular diameter 2α and of uniform intensity [$I(u, v)$ is constant across the disk], then **Eq. (61)** reduces to

$$\gamma(\Delta x, \Delta y) = \frac{2J_1(v)}{v}, \quad v = \frac{2\pi\alpha}{\lambda_a} d, \quad d = \sqrt{(\Delta x)^2 + (\Delta y)^2} \quad (62)$$

The smallest separation of the mirrors for which the visibility γ vanishes corresponds to $v=3.832$, i.e., $d_0 = 0.61 \lambda_a / \alpha$, which is in agreement with Michelson's result.

Interference Spectroscopy

Another contribution of Michelson, which was subsequently identified as an application of the coherence theory, was the use of his interferometer (Fig. 2) to determine the energy distribution in the spectral lines. The method he developed is capable to resolving spectral lines that are too narrow to be analyzed by the spectrometer. The visibility of the interference fringes depends on the energy distribution in the spectrum of the light and its measurement can give information about the spectral lines. In particular, if the energy distribution in a spectral line is symmetric about some frequency ν_0 , its profile is simply the Fourier transform of the visibility variation as a function of the path difference between the two interfering beams. This method is the basis of interference spectroscopy or the Fourier transform spectroscopy.

Within the framework of second-order coherence theory, if the mean intensity of the two beams in a Michelson's interferometer is the same, then the visibility of the interference fringes in the observation plane is related to the complex degree of coherence of light at a point at the beamsplitter where the two beams superimpose (Eq. (15)). The two quantities are related as $v(\tau) = |\gamma(\tau)|$ where $\gamma(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_1, \tau)$ is the complex degree of self-coherence at the point $\mathbf{r}_1$ on the beamsplitter. Following Eq. (19), $\gamma(\tau)$ can be represented by a Fourier integral as

$$\gamma(\tau) = \int_0^\infty s(\nu) \exp(-i2\pi\nu\tau) d\nu \quad (63)$$

where $s(\nu)$ is the normalized spectral density of the light defined as $s(\nu) = S(\nu) / \int_0^\infty S(\nu) d\nu$ and $S(\nu) \equiv S(\mathbf{r}_1, \nu) = W(\mathbf{r}_1, \mathbf{r}_1, \nu)$ is the spectral density of the beam at the point $\mathbf{r}_1$. The method is usually applied to very narrow spectral lines for which one can assume that the peak that occurs at ν_0 , and $\gamma(\tau)$ can be represented as

$$\begin{aligned} \gamma(\tau) &= \bar{\gamma}(\tau) \exp(-2\pi i \nu_0 \tau) \quad \text{with} \\ \bar{\gamma}(\tau) &= \int_{-\infty}^{\infty} \bar{s}(\mu) \exp(-i2\pi\mu\tau) d\mu \end{aligned} \quad (64)$$

where $\bar{s}(\mu)$ is the shifted spectrum such that

$$\begin{aligned} \bar{s}(\mu) &= s(\nu_0 + \mu) \quad \mu \geq -\nu_0 \\ &= 0 \quad \mu < -\nu_0 \end{aligned}$$

From the above one readily gets

$$v(\tau) = |\gamma(\tau)| = \left| \int_{-\infty}^{\infty} \bar{s}(\mu) \exp(-i2\pi\mu\tau) d\mu \right| \quad (65)$$

If the spectrum is symmetric about ν_0 , then $v(\tau)$ would be an even function of τ and the Fourier inversion would give:

$$\bar{s}(\mu) = s(\nu_0 + \mu) = 2 \int_0^\infty v(\tau) \cos(2\pi\mu\tau) d\tau \quad (66)$$

which can be used to calculate the spectral energy distribution for a symmetric spectrum about ν_0 from the visibility curve. However, for an asymmetric spectral distribution, the visibility and the phase of the complex degree of coherence must be determined as the Fourier transform of the shifted spectrum is no longer real everywhere.

Higher-Order Coherence

So far we have considered correlations of the fluctuating field variables at two space-time $(\mathbf{r}, t)$ points, as defined in Eq. (10). These are termed as second-order correlations. One can extend the concept of correlations to more than two space-time points, which will involve higher-order correlations. For example, one can define the space-time cross correlation function of order (M, N) of the random field $V(\mathbf{r}, t)$, represented by $\Gamma^{(M, N)}$, as an ensemble average of the product of the field $V(\mathbf{r}, t)$ values at N space-time points and $V^*(\mathbf{r}, t)$ at other M points. In this notation, the mutual coherence function as defined in Eq. (10) would now be $\Gamma^{(1, 1)}$. Among higher-order correlations, the one with $M = N = 2$, is of practical significance and is called the fourth-order correlation function, $\Gamma^{(2, 2)}(\mathbf{r}_1, t_1, \mathbf{r}_2, t_2; \mathbf{r}_3, t_3, \mathbf{r}_4, t_4)$. The theory of Gaussian random variables tells us that any higher correlation can be written in terms of second-order correlations over all permutations of pairs of points. In addition, if we assume that $(\mathbf{r}_3, t_3) = (\mathbf{r}_1, t_1)$ and $(\mathbf{r}_4, t_4) = (\mathbf{r}_2, t_2)$, and that the field is stationary, then $\Gamma^{(2, 2)}$ is called the intensity-intensity correlation and is given as

$$\begin{aligned} \Gamma^{(2, 2)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) &= \langle V(\mathbf{r}_1, t_1) V(\mathbf{r}_2, t_2) V^*(\mathbf{r}_1, t_1) V^*(\mathbf{r}_2, t_2) \rangle \\ &= \langle I(\mathbf{r}_1, t_1) I(\mathbf{r}_2, t_2) \rangle \\ &= \langle I(\mathbf{r}_1, t_1) \rangle \langle I(\mathbf{r}_2, t_2) \rangle (1 + |\gamma^{(1, 1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2) \end{aligned} \quad (67)$$

where

$$\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) = \frac{\Gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)}{[\langle I(\mathbf{r}_1, t_1) \rangle]^{1/2} [\langle I(\mathbf{r}_2, t_2) \rangle]^{1/2}} \quad (68)$$

We now define fluctuations in intensity at $(\mathbf{r}_j, t_j)$ as

$$\Delta I_j = I(\mathbf{r}_j, t_j) - \langle I(\mathbf{r}_j, t_j) \rangle$$

and then the correlation of intensity fluctuations becomes

$$\begin{aligned} \langle \Delta I_1 \Delta I_2 \rangle &= \langle I(\mathbf{r}_1, t_1) I(\mathbf{r}_2, t_2) \rangle - \langle I(\mathbf{r}_1, t_1) \rangle \langle I(\mathbf{r}_2, t_2) \rangle \\ &= \langle I(\mathbf{r}_1, t_1) \rangle \langle I(\mathbf{r}_2, t_2) \rangle |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2 \end{aligned} \quad (69)$$

where we have used Eq. (67). Eq. (69) forms the basis for intensity interferometry.

Hanbury-Brown and Twiss Experiment

In this landmark experiment conducted both on the laboratory scale and astronomical scale, Hanbury-Brown and Twiss demonstrated the existence of intensity-intensity correlations in terms of the correlations in the photocurrents in the two detectors and thus measured the squared modulus of complex degree of coherence. In the laboratory experiment, the arc of a mercury lamp was focused onto a circular hole to produce a secondary source. The light from this source was then divided equally into two parts through a beamsplitter. Each part was, respectively, detected by two photomultiplier tubes, which had identical square apertures in front. One of the tubes could be translated normally to the direction of propagation of light and was so positioned that its image through the splitter could be made to coincide with the other tube. Thus, by suitable translation a measured separation d between the two square apertures could be introduced. The output currents from the photomultiplier tubes were taken by cables of equal length to a correlator. In the path of each cable a high-pass filter was inserted, so that only the current fluctuations could be transmitted to the correlator. Thus, the normalized correlations between the two current fluctuations:

$$C(d) = \frac{\langle \Delta J_1(t) \Delta J_2(t) \rangle}{\langle [\Delta J_1(t)]^2 \rangle^{1/2} \langle [\Delta J_2(t)]^2 \rangle^{1/2}} \quad (70)$$

as a function of detector separation d could be measured. Now when the detector response time is much larger than the time-scale of the fluctuations in intensity, then it can be shown that the correlations in the fluctuations of the photocurrent are proportional to the correlations of the fluctuations in intensity of the light being detected. Thus, we would have

$$C(d) \approx \delta |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, 0)|^2 \quad (71)$$

where δ is the average number of photocounts of the light of one polarization during the time-scale of the fluctuations (for general thermal sources, this is much less than one). Eq. (71) represents the Hanbury-Brown-Twiss effect.

Stellar Intensity Interferometry

Michelson stellar interferometry can resolve stars which have angular sizes of the order of $0.01''$, since for smaller stars, the separation between the primary mirrors runs into several meters and maintaining stability of mirrors such that the optical paths do not change, even by fraction of a wavelength, is extremely difficult. The atmospheric turbulence further adds to this problem and obtaining stable fringe pattern becomes next to impossible for very small stars. Hanbury-Brown and Twiss applied the intensity interferometry based on their photoelectric correlation technique for determining the angular sizes of such stars. Two separate parabolic mirrors collected light from the star and the output of the photodetectors placed at the focus of each mirror was sent to a correlator. The cable lengths were made unequal so as to compensate for the time difference of the light arrival at the two mirrors. The normalized correlation of the fluctuations of the photocurrents was determined as described above. This would give the variation of the modulus-squared degree of coherence as a function of the mirror separation d from which the angular size of the stars can be estimated. The advantage of the stellar intensity interferometer over the stellar (amplitude) interferometer is that the light need not interfere as in the latter, since the photodetectors are mounted directly at the focus of the primary mirrors of the telescope. Thus, the constraint on the large path difference between the two beams is removed and large values of d can now be used. Moreover, the atmosphere turbulence and the mirror movements have very small effect. Stellar angular diameters as small as $0.0004''$ of arc with resolution of $0.00003''$ could be measured by such interferometers.

Further Reading

- Beran, M.J., Parrent, G.B., 1964. *Theory of Partial Coherence*. Englewood Cliffs, NJ: Prentice-Hall.
 Born, M., Wolf, E., 1999. *Principles of Optics*. New York: Pergamon Press.
 Carter, W.H., 1996. *Coherence theory*. In: Bass, M. (Ed.), *Hand Book of Optics*. New York: McGraw-Hill.
 Davenport, W.B., Root, W.L., 1960. *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill.
 Goodman, J.W., 1985. *Statistical Optics*. Chichester: Wiley.

- Hanbury-Brown, R., Twiss, R.Q., 1957. Interferometry of intensity fluctuations in light: I basic theory: the correlation between photons in coherent beams of radiation. *Proceedings of the Royal Society* 242, 300–324.
- Hanbury-Brown, R., Twiss, R.Q., 1957. Interferometry of intensity fluctuations in light: II an experimental test of the theory for partially coherent light. *Proceedings of the Royal Society* 243, 291–319.
- Kandpal, H.C., Vaishya, J.S., Joshi, K.C., 1994. Correlation-induced spectral shifts in optical measurements. *Optical Engineering* 33, 1996–2012.
- Mandel, L., Wolf, E., 1965. Coherence properties of optical fields. *Review of Modern Physics* 37, 231–287.
- Mandel, L., Wolf, E., 1995. *Optical Coherence and Quantum Optics*. Cambridge: Cambridge University Press.
- Marathay, A.S., 1982. *Elements of Optical Coherence*. New York: John Wiley.
- Perina, J., 1985. *Coherence of Light*. Boston: M.A. Reidel.
- Santarsiero, M., Gori, F., 1992. Spectral changes in Young interference pattern. *Physics Letters* 167, 123–128.
- Schell, A.C., 1967. A technique for the determination of the radiation pattern of a partially coherent aperture. *IEEE Transactions on Antennas and Propagation* AP-15, 187.
- Thompson, B.J., 1958. Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48, 95–97.
- Thompson, B.J., Wolf, E., 1957. Two beam interference with partially coherent light. *Journal of the Optical Society of America* 47, 895–902.
- Wolf, E., James, D.F.V., 1996. Correlation-induced spectral changes. *Report on Progress in Physics* 59, 771–818.

Diffraction Gratings

J Turunen and T Vallius, University of Joensuu, Joensuu, Finland

© 2018 Elsevier Inc. All rights reserved.

Nomenclature

U	Complex field amplitude
θ, ϕ	Diffraction angles [rad]
η	Diffraction efficiency
f, f_0	Focal length [m]
d	Grating period [m]

H	Grating thickness [m]
Q	Number of quantization levels
$\phi(r)$	Phase function [rad]
n	Refractive index
λ, λ_0	Wavelength [m]
k	Wave vector [m^{-1}]

Introduction

The classic textbook example of a diffraction grating is a periodic arrangement of parallel opaque wires: if a plane wave is incident upon such a structure, it is divided into a multitude of plane waves propagating in well-defined directions. In particular, the dramatic wavelength-dependence of these directions has attracted great interest ever since the potential of diffraction gratings in spectroscopy was realized in the late 1700s. Today spectroscopy is only one, though important, application of diffraction gratings. Gratings of many different forms belong to the basic building blocks in modern wave-optics-based design of optical elements and systems.

Grating Equations

Some examples of the wide variety of possible grating structures are illustrated in Fig. 1, which shows the profiles of some commonly encountered diffraction gratings that are invariant in the y direction. These are known as linear gratings. Each grating may be of reflection type (the refractive index n_{III} is complex) or of transmission type (n_{III} is real). In all cases n_1 and n_2 can both be real or one of them can be complex. Here we assume that the incident field is a plane wave with wave vector k in the x - z plane; more general illumination waves can be represented as superpositions of plane waves.

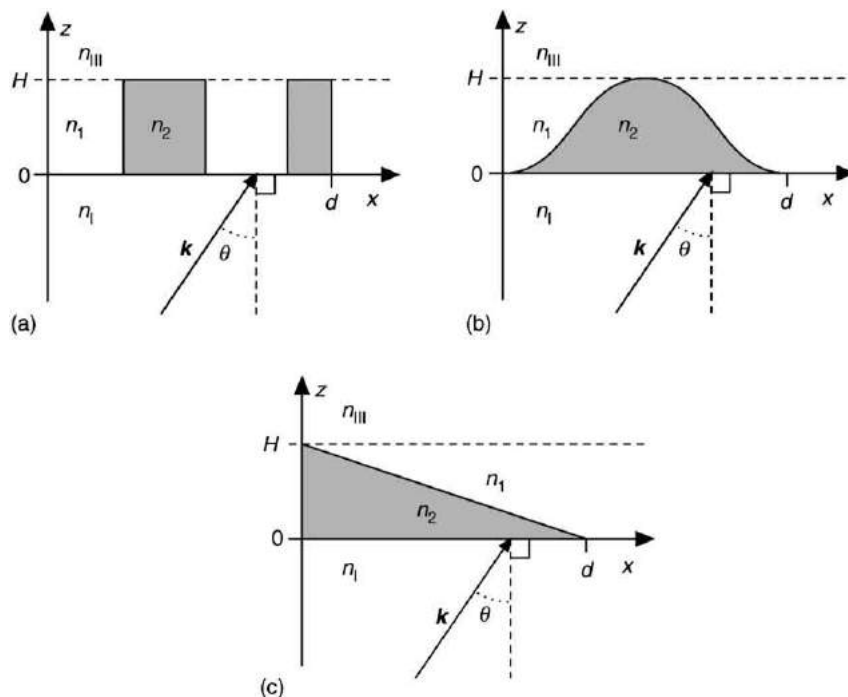


Fig. 1 Examples of linear gratings. (a) Binary grating; (b) sinusoidal grating; (c) triangular grating. Here d is the grating period, H is the thickness of the modulated region, θ is the angle of incidence, and n denotes refractive index.

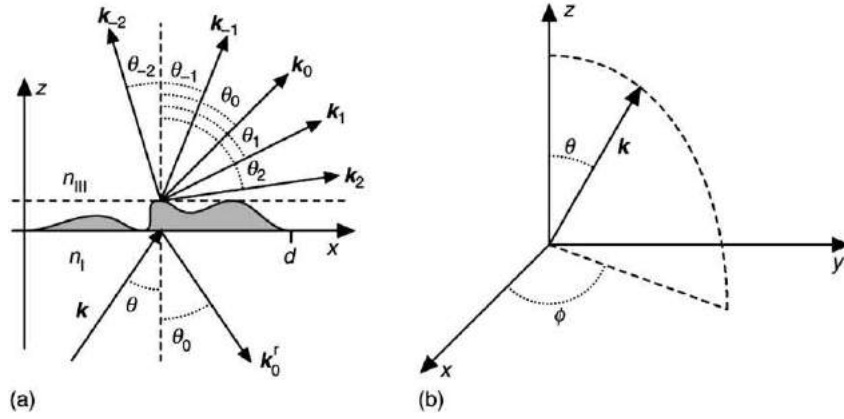


Fig. 2 Diffraction by gratings. (a) Linear grating; (b) definitions of direction angles in the case of a biperiodic grating.

Fig. 1(a) illustrates a binary grating: in the modulated region $0 < z < H$ the refractive index alternates between n_1 and n_2 , changing from one value to another at certain transition points within the period d . The structure in **Fig. 1(b)** represents a sinusoidal surface-relief grating as an example of smooth grating profiles. Finally, **Fig. 1(c)** shows a triangular grating profile.

The periodicity of a grating, illuminated by a plane wave, usually leads to the appearance of a set of reflected and transmitted plane waves (diffraction orders of the grating) that propagate in directions determined by the grating period d , the illumination wavelength λ , and the refractive indices n_1 and n_{III} (see **Fig. 2(a)**). These directions are obtained from the transmission grating equation:

$$n_{III} \sin \theta_m = n_1 \sin \theta + m \lambda / d \quad (1)$$

where m is the index of the diffraction order and θ_m is its propagation angle. For reflected orders (superscript r) one simply replaces n_{III} by n_1 in **Eq. (1)**.

Obviously only a limited number of orders satisfy the condition $|\theta_m| < \pi/2$ and thereby represent conventional propagating plane waves. Higher, so-called evanescent orders are inhomogeneous waves that propagate along the grating surface and decay exponentially in the z direction. However, they cannot be ignored in grating analysis, as we shall see.

In the case of a biperiodic grating with a period $d_x \times d_y$, the propagation direction of the incident beam is defined by two angles θ and ϕ illustrated in **Fig. 2(b)**. The grating equations now read as:

$$n_{III} \sin \theta_{mn} \cos \phi_{mn} = n_1 \sin \theta \cos \phi + m \lambda / d_x \quad (2)$$

$$n_{III} \sin \theta_{mn} \sin \phi_{mn} = n_1 \sin \theta \sin \phi + n \lambda / d_y, \quad (3)$$

where θ_{mn} and ϕ_{mn} define the propagation directions of the diffraction order with indices m and n . For reflected orders n_{III} is again replaced by n_1 . These expressions apply also to the special case of linear gratings (limit $d_y \rightarrow \infty$) illuminated by a plane wave not propagating in the x - z plane (conical incidence).

Some two-dimensionally periodic gratings are illustrated in **Fig. 3**. The structure illustrated in **Fig. 3(a)** is an extension of a binary grating to the biperiodic geometry: here a single two-dimensional feature is placed within the grating period but, of course, more features with different outlines could be included as well. Moreover, each period could contain a continuous 'mountain-range' or consist of a 'New-York'-style pixel structure. **Fig. 3(b)** illustrates a structure consisting of an array of rectangular-shaped holes pierced in a flat screen, while **Fig. 3(c)** defines an array of isolated particles on a flat substrate. If the grid or the particles are metallic, surrounded by dielectrics from all sides, one speaks of inductive and capacitive structures, respectively: in the former case relatively free-flowing currents are induced in the metallic grid structure, but in the latter case the isolated metal particles become polarized (and therefore capacitive) when illuminated by an electromagnetic field.

Grating Theory

Gratings equations provide exact knowledge of the propagation directions of the diffraction orders, but they give no information about the distribution of light among different orders. To obtain such information, one must in general solve Maxwell's equations exactly (rigorous diffraction theory), but within certain limits simpler methods are available. In all cases, however, the transmitted and reflected fields are represented as superpositions of plane waves that propagate in the directions allowed by the grating equations (these are known as Rayleigh expansions).

The most popular method for rigorous grating analysis is the Fourier Modal Method (FMM), though many alternative analysis methods exist. To understand the principle of FMM, consider first binary gratings. The field inside the modulated region $0 < z < H$ is represented as a superposition of waveguide modes, determined by imagining that the modulated region would extend from $x = -\infty$ to $x = \infty$. With given transition points the electric permittivity, its inverse, and the electric field inside the grating, can be represented in the form of Fourier series and one ends up with a matrix eigenvalue problem with finite dimensions when the

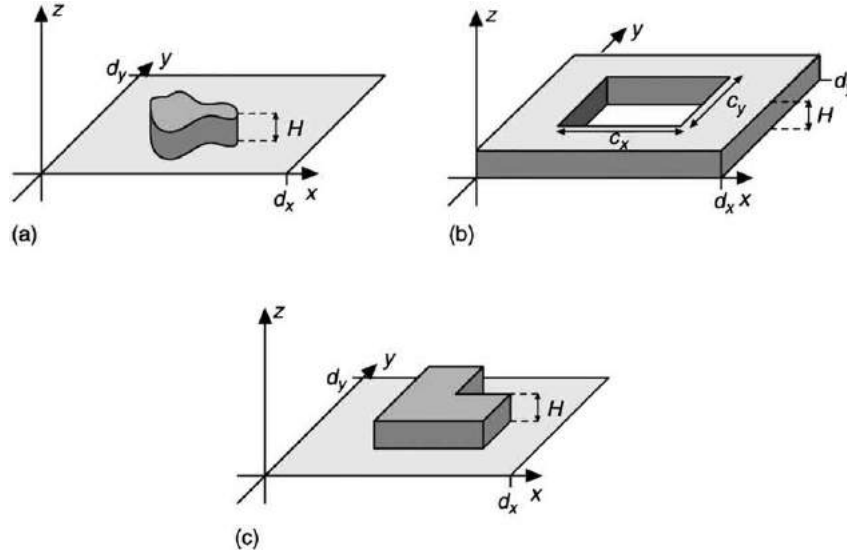


Fig. 3 Examples of biperiodic gratings. (a) General binary grating; (b) inductive structure; (c) capacitive structure.

matrices are truncated. A numerical solution of this problem provides a set of orthogonal modes that can propagate in the structure with certain propagation constants. The relative weights of the modes are then determined by applying the electromagnetic boundary conditions at the planes $z=0$ and $z=H$ to match the modal expansions inside the grating with the plane-wave (Rayleigh) expansions of the fields in the homogeneous materials on both sides of the grating. This matching procedure, which is numerically implemented by solving a set of simultaneous linear equations, yields the complex amplitudes of the electric and magnetic field components associated with the reflected and transmitted diffraction orders of the grating.

Of main interest are the diffraction efficiencies, η_m (or η_{mn} for biperiodic gratings) of the orders, which tell the distribution of energy between different propagating orders. In electromagnetic theory the diffraction efficiency is defined as the z -component of the time-averaged Poynting vector associated with the particular order, divided by that of the incident plane wave.

Gratings with z -dependent surface profiles can be handled similarly if they are 'sliced' into a sufficient number of layers, in each of which the grating is modeled as a binary structure. The electromagnetic boundary conditions are applied at each slice boundary and the approach models the real continuous structure with arbitrary degree of accuracy when the number of the slices is increased; in practice it is usually sufficient to use approximately 20–30 slices for sufficiently accurate modeling of any continuous grating profile by FMM.

The main problem with rigorous techniques such as FMM is bad scaling. If, in the case of 1D modulated (or y -invariant) gratings, the period is increased by a factor of two, the computation time required to solve the eigenvalue problem is increased by a factor of around eight. Thus we have a d^3 type scaling law. The situation is much worse for bi-periodic gratings, for which the exponent is around six. Thus no foreseeable developments in computer technology will solve the problem, and approximate analysis methods have to be sought. A multitude of such methods are under active development, but in this context we discuss the simplest example only.

If $d \gg \lambda$ and the grating profile does not contain wavelength-scale features, one can describe the effects of a grating on the incident field merely by considering it as a complex-amplitude-modulating object, which imposes a position-dependent phase shift and amplitude transmission (or reflection) coefficient on the incident field. This model is known as the thin-element approximation (TEA).

At normal incidence the plane-wave expansion of any scalar field component U immediately behind the grating (at $z=H$) takes the form:

$$U(x, y, H) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{mn} \times \exp[i2\pi(mx/d_x + ny/d_y)] \quad (4)$$

where

$$T_{mn} = \frac{1}{d_x d_y} \int_0^{d_x} \int_0^{d_y} U(x, y, H) \times \exp[-i2\pi(mx/d_x + ny/d_y)] dx dy \quad (5)$$

are the complex amplitudes of the transmitted diffraction orders. In TEA the field at $z=H$ may be connected to the field at $z=0$ (a unit-amplitude plane wave is assumed here) by optical path integration along a straight line through the modulated region of the grating:

$$U(x, y, H) = \int_0^H \exp\left[i\frac{2\pi}{\lambda} \hat{n}(x, y, z)\right] dz \quad (6)$$

where $\hat{n}(x, y, z) = n(x, y, z) + i\kappa(x, y, z)$ is the complex refractive-index distribution in the region $0 < z < H$. With TEA the diffraction efficiencies can be calculated from a simple formula:

$$\eta_{mn} = |T_{mn}|^2 \quad (7)$$

which, however, ignores Fresnel reflections.

Eqs. (5)–(7) allow one to obtain closed-form expressions for the diffraction efficiencies η_{mn} in many grating geometries. For example, a y -invariant triangular profile with $H = \lambda/|n_1 - n_{III}|$ gives $\eta_{-1} = 100\%$ if $n_1 = n_{III}$, $n_2 = n_1$, and $\theta = 0$. For a Q -level stair-step-quantized version of this grating we obtain:

$$\eta_m = \text{sinc}^2(1/Q) \quad (8)$$

where $c x = \sin(\pi x)/(\pi x)$. It should be noted, however, that these results are valid only for gratings with $d \gg \lambda$, even if they are corrected by Fresnel transmission losses at the interface between the two media.

Fabrication of Gratings

Gratings can be manufactured in a number of different ways, perhaps the simplest being to plot an equidistant array of straight black and white lines and to reduce the scale of the pattern photographically. This method can be readily adapted to the generation of more complex grating structures and diffractive elements, but it is not suitable for the production of high-quality gratings required in, for example, spectroscopy.

Ruling engines represent the traditional way of making gratings with periods of the order of the wavelength of light, especially for spectroscopic applications. These are machines equipped with a triangularly shaped diamond ruling edge and an interferometrically controlled two-dimensional mechanical translation stage that moves the substrate with respect to the ruling edge. Triangular-profile gratings of high quality can be produced with such machines, but the fabrication process is slow and the cost of the manufactured gratings is therefore high. In addition, even with real-time interferometric control of the ruling-edge position, it is impossible to avoid small local errors in the grating geometry. Such errors are particularly harmful if they are periodic since, according to the grating equation, these errors will generate spurious spectral lines known as ghosts. Such lines generated by the grating ‘super-period’ are weak but they may be confused with real spectral lines in the source spectrum. If the ruling errors are random, they result in ‘pedestal-type’ background noise. While such errors do not introduce ghosts, they nevertheless reduce the visibility of weak original spectral lines.

The introduction of lasers in the early 1960s made it possible to form stable and high-contrast interference patterns over a sufficient time-scale to allow the exposure of photographic films and other photosensitive materials such as polymeric photoresists. This interferometric technique offers a direct method for the fabrication of large diffraction gratings without appreciable periodic errors. With this technique it is possible to make highly regular gratings for spectroscopy, but it does not permit easy generation of arbitrarily shaped grating structures. The two intersecting plane waves can also be generated with a linear grating using diffraction orders $m = +1$ and $m = -1$, suppressing the zero order by appropriate design of the grating profile and choosing the period in such a way that all higher transmitted orders are evanescent. This so-called phase mask technique is suitable, for example, for the exposure of Bragg gratings in fibers.

To generate arbitrary gratings one usually employs lithographic techniques. Some of the possible processes are illustrated in **Fig. 4**. The exposure can be performed with scanning photon, electron, or ion beams, etc. The process starts with substrate preparation, i.e., coating it with a beam-sensitive layer known as resist, and in the case of electron beam exposure with a thin conductive layer to prevent charging. Following the exposure the resist is developed chemically, and in this process either the exposed or nonexposed parts of the resist are dissolved (depending on whether the resist is of positive or negative type), leading to a surface-relief profile. Several alternative fabrication steps may follow the development, depending on the desired grating type. If a high-contrast resist is used, a slightly ‘undercut’ profile as illustrated in **Fig. 4(c)** is obtained. This can be coated with a metal layer and the resist can be ‘lifted off’ chemically. Thus a binary metallic grating is formed. To produce a binary dielectric profile one bombards the substrate with ions either purely mechanically (ion beam milling) or with chemical assistance (reactive ion etching), and finally the residual metal is dissolved.

Alternatively, the particle beam dose can be varied with position to obtain an analog profile in the resist after development. This resist profile can be transformed into the substrate by proportional etching, which leads to an analog dielectric surface-relief structure (see **Fig. 4(f)**). This profile can be made reflective by depositing a thin metal film on top of it. If large-scale fabrication of identical gratings is the goal, a thick ($\sim 100 \mu\text{m}$) metal layer (for example nickel) can be grown on the surface by electroplating. After separation from the substrate, the electroplated metal film, which is a negative of the original profile, can be used as a stamp in replicating the original profile in plastics using methods such as hot embossing, ultraviolet radiation curing, and injection molding. In particular, the latter method provides the key to high-volume and low-cost production of various types of diffraction gratings and other grating-based diffractive structures in the industrial scale.

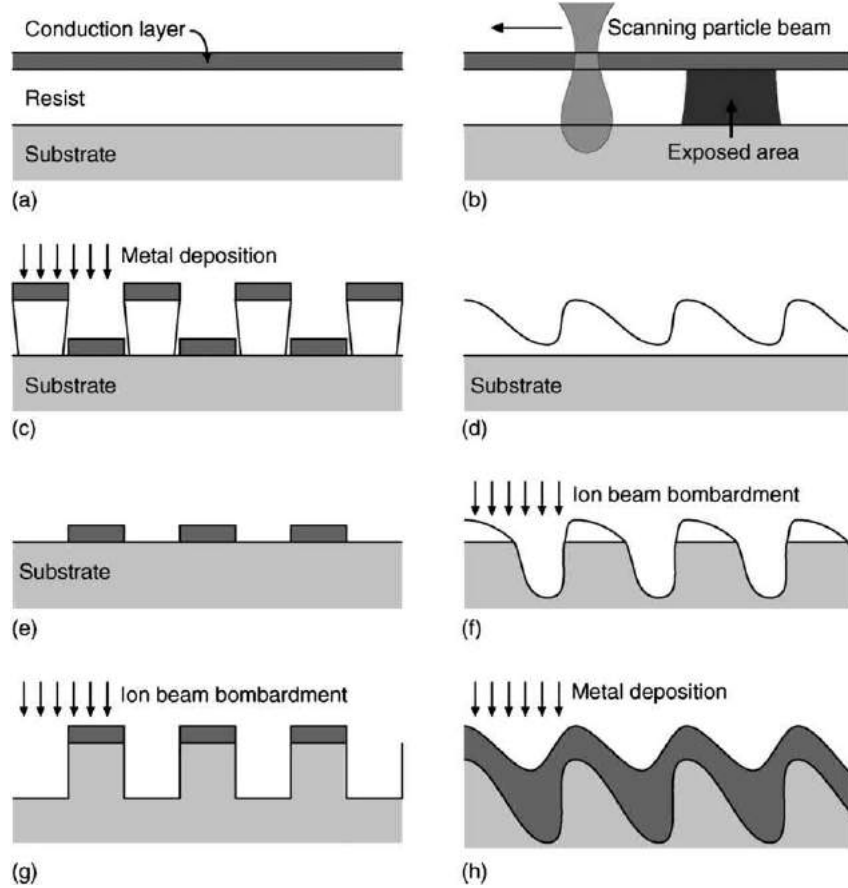


Fig. 4 Lithographic processes for grating fabrication. (a) Substrate after preparation for electron beam exposure; (b) exposure by a scanning electron beam; (c) slightly undercut profile in high-contrast resist after development, under metal film evaporation; (d) analog resist profile after variable-dose exposure and resist development; (e) finished binary metallic grating after resist lift-off; (f) transfer of the analog surface-relief profile into the substrate by proportional etching; (g) etching of a binary profile into the substrate with a metal mask; (h) growth of a metal film on an analog surface profile.

Some Grating Applications

We now proceed to describe a collection of representative grating applications. More exhaustive coverage of grating applications can be found in the Further Reading section at the end of this article.

Spectroscopic Gratings

The purpose of a spectroscopic grating is to divide the incident light into a spectrum, i.e., to direct different frequency components of the incident beam into different directions given by the grating (Eq. (1)). When the grating period is reduced, the angular dispersion $\partial\theta_m/\partial\lambda$, and therefore the resolution of the spectrograph, increases.

In the numerical examples presented in Fig. 5 we assume a reflection grating in a Littrow mount, i.e., $\theta_m^r = -\theta$. This mount is common in spectroscopy; it obviously requires that the grating is rotated when the spectrum is scanned. We see from Eq. (1) that in the Littrow mount:

$$\sin\theta = \frac{m\lambda}{2d} \quad (9)$$

By inserting this result back into Eq. (1) and differentiating we obtain a quantitative measure for dispersion in the Littrow mount:

$$\frac{\partial\theta_m}{\partial\lambda} = -\frac{m}{2d\cos\theta} \quad (10)$$

We consider TE polarization (electric field is linearly polarized in the y -direction) and order $m = -1$, optimize H to give maximum efficiency at $\lambda_0 = 550$ nm, and plot the diffraction efficiency η_{-1} over the visible wavelength range for two values of the grating period, $d = \lambda_0$ and $d = 2\lambda_0$. Rigorous diffraction theory is used with $n_1 = n_2 = 1$, and the metal is assumed to be aluminum (n_1 and n_{III} are complex-valued).

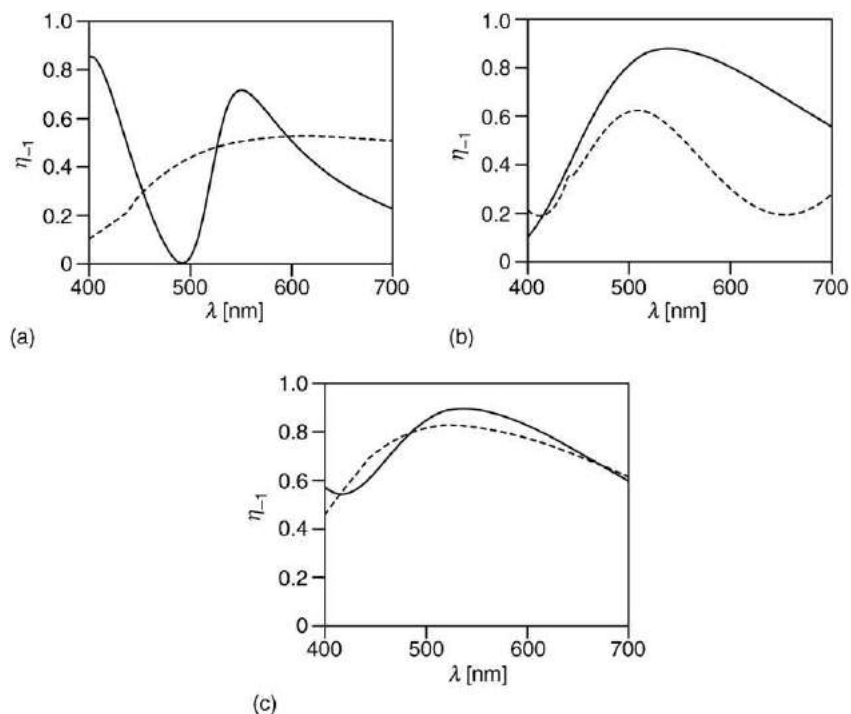


Fig. 5 Spectral diffraction efficiency curves of spectroscopic gratings in TE polarization. (a) Binary; (b) sinusoidal; and (c) triangular reflection gratings illuminated at Bragg angle. Solid lines: $d = \lambda_0$. Dashed lines: $d = 2\lambda_0$.

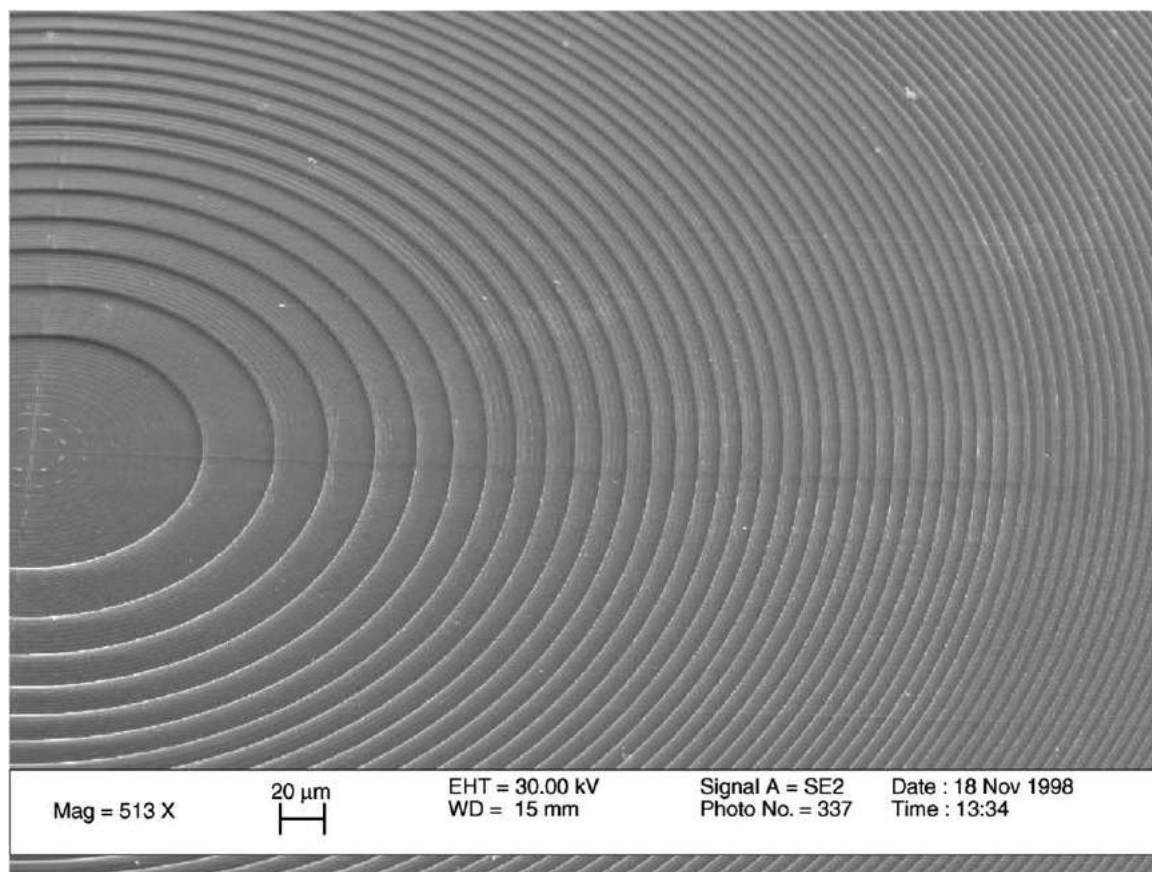


Fig. 6 A scanning electron micrograph of a part of a diffractive lens.

In Fig. 5(a) binary gratings with one groove (width $d/2$) per period is considered. Parametric optimization gives $H=422$ nm if $d=1\lambda$, and $H=168$ nm if $d=2\lambda$. Fig. 5(b) illustrates the results for sinusoidal gratings with $H=345$ nm if $d=1\lambda$, and $H=520$ nm if $d=2\lambda$. Finally, triangular gratings with $H=431$ nm if $d=1\lambda$, and $H=316$ nm if $d=2\lambda$, are considered. The triangular grating profile is seen to be the best choice in general.

Diffraction Lenses

Diffraction lenses are gratings with locally varying period and groove orientation. Fig. 6 illustrates the central part of a diffractive microlens fabricated by lithographic methods. It is seen to consist of concentric zones with radially decreasing width.

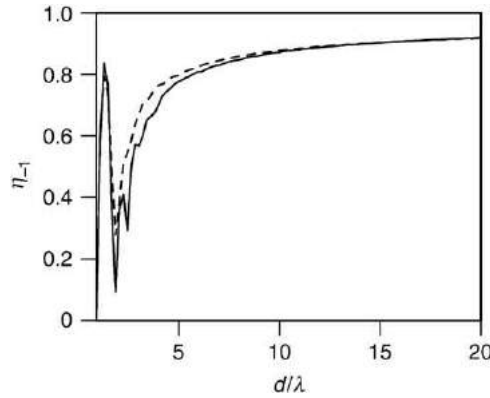


Fig. 7 Diffraction efficiency versus wavelength for triangular transmission gratings illuminated at normal incidence. Solid lines: TE polarization. Dashed lines: TM polarization.

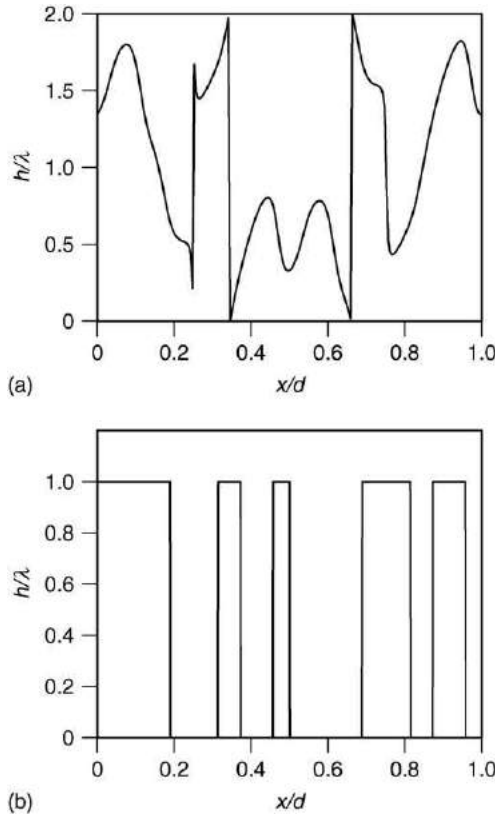


Fig. 8 Surface-relief profiles of (a) continuous and (b) binary 1 → 8 beamsplitter gratings.

The phase transmission function of an ideal diffractive focusing lens can be obtained by a simple optical path calculation, and is given by

$$\phi(r) = \frac{2\pi}{\lambda_0} \left(f - \sqrt{f^2 + r^2} \right) \approx -\frac{\pi}{\lambda_0 f} r^2 \quad (11)$$

where f is the focal length, r is the radial distance from the optical axis, and the approximate expression is valid in the paraxial region $r \ll f$. Here λ_0 is the design wavelength, which determines the modulation depth of the locally blazed grating profile:

$$H = \frac{\lambda_0}{n(\lambda_0) - 1} \quad (12)$$

and $n(\lambda_0)$ is the refractive index of the dielectric material at $\lambda = \lambda_0$.

The zone boundaries r_n are defined by the condition $\phi(r_n) = -2\pi n$, and therefore we obtain from Eq. (11):

$$r_n = \sqrt{2nf\lambda_0 + (n\lambda_0)^2} \approx \sqrt{2nf\lambda_0} \quad (13)$$

Obviously, when the numerical aperture of the lens is large, the zone width in the outer regions of the lens is reduced to values of the order of λ . Fig. 7 illustrates the local diffraction efficiency in different part of a diffractive lens: we use the order $m = -1$ of a locally triangular diffraction grating of the type shown in Fig. 1(c) with $\theta = 0$, assuming that $n_I = n_2 = 1.5$ (fused silica) and $n_1 = n_{III} = 1$. The efficiency η_{-1} is calculated as a function of λ using rigorous diffraction theory for both TE and TM polarization (the electric vector points in the groove direction in the TE case and is perpendicular to it in the TM case). It is seen that the local efficiency of the lens becomes poor in regions where the local grating period is small, and thus diffractive lenses operate best in the paraxial region (it is, however, possible to obtain improved non-paraxial performance by optimizing the local grating profile).

A factor that limits the use of diffractive lenses with polychromatic light is their large chromatic aberration: in the paraxial limit the focal length depends on the wavelength according to the formula:

$$f(\lambda) = \frac{\lambda_0}{\lambda} f \quad (14)$$

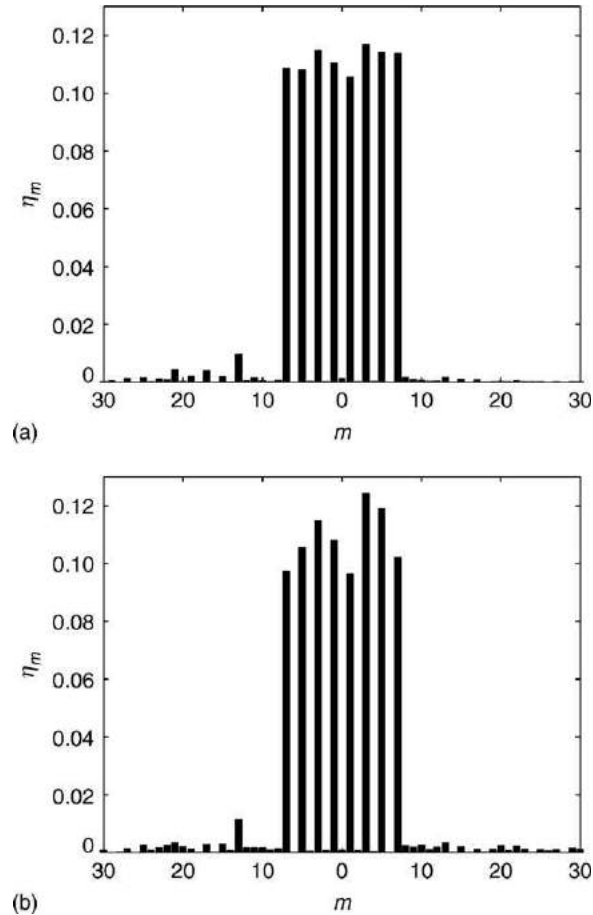


Fig. 9 Diffraction pattern produced by a continuous-profile beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

However, since the dispersion of a grating has an opposite sign compared to the diffraction of a prism, a weak positive diffractive lens can be combined with a relatively strong positive refractive lens to form a single-element achromatic lens. Such a hybrid lens is substantially thinner and lighter than a conventional all-refractive achromatic lens made of crown and flint glasses.

Beam Splitter Gratings

Multiple beamsplitters, also known as array illuminators, are gratings with sophisticated periodic structure that are capable of transforming an incident plane wave into a set of diffraction orders with a specified distribution of diffraction efficiencies. Most often one wishes to obtain a one- or two-dimensional array of diffracted plane waves with equal efficiency. This goal can be achieved by appropriate design of the grating profile, which may be of binary, multilevel, or continuous form.

Fig. 8 illustrates binary and continuous grating profiles capable of dividing one incident plane wave into eight diffracted plane waves (central odd-numbered orders of the grating) with equal efficiency in view of TEA. **Fig. 9** shows the distribution of energy among the different diffraction orders in and around the array produced by the continuous profile, evaluated by rigorous theory for two different grating periods, while **Fig. 10** gives corresponding results for the binary grating.

The fraction of incident energy that ends up in the desired eight orders is 72% for the binary grating and 96% for the continuous grating according to TEA, making the latter more attractive from a theoretical viewpoint. However, the continuous profile is more difficult to fabricate accurately than the binary one, especially for small values of the ratio d/λ . Moreover, the array uniformity produced by the continuous profile degrades faster than that of the binary profile when the period is reduced.

An example of the structure of a biperiodic grating that is capable of forming a large, shaped array of diffraction orders with equal efficiency is presented in **Fig. 11**. **Fig. 12** shows the far-field diffraction pattern produced by this element: the shape of Finland with the location of the town of Joensuu highlighted by placing the zero diffraction order on the appropriate position on the map (the apparent nonuniformity in the array is mostly due to the finite resolution of the CCD detector).

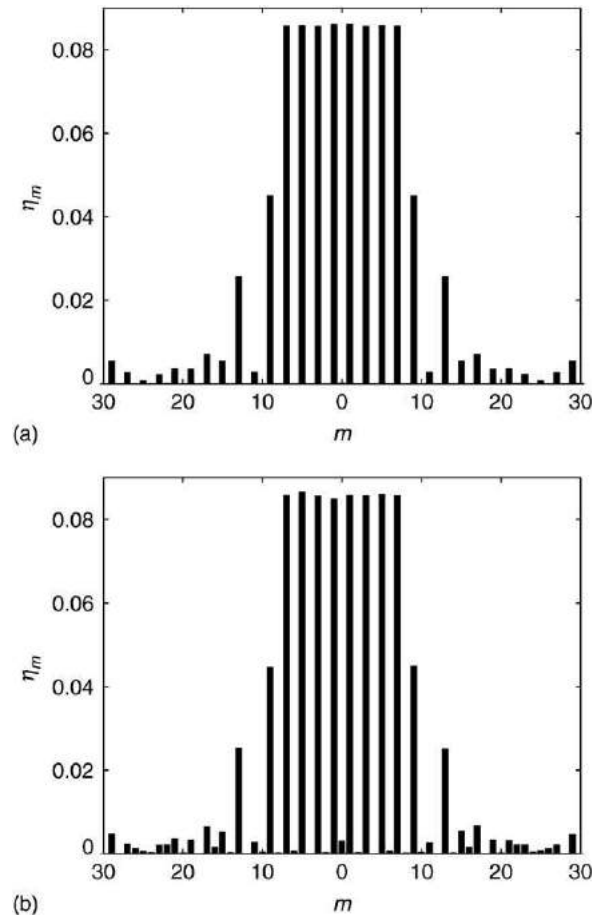


Fig. 10 Diffraction pattern produced by a binary beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

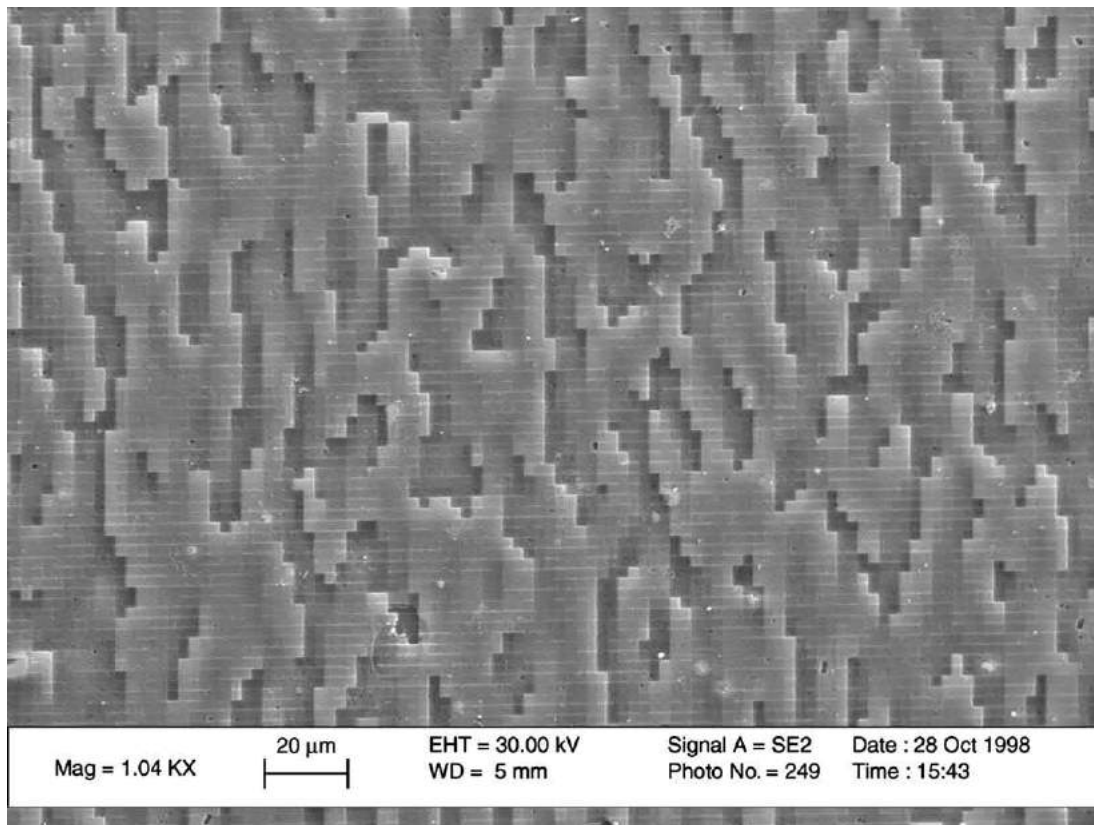


Fig. 11 Scanning electron micrograph of a sophisticated beamsplitter grating.

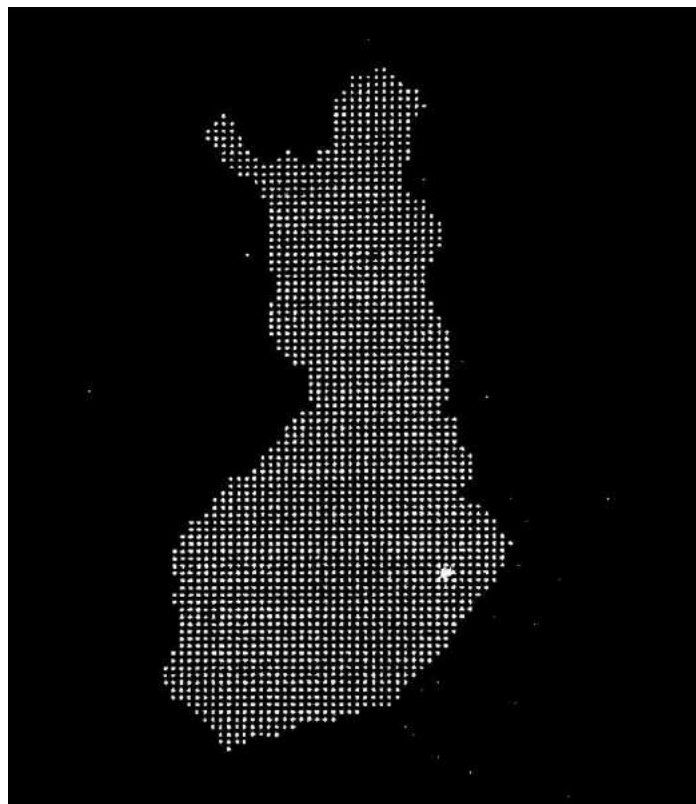


Fig. 12 The far-field diffraction pattern produced by the grating in Fig. 11.

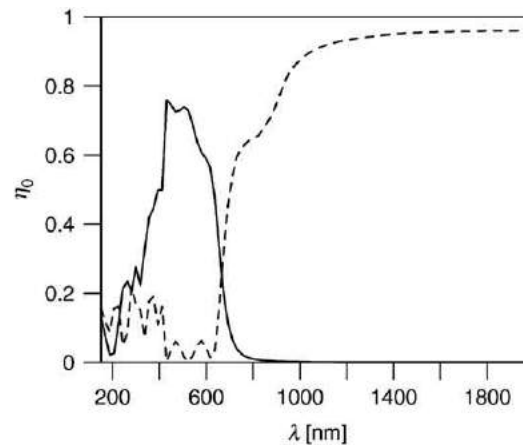


Fig. 13 Zero order spectral transmission (solid line) and reflection (dashed line) of an aluminium inductive grid filter with 400 nm period.

Inductive Grid Filters

As a final example we consider a scaled-down version of a device well known from the door of a microwave oven: one can see in through the holes pierced in the metallic screen in the oven door, but the microwaves are trapped inside because their wavelengths are large in comparison with the dimensions of the holes. Another example of these inductive grid filters is found in large radio telescope mirrors for centimeter-scale wavelengths, which are wide-grid structures rather than smoothly curved metal surfaces.

Fig. 13 illustrates the spectral transmission and reflection curves of an inductive grid filter of the type illustrated in **Fig. 3(b)**. We assume that $d_x = d_y = d = 400$ nm, $c_x = c_y = c = 310$ nm, $H = 500$ nm, and that the grid is a self-supporting structure ($n_I = n_{III} = 1$) made of aluminum. Only the zero reflected and transmitted orders propagate when $\lambda > d$, but other orders (with rather low efficiency) appear for smaller wavelengths. We see that the filter reflects infrared radiation effectively, but transmits a substantial part of radiation at visible and shorter wavelength regions.

See also: Fraunhofer Diffraction

Further Reading

- Chandezon, J., Maystre, D., Raoult, G., 1980. A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics* 11, 235–241.
- Gaylord, T.K., Moharam, M.G., 1985. Analysis and applications of optical diffraction by gratings. *Proceedings of IEEE* 73, 894–937.
- Herzig, H.P., 1997. *Micro-optics: Elements, Systems and Applications*. London: Taylor & Francis.
- Hutley, M.C., 1982. *Diffraction Gratings*. London: Academic Press.
- Li, L., 2001. Mathematical reflections on the fourier modal method in grating theory. In: Bao, G., Cowsar, L., Masters, W. (Eds.), *Mathematical Modelling in Optical Science*. Philadelphia: SIAM, pp. 111–139.
- Li, L., Chandezon, J., Granet, G., Plumey, J.-P., 1999. Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied Optics* 38, 304–313.
- Petit, R. (Ed.), 1980. *Electromagnetic Theory of Gratings*. Berlin: Springer-Verlag.
- Rayleigh, J.W.S., 1907. On the dynamical theory of gratings. *Proceeding of the Royal Society A* 79, 399.
- Solymar, L., Cooke, D.J., 1981. *Volume Holography and Volume Gratings*. London: Academic Press.
- Turunen, J., Kuittinen, M., Wyrowski, F., 2000. Diffractive optics: electromagnetic approach. In: Wolf, E. (Ed.), *Progress in Optics XL*. Amsterdam: Elsevier.
- Turunen, J., Wyrowski, F. (Eds.), 2000. *Industrial and Commercial Applications of Diffractive Optics*. Berlin: Akademie Verlag.
- van Renesse, R.L., 1993. *Optical Document Security*. Boston, MA: Artech House.

Fraunhofer Diffraction

BD Guenther, Duke University, Durham, NC, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Geometrical optics does a reasonable job of describing the propagation of light in free space unless that light encounters an obstacle. Geometrical optics predicts that the obstacle would cast a sharp shadow. On one side of the shadow's edge would be a bright uniform light distribution, due to the incident light, and on the other side there would be darkness. Close examination of an actual shadow edge reveals, however, dark fringes in the bright region and bright fringes in the dark region. This departure from the predictions of geometrical optics was given the name diffraction by Grimaldi.

There is no substantive difference between diffraction and interference. The separation between the two subjects is historical in origin and is retained for pedagogical reasons. Since diffraction and interference are actually the same physical process, we expect the observation of diffraction to be a strong function of the coherence of the illumination. With incoherent sources, geometrical optics is usually all that is needed to predict the performance of an optical system and diffraction can be ignored except at dimensions on the scale of a wavelength. With light sources having a large degree of spatial coherence, it is impossible to neglect diffraction in any analysis.

Even though diffraction produced by most common light sources and objects is a small effect, it is possible to observe diffraction without special equipment. By viewing a light source through a slit formed by two fingers that are nearly touching, fringes can be observed. Diffraction is used by photographers and television cameramen to provide artistic highlights in an image containing small light sources. They accomplish this by placing a screen in front of the camera lens. Light from point sources in the field of view is diffracted by the screen and produces 'stars' in the image. A third place where diffraction effects are easily observed is when a mercury or sodium streetlight, a few hundred meters away, is viewed through a sheer curtain.

Fresnel originally developed an approximate scalar theory based on Huygens' principle which states that:

Each point on a wavefront can be treated as a source of a spherical wavelet called a secondary wavelet or a Huygens' wavelet. The envelope of these wavelets, at some later time, is constructed by finding the tangent to the wavelets. The envelope is assumed to be the new position of the wavefront.

Rather than simply using the wavelets to construct an envelope, Fresnel's scalar theory assumes that the Huygens' wavelets interfere to produce a new wavefront.

A rigorous diffraction theory is based on Maxwell's equations and uses the boundary conditions associated with the obstacle to calculate a field scattered by the obstacle. The origins of this scattered field are currents induced in the obstacle by the incident field. The scattered field is allowed to interfere with the incident field to produce a resultant diffracted field. The application of the rigorous theory is very difficult and for most problems an approximate scalar theory developed by Kirchhoff and Sommerfeld is used.

Kirchhoff Theory of Diffraction

Kirchhoff's theory was based on the elastic theory of light but can be reformulated into a vector theory. Here we will limit our discussion to the scalar formulation. Given an incident wave, φ , we wish to calculate the optical wave at point P_0 , in Fig. 1, located at $\mathbf{r}_0$, in terms of the wave's value on a surface, S , that we construct about the observation point.

To obtain a solution to the wave equation at the point, P_0 , we select as a Green's function the one proposed by Kirchhoff (Box 1), a unit amplitude spherical wave, expanding about the point at $\mathbf{r}_0$, denoted by Ψ , and then apply Green's theorem (see Box 2). We will run into boundary condition problems as we proceed with the derivation of the solution. Sommerfeld removed the problems by assuming a slightly more complex Green's function (see Box 1). We will limit our discussion to the Kirchhoff Green's function.

Green's theorem requires that there be no sources (singularities) inside the surface S but our Green's function [1] is based on a source at $\mathbf{r}_0$. We eliminate the source by constructing a small spherical surface S_ε , of radius ε , about $\mathbf{r}_0$, excluding the singularity at $\mathbf{r}_0$ from the volume of interest, shown in gray in Fig. 1. The surface integration that must be performed in Green's theorem is over the surface $S' = S + S_\varepsilon$.

Within the volume enclosed by S' , the Green's function, Ψ , and the incident wave, φ , satisfy the scalar Helmholtz equation so that the volume integral can be written as

$$\int_V \int (\Psi \nabla^2 \varphi - \varphi \nabla^2 \Psi) dv = - \int_V \int (\Psi \varphi k^2 - \varphi \Psi k^2) dv \quad (7)$$

The right side of (7) is identically equal to zero. This fact allows us to use [6] to produce a simplified statement of Green's theorem:

$$\int_{S'} \left(\Psi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \Psi}{\partial n} \right) ds = 0 \quad (8)$$

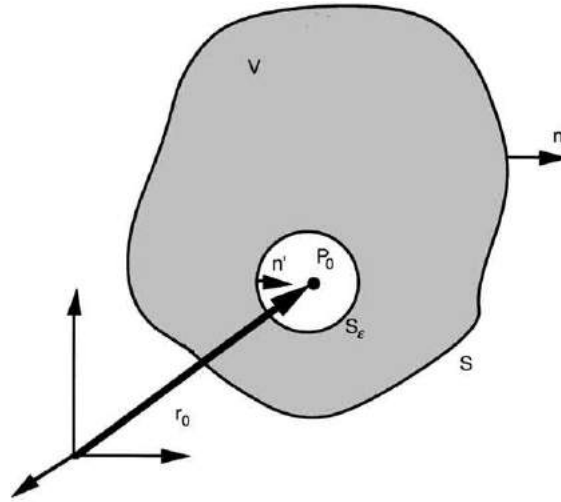


Fig. 1 Region of integration for solution of Kirchhoff's diffraction integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Box 1 Green's Function

1. **Kirchhoff Green's function:** Assume that the wave, Ψ , is due to a single point source at $\mathbf{r}_0$. At an observation point, $\mathbf{r}_1$, the Green's function is given by

$$\Psi(\mathbf{r}_1) = \frac{e^{-ikr_{01}}}{r_{01}}, \quad (1)$$

where $r_{01} = |\mathbf{r}_{01}| = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from the source of the Green's function, at $\mathbf{r}_0$, to the observation position, at $\mathbf{r}_1$.

2. **Sommerfeld Green's function:** Assume that there are two point sources, one at P_0 and one at P'_0 (see Fig. 2 where the source, P'_0 , is positioned to be the mirror image of the source at P_0 on the opposite side of the screen, thus

$$r_{01} = r'_{01} \quad \cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) = -\cos(\hat{\mathbf{n}}, \mathbf{r}'_{01}) \quad (2)$$

A Green's function that could be used is

$$\Psi' = \frac{e^{-ikr_{01}}}{r_{01}} + \frac{e^{-ikr'_{01}}}{r'_{01}} \quad (3a)$$

or

$$\Psi' = \frac{e^{-ikr_{01}}}{r_{01}} + \frac{e^{-ikr'_{01}}}{r'_{01}} \quad (3b)$$

Box 2 Green's Theorem To calculate the complex amplitude, $\tilde{E}(\mathbf{r})$ of a wave at an observation point defined by the vector $\mathbf{r}$, we need to use a mathematical relation known as Green's Theorem. Green's theorem states that, if φ and Ψ are two scalar functions that are well behaved, then

$$\int_S (\varphi \nabla \Psi \cdot \hat{\mathbf{n}} - \Psi \nabla \varphi \cdot \hat{\mathbf{n}}) dS = \int_V (\varphi \nabla^2 \Psi - \Psi \nabla^2 \varphi) dV \quad (4)$$

The vector identities

$$\nabla \Psi \cdot \hat{\mathbf{n}} = \frac{\partial \Psi}{\partial n} \quad \nabla \varphi \cdot \hat{\mathbf{n}} = \frac{\partial \varphi}{\partial n} \quad (5)$$

allow Green's theorem to be written as

$$\int_S \left[\varphi \frac{\partial \Psi}{\partial n} - \Psi \frac{\partial \varphi}{\partial n} \right] dS = \int_V [\varphi \nabla^2 \Psi - \Psi \nabla^2 \varphi] dV \quad (6)$$

This equation is the prime foundation of scalar diffraction theory but only the proper choice of Ψ , φ and the surface, S , allows direct application to the diffraction problem.

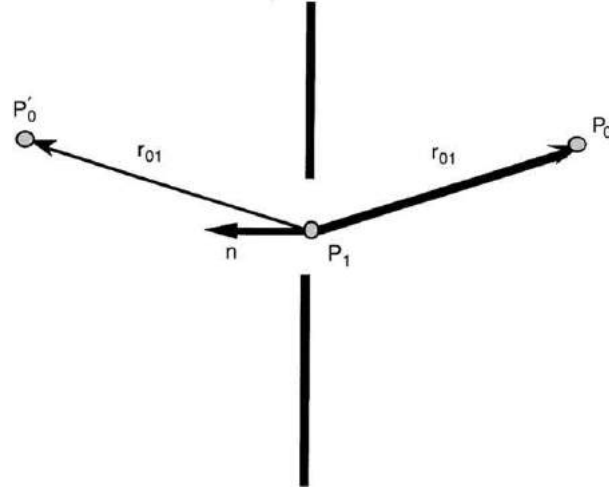


Fig. 2 Geometry for the Sommerfeld Green's function. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the integral over the combined surfaces, S and S_e , is equal to zero, the integral over the surface S must equal the negative of the integral over the surface, S_e :

$$-\int \int_{S_e} \left(\Psi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \Psi}{\partial n} \right) ds = \int \int_S \left(\Psi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \Psi}{\partial n} \right) ds \quad (9)$$

To perform the surface integrals, we must evaluate the Green's function on the surface, S' . We therefore select $\mathbf{r}_1$, the vector defining the observation point, to be on either of the two surfaces that make up S' . The derivative, $\partial \Psi / \partial n$, to be evaluated is

$$\frac{\partial \Psi}{\partial n} = \frac{\partial \Psi}{\partial r} \frac{\partial r}{\partial n} = \left(-ik \frac{e^{-ikr_{01}}}{r_{01}} - \frac{e^{-ikr_{01}}}{r_{01}^2} \right) \cos(\mathbf{n}, \mathbf{r}_{01}) \quad (10)$$

where $\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})$ is the cosine of the angle between the outward normal $\hat{\mathbf{n}}$ and the vector $\mathbf{r}_{01}$, the vector between points P_0 and P_1 . (Note from Fig. 1 that the outward normal on S_e is inward, toward P_0 , while the normal on S is outward, away from P_0 .)

For example if $\mathbf{r}_1$ were on S_e , then

$$\cos(\mathbf{n}, \mathbf{r}_{01}) = -1, \quad (11)$$

$$\frac{\partial \Psi(\mathbf{r}_1)}{\partial n} = \left[\frac{1}{\varepsilon} + ik \right] \frac{e^{-ik\varepsilon}}{\varepsilon}, \quad (12)$$

where ε is the radius of the small sphere we constructed around the point at P_0 .

We now use (10) to rewrite the two integrals in (9). The integral over the surface S_e is

$$\int \int_{S_e} \left(\Psi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \Psi}{\partial n} \right) ds = \int \int_{S_e} \left[\frac{\partial \varphi}{\partial n} \frac{e^{-ik\varepsilon}}{\varepsilon} - \varphi \frac{e^{-ik\varepsilon}}{\varepsilon} \left(\frac{1}{\varepsilon} + ik \right) \right] \times \varepsilon^2 \sin \theta d\theta d\phi \quad (13)$$

while the integral over the surface S is

$$\int \int_S \left[\frac{e^{-ikr_{01}}}{r_{01}} \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial}{\partial n} \left(\frac{e^{ikr_{01}}}{r_{01}} \right) \right] r^2 \sin \theta d\theta d\phi \quad (14)$$

The omitted volume contained within the surface S_e is allowed to shrink to zero by taking the limit as $\varepsilon \rightarrow 0$. Eq. (14) will not be affected by taking the limit. The first and last terms of the right side of (13) go to zero as $\varepsilon \rightarrow 0$ because they contain $\varepsilon e^{-ik\varepsilon}$. The second term in (13) contains $e^{-ik\varepsilon}$ which goes to 1 as $\varepsilon \rightarrow 0$. Therefore, in the limit as $\varepsilon \rightarrow 0$, (9) becomes

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi} \int \int_S \left[\frac{e^{-ikr_{01}}}{r_{01}} \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial}{\partial n} \left(\frac{e^{-ikr_{01}}}{r_{01}} \right) \right] ds \quad (15)$$

which is sometimes called the integral theorem of Kirchhoff. By carefully selecting the surface of integration in (15), we can calculate the diffraction field observed at P_0 produced by an aperture in an infinite opaque screen.

Assume that a source at P_2 in Fig. 3 produces a spherical wave that is incident on an infinite, opaque screen from the left. To find the field at P_0 in Fig. 3, we apply (15) to the surface $S_1 + S_2 + \Sigma$, where S_1 is a plane surface adjacent to the screen, Σ is that portion of S_1 in the aperture and S_2 is a large spherical surface, of radius R , centered on P_0 , see Fig. 3. The first question to address is

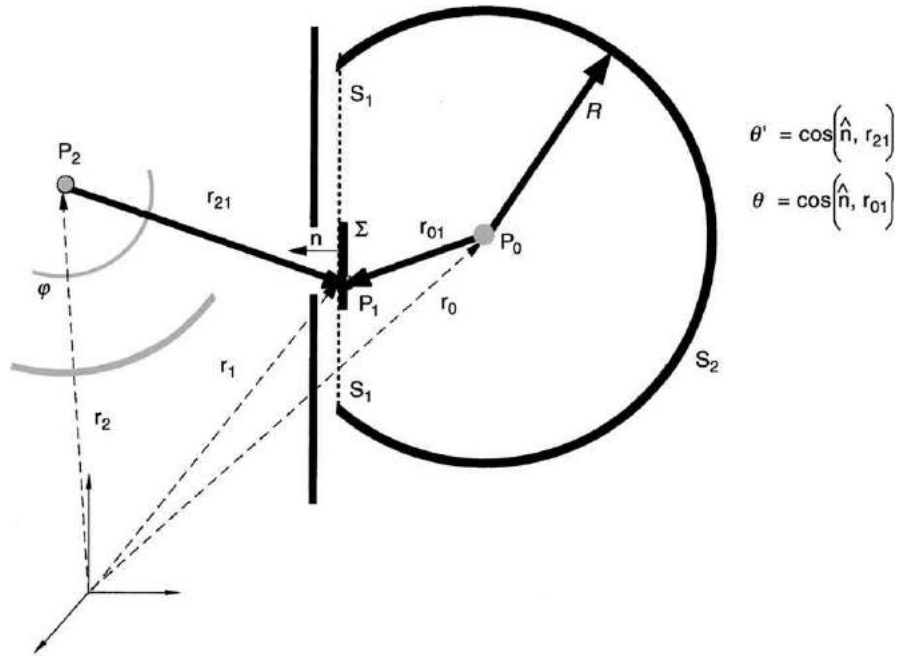


Fig. 3 Integration surface for diffraction calculation. P_2 is the source and P_0 is the point where the field is to be calculated. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

how to calculate the contribution from S_2 or better yet how do we show we can ignore contributions from S_2 . As R increases, Ψ and ϕ will decrease as $1/R$. However, the area of integration increases as R^2 so the $1/R$ fall-off is not a sufficient reason for neglecting the contribution of the integration over S_2 .

On the surface S_2 , the Green's function and its derivative are

$$\Psi = \frac{e^{-ikR}}{R} \quad (16)$$

$$\frac{\partial \Psi}{\partial n} \left(-ik - \frac{1}{R} \right) \frac{e^{-ikR}}{R} \approx -ik\Psi \quad (17)$$

where the approximate equality is for large R . Thus, the integral over S_2 for large R is

$$\iint_{S_2} \left[\Psi \frac{\partial \phi}{\partial n} + \phi (ik\Psi) \right] ds = \iint_{S_2} \Psi \left(\frac{\partial \phi}{\partial n} + ik\phi \right) R^2 \sin \theta d\theta d\phi \quad (18)$$

The quantity $R\Psi = e^{-ik \cdot R}$ is finite as $R \rightarrow \infty$ so for the integral to vanish, we must have

$$\lim_{R \rightarrow \infty} R \left(\frac{\partial \phi}{\partial n} + ik\phi \right) = 0 \quad (19)$$

This requirement is called the Sommerfeld radiation condition and is satisfied if $\phi \rightarrow 0$ as fast as $1/R$ (for example, a spherical wave). Since the illumination of the screen will be a spherical wave, or at least a linear combination of spherical waves, we should expect the integral over S_2 to make no contribution (it is exactly zero).

Another way to insure that surface S_2 makes no contribution is to assume that the light turns on at time t_0 . At a later time, t , when we desire to know the field at P_0 , the radius of S_2 is $R > c(t - t_0)$ which physically means that the light has not had time to reach S_2 . In this situation, there can be no contribution from S_2 . This is not a perfect solution because when the wave is of finite duration it can no longer be considered monochromatic.

We should expect the major contribution in the integral over S_1 to come from the portion of the surface in the aperture, Σ . We make the following assumptions about the incident wave, ϕ (these assumptions are known as St. Venaut's hypothesis or Kirchhoff's boundary conditions)

1. We assume that, in the aperture, ϕ and $\partial \phi / \partial n$ have the values they would have if the screen were not in place.
2. On the portion of S_1 not in the aperture, ϕ and $\partial \phi / \partial n$ are identically zero.

The Kirchhoff boundary conditions allow the screen to be neglected, reducing the problem to an integration only over the aperture. It is surprising that the assumptions about the contribution of the surface S_2 and the screen yield very accurate results

(as long as polarization is not important, the aperture is large with respect to the wavelength, and we don't get too close to the aperture).

Mathematically the Kirchhoff boundary conditions are incorrect. The two Kirchhoff boundary conditions imply that the field is zero everywhere behind the screen, except in the aperture, which makes Ψ and $\partial\Psi/\partial n$ discontinuous on the boundary of the aperture. Another problem with the boundary conditions is that Ψ and $\partial\Psi/\partial n$ are known only if the problem has already been solved.

The Sommerfeld Green's functions will remove the inconsistencies of the Kirchhoff theory. If we use the Green's function [3a] then Ψ vanishes in the aperture while $\partial\Psi/\partial n$ can assume the value required by Kirchhoff's boundary conditions. If we use the Green's functions [3b] then the converse holds, i.e. $\partial\Psi/\partial n$ is zero in the aperture. This improved Green's function makes the problem harder with little gain in accuracy so we will retain the Kirchhoff formalism.

As a result of the above discussion, the surface integral (15) is only performed over the surface, Σ , in the aperture. The evaluation of the Green's function on this surface can be simplified by noting that normally $r_{01} \gg \lambda$, i.e., $k \gg 1/r_{01}$, thus on Σ ,

$$\frac{\partial\Psi}{\partial n} = \cos(\mathbf{n}, \mathbf{r}_{01}) \left(-ik - \frac{1}{r_{01}} \right) \frac{e^{-ikr_{01}}}{r_{01}} \approx -ik \cos(\mathbf{n}, \mathbf{r}_{01}) \frac{e^{-ikr_{01}}}{r_{01}} \quad (20)$$

Substituting this approximate evaluation of the derivative into (15) yields

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi} \int \int_{\Sigma} \frac{e^{-ikr_{01}}}{r_{01}} \left[\frac{\partial\varphi}{\partial n} + ik\varphi \cos(\mathbf{n}, \mathbf{r}_{01}) \right] d\mathbf{s} \quad (21)$$

The source of the incident wave is a point source located at P_2 , with a position vector, $\mathbf{r}_2$, measured with respect to the coordinate system and a distance $|\mathbf{r}_{21}|$ away from P_1 , a point in the aperture (see Fig. 3). The incident wave is therefore a spherical wave of the form

$$\varphi(\mathbf{r}_{21}) = A \frac{e^{-ikr_{21}}}{r_{21}} \quad (22)$$

which fills the aperture. Here also we will assume that $r_{21} \gg \lambda$ so that the derivative of the incident wave assumes the same form as (20). Then (21) can be written

$$\varphi(\mathbf{r}_0) = \frac{iA}{\lambda} \int \int_{\Sigma} \frac{e^{-ik(\mathbf{r}_{21} + \mathbf{r}_{01})}}{r_{21}r_{01}} \times \left[\frac{\cos(\mathbf{n}, \mathbf{r}_{01}) - \cos(\mathbf{n}, \mathbf{r}_{21})}{2} \right] d\mathbf{s} \quad (23)$$

This relationship is called the Fresnel–Kirchhoff diffraction formula. It is symmetric with respect to r_{01} and r_{21} , making the problem identical when the source and measurement point are interchanged.

A physical understanding of (23) can be obtained if it is rewritten as

$$\varphi(\mathbf{r}_0) = \int \int_{\Sigma} \Phi(\mathbf{r}_{21}) \frac{e^{-ikr_{01}}}{r_{01}} d\mathbf{s} \quad (24)$$

The field at P_0 is due to the sum of an infinite number of secondary Huygens sources in the aperture Σ . The secondary sources are point sources radiating spherical waves of the form

$$\Phi(\mathbf{r}_{21}) \frac{e^{-ikr_{01}}}{r_{01}} \quad (25)$$

with amplitude $\Phi(\mathbf{r}_{21})$, defined by

$$\Phi(\mathbf{r}_{21}) = \frac{i}{\lambda} \left[A \frac{e^{-ikr_{21}}}{r_{21}} \right] \left[\frac{\cos(\mathbf{n}, \mathbf{r}_{01}) - \cos(\mathbf{n}, \mathbf{r}_{21})}{2} \right] \quad (26)$$

The imaginary constant, i , causes the wavelets from each of these secondary sources to be phase shifted with respect to the incident wave. The obliquity factor, in the amplitude (26)

$$\frac{1}{2} [\cos(\mathbf{n}, \mathbf{r}_{01}) - \cos(\mathbf{n}, \mathbf{r}_{21})] \quad (27)$$

causes the secondary sources to have a forward directed radiation pattern.

If we had used the Sommerfeld Green's function, the only modification to our result would be a change in the obliquity factor to $\cos(\mathbf{n}, \mathbf{r}_{01})$. In our discussions below we will assume that the angles are all small, allowing the obliquity factor to be replaced by 1 so that in the remaining discussion the choice of Green's function has no impact.

We will assume the source of light is at infinity, $z_1 = \infty$ in Fig. 3, so that the aperture is illuminated by a plane wave, traveling parallel to the z -axis. With this assumption

$$\theta' = \cos(\mathbf{n}, \mathbf{r}_{21}) \approx -1 \quad (28)$$

We will also make the paraxial approximation which assumes that the viewing position is close to the z -axis, leading to

$$\theta = \cos(\mathbf{n}, \mathbf{r}_{01}) \approx 1 \quad (29)$$

With these assumptions [Eq. \(24\)](#) becomes identical to the Huygens–Fresnel integral discussed in the section on Fresnel diffraction:

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda} \int \int_{\Sigma} \frac{\tilde{E}_i(\mathbf{r})}{R} e^{-ik \cdot \mathbf{R}} d\mathbf{s} \quad (30)$$

The modern interpretation of the Fresnel–Kirchhoff (or now in the small-angle approximation, the Fresnel–Huygens) integral is to view it as a convolution integral. By considering free space to be a linear system, the result of propagation can be calculated by convolving the incident (input) wave with the impulse response of free space (our Green’s function),

$$\frac{i}{\lambda} \frac{e^{-ik \cdot R}}{R} \quad (31)$$

The job of calculating diffraction has only begun with the derivation of the Fresnel–Kirchhoff integral. In general, an analytic expression for the integral cannot be found because of the difficulty of performing the integration over R . There are two approximations that will allow us to obtain analytic expressions of the Fresnel–Kirchhoff integral. In both approximations, all dimensions are assumed to be large with respect to the wavelength. In one approximation, the viewing position is assumed to be far from the obstruction; the resulting diffraction is called Fraunhofer diffraction and will be discussed here. The second approximation, which leads to Fresnel diffraction, assumes that the observation point is nearer the obstruction, to be quantified below.

Fraunhofer Approximation

In Fraunhofer diffraction, we require that the source of light and the observation point, P_0 , be far from the aperture so that the incident and diffracted wave can be approximated by plane waves. A consequence of this requirement is that the entire waveform passing through the aperture contributes to the observed diffraction.

The geometry to be used in this derivation is shown in Fig. 4. The wave incident normal to the aperture, Σ , is a plane wave and the objective of the calculation is to find the departure of the transmitted wave from its geometrical optical path. The calculation will provide the light distribution, transmitted by the aperture, as a function of the angle the light is deflected from the incident direction. We assume that diffraction makes only a small perturbation on the predictions of geometrical optics. The deflection angles encountered in this derivation are, therefore, small, as assumed above, and we will be able to use the paraxial approximation.

The distance from a point P, in the aperture, to the observation point P₀, of Fig. 4, is

$$R^2 = (x - \xi)^2 + (y - \eta)^2 + Z^2 \quad (32)$$

From Fig. 4 we see R_0 is the distance from the center of the screen to the observation point, P_0 ,

$$R_0^2 = \xi^2 + \eta^2 + Z^2 \quad (33)$$

The difference between these two vectors is

$$R_0^2 - R^2 = (R_0 - R)(R_0 + R) = \xi^2 + \eta^2 + Z^2 - Z^2 - (x^2 - 2x\xi + \xi^2) - (y^2 - 2y\eta + \eta^2) = 2(x\xi + y\eta) - (x^2 + y^2) \quad (34)$$

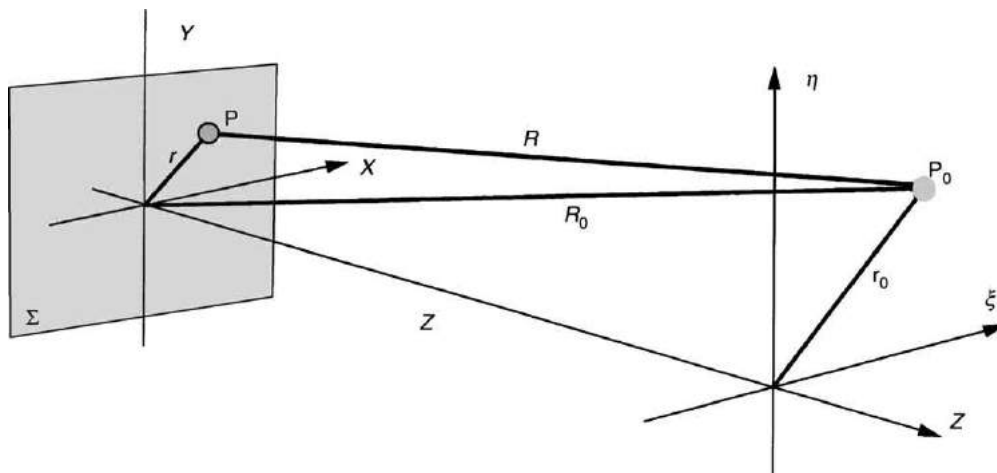


Fig. 4 Geometry for Fraunhofer diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The equation for the position of point P in the aperture can be written in terms of (34):

$$r = R_0 - R = \frac{R_0^2 - R^2}{R_0 + R}, = [2(x\xi + y\eta) - (x^2 + y^2)] \frac{1}{R_0 + R} \quad (35)$$

The reciprocal of $(R_0 + R)$ can be written as

$$\frac{1}{R_0 + R} = \frac{1}{2R_0 + R - R_0} = \frac{1}{2R_0} \left(1 + \frac{R - R_0}{2R_0}\right)^{-1} \quad (36)$$

Now using (36)

$$R_0 - R = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0} \right] \left[1 - \frac{R_0 - R}{2R_0} \right]^{-1} = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0} \right] \left[1 - \frac{r}{2R_0} \right]^{-1} \quad (37)$$

If the diffraction integral (30) is to have a finite (nonzero) value, then

$$k|R_0 - R| \ll kR_0 \quad (38)$$

This requirement insures that all the Huygens's wavelets, produced over the aperture from the center out to the position r , will have similar phases and will interfere to produce a nonzero amplitude at P_0 . This requirement results in

$$\frac{1}{1 - \frac{r}{2R_0}} \approx 1 \quad (39)$$

From this equation we are tempted to state that the aperture dimension must be small relative to the observation distance. However, for the exponent to provide a finite contribution in the integral, it is kr that is small, not the aperture dimension, r .

Using the approximation for $(R_0 - R)$, we obtain the diffraction integral

$$\tilde{E}_P = \frac{i\alpha e^{-ikR_0}}{\lambda R_0} \int \int_{\Sigma} f(x, y) e^{+ik \left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0} \right)} dx dy \quad (40)$$

The change in the amplitude of the wave due to the change in r , as we move across the aperture, is neglected, allowing R in the denominator of the Huygens–Fresnel integral to be replaced by R_0 and moved outside of the integral. We have introduced the complex transmission function, $f(x, y)$, of the aperture to allow a very general aperture to be treated. If the aperture function described the variation in absorption of the aperture as a function of position, as would be produced by a photographic negative, then $f(x, y)$ would be a real function. If the aperture function described the variation in transmission of a biological sample, it might be entirely imaginary.

The argument of the exponent in (40) is

$$ik \left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0} \right) = i2\pi \left(\frac{x\xi + y\eta}{\lambda R_0} - \frac{x^2 + y^2}{2\lambda R_0} \right) \quad (41)$$

If the observation point, P_0 , is far from the screen, we can neglect the second term and treat the phase variation across the aperture as a linear function of position. This is equivalent to assuming that the diffracted wave is a collection of plane waves. Mathematically, the second term in (41) can be neglected if

$$\frac{x^2 + y^2}{2\lambda R_0} \ll 1 \quad (42)$$

This is called the far-field approximation and the theory yields Fraunhofer diffraction. If the quadratic term of (41) is on the order of 2π then the fraction is

$$\frac{x^2 + y^2}{2\lambda R_0} \approx \mathcal{O}(1) \quad (43)$$

and we must retain the quadratic term and the theory yields Fresnel diffraction.

We define spatial frequencies in the x and y direction by

$$\begin{aligned} \omega_x &= \frac{2\pi}{\lambda} \sin \theta_x = -\frac{2\pi\xi}{\lambda R_0} \\ \omega_y &= \frac{2\pi}{\lambda} \sin \theta_y = -\frac{2\pi\eta}{\lambda R_0} \end{aligned} \quad (44)$$

We define the spatial frequencies with a negative sign to allow equations involving spatial frequencies to have the same form as those involving temporal frequencies. The negative sign is required because ωt and $\mathbf{k} \cdot \mathbf{r}$ appear in the phase of the wave with opposite signs and we want the Fourier transform in space and in time to have the same form.

With the variables defined in (44), the integral becomes

$$\tilde{E}_P(\omega_x, \omega_y) = \frac{i\alpha e^{-ikR_0}}{\lambda R_0} \int \int_{\Sigma} f(x, y) e^{-i(\omega_x x + \omega_y y)} dx dy \quad (45)$$

The result of our derivation is that the Fraunhofer diffraction field, $\tilde{E}_p$ can be calculated by performing the two-dimensional Fourier transform of the aperture's transmission function.

Definition of Fourier Transform

In this discussion the function $F(\omega)$ is defined as the *Fourier transform* of $f(t)$:

$$\mathcal{F}\{f(t)\} \equiv F(\omega) \equiv \int_{-\infty}^{\infty} f(\tau) e^{-i\omega\tau} d\tau \quad (46)$$

The transformation from a temporal to a frequency representation given by (46) does not destroy information; thus, the inverse transform can also be defined

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \quad (47)$$

By assuming that the illumination of an aperture is a plane wave and that the observation position is in the far field, the diffraction of an aperture is found to be given by the Fourier transform of the function describing the amplitude transmission of the aperture. The amplitude transmission of the aperture, $f(x, y)$, may thus be interpreted as the superposition of mutually coherent plane waves.

Diffraction by a Rectangular Aperture

We will now use Fourier transform theory to calculate the Fraunhofer diffraction pattern from a rectangular slit and point out the reciprocal relationship between the size of the diffraction pattern and the size of the aperture.

Consider a rectangular aperture with a transmission function given by

$$f(x, y) = \begin{cases} 1 & |x| \leq x_0, |y| \leq y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \quad (48)$$

Because the aperture is two-dimensional, we need to apply a two-dimensional Fourier transform but it is not difficult because the amplitude transmission function is separable, in x and y . The diffraction amplitude distribution from the rectangular slit is simply one-dimensional transforms carried out individually for the x and y dimensions:

$$\tilde{E}_p \frac{i\alpha}{\lambda R_0} e^{-ik \cdot R_0} \int_{-x_0}^{x_0} f(x) e^{-i\omega_x x} dx \int_{-y_0}^{y_0} f(y) e^{-i\omega_y y} dy \quad (49)$$

Since both $f(x)$ and $f(y)$ are defined as symmetric functions, we need only calculate the cosine transforms to obtain the diffracted field:

To calculate the Fourier transform defined by (46) we can rewrite the transform as

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau) \cos \omega\tau d\tau - i \int_{-\infty}^{\infty} f(\tau) \sin \omega\tau d\tau \quad (50)$$

If $f(\tau)$ is a real, even function then the Fourier transform can be obtained by simply calculating the cosine transform:

$$\int_{-\infty}^{\infty} f(\tau) \cos \omega\tau d\tau \quad (51)$$

$$\tilde{E}_p = i \frac{4x_0 y_0 \alpha}{\lambda R_0} e^{-ik \cdot R_0} \frac{\sin \omega_x x_0}{\omega_x x_0} \frac{\sin \omega_y y_0}{\omega_y y_0} \quad (52)$$

The intensity distribution of the Fraunhofer diffraction produced by the rectangular aperture is

$$I_P = I_0 \frac{\sin^2 \omega_x x_0}{(\omega_x x_0)^2} \frac{\sin^2 \omega_y y_0}{(\omega_y y_0)^2} \quad (53)$$

The maximum intensities in the x and y directions occur at

$$\omega_x x_0 = \omega_y y_0 = 0 \quad (54)$$

The area of this rectangular aperture is defined as $A = 4x_0 y_0$, resulting in an expression for the maximum intensity of

$$I_0 = \left[\frac{i 4x_0 y_0 \alpha e^{-ik \cdot R_0}}{\lambda R_0} \right]^2 = \frac{A^2 \alpha^2}{\lambda^2 R_0^2} \quad (55)$$

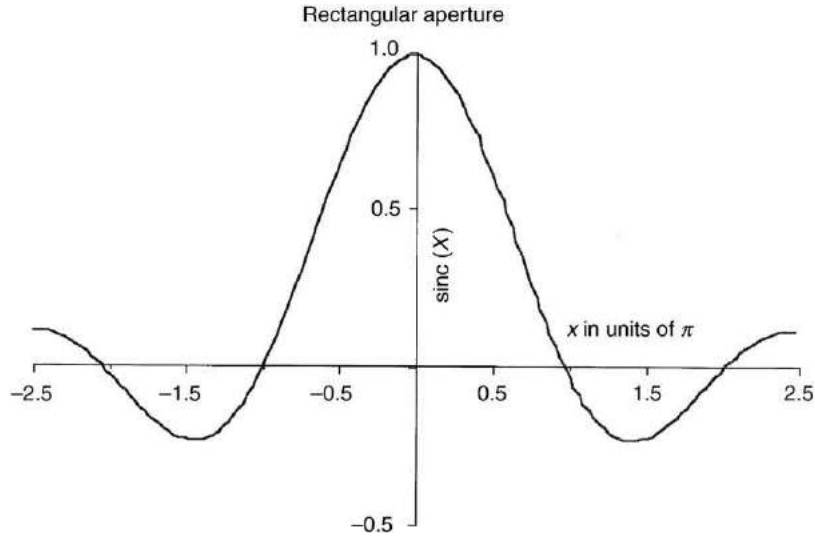


Fig. 5 The amplitude of the wave diffracted by a rectangular slit. This sinc function describes the light wave's amplitude that would exist in both the x and y directions. The coordinate would have to be scaled by the dimension of the slit. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The minima of (53) occur when $\omega_x x_0 = n\pi$ or when $\omega_y y_0 = m\pi$. The location of the zeros can be specified as a dimension in the observation plane or, using the paraxial approximation, in terms of the angle defined by (44)

$$\begin{aligned}\sin \theta_x \approx \theta_x &\approx \frac{\xi}{R_0} = \frac{n\lambda}{2x_0} \\ \sin \theta_y \approx \theta_y &\approx \frac{\eta}{R_0} = \frac{m\lambda}{2y_0}\end{aligned}\quad (56)$$

The dimensions of the diffraction pattern are characterized by the location of the first zero, i.e., when $n=m=1$, and are given by the observation plane coordinates, ξ and η . The dimensions of the diffraction pattern are inversely proportional to the dimensions of the aperture. As the aperture dimension expands, the width of the diffraction pattern decreases until, in the limit of an infinitely wide aperture, the diffraction pattern becomes a delta function.

Fig. 5 is a theoretical plot of the amplitude of the diffraction wave from a rectangular slit.

Diffraction from a Circular Aperture

To obtain the diffraction pattern from a circular aperture we first convert from the rectangular coordinate system we have used to a cylindrical coordinate system. The new cylindrical geometry is shown in Fig. 6. At the aperture plane

$$\begin{aligned}x &= s \cdot \cos \varphi & y &= s \cdot \sin \varphi \\ f(x, y) &= f(s, \varphi) & dx dy &= s ds d\varphi\end{aligned}\quad (57)$$

At the observation plane

$$\xi = \rho \cdot \cos \theta \quad \eta = \rho \cdot \sin \theta \quad (58)$$

In the new, cylindrical, coordinate system at the observation plane, the spatial frequencies are written as

$$\begin{aligned}\omega_x &= -\frac{k\xi}{R_0} = -\frac{k\rho}{R_0} \cos \theta \\ \omega_y &= -\frac{k\eta}{R_0} = -\frac{k\rho}{R_0} \sin \theta\end{aligned}\quad (59)$$

From Fig. 6 we see that the observation point, P, can be defined in terms of the angle ψ , where

$$\sin \psi = \frac{\rho}{R_0} \quad (60)$$

This allows an angular representation of the size of the diffraction pattern if it is desired.

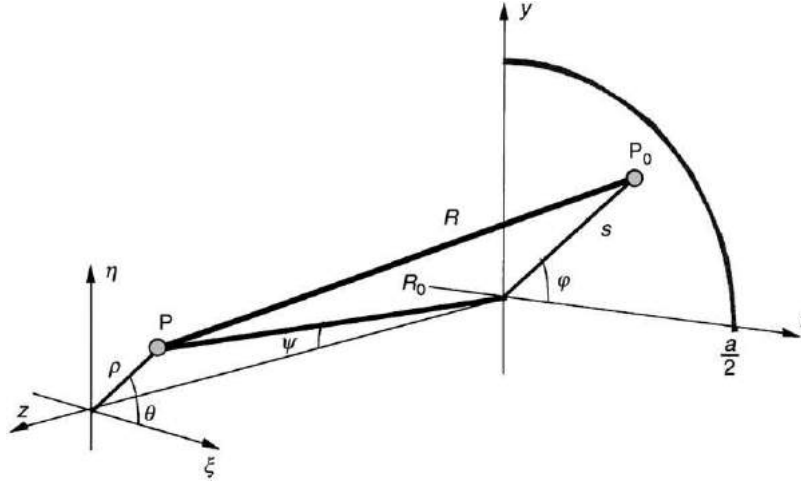


Fig. 6 Geometry for calculation of Fraunhofer diffraction from a circular aperture. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Using (57) and (59), we may write

$$\begin{aligned}\omega_x x + \omega_y y &= -\frac{ks\rho}{R_0} (\cos \theta \cos \varphi + \sin \theta \sin \varphi), \\ &= -\frac{ks\rho}{R_0} \cos(\theta - \varphi)\end{aligned}\quad (61)$$

The Huygens–Fresnel integral for Fraunhofer diffraction can now be written in terms of cylindrical coordinates:

$$\tilde{E}_P = \frac{i\alpha}{\lambda R_0} e^{-ikR_0} \int_0^{\frac{a}{2}} \int_0^{2\pi} f(s, \varphi) e^{-ik\frac{s\rho}{R_0} \cos(\theta - \varphi)} s ds d\varphi \quad (62)$$

We will use this equation to calculate the diffraction amplitude from a clear aperture of diameter a , defined by the equation

$$f(s, \varphi) = \begin{cases} 1 & s \leq \frac{a}{2}, \text{ all } \varphi \\ 0 & s > \frac{a}{2} \end{cases} \quad (63)$$

The symmetry of this problem is such that

$$f(s, \varphi) = f(s)g(\varphi) = f(s) \quad (64)$$

$$\mathcal{F}\{f(s, \varphi)\} = \mathcal{F}\{f(s)\} = F(\rho) \quad (65)$$

$$F(\rho, \theta) = \int_0^\infty f(s)s ds \int_0^{2\pi} e^{-i\rho \cdot s \cos(\varphi - \theta)} d\varphi \quad (66)$$

$$F(\rho) = \int_0^\infty f(s)s ds \int_0^{2\pi} e^{-i\rho \cdot s \cos \varphi} d\varphi \quad (67)$$

The second integral in (67) belongs to a class of functions called the Bessel function defined by the integral

$$J_n(s\rho) = \frac{1}{2\pi} \int_0^{2\pi} e^{-i[s\rho \sin \varphi - n\varphi]} d\varphi \quad (68)$$

In (67) the integral corresponds to the $n=0$, zero-order, Bessel function. Using this definition, we can rewrite (67) as

$$F(\rho) = \int_0^\infty f(s)J_0(s\rho)s ds \quad (69)$$

This transform is called the Fourier–Bessel transform or the Hankel zero-order transform.

Using these definitions, the transform of the aperture function, $f(s)$, is

$$F(\rho) = \int_0^{\frac{a}{2}} J_0(s\rho)s ds \quad (70)$$

We use the identity

$$xJ_1(x) = \int_0^x \chi J_0(\chi) d\chi \quad (71)$$

to obtain

$$F(\rho) = \frac{a}{2\rho} J_1\left(\frac{\rho a}{2}\right) \quad (72)$$

We can now write the amplitude of the diffracted wave as

$$\tilde{E}_P = \frac{i\alpha}{\lambda} e^{-ikR_0} \left[\frac{\pi a}{k\rho} J_1\left(\frac{ka\rho}{2R_0}\right) \right] \quad (73)$$

A plot of the function contained within the bracket of (73) is given in Fig. 7. If we define

$$u = \frac{ka\rho}{2R_0} \quad (74)$$

then the spatial distribution of intensity in the diffraction pattern can be written in a form known as the Airy formula,

$$I = I_0 \left[\frac{2J_1(u)}{u} \right]^2 \quad (75)$$

where we have defined

$$I_0 = \left(\frac{\alpha A}{\lambda R_0} \right)^2 \quad (76)$$

with A representing the area of the aperture, $A = \pi\left(\frac{a}{2}\right)^2$.

The intensity pattern described by (75) is called the Airy pattern. The intensity at $u=0$ is the same as was obtained for a rectangular aperture of the same area, because, in the limit,

$$\lim_{u \rightarrow 0} \left[\frac{2J_1(u)}{u} \right] = 1 \quad (77)$$

For the Airy pattern, 84% of the total area is contained in the disk between the first zeros of (75). Those zeros occur at $u = \pm 1.22\pi$, which corresponds to a radius in the observation plane of

$$\rho = \frac{1.22(\lambda R_0)}{a} \quad (78)$$

91% of the light intensity is contained within the circle bounded by the second minimum at $u = 2.233\pi$. The intensities in the secondary maxima of the diffraction pattern of a rectangular aperture (53) are much larger than the intensities in the secondary maxima of the Airy pattern of a circular aperture of equal area. The peak intensities, relative to the central maximum, of the first three secondary maxima of a rectangular aperture are 4.7%, 1.6%, and 0.8%, respectively. For a circular aperture, the same quantities are 1.7%, 0.04%, and 0.02%, respectively.

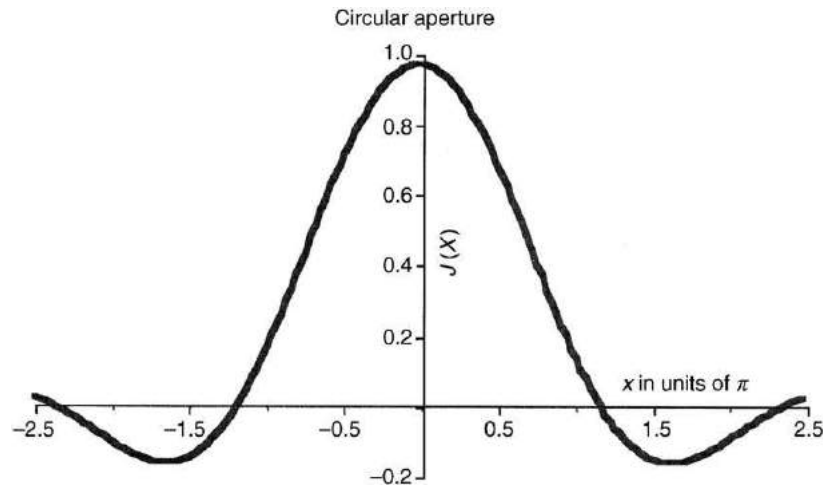


Fig. 7 Amplitude of a wave diffracted by a circular aperture. The observed light distribution is constructed by rotating this Bessel function around the optical axis. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Array Theorem

There is an elegant mathematical technique for handling diffraction from multiple apertures, called the array theorem. The theorem is based on the fact that Fraunhofer diffraction is given by the Fourier transform of the aperture function and utilizes the convolution integral.

The convolution of the functions $a(t)$ and $b(t)$ is defined as

$$g(t) = a(t) \otimes b(t) = \int_{-\infty}^{\infty} a(t)b(\tau - t) dt \quad (79)$$

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions:

$$\mathcal{F}\{a(t) \otimes b(t)\} = \mathcal{F}\left\{\int_{-\infty}^{\infty} a(t)b(\tau - t) dt\right\} = A(\omega)B(\omega) \quad (80)$$

We will demonstrate the array theorem for one dimension, where the functions represent slit apertures. The results can be extended to two dimensions in a straightforward way.

Assume that we have a collection of identical apertures, shown on the right of Fig. 8. If one of the apertures is located at the origin of the aperture plane, its transmission function is $\psi(x)$. The transmission function of an aperture located at a point, x_n , can be written in terms of a generalized aperture function, $\psi(x - \alpha)$, by the use of the sifting property of the delta function

$$\psi(x - x_n) = \int \psi(x - \alpha)\delta(\alpha - x_n) d\alpha \quad (81)$$

The aperture transmission function representing an array of apertures will be the sum of the distributions of the individual apertures, represented graphically in Fig. 8 and mathematically by the summation

$$\Psi(x) = \sum_{n=1}^N \psi(x - x_n) \quad (82)$$

The Fraunhofer diffraction from this array is $\Phi(\omega_x)$, the Fourier transform of $\Psi(x)$,

$$\Phi(\omega_x) = \int_{-\infty}^{\infty} \Psi(x)e^{-i\omega_x x} dx \quad (83)$$

which can be rewritten as

$$\Phi(\omega_x) = \sum_{n=1}^N \int_{-\infty}^{\infty} \psi(x - x_n)e^{-i\omega_x x} dx \quad (84)$$

We now make use of the fact that $\psi(x - x_n)$ can be expressed in terms of a convolution integral. The Fourier transform of $\psi(x - x_n)$ is, from the convolution theorem (80), the product of the Fourier transforms of the individual functions that make up the convolution:

$$\Phi(\omega_x) = \sum_{n=1}^N \mathcal{F}\{\psi(x - \alpha)\}\mathcal{F}\{\delta(\alpha - x_n)\} = \mathcal{F}\{\psi(x - \alpha)\} \sum_{n=1}^N \mathcal{F}\{\delta(\alpha - x_n)\} = \mathcal{F}\{\psi(x - \alpha)\}\mathcal{F}\left\{\sum_{n=1}^N \delta(\alpha - x_n)\right\} \quad (85)$$

The first transform in (85) is the diffraction pattern of the generalized aperture function and the second transform is the diffraction pattern produced by a set of point sources with the same spatial distribution as the array of identical apertures. We will call this second transform the array function.

To summarize, the array theorem states that the diffraction pattern of an array of similar apertures is given by the product of the diffraction pattern from a single aperture and the diffraction (or interference) pattern of an identically distributed array of point sources.

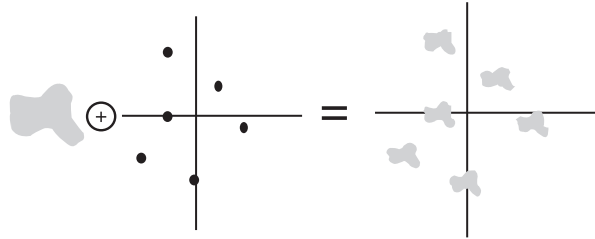


Fig. 8 The convolution of an aperture with an array of delta functions will produce an array of identical apertures, each located at the position of one of the delta functions. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

***N* Rectangular Slits**

An array of N identical apertures is called a diffraction grating in optics. The Fraunhofer diffraction patterns produced by such an array have two important properties; a number of very narrow beams are produced by the array and the beam positions are a function of the wavelength of illumination of the apertures and the relative phase of the waves radiated by each of the apertures.

Because of these properties, arrays of diffracting apertures have been used in a number of applications.

- At radio frequencies, arrays of dipole antennas are used to both radiate and receive signals in both radar and radio astronomy systems. One advantage offered by diffracting arrays at radio frequencies is that the beam produced by the array can be electrically steered by adjusting the relative phase of the individual dipoles.
- An optical realization of a two-element array of radiators is Young's two-slit experiment and a realization of a two-element array of receivers is Michelson's stellar interferometer. Two optical implementations of arrays, containing more diffracting elements, are diffraction gratings and holograms. In nature, periodic arrays of diffracting elements are the origin of the colors observed on some invertebrates.
- Many solids are naturally arranged in three-dimensional arrays of atoms or molecules that act as diffraction gratings when illuminated by X-ray wavelengths. The resulting Fraunhofer diffraction patterns are used to analyze the ordered structure of the solids.

The array theorem can be used to calculate the diffraction pattern from N rectangular slits, each of width a and separation d . The aperture function of a single slit is equal to

$$\psi(x, y) = \begin{cases} 1 & |x| \leq \frac{a}{2}, |y| \leq y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \quad (86)$$

The Fraunhofer diffraction pattern of this aperture has already been calculated and is given by (52):

$$\mathcal{F}\{\psi(x, y)\} = K \frac{\sin \alpha}{\alpha} \quad (87)$$

where the constant K and the variable α are

$$K = \frac{i2ay_0\alpha}{\lambda R_0} e^{-ikR_0} \frac{\sin \omega_y y_0}{\omega_y y_0} \quad (88)$$

$$\alpha = \frac{ka}{2} \sin \theta_x$$

The array function is

$$A(x) = \sum_{n=1}^N \delta(x - x_n) \quad \text{where } x_n = (n-1)d \quad (89)$$

and its Fourier transform is given by

$$\mathcal{F}\{A(x)\} = \frac{\sin N\beta}{\sin \beta} \quad (90)$$

$$\beta = \frac{kd}{2} \sin \theta_x$$

The Fraunhofer diffraction pattern's intensity distribution in the x -direction is thus given by

$$I_\theta = I_0 \underbrace{\frac{\sin^2 \alpha}{\alpha^2}}_{\text{shape factor}} \underbrace{\frac{\sin^2 N\beta}{\sin^2 \beta}}_{\text{grating factor}} \quad (91)$$

We have combined the variation in intensity in the y -direction into the constant I_0 because we assume that the intensity variation in the x -direction will be measured at a constant value of y .

A physical interpretation of (91) views the first factor as arising from the diffraction associated with a single slit; it is called the shape factor. The second factor arises from the interference between light from different slits; it is called the grating factor. The fine detail in the spatial light distribution of the diffraction pattern is described by the grating factor and arises from the coarse detail in the diffracting object. The coarse, overall light distribution in the diffraction pattern is described by the shape factor and arises from the fine detail in the diffracting object.

Young's Double Slit

The array theorem makes the analysis of Young's two-slit experiment a trivial exercise. This application of the array theorem will demonstrate that the interference between two slits arises naturally from an application of diffraction theory. The result of this analysis will support a previous assertion that interference describes the same physical process as diffraction and the division of the two subjects is an arbitrary one.

The intensity of the diffraction pattern from two slits is obtained from (91) by setting $N=2$:

$$I_\theta = I_0 \frac{\sin^2 \alpha}{\alpha^2} \cos^2 \beta \quad (92)$$

The sinc function describes the energy distribution of the overall diffraction pattern, while the cosine function describes the energy distribution created by interference between the light waves from the two slits. Physically, α is a measure of the phase difference between points in one slit and β is a measure of the phase difference between similar points in the two slits. Zeros in the diffraction intensity occur whenever $\alpha = n\pi$ or whenever $\beta = \frac{1}{2}(2n+1)\pi$. Fig. 9 shows the interference maxima, from the grating factor, under the central diffraction peak, described by the shape factor. The number of interference maxima contained under the central maximum is given by

$$\frac{2d}{a} - 1 \quad (93)$$

In Fig. 9 three different slit spacings are shown with the ratio d/a equal to 3, 6, and 9, respectively.

The Diffraction Grating

In this section we will use the array theorem to derive the diffraction intensity distribution of a large number of identical apertures. We will discover that the positions of the principal maxima are a function of the illuminating wavelength. This functional relationship has led to the application of a diffraction grating to wavelength measurements.

The diffraction grating normally used for wavelength measurements is not a large number of diffracting apertures but rather a large number of reflecting grooves cut in a surface such as gold or aluminum. The theory to be derived also applies to these

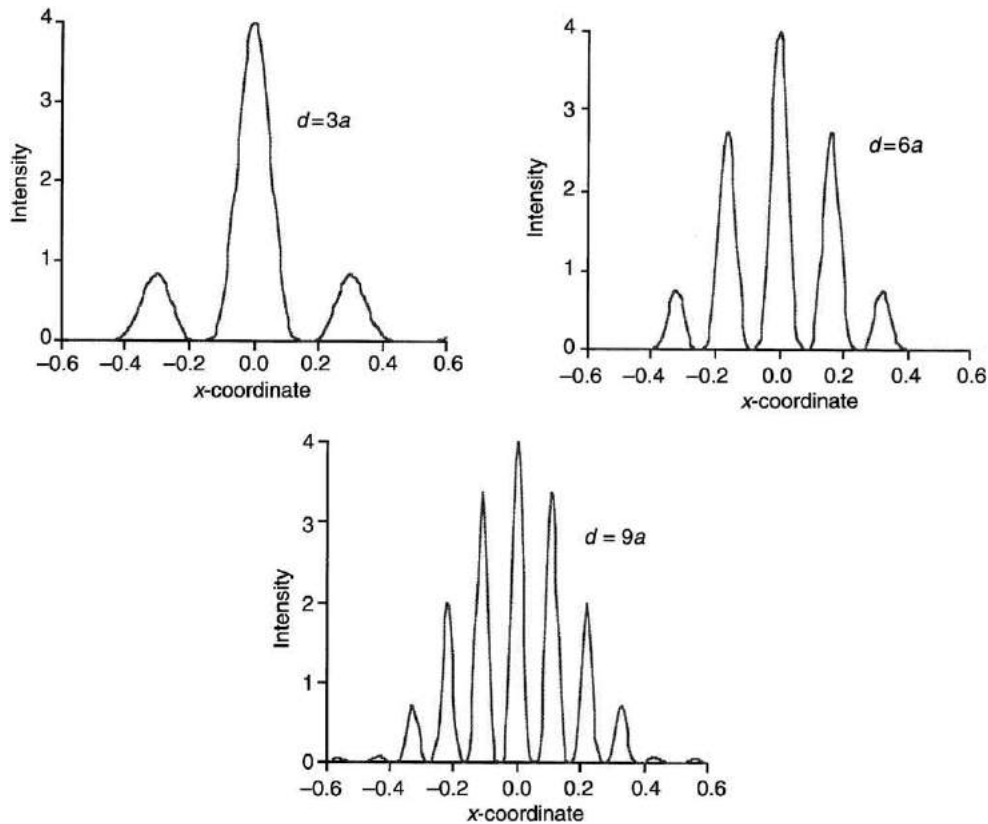


Fig. 9 The number of interference fringes beneath the main diffraction peak of a Young's two-slit experiment with rectangular apertures. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

reflection gratings but a modification must be made to the theory. The shape of the grooves in the reflection grating can be used to control the fraction of light diffracted into a principal maximum and we will examine how to take this into account. A grating whose groove shape has been controlled to enhance the energy contained in a particular principal maximum is called a blazed grating. The use of special groove shapes is equivalent to the modification of the phase of individual elements of an antenna array at radio frequencies.

The construction of a diffraction grating for use in an optical device for measuring wavelength was first suggested by David Rittenhouse, an American astronomer, in 1785, but the idea was ignored until Fraunhofer reinvented the concept in 1819. Fraunhofer's first gratings were fine wires spaced by wrapping the wires in the threads of two parallel screws. He later made gratings by cutting (*ruling*) grooves in gold films deposited on the surface of glass. H.A. Rowland made a number of well designed ruling machines, which made possible the production of large-area gratings. Following a suggestion by Lord Rayleigh, Robert Williams Wood (1868–1955) developed the capability to control the shape of the individual grooves.

If N is allowed to assume values much larger than 2, the appearance of the interference fringes, predicted by the grating factor, changes from a simple sinusoidal variation to a set of narrow maxima, called principal maxima, surrounded by much smaller, secondary maxima. To calculate the diffraction pattern we must evaluate (91) when N is a large number.

Whenever $N\beta = m\pi$, $m = 0, 1, 2, \dots$, the numerator of the second factor in (91) will be zero, leading to an intensity that is zero, $I_\theta = 0$. The denominator of the second factor in (91) is zero when $\beta = l\pi$, $l = 0, 1, 2, \dots$. If both conditions occur at the same time, the ratio of m and N is equal to an integer, $m/N = l$, and instead of zero intensity, we have an indeterminate value for the intensity, $I_\theta = 0/0$. To find the actual value for the intensity, we must apply L'Hospital's rule

$$\lim_{\beta \rightarrow l\pi} \frac{\sin N\beta}{\sin \beta} = \lim_{\beta \rightarrow l\pi} \frac{N \cos N\beta}{\cos \beta} = N \quad (94)$$

L'Hospital's rule predicts that whenever

$$\beta = l\pi = \frac{m}{N}\pi \quad (95)$$

where l is an integer, a principal maximum in the intensity will occur with a value given by

$$I_{\theta p} = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2} \quad (96)$$

Secondary maxima, much weaker than the principal maxima, occur when

$$\beta = \left(\frac{2m+1}{2N} \right) \pi \quad m = 1, 2, \dots \quad (97)$$

(When $m=0$, m/N is an integer, thus the first value that m can have in (97) is $m=1$.) The intensity of each secondary maximum is given by

$$I_{\theta s} = I_0 \frac{\sin^2 \alpha}{\alpha^2} \frac{\sin^2 N\beta}{\sin^2 \beta} = I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{\sin^2 \left(\frac{2m+1}{2} \right) \pi}{\sin^2 \left(\frac{2m+1}{2N} \right) \pi} \right] = I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{1}{\sin \left(\frac{2m+1}{2N} \right) \pi} \right]^2 \quad (98)$$

The quantity $(2m+1)/2N$ is a small number for large values of N , allowing the small angle approximation to be made:

$$I_{\theta s} \approx I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{2N}{\pi(2m+1)} \right]^2 \quad (99)$$

The ratio of the intensity of a secondary maximum and a principal maximum is given by

$$\frac{I_{\theta s}}{I_{\theta p}} = \left[\frac{2}{\pi(2m+1)} \right]^2 \quad (100)$$

The strongest secondary maximum occurs for $m=1$ and, for large N , has an intensity that is about 4.5% of the intensity of the neighboring principal maximum.

The positions of principal maxima occur at angles specified by the grating formula.

$$\beta = \frac{kd \sin \theta}{2} = l\pi = \frac{m}{N}\pi \quad (101)$$

The angular position of the principal maxima (the diffracted angle θ_d) is given by the Bragg equation

$$\sin \theta_d = \frac{l\lambda}{d} \quad (102)$$

where l is called the grating order. This simple relationship between the angle and the wavelength can be used to construct a device to measure wavelength, the grating spectrometer.

The model we have used to obtain this result is based on a periodic array of identical apertures. The transmission function of this array is a periodic square wave. If, for the moment, we treat the grating as infinite in size, we discover that the principal maxima in the diffraction pattern correspond to the terms of the Fourier series describing a square wave.

The zero order, $m=0$, corresponds to the temporal average of the periodic square wave and has an intensity proportional to the spatially averaged transmission of the grating. Because of its equivalence to the temporal average of a time-varying signal, the zero-order principal maximum is often called the dc term.

The first-order principal maximum corresponds to the fundamental spatial frequency of the grating and the higher orders correspond to the harmonics of this frequency.

The dc term provides no information about the wavelength of the illumination. Information about the wavelength of the illuminating light can only be obtained by measuring the angular position of the first or higher order principal maxima.

Grating Spectrometer

The curves shown in Fig. 10 display the separation of the principal maxima as a function of $\sin \theta_d$. The angular separation of principal maxima can be converted to a linear dimension by assuming a distance, r , from the grating to the observation plane. In the lower right-hand curve of Fig. 10, a distance of 2 meters was assumed. Grating spectrometers are classified by the size in meters of R used in their design. The larger r , the easier it is to resolve wavelength differences. For example, a 1 meter spectrometer is a higher-resolution instrument than a 1/4 meter spectrometer.

The fact that the grating is finite in size causes each of the orders to have an angular width that limits the resolution with which the illuminating wavelength can be measured. To calculate the resolving power of the grating, we first determine the angular width of a principal maximum. This is accomplished by measuring the angular change, of the principal maximum's position, when β changes from $\beta = l\pi = m\pi/N$ to $\beta = (m+1)\pi/N$, i.e., $\Delta\beta = \pi/N$. Using the definition of β

$$\beta = \frac{kd \sin \theta_d}{2} \quad (103)$$

$$\Delta\beta = \frac{\pi d \cos \theta_d \Delta\theta}{\lambda}$$

and the angular width is

$$\Delta\theta = \frac{\lambda}{Nd \cos \theta_d} \quad (104)$$

The derivative of the grating formula gives

$$\Delta\lambda = \frac{d}{l} \cos \theta_d \Delta\theta \quad (105)$$

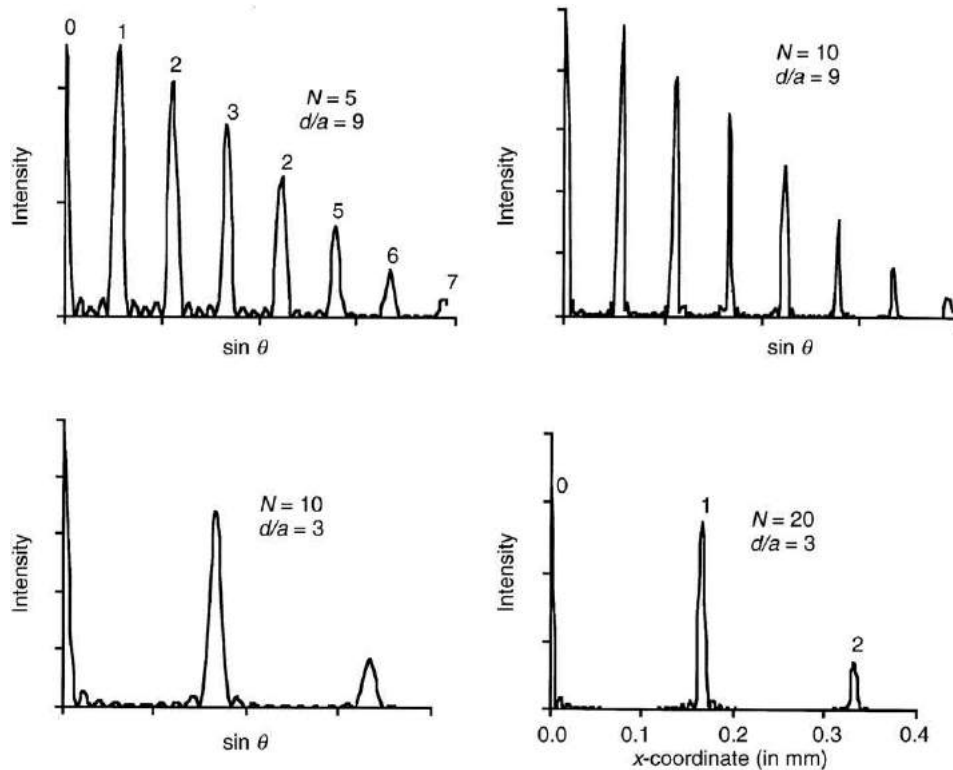


Fig. 10 Decrease in the width of the principal maxima of a transmission grating with an increasing number of slits. The various principal maxima are called orders, numbering from zero, at the origin, out to as large as seven in this example. Also shown is the effect of different ratios, d/a , on the number of visible orders. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\Delta\theta = \frac{l\Delta\lambda}{d \cos \theta_d}$$

Equating this spread in angle to Eq. (104) yields

$$\frac{l\Delta\lambda}{d \cos \theta} = \frac{\lambda}{Nd \cos \theta} \quad (106)$$

The resolving power of a grating is therefore

$$\frac{\lambda}{\Delta\lambda} = Nl \quad (107)$$

The improvement of resolving power with N can be seen in Fig. 10. A grating 2 inches wide and containing 15 000 grooves per inch would have a resolving power in second order ($l=2$) of 6×10^4 . At a wavelength of 600 nm this grating could resolve two waves, differing in wavelength by 0.01 nm.

The diffraction grating is limited by overlapping orders. If two wavelengths, λ and $\lambda + \Delta\lambda$, have successive orders that are coincident, then from Eq. (102)

$$(l+1)\lambda = l(\lambda + \Delta\lambda) \quad (108)$$

The minimum wavelength difference for which this occurs is defined as the free spectral range of the diffraction grating

$$(\Delta\lambda)_{\text{SR}} = \frac{\lambda}{l} \quad (109)$$

Blazed Gratings

We have been discussing amplitude transmission gratings. Amplitude transmission gratings have little practical use because they waste light. The light loss is from a number of sources.

1. Light is diffracted simultaneously into both positive and negative orders (the positive and negative frequencies of the Fourier transform). The negative diffraction orders contain redundant information and waste light.
2. In an amplitude transmission grating, light is thrown away because of the opaque portions of the slit array.
3. The width of an individual aperture leads to a shape factor for a rectangular aperture of $\text{sinc}^2 \alpha = (\sin^2 \alpha) / \alpha^2$, which modulates the grating factor and causes the amplitude of the orders to rapidly decrease. This can be observed in Fig. 10, where the second order is very weak. Because of the loss in intensity at higher orders, only the first few orders ($l=1,2$ or) are useful. The shape factor also causes a decrease in diffracted light intensity with increasing wavelength for higher-order principal maxima.
4. The location of the maximum in the diffracted light, i.e., the angular position for which the shape factor is a maximum, coincides with the location of the principal maximum due to the zero-order interference. This zero-order maximum is independent of wavelength and is not of much use.

One solution to the problems created by transmission gratings would be the use of a grating that modified only the phase of the transmitted wave. Such gratings would operate using the same physical processes as a microwave phased array antenna, where the location of the shape factor's maximum is controlled by adding a constant phase shift to each antenna element. The construction of an optical transmission phase grating with a uniform phase variation across the aperture of the grating is very difficult. For this reason, a second approach, based on the use of reflection gratings, is the practical solution to the problems listed above.

By tilting the reflecting surface of each groove of a reflection grating, Fig. 11, the position of the shape factor's maximum can be controlled. Problems 1, 3, and 4 are eliminated because the shape factor maximum is moved from the optical axis out to some angle with respect to the axis. The use of reflection gratings also removes Problem 2 because all of the incident light is reflected by the grating.

Robert Wood, in 1910, developed the technique of producing grooves of a desired shape in a reflective grating by shaping the diamond tool used to cut the grooves. Gratings, with grooves shaped to enhance their performance at a particular wavelength, are said to be blazed for that wavelength. The physical properties on which blazed gratings are based can be understood by using Fig. 11. The groove faces can be treated as an array of mirror surfaces. The normal to each of the groove faces makes an angle θ_B with the normal to the grating surface. We can measure the angle of incidence and the angle of diffraction with respect to the grating normal or with respect to the groove normal, as shown in Fig. 11. From Fig. 11, we can write a relationship between the angles

$$\theta_i = \varphi_i - \theta_B \quad -\theta_d = -\varphi_d + \theta_B \quad (110)$$

(The sign convention used defines positive angles as those measured in a counterclockwise rotation from the normal to the surface. Therefore, θ_B is a negative angle.) The blaze angle provides an extra degree of freedom that will allow independent adjustment of the angular location of the principal maxima of the grating factor and the zero-order, single-aperture, diffraction maximum. To see how this is accomplished, we must determine, first, the effect of off-axis illumination of a diffraction grating.

Off-axis illumination is easy to incorporate into the equation for the diffraction intensity from an array. To include the effect of an off-axis source, the phase of the illuminating wave is modified by changing the incident illumination from a plane wave of amplitude, E , traveling parallel to the optical axis, to a plane wave with the same amplitude, traveling at an angle θ_i to the optical

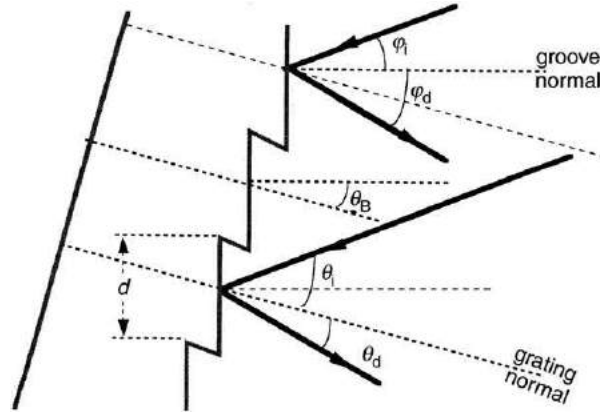


Fig. 11 Geometry for a blazed reflection grating. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

axis: $\tilde{E}e^{-ikx \sin \theta_i}$. (Because we are interested only in the effects in a plane normal to the direction of propagation, we ignore the phase associated with propagation along the z -direction, $kz \cos \theta_i$.) The off-axis illumination results in a modification of the parameter for single-aperture diffraction from

$$\alpha = \frac{ka}{2} \sin \theta_d \quad (111)$$

to

$$\alpha = \frac{ka}{2} (\sin \theta_i + \sin \theta_d) \quad (112)$$

and for multiple aperture interference from

$$\beta = \frac{kd}{2} \sin \theta_d \quad (113)$$

to

$$\beta = \frac{kd}{2} (\sin \theta_i + \sin \theta_d) \quad (114)$$

The zero-order, single-aperture, diffraction peak occurs when $\alpha = 0$. If we measure the angles with respect to the groove face, this occurs when

$$\alpha = \frac{ka}{2} (\sin \varphi_i + \sin \varphi_d) = 0 \quad (115)$$

The angles are therefore related by

$$\sin \varphi_i = -\sin \varphi_d \quad (116)$$

$$\varphi_i = -\varphi_d$$

We see that the single-aperture, diffraction maximum (the shape factor's maximum) occurs at the same angle that reflection from the groove faces occurs. We can write this result in terms of the angles measured with respect to the grating normal

$$\theta_i = -(\theta_d + 2\theta_B) \quad (117)$$

The blaze condition requires the single aperture diffraction maximum to occur at the l th principal maximum, for wavelength λ_B . At that position

$$l\pi = \frac{2\pi d}{\lambda_B} (\sin \theta_i + \sin \theta_d) \quad (118)$$

$$l\lambda_B = 2d \sin \frac{1}{2}(\theta_i + \theta_d) \cos \frac{1}{2}(\theta_i - \theta_d)$$

For the special geometrical configuration called the Littrow condition, where $\theta_i = \theta_d$, we find that (117) leads to the equation

$$l\lambda_B = 2d \sin \theta_B \quad (119)$$

By adjusting the blaze angle, the single-aperture diffraction peak can be positioned on any order of the interference pattern. Typical blaze angles are between 15° and 30° but gratings are made with larger blaze angles.

Further Reading

- Born, M., Wolf, E., 1970. Principles of Optics. New York: Pergamon Press.
- Guenther, R.D., 1990. Modern Optics. New York: Wiley.
- Haus, H.A., 1984. Waves and Fields in Optoelectronics. Englewood Cliffs, NJ: Prentice-Hall.
- Hecht, E., 1998. Optics, 3rd edn Reading, MA: Addison-Wesley.
- Jackson, J.D., 1962. Classical Electrodynamics. New York: Wiley.
- Klein, M.V., 1970. Optics. New York: Wiley.
- O'Neill, E.L., 1963. Introduction to Statistical Optics. Reading, MA: Addison-Wesley.
- Rossi, B., 1957. Optics. Reading, MA: Addison-Wesley.
- Rubinowicz, A., 1984. In: Wolf, E. (Ed.), Progress in Optics, Take Miyamoto–Wolf diffraction wave, vol. IV. North-Holland: Amsterdam, p. 201.
- Sommerfeld, A., 1964. Optics. New York: Academic Press.
- Stone, J.M., 1963. Radiation and Optics. New York: McGraw-Hill.

Fresnel Diffraction

BD Guenther, Duke University, Durham, NC, USA

© 2005 Elsevier Ltd. All rights reserved.

Fresnel was a civil engineer who pursued optics as a hobby, after a long day of road construction. Based on his own very high-quality observations of diffraction, Fresnel used the wave propagation concept developed by Christiaan Huygens (1629–1695) to develop a theoretical explanation of diffraction. We will use a descriptive approach to obtain the Huygens–Fresnel integral by assuming an aperture can be described by N pinholes which act as sources for Huygens' wavelets. The interference between these sources will lead to the Huygens–Fresnel integral for diffraction.

Huygens' principle views wavefronts as the product of wavelets from various luminous points acting together. To apply Huygens' principle to the propagation of light through an aperture of arbitrary shape, we need to develop a mathematical description of the field from an array of Huygens' sources filling the aperture. We will begin by obtaining the field from a pinhole that is illuminated by a plane wave,

$$\tilde{E}_i(\mathbf{r}, t) = \tilde{E}_i(\mathbf{r})e^{i\omega t}$$

Following the lead of Fresnel, we will use theory of interference to combine the fields from two pinholes and then generalize to N pinholes. Finally by letting the areas of each pinhole approach an infinitesimal value, we will construct an arbitrary aperture of infinitesimal pinholes. The result will be the Huygens–Fresnel integral.

We know that the wave obtained after propagation through an aperture must be a solution of the wave equation,

$$\nabla^2 \tilde{E} = \mu\epsilon \frac{\partial^2 \tilde{E}}{\partial t^2}$$

We will be interested only in the spatial variation of the wave so we need only look for solutions of the Helmholtz equation

$$(\nabla^2 + k^2)\tilde{E} = 0$$

The problem is further simplified by replacing this vector equation with a scalar equation,

$$(\nabla^2 + k^2)\tilde{E}(x, y, z) = 0$$

This replacement is proper for those cases where $\mathbf{n}E(x, y, z)$ [where $\mathbf{n}$ is a unit vector] is a solution of the vector Helmholtz equation. In general, we cannot substitute $\mathbf{n}E$ for the electric field $\mathbf{E}$ because of Maxwell's equation

$$\nabla \cdot \mathbf{E} = 0$$

Rather than working with the magnitude of the electric field, we should use the scalar amplitude of the vector potential. We will neglect this complication and assume the scalar, E , is a single component of the vector field, $\mathbf{E}$. A complete solution would involve a scalar solution for each component of $\mathbf{E}$.

The pinhole is illuminated by a plane wave, and the wave that leaves the pinhole will be a spherical wave which is written in complex notation as

$$\tilde{E}(\mathbf{r})e^{i\omega t} = A \frac{e^{-i\delta} e^{-ik \cdot \mathbf{r}}}{r} e^{i\omega t}$$

The complex amplitude

$$\tilde{E}(\mathbf{r}) = A \frac{e^{-i\delta} e^{-ik \cdot \mathbf{r}}}{r} \quad (1)$$

is a solution of the Helmholtz equation. The field at P_0 , in Fig. 1, from two pinholes: one at P_1 , located a distance $r_{01} = |\mathbf{r}_0 - \mathbf{r}_1|$ from P_0 , and one at P_2 , located a distance $r_{02} = |\mathbf{r}_0 - \mathbf{r}_2|$ from P_0 , is a generalization of Young's interference. The complex amplitude is

$$\tilde{E}(\mathbf{r}_0) = \frac{A_1}{r_{01}} e^{-ik \cdot \mathbf{r}_{01}} + \frac{A_2}{r_{02}} e^{-ik \cdot \mathbf{r}_{02}} \quad (2)$$

We have incorporated the phase δ_1 and δ_2 into the constants A_1 and A_2 to simplify the equations.

The light emitted from the pinholes is due to a wave, E_i , incident onto the screen from the left. The units of E_i are per unit area so to obtain the amount of light passing through the pinholes, we must multiply E_i by the areas of the pinholes. If $\Delta\sigma_1$ and $\Delta\sigma_2$ are the areas of the two pinholes, respectively, then

$$A_1 \propto \tilde{E}_i(\mathbf{r}_1) \Delta\sigma_1 \quad A_2 \propto \tilde{E}_i(\mathbf{r}_2) \Delta\sigma_2$$

$$\tilde{E}(\mathbf{r}_0) = C_1 \frac{\tilde{E}_i(\mathbf{r}_1)}{r_{01}} e^{-ik \cdot \mathbf{r}_{01}} \Delta\sigma_1 + C_2 \frac{\tilde{E}_i(\mathbf{r}_2)}{r_{02}} e^{-ik \cdot \mathbf{r}_{02}} \Delta\sigma_2 \quad (3)$$

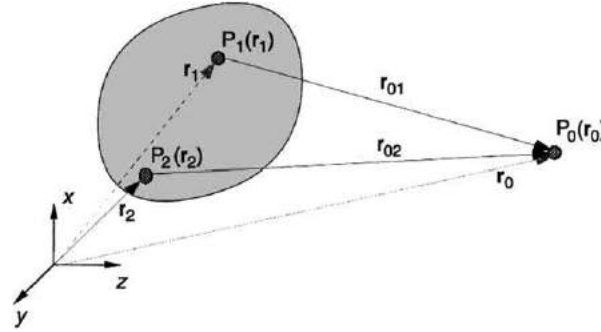


Fig. 1 Geometry for application of Huygens' principle to two pinholes. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

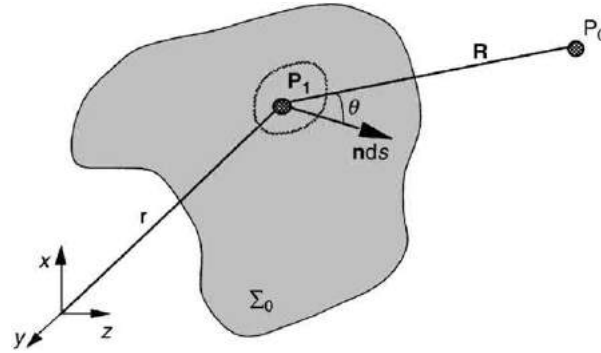


Fig. 2 The geometry for calculating the field at P_0 using (9). Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

where C_i is the constant of proportionality. The constant will depend on the angle that $\mathbf{r}_{0i}$ makes with the normal to $\Delta\sigma_i$. This geometrical dependence arises from the fact that the apparent area of the pinhole decreases as the observation angle approaches 90° .

We can generalize (3) to N pinholes

$$\tilde{E}(\mathbf{r}_0) = \sum_{j=1}^N C_j \frac{\tilde{E}_i(\mathbf{r}_j)}{r_{0j}} e^{-ikr_{0j}} \Delta\sigma_j \quad (4)$$

The pinhole's diameter is assumed to be small compared to the distances to the viewing position but large compared to a wavelength. In the limit as $\Delta\sigma_j$ goes to zero, the pinhole becomes a Huygens source. By letting N become large, we can fill the aperture with these infinitesimal Huygens' sources and convert the summation to an integral.

It is in this way that we obtain the complex amplitude, at the point P_0 , see Fig. 2, from a wave exiting an aperture, Σ_0 , by integrating over the area of the aperture. We replace $\mathbf{r}_{0j}$ in (4), by R , the position of P_1 , the infinitesimal Huygens' source of area, ds , measured with respect to the observation point, P_0 . We also replace $\mathbf{r}_j$ by $\mathbf{r}$, the position of the infinitesimal area, ds , with respect to the origin of the coordinate system. The discrete summation (4) becomes the integral

$$\tilde{E}(\mathbf{r}_0) = \int \int_A C(\mathbf{r}) \frac{\tilde{E}_i(\mathbf{r})}{R} e^{-ikR} d\mathbf{s} \quad (5)$$

This is the Fresnel integral. The variable $C(\mathbf{r})$ depends upon θ , the angle between $\mathbf{n}$, the unit vector normal to the aperture, and $\mathbf{R}$, shown in Fig. 2. We now need to determine how to treat $C(\mathbf{r})$.

The Obliquity Factor

When using Huygens' principle, a problem arises with the spherical wavelet produced by each Huygens' source. Part of the spherical wavelet propagates backward and results in an envelope propagating toward the light source, but such a wave is not observed in nature. Huygens neglected this problem. Fresnel required the existence of an obliquity factor to cancel the backward wavelet but was unable to derive its functional form. When Kirchhoff placed the theory of diffraction on a firm mathematical foundation, the obliquity factor was generated quite naturally in the derivation. In this intuitive derivation of the Huygens–Fresnel

integral, we will present an argument for treating the obliquity factor as a constant at optical wavelengths. We will then derive the constant value it must be assigned.

The obliquity factor, $C(\mathbf{r})$, in (5), causes the amplitude per unit area of the transmitted light to decrease as the viewing angle increases. This is a result of a decrease in the effective aperture area with viewing angle. When Kirchhoff applied Green's theorem to the scalar diffraction problem, he found the obliquity factor to have an angular dependence given by

$$\frac{\cos(\hat{\mathbf{n}}, \mathbf{R}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_s)}{2} \quad (6)$$

which includes the geometrical effect of the incident wave arriving at the aperture, at an angle with respect to the normal to the aperture from a source located at a position $\mathbf{r}_s$, with respect to the aperture. If the source is at infinity, then the incident wave can be treated as a plane wave, incident normal to the aperture, and we may simplify the angular dependence of the obliquity factor to

$$\frac{1 + \cos\theta}{2}$$

where $\theta \rightarrow (\hat{\mathbf{n}}, \mathbf{R})$ is the angle between the normal to the aperture and the vector $\mathbf{R}$. This is the configuration shown in Fig. 2. The obliquity factor provides an explanation of why it is possible to ignore the backward-propagating wave that occurs in application of Huygens' principle. For the backward wave, $\theta = \pi$, and the obliquity factor is zero.

The obliquity factor increases the difficulty of working with the Fresnel integral and it is to our benefit to be able to treat it as a constant. We can neglect the angular contribution of the obliquity factor by making an assumption about the resolving power of an optical system operating at visible wavelengths.

Assume we are attempting to resolve two stars that produce plane waves at the aperture of a telescope with an aperture diameter, a . The wavefronts from the two stars make an angle ψ with respect to each other, Fig. 3:

$$\tan\psi \approx \psi = \frac{\Delta x}{a}$$

The smallest angle, ψ , that can be measured is determined by the smallest length, Δx , that can be measured. We know we can measure a fraction of wavelength with an interferometer but, without an interferometer, we can measure a length no smaller than λ , leading to the assumption that $\Delta x \geq \lambda$. This reasoning leads to the assumption that the smallest angle we can measure is

$$\psi \geq \frac{\lambda}{a} \quad (7)$$

The resolution limit established by the above reasoning places a limit on the minimum separation that can be produced at the back focal plane of the telescope. The minimum distance on the focal plane between the images of star 1 and 2 is given by

$$d = f\psi \quad (8)$$

From (7)

$$d = \frac{\lambda f}{a}$$

The resolution limit of the telescope can also be expressed in terms of the cone angle produced when the incident plane wave is focused on the back focal plane of the lens. From the geometry of Fig. 3, the cone angle is given by

$$\tan\theta = \frac{a}{2f}$$

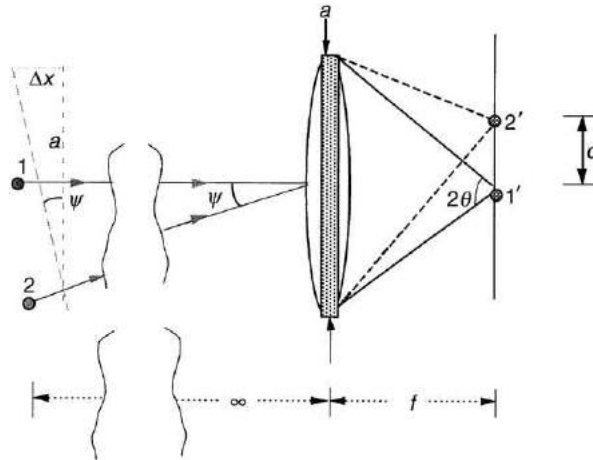


Fig. 3 Telescope resolution. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The separation between the stars 1 and 2 at the back focal plane of the lens in Fig. 3 is thus

$$d = \frac{\lambda}{2 \tan \theta} \quad (9)$$

If we assume that the minimum d is 3λ (this is four times the resolution of a typical photographic film at visible wavelengths), then the largest obliquity angle that should be encountered in a visible imaging system is

$$\tan \theta = \frac{\lambda}{2d} = \frac{\lambda}{6\lambda} = \frac{1}{6} = 0.167$$

$$\theta = 9.5^\circ \approx 10^\circ$$

yielding a value for $\cos \theta = 0.985$ and $(1 + \cos \theta)/2 = 0.992$. The obliquity factor has only changed by 0.8% over the angular variation of 0° to 10° . The obliquity factor as a function of angle is shown in Fig. 4 for an incident plane wave. The obliquity factor undergoes a 10% change when θ varies from 0° to about 40° ; therefore, the obliquity factor can safely be treated as a constant in any optical system that involves angles less than 40° .

While we have shown that it is safe to ignore the variability of C , we still have not assigned a value to the obliquity factor. To find the proper value for C , we will compare the result obtained using (5) with the result predicted by using geometric optics.

We illuminate the aperture Σ_0 in Fig. 5 with a plane wave of amplitude α , traveling parallel to the z -axis. Geometrical optics (equivalently the propagation of an infinite plane wave) predicts a field at P_0 , on the z -axis, a distance z_0 from the aperture, given by

$$\tilde{E}_{\text{geom}} = \alpha e^{-ikz_0} \quad (10)$$

The area of the infinitesimal at P_1 (the Huygens source) is

$$ds = r \, dr \, d\phi$$

Based on our previous argument, the obliquity factor can be treated as a constant, C , that can be removed from under the integral. The incident wave is a plane wave whose value at $z=0$ is $E(r) = \alpha$. Using these parameters, the Fresnel integral (5) can be written as

$$\tilde{E}(z_0) = C\alpha \iint \frac{e^{-ikR}}{R} r \, dr \, d\phi \quad (11)$$

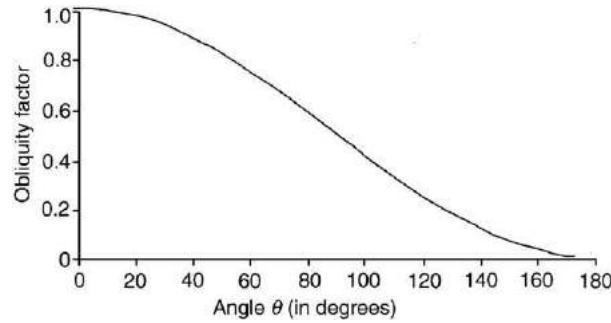


Fig. 4 The obliquity factor as a function of the angle defined in Eq. (6) for an incident plane wave. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

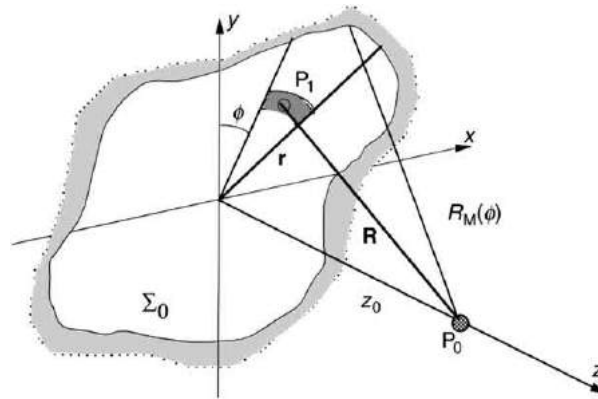


Fig. 5 Geometry for evaluating the constant in the Fresnel integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

From the geometry in Fig. 5, the distance from the Huygens source to the observation point is

$$z_0^2 + r^2 = R^2$$

where z_0 is a constant equal to the distance from the observation plane to the aperture plane. The variable of integration can be written in terms of R :

$$r \, dr = R \, dR$$

The limits of integration over the aperture extend from $R=z_0$ to the maximum value of R , $R=R_M(\phi)$.

$$\tilde{E}(z_0) = C\alpha \int_0^{2\pi} \int_{z_0}^{R_M(\phi)} e^{-ikR} dR \, d\phi \quad (12)$$

The integration over R can now be carried out to yield

$$\tilde{E}(z_0) = \underbrace{\frac{C\alpha}{ik} e^{-ikz_0} \int_0^{2\pi} d\phi}_{\text{Geometrical Optics}} - \underbrace{\frac{C\alpha}{ik} \int_0^{2\pi} e^{-ikR_M} d\phi}_{\text{Diffraction}} \quad (13)$$

The first integration in (13) is easy to perform and contains the amplitude at the observation point due to geometric optics. The second term may be interpreted as interference of the waves diffracted by the boundary of the aperture. An equivalent statement is that the second term is the interference of the waves scattered from the aperture's boundary, an interpretation of diffraction that was first suggested by Young.

The aperture is irregular in shape, at least on the scale of a wavelength, thus in general, kR_M will vary over many multiples of 2π as we integrate around the aperture. For this reason, we should be able to ignore the second integral in (13) if we confine our attention to the light distribution on the z -axis. After neglecting the second term, we are left with only the geometrical optics component of (13):

$$\tilde{E}(z_0) = \frac{2\pi C}{ik} \alpha e^{-ikz_0} \quad (14)$$

For (14) to agree with the prediction of geometric optics (10) the constant C must be equal to

$$C = \frac{ik}{2\pi} = \frac{i}{\lambda} \quad (15)$$

The Huygens–Fresnel integral can be written, using the value for C just derived,

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda} \int \int_{\Sigma} \frac{\tilde{E}_i(\mathbf{r})}{R} e^{-ikR} d\mathbf{s} \quad (16)$$

The job of calculating diffraction has only begun with the derivation of the Huygens–Fresnel integral (16). Rigorous solutions exist for only a few idealized obstructions. To allow discussion of the general properties of diffraction, it is necessary to use approximate solutions. To determine what type approximations we can make let's look at the light propagation path shown in Fig. 6.

For the second term in (13) to contribute to the field at point P_0 , in Fig. 6, the phase of the exponent that we neglected in (13) must not vary over 2π when the integration is performed over the aperture, i.e.,

$$\Delta = k \cdot (R_1 - R'_1 + R_2 - R'_2) < 2\pi$$

From the geometry of Fig. 6, the two paths from the source, S , to the observation point, P_0 , are

$$R_1 + R_2 = \sqrt{z_1^2 + (x_1 + b)^2} + \sqrt{z_2^2 + (x_2 + b)^2}$$

$$R'_1 + R'_2 = \sqrt{z_1^2 + x_1^2} + \sqrt{z_2^2 + x_2^2}$$

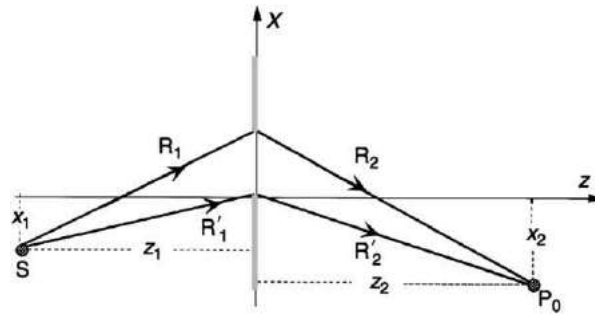


Fig. 6 One-dimensional slit used to establish Fresnel approximations. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

By assuming that the aperture, of width b , is small compared to the distances z_1 and z_2 , the difference between these distances, Δ , can be rewritten as a binomial expansion of a square root

$$\sqrt{1+b} = 1 + \frac{b}{2} - \frac{b^2}{8} + \dots \quad (17)$$

thus

$$\Delta \approx k \left(\frac{x_1}{z_1} + \frac{x_2}{z_2} \right) b + \frac{k}{2} \left(\frac{1}{z_1} + \frac{1}{z_2} \right) b^2 + \dots \quad (18)$$

If we assume that the second term of this expansion is small, i.e.,

$$\frac{k}{2} \left(\frac{1}{z_1} + \frac{1}{z_2} \right) b^2 \ll 2\pi$$

$$\frac{b^2}{2} \frac{1}{z_1} + \frac{1}{z_2} < \lambda$$

we see that the phase of the wavefront in the aperture is assumed to have a quadratic dependence upon aperture coordinates. Diffraction predicted by this assumption is called Fresnel diffraction.

To discover the details of the approximation consider Fig. 7. In the geometry of Fig. 7, the Fresnel integral (16) becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda} \iint_{\Sigma} f(x, y) \frac{e^{-ik(R+R')}}{RR'} dx dy \quad (19)$$

As we mentioned in our discussion of (13) the integral that adds diffractive effects to the geometrical optics field is nonzero only when the phase of the integrand is stationary. For Fresnel diffraction, we can insure that the phase is nearly constant if R does not differ appreciably from Z , or R' from Z' . This is equivalent to stating that only the wave in the aperture, around a point in Fig. 7 labeled S , called the stationary point, will contribute to E_p . The stationary point, S , is the point in the aperture plane where the line, connecting the source and observation positions, intersects the plane. Physically, only light propagating over paths nearly equal to the path predicted by geometrical optics (obtained from Fermat's principle) will contribute to E_{P_0} .

The geometry of Fig. 7 allows the distances R and R' to be written as

$$R^2 = (x - \xi)^2 + (y - \eta)^2 + Z^2 \quad (20)$$

$$R'^2 = (x_s - x)^2 + (y_s - y)^2 + Z'^2$$

To solve these expressions for R and R' , we apply the binomial expansion (17) and retain the first two terms:

$$R + R' \approx Z + Z' + \left[\frac{(x - \xi)^2}{2Z} + \frac{(x_s - x)^2}{2Z'} \right] + \left[\frac{(y - \eta)^2}{2Z} + \frac{(y_s - y)^2}{2Z'} \right] \quad (21)$$

In the denominator of (19), R is replaced by Z and R' by Z' . (By making this replacement, we are implicitly assuming that the amplitudes of the spherical waves are not modified by the differences in propagation distances over the area of integration. This is a reasonable approximation because the two propagation distances are within a few hundred wavelengths of each other). In the exponent, we must use (21), because the phase changes by a large fraction of a wavelength, as we move from point to point in the aperture.

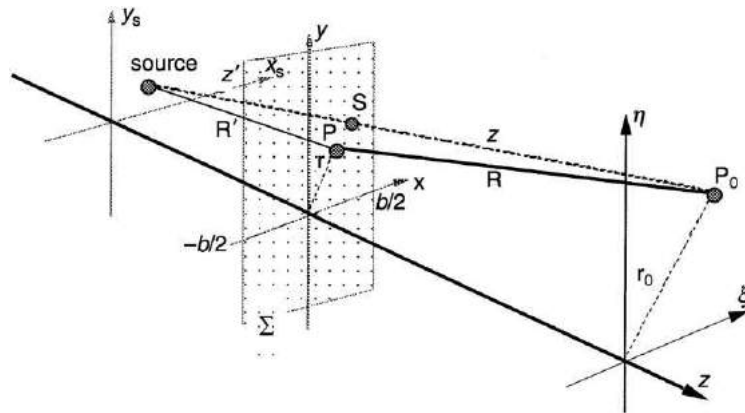


Fig. 7 Geometry for Fresnel diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the wave, incident on the aperture, were a plane wave we can assume $R' = \infty$ and (19) becomes

$$\tilde{E}_{P_0} = \frac{iz}{\lambda} \frac{e^{-ikZ}}{Z} \int \int_{\Sigma} f(x, y) e^{-\frac{ik}{2Z}[(x-\xi)^2 + (y-\eta)^2]} dx dy \quad (22)$$

The physical interpretation of (22) states that when a plane wave illuminates the obstruction, the field at point P_0 is a spherical wave, originating at the aperture, a distance Z away from P_0 :

$$\frac{e^{-ikZ}}{Z}$$

The amplitude and phase of this spherical wave are modified by the integral containing a quadratic phase dependent on the obstruction's spatial coordinates.

By defining three new parameters,

$$\rho = \frac{ZZ'}{Z + Z'} \quad \text{or} \quad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'} \quad (23a)$$

$$x_0 = \frac{Z'\xi + Zx_s}{Z + Z'} \quad (23b)$$

$$y_0 = \frac{Z'\eta + Zy_s}{Z + Z'} \quad (23c)$$

the more general expression for Fresnel diffraction of a spherical wave can be placed in the same format as (22). The newly defined parameters x_0 and y_0 are the coordinates, in the aperture plane, of the stationary point, S. The parameters in (23) can be used to express, after some manipulation, the spatial dependence of the phase, in the integrand of (19)

$$R + R' = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} + \left[\frac{(x - x_0)^2 + (y - y_0)^2}{2\rho} \right]$$

A further simplification to the Fresnel diffraction integral can be made by obtaining an expression for the distance, D , between the source and the observation point:

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} \quad (24)$$

We use this definition of D to also write

$$\frac{1}{ZZ'} = \frac{Z + Z'}{ZZ'} \frac{1}{Z + Z'} \approx \frac{1}{\rho D}$$

Using the parameters we have just defined, we may rewrite (19) as

$$\tilde{E}_{P_0} = \frac{iz}{\lambda \rho D} e^{-ikD} \int \int f(x, y) e^{-\frac{ik}{2\rho}[(x-x_0)^2 + (y-y_0)^2]} dx dy \quad (25)$$

With the use of the variables defined in (23), the physical significance of the general expression of the Huygens–Fresnel integral can be understood in an equivalent fashion as was (22). At point P_0 , a spherical wave

$$\frac{e^{-ikD}}{D}$$

originating at the source, a distance D away, is observed, if no obstruction are present. Because of the obstruction, the amplitude and phase of the spherical wave are modified by the integral in (25). This correction to the predictions of geometrical optics is called Fresnel diffraction.

Mathematically, (25) is an application of the method of stationary phase, a technique developed in 1887 by Lord Kelvin to calculate the form of a boat's wake. The integration is nonzero only in the region of the critical point we have labeled S. Physically, the light distribution, at the observation point, is due to wavelets from the region around S. The phase variations of light, coming from other regions in the aperture, are so rapid that the value of the integral over those spatial coordinates is zero. The calculation of the integral for Fresnel diffraction is complicated because, if the observation point is moved, a new integration around a different stationary point in the aperture must be performed.

Rectangular Apertures

If the aperture function, $f(x, y)$, is separable in the spatial coordinates of the aperture, then we can rewrite (25) as

$$\begin{aligned} \tilde{E}_{P_0} &= \frac{iz}{\lambda \rho D} e^{-ikD} \int_{-\infty}^{\infty} f(x) e^{-ig(x)} dx \int_{-\infty}^{\infty} f(y) e^{-ig(y)} dy \\ \tilde{E}_{P_0} &= A[C(x) - iS(x)][C(y) - iS(y)] \end{aligned}$$

A represents the spherical wave from the source, a distance D away,

$$A = \frac{i\alpha}{2D} e^{-ikD}$$

If we treat the aperture function as a simple constant, C and S are integrals of the form

$$C(x) = \int_{x_1}^{x_2} \cos[g(x)] dx \quad (26a)$$

$$S(x) = \int_{x_1}^{x_2} \sin[g(x)] dx \quad (26b)$$

The integrals, $C(x)$ and $S(x)$, have been evaluated numerically and are found in tables of Fresnel integrals. To use the tabulated values for the integrals, (26) must be written in a general form

$$C(w) = \int_0^w \cos\left(\frac{\pi u^2}{2}\right) du \quad (27)$$

$$S(w) = \int_0^w \sin\left(\frac{\pi u^2}{2}\right) du \quad (28)$$

The variable u is an aperture coordinate, measured relative to the stationary point, $S(x_0, y_0)$, in units of

$$\sqrt{\frac{\lambda \rho}{2}} \quad u = \sqrt{\frac{2}{\lambda \rho}}(x - x_0) \quad \text{or} \quad u = \sqrt{\frac{2}{\lambda \rho}}(y - y_0) \quad (29)$$

The parameter w in (27) and (28) specifies the location of the aperture edge relative to the stationary point S . The parameter w is calculated through the use of (29) with x and y replaced by the coordinates of the aperture edge.

The Cornu Spiral

The plot of $S(w)$ versus $C(w)$, shown in Fig. 8, is called the Cornu spiral in honor of M. Alfred Cornu (1841–1902) who was the first to use this plot for graphical evaluation of the Fresnel integrals.

To use the Cornu spiral, the limits w_1 and w_2 of the aperture are located along the arc of the spiral. The length of the straight line segment drawn from w_1 to w_2 gives the magnitude of the integral. For example, if there were no aperture present, then, for the x -dimension, $w_1 = -\infty$ and $w_2 = \infty$. The length of the line segment from the point $(-1/2, 1/2)$ to $(1/2, 1/2)$ would be the value of E_{P_0} , i.e., $\sqrt{2}$. An identical value is obtained for the y -dimension, so that

$$\tilde{E}_{P_0} = 2A = \frac{\alpha e^{-ikD}}{D}$$

This is the spherical wave that is expected when no aperture is present.

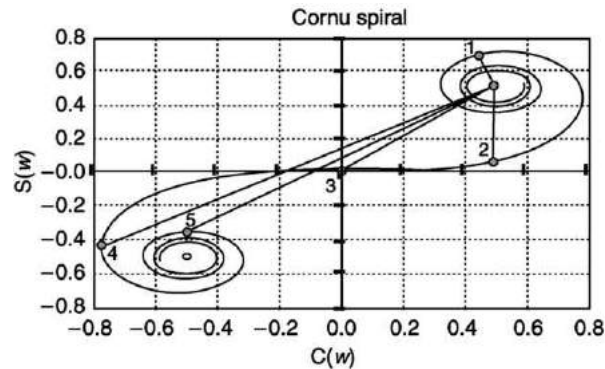


Fig. 8 Cornu spiral obtained by plotting the values for $C(w)$ versus $S(w)$ from a table of Fresnel integrals. The numbers indicated along the arc of the spiral are values of w . The points 1–5 correspond to observation points shown in Fig. 9 and are used to calculate the light intensity diffracted by a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

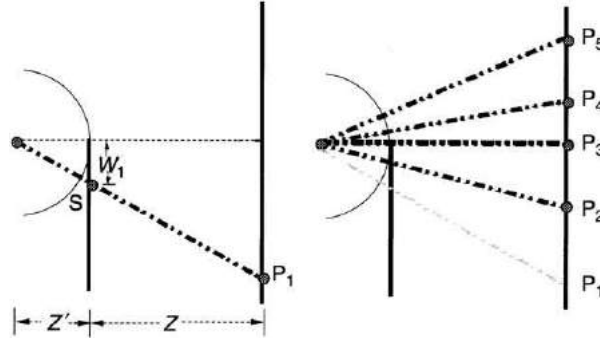


Fig. 9 A geometrical construct to determine w_1 for the stationary point S associated with five different observation points. The values of w are then used in Fig. 8 to calculate the intensity of light diffracted around a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

As an example of the application of the Fresnel integrals, assume the diffracting aperture is an infinitely long straight edge reducing the problem to one dimension:

$$f(x, y) = \begin{cases} 1 & x \geq x_1 \\ 0 & x < x_1 \end{cases}$$

The value of E_{P_0} in the y -dimension is from above $\sqrt{2}$ since $w_1 = -\infty$ and $w_1 = \infty$. In the x -dimension, the value of w at the edge is

$$w_{x_1} = \sqrt{\frac{2}{\lambda \rho}} (x_1 - x_0)$$

We will assume that the straight edge blocks the negative half-plane so that $x_1 = 0$. The straight edge is treated as an infinitely wide slit with one edge located at infinity so that $w_2 = \infty$ in the upper half-plane. When the observation point is moved, the coordinate x_0 of S changes and the origin of the aperture's coordinate system moves. New values for the w 's must therefore be calculated for each observation point.

Fig. 9 shows the assumed geometry. A set of observation points, P_1 – P_5 , on the observation screen are selected. The origin of the coordinate system in the aperture plane is relocated to a new position (given by the position of S) when a new observation point is selected. The value of w_1 , the position of the edge with respect to S , must be recalculated for each observation point. The distance from the origin to the straight edge, w_1 , is positive for P_1 and P_2 , zero for P_3 , and negative for P_4 and P_5 .

Fig. 8 shows the geometrical method used to calculate the intensity values at each of the observation points in Fig. 9 using the Cornu spiral. The numbers labeling the straight line segments in Fig. 8 are associated with the labels of the observation points in Fig. 9.

To obtain an accurate calculation of Fresnel diffraction from a straight edge, a table of Fresnel integrals provides the input to the following equation

$$I_P = I_0 \left\{ \left[\frac{1}{2} - C(w_1) \right]^2 + \left[\frac{1}{2} - S(w_1) \right]^2 \right\}$$

where $I_0 = 2A^2$. Table 1, shows the values extracted from the table of Fresnel integrals to find the relative intensity at various observation points. The result obtained by using either method for calculating the light distribution in the observation plane, due to the straight edge in Fig. 9, is plotted in Fig. 10. The relative intensities at the observation points depicted in Fig. 9 are labeled on the diffraction curve of Fig. 10.

Fresnel Zones

Fresnel used a geometrical construction of zones to evaluate the Huygens–Fresnel integral. The Fresnel zone is a mathematical construct, serving the role of a Huygens source in the description of wave propagation. Assume that at time t , a spherical wavefront from a source at P_1 has a radius of R' . To determine the field at the observation point, P_0 , due to this wavefront, a set of concentric spheres of radii, Z , $Z + (\lambda/2)$, $Z + 2(\lambda/2)$, ..., $Z + j(\lambda/2)$, ... are constructed, where Z is the distance from the wavefront to the observation point on the line connecting P_1 and P_0 (see Fig. 11). These spheres divide the wavefront into a number of zones, $\zeta_1, \zeta_2, \dots, \zeta_j, \dots$, called Fresnel zones, or half-period zones.

We treat each zone as a circular aperture illuminated from the left by a spherical wave of the form

$$\tilde{E}_j(R') = \frac{Ae^{-ikR'}}{R'} = \frac{Ae^{-ikR'}}{R'}$$

Table 1 Fresnel integrals for straight edge

w_f	$C(w_f)$	$S(w_f)$	I_P/I_0	P_f
∞	0.5	0.5	0	
2.0	0.4882	0.3434	0.01	
1.5	0.4453	0.6975	0.021	P_1
1.0	0.7799	0.4383	0.04	
0.5	0.4923	0.0647	0.09	P_2
0	0	0	0.25	P_3
-0.5	-0.4923	-0.0647	0.65	
-1.0	-0.7799	-0.4383	1.26	P_4
-1.2	-0.7154	-0.6234	1.37	
-1.5	-0.4453	-0.6975	1.16	
-2.0	-0.4882	-0.3434	0.84	P_5
-2.5	-0.4574	-0.6192	1.08	
$-\infty$	-0.5	-0.5	1.0	

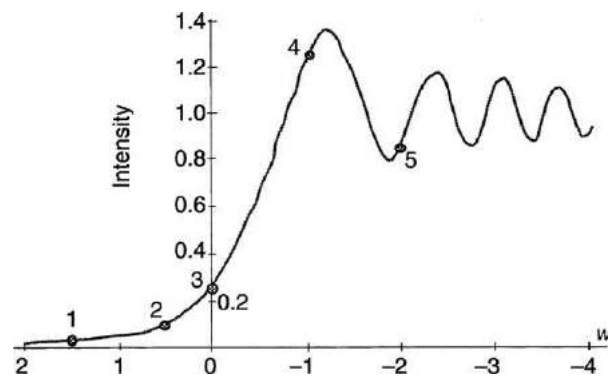


Fig. 10 Light diffracted by a straight edge with its edge located at $w=0$. The points labeled 1–5 were found using the construction in Fig. 9 to obtain w . This was then used in Fig. 8 to find a position on the Cornu spiral. The length of the lines shown in Fig. 8 from the $(1/2, 1/2)$ point to the numbered positions led to the intensities shown. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

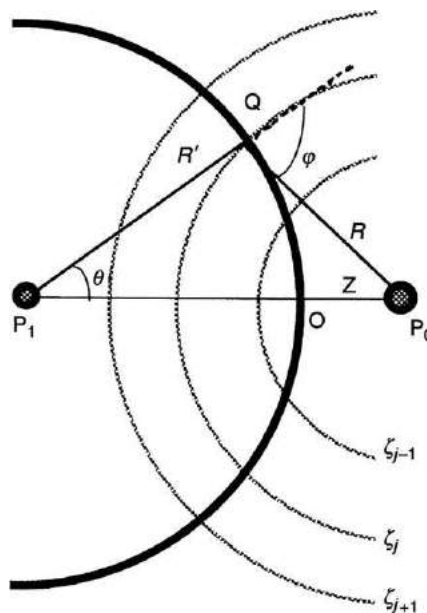


Fig. 11 Construction of Fresnel zones observed from a position P_0 on a spherical wave originating from a source at point P_1 . Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

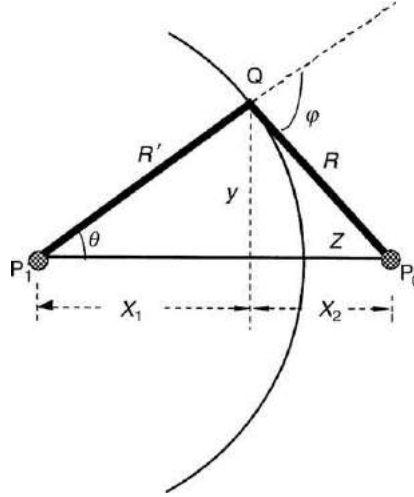


Fig. 12 Geometry for finding the relationship between R and R' yielding Eq. (32). Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

R' is the radius of the spherical wave. The field at P_0 due to the j th zone is obtained by using (5)

$$\tilde{E}_j(P_0) = \frac{A}{R} e^{-ik \cdot R'} \int \int_{\zeta_j} C(\varphi) \frac{e^{-ik \cdot R}}{R} ds \quad (30)$$

For the integration over the j th zone, the surface element is

$$ds = R' 2 \sin \theta d\theta d\phi \quad (31)$$

The limits of integration extend over the range

$$Z + (j-1)\frac{\lambda}{2} \leq R \leq Z + j\frac{\lambda}{2}$$

The variable of integration is R ; thus, a relationship between R' and R must be found. This is accomplished by using the geometrical construction shown in Fig. 12.

A perpendicular is drawn from the point Q on the spherical wave to the line connecting P_1 and P_0 , in Fig. 12. The distance from the source to the observation point is $x_1 + x_2$ and the distance from the source to the plane of the zone is the radius of the incident spherical wave, R' . The distance from P_1 to P_0 can be written

$$x_1 + x_2 = R' + Z$$

The distance from the observation point to the zone is

$$\begin{aligned} R^2 &= x_2^2 + y^2 = [(R' + Z) - x_1]^2 + (R' \sin \theta)^2 \\ &= R'^2 + (R' + Z)^2 - 2R'(R' + Z) \cos \theta \end{aligned}$$

The derivative of this expression yields

$$R dR = R'(R' + Z) \sin \theta d\theta \quad (32)$$

Substituting (32) into (31) gives

$$ds = \frac{R'}{R' + Z} R dR d\phi \quad (33)$$

The integration over ϕ is accomplished by rotating the surface element about the P_1P_0 axis. After integrating over ϕ between the limits of 0 and 2π , we obtain

$$\tilde{E}_j(P_0) = \frac{2\pi A}{R' + Z} e^{-ik \cdot R'} \int_{Z+(j-1)\frac{\lambda}{2}}^{Z+j\frac{\lambda}{2}} e^{-ik \cdot R} C(\varphi) dR \quad (34)$$

We assume that $R', Z \gg \lambda$ so that the obliquity factor is a constant over a single zone, i.e., $C(\varphi) = C_j$. The obliquity factor is not really a constant as we earlier assumed but rather a very slowly varying parameter of j , changing by less than two parts in 10^4 when j changes by 500 (for $Z=1$ meter and $\lambda=500$ nm). In fact it only changes by 2 parts in 10^{-7} across one zone. We are safe in applying the assumption that the obliquity factor is a constant over a zone and this allows the integral in (34) to be calculated:

$$\tilde{E}_j(P_0) = \frac{2\pi i C_j A}{k(R' + Z)} e^{-ik(R' + Z)} e^{-ikj\frac{\lambda}{2}} \left[1 - e^{-ik\frac{\lambda}{2}} \right] \quad (35)$$

Using the identity $k\lambda=2\pi$ and the definition for the distance between the source and observation point (24) modified for this geometry, $D=R'+Z$ (35) can be simplified to

$$\tilde{E}_j(P_0) = 2i\lambda(-1)^j \frac{C_j A}{D} e^{-ikD} \quad (36)$$

The physical reasons for the behavior predicted by (36) are quite easy to understand. The distance from P_0 to a zone changes by only $\lambda/2$ as we move from zone to zone and the area of a zone is almost a constant, independent of the zone number; thus, the amplitudes of the Huygens wavelets from each zone should be approximately equal. The alternation in sign, from zone to zone, is due to the phase change of the light wave from adjacent zones because the propagation paths for adjacent zones differ by $\lambda/2$.

To find the total field strength at P_0 due to N zones, the collection of Huygens' wavelets is added:

$$E(P_0) = \sum_{j=1}^N \tilde{E}_j(P_0) = \frac{2i\lambda A}{D} e^{-ikD} \sum_{j=1}^N (-1)^j C_j \quad (37)$$

To evaluate the sum, the elements of the sum are regrouped and rewritten as

$$-\sum_{j=1}^N (-1)^j C_j = \frac{C_1}{2} + \left(\frac{C_1}{2} - C_2 + \frac{C_3}{2}\right) + \left(\frac{C_3}{2} - C_4 + \frac{C_5}{2}\right) + \dots$$

Because the C 's are very slowly varying functions of j , even out to 500 zones, we are justified in setting the quantities in parentheses equal to zero. With this approximation, the summation can be set equal to one of two values, depending upon whether there is an even or odd number of terms in the summation

$$-\sum_{j=1}^N (-1)^j C_j = \begin{cases} \frac{1}{2}(C_1 + C_N) & N \text{ odd} \\ \frac{1}{2}(C_1 - C_N) & N \text{ even} \end{cases}$$

For very large N , the obliquity factor approaches zero, $C_N \rightarrow 0$, as was demonstrated in Fig. 4. Thus, the theory has led us to the conclusion that the total field produced by an unobstructed wave is equal to one half the contribution from the first Fresnel zone, i.e., $E=E_1/2$. Stating this result in a slightly different way, we obtain a surprising result – the contribution from the first Fresnel zone is twice the amplitude of the unobstructed wave!

The zone construction can be used to analyze the effect of the obstruction of all or part of a zone. For example, by constructing a circular aperture with a diameter equal to the diameter of the first Fresnel zone, we have just demonstrated that it is possible to produce an intensity at the point P_0 equal to four times the intensity that would be observed if no aperture were present. To analyze the effects of a smaller aperture, we subdivide a half-period zone into a number of subzones such that there is a constant phase difference between each subzone. The individual vectors form a curve known as the vibrational curve. Fig. 13 shows a vibrational curve produced by the addition of waves from nine subzones of the first Fresnel zone. The vibrational curve in Fig. 13 is an arc with the appearance of a half-circle. If the radius of curvature of the arc were calculated, we would discover that it is a constant, except for the contribution of the obliquity factor,

$$\rho = \frac{\lambda R'}{R' + Z} C(\varphi)$$

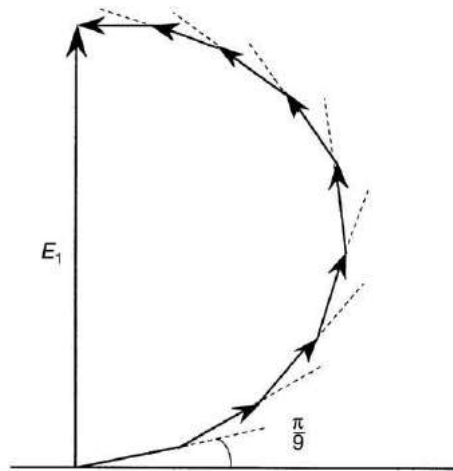


Fig. 13 Vector addition of waves from nine subzones constructed in the first Fresnel zone. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

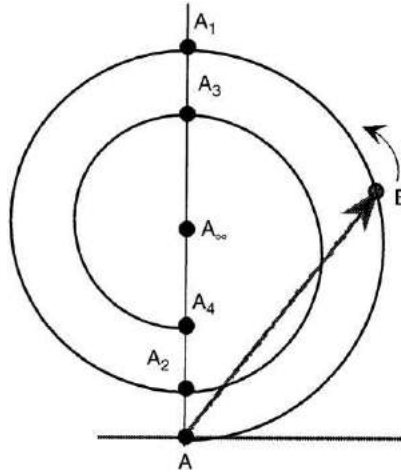


Fig. 14 Vibration curve for determining the Fresnel diffraction from a circular aperture. The change of the diameter of the half-circles making up the spiral has been exaggerated for easy visualization. The actual changes are one part in 10^{-5} . The position of point B is determined by the aperture size. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the obliquity factor for a single zone is a constant to 2 parts in 10^{-7} , the radius of curvature of the vibration curve can be considered a constant over a single zone. If we let the number of subzones approach infinity, the vibrational curve becomes a semicircle whose chord is equal to the wavelet produced by zone one, i.e., E_1 . If we subdivide additional zones, and add the subzone contributions, we create other half-circles whose radii decrease at the same rate as the obliquity factor. The vibrational curve for the total wave is a spiral, constructed of semicircles which converges to a point halfway between the first half-circle, see Fig. 14. The length of the vector from the start to the end of the spiral is $E = E_1/2$, as we derived above. When this same construction technique is applied to a rectangular aperture, the vibrational curve generated is the Cornu spiral.

Circular Aperture

The vector addition technique, described in Fig. 13, can be used to evaluate Fresnel diffraction, at a point P_0 , from a circular aperture and yields the intensity distribution along the axis of symmetry of the circular aperture. The zone concept will also allow a qualitative description of the light distribution normal to this axis.

To develop a quantitative estimate of the intensity at an observation point on the axis of symmetry of the circular aperture, we construct a spiral, Fig. 14, to represent the Fresnel zones of a spherical wave incident on the aperture. The point B on the spiral shown in Fig. 14 corresponds to the portion of the spherical wave unobstructed by the screen. The length of the chord, AB, represents the amplitude of the light wave at the observation point P_0 . As the diameter of the aperture increases, B moves along the spiral, in a counterclockwise fashion, away from A. The first maximum occurs when B reaches the point labeled A_1 in Fig. 14; the aperture has then uncovered the first Fresnel zone. At this point, the amplitude is twice what it would be with no obstruction. Four times the intensity!

If the aperture's diameter continues to increase, B reaches the point labeled A_2 in Fig. 14 and the amplitude is very nearly zero; two zones are now exposed in the aperture. Further maxima occur when an odd number of zones are in the aperture and further minima when an even number of zones are exposed. Fig. 15 shows an aperture containing four exposed Fresnel zones. The amplitude at the observation point would correspond to the chord drawn from A to A_4 in Fig. 14.

The aperture diameter can be fixed and the observation point P_0 can move along, or perpendicular to, the axis of symmetry of the circular aperture. As P_0 is moved away from the aperture, along the symmetry axis, i.e., as Z increases, the radius of the Fresnel zones increase without limit. For small values of n , the radius of the n th zone can be approximated by

$$r_n \approx \sqrt{nZ\lambda} \quad (38)$$

At $Z_{\max}$ the light intensity is a maximum, given by the chord length from A to A_1 in Fig. 14. If $\lambda = 500$ nm and $a = 0.5$ mm then this maximum occurs when $Z = 0.5$ m

$$Z_{\max} = \frac{a^2}{\lambda} \quad (39)$$

If we start at $Z_{\max}$ and move toward the aperture, along the axis, as Z decreases in value, a point will be reached when the intensity on the axis becomes a minimum. The value of Z where the first minimum in intensity is observed is equal to

$$Z_{\min} = \frac{a^2}{2\lambda}$$

In Fig. 14 the chord would extend from A to A_2 .

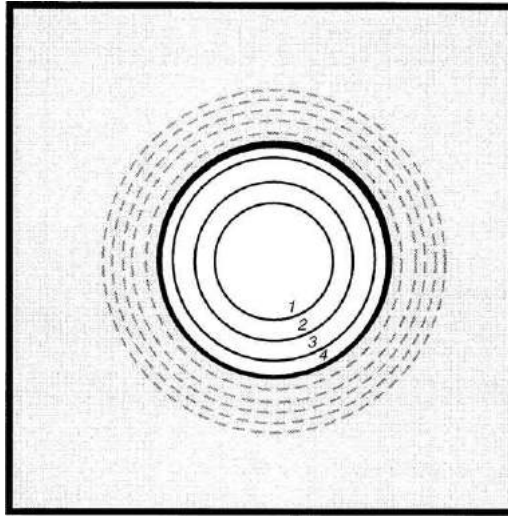


Fig. 15 Aperture with four Fresnel zones exposed. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

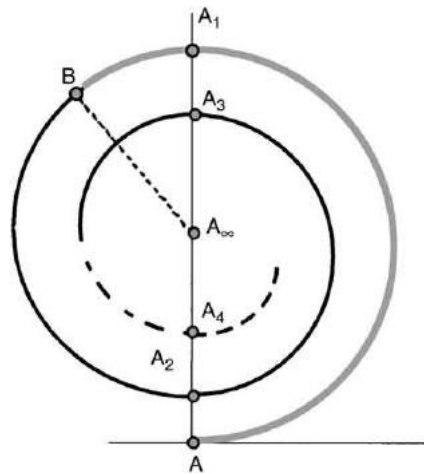


Fig. 16 Vibration curve for opaque disk. The shaded region makes no contribution because it is associated with the portion of the wave obstructed by the opaque object. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

As the observation point, P_0 , is moved along the axis toward the aperture, and Z assumes values less than $Z_{\min}$, the point B in Fig. 14 spirals inward, toward the center of the spiral and the intensity cycles through maximum and minimum values. The cycling of the intensity, as the observation point moves toward the aperture, will not continue indefinitely. At some point, the field on the axis will approach the field observed without the aperture because the distance between $Z_{\max}$ and $Z_{\min}$ shrinks to the wavelength of light making the intensity variation unobservable.

For values of Z that exceed $Z_{\max}$ of (39), the aperture radius, a , will be smaller than the radius of the first zone, and Fraunhofer diffraction will be observed because the aperture contains only one zone and the phase in the aperture is a constant.

Opaque Screen

If the screen containing a circular aperture, of radius a , is replaced by an opaque disk of radius a , the intensity distribution on the symmetry axis, behind the disk, is found to be equal to the value that would be observed with no disk present. This prediction was first derived by Poisson, to demonstrate that wave theory was incorrect; however, experimental observation supported the prediction and verified the theory.

We construct a spiral, shown in Fig. 16, similar to the one for the circular aperture in Fig. 14. The point B on the spiral represents the edge of the disk. The shaded portion of the spiral from A to B does not contribute because that portion of the wave is covered by the disk and the zones, associated with that portion of the wave, cannot be seen from the observation point.

The amplitude at P_0 is the length of the chord from B to A_∞ shown in Fig. 16. If the observation point moves toward the disk, then B moves along the spiral toward A_∞ . There is always intensity on the axis for this configuration, though it slowly decreases until it reaches zero when the observation point reaches the disk; this corresponds to point B reaching point A_∞ on the spiral. Physically, zero intensity occurs when the disk blocks the entire light wave. There are no maxima or minima observed as the disk diameter, a , increases or as the observation distance changes. If the observation point is moved perpendicular to the symmetry axis, a set of concentric bright rings are observed. The origin of these bright rings can be explained using Fresnel zones, in a manner similar to the one used to explain the bright rings observed in a Fresnel diffraction pattern from a circular aperture.

Zone Plate

In the construction of Fresnel zones, each zone was assumed to produce a Huygens wavelet, out of phase with the wavelets produced by its nearest neighbors. If every other zone were blocked, then there would be no negative contributions to (37). The intensity on-axis would be equal to the square of the sum of the amplitudes produced by the in-phase zones – exceeding the intensity of the incident wave. An optical component made by the obstruction of either the odd or the even zones could therefore be used to concentrate the energy in a light wave.

The boundaries of the opaque zones, used to block out-of-phase wavelets, are seen from (38) to increase as the square root of the integers. An array of opaque rings, constructed according to this prescription is called a zone plate, see Fig. 17.

A zone plate will perform like a lens with focal length

$$f = \pm \frac{r_1^2}{\lambda} \quad (40)$$

The zone plate, shown in Fig. 17, will act as both a positive and negative lens. What we originally called the source can now be labeled the object point, O, and what we called the observation point can now be labeled the image point, I. The light passing through the zone plate is diffracted into two paths, labeled C and D in Fig. 17. The light waves, labeled C, converge to a real image point I. For these waves the zone plate performs the role of a positive lens with a focal length given by the positive value of (40). The light waves, labeled D in Fig. 17, appear to originate from the virtual image point labeled I. For these waves the zone plate performs the role of a negative lens with a focal length given by the negative value of (40).

The zone plate will not have a single focus, as is the case for a refractive optical element, but rather will have multiple foci. As we move toward the zone plate, from the first focus, given by (40), the effective Fresnel zones will decrease in diameter. The zone plate will no longer obstruct out-of-phase Fresnel zones and the light intensity on the axis will decrease. However, additional maxima, of the on-axis intensity, will be observed at values of Z for which the first zone plate opening contains an odd number of zones. These positions can also be labeled as foci of the zone plate; however, the intensity at each of these foci will be less than the intensity at the primary focus.

Lord Rayleigh suggested that an improvement of the zone plate design would result if, instead of blocking every other zone, we shifted the phase of alternate zones by 180° . The resulting zone plate, called a phase-reversal zone plate, would, more efficiently, utilize the incident light. R.W. Wood was the first to make such a zone plate. Holography provides an optical method of

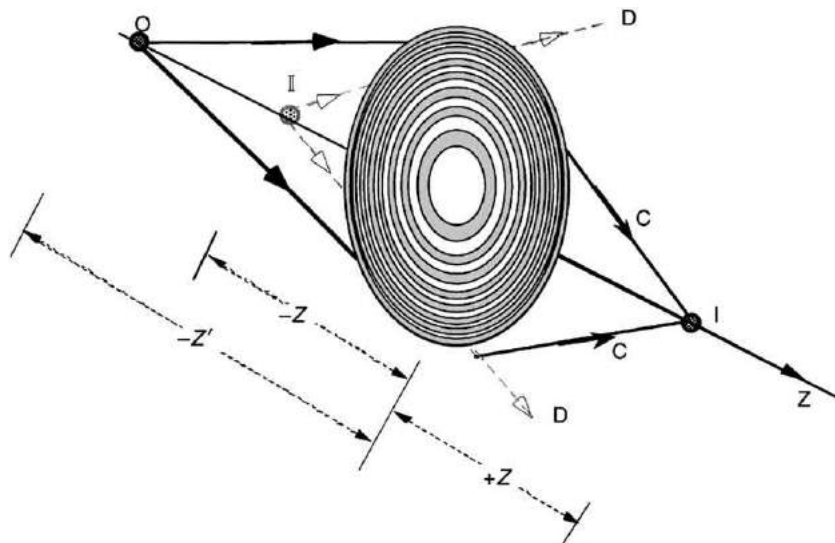


Fig. 17 A zone plate acts as if it were both a positive and a negative lens. Light from the object, O, is diffracted into waves traveling in both the D and the C directions. The light traveling in the C direction produces a real image of O at I. The light traveling in the D direction appears to originate from a virtual image at I. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

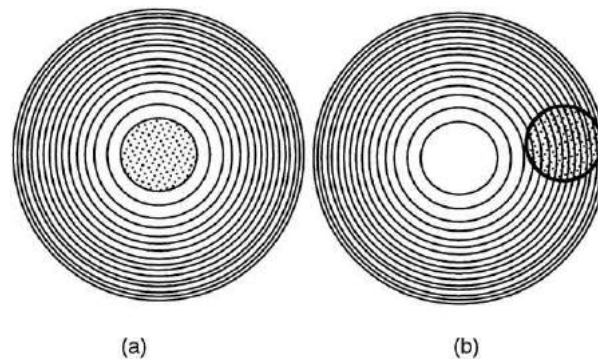


Fig. 18 (a) A set of Fresnel zones has been constructed about the optical path taken by some light ray. If the optical path of the ray is varied over the cross-hatched area shown in the figure, then the optical path length does not change. This cross-hatched area is equal to the first Fresnel zone and is described as the neighborhood of the light ray. (b) The neighborhood defined in (a) is moved so that it surrounds an incorrect optical path for a light ray. We see that this region of space would contribute no wave amplitude at the observation point because of the destructive interference between the large number of partial zones contained in the neighborhood. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

constructing the phase-reversal zone plates in the visible region of the spectrum. Semiconductor lithography has been used to produce zone plates in the x-ray region of the spectrum.

The resolving power of a zone plate is a function of the number of zones contained in the plate. When the number of zones exceeds 200, the zone plate's resolution approaches that of a refractive lens.

Fermat's Principle

The Fresnel zone construction provides physical insight into the interpretation of Fermat's principle which states that, if the optical path length of a light ray is varied in a neighborhood about the true path, there is no change in path length. By constructing a set of Fresnel zones about the optical path of a light ray, we can discover the physical meaning of a neighborhood.

The rules for constructing a Fresnel zone require that all rays passing through a given Fresnel zone have the same optical path length. The true path will pass through the center of the first Fresnel zone constructed on the actual wavefront. A neighborhood must be the area of the first Fresnel zone, for it is over this area that the optical path length does not change. **Fig. 18(a)** shows the neighborhood, about the true optical path, as a cross-hatched region equal to the first Fresnel zone.

Light waves do travel over wrong paths but we do not observe them because the phase differences for those waves that travel over the 'wrong' paths are such that they destructively interfere. By moving the neighborhood, defined in the previous paragraph as the first Fresnel zone, to a region displaced from the true optical path, we can use the zone construction to see that this statement is correct. In **Fig. 18(b)** the neighborhood is constructed about a ray that is improper according to Fermat's principle. We see that this region of space would contribute no energy at the observation point because of destructive interference between the large number of partial zones contained in the neighborhood.

This leads us to another interpretation of diffraction. If an obstruction is placed in the wavefront so that we block some of the normally interfering wavelets, we see light not predicted by Fermat's principle, i.e., we see diffraction.

Further Reading

- Backer, B.B., Copson, E.J., 1969. *The Mathematical Theory of Huygens' Principle*. London: Oxford University Press.
- Born, M., Wolf, E., 1970. *Principles of Optics*. New York: Pergamon Press.
- Guenther, R.D., 1990. *Modern Optics*. New York: Wiley.
- Haus, H.A., 1984. *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.
- Hecht, E., 1998. *Optics*, 3rd edn. Reading, MA: Addison-Wesley.
- Klein, M.V., 1970. *Optics*. New York: Wiley.
- Rubinowicz, A., 1984. In: Wolf, E. (Ed.), *Tahe Miyamoto-Wolf diffraction wave*, *Progress in Optics*, vol. IV. .
- Sayanagi, I., 1967. Pinhole Imagery. *Journal of the Optical Society of America*. 57, 1091–1099.
- Sears, F.W., 1949. *Optics*. Reading, MA: Addison-Wesley.
- Sommerfeld, A., 1964. *Optics*. New York: Academic Press.
- Stegaliani, D.J., Mittra, R., Semonin, R.G., 1967. Resolving power of a zone plate. *Journal of the Optical Society of America* 57, 610–613.
- Stone, J.M., 1963. *Radiation and Optics*. New York: McGraw-Hill.

Early History of Quantum Electronics

CH Townes, University of California, Berkeley, CA, USA

© 2005 C H Townes. Published by Elsevier Science Ltd. All rights reserved

All the ideas essential to making a laser were known before 1930, but there was no operating laser before 1960. So why didn't the laser come sooner? There are several reasons. One important impediment to the laser invention was that a combination of ideas from quantum mechanics and from electrical engineering was needed, and these two fields were not well mixed in the early days. Another is that, while some physicists recognized that amplification could occur if there was a population inversion of states, they did not consider the coherence of such amplifications and did not recognize its importance or usefulness. It was just something that in principle could happen, but was not very interesting. What was required was a combination of physics and electrical engineering thinking and recognition of the coherence of such amplification. In addition the importance of such a development had to be visualized and recognized in a way that it led to very devoted work towards its achievement.

It is striking that lasers grew out of the study of microwave spectra of molecules, for which engineering and quantum mechanics were both important and which helped orient thinking in the appropriate direction. That this origin was not accidental is convincingly demonstrated by the fact that three independent ideas for amplification by stimulated emission were generated about 1950. They were from Joe Weber, at the University of Maryland, from Nikolai Basov and Alexander Prokhovov at the Soviet Academy, and myself at Columbia University. Weber primarily wanted to point out the possibility, but didn't try to do it. In addition, his numbers were a bit off and no practical system was suggested. Basov and Prokhovov actually worked towards microwave amplification using a beam of molecules, as did I.

That stimulated amplification was recognized early, but not thought through, is illustrated by the fact that Prof. Richard Tolman, a theoretical physicist, wrote a discussion in 1924 of the net absorption of light by molecules, pointing out in particular that induced emission counteracted absorption and noting that if there were more molecules in the upper than in the lower state there could be 'negative absorption'. But, he wrote, 'This would usually be very small'. The Russian physicist Vitaly Ginsburg wrote me, after the maser and laser had appeared, that his professor, S.M. Levi, had been well aware of such effects back in the 1930s and had told him 'create an overpopulation at higher atomic levels and you will obtain an amplifier; the whole trouble is that it is difficult to create a substantial overpopulation of levels'.

The German physicist F.G. Houtermans said to me that in 1932, when told by a colleague of an unusual light intensity in a gaseous discharge, he had thought it might be a 'photon avalanche', i.e. multiplication of photons by stimulated emission.

Another Russian physicist, V.A. Fabricant, wrote a thesis in 1939 in which he discussed absorption and emission of light radiation in a gas and looked for 'negative absorption', or amplification. He did not discuss coherence or a resonant cavity, and was not able to achieve any amplification so his work was quickly forgotten. None of these early mentions of stimulated emission proposed how to actually get amplification, that it would be useful, nor clearly noted its coherence. Tolman did, however, write in 1927 that, 'We should expect radiation induced by an external field to be coherent with the radiation associated with that field'. I know of no proposal to actually make use of such amplification, until those made by microwave spectroscopists in the early 1950s.

Another clear indicator that even in the 1950s physicists and engineers did not think amplification by stimulated emission was particularly interesting nor useful is that during the 2½ years when Jim Gordon, Herb Zeiger, and I were working on trying to obtain microwave amplification (the maser), a large number of scientists visited my laboratory, saw what we were doing, but no one bothered to also try to obtain such amplification. After the maser worked, it hit the newspapers and then very quickly became a popular and intense field of interest for a number of physicists.

My own drive to produce oscillators by stimulated emission came from my strong interest in obtaining sources of waves shorter than the few millimeters wavelength which could be produced by electronic devices, in order to extend the high resolution study of molecular spectra to wavelengths shorter than microwaves, down into the far infrared. My students and I worked on several possible schemes for producing waves shorter than those produced by klystrons or magnetrons – frequency multiplication by nonlinearities, electronic beams passing over surfaces of solid materials with resonances, and anything else I could think of. None worked very well.

In early 1950, I was asked by the Office of Naval Research to form and chair a committee which would examine possible research towards obtaining wavelengths down to one millimeter and shorter. I chose outstanding scientists and engineers in a variety of fields which might touch on this problem. We met together, and visited many pertinent laboratories and individuals interested in such research. Nothing very promising seemed to turn up. But we wanted, of course, to at least provide a report summarizing the situation as we saw it. Our last meeting was April 26, 1951, in Washington, D.C. Worrying over our lack of success, I woke up early before the meeting. Breakfast was not ready, so I walked over to nearby Franklin Park, sat on a bench in front of beautiful blooming azaleas, and puzzled over why neither I nor the Committee had found any promising solutions.

I went over all the ideas I had had previously. Molecules, of course, can produce high frequencies. But I had previously concluded, I thought wisely and rigorously, that one could not obtain intense radiation from them because radiation intensity was a function of temperature, and the temperature could not be very high without destroying the molecules. Suddenly I realized that they did not need to have a temperature in the usual sense – they need not be in temperature equilibrium. There could be

more molecules in an upper than a lower excitation state, which could in principle produce indefinitely intense radiation. I pulled a piece of paper out of my pocket and wrote down the equations and numbers for such a case, using a molecule beam sent into a resonant cavity and ammonia molecules with which I was very familiar. My equations said one could get enough excited molecules, low enough loss in a resonant cavity, and it would work! Why hadn't I thought of it before?!

Back at Columbia, where I worked, and about 4 months later, the graduate student Jim Gordon agreed to work on trying to obtain such an oscillator using a beam of ammonia. I assured him that if it didn't work he could modify the experiment to do interesting spectroscopy and thus complete a thesis. But we both thought he had a good chance of making it work. And a young post doc working with me, Herb Zeiger, joined the effort.

Actually, I had first thought of obtaining stimulated emission from a molecular beam back in 1948, but simply as a demonstration of physical principles rather than as a useful amplification. Also, about a year later a young post doc, J.W. Trischka, working on molecular beams with Professors Rabi and Kusch, had also thought of demonstrating stimulated emission. We talked about it together, and he decided it wasn't worthwhile just demonstrating the effect because it wouldn't really prove any new physics. Neither of us at that time had recognized the real point and the possibility of useful amplification, which is one of the reasons mentioned above that the idea was delayed as long as it was.

It is perhaps important to emphasize again and to illustrate how out-of-the-way the use of stimulated emission was at that time for physicists, and how distant stimulated emission was from engineers. While we were working on the ammonia beam maser, Prof. L.H. Thomas, an outstanding physicist known for the 'Thomas Effect', frequently would run into me in the hallway at Columbia University and say that I didn't understand, and that my proposed ammonia oscillator could not work. However, I never got a clear explanation from him of what it was I didn't understand. And after we had been working on the ammonia maser system for about 2 years (a more-or-less normal time for a student thesis project), the Physics Department Chairman, Prof. Polycarp Kusch, and the previous chairman, Prof. I.I. Rabi, came into my office to object. They were excellent physicists, both received Nobel Prizes, and both were experts on molecular beams. They sat down in my office and said 'Look, Charlie, that is not going to work. We know it won't work and you know it won't work. You are wasting departmental money, and must stop!' Other people had also questioned what I was doing, frequently in particular whether the stimulated radiation would be coherent. So I had already thought over the situation many times. I had full notes from the quantum mechanics course I took as a student back in 1938, and could derive from them a proof of coherence. I felt also that I knew the quantitative numbers, such as the molecular beam intensity and the possible 'Q', or loss, in the cavity resonator, very well. I was also by then an Associate Professor, and the department chairman could not fire me simply because of stupidity. I replied to Rabi and Kusch 'No, I believe it has a reasonable chance of working and I'm going to continue!' Annoyed, they stomped out of the room.

About 2 months after the Rabi/Kusch incident, Jim Gordon dashed into the classroom where I was lecturing and said loudly 'It's working!' Most of the class then went up to the lab to see the new device. Rabi and Kusch were not against me; even though they were outstanding physicists, they probably just didn't quite comprehend the device. A couple of months after its successful operation, Kusch more or less apologized by saying 'Well, I should have realized that you probably know more about what you are doing than I do.'

After the maser successfully operated, there were other incidences showing the lack of appropriate focus of physicists on stimulated emission. I was a friend of Aage Bohr, the son of Niels Bohr, and in that connection was visiting him in Denmark. Niels Bohr asked me what research I was presently doing, so I told him about our new oscillator, the maser, and its remarkably pure frequency. He looked at me and said 'Oh no, that's not possible. You must be misunderstanding something.' I emphasized again what it was really doing, but he still seemed not to believe it could function that way. I presumed he was thinking in terms of the uncertainty principle and the finite time of passage of a molecule through the cavity, though I never was quite sure just why he felt it impossible. A similar thing happened shortly after that at a cocktail party in Princeton, where John Von Neumann asked the same question – what was my research at the moment? After telling him about the maser oscillator and the purity of frequency he reacted very similarly 'Oh no,' he said, 'That can't be right. You must be misunderstanding something.' After arguing a little more, he left to get another drink. Fifteen minutes later he came back, saying 'Hey, you are right!' He had understood, and wanted to talk much more about the maser and of possibly using excited semiconductors. Only after his death did I learn from his notebooks that he had considered exciting electrons in semiconductors with neutrons from a reactor, and had written Edward Teller about whether some experiments might be done with this to obtain intense light. However, he had not considered coherence, and Teller was apparently not interested enough to respond, so the matter was dropped.

The above account reminds me a bit of the amusing comments of Arthur Clark on 'Change'. He writes:

'People go through four states before any revolutionary development:

1. It's nonsense, don't waste my time
2. It's interesting, but not important
3. I always said it was a good idea
4. I thought of it first'.

It's clear that the world of physics was thinking very little in the direction of masers or lasers, and that many preconceptions as well as lack of interest stood in the way. My own experience with engineering at Bell Telephone Labs during World War II, in designing radar and electronic systems, plus my intense interest in obtaining short-wave oscillators, were clearly important in bringing me to the right ideas.

As masers became very interesting to the physics community and the field grew rapidly, my engineering experiences continued to be of help. I was well acquainted with the theoretical examination of noise in vacuum tube amplifiers by various individuals at Bell Labs, and recognized that the maser could provide much more sensitive amplification than could common electronic amplifiers, where discrete electronic charges produce the basic noise. On sabbatical in France, I worked on electron spin masers with Jean Combrisson and Arnold Honig who had the appropriate equipment. And then in Japan, Koichi Shimoda, Hidetoshi Takahashi, and I wrote a theoretical paper on the basic quantum noise of maser (or laser) amplification. The theory showed that maser amplifiers of microwaves should be about 100 times more sensitive than the existing electronic ones.

After 2–3 years of maser experiments and development I felt I wanted to move on to the shorter wavelengths for which I had generated the maser idea. Although I had first tried the idea at microwave frequencies because that seemed the easiest way to test out the general idea and the result had been exciting, I still wanted those shorter wavelengths. I had not come up with any great ideas of just how to get to much shorter wavelengths, which is why I waited several years after the maser worked before moving on. However, in 1957, 3 years after the first successful operation of the maser, I decided it was high time to simply figure out what was the best way I could imagine to move on into the infrared region and do it. A number of physicists had concluded that of course one couldn't make masers work at much shorter wavelengths, certainly not in the visible region, because spontaneous emission becomes so much faster as the wavelength is shortened and adequate inversion of population was not practical. But that was intuition, not quantitative science. As I wrote down equations for what might be done to move towards shorter wavelengths using a model of atomic or molecular excitation by radiation and a reasonably high Q resonator, it became clear that it was quite practical to move on even into the visible region. That was exciting! Why hadn't I, or someone else, looked at it carefully and quantitatively before that time?

I was at the time consulting at Bell Labs, with the assignment to spend a day there every 2 weeks and just talk with the Lab's scientific personnel. Since my former post doc and now brother-in-law Arthur Schawlow was then at Bell Labs, I of course talked with him. On telling him of my ideas for an 'optical maser', using a resonant cavity and excitation of atoms with optical radiation, he said he too was interested in that, and we decided to work together to optimize a system. It was Art who then suggested use of a Fabry–Perot resonator rather than the cavity with large holes I had used for a model, and that was an excellent addition. Why I did not think of that is a mystery, but Art had done his thesis at the University of Toronto in Fabry–Perot spectroscopy, and that might have been why the thought came to him.

Since Art Schawlow was participating, I decided the patent for the new 'optical maser' should belong to the Bell Labs (I already claimed ownership of the basic maser patent, which covered all wavelengths). Hence we recognized that the new idea must be kept confidential until Bell Labs lawyers had worked out and applied for an appropriate patent. This delay in public information sheds some additional light on how the scientific and technical world was thinking at the time. I had written down my original ideas for an 'optical maser' in my notebook and had it witnessed by my student Joe Giordmaine at Columbia University. I had also talked with a Columbia student Gordon Gould because he had been doing research with an intense light source and I wanted to know how much intensity he had in order to be sure I could get enough excited atoms and provide an oscillator at these short wavelengths. Except for these two persons, neither Art Schawlow nor I told anyone outside of Bell Labs about the 'optical maser', or laser idea until after Bell Labs had properly prepared its patent, which was about 11 months after my first notebook entry. For that entire time, no one has produced a record of any thoughts about extending the maser to optical wavelengths except Gordon Gould, who entered some ideas in his notebook about 1 month after I talked with him about the possibility of an optical maser, and 2 months after my original notebook entry. His notes were later to be the source of a long patent case.

The striking observation is that no one outside of Bell Labs except Gordon Gould, to whom I had explained my ideas, seems to have written or noted down anything about extending masers to these shorter wavelengths during the 11 months from my recognition that it could well be done until after the Schawlow and Townes paper on how to do it became available. After that there was considerable excitement and a number of ideas.

It is also noteworthy that, because of low general interest and competition, I did not publish a paper on how a maser might be made, but waited for publication until we demonstrated its operation, about 3 years after the idea had arrived. But after the maser worked the field became exciting and competitive. Hence, Schawlow and I thought we surely should publish a theoretical paper establishing the idea before taking the time to make a laser work. And indeed, there was much competition to build the first laser.

I myself helped a couple of my graduate students start work on trying to build an 'optical maser' or laser. But at about that time (1958), I was urged to undertake a job in Washington as Vice President and Director of Research of the Institute for Defense Analysis, an organization put together largely by the presidents of several universities to try to help advise the government on matters of science and technology. Sputnik had gone up the year before, and everyone was worried about the position of the U.S. with respect to Communist Russia, which seemed to be ahead particularly in some areas important to defense. I decided I should try to help, and accepted a two-year appointment in Washington. I recognized that this seriously distracted my attention from developing a laser quickly, but knew there were many others working towards lasers and so such devices would certainly be developed and our doing it just at Columbia University was neither critically important nor highly probable.

The field of masers and 'optical masers' or lasers was becoming so exciting and a bit scrambled that The Office of Naval Research asked me if I would organize a meeting on the subject. I did, with the help of a committee of many distinguished people in the field or closely related science and technology. And it was in a meeting of the Committee that we christened the field with the name 'Quantum Electronics'. This first international meeting on the subject was at Shawanga Lodge in New York State in September 1959. It made an occasion for the Russians Basov and Prokhorov to visit the United States (and my lab and home), and

was about 6 months before Ted Maiman made the first working laser. There were many interesting discussions of masers and their coming operation at optical wavelengths.

It is significant to note that, while the maser grew out of basic research in universities (with Russian work at the Russian Academy) and industry had little to do with early masers, all the first lasers were made in industrial laboratories. This illustrates the sociology and the strengths and weaknesses of industrial and of academic laboratories. While masers (and from them lasers) originated from research on microwave spectroscopy of molecules, industrial laboratories believed the field of microwave spectroscopy had little to offer in the way of commercial results. Because of equipment available and the interest of physicists in industry, the field was initiated and pursued immediately after World War II in commercial laboratories – by myself at Bell Labs, my friends at the RCA Labs, at Westinghouse, and at General Electric. For lack of interest in industry, such work was soon shut down, except at Bell Labs, and it moved to universities. Bell Labs generously allowed me to continue such work, although they wanted me to do some ‘more useful’ engineering. Once the field had obvious commercial possibilities, commercial laboratories began to support it well and the clear importance of masers and lasers made industry very interested. Really interested industry can more easily put strong new resources and push harder on a field than can academic laboratories, where money has to be granted and professors have a variety of other assignments. The first laser was made to work, of course, by Ted Maiman at Hughes Research Laboratories. Ted had been a student of Willis Lamb at Stanford, and there worked on radio spectroscopy. The second type of laser, rather similar to Maiman’s but using a different material, was made to work at the General Electric Laboratories by Peter Sorokin and Mirek Stevenson. Sorokin had been a student of Bloembergen at Harvard in microwave or radio spectroscopy and Stevenson was one of my students in microwave spectroscopy at Columbia University. The next type of laser, and one I particularly admire, was made by Ali Javan, William Bennett, and Don Herriott. Javan had been a student with me in microwave spectroscopy at Columbia University, Bill Bennett a student in the molecular beam group at Columbia working on radio spectroscopy, and Don Herriott was an optics specialist. All of these originators, with the exception of Herriott, had been working at universities in the field which originated the idea and were recently hired by industry. The next important laser, involving semiconductors, was invented by Robert Hall and collaborators at General Electric Labs. Note that every one of these early lasers was created in industry.

After the first laser was operated, my own students at Columbia quickly turned from trying to make a laser to using lasers to explore more physics, the normal university function. And I am delighted that masers and lasers have provided such excellent tools for research, as well as for commercial and medical applications.

After a few years of exploring new physics with lasers, I decided that since there were many excellent scientists doing such work, I should move into fields which it seemed to me were being relatively neglected. I moved to the University of California at Berkeley to look for molecules in interstellar space by microwave spectroscopy, and to do infrared astronomy. Very soon after initiating work at Berkeley, one of my students, Albert Cheung, not only discovered the first polyatomic molecules in space, he found powerful water masers. A while before our discovery of water and identification of its radiation as due to maser action, it had been deduced that some microwave radiation of OH must be due to maser action. And since then, many, many masers due to a wide variety of molecules have been found in astronomical sources as well as a few powerful lasers. Since deviations from thermal equilibrium are common in the very thin gases excited by powerful sources in space, we should have expected this, but didn’t. And these masers in space could have been easily detected with radio technology available back in the 1930s if anyone had searched the microwave spectrum carefully.

Clearly, masers and lasers could have been discovered and used much sooner than they actually were. We simply were neither thinking nor looking in the right directions. And this raises the natural question – what important and more-or-less obvious ideas are we missing now because of our lack of imaginative exploration?

Further Reading

- Basov, N., Prokhorov, A., 1954. Application of molecular beams to radio spectroscopic studies of rotation spectra of molecules. *J. Experimental. Theoretical Physics U.S.S.R.* 27, 431–438.
- Bloembergen, N., 1956. Proposal for a new type of solid state maser. *Physical Review* 104, 324–327.
- Cheung, A., Rank, D., Townes, C., Thornton, D., Welch, J., 1969. Detection of water in interstellar regions by its microwave radiation. *Nature* 121, 626–628.
- Combrisson, J., Honig, A., Townes, C., 1956. Utilization de la resonance de spins electroniques pour realiser un oscillateur ou un amplificateur en hyperfréquences. *Comptes Rendues* 242, 2451.
- Gordon, J., Zeiger, H., Townes, C., 1954. Molecular microwave oscillator and new hyperfine spectrum of NH_3 . *The Physical Review* 95, 282–284.
- Hall, R., Fenner, G., Kingsley, G., Soltys, J., Carlson, R., 1962. Coherent light emission from GaAs junction. *Physical Review Letters* 9, 366–368.
- Javan, A., Bennett, W., Herriot, D., 1961. Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture. *Physical Review Letters* 6, 106–110.
- Knowles, S., Mayer, C., Cheung, A., Rank, D., Townes, C., 1969. Spectra, variability, size, and polarization of H_2O microwave emission sources in the galaxy. *Science* 163, 1055–1057.
- Maiman, T., 1960. Stimulated optical radiation in ruby. *Nature* 187, 493–494.
- Perkins, F., Gold, T., Salpeter, E., 1966. Maser action in interstellar OH. *Ap. J.* 145, 361–366.
- Schawlow, A., Townes, C., 1958. Infrared and optical masers. *Physical Review* 112, 1940–1949.
- Shimoda, K., Takahashi, H., Townes, C., 1957. Fluctuations in amplification of quanta with application to maser amplifiers. *Journal of the Physical Society of Japan* 12, 686–700.
- Sorokin, P., Stevenson, M., 1960. Stimulated infrared emission from trivalent uranium. *Physical Review Letters* 5, 557.
- Tolman, R., 1924. Weak quantization. *The Physical Review* 24, 287–295.
- Tolman, R., 1927. Statistical mechanics with applications to physics and chemistry. The Chemical Catalog Co., New York. pp. 169.
- Weber, J., 1953. Amplification of microwave radiation by substances not in thermal equilibrium. *Trans. Inst. Radio Engineers Professional Group on Electron Devices* 3, 1–4.

Lenses and Mirrors

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Lenses are carefully shaped pieces of transparent material used to form images. The term 'transparent' does not necessarily mean that the lens is transmitting visible light. The element germanium, well-known in the transistor industry, is a light-gray opaque crystal which forms images with infrared radiation, and provides the optics for night-vision instruments. The lenses used in cameras, telescopes, microscopes, and other well-known applications, are made from special glasses or – more recently – from plastics. Hundreds of different optical glass formulas are listed in the manufacturers' catalogues. Plastics have come into use in recent years because it is now possible to mold them with high precision. The way in which lenses form images has been studied for several hundred years, resulting in an enormous and very complicated literature. In this article, we show how this theory can be simplified by introducing the restriction that lenses are perfect. This limitation provides an easy introduction to the image forming process.

The Cardinal Points and Planes of an Optical System

The location, size and orientation of an image are items of information which a designer needs to know. The nature of the image can be determined by a simple model of an optical system, based on two experimental observations. The first is well-known and is pictured in Fig. 1. Parallel rays of light passing through a double convex lens (a simple magnifying glass, for example) should meet at the focal point or focus, designated as F' . For rays from the right side of the lens, another such point F exists at an equal distance from the lens. The other experiment uses the lens to magnify (Fig. 2(a)). The amount of magnification decreases as the lens is brought closer to 'ISAAC NEWTON' (Fig. 2(b)), and when the lens touches the paper, object and image are approximately the same size. The locations of object and image corresponding to equality in size are called the unit or principal points H and H' , respectively, and the set of four points F , F' , H , and H' are the cardinal points. Therefore, the planes through these points normal to the axis of the lens are the cardinal planes. Fig. 3 shows an object whose image we can now determine. The coordinate orientation in this figure may appear strange at first. It is customary in optics to designate the lens axis as OZ , and the other two axes form a right-handed system. The x -axis then lies in the plane of the figure; the y -axis, coming out of the paper and perpendicular to it, is not shown. The object, of height x , is placed to the left of the focal point F . The cardinal points occur in the order F , H , H' , F' , since we expect – based on the experiment corresponding to Fig. 2 – that the unit points are very close to the lens. Another fact to be demonstrated later is that H and H' are actually inside the lens. Two rays are shown leaving the point P at the top of the object. The ray which is parallel to the axis strikes the unit plane through H and continues to the right, completely missing the lens. The ray then meets the unit plane through H' at a point which must be a distance x from the axis. To understand why, we recognize that if the object were relocated so as to lie on the object-space unit plane, the ray emerging at the other unit plane will be at a distance x from the axis; as required by the definition of these planes. We then assume that the ray is bent or refracted at the unit plane at H' and proceeds to, and beyond, the focal point F' . The second ray, from P through the focal point F , is easily traced since the lens is indifferent to the direction of the light. Assume temporarily that we know the location of the image point P' . Let a parallel ray leave this point and travel to the left. The procedure just given indicates that this ray will strike the unit plane at H' , emerge at the other unit plane, be refracted so as to pass through F and reach the object point P . Then reversing the direction of this ray, it will start at

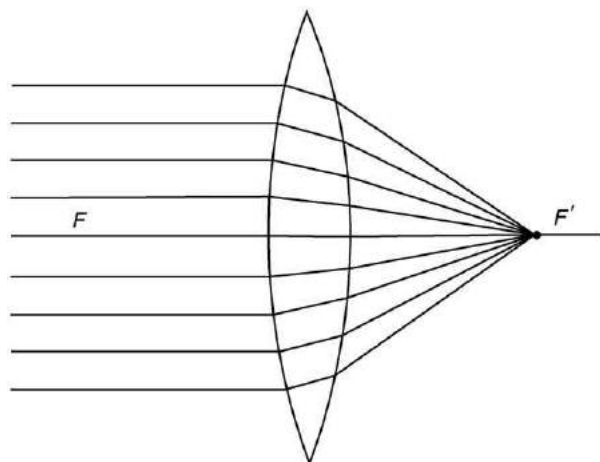


Fig. 1 Focal points of a double convex lens.

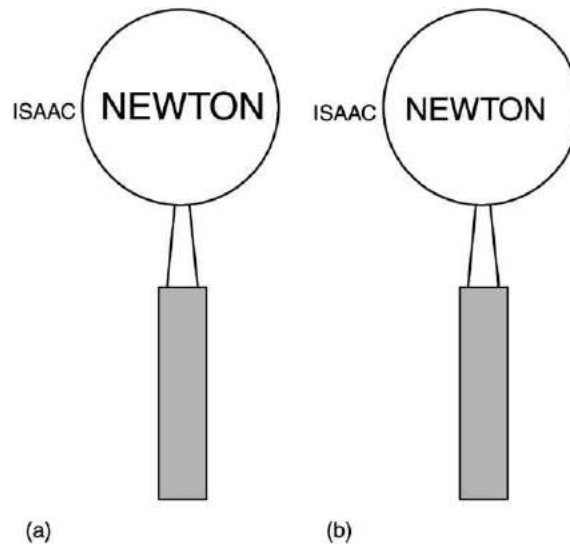


Fig. 2 Magnification by a double convex lens. The magnification in (a) is reduced when the lens is brought closer to the image (b).

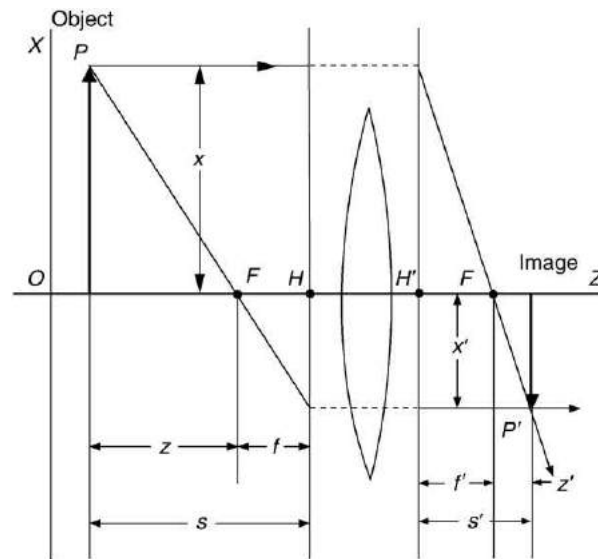


Fig. 3 Ray tracing using cardinal points and planes.

P , emerge at the unit plane H' and travel parallel to the axis, intersecting the other ray at P' . The distance between F and H is called the object space focal length f , with f' being the corresponding image space quantity. Simple geometry, making use of the similar triangles and the distances indicated in this diagram, can be used to derive the Newton lens equation or the equivalent Gauss lens equation. Detailed derivations of these two equations will be found in either of the sources listed in the Further Reading at the end of this article. It is not usually made clear in physics texts that these equations apply only to the special case of a single, thin lens, and we shall not apply them here. The image in this figure is real, inverted, and reduced. That is, it is smaller than the object, oriented in the opposite direction, and can be projected onto a screen located at the image position. This result can be verified by using an ordinary magnifier to form the image of a distant light source on a sheet of white paper.

Paraxial Matrices

To consider how light rays behave in optical systems, we introduce the index of refraction n of a material medium, defined as the ratio of the velocity of light c in a vacuum to its velocity v in that medium, or:

$$n = \frac{c}{v} \quad (1)$$

The velocity of light in air is approximately the same as it is in a vacuum. A light ray crossing the interface between air and a denser medium such as glass will be bent towards the normal to the surface. This is the phenomenon of refraction and the amount of bending is governed by Snell's law, which has the form:

$$n \sin \theta = n' \sin \theta' \quad (2)$$

where n and n' are the indices of refraction of the two media and θ and θ' are the angles as measured from the normal. Fig. 4 shows a ray of light leaving an object point P and striking the first surface of a lens at point P_1 . It is refracted there, and proceeds to the second surface. All rays shown lie in the z, x -plane or meridional plane, which passes through the symmetry axis OZ . The amount of refraction is specified by Snell's law, Eq. (2), which we shall now simplify. The sine of an angle can be approximated by the value of the angle itself, if this value is small when expressed in radians (less than 0.1 rad), and Snell's law simplifies to

$$n_1 \theta_1 = n'_1 \theta'_1 \quad (3)$$

This is called the paraxial form of Snell's law; the word paraxial means 'close to the axis'. It turns out, however, that the Snell's law angles are not convenient to work with, and we eliminate them with the terms:

$$\theta_1 = \alpha_1 + \phi, \quad \theta'_1 = \alpha'_1 + \phi \quad (4)$$

where α_1 is the angle which the incident ray makes with the axis OZ , α'_1 is the corresponding angle for the refracted ray, and ϕ is the angle which the radius r_1 (the line from C to P_1) of the lens surface makes at the center of curvature C . This particular angle can be specified as

$$\sin \phi = \frac{x_1}{r_1} \quad (5)$$

but using the paraxial approximation simplifies this to

$$\phi = \frac{x_1}{r_1} \quad (6)$$

Substituting for the angles gives

$$n'_1 \alpha'_1 = \frac{n_1 - n'_1}{r_1} + n_1 \alpha_1 \quad (7)$$

Notice that the distance from the point P_1 to the axis is labeled as either x_1 or x'_1 . This strange notation leads to the trivial relation

$$x_1 = x'_1 \quad (8)$$

and this can be combined with Eq. (7) to obtain the matrix equation:

$$\begin{pmatrix} n'_1 \alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1 \alpha_1 \\ x_1 \end{pmatrix} \quad (9)$$

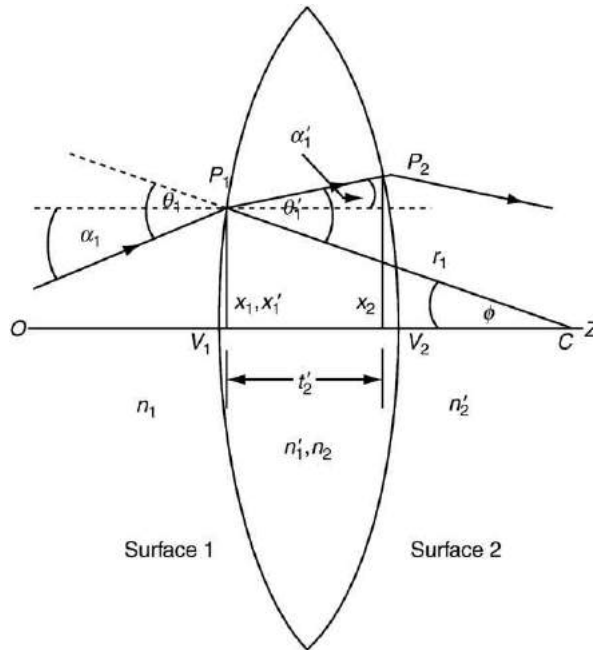


Fig. 4 Ray passing through a double convex lens.

where the constant k_1 is called the refracting power of surface 1 and is defined as

$$k_1 = \frac{n'_1 - n_1}{r_1} \quad (10)$$

To review the procedure for matrix multiplication, the first element $n'_1 \alpha'_1$ in the one-column product matrix is calculated by multiplying the upper left-hand corner of the 2×2 matrix with the first element of the one-column matrix to its right, obtaining 1 ($n_1 \alpha_1$), and to this is added the product $-k_1 x_1$ of the upper right-hand element in the 2×2 matrix and the second member of the one-column matrix. This square matrix is called the refraction matrix R_1 for surface 1, defined as

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \quad (11)$$

Next, we look at what happens to the ray as it travels from surface 1 to surface 2. As it goes from P_1 to P_2 , its distance from the axis becomes

$$x_2 = x'_1 + t'_1 \tan \alpha'_1 \quad (12)$$

or using the paraxial approximation for small angles:

$$x_2 = x'_1 + t'_1 \alpha'_1 \quad (13)$$

Another approximation we shall use is to regard t'_1 as being equal to the distance between the lens vertices V_1 and V_2 . Using the identity:

$$\alpha_2 = \alpha'_1 \quad (14)$$

leads to a second matrix equation:

$$\begin{pmatrix} n_2 \alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \begin{pmatrix} n'_1 \alpha'_1 \\ x'_1 \end{pmatrix} \quad (15)$$

where the 2×2 translation matrix T_{21} is defined as

$$T_{21} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \quad (16)$$

To get an equation which combines a refraction followed by a translation, we take the one-column matrix on the left side of [Eq. \(9\)](#) and substitute it into [Eq. \(15\)](#) to obtain

$$\begin{pmatrix} n_2 \alpha_2 \\ x_2 \end{pmatrix} = T_{21} R_1 \begin{pmatrix} n_1 \alpha_1 \\ x_1 \end{pmatrix} \quad (17)$$

Note that the 2×2 matrices appear in right to left order and that the multiplication of two square matrices is an extension of the rule given above. This equation gives the location and slope of the ray which strikes surface 2 after a refraction and a translation. To continue the ray trace at surface 2, we introduce a second refraction matrix R_2 by expressing k_2 in terms of the two indices at this surface and the radius. Then [Eq. \(17\)](#) is extended to give the relation:

$$\begin{pmatrix} n'_2 \alpha'_2 \\ x'_2 \end{pmatrix} = R_2 T_{21} R_1 \begin{pmatrix} n_1 \alpha_1 \\ x_1 \end{pmatrix} \quad (18)$$

This process can obviously be applied to any number of components. The product of the three 2×2 matrices in the above equation is known as the system matrix S_{21} and it completely specifies the effect of the lens on the incident ray passing through it. It is also written as

$$S_{21} = R_2 T_{21} R_1 = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \quad (19)$$

where the four quantities a , b , c , and d are known as the Gaussian constants. They are used extensively in specifying the behavior of a lens or an optical system. We now state the sign and notation conventions which are necessary for the application of paraxial matrix methods:

1. Light normally travels from left to right.
2. The optical systems we deal with are symmetrical about the z -axis. The intersections of the refracting or reflecting surfaces with this axis are the vertices and are designated in the order encountered as V_1 , V_2 , etc.
3. Positive directions along the axes are measured from the origin in the usual Cartesian fashion, so that horizontal distances (that is, along the z -axis) are positive if measured from left to right. Angles are positive when measured up from the z -axis.
4. Quantities associated with the incident ray are unprimed; those for the refracted ray are primed.
5. A subscript denotes the associated surface.
6. If the center of curvature of a surface is to its right, the radius is positive, and vice versa.
7. There is a special rule for mirrors to be explained below.

Using the Gaussian Constants

Fig. 5 shows the double convex lens of Fig. 4 with the assumed locations of the cardinal points indicated. These positions, which we shall now determine accurately, are at distances designated as l_F , l'_F , l_H , and l'_H and are measured from the associated vertex. The object position can be called t , a positive quantity which is measured to the right from object to the first vertex, or it can be called t_1 if measured in the opposite direction. For the first choice, the matrix specifying the translation from object to lens will have the quantity t/n_1 in its lower left-hand corner. However, it is both logical and convenient to use the first vertex as the reference point. Hence, we replace t/n_1 in the translation matrix with the quantity $-t_1/n_1$, and remember to specify t_1 as a negative number when calculating the image position. The equation connecting object and image is obtained by starting with this matrix, multiplying it by the system matrix of Eq. (19), and finally by a translation matrix corresponding to the translation t'_2 from lens to image to obtain:

$$\begin{pmatrix} n'_2 \alpha'_2 \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_2/n'_2 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ -t_1/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1 \alpha_1 \\ x \end{pmatrix} \quad (20)$$

where α or α_1 is the angle that the ray from the top of the object makes with the z -axis. This equation takes the initial value of the inclination α and of the position x of the ray leaving the object, and determines the final values, α' and x' at the image. The product of the three 2×2 matrices on the right-hand side can be consolidated into a single matrix called the object-image matrix $S_{P'P}$. Then Eq. (20) can be written as

$$\begin{pmatrix} n'_2 \alpha'_2 \\ x' \end{pmatrix} = S_{P'P} \begin{pmatrix} n_1 \alpha_1 \\ x \end{pmatrix} \quad (21)$$

and this matrix has the complicated form:

$$S_{P'P} = \begin{pmatrix} b + \frac{at_1}{n_1} & -a \\ \frac{bt'_2}{n'_2} + \frac{at_1 t'_2}{n_1 n'_2} - d - \frac{ct_1}{n_1} & c - \frac{at'_2}{n'_2} \end{pmatrix} \quad (22)$$

If we put this matrix into the previous equation, we obtain an unsatisfactory result: the value of x' will depend on the angle α_1 made by the incident ray, as can be seen by multiplying the two matrices on the right side. The ratio of x' to x is called the magnification m ; that is

$$m = \frac{x'}{x} \quad (23)$$

A perfect image can be formed only if the magnification is determined solely by the object distance and the constants of the lens. To eliminate this difficulty, the lower left-hand element in the matrix, Eq. (22) is required to be equal to zero, and the magnification is then:

$$m = \frac{x'}{x} = c - \frac{at'_2}{n'_2} \quad (24)$$

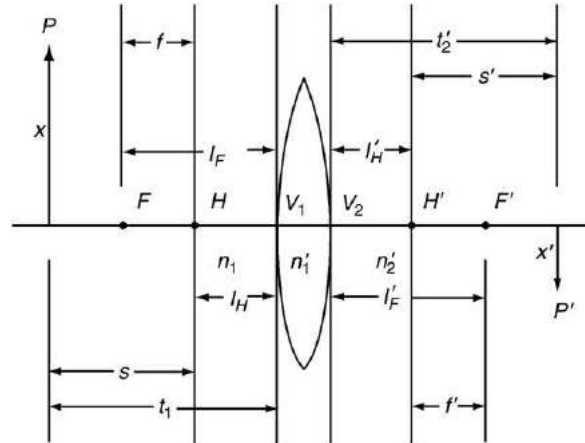


Fig. 5 Definitions of cardinal plane locations.

The determinant of this matrix is unity, since it is the product of three matrices whose individual determinants are unity. It follows that

$$\frac{1}{m} = b + \frac{at_1}{n_1} \quad (25)$$

so that

$$\begin{pmatrix} n'_2 \alpha'_2 \\ x' \end{pmatrix} = \begin{pmatrix} 1/m & -a \\ 0 & m \end{pmatrix} \begin{pmatrix} n_1 \alpha \\ x \end{pmatrix} \quad (26)$$

Using the fact that the lower left-hand element of the matrix Eq. (22) is zero shows that

$$\frac{t'_2}{n'_2} = \frac{d + ct_1/n_1}{b + at_1/n_1}, \quad \frac{t_1}{n_1} = \frac{d - bt'_2/n'_2}{-c + at'_2/n'_2} \quad (27)$$

This equation connects the object distance t_1 with the image distance t'_2 , both quantities being measured from the appropriate vertex. This relation is known as the generalized Gauss lens equation. It applies to all lens systems, no matter how complicated.

We can determine the location of the unit planes by letting $m=1$. This t'_2 in Eq. (27) is the location l'_H of the image space unit plane, and it becomes

$$l'_H = \frac{n'_2(c-1)}{a} \quad (28)$$

This relation expresses the location of the unit plane on the image side as the distance from V_2 to H' . Similarly, the location of H with respect of V_1 is

$$l_H = \frac{n_1(1-b)}{a} \quad (29)$$

To locate the focal planes, consider a set of parallel rays coming from infinity and producing a point image at F' . The terms containing t_1 in Eq. (27) are much larger than d or b , and this relation becomes

$$l'_F = \frac{n'_2 c}{a} \quad (30)$$

Letting t'_2 become infinite in the right-hand half of Eq. (27), the location of F is given by the equation

$$l_F = \frac{-n_1 b}{a} \quad (31)$$

The focal lengths f and f' shown in Fig. 5 were defined earlier as the distances between their respective focal and unit planes. Hence

$$f' = l'_F - l'_H = \frac{n'_2}{a} \quad (32)$$

and

$$f = l_F - l_H = \frac{-n_1}{a} \quad (33)$$

If we let $n_1 = n'_2 = 1$ for air, these become

$$-f = f' = \frac{1}{a} \quad (34)$$

The Gaussian constant a is thus the reciprocal of the focal length in air. Even when the lens is not in air, it is still true that $f' = -f$ if the lens is in a single medium. It is also true regardless of whether or not the lens is symmetric.

It is customary in optics to work with a consistent set of units, such as centimeters or millimeters, and the sign conventions used for x and y are the standard Cartesian rules. As mentioned above, curved surfaces which open to the right have a positive radius and vice versa. As an example, let a double convex lens have radii of 2.0 and 1.0, respectively, an index of refraction of 1.5, and a thickness of 0.5. These quantities are then denoted as

$$\begin{aligned} r_1 &= +2.0, & r_2 &= -1.0, & t'_1 &= t_2 = +0.5 \\ n_1 &= 1.0, & n'_1 &= 1.5 = n_2, & n'_2 &= 1.0 \end{aligned} \quad (35)$$

By Eq. (10):

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{1.5 - 1.0}{2.0} = 0.25 \quad (36)$$

and

$$k_2 = \frac{1.0 - 1.5}{-1.0} = 0.50 \quad (37)$$

The system matrix S_{21} is

$$\begin{aligned}
 S_{21} &= R_2 T_{21} R_1 \\
 &= \begin{pmatrix} 1 & -0.50 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5/1.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 \\ 0 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 0.83 & -0.71 \\ 0.33 & 0.92 \end{pmatrix}
 \end{aligned} \tag{38}$$

and we note that the determinant is

$$(0.83)(0.92) - (0.71)(-0.33) = 1.00 \tag{39}$$

This property of the system matrix provides a very useful check on the accuracy of matrix multiplications. The locations of the four cardinal points can be determined from the formulas given above and their locations are shown in the scale drawing of Fig. 6. Since we know how to locate the cardinal points of a lens, we are in a position to ray trace in a precise manner, using the method of Fig. 3 and adding a third ray. After locating F , H , H' , and F' , we no longer need the lens; the cardinal planes with the proper separation are fully equivalent to the lens they replace. This is very much like what electrical engineers do when they replace a complicated circuit with a black box; this equivalent circuit behaves just like the original. Ray tracing becomes more complicated when we deal with diverging lenses (Fig. 7), but the procedure gives above still applies. Parallel rays spread apart and do not come to a focus on the right-hand side of the lens. However, if the refracted rays are extended backwards, these extensions will meet as indicated. This behavior can be verified quantitatively by using the formulas developed above. It will be found that F and F' have exchanged places, but H and H' retain their original positions inside the lens. If an object is placed between the focal point F and the first vertex of a convex lens (Fig. 8), then our usual procedure produces two rays which do not intersect in image space. The ray from the object which is parallel to the axis is refracted downward so that it does not intersect the other ray in image space. However, the extensions to the left of these two rays meet to form an image which is erect, magnified, and virtual. This is the normal action of a magnifying lens; the image can be seen but cannot be projected onto a screen, unlike the real, inverted image of Fig. 3. This ray tracing diagram confirms what would be seen when a double convex lens is used to magnify an object lying close to the lens. For the double concave lens of Fig. 9, a parallel ray leaves an object point P and is refracted upward at the unit plane H' in such a way that its extension, rather than the ray itself, passes through F' . The ray headed for F is refracted at H before it can reach the object-space focal plane and becomes parallel to the axis. The third ray, going from P to H , emerges at H' parallel to its original direction and its extension to the left passes through the image point P' already determined by the intersection of the other two rays. The resulting virtual image is upright and reduced. All three rays in this diagram behave exactly like the corresponding rays in Fig. 3; the only change is the use of the extensions to locate the image. The eye receives the diverging rays from P and believes that they are coming from P' .

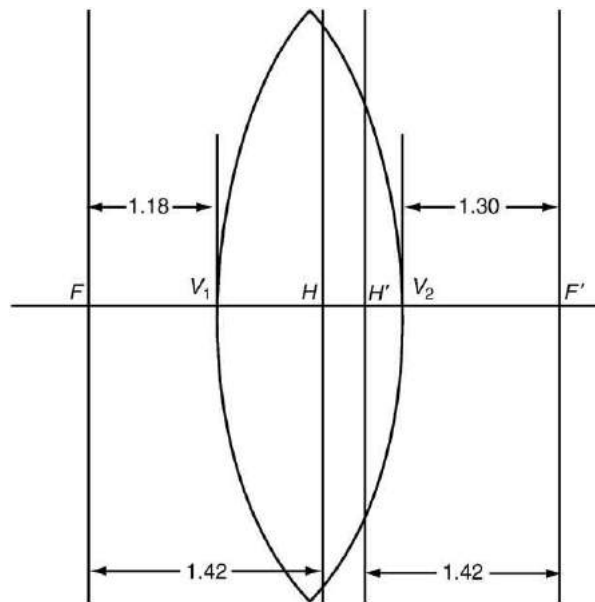


Fig. 6 Positions of cardinal planes for an asymmetric lens.

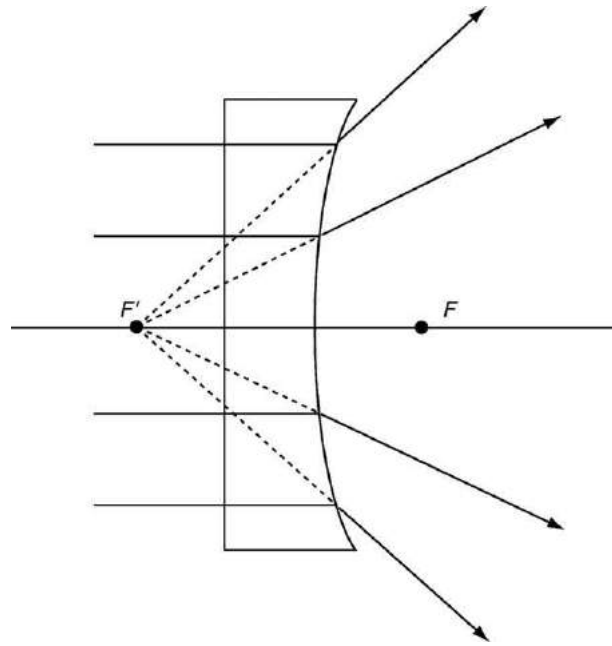


Fig. 7 Ray tracing for a planar-concave lens.

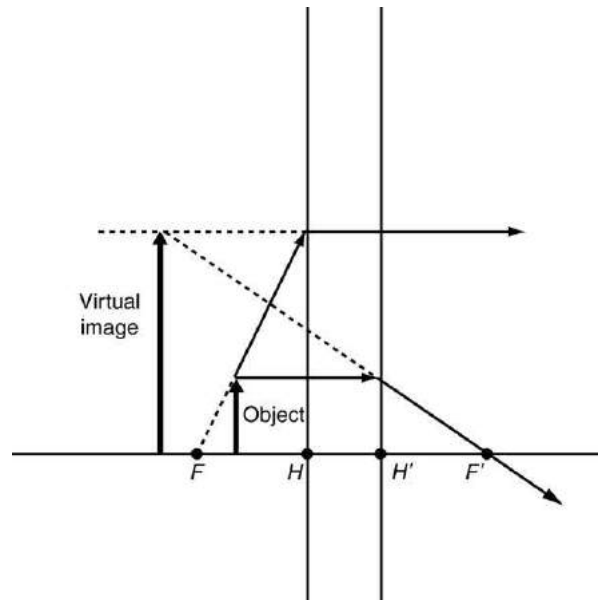


Fig. 8 Formation of a virtual image by a convex lens.

Nodal Points and Planes

Let a ray leave a point on the z -axis at an angle α_1 , and have an angle α'_2 when it reaches image space. Since $x=0$ at the starting point, Eq. (26) shows that

$$n'_2 \alpha'_2 = n_1 \alpha_1 / m \quad (40)$$

The ratio of the final angle to the initial angle in this equation is the angular magnification μ , which is then

$$\mu = n_1 / m n'_2 \quad (41)$$

The locations of object and image for unit angular magnification are called the nodal points, labeled as N and N' . The above

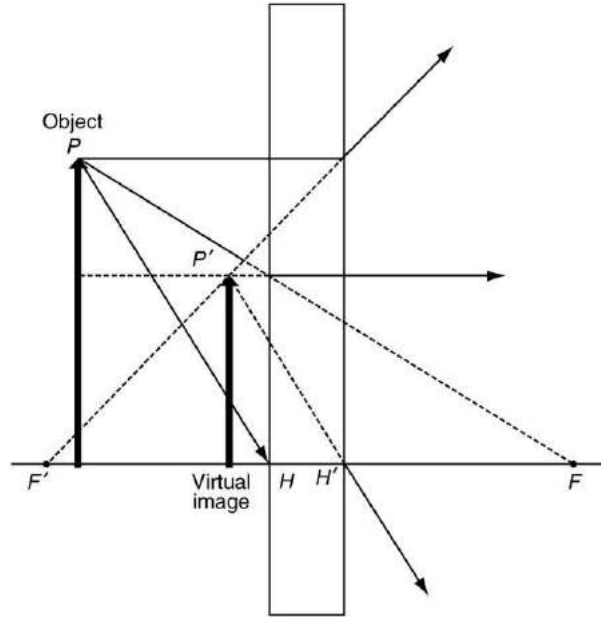


Fig. 9 Formation of a virtual image by a concave lens.

definition shows that the linear and angular magnification are reciprocals of one another when object and image are in air or the same medium. The procedure that located the unit points will show that the nodal points have positions given by

$$\frac{l_N}{n_1} = \frac{(n'_2/n_1) - b}{a} \quad (42)$$

and

$$\frac{l'_N}{n'_2} = \frac{c - (n_1/n'_2)}{a} \quad (43)$$

When the two indices are identical, these relations are identical to those for the points H and H' . That is why the rays from P to H and H' to P' in [Fig. 3](#) are parallel.

Compound Lenses

A great advantage of the paraxial matrix method is the ease of dealing with systems having a large number of lenses. To start simply, consider the cemented doublet of [Fig. 10](#): a pair of lenses for which surface 2 of the first element and surface 1 of the second element match perfectly. The parameters of this doublet are given in the manner shown below. It is understood that the first entry under r is r_1 , the second entry is r_2 , and so on. Note that this doublet has only three surfaces; if the two parts were separated by an air space, producing an air-spaced doublet, then there would be four surfaces to specify. The values of the indices n' and the spacings t' are placed between the values of r . The constants of the doublet are then

r	n'	t'	
1.0			
	1.500	0.5	
-2.0			
	1.632	0.4	
∞			

(44)

Note that surface 3, which is flat, has a radius of infinity. Numbering the vertices in the usual manner, the system matrix is

$$S_{31} = R_3 T_{32} R_2 T_{21} R_1 \quad (45)$$

where

$$k_2 = (n'_2 - n_2)/r_2 = (1.632 - 1.500)/-2.0 \quad (46)$$

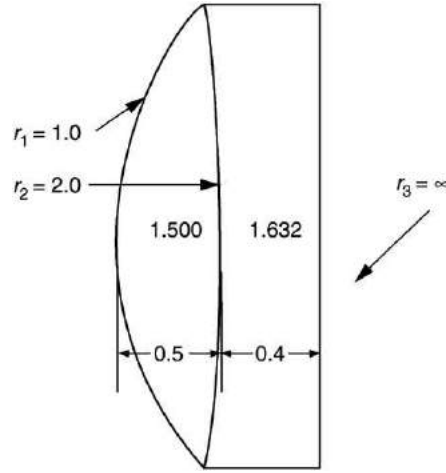


Fig. 10 Specification for a cemented doublet.

Approximations for Thin Lenses

An advantage of paraxial matrix optics is the ease of obtaining useful relations. For example, combining the several expressions for the Gaussian constant a , we obtain:

$$a = -\frac{n_1}{f} = \frac{n_2'}{f'} = k_1 + k_2 - \frac{k_1 k_2 t_1}{n_1'} \quad (47)$$

and using the definitions of k_1 and k_2 with the lens in air leads to

$$\frac{1}{f'} = (n_1' - 1) \left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n_1' - 1)t_1}{n_1' r_1 r_2} \right] \quad (48)$$

which is the lensmakers' equation. It tells how to find the focal length of a lens from a knowledge of its material and geometry. This derivation is much more direct than what you will usually find. This equation becomes simpler when we are dealing with thin lenses – those for which the third term in brackets can be neglected by assuming that the lens thickness is approximately zero. Then

$$\frac{1}{f'} = (n_1' - 1) \left[\frac{1}{r_1} - \frac{1}{r_2} \right] \quad (49)$$

which is the form often seen in texts. We then realize that the original form is valid for lenses of any thickness, but the thin lens form can be used if the spacing is much less than the two radii. It is also seen that when the thickness is approximately zero, then

$$b = c = 1 \quad (50)$$

and the unit planes have locations

$$l_H = l_H' = 0 \quad (51)$$

Thus, they are now coincident at the center of the lens. A ray from any point on the object would pass undeviated through the geometrical center of the lens, as is often shown in the elementary books. These new values of the Gaussian constants give a system matrix of the form:

$$S_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix} \quad (52)$$

This matrix contains a single constant, the focal length of the lens, so that the preliminary design of an optical system is merely a matter of specifying the focal length of the lenses involved and using the resulting matrices to determine the object-image relationship. We also note that this matrix looks like a refraction matrix; the non-zero element is in the upper right-hand corner. This implies that the refraction–translation–refraction procedure that actually occurs can be replaced, for a thin lens, by a single refraction occurring at the coinciding unit planes. Ophthalmologists take advantage of the thin lens approximation by specifying focal lengths in units called diopters. The reciprocal of the focal length in meters determines its refraction power in diopters. For example, a lens with $f' = 10$ cm will be a 10 diopter lens. If two lenses are placed in contact, the combined power is the sum of the individual powers, for multiplying matrices of the above form gives the upper right-hand element as $(-1/f_1' - 1/f_2')$.

The Paraxial System for Design or Analysis

The material given in this article can serve as the basis for an organized way of taking the specifications of an optical system and using them to gain a full understanding of this system. We shall demonstrate it with a practical example which has some interesting features. Chemical engineers have long known that the index of refraction of a liquid can be determined by observing the empty bore of a thick glass tube with a telescope having a calibrated eyepiece and then measuring the magnification of the bore when the liquid flows through it. Consider a liquid with an index of 1.333 and a tube with bore radius of 3 cm, outer radius of 4 cm, and glass of index equal to 1.500. The first step is to specify the parameters of the optical system. These are the radii $r_1 = -3$, $r_2 = -4$ and the thickness $t'_1 = 1$, regarding the tube as a concentric lens with an object to its left; the indices $n_1 = 4/3$ (fractions are convenient), $n'_1 = 3/2$, $n'_2 = 1$. Calculating the elements of the two refraction matrices and the translation matrix, the system matrix is

$$S_{21} = \begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2/3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1/18 \\ 0 & 1 \end{pmatrix} \quad (53)$$

Multiplying, the Gaussian constants are

$$a = 2/27, \quad b = 11/12, \quad c = 28/27, \quad d = -2/3 \quad (54)$$

An important check is to verify that $bc - ad = 1$, as is the case here. The expressions given previously for the positions of the cardinal points then lead to the values

$$\begin{aligned} l_H &= 3/2, & l'_H &= 1/2, & l_F &= -16\frac{1}{2}, \\ l'_F &= 14, & l_N &= -3, & l'_N &= -4 \end{aligned} \quad (55)$$

These points are shown in Fig. 11. Although the same scale has been used for both horizontal and vertical distances, this is usually not necessary; the vertical scale, which merely serves to verify the magnification, can be arbitrary. The entire liquid column is the object in question, but it is simpler to use a vertical radius as the object. Then the image, generated as described in all the previous examples, coincides with the object and it is enlarged, erect, and virtual. We confirm the image location shown by using the generalized Gaussian lens equation, giving

$$\begin{aligned} t' &= [(-2/3) + (28/27)(-3)/(4/3)]/[11/12] \\ &\quad + (2/27)(-3)/(4/3) \\ &= -4 \end{aligned} \quad (56)$$

and this agrees with the sketch. The denominator of this expression was shown to be the reciprocal of the magnification $1/m$ with a value of $3/4$, and m itself can be calculated from t' as $c - at' = 4/3$; this is another important check on the accuracy of the calculation; the sketch obeys this conclusion, and the purpose of this analysis is confirmed.

Reflectors

To show how reflecting surfaces are handled with matrices, consider a plane mirror located at the origin and normal to the z -axis (Fig. 12). The ray which strikes it at an angle α_1 leaves at an equal angle, resulting in specular reflection. Since $k=0$ for a flat surface, the refraction matrix reduces to the unit matrix and by Eq. (9):

$$n'_1 \alpha'_1 = n_1 \alpha_1 \quad (57)$$

This contradicts Fig. 12, since α_1 and α'_1 are opposite in sign.

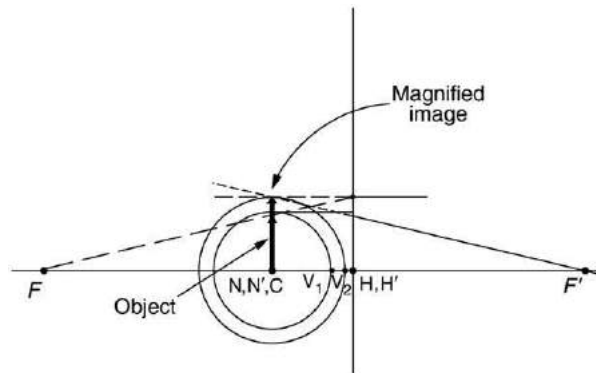


Fig. 11 Ray tracing diagram for a thick-walled glass tube.

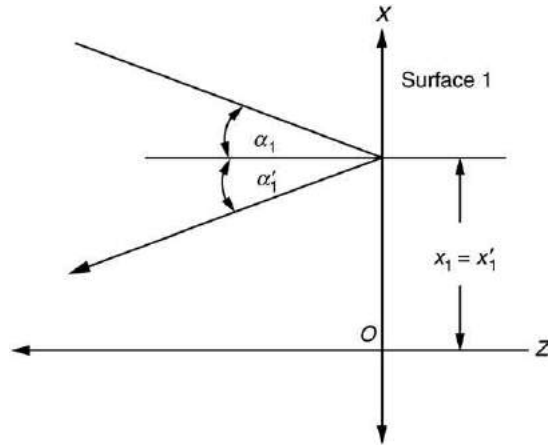


Fig. 12 Specular reflection.

If, however, we specify that

$$n'_1 = -n_1 \quad (58)$$

then the difficulty is removed. Hence, reflections involve a reversal of the sign on the index of refraction, and two successive reflections restore the original sign. For a mirror, the refraction power is

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{-1 - (1)}{r_1} = -\frac{2}{r_1} \quad (59)$$

The system matrix for a single refracting surface thus becomes

$$S_{11} = R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2/r_1 \\ 0 & 1 \end{pmatrix} \quad (60)$$

with

$$a = -2/r_1, \quad b = c = 1, \quad d = 0 \quad (61)$$

The connection between object and image can be expressed as

$$\begin{pmatrix} n'_1 \alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'/n'_1 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -t/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1 \alpha_1 \\ x_1 \end{pmatrix} \quad (62)$$

and using the same procedure that was applied to Eqs. (28)–(31), the locations of the six cardinal points are

$$\begin{aligned} l_F &= -n_1 b/a, & l'_F &= n'_1 c/a \\ l_H &= n_1(1-b)/a, & l'_H &= n'_1(c-1)/a \\ \frac{l_N}{n_1} &= \frac{(n'_1/n_1) - b}{a}, & \frac{l'_N}{n'_1} &= \frac{c - (n_1/n'_1)}{a} \end{aligned} \quad (63)$$

so that for the spherical mirror:

$$l_F = \frac{(-1)(1)}{-2/r_1} = \frac{r_1}{2} = l'_F \quad (64)$$

$$l_H = l'_H = 0 \quad (65)$$

and

$$l_N = \frac{-1-1}{-2/r_1} = r_1 \quad (66)$$

but

$$\frac{l'_N}{-1} = \frac{1-(-1)}{-2/r_1}, \quad l'_N = r_1 \quad (67)$$

Since r_1 is negative, these equations show that the unit points lie on the vertex, the foci coincide halfway between the center of curvature and the vertex, and the nodal points are at this center. The perfect focusing is a consequence of the paraxial approximation. Ray tracing for this mirror uses the procedure developed for lenses, which applies to optical systems of any complexity. Fig. 13 shows an object to the left of the center of curvature. The ray from P parallel to the axis goes to H' and then through F' , while the ray through F goes to H and then becomes parallel to the axis. Their intersection at P' produces a real, inverted image. The

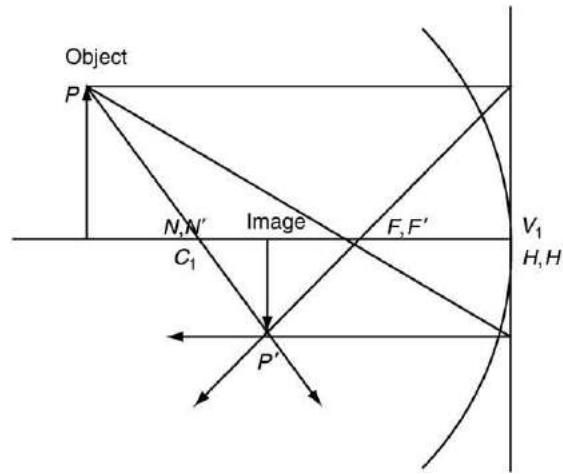


Fig. 13 Ray tracing for a concave mirror.

third set of rays represents something different: it requires a knowledge of the nodal point locations. The ray from P to N should be parallel to the ray from N' to P' , by definition; in this example, they meet this requirement by being colinear.

As an example of the power of matrix optics to simplify the design or analysis of an optical system involving lenses and mirrors, consider an object that is 15 units to the left of a converging lens, with focal length $f' = 10$. A concave mirror, of radius $r_1 = 16$, is 20 units to the right of the lens. The refraction power of the mirror is $k_1 = (-1 - 1)/-16 = 1/8$ and using the thin lens form of the system matrix, the combined matrix is

$$\begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1/10 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -3/2 & 1/40 \\ 20 & -1 \end{pmatrix} \quad (68)$$

This gives the constants a , b , c , and d , from which we find that the image is $-4 \frac{4}{9}$ units to the left of the mirror, its magnification is $-8/9$, and it is inverted. Compare the simplicity of this procedure with the usual way of doing this calculation, which involves first finding the image in the lens and then using that image as a virtual object.

Conclusion

It has been shown that the use of paraxial matrices provides a simple approach to an understanding of the kind of optical systems that provide perfect images. Further details and derivations of expressions given here will be found in the Further Reading below.

Further Reading

Brouwer, W., 1964. *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.
Nussbaum, A., 1998. *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Aberrations

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The images may be magnified or reduced and they may be erect or inverted, but they are otherwise faithful replicas of the object. This behavior is a consequence of ray tracing based on an approximate or paraxial form of Snell's law which we write as

$$n\theta = n'\theta' \quad (1)$$

where n and n' are the indices of refraction on either side of an interface and θ and θ' are the angles of incidence and refraction. If, however, the following exact form of Snell's law:

$$n\sin\theta = n'\sin\theta' \quad (2)$$

governs the ray behavior, as is the case when the rays have large inclinations with respect to the symmetry axis, then it is necessary to devise a more involved ray tracing procedure. Fig. 1, which shows the first surface of a lens, will be used to explain how this procedure works. A ray leaves the object point P and strikes the lens surface at the point P_1 . Rays which lie completely in the plane ZOX are said to be meridional; this is the plane which passes through the symmetry axis, just as meridians on the Earth's surface are determined by planes through the geographic axis. By regarding the various distances in the figure as vectors, it is possible to start with the initial coordinates of the ray and with the direction-cosines which specify its slope and find the coordinates of the point P' where the ray strikes the lens surface. A derivation of the equation governing this process will be found in the text of A Nussbaum (see Further Reading). This derivation involves only elementary algebra and vector analysis, and will be found simpler than the usual textbooks treatment of ray tracing procedures. The equation is

$$T_1 = \frac{F}{-E + \sqrt{E^2 - c_1 F}} \quad (3)$$

where

$$E = Bc_1/2 = c_1[(z - v_1)N + xL] - N \quad (4)$$

and

$$F = Cc_1 = c_1[(z - v_1)^2 + x^2] - 2(z - v_1) \quad (5)$$

Knowing the coordinates z and x of the starting point P of the ray, as well as its starting cosines N and L , plus the curvature c_1 or radius $r_1 = 1/c_1$ of the lens surface and the location v_1 of the surface's vertex, we can readily calculate the length T_1 of the ray. Then its components on the two axes give the coordinates of P' . Next, we need to find the slope of the ray after it is refracted at the lens surface, and this can be calculated from Snell's law combined with the behavior of the incident and refracted rays regarded as vectors. Again we invoke the reference cited above for a derivation of the equations giving the value of the direction-cosines after refraction; these expressions are

$$n'_1 N'_1 = K_1/c_1 + n_1 N_1 - K_1 z_1 \quad (6)$$

and

$$n'_1 L'_1 = n_1 L_1 - K_1 x_1 \quad (7)$$

where K_1 is known as the refracting power and has the definition

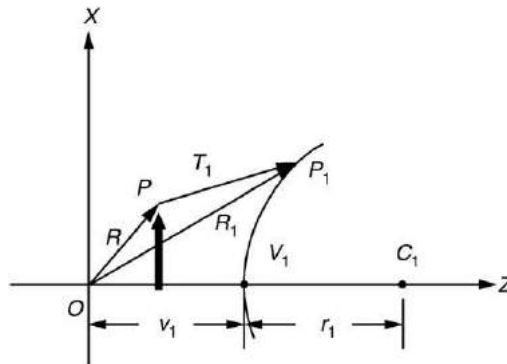


Fig. 1 Ray tracing in the meridional plane.

$$K_1 = c_1 (n_1' \cos \theta' - n_1 \cos \theta) \quad (8)$$

The ray can now be traced to the next surface knowing the lens thickness, and then to the image plane. These two sets of equations, when incorporated into a computer program of about two dozen lines, will trace meridional rays to any desired accuracy through an optical system containing an arbitrary number of lenses and spherical reflectors. As an example, consider a symmetric double convex lens with radii of 50 units at surface 1, -50 units at surface 2, a thickness of 15 units, and an index of 1.50. Rays parallel to the axis – one at 0.2 units above the axis, a second one at 2 units above the axis, and a third one at 20 units above the axis – are started well to the left and traced to the paraxial focal plane, which is at 47.368 units from the second surface. The ray that is 0.2 units above the axis will cross the focal plane at the axis; this is a paraxial ray, passing through the paraxial focal point F' . The ray which is 2 units above the axis crosses the focal plane slightly below the focal point, and is therefore not quite paraxial, while the ray which starts at 20 units above the axis falls well short of the focal point, intersecting the axis before it reaches the focal plane. This behavior indicates that the lens possesses a defect known as spherical aberration and these calculations imply that spherical aberration should be defined as the failure of nonparaxial rays to conform to the behavior of paraxial rays. This definition does not appear in optics texts, and the few definitions that are given appear to be incorrect. It is meaningless to speak of a focal point as defined by a pair of nonparaxial rays; this infinite set of ray-pairs determines a corresponding number of intersection points, all of which fall short of the true focal point F' . This situation is nicely illustrated by the next figure, which shows a large number of parallel rays, the central ones being paraxial and the others behaving differently. The rays which are very close to the axis will meet as expected at the focal point F' . The rays a little farther out – the intermediate rays – will cross the axis just a little to the left of F' , and those near the edge of the lens – the *marginal* rays – will fall very short, as indicated. Rotating this diagram about the axis, we realize that spherical aberration produces a very blurry circular image at the paraxial plane. The three-dimensional envelope of the group of rays produced by this rotation is known as the caustic surface and the narrowest cross-section of this surface, slightly to the left of the paraxial focal plane, is called the circle of least confusion. If this location is chosen as the image plane, then the amount of spherical aberration can be reduced somewhat. Other methods of improving the image will be considered below. **Fig. 2** was obtained by adding a graphics step to the computer program described above. **Fig. 3** shows how to define spherical aberration

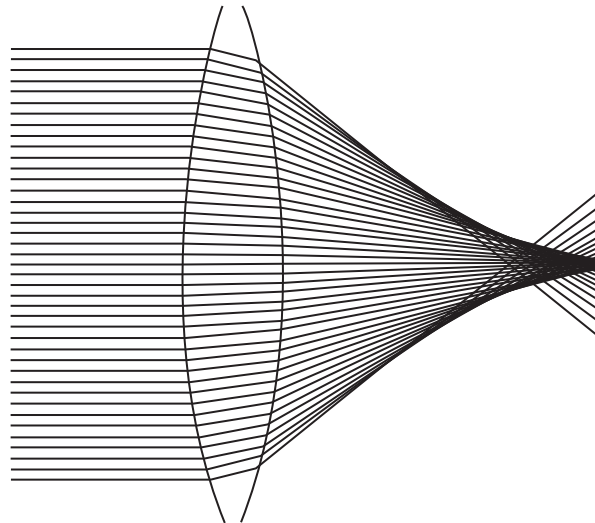


Fig. 2 Spherical aberration for parallel meridional rays.

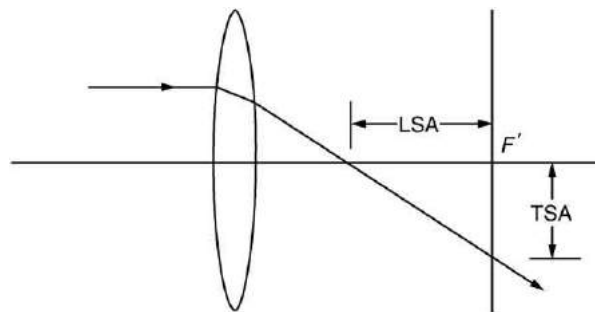


Fig. 3 Definitions of transverse and longitudinal spherical aberration.

quantitatively. Two kinds of aberration are specified in this figure. The place where the ray crosses the axis, lying to the left of F' , is at a distance, measured from the paraxial focus, called the longitudinal spherical aberration (LSA), while the distance below the axis at the focal plane is the transverse spherical aberration (TSA). To reduce this aberration, note in Fig. 2 that the amount of refraction at the two surfaces of the lens for marginal rays is not the same. Equalizing the refraction at the two surfaces will improve the situation. Suppose we alter the lens while keeping the index, thickness, and focal length constant; only the radii will change. This can be done if the new radii satisfy the paraxial equation giving the focal length in terms of the lens constants. The process of varying the shape of lens, while holding the other parameters constant, is called *bending the lens*. This can be easily accomplished by a computer program, obtaining what are known as optimal shape lenses, which are commercially available. To study their effect on spherical aberration, define the shape factor σ of a lens as

$$\sigma = \frac{r_2 + r_1}{r_2 - r_1} \quad (9)$$

from which $\sigma=0$ for a symmetric lens ($r_1 = -r_2$). Fig. 4(a) shows the monitor screen for a symmetric lens, and Fig. 4(b) illustrates the result of lens bending. This lens has had the curvature of the first surface raised and the second surface has been flattened to keep the focal length constant. Let us now consider what lens designers can do about spherical aberration. Brouwer in his book (see Further Reading) gives the specifications for a cemented doublet consisting of a double convex lens followed by a lens with two concave surfaces. The longitudinal spherical aberration of the front lens alone varies as shown in Fig. 5. The way that the designer arranged for the marginal rays to meet at the paraxial focus, rather than falling short, was to recognize that a diverging lens following the front element would refract the rays away from – rather than closer to – the axis. When this second component is added, we would expect the behavior shown in Fig. 6. Note that the value of LSA for the doublet has been enormously reduced, as indicated by the horizontal scale change. Spherical aberration can be reduced in more elaborate systems as part of the design process. Telescopes, in particular, make use of corrector plates which are placed in front of the large mirrors.

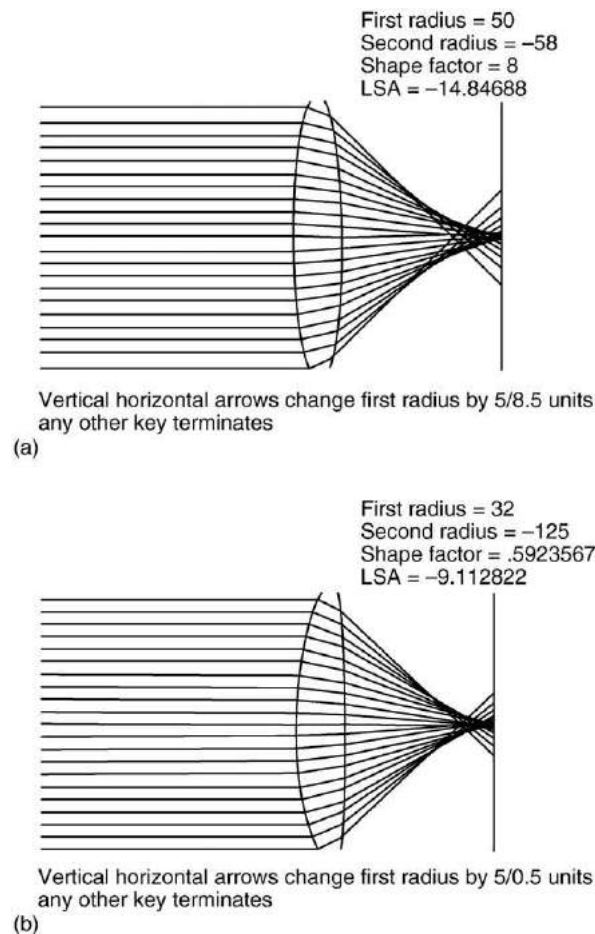


Fig. 4 Lens bending as performed numerically.

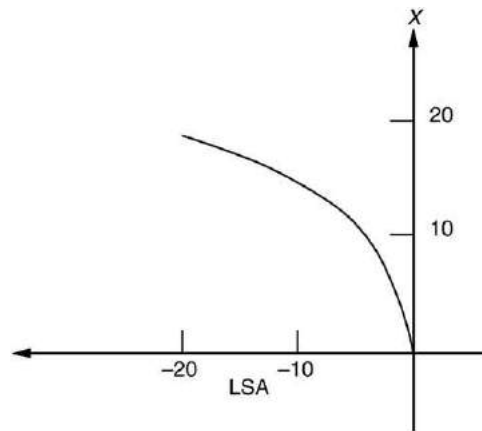


Fig. 5 Spherical aberration for a single lens.

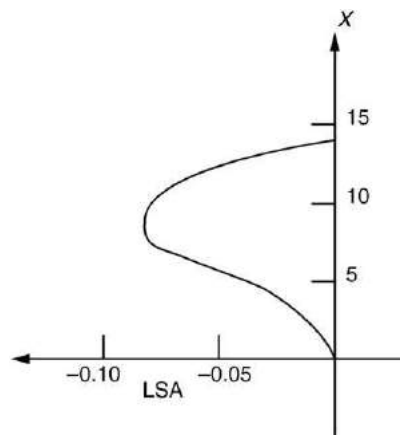


Fig. 6 Reduction of spherical aberration by a doublet.

Coma

We have seen how to trace nonparaxial, meridional rays and found that spherical aberration appears because the intermediate and marginal rays do not behave like the central rays. We now wish to deal with rays that do not lie in a meridional plane; these are nonmeridional or skew rays. Calculating the behavior of such a ray would appear to be a complicated problem, but it turns out to be a simple extension of what has already been accomplished. Let a skew ray start at an arbitrary point in space with coordinates x , y , and z and let it have direction cosines L , M , and N with respect to OX , OY , and OZ . The points P and P_1 in Fig. 1 are now out of the plane ZOX and when the vectors in this figure are expressed as the sum of three, rather than two, components, the equations given above for meridional rays acquire terms in y and M which have the same form as those for x and L . Now trace a set of skew rays chosen to lie on a cylinder (Fig. 7) which is centered about the z -axis of a lens, so that all the rays are meridional. These rays are fairly far from the axis, so that they are also nonparaxial, but they all meet at a common image point, forming a cone whose apex is this image point. Then give this cylinder a downward displacement while holding fixed the intersection point of each ray with the dotted circle on the front of the lens. The tilting of the cylinder changes all its rays – except the top and the bottom rays – from meridional to skew. In addition, the cone on the image side will then tilt upwards; it should not change in any other way if the skew rays continue to meet at a well-defined apex. But this is not what happens in a simple lens, as will now be explained.

Let the top and bottom meridional rays meet at a point P in image space (Fig. 8) and use this point to determine the location of an image plane. The skew ray just below the uppermost ray in Fig. 7 will pass through this plane fairly close to the intersection point. Let us guess (and confirm later) that it will pass through a point slightly to the right of P and a little below it, as we observe this plane from the lens position. This is point 1 in Fig. 8. The next skew ray on the front of the lens will strike the image plane still farther below and to the right, producing point 2 in the figure. As we continue to trace these rays down the front half of the circle, we realize that a ray almost at the bottom of Fig. 7 will have to be very close to the lower meridional ray and that its image point is just below but to the left of P ; this is the point labeled ‘next-to-last’ in Fig. 8. All of these points join to form a closed curve. The

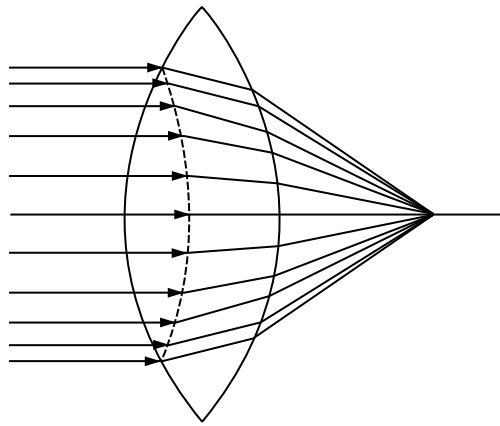


Fig. 7 Configuration used to illustrate coma.

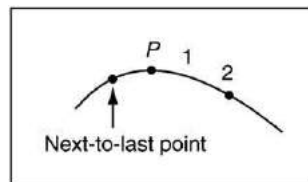


Fig. 8 Starting a coma plot.

back halves of the dotted circle and of the ray cylinder are not shown in the figure, but since we are again starting at P in **Fig. 8**, the pattern will repeat itself. That is, we get a second closed curve which – by symmetry – should be identical to the first curve. This is a most unusual effect; double-valued phenomena are quite rare. To confirm this prediction, we use the skew ray equations, and trace the pattern formed by a set of concentric cylinders of different sizes. The appearance of the coma pattern generated by the computer is shown in **Fig. 9**. Each half-circle on the front side of the lens produces a closed figure (an approximation to a circle) as an image. If coma were the only aberration (when it is called ideal coma), the plots would be the perfect, coinciding circles of **Fig. 10**. The comet-like shape explains where the name comes from. We can picture the generation of these circles in the way shown in this figure. The ray through the geometrical center of the lens produces a point in the paraxial focal plane. The top and bottom rays (the meridional rays) from the circles of increasing size produce points which define image planes that are successively closer to the lens; this is a consequence of spherical aberration. Moving the top ray parallel to itself and along the half-circle generates the image plane pattern. The lines tangential to the coma circles make an angle of 60° with each other, as can be seen by measurement on the computer output. Most optics books look at coma in a way different from that given here. Coma has been defined by WJ Smith (see Further Reading) as the variation of magnification with aperture, analogous to the definition of spherical aberration as the variation of focal length with aperture. And as was previously mentioned, focal length and magnification are purely paraxial concepts, not applicable to the description of aberrations. The definition that comes from the arguments given above is that coma is the failure of skew rays to match the behavior of meridional rays, just as spherical aberration is the failure of meridional rays to match the behavior of paraxial rays. Most texts show coma in its ideal form.

Astigmatism

We have shown that spherical aberration is the failure of meridional rays to obey the paraxial approximation and coma is the failure of skew rays to match the behavior of meridional rays. To see the connection between these two aberrations, consider the object point P of **Fig. 11**. This very complicated diagram can actually be very helpful in understanding the third of the five geometrical aberrations. Let P be the source of a meridional fan: this is the group of rays with the top ray labeled PA and the bottom ray is PB . If the lens is completely corrected for spherical aberration, this fan will have a sharp image point P'_T lying directly below the z -axis. Now let P also be the source of a fan bounded by the rays PC and PD . This fan is at right-angles to the other one; we can think of these rays as having the maximum skewness possible. If the lens has no coma, this fan also will produce a sharp image point P'_s , which, for a converging lens, will be farther from the lens than the tangential focal point. In other words, even though the lens has been fully corrected for both aberrations, the two corrections will not necessarily produce a common image. In the figure, the meridional fan is called a tangential fan and the skew fan is called a sagittal fan; these are the terms commonly used by lens designers. The failure of the sagittal and tangential rays to produce a single image in a lens corrected for both spherical

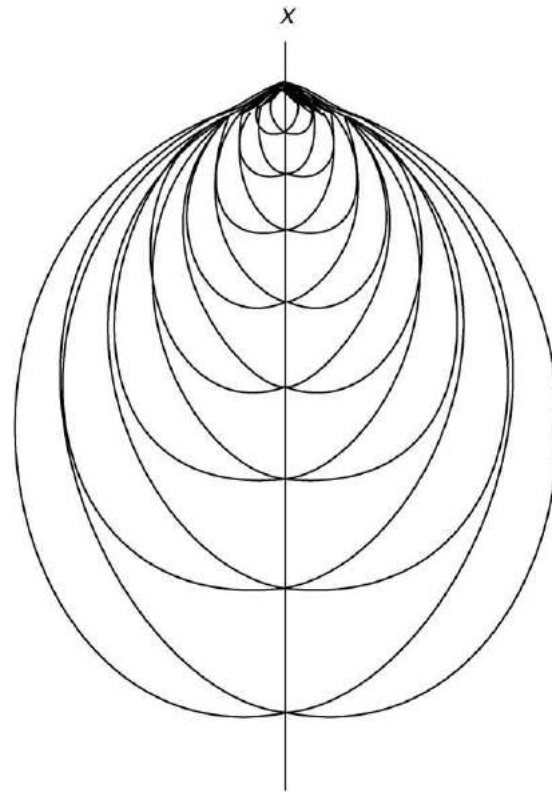


Fig. 9 Coma as calculated numerically.

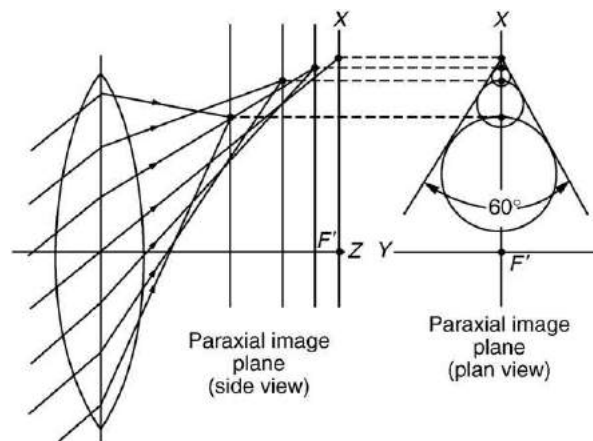


Fig. 10 Customary way of illustrating coma.

aberration and coma is known as astigmatism. Astigmatism, coma, and spherical aberration are the point aberrations that an optical system can have. We have already noted that spherical aberration is the only one that can be associated with a point on the z -axis; the others require an off-axis object. Astigmatism is very common in the human eye; to see how it shows up, look at the sagittal image point P' , of [Fig. 11](#). This point is a distance z_s from the x,y -plane. A plane through this point will have a line image on it, the sagittal line, due to the tangential fan, which has already come to a focus and is now diverging. The converse effect, producing a tangential line, occurs at the other image plane. If we locate an image plane halfway between the two, then both fans contribute to the image and in the ideal case it will be a circle. As the image plane is moved forwards or backwards, these images become ellipses and eventually reduce to a sagittal or a tangential line, as shown in [Fig. 12](#). Because there are two different image planes, an object with spokes ([Fig. 13](#)) will have an image for which the vertical lines are sharp and the others get gradually poorer as they become more horizontal, or vice versa, depending on which image plane is chosen. Eye examinations detect astigmatism if the spokes appear to go from black to gray. This example of the effect of astigmatism was often based on a radial pattern of arrows, which explains the origin of the word sagittal, derived from the Latin 'sagitta' for arrow.

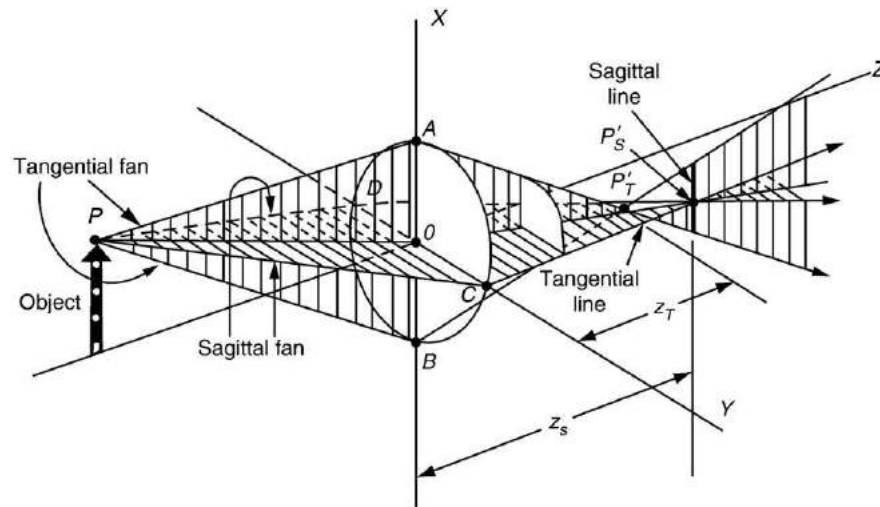


Fig. 11 Tangential and sagittal fans displaying astigmatism.

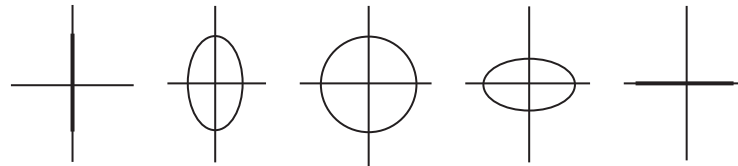


Fig. 12 Astigmatic image patterns.

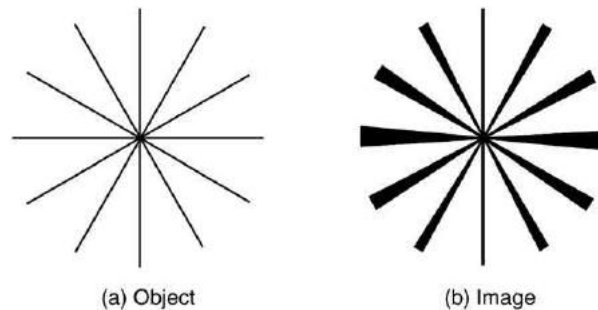


Fig. 13 Image degraded by astigmatism.

Curvature of Field and Distortion

Having covered the three point aberrations, there are two lens defects associated with extended objects. If we move the object point P in Fig. 11 closer to or farther from the z -axis, we would expect the positions of the tangential and sagittal focal planes to shift, for it is only when the paraxial approximation holds that these image points are independent of x . Hence, we obtain the two curves of Fig. 14, which shows what astigmatism does to the image of a two-dimensional object. If the astigmatism could be eliminated, the effect would be to make these curved image planes coincide, but we have no guarantee that the common image will be flat, or paraxial. The resulting defect is called Petzval curvature or curvature of field. For a single lens, the Petzval surface can be flattened by a stop in the proper place, and this is usually done in inexpensive cameras. Petzval curvature is associated with the z -axis. If we take the object in Fig. 11 and move it along the y -axis, then all rays leaving it are skew and this introduces distortion, the aberration associated with the coordinates normal to the symmetry axis. Distortion is what causes vertical lines to bulge outward (barrel distortion) or inward (pincushion distortion) at the image plane, as shown in Fig. 15. A pinhole camera will have no distortion, since there are no skew rays, and a single thin lens with a small aperture will have very little. Placing a stop near the lens to reduce astigmatism and curvature of field introduces distortion because, as shown in Fig. 16(a), the rays for object points far from the axis are limited to off-center sections of the lens. The situation in this figure corresponds to barrel distortion; placing the stop on the

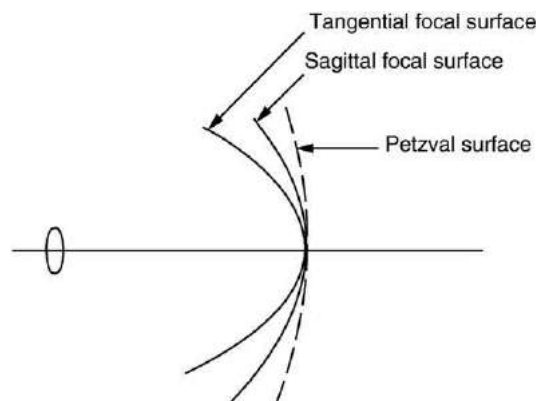


Fig. 14 Petzval of field curvature.

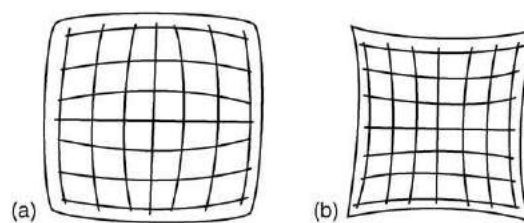


Fig. 15 Barrel and pincushion distortion.

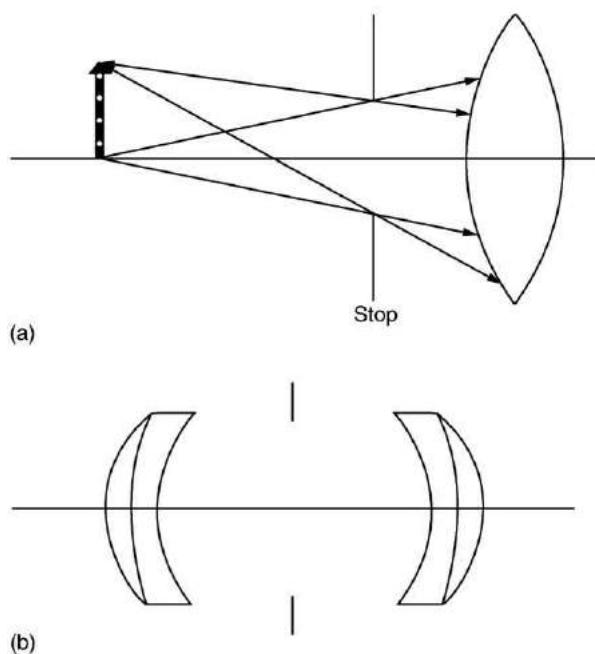


Fig. 16 (a) Distortion-creating stop. (b) Symmetric-component lens that reduces distortion.

other side of the lens produces pincushioning. Distortion can be therefore reduced by a design which consists of two identical groupings (**Fig. 16(b)**), with the iris diaphragm placed between them; the distortion of the front half cancels that of the back half. **Fig. 17** shows the output for a calculation using approximate formulas rather than exact ray tracing. Note that the designer has arranged for astigmatism to vanish at the lens margin. This example indicates the nature of astigmatism in a system for which the coma and the spherical aberration are low but not necessarily zero; it is the failure of the tangential and sagittal image planes to coincide.

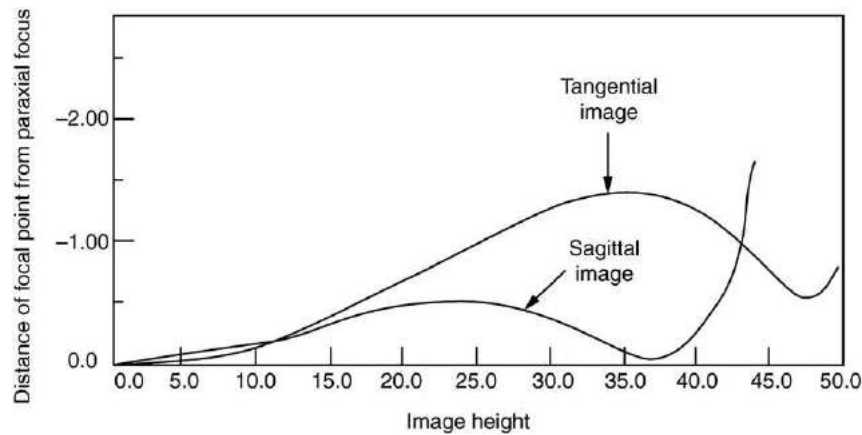


Fig. 17 Reduction of astigmatism through design.

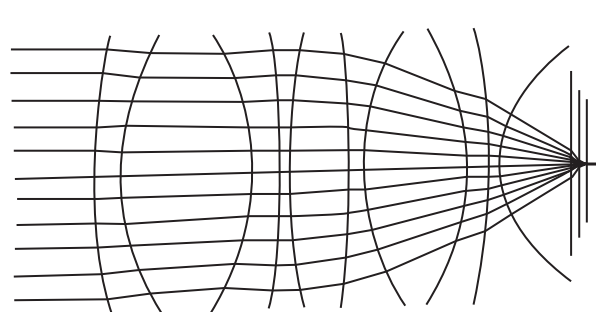


Fig. 18 High-power high-resolution microscope objective.

Nonspherical Surfaces

So far, we have considered optical systems that use only spherical or plane reflecting and refracting surfaces. However, many astronomical telescopes use parabolic and hyperbolic surfaces. In recent years, nonspherical or aspheric glass lenses have become more common because they can be mass produced by computer-controlled machinery and plastic lenses can be molded with very strict tolerances. Ray tracing procedures given above can easily be extended to the simplest kind of aspheric surfaces: the conic sections. Iterative procedures may be applied to surfaces more complicated than the conic sections. A well-known example is the curved mirror; a spherical reflecting surface will have all the aberrations mentioned here, but a parabolic mirror will bring parallel rays to a perfect focus at the paraxial focal point, thus eliminating the spherical aberration. Headlamp reflectors in automobiles are a well-known example. A plano-convex lens with this defect corrected will be one whose first surface is flat and whose second surface is a hyperbola with its eccentricity equal to the index of refraction of the glass composing the lens. One form of astronomical telescope, which has a large primary concave mirror to collect the light and reflect it back to a smaller secondary convex mirror, is the Ritchey–Chrétien design of 1924. Both mirrors are hyperbolic in shape, completely eliminating spherical aberration. This is the design of the Hubble space telescope. The two-meter diameter primary mirror, formed by starting with a spherical mirror and converted to a hyperbola by computer-controlled polishing, was originally fabricated incorrectly and caused spherical aberration. From a knowledge of the computer calculations, it was possible to design, build, and place in orbit a large correcting optical element which removed all the aberrations from the emerging image.

Conclusion

The three-point aberrations and the two extended object aberrations have been defined and illustrated. Simple ways of reducing them have been mentioned, but the best way of dealing with aberrations is in the process of lens design. In addition, there is a completely separate defect known as chromatic aberration. All the above calculations are based on ray tracing for a single wavelength (or color) of visible light. However, the index of refraction of a transparent material varies with the wavelength of the light passing through it, and a design good for one value of the wavelength will give an out-of-focus image for any other wavelength. It has been known for many years that an optical system composed of two different kinds of glass will provide a

complete correction for two colors – usually red and blue – and other colors will be partially corrected. Such lenses are said to be achromatic. For highly demanding applications, such as faithful color reproductions of works of art, it is possible to design elaborate and expensive lenses which correct for yellow light as well. This kind of optical system is said to be apochromatic. With perfect imaging at the two ends and the middle of the visible spectrum, the color error elsewhere is virtually nonexistent. As an example of a design which is free of all aberrations – geometric and chromatic – **Fig. 18** shows the ray trace for a high-power aberration-free and apochromatic microscope objective. Note that the trace shows parallel rays brought to a focus inside the cover glass; in use, the light source would be a point on the slide and the light would converge at the eyepiece.

Further Reading

- Brouwer, W., 1964. *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.
- Laikin, M., 1990. *Lens Design*. New York: Marcel Dekker.
- Nussbaum, A., 1998. *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.
- Nussbaum, A., Phillips, R.A., 1976. *Contemporary Optics for Scientists and Engineers*. Englewoods Cliffs, NJ: Prentice Hall.
- Smith, W.J., 1990. *Modern Optical Engineering*, 2nd edn. New York: McGraw-Hill Book Co.

Prisms

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Prisms are solid structures made from a transparent material (usually glass). **Fig. 1** shows the simplest type of prism – one that has many uses. The vertical sides are two identical rectangles at right angles to each other and a larger third rectangle is at 45° to the other two. The top and bottom are 45° – 45° – 90° triangles. The most common applications of prisms are:

1. To bend light in a specified direction
2. To fold an optical system into a smaller space
3. To provide proper image orientation
4. To combine or split optical beams, using partially reflecting surfaces
5. To disperse light in optical instruments such as spectrographs.

Single Prisms as Reflectors

A prism like that of **Fig. 1** can act as a very effective reflector; it is more efficient than a mirror since there is no metal coating to absorb light. **Fig. 2** shows a light ray entering the prism at right angles to one of the smaller sides, reflected at the larger side, and emerging at right angles to the other small side. The path shown is determined by Snell's law. This law has the form:

$$n \sin \theta = n' \sin \theta' \quad (1)$$

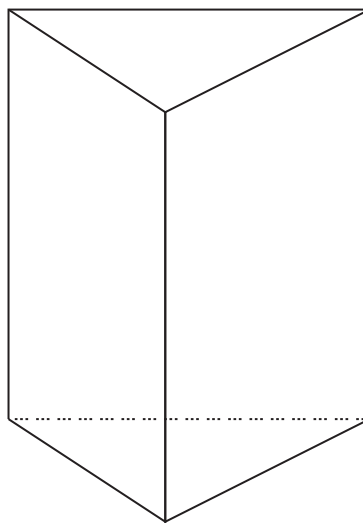


Fig. 1 A 45–45–90 degree prism.

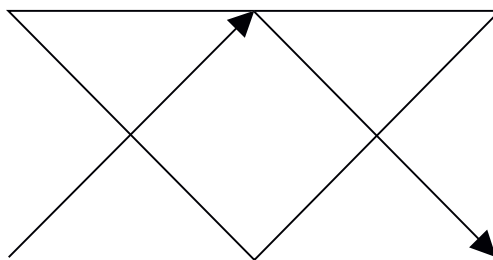


Fig. 2 Total internal reflection by a prism.

This indicates that a ray of light making an angle θ to the normal at the interface between media of indices of refraction n and n' will be refracted and the angle to the normal becomes θ' . Notice that the light ray strikes the surface where the reflection takes place at an angle of 45° to the normal; this angle is too large to permit it to emerge from the prism after refraction. To understand why, we recognize that when the ray is able to leave the glass, Snell's law requires that the emerging angle be larger than the incident angle. The largest possible incident angle – known as the critical angle – will correspond to an emerging angle of 90° ; that is, this ray will be parallel to the interface. For an index of 1.50, as an example, the critical angle θ_c satisfies Snell's law in the form:

$$1.5 \sin \theta_c = 1.0 \sin 90^\circ \quad (2)$$

from which $\theta_c = 41.8^\circ$. The ray shown in the figure will therefore remain in the glass. Furthermore, since the indices appearing on both sides of Snell's law are now those of the glass, then the incident and refracted angles must be identical, and the prism reflects the ray in the same way that a mirror would. This phenomenon is known as total internal reflection.

This prism has an important application as an optical component of a fingerprint machine, whose action depends on modifying the total internal reflection which normally occurs. Fig. 3 shows the prism and the reflecting behavior of Fig. 2, with a person's finger pressed against the long side. When the ray strikes a place on the surface corresponding to a valley in the fingerprint pattern, the index at this region is that of air, and the ray is reflected in the usual manner. But a ray striking a portion of the prism which lies immediately below a raised section of the skin will make an angle with the normal which is now smaller than the critical angle, since the index is much greater than that of air, so that the ray will cross the interface and be absorbed by the finger. The reflected beam, with the fingerprint pattern sharply reproduced, is sent to a detector and the resulting signal goes to a computer, where it is processed and used to drive a printer. The fingerprint cards generated by this kind of equipment are sharper than those done by the traditional smeared-ink method. It is possible to observe the phenomenon just described by looking at your fingers while holding a glass of water – the fingerprints will show up in an enhanced manner. In addition to the better quality, the manufacturer of one version of this equipment has incorporated a feature in their computer program which can detect attempts by the person being fingerprinted to degrade the image.

A more elaborate single prism – one with five sides – is the pentaprism of Fig. 4, used extensively in single lens reflex cameras. This prism does not take advantage of the total internal reflection process, as described above. Below the bottom face is a mirror at

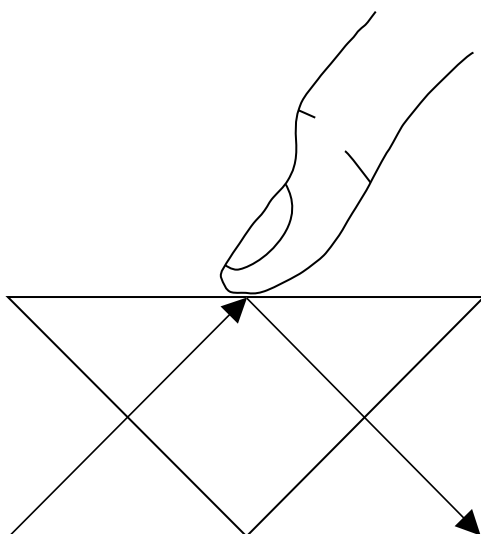


Fig. 3 Use of a prism to produce a fingerprint.

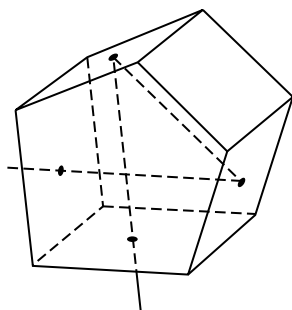


Fig. 4 A pentaprism.

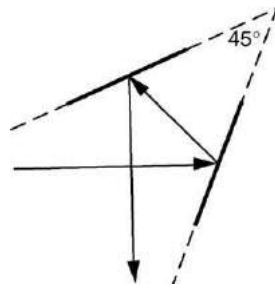


Fig. 5 Action of coated faces of a pentaprism.

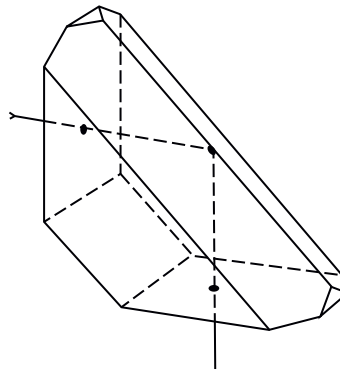


Fig. 6 Amici prism.

a 45° angle. Light from the camera's lens is reflected upward and emerges from the left-hand face where the scene being photographed is observed by the eye. Since there are two reflections, the reversal caused by a single mirror is eliminated. The camera's mirror is hinged and when the exposure button is pressed, it flies up and the incoming light can reach the film. The two reflecting faces, which lie at an angle of 45° to one another (**Fig. 5**), are coated. To see why this is necessary, note that the two angles in the isosceles triangle formed by the light ray and the face extensions must equal $(180^\circ - 45^\circ)/2$ or 67.5° . This ray must then lie at an angle of $90^\circ - 67.5^\circ$ to the normal, or 22.5° , which is much less than the critical angle needed for reflection.

There are at least a dozen other forms of prism which have been designed and used. For example, the Amici prism (**Fig. 6**) is a truncated right-angle prism with a roof section added to the hypotenuse face. It splits the image down the middle and thereby interchanges the right and left portions. These prisms must be very carefully made to avoid producing a double image. One use is in telescopes to correct the reversal of an image.

Double Prism Reflectors

Another extensive application of prisms is their use in pairs as components of telescopes and binoculars. The most widely used is a pair of Porro prisms: two right angle prisms arranged as shown in **Fig. 7**. The double change in direction has two effects; it eliminates image reversal and it shortens the spacing between objective and eyepiece lenses, making the binoculars lighter and more compact. It is possible to calculate how light rays behave in Porro prisms with a ray tracing procedure. This method involves the use of 4×4 matrices and its development is rather complicated. A simple understanding of the way light passes through this optical component can be based on **Fig. 8**, which shows the two prisms separated for clarity. An arbitrary ray (the heavy line) coming from the object will be reflected at a 45° angle at each of the four faces that it strikes. Using two line segments at right angles as an object, the ray from the top of the vertical arrow and the ray from the end of the horizontal arrow will behave as indicated. The image is then identical to the object but rotated by 180° . Since the optics of the binoculars – the combined effect of the objective and the eyepiece – result in an inverted image, the Porro prism corrects this problem.

Prisms as Instruments

Prisms are the crucial optical element in many laboratory instruments. To show how they can be used to measure the index of refraction of transparent materials – solids or liquids – consider the prism of **Fig. 9**, with apex angle A . Light comes to the prism at

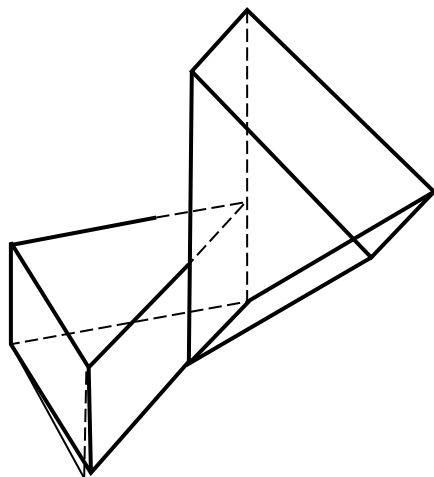


Fig. 7 A pair of Porro prisms.

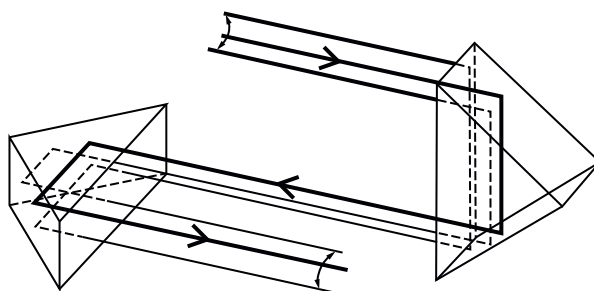


Fig. 8 Action of Porro prisms.

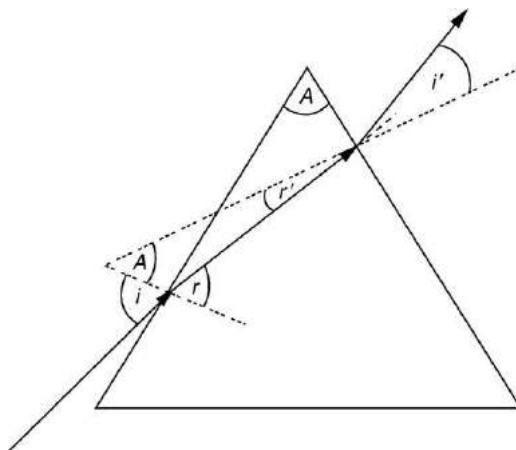


Fig. 9 Calculation of prism deviation.

an incident angle i , enters the prism at the angle of refraction r , crosses the prism and is refracted a second time when it emerges, the associated angles being designated as r' and i' . When the ray leaves the prism, it is traveling in a direction which represents an angular deviation δ from its original orientation. This deviation is the sum of the individual changes in direction at each face, or

$$\delta = (i - r) + (r' - i') \quad (3)$$

For the triangle formed by the projected normals, which meet at the angle A :

$$r = r' + A \quad (4)$$

Hence

$$\delta = i + i' - A \quad (5)$$

The minimum value of the deviation is an important property of the prism; we find it by differentiation to obtain:

$$d\delta = di + di' = 0 \quad (6)$$

Differentiation of the form of Snell's law valid at each interface, namely:

$$\sin i = n \sin r, \quad \sin i' = n \sin r' \quad (7)$$

gives

$$di = n \cos r \, dr / \cos i \quad (8)$$

$$\begin{aligned} di' &= n \cos r' \, dr' / \cos i' \\ &= n \cos r' \, dr / \cos i' \end{aligned} \quad (9)$$

where dr' can be replaced by dr , as obtained from the equation connecting r , r' , and A . It now follows that:

$$\begin{aligned} d\delta &= n \left[\cos r / \sqrt{1 - n^2 \sin^2 r} \right] \\ &\quad - \left[\cos(A - r) / \sqrt{1 - n^2 \sin^2(A - r)} \right] dr = 0 \end{aligned} \quad (10)$$

from which

$$\begin{aligned} \cos^2 r \{1 - n^2 \sin^2(A - r)\} \\ = \cos^2(A - r) \{1 - n^2 \sin^2 r\} \end{aligned} \quad (11)$$

or:

$$\cos^2 r = \cos^2(A - r) \quad (12)$$

and finally:

$$r = A/2, \quad r' = A/2, \quad i' = i \quad (13)$$

The differentiation thus leads to the conclusion that the ray enters and leaves the prism in a symmetrical manner. The amount of deviation is now:

$$\delta = 2i - A \quad (14)$$

or:

$$i = (\delta + A)/2 \quad (15)$$

and:

$$n = \sin i / \sin r = \sin\{(\delta + A)/2\} / \sin(A/2) \quad (16)$$

Using this expression to calculate n as a function of δ , it is found that δ is a minimum, so that a measurement of the minimum deviation produced by a prism gives the value of its index of refraction in a direct and accurate way. This method can also be used for liquids; a triangular glass cell with walls of reasonable thickness is made very accurately and used to measure δ . The refraction due to the walls does not affect the measurement, since the refraction occurring at the entrance wall is equal but opposite to that at the exit wall.

The description of prism behavior given so far assumes that we are dealing with monochromatic light; that is light of a single wavelength or single color in the visible spectrum. However, the index n of a prism varies with wavelength and white light passing through a prism will undergo dispersion – each component of a beam will emerge at a different angle and can be observed. The prisms that are part of expensive glass chandeliers display such an effect. This property of a prism is used in spectrometers – instruments which can measure the wavelength of light. **Fig. 10** shows a schematic diagram of this kind of instrument of high quality. Note that the light is brought into and taken out of the spectrometer with doublets. Various kinds of spectrometers have been put into use, with the ability to resolve closely spaced wavelengths being a crucial factor in the design.

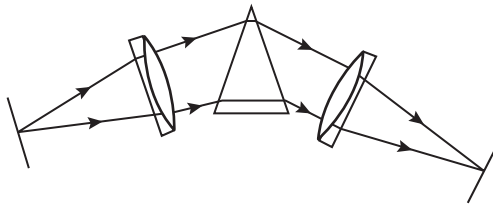


Fig. 10 A prism spectrometer.

Further Reading

Brouwer, W., 1964. *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Longhurst, R.S., 1967. *Geometrical and Physical Optics*, 2nd edn. New York: John Wiley & Sons.

MIL-HDBK-141, 1962. *Military Standardization Handbook*. Washington, DC: US Government Printing Office.

Acousto-Optics

M Gottlieb and D Suhre, Carnegie Mellon University, Pittsburgh, PA, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

b	dispersion constant (–)	p	photo-elastic constant (–)
c	elastic stiffness (dynes cm ^{–2})	P	pressure (dynes cm ^{–2})
D, d	optical beam diameters (cm)	P_a	acoustic power (watts cm ^{–2})
e	elastic strain (–)	V	acoustic velocity (cm sec ^{–1})
f	acoustic frequency (hertz)	β	compressibility (cm ² dyne ^{–1})
F	optical focal length (cm)		diffraction efficiency (–)
F_ϕ, F_θ	angular form factors (–)	ϵ	dielectric constant (–)
k	optical momentum wavevector (cm ^{–1})	ω	optical frequency (hertz)
K	acoustic momentum wavevector (cm ^{–1})	Ω	solid angle (steradians)
M_2	acousto-optic figure of merit (sec ³ gram ^{–1})	ρ	density (grams cm ^{–3})
n	refractive index	θ	Bragg angles (radians)
		τ	acoustic wave travel time (sec)

Acousto-optics deals with the interaction of light waves with sound waves and has given rise to a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The basic principles and phenomena controlling these interactions were largely understood by the mid-1930s, but had little practical significance because very high acoustic power levels were required to attain good optical efficiency and there were few good sources of well-collimated monochromatic light. During the period from 1960 to the present, several key technologies were developing rapidly, at the same time that many applications of the laser were being suggested which require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic acousto-optic crystals with low-drive-power requirements and low acoustic losses at high frequencies. This combination of properties makes acousto-optics feasible for many systems, and for several is the method of choice to satisfy demanding requirements. This chapter will summarize the basic features of acousto-optic interactions and the operating principles of the most common acousto-optic devices.

The Photo-Elastic Effect

The change induced in the refractive index of a transparent material by the pressure change produced by an acoustic disturbance is the underlying mechanism of all acousto-optic interactions. An acoustic wave produces periodic regions of compression and rarefaction in the material, which modulates the density. The Lorentz–Lorenz relation relates the refractive index to the density, for the simplest case of an ideal gas

$$\frac{(n^2 - 1)}{(n^2 + 2)} \propto \rho \quad (1)$$

where n is the refractive index and ρ is the density. This relation is followed to a good approximation for most simple solid materials as well. The elasto-optic coefficient is

$$\frac{\rho}{dn} \frac{dn}{d\rho} = \frac{(n^2 - 1)/(n^2 + 2)}{6n} = p \quad (2)$$

The fundamental quantity given by Eq. (2), also known as the photo-elastic constant p , is related to the pressure applied by

$$p = \frac{1}{\beta} \left(\frac{dn}{dp} \right) \quad (3)$$

where P is the applied pressure and β is the compressibility of the material. The photo-elastic constant of an ideal material with refractive index of 1.5 is 0.59. The photo-elastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6.

The relation in Eq. (3) follows from the usual definition of the photo-elastic constant:

$$\Delta(1/\epsilon) = \Delta(1/n^2) = pe \quad (4)$$

where ϵ is the dielectric constant, $\epsilon = n^2$, and e is the acoustic strain amplitude. The induced change in refractive index, Δn , is

$$\Delta n = -n^3 pe \quad (5)$$

Strain amplitudes typically lie in the range of 10^{-8} to 10^{-5} , with Δn in the range of about 10^{-8} to 10^{-5} (for $n = 1.5$). Devices based upon such a small change in refractive index are capable of generating large effects because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The more complete relation defining the photoelastic interaction is more complex than the simple scalar Eq. (5) in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A tensor relation between the dielectric properties, the elastic strain, and the photo-elastic coefficient describes the interaction, particularly for anisotropic materials. The tensor equation is

$$\Delta(1/n^2) = \sum_{ijkl} p_{ijkl} e_{kl} = \Delta(1/\epsilon)_{ij} \quad (6)$$

where $(1/\epsilon)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the Cartesian strain components, and p_{ijkl} are the components of the photo-elastic tensor. The crystal symmetry of the photoelastic material places limits on the possible configurations of interaction geometry.

Diffraction by Acoustic Waves

For typical acousto-optic devices the acoustic wave acts as a diffraction grating, made up of periodic changes in optical phase, moving at sound velocity. These features determine the properties of acousto-optic diffraction. Using a quantum-based model, the light and sound may be thought of as particles, photons and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. A description of both is given here. Consider Fig. 1 in which the light wave, of frequency ω and wavelength λ , is incident into the material with an acoustic wave of frequency f and wavelength Λ . If the refractive index of the medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \quad (7)$$

where L is the length of the interaction region. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm and $\lambda = 0.5$ μm , with Δn reaching a peak value of 10^{-5} . This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ , the number of optical wavelengths, is 50 000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0 e^{i\omega t} \quad (8)$$

then the field of the phase-modulated emerging light will be

$$E = E_0 e^{i\omega t + \Delta\phi} = e^{i\omega t} e^{2\pi i(L/\lambda)\sin(\omega t)} \quad (9)$$

where f is the acoustic frequency.

A detailed derivation of the resulting temporal and spatial distribution of the light field is mathematically complex, but, analogy with radio-wave modulation can be used to arrive at the resultant fields. The spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency f , as shown in Fig. 2. Sidebands are produced about the carrier frequency, such that the frequency of the m th sideband is $\omega + mf$, where m may be positive or negative. The amplitude of each of the sidebands is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. The odd-numbered negative orders are 180° out of phase with the even-numbered ones, a feature that may be useful for certain signal processing applications. The light emerging from the delay line is composed of a

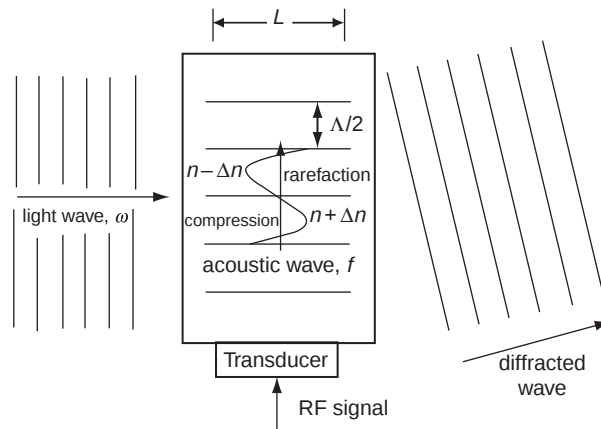


Fig. 1 Interaction of light waves with acoustic waves.

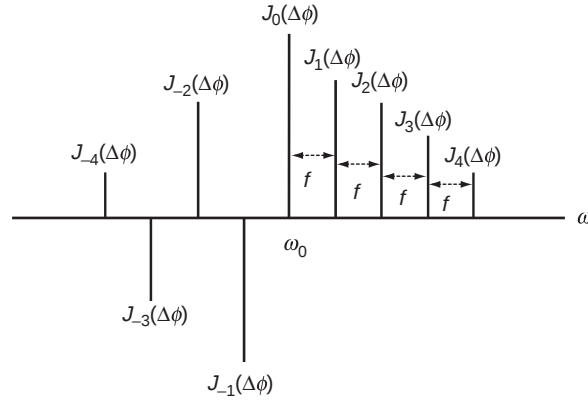


Fig. 2 Spectrum of a phase-modulated wave of carrier frequency ω_0 , and modulation index $\Delta\phi$.

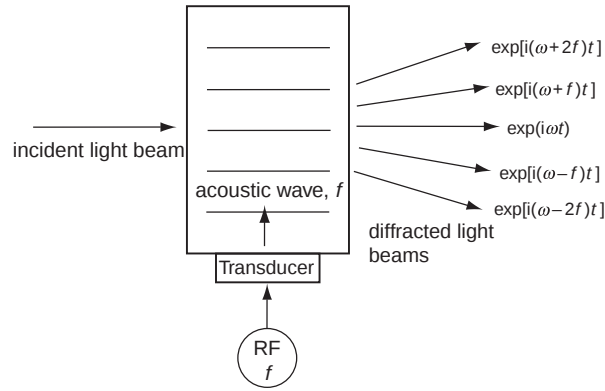


Fig. 3 Diffraction of light in the Raman-Nath limit.

number of light waves whose frequencies have been shifted by mf from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index. The two extremes of the diffraction process are simplified in the 'thin grating' and the 'thick grating' models. The ratio of the interaction length, L , to the acoustic wavelength, Λ , will determine the character of the diffraction process. The plane wave approximation in which the acoustic energy propagates as a plane wave is valid when this ratio is very large. However, when this ratio is small, the acoustic propagation can be described in terms of a sum of plane waves, the angular spectrum of which increases as the ratio decreases. In this phenomenological model the partial wave which is propagating at an angle λ/Λ to the forward direction may diffract the light second time into an angle $2\theta = 2\lambda/\Lambda$, and the frequency of this light will once again be up-shifted, for a total frequency shift of $2f$. If the spectrum of acoustic waves contains sufficient power of still higher orders, then this rediffraction process can be repeated, so that light will be multiply diffracted m times into higher-order angles, $m\theta = m\lambda/\Lambda$ each with a frequency shift mf . A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Fig. 3, where the deflection angle corresponding to the m th order is given by $\sin\theta_m = m\lambda/\Lambda$ and the frequency of the light deflected into the m th order is $\omega_m = \omega + mf$. The intensity of the carrier wave, or zero order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The first order will be a maximum for $\Delta\phi = 1.8$, corresponding to the maximum of the first-order Bessel function B_1 , and decreasing for larger modulation indices. These phenomena were described by Debye and Sears and so are often referred to as Debye-Sears diffraction. An extensive theoretical analysis of the effect was given by Raman and Nath and is alternatively referred to as Raman-Nath diffraction. As the interaction length is increased the Raman-Nath diffraction gradually weakens. The weakening begins around an interaction length $L \sim \Lambda^2/4\lambda$. This value of L is expressed in the Q parameter

$$Q = 4L\lambda/\Lambda^2 \quad (10)$$

which is known as the Raman-Nath parameter. A different regime of diffraction takes effect for values of the interaction length $Q \gg 1$.

The phenomena in this regime can be more easily understood in terms of the quantum description of the light and sound waves as collisions between photons and phonons. In this model, the dynamics of the collisions of the light and sound are governed by the laws of conservation of energy and momentum. The momentum vectors magnitudes of the light and sound, k and

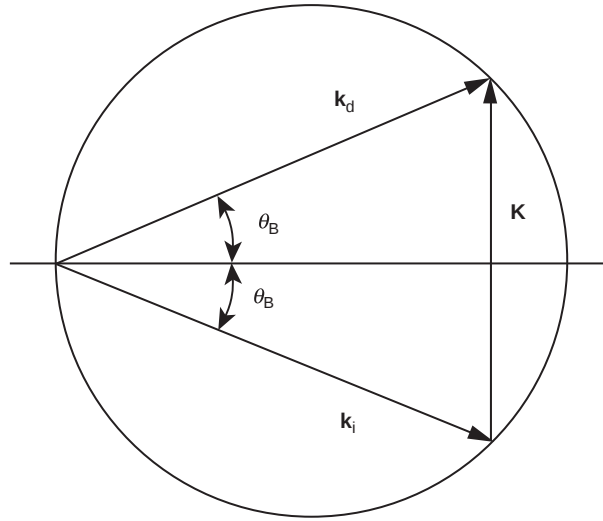


Fig. 4 Momentum diagram for diffraction of light.

K , are given by the well-known expressions

$$|k| = 2\pi n/\lambda \text{ and } |K| = 2\pi/\Lambda \quad (11)$$

where the acoustic wavelength, Λ , is related to the acoustic velocity, V , by $\Lambda = V/f$.

Conservation of momentum is expressed by the vector relation

$$k_i + K = k_d \quad (12)$$

the diagram for which is shown in **Fig. 3**, where k_i , and k_d represent the momentum of the incident photon and the diffracted photon, respectively. In this notation

$$k_i = 2\pi n_i/\lambda, \quad (13a)$$

and

$$k_d = 2\pi n_d/\lambda \quad (13b)$$

If the material is anisotropic (birefringent), n_i and n_d may be different.

Conservation of energy requires that

$$h\omega_d = h\omega_i + hf \quad (14)$$

in which h is Planck's constant. Since ω_i lies in the optical frequency range, $\sim 10^{13}$ Hz, and f lies in the RF or microwave range, 10^6 – 10^9 Hz, then $\omega_d \sim \omega_i$. This results, for isotropic materials, in the magnitudes of k_i and k_d being equal and resulting in the isosceles momentum triangle of **Fig. 4**. The angles of incidence and diffraction (with respect to the planar acoustic wavefronts), called the Bragg angles, are equal for this case, and are

$$\theta_B = (1/2)\lambda/\Lambda \quad (15)$$

A schematic diagram for this process is shown in **Fig. 5**. The interaction will be large only if the light is well-collimated, and incident at this angle. Unlike the Debye–Sears regime, there is only a single diffracted beam. The energy of the phonon may either be added to that of the incident photon, increasing its frequency to $\omega_d = \omega_i + f$, or the reverse, resulting in $\omega_d = \omega_i - f$. The sense of the momentum vectors determines which of these occurs. The Debye–Sears effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ , Λ , and L .

Anisotropic Bragg Diffraction

In optically anisotropic materials, acousto-optic Bragg diffraction can occur between an ordinary and extraordinary optical beam, and vice versa. This is generally referred to as birefringent diffraction, and it offers additional design capabilities compared to the isotropic case, such as enhancing the angular aperture, and extending the aperture–bandwidth product.

The theory of diffraction of light so far described has assumed that the optical medium is isotropic. A number of important acousto-optic devices make use of the properties of birefringent materials. It is different from diffraction in isotropic media in that the magnitude of the momentum vector of the light will in general be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of **Fig. 4**. The momentum vector for ordinary polarized light will, in general, be different from the momentum vector for light that is extraordinary polarized.

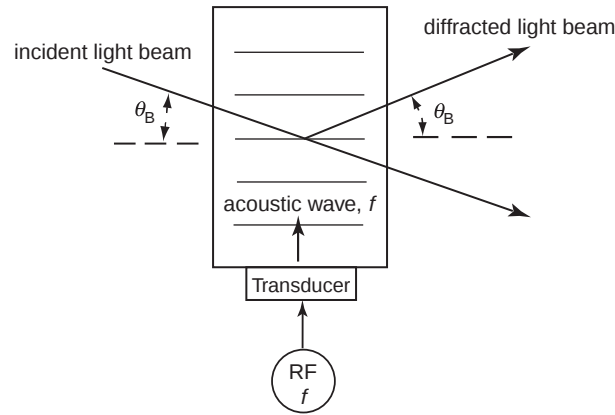


Fig. 5 Acousto-optic diffraction cell.

To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon which occurs when light interacts with shear acoustic waves, i.e., waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90° . The underlying reason for this is that the shear disturbance induces a birefringence which causes the plane of polarization to be rotated. The acousto-optic tunable filter (AOTF) is a particularly interesting birefringent device. It was first developed for the case of collinear optical and acoustic beams, and was used for wavelength selection in a dye laser.

Diffraction Efficiency

The acoustic power required to diffract light will be determined by the geometric parameters as well as by the properties of the medium. A simplified calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave shown in [Fig. 2](#), we can see that the ratio of the intensity in the first order to that in the zero order is

$$I_1/I_0 = [B_1(\Delta\phi)/B_0(\Delta\phi)]^2 \quad (16)$$

By analogy this same result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$P_a = \frac{1}{2} c V e^2 \quad (17)$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β , the density ρ , and acoustic velocity v through the expression

$$c = \frac{1}{\beta} = \rho V^2 \quad (18)$$

and the acoustic power density is

$$P_a = \frac{1}{2} \rho V^3 e^2 \quad (19)$$

The optical phase modulation of the medium resulting from the change in refractive index, Δn , is

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \quad (20)$$

Using [Eq. \(5\)](#) for Δn , the phase modulation is related to the acoustic power density by

$$\Delta\phi = -\pi(L/\lambda)n^3p(2P_a/\rho V^3)^{1/2} \quad (21)$$

The diffraction efficiency can now be calculated by using this value for the optical phase change in Eq. (16). The first- and second-order Bessel functions can be approximated for small modulation index by

$$\begin{aligned} B_0(\Delta\phi) &\sim \cos(\Delta\phi) \sim (\Delta\phi)^2, \text{ and} \\ B_1(\Delta\phi) &\sim \sin(\Delta\phi) \sim \Delta\phi \end{aligned} \quad (22)$$

The small signal approximation to the diffraction efficiency is then

$$I_1/I_0 \sim (\Delta\phi)^2 = (\pi^2 L^2 / 2\lambda^2) (n^6 p^2 / \rho V^3) P_a \quad (23)$$

The expression $n^6 p^2 / \rho V^3$ is known as the figure of merit of the material, and is designated as M_2 ; it is comprised entirely of intrinsic material properties.

Acousto-optic Materials

Acousto-optic device technology has matured to the point that performance is chiefly limited by material parameters, particularly the figure of merit and acoustic attenuation. Nature has arranged that materials with high figures of merit usually have high attenuation and vice versa. The widely used acousto-optic materials are fused quartz, tellurium dioxide, and lithium niobate. Development work on new infrared materials has been reported recently. A list of commonly used acousto-optic materials is given in Table 1. For materials with a low figure of merit, a higher drive power can be used to obtain the required efficiency.

Experience has indicated that the upper limit for very small devices (active area $\sim 0.1 \text{ mm}^2$) is a drive power density of $100\text{--}500 \text{ mW/mm}^2$, provided there is proper heat sinking to transfer the heat energy. For larger devices, sizes greater than $\sim 1 \text{ cm}^2$, the limit is closer to a few W/cm^2 . At the high drive power levels, the acoustic attenuation may cause significant optical distortion.

AO Devices

Resolution, bandwidth, and speed are the important characteristics of acousto-optic scanners, shared by all types of scanning devices. Depending upon the application, only one, or all, of these characteristics may have to be optimized. Consider the acousto-optic scanner in Fig. 6 with a collimated incident beam of width D , diffracted to an angle θ_B at its RF bandcenter, and whose bandwidth is Δf . If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$\Delta x = F\Delta\phi \sim F - \lambda/D \quad (24)$$

Table 1 Selected acousto-optic materials

Material	Transmission range (μm) mission	Acoustic mode & propagation direction	v (cm/s $\times 10^5$)	Acoustic attenuation (dB/cm GHz ²) ^c	n	M_2 (s ³ /g $\times 10^{-18}$)
Visible–near-infrared (VIS–NIR)						
LiNbO ₃	0.04–4.5	L[100] ^a	6.57	0.15	2.20	7.0
		S[001] ^b	3.59	2.6	2.29	2.92
TiO ₂ ,	0.45–6	L[001]	10.3	0.55	2.58	1.52
α -SiO ₂ ;	0.12–4.5	L[001]	6.32	2.1	1.54	1.48
		L[100]	5.72	3.0	1.55	2.38
TeO ₂ .	0.35–5	L[001]	4.20	15	2.26	34.5
		S[110]	0.616	90	2.26	793
Far IR						
Ge	2–20	L[111]	5.50	30	4.00	840
Tl ₃ AsS ₄ (TAS)	0.6–12	L[001]	2.5	29	2.63	510
GaAs	1–11	L[110]	5.15	30	3.37	104
ZnTe	0.55–20	L[110]	3.37	130	2.77	18
GaP	0.6–10	L[110]	6.32	60	3.31	30
Tl ₃ PSe ₄	0.85–9	L[100]	2.0	150	2.9	2069
Te	5–20	L[100]	2.2	60	4.8	4400
CdS	0.5–11	L[100]	4.17	90	2.44	12
GaP	0.6–10	L[110]	6.32	60	3.31	30
ZnS	0.4–12	L[001]	5.82	27	2.35	3.4
		S[001]	2.63	130	2.35	8.4

^aLongitudinal acoustic waves propagating along the [100] crystallographic direction.

^bShear acoustic waves propagating along the [001] crystallographic direction.

^cAttenuation constant at 1 GHz; the frequency dependence of the attenuation for most crystals is quadratic.

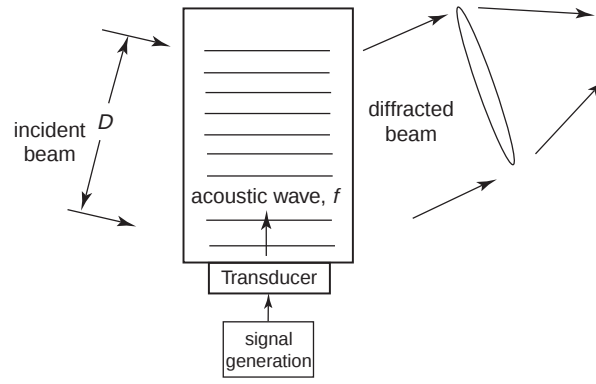


Fig. 6 Acousto-optic scanner.

where F is the focal length of the lens. The light intensity will be distributed in the focal plane with a sinc^2 distribution by aperture diffraction. The number of resolvable spots, N , will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta\theta}{\Delta\phi} \quad (25)$$

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta\theta = (\lambda/V \cos \theta_0) \Delta f \quad (26)$$

and

$$N = \Delta f (D/V \cos \theta_0) = \Delta f \tau \quad (27)$$

where θ_0 is the Bragg angle at band center, and τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the acousto-optic scanner, a concept applied to a variety of electronic devices as a measure of information handling capacity. The time–bandwidth product of an acousto-optic Bragg cell is equivalent to the number of bits of information which may be instantaneously processed by the system. In order to maximize the number of resolution elements, it is desirable to have as large a bandwidth as possible (i.e., large frequency range) and also as large an aperture delay time as possible. There are two factors limiting the bandwidth of an acousto-optic device: the bandwidth of the transducer structure (discussed later), and acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to 3 dB. A reasonable approximation of the maximum attainable bandwidth is

$$\Delta f = 0.7 f_{\max} \quad (28)$$

Birefringent Scanners

The birefringent scanner is a significant use for anisotropic Bragg diffraction. There are a number of advantages with the birefringent scanner over the anisotropic scanner, such as a larger angular scan range along with lower frequencies. There are also disadvantages, such as lower speed due to generally lower shear wave velocities, and the particular application will dictate whether an isotropic or birefringent scanner is better. Scanners represent a fairly important aspect of acousto-optic devices due to the widespread commercial use of laser beam deflectors for displays and laser printers. For applications where the required speed is beyond that of mechanical scanners, the acousto-optic scanner is an ideal candidate. However, unlike mirrors, acousto-optic deflectors are wavelength sensitive, and can only be used with single-wavelength laser beams.

The birefringent scanner can be described with the wavevector diagram, as shown in Fig. 7 for a positive uniaxial crystal where the extraordinary index of refraction is larger than the ordinary. The acoustic wavevector is tangent to the diffracted surface, which produces the largest scan angle for a given acoustic bandwidth. This is also the degenerate case, where only a single diffracted beam results, whereas two beams at two different acoustic frequencies will generally result from arbitrary input and acoustic propagation directions. The azimuthal acceptance angles (angles normal to the plane of incidence) of the acoustic and optical wavevectors can be different, although propagation and polarization directions must be selected that will produce high efficiency.

For the positive uniaxial case, the acoustic and diffracted wavevectors will be perpendicular at the design point, which allows the center frequency to be calculated from geometry as

$$f_0 = (V/\lambda)(n_i^2 - n_o^2)^{1/2} \quad (29)$$

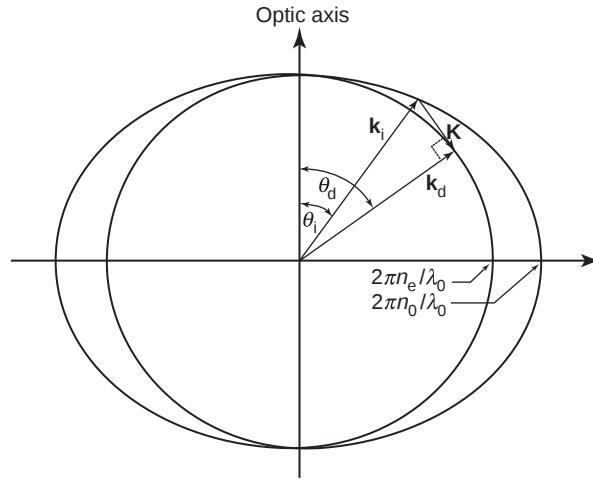


Fig. 7 Wavevector diagram for birefringent scanner using a positive uniaxial crystal.

where n_i is the index of refraction at the incidence angle. Due to the typically low velocity of shear waves, this frequency can be much lower than an isotropic scanner, and the incidence angle can also be chosen to adjust the frequency. It is also possible to use the optical activity of certain materials such as TeO_2 , along with circularly polarized light to further reduce the acoustic frequency. It is important to maintain as low a frequency as possible due to acoustic attenuation, which is especially high with the soft materials typically used for birefringent applications.

The bandwidth over which the scanner can efficiently operate is fairly large due to having the acoustic and diffracted wavevectors perpendicular, and is approximately $\Delta f = 2f_0$, assuming an octave of bandwidth for the isotropic scanner. This will produce a larger scan angle than an isotropic scanner, and more spots of resolution. The number of resolution spots is determined through diffraction by the aperture size and diffraction, along with the scan angle. Since a larger aperture requires a longer time for the acoustic waves to propagate across the aperture, the response time τ to access the scanner will increase, and the product of τ and f is related to the number of spots by Eq. (27). The advantage of the birefringent scanner is that it can operate at a lower frequency f_0 with better performance. However, since other factors such as acoustic attenuation are important in designing a scanner, for some applications it might be better to operate an isotropic scanner at a higher frequency.

AOTFs

With anisotropic Bragg diffraction, the magnitude of the diffracted wavevector k_d will differ from the incident wavevector k_i , which cannot occur for the isotropic case. This is illustrated in Fig. 8 for an AOTF utilizing a negative uniaxial crystal where an extraordinary input wave propagating at an incident angle θ_i relative to the crystal axis is diffracted into an ordinary output wave at an angle θ_d . This occurs through an acoustic wave propagating at an angle θ_a with wavevector K_a , where all the wavevectors lie in a plane through the optic axis of the crystal. This diagram is identical to the index ellipsoid for the crystal, scaled by $2\pi/\lambda$, where λ is the vacuum wavelength. The acoustic wavevector is shown as adding to the incident wavevector, which increases the frequency of the optical wave by the acoustic frequency. It can also be represented in the reverse direction, which would reduce the optical frequency by the same amount.

An AOTF spectrally filters the optical input, and maintains its spectral purity over a large angular aperture. These conditions are maximized when the power flow or ray directions of the input and output beams are collinear. This produces a parallel-tangents condition, where lines drawn tangent to the wavevector surfaces and connecting the input and diffracted wavevectors are parallel. For beams at 90° to the optical axis, it is referred to as the collinear case, whereas all other parallel-tangents conditions are noncollinear. The diffracted beam is rotated by 90° during anisotropic Bragg interaction, so that crossed polarizers can be used to separate the input and diffracted beams. For the collinear case polarization separation must be used, whereas angular separation can also be used for the noncollinear case.

The AOTF design equations can be derived from the geometrical conditions imposed by the wavevector matching condition

$$k_d = k_i + K \quad (30)$$

along with decomposing the acoustic wave into its Fourier components which result from the finite interaction length produced by the transducer. These components allow for phase matching over a range of input angles, and they also allow for a spectral spread of the interaction. The full width half maximum (FWHM) resolution is given by

$$\Delta\lambda = \frac{1.8\lambda^2}{bL \sin 2\theta_i} \quad (31)$$

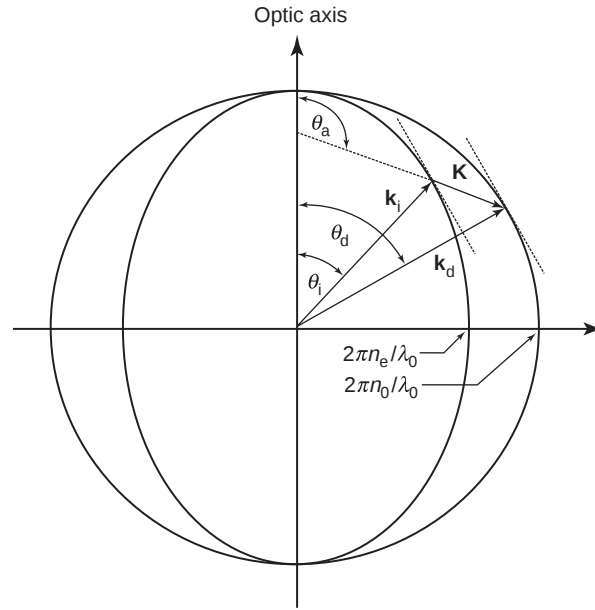


Fig. 8 Wavevector diagram for an AOTF using a negative uniaxial crystal.

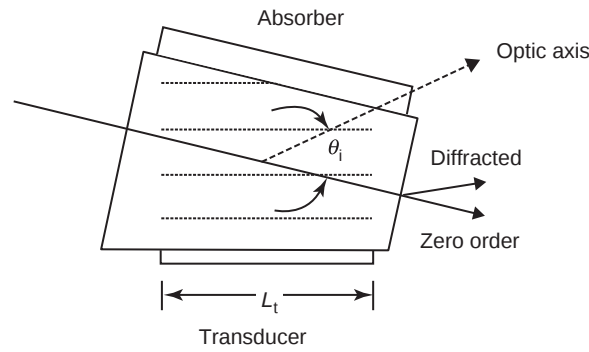


Fig. 9 Noncollinear AOTF orientation of optical and acoustic beams.

where λ is the vacuum wavelength, b is the dispersion term (essentially $2\pi\Delta n$, where the birefringence Δn is the difference between the ordinary index n_o and extraordinary index n_e of refraction), L is the interaction length of the input beam defined geometrically by the acoustic beam, and θ_i refers to the angle of the central or cardinal ray of the input beam, and the AOTF is designed around this angle. It is therefore a sensitive function of the incidence design angle, with the highest resolution occurring for the collinear case. The resolution is also narrower for larger birefringence.

The geometry of a noncollinear AOTF is shown schematically in **Fig. 9**. The acoustic waves propagate in the correct direction and are generated by a transducer of length L_t , which is related geometrically to the interaction length. These waves are generally absorbed at the other side of the AOTF to prevent interfering reflections, and the sides may also be wedged to eliminate reflections into the interaction region. The optical beam enters the AOTF at the correct angle to the crystal optic axis, and either diffracts or passes through as the zero order. The input beam will have an angular spread, producing an acceptance angle that is a function of the interaction length, wavelength, and crystal parameters, and the external solid angular aperture is roughly given by

$$\Delta\Omega = \frac{\pi n^2 \lambda}{\Delta n L} \quad (32)$$

where n refers to the diffracted beam index of refraction, which can be either n_o or n_e depending upon the design. The resolution and solid angle are therefore related through the transducer length, and the product of the resolving power times the solid angle is given by

$$(\lambda/\Delta\lambda)(\Delta\Omega) = 2\pi n^2 \sin^2 \theta_i / |F_\theta F_\phi|^{1/2} \quad (33)$$

where

$$F_\theta = 2\cos^2 \theta_i - \sin^2 \theta_i \quad (34a)$$

$$F_\phi = 2\cos^2 \theta_i + \sin^2 \theta_i \quad (34b)$$

are form factors for the polar and azimuthal components of the optical field. Since all the wavevectors lie in a plane containing the optic axis, the azimuthal components of the incident, diffracted, and acoustic wavevectors must always be identical under the parallel-tangents condition.

The product $(\lambda/\Delta\lambda)(\Delta\Omega)$ forms a figure of merit for spectrometers, and for the collinear case, the product is identical to a Fabry-Perot etalon having an index of n , and at other angles it is more complicated due to the angular dependence. The angular field is also symmetrical for the collinear AOTF, but nowhere else other than along the optic axis, where the resolving power becomes zero.

The great advantage of the AOTF designed under the parallel-tangents condition is the large angular aperture compared to the general case. Under this condition, the dependence of the angular aperture on resolving power is second order rather than first order, and the angular aperture can be tens of degrees for a typical resolution, which is useful for imaging applications. At the special angle of 54.7° , the second-order dependence is also zero, and the aperture can be extremely large relative to the resolving power, although this requires a high-resolution device with a corresponding narrow field, since the condition would soon be violated at larger field angles.

The efficiency of the AOTF under phase matching is given by

$$\eta = \sin^2 [\pi^2 M_2 P_a L^2 / (2\lambda^2)]^{1/2} \quad (35)$$

which depends on θ_i and various material properties. Since the interaction is anisotropic, M_2 must be taken as a tensor quantity, in which both the polarization and propagation directions of the acoustic and optical waves must be accounted for. For acoustic waves, the polarization direction is the particle motion, which is perpendicular to the propagation direction for shear waves. In general, M_2 is much larger for a specific configuration of a particular material, and AOTFs are designed around this condition. The range of useful design angles is also usually limited since M_2 is generally a sensitive function of θ_i . The angular dependence on M_2 is important in designing an efficient device since it can be near zero at specific design angles for some materials, such as with TeO_2 for the collinear case.

The AOTF must be designed to optimize performance. The resolution is generally given as a system requirement, and the design angle along with the transducer length must be adjusted to optimize the throughput, or total optical power through the AOTF. This requires maximizing the efficiency, angular aperture, and aperture dimension. The maximum crystal size that can be grown ultimately limits these parameters, both in the interaction length and in the aperture size.

AOTFs have been used for a wide range of spectral filtering applications for spectroscopy and laser applications. Both collinear and noncollinear devices have electronically tuned a variety of lasers, including dye, semiconductor, and Ti:sapphire lasers. The AOTF is placed within the laser cavity, which requires high transmission on the laser line, with enough resolution to separate nearby laser transitions. A related application is the multiplexing of optical communications systems, where the AOTF is used to separate the various laser diode wavelengths. Perhaps the most widely used application is spectroscopy, and single-point detection systems have been used for a variety of biological and chemical applications in the visible and infrared. These techniques have been extended to spectral imaging, again in the visible and the infrared. Due to a variety of AOTF aberrations the image quality will be somewhat degraded. Since the AOTF is electronically tunable, it can be rapidly modulated both in amplitude and wavelength, which allows for modulation spectroscopy to detect small signals in a large background. By applying multiple frequencies, the AOTF has also been used as a rejection filter, in which all wavelengths other than that selected are allowed to pass.

Modulators

Acousto-optic modulators are used to vary and control laser beam intensity. A Bragg configuration gives a single first-order output beam, where intensity is directly linked to the power of RF control signal. The rise time of the modulator is the time required for the acoustic wave to traverse the laser beam. For minimum rise time the laser beam will be focused down, forming a beam waist as it passes through the modulator.

A Bragg configuration gives a single first-order output beam, the intensity of which is controlled by the RF signal power, and diffraction angle controlled by the drive frequency. Two deflectors can be used in series and at right angles to each other to give full two-dimensional scanning.

One requirement in the design of acousto-optic modulators is that the diffracted beam and undiffracted beam must be well separated. For an adequate extinction ratio, the Bragg angle should be at least as large as the divergence of the optical beam. This condition puts a minimum value on the center frequency. Equating the Bragg angle

$$\theta_B = \frac{\lambda f_0}{2nV} \quad (36)$$

and the diffraction angle of the Gaussian beam

$$\delta\theta_0 = \frac{4\lambda}{\pi(nd)} \quad (37)$$

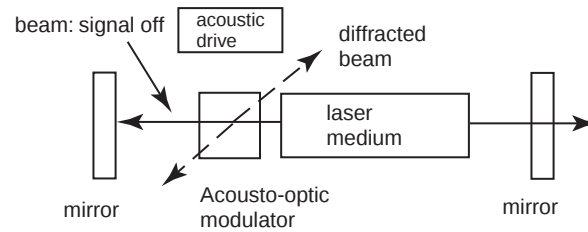


Fig. 10 An acousto-optic *Q*-switch in a laser cavity.

where d is the beam diameter at the minimum, it follows that the lower limit of the acoustic frequency is given by

$$f_0 = \frac{8}{\pi\tau} \quad (38)$$

We may combine these expressions to determine a limit of modulator bandwidth of acousto-optic modulators, with the result

$$\Delta f \sim \frac{1}{4}f_m \quad (39)$$

i.e., the modulation bandwidth is approximately equal to 25% of the midband acoustic frequency, f_m . In view of the present status of transducer technology, the modulation bandwidth of acousto-optic modulators is limited to several hundred MHz.

A frequency shifter uses the shift inherent in the acousto-optic interaction to up- or down-shift a laser's frequency. Two kinds of shifters can be distinguished: the fixed frequency shifter, and the variable frequency shifter. The frequency shift is equal to the signal frequency applied to the transducer. Various device configurations can be used to shift the laser beam, such as multiple travels inside the shifter to double or triple the frequency shift, or a combination of two frequency shifters in series. Acousto-optic frequency shifters can be used, for example, in optical heterodyning and interferometry, or in laser Doppler velocimetry for particle velocity analysis.

Q-Switches

A *Q*-switch is a special modulator which introduces high repetition rate losses inside a laser cavity (typically 1 to 100 KHz). They are designed for minimum insertion loss and to be able to withstand very high laser powers. In normal use an RF signal is applied to diffract a portion of the laser cavity flux out of the cavity. This increases the cavity losses and prevents oscillation. When the RF signal is switched off, the cavity losses decrease rapidly and an intense laser pulse evolves. A diagram of a typical acousto-optic *Q*-switched laser cavity is shown in [Fig. 10](#). *Q*-switches are the preferred method of intracavity modulation where high speed, high stability, and modest cost are important design factors.

Further Reading

- Chang, I.C., 1976. Acousto-optic devices and applications. *IEEE Transactions: Sonics and Ultrasonics* SU-23, 2–22.
- Chang, I.C., 1981. Acousto-optic tunable filter. *Optical Engineering* 20, 824–829.
- Dixon, R.W., 1967. Photo-elastic properties of selected materials and their relevance for applications to acoustic light modulators and scanners. *Journal of Applied Physics* 38, 5149–5153.
- Gordon, E.I., 1966. A review of acousto-optical deflection and modulation devices. *Proceedings of the IEEE* 54, 1391–1401.
- Gottlieb, M., Ireland, C.L.M., Ley, J.M., 1983. *Electro-optic and Acousto-optic Scanning and Deflection*. New York: Marcel Dekker.
- Goutzoulis, A.P., Pape, D.R., 1994. *Design and Fabrication of Acousto-optic Devices*. New York: Marcel Dekker.
- Uchida, N., Niizeki, N., 1973. Acousto-optic deflection materials and techniques. *Proceedings of the IEEE* 61, 1073–1092.
- Xu, J., Stroud, R., 1992. *Acousto-optic Devices: Principles, Design, and Applications*. New York: Wiley.

The electro-optic (EO) effect can be described as any one of a number of phenomena that occur when an electromagnetic wave in the optical spectrum (e.g., characterized by wavelengths in the range 200 to 2000 nm) interacts with an electric field, or with matter under the influence of an electric field. Two of the most important electro-optic phenomena are the Kerr effect (discovered by John Kerr in 1875) and the Pockels effect (discovered by Friedrich Pockels in 1893), in which birefringence is induced or modified in a liquid (the Kerr effect) or a solid (the Pockels effect). Birefringence is the difference in refractive indices for light of orthogonal line polarizations, one of which is parallel to the induced optical axis. The Kerr effect involves creation of birefringence in a liquid that is otherwise not birefringent. The degree of birefringence is quadratically proportional to the applied electric field strength. Hence, the Kerr effect is frequently referred to as the quadratic electro-optic effect while the Pockels effect involves a linear dependence on the applied electric field and is referred to as the linear electro-optic effect. In considering practical applications of the electro-optic effect, a commonly encountered term is that of 'electro-optic modulator', which can be described as a device wherein a signal control element is used to modulate a beam of light. The control element is typically an electric field with frequency components in the zero (DC) to hundreds of gigahertz (1GHz = 10^9 Hz) range (or even tens of terahertz (1THz = 10^{12} Hz)). The modulation may be imposed on the phase, frequency, amplitude, or direction of the modulated optical beam.

Electro-optic effects are one class of second-order nonlinear optical phenomena. Other important second-order nonlinear optical phenomena include sum (e.g., second-harmonic) and difference frequency generation. Such phenomena derive from the second-order term in the power series expansion of macroscopic (material) polarization, P , in terms of applied electric fields:

$$P_i = \chi_{ij}^{(1)} E_j + \chi_{ijk}^{(2)} E_j E_k + \chi_{ijkl}^{(3)} E_j E_k E_l + \dots \quad (1)$$

where $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$, and $\chi_{ijkl}^{(3)}$ are the linear, second-, and third-order optical susceptibilities, respectively. Second-harmonic generation (SHG) or frequency doubling (2ω) effects, where a beam at twice the frequency of the input beam is generated, can be seen by substituting a sinusoidal field, $E_\omega \cos(\omega t - kz)$; for E_j and E_k . After using a well-known trigonometric identity, the equation for macroscopic polarization becomes (through second-order and dropping the subscript i):

$$P = \chi^{(1)} E_\omega \cos(\omega t - kz) + (1/2) \chi^{(2)} E_\omega^2 [1 + \cos(2\omega t - 2kz)] + \dots \quad (2)$$

The electro-optic effect can be appreciated by considering interaction of the medium with an optical field $E_\omega \cos(\omega t - kz)$ and a low-frequency field E_0 . Note that although we have used the symbol zero to denote the frequency of the low-frequency electrical field, actual frequencies can extend to hundreds of gigahertz or even tens of terahertz. Again substituting (and now keeping terms to third order), the equation for polarization can be written as:

$$P = \chi^{(1)} E_\omega \cos(\omega t - kz) + 2\chi^{(2)} E_0 E_\omega \cos(\omega t - kz) + 3\chi^{(3)} E_0^2 E_\omega \cos(\omega t - kz) + (3/4) \chi^{(3)} E_\omega^3 \cos(\omega t - kz) = \chi_{\text{eff}} E_\omega \cos(\omega t - kz) \quad (3)$$

The equation for the nonlinear index of refraction, n , can be expressed as

$$\begin{aligned} n^2 &= 1 + 4\pi\chi_{\text{eff}} \\ &= 1 + 4\pi[\chi^{(1)} + 2\chi^{(2)} E_0 + 3\chi^{(3)} E_0^2 + (3/4)\chi^{(3)} E_\omega^2] \end{aligned} \quad (4)$$

Denoting the linear index of refraction as n_0 , the above equation can be rewritten as

$$n^2 - n_0^2 = 8\pi\chi^{(2)} E_0 + 12\pi\chi^{(3)} E_0^2 + 3\pi\chi^{(3)} E_\omega^2 \quad (5)$$

which in turn leads to

$$n = n_0 + (4\pi/n_0)\chi^{(2)} E_0 + (6\pi/n_0)\chi^{(3)} E_0^2 + (3\pi/2n_0)\chi^{(3)} E_\omega^2 \quad (6)$$

The definition of light intensity in cgs units is $E_\omega^2 = (8\pi/cn_0)I(\omega)$, which when substituted into the above equation gives

$$n = n_0 + (4\pi/n_0)\chi^{(2)} E_0 + (6\pi/n_0)\chi^{(3)} E_0^2 + (12\pi/cn_0^2)\chi^{(3)} I(\omega) \quad (7)$$

The index of refraction can now be written as

$$n(\omega) = n_0(\omega) + n_1(\omega) E_0 + n_2(0) E_0^2 + n_3(\omega) I(\omega) \quad (8)$$

where the terms $n_1(\omega)$, $n_2(0)$, $n_3(\omega)$ correspond to the linear electro-optic effect, the quadratic electro-optic effect, and the optical Kerr effect, respectively. The above equation can also be expressed as

$$n = n_0 - (1/2)n^3 r E_0 - (1/2)n^3 s E_0^2 - \dots \quad (9)$$

where r and s are the linear and quadratic electro-optic coefficients, respectively.

Unfortunately, a great deal of confusion exists concerning the use of the terms electro-optic effect and electro-optic modulator. First of all, the terms 'electro-optic' and 'opto-electronic' are frequently confused and used interchangeably. Opto-electronic devices function as electrical-to-optical or optical-to-electrical signal transducers. Examples of commonly used opto-electronic devices include light-emitting diodes (LEDs), photodiodes (PDs), injection laser diodes (ILDs), and integrated optical circuit (IOC)

elements. An electro-optic device can also function as an electrical-to-optical signal transducer (e.g., a Mach–Zehnder interferometer or a birefringent modulator employing polarizers at the input and the output); however, the mode of operation is entirely different for electro-optic and opto-electronic devices. For example, with LEDs or modulated lasers, the applied electric field produces modulation of transmission of light by altering the excited state population of the light emitting (lasing) state of the material. With electro-optic materials, no actual excited state population is generated. Electro-optic devices typically operate in regions of the optical spectrum removed from resonant transitions, i.e., regions of relatively high transparency. The applied electric field acts only to perturb the charge distribution of the material (the spatial positions of the electrons and nuclei). The altered charge distribution interacts with the optical beam transmitting the material with the result that the speed of light in the material (i.e., the index of refraction or birefringence) is altered. The response times of opto-electronic and electro-optic phenomena (and hence bandwidths of devices exploiting these phenomena) are quite different. In the former case, response will be limited by the lifetime of the relevant (emitting) excited state while in the latter case, response will be defined by the relaxation time of the material (e.g., reorientation time of a liquid or lattice relaxation time of a solid) following removal of the perturbing electric field.

An even greater confusion can arise due to the jargon used in particular technologies, such as telecommunications. A telecommunications engineer frequently refers to ‘electro-optic switching’ when describing opto-electronic transduction of an optical signal to an electrical signal, followed by re-routing in the electrical domain, and finally opto-electronic transduction of the electrical signal back to the optical domain. Electro-optic switching (in the sense used in this article) involves quite a different operation and would be described by that same telecommunications engineer as ‘all-optical switching’—a term reserved by physicists for the optical Kerr effect (control of one light beam by another light beam).

A second point of confusion involves distinguishing between the terms ‘electro-optic’ and ‘electro-absorptive’ modulation. Again, the terms are frequently used interchangeably. However, they involve quite different physical mechanisms. An electro-absorptive modulator is, like a modulated laser or LED, a ‘resonant’ device. In a device such as a gallium arsenide (GaAs) or indium phosphide (InP) quantum dot electro-absorptive modulator, the position and width of an optical absorption (resonant transition) are defined by the physical dimensions of the quantum dot. Application of an electric field shifts the spectral wavelength of the quantum dot absorption. This phenomenon can be used to dramatically change optical transmission through (or equivalently, absorption by) the material by choosing the wavelength of the propagating beam of light to be near the resonant absorption. With the electric field off, the material is reasonably transparent but becomes strongly absorbing when the electric field is applied, shifting the resonant absorption to the wavelength of the propagating beam. Obviously, the voltage required to achieve a desired change in optical transmission will depend on control of quantum dot dimensions in fabrication, which has been and continues to be a topic of research focus for electro-absorptive materials. Likewise, control of insertion losses of such devices is a concern as the wavelength used for the transmissive state of the device must be sufficiently close to the optical resonance to achieve the desired attenuation with application of a modest control voltage. Another issue to be faced with electro-absorptive devices is that of ‘chirp’ (optical signal distortion arising from the fact that both absorption and index of refraction – the imaginary and real parts of the optical susceptibility – change with application of the electric control field). On the positive side, electro-absorptive modulators are ‘polarization-insensitive’ modulators and are thus conveniently used with multimode (as well as single mode) optical transmission.

Further confusion exists because widely different types of materials can be used in electro-optic and electro-absorptive devices. As these materials are frequently competing for the same applications (signal transduction, optical switching, etc.), the advantages and disadvantages of different materials are often compared without maintaining the context that quite different phenomena are involved. Even restricting our discussion to electro-optic materials, we note that materials can range from organic liquid crystalline materials (nematic, smetic, ferroelectric, etc.) to organic electro-optic materials (both crystalline and ‘macromolecular’) to inorganic crystalline materials (lithium niobate, LiNbO_3 ; barium titanate, BaTiO_3 ; barium borate, BBO; potassium dihydrogen phosphate, KDP; lithium tantalate, LiTaO_3 ; zinc telluride, ZnTe , etc.). The response times of these materials to transient application of an applied electric field will be quite different, reflecting the different types of lattice motion involved. With liquid crystalline materials, particularly nematic materials, considerable molecular reorientation is involved and due to the mass that must be moved, response times are quite slow although the index change will be relatively large. With use of liquid crystalline materials, device bandwidths will typically be limited to tens of megahertz (MHz) or less. In contrast, conjugated π -electron organic materials typically exhibit response times of tens to hundreds of femtoseconds, which translates to potential device bandwidths of tens of terahertz (actual device bandwidths may be less due to factors other than material response). For conjugated π -electron organics, the response time is the phase relaxation time of the π -electron system. Because only a slight change in bond length alternation of the conjugated π -electron system occurs with application of an applied electric field and because strong electron–phonon coupling and resonance stabilization of the π -electron system act to reduce the barrier to lattice relaxation, the response times of π -electron organic materials are among the fastest observed in nature. With crystalline inorganic materials, the ionic constituents move to new locations with application of an electric field with the exact movement determined by the field strength, the charge on the ions, and the restoring force. Unequal restoring forces along the three mutually orthogonal (perpendicular) axes of the crystal lead to anisotropy in the optical properties of the material. The applied electric field will induce a change in the anisotropy (the principal refractive indices). The symmetry of the electro-optic tensor will be closely related to the symmetry of the piezoelectric tensor. The linear electro-optic effect requires that the crystal exhibit noncentrosymmetric (acentric or ferroelectric) symmetry. A centrosymmetric crystal possesses a center of symmetry defined by identical particles existing in the lattice at vectors $\mathbf{r}$ and $-\mathbf{r}$, where $\mathbf{r}$ is a position vector measured from an arbitrary origin. A centrosymmetric crystal, like an isotropic liquid or gas, can exhibit a quadratic electro-optic effect.

For the sake of completeness, it can also be noted that index of refraction changes can be induced by acoustic waves (acousto-optic modulation) and by heating (thermo-optic modulation). Also, elasto-optic and photo-elastic effects can produce index of refraction changes.

To keep this article to a manageable length, discussion is limited to solid-state electro-optic materials. Design principles being used to produce new organic electro-optic materials will be illustrated. Since inorganic materials exist as crystals, limited chemical modification of such materials is possible, although processing techniques for fabricating thin films of such crystalline materials have been developed in a number of cases. For organic macromolecular materials, development of new materials is an on-going activity. Several representative devices being fabricated, using electro-optic materials, will be discussed. Four of the most commonly encountered types of EO devices include: (i) stripline devices; (ii) prism, cascaded prism, and superprism devices; (iii) resonated devices, such as ring microresonator and photonic bandgap devices, and (iv) waveplates. Of course, devices such as stripline devices can be configured in a variety of ways to produce polarization-sensitive and polarization-insensitive electrical-to-optical signal transducers, optical switches either using simple 1×2 or 2×2 switch architectures or multimode interference (MMI) switches, optical gyroscopes, photonically controlled phased array radar systems, spectrum analyzers, optical digital signal processors, analog-to-digital (A/D) and digital-to-analog (D/A) signal converters, electromagnetic (radiofrequency to millimeter wave) signal generators, voltage sensors, etc. Prism-type devices are typically used for spatial light modulation or laser beam deflection. Ring microresonator devices afford a rich array of applications ranging from active wavelength division multiplexing (WDM) to active optical interconnect reconfiguration, to voltage-controlled wavelength tuning of laser outputs.

Basic Relationships

The effect of the applied field is to deform the index of refraction ellipsoid, which can be represented as

$$\left(\frac{1}{n^2}\right)_1 X^2 + \left(\frac{1}{n^2}\right)_2 Y^2 + \left(\frac{1}{n^2}\right)_3 Z^2 + 2\left(\frac{1}{n^2}\right)_4 YZ + 2\left(\frac{1}{n^2}\right)_5 XZ + 2\left(\frac{1}{n^2}\right)_6 XY = 1 \quad (10)$$

leading to the corrections

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^3 r_{ij} E_{0j} \quad (11)$$

where the electro-optic tensor components are related to the tensor components of the second-order nonlinear optical susceptibility by $r_{ij} = -8\pi/n_{0i}^2 n_{0j}^2 \chi_{ji}^{(2)}$. The full tensorial equation for change of the 'indicatrix' (index of refraction ellipsoid) can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_{0X} \\ E_{0Y} \\ E_{0Z} \end{pmatrix} \quad (12)$$

The electro-optic tensor reduces considerably for specific materials reflecting the symmetry of the particular material. For inorganic crystalline materials, the electro-optic tensor can be reasonably complex (containing many nonzero elements). It is also helpful to note that for crystalline inorganic materials, the electric-field-induced change in crystal shape leads to strain and the orientation of the indicatrix is altered, leading to an additional contribution to the change in index coefficients:

$$\Delta(1/n^2)_i = \sum_k p_{ik} S_k + \sum_j r_{ij} E_j \quad (13)$$

where S_k is a component of the strain and p_{ik} is the elasto-optic tensor. At high frequencies, the inertia of the crystal prevents macroscopic straining so that the first term of the above equation vanishes. At low frequencies, elasto-optic effects cannot be ignored. Because the deformation leading to strain is generally caused by the inverse piezoelectric effect, it is linearly related to the applied electric field (the same dependence as the linear electro-optic effect). This can lead to a frequency dependence of the 'effective' electro-optic coefficient for crystalline materials.

A brief comment on the photo-elastic effect (the change in index coefficients produced directly by applied stress) is warranted. This effect has the form:

$$\Delta(1/n^2)_i = \sum_l \pi_{il} \sigma_l \quad (14)$$

where σ_l are the components of the stress and π_{il} are the piezo-optical coefficients. Note that this effect is independent of the applied electric field.

For axially symmetric 'charge-transfer type' organic chromophores prepared by deposition or electric field poling, only two unique tensor elements, r_{33} and r_{13} , survive:

$$\tilde{r} = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{13} & 0 \\ r_{13} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (15)$$

The electro-optic effect for organic macromolecular materials derives from molecules with conjugated π -electrons confined within molecules (chromophores). The electro-optic tensor elements can be related to molecular first hyperpolarizability, β , and the chromophore number density, N (molecules/cm³), by

$$\begin{aligned} r_{33} &= -2Nf(0)\beta\langle\cos^3\theta\rangle/n_0e^4 \\ r_{13} &= -Nf(0)\beta\langle\sin^2\theta\cos\theta\rangle/n_0^2n_{0e}^2 \end{aligned} \quad (16)$$

The expressions in brackets are 'order parameters' describing the degree of noncentrosymmetric order. $\langle\cos^3\theta\rangle = 1$ corresponds to all 'dipolar' (charge transfer type) molecules pointing in the same direction (perfect noncentrosymmetric or ferroelectric order) while $\langle\cos^3\theta\rangle = 0$ corresponds to complete disorder and no electro-optic effect. The reason that three angles appear in the order parameter is that an angle is required to represent the principal symmetry axis for the chromophore, the principal symmetry axis of the optical field, and the principal symmetry axis of the electric control field. Thus, averaging (denoted by brackets) must be carried out over each of these three angles. The tensor element r_{33} corresponds to the electric control field applied along the principal axis of the ordered chromophores and the principal axis of the optical field vector, transverse magnetic (TM) polarized light. The tensor element r_{13} corresponds to the control electric field applied along the principal axis of the optical field, transverse electric (TE) polarized light, and orthogonal to the principal chromophore axis. The notation r_{33} is contracted notation for r_{333} . The factor $f(0)$ takes into account the dielectric nature of the medium into which the chromophores (molecules with large first hyperpolarizability) are embedded.

For organic electro-optic materials, elasto-optic effects do not appear to make significant contributions. Moreover, electro-optic activity can be systematically improved by improving β and by optimizing the product of chromophore number density and order parameter. Due to the presence of intermolecular electrostatic interactions, W , among chromophores, order parameters and number density are not independent, e.g., for materials that are prepared by electric field poling the order parameter relevant to the principal component of the electro-optic tensor can be expressed approximately as

$$\langle\cos^3\theta\rangle = L_3(\mu f(0)E_p/kT)[1 - L_1(W/kT)^2] \quad (17)$$

where L_3 , and L_1 are Langevin functions, μ is the chromophore dipole moment, E_p is the strength of the electric poling field, k is the Boltzmann constant, and T is the Kelvin poling temperature. W is a quadratic function of chromophore number density, N . The above equation indicates that a maximum will be observed in the graph of r_{33} versus N .

Materials

With inorganic and organic crystalline materials, very little can be done to optimize electro-optic activity other than discovering new EO crystalline materials or employing isotopic (e.g., deuterium for hydrogen) or ion substitution with existing materials.

As with organic liquid crystalline materials, the electro-optic activity of macromolecular organic second-order nonlinear materials can be systematically improved by molecular modification. As shown in Fig. 1, quantum mechanical calculations can provide useful guidance as to which structural modifications will lead to improvements in molecular (first) hyperpolarizability and ultimately electro-optic activity. The molecular hyperpolarizability (long wavelength limit) can be increased by a factor of two by simple variation of the acceptor structure. The calculated values of wavelengths for the interband charge transfer transitions are 390 nm, 403 nm, and 430 nm. Since dipole moments do not change significantly with structure, intermolecular electrostatic interactions should be similar for these three chromophores and the improvement in β should translate to an improvement in electro-optic activity. This theoretical prediction has been experimentally verified. This is just one example of use of quantum mechanics to guide the improvement of electro-optic activity for organic materials. This figure also illustrates the typical structure of an organic electro-optic chromophore, which consists of an electron donor region (an amine donor in the example shown), a

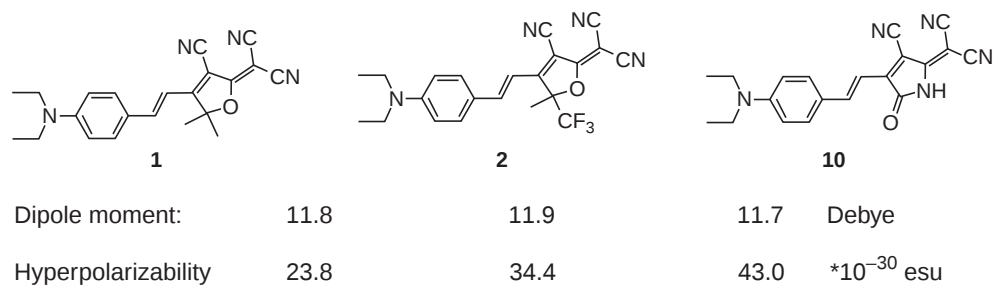


Fig. 1 Density functional calculations (DFT) of molecular first hyperpolarizability (β) and dipole moment (μ) for three chromophore structures.

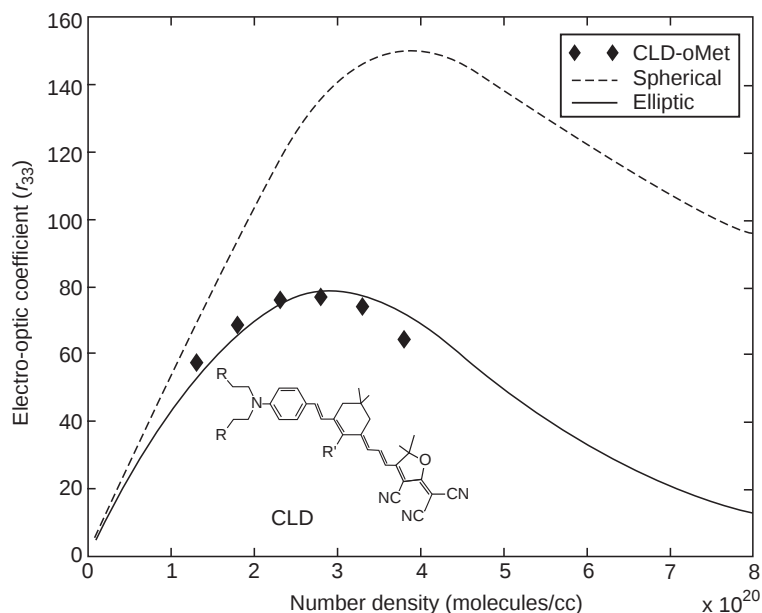
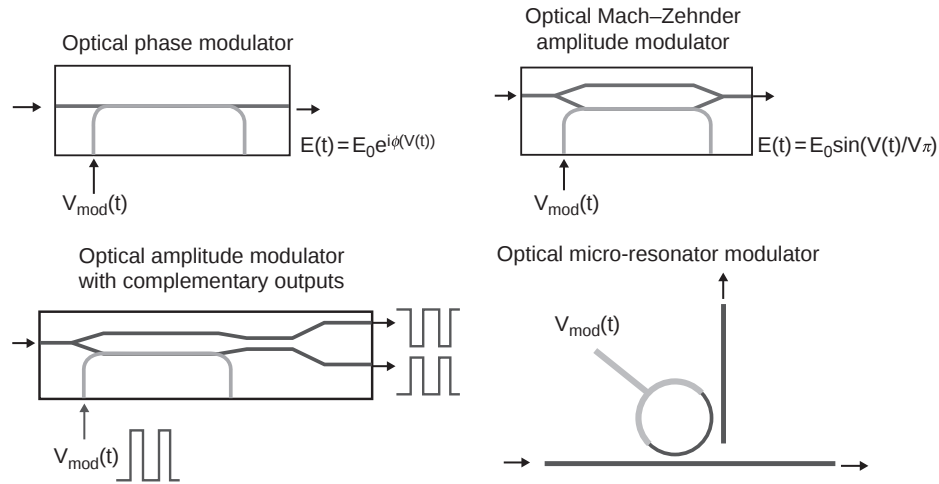


Fig. 2 Two theoretical (statistical mechanical) simulations of the variation of electro-optic activity with chromophore number density are shown and compared with experimentally observed results for the CLD chromophore (structure shown) dissolved in an amorphous polycarbonate host polymer.

connective conjugated π -electron bridge, and an electron deficient electron acceptor region. The theoretically predicted (by density functional theory, DFT, calculations) variation of molecular first hyperpolarizability with molecular structure has been verified for the structures shown. However, the critical point is that electro-optic activity of organic materials (currently exhibiting a maximum value of 182 pm/V at 1.3 microns wavelength) is likely to be improved for some time to come. This value is to be compared to a value of 32 pm/V for lithium niobate. Another factor influencing the value of macroscopic electro-optic activity realized for electrically poled organic materials is the effect of intermolecular electrostatic interactions on optimization of $N\langle\cos^3\theta\rangle$. Statistical mechanical calculations can provide guidance as to how this is achieved (see Fig. 2 where the role of chromophore shape is illustrated). Approximating the chromophore shape by a rigid prolate ellipsoid with embedded dipole moment yields nearly quantitative reproduction of the experimental data. If the π -electron structure (dipole moment) is left unchanged but the shape of the chromophore is altered to a spherical shape, a significant improvement in electro-optic activity is predicted. The shape effect arises because there are two components to the electronic dipole-dipole intermolecular electrostatic potential function describing the many body interaction of chromophores. One component favors centrosymmetric ordering while the second component favors noncentrosymmetric ordering. Their relative weighting in defining macroscopic order depends on chromophore shape. Theoretical calculations have also shown that embedding chromophores in nanoscopic macromolecular objects (such as dendrimers and dendronized polymers) can dramatically enhance electro-optic activity. This type of engineering can also be used to control auxiliary properties such as optical loss and stability. The exceptional processability of organic electro-optic materials has permitted conformal and flexible devices to be fabricated and has permitted electro-optic circuitry to be integrated vertically on top of semiconductor VLSI (very large-scale integration) electronics. Mach-Zehnder modulators, with 3 dB operational bandwidths of 200 GHz, have been fabricated from organic electro-optic materials as have Mach-Zehnder devices with operational drive voltages of less than one volt. The major advantage that crystalline inorganic materials possess, relative to macromolecular organic EO materials, is superior thermal and photostability. A comparison of lithium niobate and the best organic EO materials, is given

Table 1 Comparison of electro-optic performance for inorganic and organic EO materials at 1.3 microns wavelength

Parameter	<i>LiNbO₃</i>	<i>NLO polymer</i>
Effective EO coefficient, r_{eff} , pm/V	32	182
Index of refraction (n)	2.2	1.6–1.7
Dielectric permittivity (ϵ)	30	3
Material Figure of Merit, $n^3(r_{\text{eff}})/\epsilon$	10	270
Bandwidth $\times$ length product, GHz cm	7	> 100
$V_\pi L$, V cm	5	1.5
Optical loss, dB/cm	0.2	0.2–1.0
Processing temperature ($^\circ\text{C}$)	1000	< 200
Multiple layers possible	No	Yes

**Fig. 3** Schematic representation of simple electro-optic device structures are shown: Upper left (birefringent modulator), upper right (Mach-Zehnder interferometer), lower left (1×2 optical switch), lower right (ring microresonator).

in **Table 1**. Cost is a significant factor if electro-optic materials are to compete with modulated lasers and other opto-electronic devices for applications in communications, computing, transportation, and defense. Currently, manufacturing costs for both crystalline and macromolecular materials are very high (several thousand dollars per modulator). Recently, it has been demonstrated that organic macromolecular materials can be used with soft lithography techniques to mass-produce electro-optic circuitry. This could conceivably dramatically reduce manufacturing costs. Also, the fact that organic macromolecular materials can be directly integrated with silicon photonics and VLSI semiconductor electronics, may influence future commercial utilization of electro-optic materials, which is currently dominated by lithium niobate.

Devices

Representative simple device structures are shown in **Fig. 3**. For a simple stripline phase (birefringence) modulator (**Fig. 3**, upper left), the phase retardation produced by application of an electric field is given by

$$\Delta\phi; = 2\pi\Delta nL/\lambda = \pi n^3 r_{\text{eff}} EL/\lambda = \pi n^3 r_{\text{eff}} VL/\lambda d \quad (18)$$

where L is the electrode length (the distance over which the electrical and optical fields co-propagate), λ is the optical wavelength, $V = E/d$ is the electric field felt by the material, and d is the electrode spacing. Electro-optic device performance is frequently characterized by the voltage, V_π , required to produce a phase shift of π radians. For a TM polarized light propagating through a birefringent modulator:

$$V_\pi = \lambda d / (n_{0e}^3 r_{33} L \Gamma) \quad (19)$$

where Γ characterizes the overlap of the optical and electrical waves.

A simple Mach-Zehnder interferometer (electro-to-optical signal transducer) is shown in the upper right of **Fig. 3**. When such a device is constructed from electrically poled organic electro-optic chromophores and a voltage is applied to one arm, a retardation of light propagating in that arm will be effected. This, in turn, will result in a change in the constructive/destructive interference of

the beams at the output of the device. The consequence of applying an electrical signal to one arm of a Mach–Zehnder interferometer is that the applied voltage is transduced onto the optical output beam as an amplitude modulation of that beam. The voltage required to achieve full wave modulation for an optical beam of TM polarization is

$$V_{\pi} = \lambda d / (n_{0e}^3 r_{33} L \Gamma) \quad (20)$$

Amplitude modulation can also be achieved using a birefringent modulator by placing polarizers at the input and output of the device. If an input polarizer is used to launch TM and TE modes, the birefringent modulator functions to produce a rotation of light, which is turned into an amplitude modulation by inserting a polarizer at the output. For such a device, V_{π} depends upon the difference in r_{33} and r_{13} . Note that for poled organic chromophores, $r_{33} \approx 3r_{13}$, $(V_{\pi})_{\text{BRM}} \approx 1.5(V_{\pi})_{\text{MZM}}$. For the above devices, it is clear that V_{π} will depend on the length over which the electrical and optical fields co-propagate in phase. For low electrical field frequencies, L will be the electrode length. The electrode spacing d , that can be used, will depend on the index of refraction difference between the core (electro-optic material) and a cladding material used to prevent the optical field experiencing the metal electrodes (an event that would dramatically increase propagation loss). For poled organic macromolecular materials and standard commercial claddings, d is typically in the order of 10 microns. Electrode lengths usually range from 1 to 3 cm. Two factors limit bandwidth in stripline devices: (i) velocity mismatch between the electrical and optical waves (which relates to the difference between n_0^2 and ϵ , where ϵ is the dielectric permittivity of the material – see Table 1); and (ii) microwave and millimeter wave loss in the metal electrodes. For organic electro-optic materials, the latter defines operational 3 dB bandwidth of devices.

Optical routing switches (see Fig. 3, lower left) are more complicated to understand and typically require somewhat larger V_{π} voltages to effect a switching operation (e.g., $(V_{\pi})_{\text{simple2} \times 2 \text{ crossbar}} \approx 1.7(V_{\pi})_{\text{MZM}}$).

Two conditions must be satisfied for light to be coupled into a ring microresonator device (see Fig. 3, lower right): (i) the circumference of the ring must be a multiple of the wavelength of light, and (ii) the velocity of light in the ring must equal the velocity of light at the input. If a ring microresonator is fabricated from an electro-optic material, then an applied electrical voltage can be used to control coupling of light into and out of the ring microresonator. The same electrode structure, that is used to achieve voltage-controlled wavelength selective filtering, can also be used to transduce electrical information onto the optical beam while it is in the ring resonator. The quality (Q) factor of a ring microresonator defines both the wavelength selectivity and the bandwidth of the device. Ring microresonator devices also have the advantage of permitting a long interaction length to be achieved in a very compact device structure. The dimensions (radii) of ring microresonator structures are defined by bending loss and thus by the difference between core and cladding materials. Typical radii can range from hundreds of nanometers to hundreds of microns. With ring microresonators, a reduction in drive voltage requirements can be achieved by trading off bandwidth. Another price paid in the use of ring microresonator device structures is that of bending loss, which is not present in the other device structure shown. Ring microresonators can be used for active wavelength division multiplexing (WDM), voltage-controlled optical signal routing, and wavelength tuning of optical sources.

The interaction length of a prism optical beam deflector can be increased by cascading prisms together. For such a device structure, the angle of deflection becomes

$$\theta = n_0^3 r_{33} (V/d)(L/h) \quad (21)$$

where h is the height of an individual prism and L is the length of the prism array. For currently available electro-optic materials only small deflection angles (a few degrees or less) can be achieved with practical voltage levels. Superprism structures may permit large deflection angles to be achieved but such devices have yet to be demonstrated in a convincing manner.

Further Reading

- Butcher, P.N., Cotter, D., 1990. *The Elements of Nonlinear Optics*. New York: Cambridge University Press.
- Dagli, N., 1999. Wide-bandwidth lasers and modulators for RF photonics. *IEEE Trans. Microwave Theor. Techniques* 47, 1151–1171.
- Dalton, L.R., 2003. Rational design of organic electro-optic materials. *Journal of Physical Condensed Materials* 15, R897–R934.
- Dalton, L., Harper, A., Ren, A., *et al.*, 1999. Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fiber optics. *Ind. Eng. Chem. Res.* 38, 8–33.
- Hornak, L.A., 1992. *Polymers for Lightwave and Integrated Optics*. New York: Marcel Dekker.
- Khazaei, H.R., Wang, W.J., Berolo, E., Ghannouchi, F., 1998. High speed GaAs/AlGaAs traveling wave electro-optic modulators. *Proceedings of SPIE* 3278, 94–100.
- Lawetz, C., Cartledge, J.C., Rolland, C., Yu, J., 1997. Modulation characteristics of semiconductor mach-zehnder optical modulators. *IEEE Journal of Lightwave Technology* 15, 697–702.
- Lee, K.S., 2002. *Advances in Polymer Science*, vol. 158. Heidelberg: Springer.
- Lee, M., Katz, H.E., Erben, C., *et al.*, 2002. Broadband modulation of light using an electro-optic polymer. *Science* 298, 1401–1403.
- Nalwa, H.S., Miyata, S., 1997. *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton, FL: CRC Press.
- Noguchi, K., Mitomi, O., Miyazawa, H., 1998. Millimeter-wave Ti:LiNbO₃ optical modulators. *IEEE Journal of Lightwave Technology* 16, 615–619.
- Shen, Y.R., 1984. *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.
- Shi, Y., Zhang, C., Zhang, H., *et al.*, 2000. Low (sub-1 volt) halfwave voltage polymeric electro-optic modulators achieved by controlling chromophore shape. *Science* 288, 119–122.
- Wise, D.L., Wnek, G.E., Trantolo, D.J., Cooper, T.M., Gresser, J.D., 1998. *Electrical and Optical Polymer Systems. Fundamentals, Methods and Applications*. New York: Marcel Dekker.
- Yariv, A., Yeh, P., 1984. *Optical Waves in Crystals*. New York: John Wiley & Sons.

Polarization Introduction

JM Bennett, Michelson Laboratory, China Lake, CA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Polarization is one property of light waves and will be defined after a brief overview of the properties of light. Light is part of the electromagnetic spectrum of waves that have both particle-like and wave-like properties. Light waves carry energy in the form of photons which act like particles; the photon energy increases in proportion to the frequency of the wave. The particle-like properties of light and other electromagnetic waves are described by quantum mechanics. Light also acts like transverse waves that travel in straight lines in air and vacuum and can be described by classical electromagnetic theory. Polarization deals with the wave-like properties of light, and is described mathematically by Maxwell's equations, the key relations in electromagnetic theory. Polarization effects occur when light interacts with matter. In order to understand polarization, there will be a brief introduction to the subject of electromagnetic waves.

For a more in-depth treatment of the material in this article, the reader is directed to the Further Reading list at the end of this article and in particular to the two chapters on 'Polarization' and 'Polarizers' written by the author in Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chapter 5 and vol. II, chapter 3. New York: McGraw-Hill Inc.

Electromagnetic waves have both electric and magnetic fields associated with them. These are vibrations in directions perpendicular to the direction the wave is traveling, i.e., the direction of propagation. $\vec{E}$ represents the vector of the electric field and $\vec{H}$ perpendicular to $\vec{E}$ represents the magnetic field vector. Both these vectors are complex and have real and imaginary parts. Polarization effects are always associated with the $\vec{E}$ vector. Specifically, the plane of polarization is defined as the plane in which the $\vec{E}$ vector is vibrating. Waves having different amplitudes, phases, or angular orientations (azimuths) of their electric or magnetic vectors can be combined by conventional vector addition methods. Also the $\vec{E}$ vector of a particular vibration can be resolved into two components in mutually perpendicular directions that are vibrating in phase.

If a light source such as the sun, a candle flame, or an electric light bulb is considered on a microscopic scale, each vibrating atom or molecule emits linearly polarized light (see the definition below). But the individual atoms or molecules do not act together, so their vibrations have no fixed phase relationships to each other and they cannot be added into a single linearly polarized beam. Thus, we call light from these sources unpolarized. In an unpolarized light beam, the $\vec{E}$ vector vibrates in all directions perpendicular to the direction of propagation. If a snapshot is taken at a particular instant of time, different parts of the beam will have $\vec{E}$ vectors vibrating with different amplitudes and phases at different angles to each other, but all in a plane perpendicular to the direction of propagation. In the most common convention used in optics, the wave travels in the $+z$ direction in a right hand coordinate system and the $\vec{E}$ vectors are all vibrating at various angles in the x - y plane. The angle of vibration is measured from the positive x axis in a counterclockwise direction when the observer is looking against the direction of propagation of the light beam.

For linearly polarized or plane polarized light, if a snapshot of a light beam is taken at a particular instant, the $\vec{E}$ vector will be vibrating at a certain angle in the x - y plane. As time (or position on the traveling wave) varies, the amplitude of the $\vec{E}$ vector will vary in a sinusoidal manner, but the vibration will remain at the same angle in the x - y plane. As an example, the real part of the electric vector of a linearly polarized beam that is vibrating in the $+x$ and $-x$ directions and traveling in the $+z$ direction is given by the relation:

$$\vec{E}_x(z, t) = \hat{i}E_0 \cos(kz - \omega t) \quad (1)$$

where t is the time, $\hat{i}$ is a unit vector in the $+x$ direction, E_0 is the amplitude of the vibration, $\vec{k}$ is the propagation vector in the direction the wave is traveling, i.e., the z direction (its magnitude is $2\pi/\lambda$), and $\omega = 2\pi f$ where f is the vibration frequency of the wave. In free space, $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the wavelength of the vibration.

Light can also be circularly or elliptically polarized. Circularly polarized light is produced by adding the electric vectors of two waves, each having the same amplitude but which are 90° out of phase; the resulting vibration sweeps out a circle in the x - y plane. When the first wave is vibrating in the $+x$ and $-x$ directions (Eq. (1)), the second wave is vibrating in the $+y$ and $-y$ directions:

$$\vec{E}_y(z, t) = \hat{j}E_0 \sin(kz - \omega t) \quad (2)$$

and is added to the first wave. The resultant wave is the sum of Eqs. (1) and (2):

$$\vec{E} = E_0[\hat{i} \cos(kz - \omega t) + \hat{j} \sin(kz - \omega t)] \quad (3)$$

The direction of vibration of this wave will rotate in a circle as either the time increases or as the distance along the wave increases, but the amplitude of the vibration will not change. As time increases, the vibration will make a circle in the clockwise direction (to negative angles). This light is defined to be right circularly polarized. If the $+$ sign between the two parts of Eq. (3) is changed to a minus sign, as time increases, the vibration will make a circle in the counterclockwise direction (to positive angles) and the light is now defined to be left circularly polarized. These definitions are for conventional (traditional) optics. However, the opposite

definitions with right (left) circularly polarized light defining circles in the counterclockwise (clockwise) direction are also in use. In modern physics, there is still another convention that defines right (left) circularly polarized light as having negative (positive) helicity.

Right and left elliptically polarized light beams have the same angle conventions as for circularly polarized beams but are produced by adding two electric vectors that have different amplitudes. In Eqs. (1) and (2), the amplitude terms will be, for example, E_1 and E_2 instead of E_0 and Eq. (3) can no longer be used. There is now a major axis and a minor axis of the ellipse. If $E_1 > E_2$, the major axis will be along the x axis; for $E_1 < E_2$, the major axis will coincide with the y axis.

The preceding discussion has dealt entirely with the electric vector of the electromagnetic field. However, one cannot observe the $\vec{E}$ vector. The irradiance (energy per unit area per unit time), $\vec{E} \cdot \vec{E}^*$ is what can be observed visually and measured by electronic detectors. Thus, measurements of the polarization are irradiance measurements. Light transmitted by a polarizer is called the transmittance of the polarizer; similarly, light reflected from a polarizer is called the reflectance of the polarizer.

One of the most important parts of the subject of polarization is how to produce linearly polarized light starting with unpolarized light. This is done by using polarizers, as discussed in the section on polarizers below. Certain materials have special properties that make them able to polarize light. Depending on the application, different kinds of polarizers are preferred. Optics textbooks by Hecht and Guenther discuss the most important polarizers and two articles by Bennett describe many kinds of polarizers including special ones (see the Further Reading at the end of this article for full references). Only the basic principles will be described here.

Sunlight scattered by air molecules in the atmosphere (Rayleigh scattering) is also partially linearly polarized. The air molecules act like tiny dipoles and vibrate when they absorb sunlight. They emit radiation that is polarized in certain directions relative to the vibration direction. When the viewer is at a 90° angle with respect to the sun, the polarization of the skylight is a maximum. Rayleigh scattering is the subject of several books and will not be further discussed here.

Retarders are used to change linearly polarized light into circularly or elliptically polarized light and compensators, which are a form of retarders, can analyze an unknown type of polarized light and determine its composition. They are discussed below.

Polarimetry and ellipsometry are closely related techniques that are used to determine the optical properties of a material by shining a known type of polarized light on the material at non-normal incidence and analyzing the polarization properties of the light after it has been reflected from the material. These subjects are treated in other articles in this encyclopedia and in several references at the end of this article.

Changes can be produced in the optical properties of some materials by applying an electric field, a magnetic field, an acoustic field, or another form of variable pressure. The materials changed in these ways are said to be electro-optic, magneto-optic, or piezo-optic. The changes in the optical properties modify some parameter of a light wave passing through a material or reflecting from it. The amplitude, phase, frequency, polarization, or direction of the light wave can be modified. In this article, we are only concerned with modifications that produce changes in the polarization; these are discussed towards the end of this article.

Matrix methods have been developed to handle problems involving polarization and there is also a visual representation of the matrix algebra that is based on the Poincaré sphere. Both topics are covered at the end of this article.

Polarizers

Basic Relations for Polarizers

A linear polarizer is anything which, when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one direction with only a small component vibrating in the direction perpendicular to it. The transmittance T of the linear polarizer is

$$T = \frac{1}{2}(T_1 + T_2) \quad (4)$$

where T_1 , the principal transmittance of the polarizer, is $\gg T_2$, the transmittance of the polarizer at 90° to the principal transmittance. Thus, a perfect polarizer would transmit only 50% of an incident unpolarized beam.

If a linear polarizer is placed in a linearly polarized beam and is rotated about an axis parallel to the beam direction, the transmittance will vary between a maximum value T_1 and a minimum value T_2 according to the law:

$$T = (T_1 - T_2)\cos^2\theta + T_2 \quad (5)$$

where θ is the angle between the plane of vibration of the principal transmittance and the plane of vibration of the electric vector in the incident beam.

The extinction ratio ρ_P of a polarizer is defined as

$$\rho_P = \frac{T_2}{T_1} \quad (6)$$

and the degree of polarization of a polarization P is

$$P = \frac{T_1 - T_2}{T_1 + T_2} \quad (7)$$

When two identical polarizers are placed in an unpolarized beam, and the directions of their principal transmittances, T_A and T_B , are inclined at an angle θ to each other, the transmittance of the pair will be

$$T_\theta = \frac{1}{2} (T_1^2 + T_2^2) \cos^2 \theta + T_1 T_2 \sin^2 \theta \quad (8)$$

Thus, when the directions of the principal transmittances are aligned, $T_\parallel = \frac{1}{2} (T_1^2 + T_2^2)$, and when they are perpendicular, $T_\perp = T_1 T_2$.

Birefringent Materials (Calcite)

The majority of high-quality polarizers are made from calcite. This is a birefringent (doubly refracting) crystalline material that is uniaxial (i.e., there is one preferred direction in the crystal). A birefringent material acts differently for light going in different directions through the crystal. For example, if an unpolarized light beam passes through the crystal in a certain direction, it will be split into two spatially separated beams that are parallel but are linearly polarized at right angles to each other. A uniaxial crystal has an optic axis (i.e., a certain direction in the crystal). When light rays travel parallel to the optic axis, they travel at the same velocity and there is no difference between them. When light passes through the crystal in other directions, the ray whose vibration direction is perpendicular to the optic axis is governed by the ordinary laws of geometrical optics (the same as for isotropic materials) and is called the ordinary ray. The ray whose vibration direction is parallel to the optic axis does not follow the normal geometrical optics laws and is called the extraordinary ray. One ray travels faster than the other, so there is a phase retardation for one ray relative to the other. This is the principle of a retarder or retardation plate (see the next section).

Calcite is a negative uniaxial crystal which means that the refractive index for the ordinary ray is larger than the refractive index for the extraordinary ray. When the ordinary ray enters a block of calcite from air at non-normal incidence, it is bent more than the extraordinary ray.

Calcite can be easily cleaved along three distinct planes, making it possible to produce rhombs of the form shown in Fig. 1. The optic axis, going in the HI direction, makes equal angles with all three faces at point H. Any plane, such as DBFH, which contains the optic axis and is perpendicular to the two opposite faces of the rhomb ABCD and EFGH is called a principal section. If the plane of incidence of light on the rhomb coincides with a principal section, the light entering the crystal will be split into two components polarized at right angles to each other which travel in slightly different directions and leave the crystal as two beams slightly displaced but parallel to each other.

The large birefringence of calcite and its excellent transmission through the visible spectral region and into the ultraviolet and infrared regions has made it possible to make excellent high extinction ratio polarizers with different designs. Some of these are shown in Fig. 2. There are two main types: Glan types with rectangular shapes and Nicol types with rhombohedral shapes. The polarizers are made of two pieces of calcite cemented together. Glan types have their optic axes in the plane of the entrance face. In the Nicol types, the principal section is perpendicular to the entrance face, but the optic axis is neither parallel nor perpendicular to the face. The two halves of conventional polarizing prisms are cemented together with cement that has a refractive index intermediate between the ordinary and extraordinary refractive indices. This enables one ray (generally the extraordinary, or e ray) to be transmitted and the other to be reflected at the cut, so that only one ray exits from the prism in the direction of the

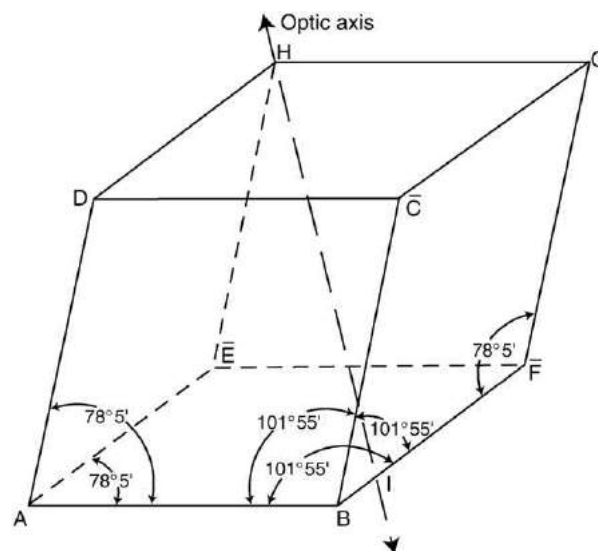


Fig. 1 Schematic representation of a rhombohedral calcite crystal showing the angles between faces. The optic axis passes through corner H and point I on side BF. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

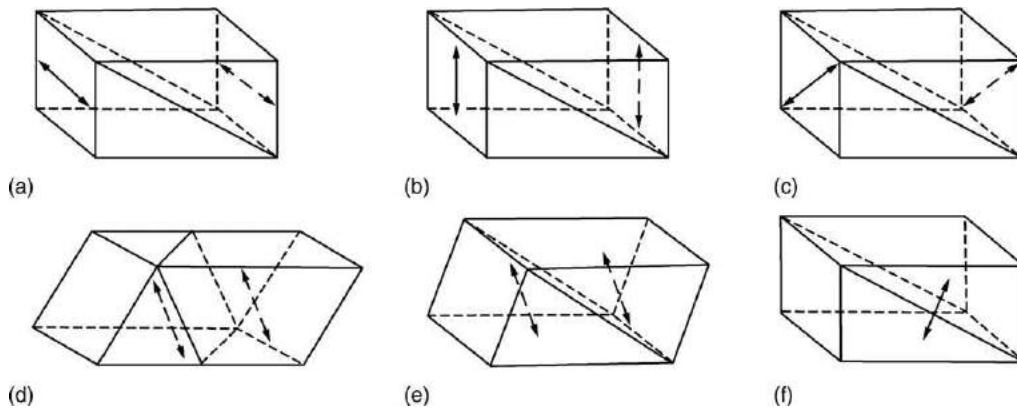


Fig. 2 Types of polarizing prisms. Glan types: (a) Glan-Thompson, (b) Lippich, and (c) Frank-Ritter; Nicol types: (d) conventional Nicol, (e) Nicol-Halle form, and (f) Hartnack-Prazmowsky. The optic axes are indicated by the double-pointed arrows. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

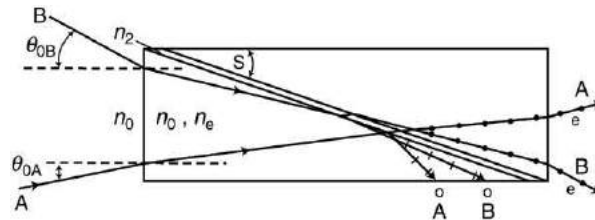


Fig. 3 Extreme rays (A and B) that can pass through a cemented Glan-Thompson prism. Both rays change to ordinary and extraordinary rays in the calcite. Both ordinary rays are reflected at the boundary between the two halves of the prism and the extraordinary rays exit in the directions indicated. Rays entering the prism between rays A and B would be transmitted by the prism. The field angle of the prism is twice the smaller of the two angles of incidence (ray A).

incoming ray. The Glan types are also used without cement with only an air space between the two halves. They can be used at shorter wavelengths in the ultraviolet where the cement absorbs. The extinction ratio can be very high for air-spaced prisms. For example, for a Glan-Foucault prism (an air-spaced Glan-Thompson prism, Fig. 2), the extinction ratio can be better than 1×10^{-5} and the prism is usable from about $0.214 \mu\text{m}$ in the ultraviolet to $2.3 \mu\text{m}$ in the infrared. However, air-spaced polarizers have very small field angles, so they are mainly used with laser sources where the beam is parallel.

The extreme paths of light through a cemented Glan-Thompson prism are shown in Fig. 3. This prism has a length that is three times the width of the entrance aperture (i.e., an L/A ratio of 3). The optic axis is perpendicular to the plane of incidence which is in the plane of the paper. In the first half of the polarizer, the paths of the ordinary and extraordinary rays, both of which follow the conventional law of refraction (Snell's Law, Eq. (8) above) nearly coincide. Ray A is incident on the entrance face of the polarizer at an angle such that the angle of incidence on the cut is the smallest angle for which the o ray is totally internally reflected. Ray B is incident at an angle such that the angle of refraction in the first half of the prism is essentially equal to the cut angle, S , so that the e ray just passes through the cut. The field angle of the polarizer is twice the smaller of angles θ_A and θ_B . Thus, all rays having angles of incidence between rays A and B will be polarized when exiting the prism. The paths of similar rays can be traced through the other prism designs.

In addition to conventional polarizing prisms, there are also polarizing beamsplitter prisms and Feussner prisms. In polarizing beamsplitter prisms, the two beams that are polarized at right angles to each other both emerge from the prism but are separated spatially. Fig. 4 shows a diagram of several types of polarizing beamsplitter prisms and Fig. 5 shows the paths of rays through side views of these prisms.

A Feussner prism is made of isotropic material but the film separating the two halves is birefringent. These prisms have the advantage that much less birefringent material is required but they have a more limited wavelength range when calcite or sodium nitrite (another birefringent material) is used because the transmitted ordinary ray has a shorter transmission range than the extraordinary ray which is transmitted by the conventional and air-spaced polarizing prisms.

Dichroic Absorbers

A dichroic material is one that absorbs light polarized in one direction more strongly than light polarized at right angles to that direction. Dichroic materials are different from birefringent materials because the latter usually have negligible absorption coefficients for both the ordinary and extraordinary rays. Stretched polyvinyl alcohol sheets treated with absorbing dyes or

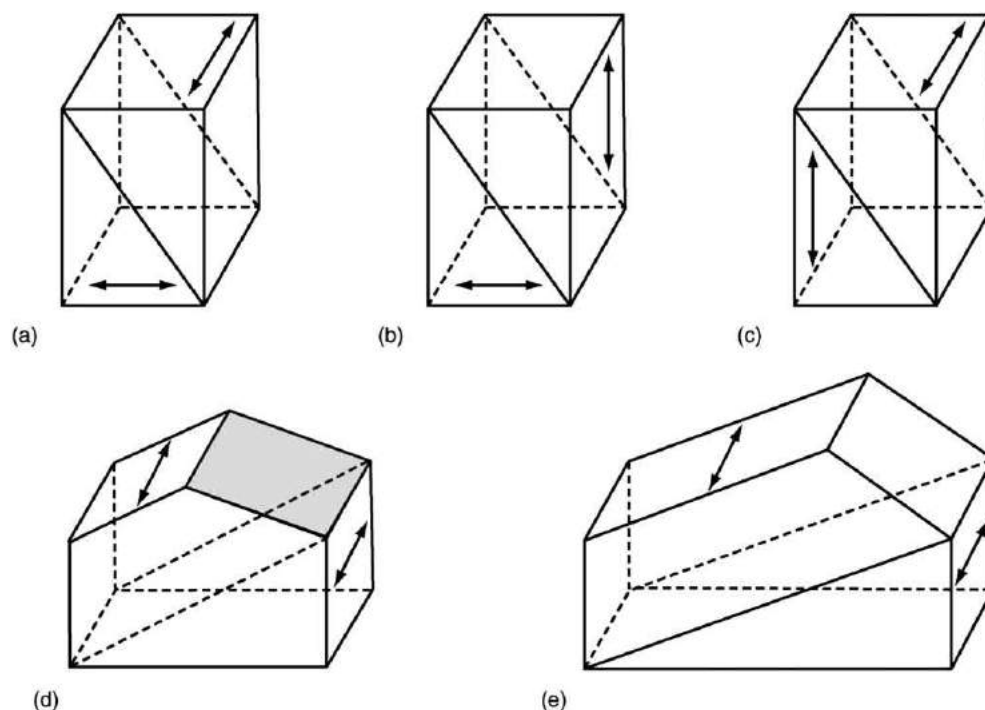


Fig. 4 Three-dimensional views of various types of polarizing beamsplitter prisms: (a) Rochon; (b) Sénarmont; (c) Wollaston; (d) Foster (shaded face is silvered); and (e) beamsplitting Glan-Thompson. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

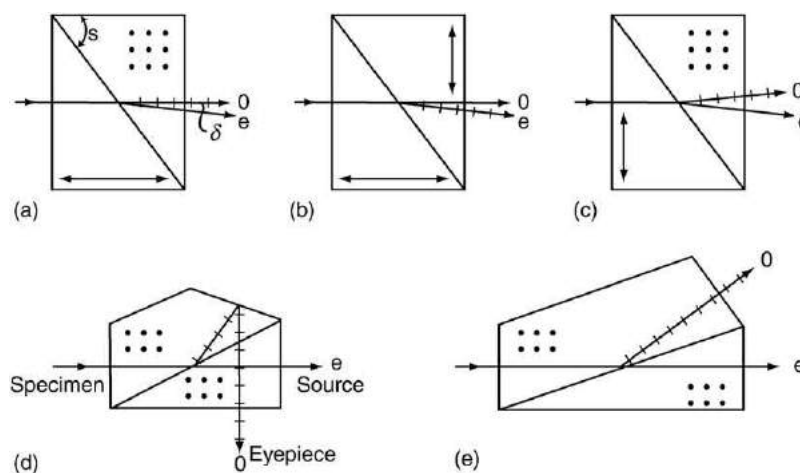


Fig. 5 Side views of the polarizing beamsplitter prisms in Fig. 4. The directions of the optic axes are indicated by the dots and the heavy double-pointed arrows. When the Foster prism is used as a microscope illuminator, the source, specimen, and eyepiece are in the positions indicated. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

polymeric iodine are the most common type of dichroic absorbers and are primarily sold under the tradename Polaroid. These polarizers do not have as good an extinction ratio as the calcite prism polarizers (see above), but they are inexpensive, come in large sizes, are easily rotated, and produce negligible beam deviation. Also, they are thin, lightweight, and can be made in any desired shape. One of their main advantages is that they are insensitive to the degree of collimation of the beam and can be used in strongly convergent or divergent light. Depending on the density and type of absorbing dye used to make the polarizer, the transmission can be maximized for the visible or near infrared spectrum. The extinction ratio of the Polaroid HN-22 sheet compares favorably with that of the Glan-Thompson prisms throughout the visible spectral region, but the transmission of the Glan-Thompson prism is superior. As the dichroic polarizer transmission increases, the extinction ratio becomes worse.

Transmission and extinction ratio curves for various types of dichroic polarizers are shown in Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

A wire grid polarizer is another kind of dichroic absorber. It is made of a series of equally spaced conducting bars or wires that are either free standing or deposited onto a transparent substrate (backing plate). Energy that is polarized parallel to the length of the bars is absorbed out of the incoming wave by inducing oscillations in the electrons in the metal. Thus, only light polarized perpendicular to the bars will be transmitted. In order to have an appreciable degree of polarization, the wavelength must be at least twice the spacing between grids in the polarizer. Because of the difficulty of making wire grids with small enough spacings, these polarizers are limited to the mid- and long-wavelength infrared spectral region, beyond about 5 μm . The polarizer will have the best extinction ratio if the substrate has a low refractive index; if a high refractive index substrate such as silicon or germanium is used, it must be covered with an antireflection coating before the grid is deposited.

Most of the wire grid polarizers have been used with microwaves, so the theory is in the form for transmission lines. Similar transmission line theory has been applied to infrared polarizers in the form of bars and wires.

Reflection and Transmission

Light can be polarized by reflecting it from the flat surface of a material inclined at non-normal incidence to the light beam or by transmitting it through a transparent plate at non-normal incidence. These polarizers work because light has different reflectances when the electric vector is linearly polarized parallel and perpendicular to the plane of incidence. The plane of incidence contains both the incoming light beam (incident beam) and the reflected light beam. The angles for both the incident and reflected light beams are measured relative to an axis perpendicular to the surface (the surface normal). The reflection coefficients are given by the Fresnel equations and, for nonabsorbing materials, are:

$$R_s = r_s^2 = \frac{\vec{E}_s^2(\text{reflected})}{\vec{E}_s^2(\text{incident})} = \frac{\sin^2(\theta_0 - \theta_1)}{\sin^2(\theta_0 + \theta_1)} \quad (9)$$

$$R_p = r_p^2 = \frac{\vec{E}_p^2(\text{reflected})}{\vec{E}_p^2(\text{incident})} = \frac{\tan^2(\theta_0 - \theta_1)}{\tan^2(\theta_0 + \theta_1)} \quad (10)$$

where R_s and R_p are called the reflectances (previously called the intensity reflection coefficients), and r_s and r_p are the amplitude reflection coefficients. $\vec{E}_s$ ($\vec{E}_p$) is the incident or reflected electromagnetic wave vibrating perpendicular (parallel) to the plane of incidence and θ_0 and θ_1 are the angles of incidence and refraction, respectively. The angle of refraction is the angle of the light beam in the material, measured from the surface normal. For a nonabsorbing material, it can be obtained in terms of the refractive index n_1 of the material from Snell's Law:

$$\frac{\sin \theta_0}{\sin \theta_1} = \frac{n_1}{n_0} \quad (11)$$

where n_0 and n_1 are the refractive indexes of the incident medium (usually air with $n_0=1$) and the material. Sometimes the refractive index of the material is expressed as a ratio measured relative to the refractive index of the incident medium: $n=n_1/n_0$. The reflectances of absorbing materials are similar to Eqs. (9)–(11) but involve complex refractive indexes and angles of refraction.

Fig. 6 shows curves for R_s and R_p as a function of angle of incidence for four nonabsorbing materials that have different refractive indexes. In all cases the reflectance is higher for the s component than for the p component except at 0° and 90° angles of incidence where they are the same. At the so-called Brewster angle, θ_B , $R_p=0$, $\tan \theta_B=n_1/n_0$, and incident unpolarized light is now completely linearly polarized perpendicular to the plane of incidence (s -polarized light). Note that high refractive index materials produce more intense polarized beams (i.e., have higher reflectances) than low refractive index materials.

Fig. 7 shows the reflectances and extinction ratios for materials having different refractive indexes. Each curve is for a single reflection. Since plates have two surfaces, there will be two reflections from an actual reflection polarizer so the reflectance and extinction ratios will increase. If the plates are thick, the reflection from the back surface of the plate will be displaced from the reflection from the front surface of the plate. A reflection polarizer is normally used close to the Brewster angle because a high degree of polarization (i.e., a very small extinction ratio) is desired. Because the extinction ratio changes rapidly for angles of incidence close to the Brewster angle, the incident beam must be well collimated to obtain a high degree of polarization. A main disadvantage of reflection polarizers is that the reflected beam is no longer parallel to the incident beam and two additional reflections are required to align the beam.

Light transmitted through a plate will only be partially linearly polarized at any angle of incidence including the Brewster angle because both s - and p -components are partially transmitted. Transmission polarizers are thus made of several plates to increase the degree of polarization. Fig. 8 shows the transmittance and extinction ratio for a stack of four plane parallel plates assuming multiple incoherent reflections within each plate and no reflections between plates. At the Brewster angle the p -transmittance is theoretically 1 (assuming that there is no absorption within the material) but the extinction ratio greatly depends on the refractive index of the material. A stack of low refractive index plates ($n=1.5$) has a poor extinction ratio, while a pile of high refractive index plates ($n=4.0$) has a much better extinction ratio. The plates are often inclined at small angles to each other so the light multiply reflected between plates (which decreases the polarization) is removed from the transmitted beam. The sides of each plate are plane parallel to increase the polarization, but the plates are too thick for the amplitudes of the multiple internally reflected beams

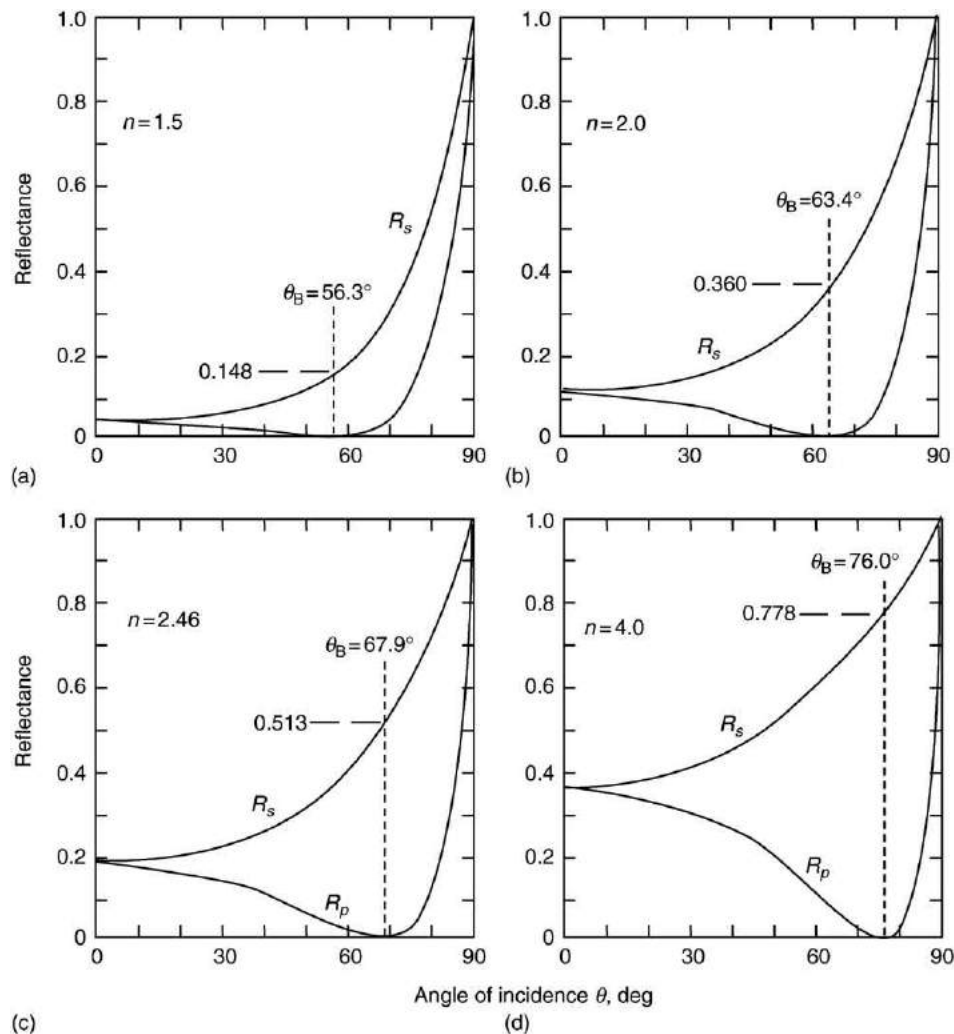


Fig. 6 Reflectance of light polarized parallel R_p and perpendicular R_s to the plane of incidence from materials of different refractive index n as a function of angle of incidence: (a) $n=1.5$ (alkali halides in the ultraviolet, glass (approximately) in the visible, and sheet plastics in the infrared), (b) $n=2.0$ (AgCl in the infrared), (c) $n=2.46$ (Se in the infrared), and (d) $n=4.0$ (Ge in the infrared). The Brewster angle θ_B (at which R_p goes to 0) and the magnitude of R_s at θ_B are also indicated. Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

to add or subtract. If the amplitudes of the beams could be added, there would be interference effects and the transmittance would vary with the thickness of each plate and with the wavelength. These are so-called interference polarizers and are mentioned below.

Transmission polarizers do not have the angle deviation problem of the reflection polarizers, but the transmitted beam may be slightly displaced parallel to the incident beam if the plates are thick. The main problem with transmission polarizers is that the light is not completely polarized even with several plates in a stack. If the plates are slightly absorbing, the transmission of the polarizer is reduced.

Miscellaneous Types

There are numerous other types of polarizers that mostly use thin films. Interference effects in thin films can increase the polarization efficiency at certain wavelengths depending on the thicknesses of the films and the design of the multilayer coating. Many of the applications involve non-normal incidence designs. For example, a single high refractive index film or a multilayer coating evaporated onto a low refractive index substrate increases the degree of polarization of reflection and transmission polarizers. Polarizing beamsplitters also use multilayer dielectric coatings. Free-standing films of different thicknesses were formerly used for infrared polarizers before wire grid polarizers became available.

Dichroic absorbers have been made from two-phase lamellar eutectics of thin needles of a conducting material embedded in a transparent matrix. Thin sheets of pyrolytic graphite material also act as dichroic absorbers. There are also numerous miniature polarizer designs for fiber optics and nanotechnology applications.

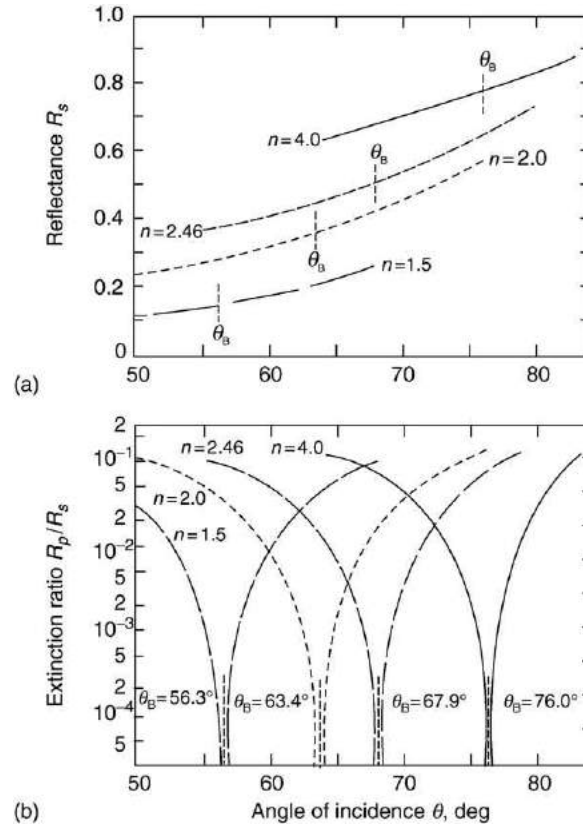


Fig. 7 (a) Reflectance R_s and (b) extinction ratio R_p/R_s for materials of different refractive index at angles near the Brewster angle θ_B . A single surface of the material is assumed. Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

Retarders or Retardation Plates

Retarders, or retardation plates, are devices that can change linearly polarized light into circularly or elliptically polarized light. They can also rotate the plane of polarization of linearly polarized light. A retardation plate is made from a uniaxial crystal that has an optic axis (as discussed above); the light travels in a direction perpendicular to the optic axis, as described below and shown in Fig. 9. The ordinary and extraordinary rays travel at different velocities through the crystal depending on their refractive indexes: $n=c/v$, where c is the velocity of light in a vacuum and v is the velocity of light in the material. The larger the refractive index, the slower is the velocity of light in the material. The ordinary ray with a refractive index n_o and the extraordinary ray with a refractive index n_e have mutually perpendicular vibration directions. If the ray has a vibration direction at another azimuth in the crystal, the refractive index is intermediate between n_o and n_e . For a positive uniaxial crystal $n_e > n_o$ so the extraordinary ray travels slower than the ordinary ray through the crystal. Thus, the designation ‘fast axis’ is often used for the ordinary ray and ‘slow axis’ is used for the extraordinary ray, as shown in Fig. 9. In a negative uniaxial crystal, $n_e < n_o$, i.e., the velocities along the two axes are reversed.

Because there is a velocity difference between the ordinary and extraordinary rays traveling through a uniaxial crystal, they get out of phase. Depending on their velocities and the optical thickness of the material (the refractive index times the physical thickness), the addition of their amplitudes results in a wave whose vibration direction is either: (a) perpendicular to the original vibration direction; (b) rotates in a circle (right or left circularly polarized light); or (c) rotates in an ellipse (right or left elliptically polarized light). The path difference N between two rays in the crystal, measured in terms of the wavelength of the light, is $N\lambda = \pm d(n_e - n_o)$, where d is the physical thickness of the material. The corresponding phase difference δ between the two rays is $\delta = 2\pi N = \pm (2\pi d/\lambda)(n_e - n_o)$. The quantity N is important because if beams of light vibrating along the fast and slow axes of a crystal get out of step by one quarter of the wavelength of the light, the device is called a quarter wave (or $\lambda/4$) retardation plate, or simply a quarter-wave plate. There are also half-wave ($\lambda/2$) plates that rotate the plane of polarization, and full-wave (λ) plates that rotate the plane of polarization through 180° . If light passes through a full-wave plate, theoretically it is indistinguishable from the original beam. However, materials are normally temperature sensitive so that as the temperature changes, the retardation is no longer exactly one wave, but may be slightly less than or greater than one wavelength. Other thicknesses of retardation plates produce elliptically polarized light.

Fig. 10 shows what happens to a beam of linearly polarized light when the axis of vibration is at 45° to the fast and slow axes of a positive uniaxial crystal. The fast axis is in the horizontal direction, the same as in Fig. 9.

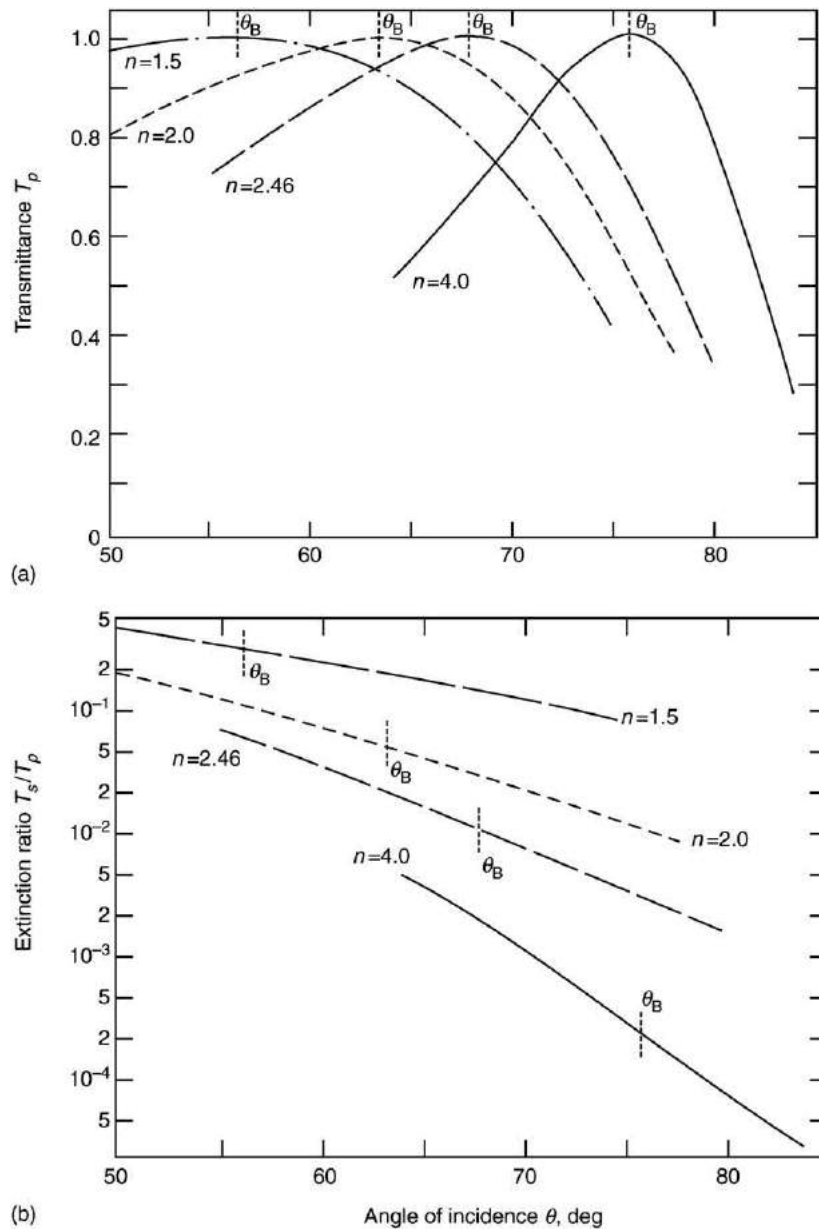


Fig. 8 (a) Transmittance and (b) extinction ratio of four plane-parallel plates of refractive index n as a function of angle of incidence, for angles near the Brewster angle. Assumptions are multiple reflections but no interference within each plate and no reflections between plates. Reproduced with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

The main purpose of a quarter-wave plate is to change linearly polarized light into circularly polarized light. However, as Fig. 11 shows, the state of polarization of the light will be quite different depending on the orientation of the plane of polarization of the incident beam relative to the fast axis of the quarter-wave plate.

A half-wave plate is used to rotate the plane of polarization of a linearly polarized beam. The plane of polarization is always rotated through an angle that is twice the angle the initial plane of polarization makes with the fast axis of the uniaxial crystal. A linearly polarized beam always remains linearly polarized.

Retardation plates are often made of mica, stretched polyvinyl alcohol, or crystal quartz, although they can also be made of other stretched plastics, sapphire, magnesium fluoride, and other materials.

Variable Retardation Plates and Compensators

Variable retardation plates are devices whose retardation can be varied in a variety of ways. They can be used to modulate or vary the phase of a beam of linearly polarized light or to analyze a beam of unknown polarization (often elliptically polarized light)

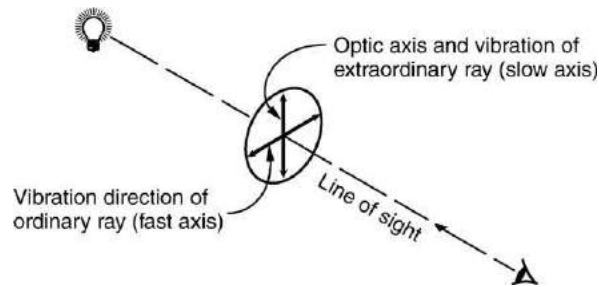


Fig. 9 Light incident normally on the front surface of a retardation plate showing the vibration directions of the ordinary and extraordinary rays. In a positive uniaxial crystal, the fast and slow axes are as indicated in parentheses; in a negative uniaxial crystal, the two axes are interchanged. Adapted with permission from Bennett JM (1995) *Polarization*. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

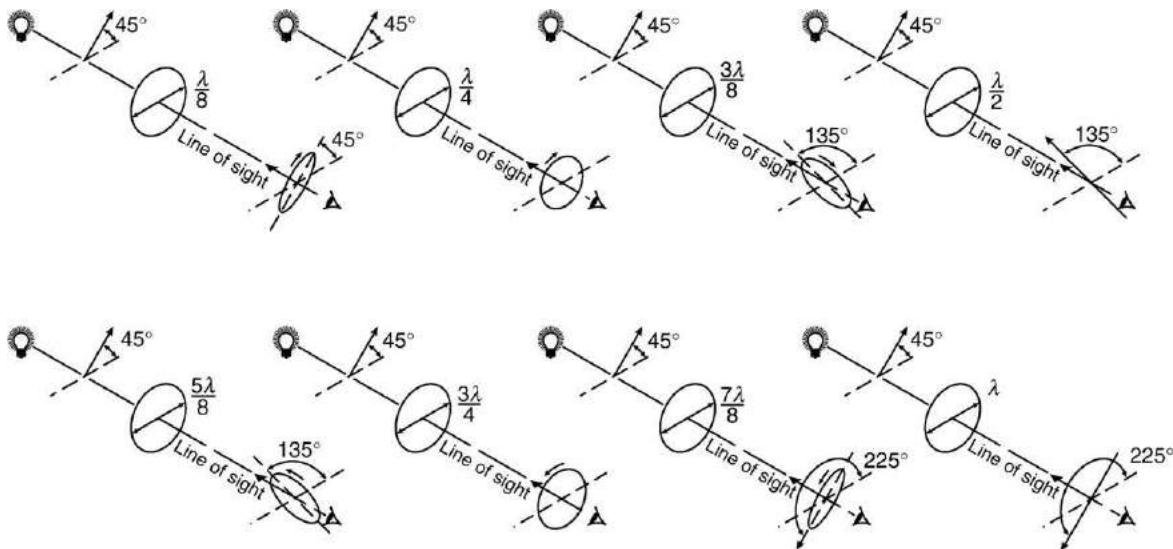


Fig. 10 State of polarization of a light wave after passing through a crystal plate whose retardation is indicated in fractions of a wavelength (phase retardation $2\pi/\lambda$ times these values) and whose fast axis is indicated by the double arrow. In all cases the incident light is linearly polarized at an azimuth of 45° to the direction of the fast axis. Adapted with permission from Bennett JM (1995) *Polarization*. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

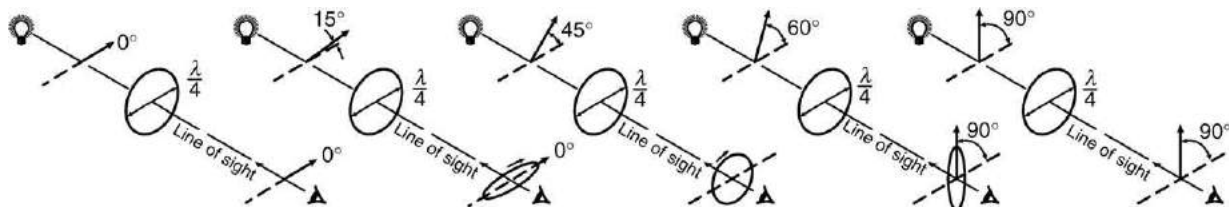


Fig. 11 State of polarization of a light wave after passing through a $\lambda/4$ plate (whose fast axis is indicated by the double arrow) for different azimuths of the incident linearly polarized beam. Adapted with permission from Bennett JM (1995) *Polarization*. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

such as might be produced by transmission through a birefringent material or by reflection from a metal or film-covered surface. The term compensator is often applied to a variable retardation plate since it can be used to compensate for the phase retardation produced by a material. Common types of compensators include the Babinet and Soleil compensators, in which the total thickness of a birefringent material in the light path is changed, the Sénarmont compensator which consists of a fixed quarter-wave plate and rotatable analyzer to compensate for varying amounts of ellipticity in a light beam, and tilting-plate compensators, which change the thickness of birefringent material in the light beam by changing the angle of incidence. Electro-optic and piezo-optic modulators can also be used as high-frequency variable retardation plates since their birefringence can be changed by varying the electric field or pressure (see next section).

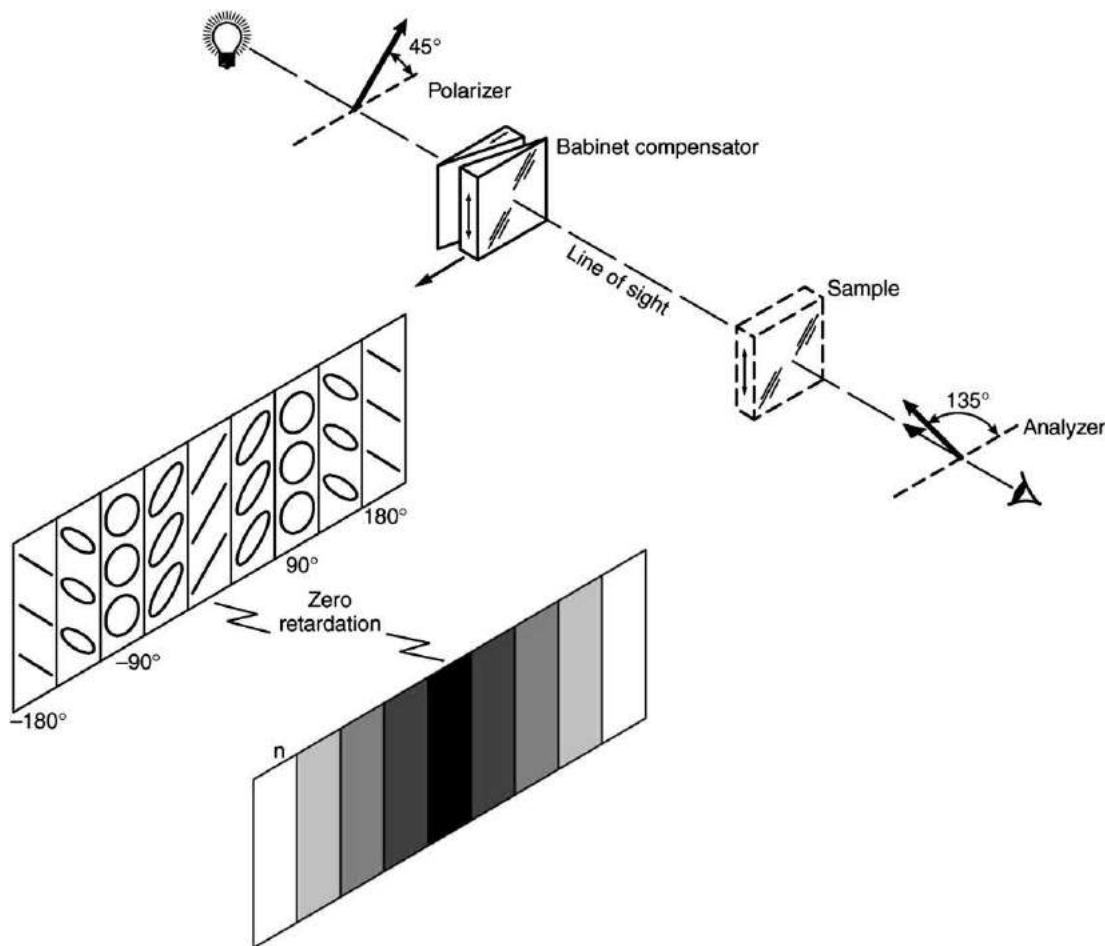


Fig. 12 Arrangement of a Babinet compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. Retardations are indicated for alternate regions. After the beam passes through the analyzer, the field is crossed by a series of dark bands, one of which is shown to the left of the analyzer. Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

The Babinet compensator, shown schematically in Fig. 12 consists of two crystal quartz wedges, each with its axis in the plane of the face but with the axes at right angles to each other. One wedge is stationary, and the other can be moved in the direction indicated by the arrow, so that the total amount of quartz through which the light passes can be varied. The total retardation is proportional to the difference in thickness between the two wedges. This type of compensator was used extensively when the light source was a vertical slit and visual measurements were made of the state of polarization of a beam. However the bands representing different phase retardations were too narrow to be used effectively with a photoelectric detector, so the Babinet compensator was replaced by a Soleil compensator (Fig. 13). This device, sometimes called a Babinet–Soleil compensator, is similar to the Babinet compensator except that the field of view has a uniform tint if the compensator is constructed correctly. This is because the ratio of the thicknesses of the two crystal quartz blocks (a movable wedge and a fixed wedge attached to a plate with the two axes perpendicular to each other) is the same over the entire field. The Soleil compensator will produce light of varying ellipticity depending on the position of the movable wedge. It is used in the same way as a Babinet compensator with the uniformly dark field of the Soleil compensator corresponding to the black zero-retardation band in the Babinet compensator.

It is sometimes necessary to accurately measure the azimuth of a beam of linearly polarized light using a photoelectric detector with a rotatable analyzer (i.e., a polarizer) directly in front of it. The obvious method is to rotate the analyzer until the detector signal is a minimum and then read the analyzer angle, which equals the extinction angle (perpendicular to the azimuth of the linearly polarized beam). However, the angle can be determined more precisely if the analyzer is offset by a small angle from the extinction angle and the transmittance noted. Then the analyzer is offset by a small angle on the other side of the extinction angle at the angle where the transmittance is the same. One half the difference between these two azimuthal angles gives a more accurate value of the extinction angle than can be obtained by measuring it directly.

Before the days of photoelectric detectors, half-shade devices were extensively used to determine the azimuth of a linearly polarized beam. The device consisted of two polarizers with their axes inclined at a small angle to each other. As the device was

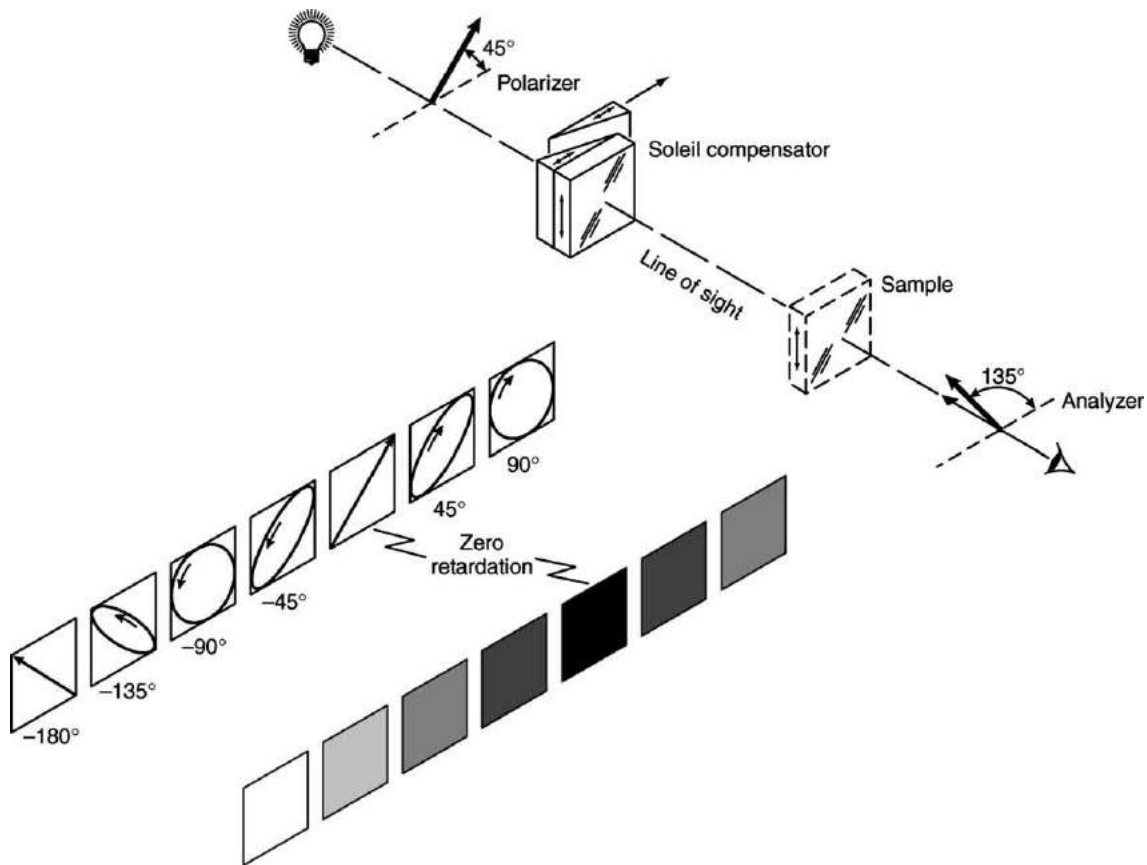


Fig. 13 Arrangement of a Soleil compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. After the beam passes through the analyzer, the field appears as one of the shades of gray shown to the left of the analyzer. Adapted with permission from Bennett JM (1995) *Polarizers*. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

rotated, one part of the field became darker and the other part became lighter. At the match position, both parts of the field appeared equally bright. There were a variety of these devices as well as ellipticity half-shade devices that could detect very small amounts of ellipticity in a nominally linearly polarized beam and hence could verify when a compensator had completely converted elliptically polarized light into linearly polarized light.

Electro-Optic, Magneto-Optic, and Piezo-Optic Devices

The state of polarization of light can be rapidly altered by passing it through a material that has electro-optic, magneto-optic or piezo-optic properties. The voltage, magnetic field, or pressure are varied to make the material birefringent (see above). Some materials are isotropic without the applied field, i.e., the refractive index is the same in all directions. Other materials have an optic axis (uniaxial materials) or two optic axes (biaxial materials). In these cases, the applied field creates further asymmetries in the material.

The most common electro-optic, magneto-optic and piezo-optic effects are the Pockel's effect (depending on the electric field), the Kerr effect (depending on the square of the electric field), the Faraday effect (depending on the magnetic field), the Cotton-Mouton and Voigt effects (depending on the square of the magnetic field), and stress birefringence or the photoelastic effect (depending on pressure changes). These effects are shown in [Table 1](#) along with order of magnitude strengths needed to produce changes in the refractive index, and materials that most strongly exhibit the effects. The mathematical descriptions of the effects involve tensors and are too involved to present here. However, simple physical descriptions will be given.

If a varying electric field acts on an electro-optic material, electrons, ions, or permanent dipoles in the material are made to reorient, inducing an electric polarization. The induced polarization creates birefringence that modifies the optical polarization of a light beam passing through the material. The electric field strength determines how the polarization is changed (see [Fig. 10](#)).

Linearly polarized light passing through a magneto-optic material will be rotated in a direction parallel to the direction of the applied magnetic field. This phenomenon, called the Faraday effect, or Faraday rotation, is similar to what happens in optically

Table 1 Electro-optic, magneto-optic, and piezo-optic effects

<i>Effect</i>	<i>Device</i>	<i>Effect proportional to</i>	<i>Strength (Δn)</i>	<i>Materials</i>
Pockels (electro-optic)	Pockels cell	E	$10^{-12}E$	BaTiO ₃ , LiNbO ₃
Kerr	Kerr cell	E^2	$10^{-19}E^2$	Nitrobenzene, Benzonitrile
Faraday (magneto-optic)	Faraday rotation	H	$10^{-14}H$	Zns, GaAs, CS ₂
Cotton–Mouton	^a	H^2	$10^{-25}H^2$	Chloroform, acetone
Photoelastic effect; stress birefringence	^b	Pressure	–	Fused silica, polystyrene, Ge, KDP, ruby

^aToo small to be of technological importance.

^bDepends on the mounting or the processing of the material. Adapted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

active materials (see the next paragraph) except that the Faraday effect depends on the direction of the magnetic field, but not on the direction the light is traveling.

There are naturally optically active materials such as camphor, nicotine, sugar solutions, and crystal quartz that can rotate the plane of linearly polarized light passing through the material. For example, the Si–O bonds in crystal quartz form a helical path around the optic axis. This crystal structure interacts with an incoming linearly polarized beam traveling parallel to the optic axis and rotates it in a clockwise direction. A 1 mm-thick piece of quartz will rotate a linearly polarized beam by 21.7°. Other materials have asymmetric structures that have mirror images. A mirror image structure cannot be obtained by simply rotating the group of atoms in space. These materials are also optically active. The amount of rotation produced by an optically active material is proportional to the thickness (optical density) of the material that the light passes through. Dextrose (a form of glucose) and levulose (also known as fructose) rotate the plane of linearly polarized light in clockwise and counterclockwise directions, respectively. Many other organic molecules have D- (right handed, dextro-rotary) and L- (left handed, levo-rotary) mirror image forms.

An isotropic material can become anisotropic when stress or an induced strain is applied because of an elasto-optic interaction known as stress birefringence, or the photoelastic effect. This effect is usually bad because it reduces the performance of optical components and devices by introducing phase distortions caused by improper mounting or unequal thermal expansion between the mounts and components. Another source of distortion is strain that has been frozen into optical components during manufacture. French curves are excellent examples that show colored strain birefringence patterns when viewed between crossed polarizers. In one application, strain birefringence has been used constructively to produce a variable phase retarder made from crystal quartz and fused silica. A block of electro-optic crystal quartz, is cemented to a block of isotropic fused silica. When a variable electric field is applied to the crystal quartz, its length changes at the resonant frequency of the block. This produces strain in the fused silica block which in turn produces a variable retardation of light passing through the strained fused silica. Depending on the magnitude of the varying electric field, the device can act like a variable quarter-wave plate, a variable half-wave plate, or have other applications.

Matrix Methods for Computing Polarization

An optical system containing various polarizing and retarding devices can be modeled using matrices. In general, there is an incident beam (in matrix form), that has some state of polarization. It interacts with a device called an instrument (also in matrix form) that alters the state of polarization, so that the exiting beam (a third matrix, the product of the first and second matrices) has another state of polarization. There can be more than one instrument matrix acting on the same incident matrix to produce the final matrix. The two most common matrix methods are the Mueller calculus and the Jones calculus.

There is also a visual representation of this process, a Poincaré sphere, on which vectors represent different states of polarization. The polarization instrument moves the vector representing the incident polarization around the sphere in a prescribed manner. The vectors are called Stokes vectors and are used in the Mueller calculus. The various states of polarization are represented on the Poincaré sphere as follows: The equator represents various forms of linear polarization, the poles represent right- and left-circular polarization, and other points on the sphere represent elliptically polarized light. Every point on the sphere corresponds to a different polarization form. The radius of the sphere represents the incident irradiance of the light beam (which is usually assumed to be unity). The effects of various polarization instruments are determined by displacements on the sphere. Partially polarized light or absorption may be dealt with approximately by ignoring the irradiance factor, since the state of polarization is generally the quantity desired. The Poincaré sphere is most useful for visualizing problems involving nonabsorbing materials, various polarization instruments including polarizers, retarders, compensators, half-shade devices, and depolarizers, and has also been applied to ellipsometric problems and stress-optical measurements.

The Poincaré sphere is used with Stokes vectors, which are often designated S_0 , S_1 , S_2 , and S_3 . S_0 is the incident irradiance of the light beam, corresponding to the radius of the Poincaré sphere. S_1 is the difference in irradiances between the horizontal and vertical polarization components of the beam; for example, when S_1 is positive, the preference is for horizontal polarization. S_2 indicates preference for +45° or –45° polarization depending on whether it is positive or negative, and S_3 gives the preference for right or left circular polarization. Thus, the Stokes vectors S_1 , S_2 , and S_3 are simply the three Cartesian coordinates of a point on the

Poincaré sphere. S_1 and S_2 are perpendicular to each other in the equatorial plane, and S_3 points toward the north pole of the sphere. Any state of polarization of a light beam can be specified by these three vectors. The irradiance vector S_0 is related to the other three by the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$ when the beam is completely polarized. If the beam is partially polarized, $S_0^2 > S_1^2 + S_2^2 + S_3^2$.

In the Mueller calculus, the incident beam is represented by the four-component Stokes vector, written as a column vector. This vector has all real elements and gives information about irradiance properties of the beam. Thus it is not able to handle problems involving phase changes or combinations of two beams that are coherent. The instrument matrix is a 4×4 matrix with all real elements.

In the Jones calculus, the Jones vector representing the incident beam is a two-component column vector that generally has complex elements. It contains information about the *amplitude* properties of the beam and hence is well suited for handling problems involving the phases of light waves. However, it cannot handle problems involving depolarization, as the Mueller calculus can. The Jones instrument matrix is a 2×2 matrix whose elements are generally complex. The Jones calculus is well suited to problems involving a large number of similar devices arranged in series in a regular manner and makes it possible to obtain a result expressed explicitly in terms of the number of such devices. The Jones instrument matrix of a train of transparent or absorbing nondepolarizing polarizers and retarders contains no redundant information. The matrix contains four elements, each of which has two parts, so that there are a total of eight constants, none of which is a function of any other. The Mueller instrument matrix of such a train contains much redundancy; there are 16 constants but only 7 of them are independent. More information about the Poincaré sphere and the Mueller and Jones calculus is available in other references.

See also: Matrix Analysis

Further Reading

- Auton, J.P., 1967. Infrared transmission polarizers by photolithography. *Applied Optics* 6, 1023–1027.
- Azzam, Rasheed, M.A., Bashara, N.M., 1977. *Ellipsometry and Polarized Light*. Amsterdam: North-Holland Publishing Co (chaps. 1 and 2).
- Azzam, Rasheed, M.A., 1995. In: Bass, M. (Ed.), *Handbook of Optics, Ellipsometry*, 2nd edn., vol. II. New York: McGraw-Hill (chap. 27).
- Bennett, J.M., 1995. In: Bass, M. (Ed.), *Handbook of Optics, Polarization*, 2nd edn., vol. I. New York: McGraw-Hill (chap. 5).
- Bennett, M., 1995. In: Bass, M. (Ed.), *Handbook of Optics, Polarizers*, 2nd edn., vol. II. New York: McGraw-Hill (chap. 3).
- Bohren, C.F., Huffman, D.R., 1983. *Absorption and Scattering of Light by Small Particles*. New York: John Wiley.
- Born, M., Wolf, E., 1999. *Principles of Optics*, 7th edn, Cambridge: Cambridge University Press (chap. 1).
- Chipman, R.A., 1995. In: Bass, M. (Ed.), *Handbook of Optics, Polarimetry*, 2nd edn., vol. II. New York: McGraw-Hill (chap. 22).
- Clark, J.R., 1956. New calculus for the treatment of optical systems. VIII. Electromagnetic theory. *Journal of the Optical Society of America* 46, 126–131.
- Guenther, R.D., 1990. *Modern Optics*. New York: John Wiley and Sons (chaps. 2, 13, and 14).
- Hass, M., O'Hara, M., 1965. Sheet infrared transmission polarizers. *Applied Optics* 4, 1027–1031.
- Hecht, E., Zajac, A., 1997. *Optics*, 3rd edn. Reading, MA: Addison-Wesley Publishing Co (chap. 8).
- Jackson, J.D., 1999. *Classical Electrodynamics*, 3rd edn. New York: John Wiley and Sons, Inc (chap. 7).
- Jaspersen, S.N., Schnatterly, S.E., 1969. An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. *Reviews of Scientific Instruments* 40, 761–767.
- Kerker, M., 1969. *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.
- Kliger, D.S., Lewis, J.W., Randall, C.E., 1990. *Polarized Light in Optics and Spectroscopy*. San Diego: Academic Press (chaps. 1–5).
- Maldonado, T.A., 1995. In: Bass, M. (Ed.), *Handbook of Optics, Electro-optic modulators*, 2nd edn., vol. II. New York: McGraw-Hill, Inc. (chap. 13).
- Markuvitz, N. (Ed.), 1951. *Waveguide Handbook*. New York: McGraw-Hill Book Co, pp. 218–229. 285–289.
- Shurcliff, W.A., 1962. *Polarized Light, Production and Use*. Cambridge, MA: Harvard University Press.
- van de Hulst, H.C., 1957. *Light Scattering by Small Particles*. New York: John Wiley.

Introduction

Hooke postulated, in the seventeenth century, that light waves must be transverse but his idea was forgotten. Young and Fresnel put forward the same idea in the nineteenth century and accompanied their postulation with a theoretical description of light based on transverse waves. Forty years later, Maxwell proved that light must be a transverse wave and that $\mathbf{E}$ and $\mathbf{H}$, for a plane wave in an isotropic medium with no free charge and no currents, are mutually perpendicular and lie in a plane normal to the direction of propagation, $\mathbf{k}$.

Convention requires that we use the electric vector to label the direction of the electromagnetic wave's polarization. The direction of the displacement vector is called the direction of polarization and the plane containing the direction of polarization and the propagation vector is called the plane of polarization. The selection of the electric field is not completely arbitrary, except in relativistic situations, when $v \approx c$, the interaction of the electromagnetic wave with matter will be dominated by the electric field.

Assume that a plane wave is propagating in the z -direction. In complex notation, the plane wave is given by

$$\tilde{\mathbf{E}} = \mathbf{E}_0 e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi)} = \mathbf{E}_0 e^{i(\omega t - kz + \phi)} \quad (1)$$

The procedure used to decompose an arbitrary polarization into components parallel to two axes of a Cartesian coordinate system, is a technique used extensively in vector algebra to simplify mathematical calculations. According to the mathematical formalism associated with this technique, the polarization is described in terms of a set of basis vectors, $\mathbf{e}_i$. An arbitrary polarization would be expressed as

$$\mathbf{E} = \sum_{i=1}^2 a_i \mathbf{e}_i \quad (2)$$

The set of basis vectors, $\mathbf{e}_i$, are orthonormal, i.e.:

$$\mathbf{e}_i \mathbf{e}_j^* = \delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \quad (3)$$

where we have assumed that the basis vectors could be complex. We mention this mathematical formalism because an identical formalism is encountered in a description of spin.

In a Cartesian coordinate system, the $\mathbf{e}_i$'s are the unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$. The summation in Eq. (2) extends over only two terms because the electromagnetic wave is transverse, confining $\mathbf{E}$ to a plane normal to the direction of propagation (according to the coordinate convention we have selected, the $\mathbf{E}$ field is in the x, y plane).

Polarization Ellipse

Following the formalism of Eq. (2), a polarized wave can be written in terms of the x and y components of $\mathbf{E}_0$

$$\tilde{\mathbf{E}} = \mathbf{E}_{0x} e^{i(\omega t - kz + \phi_1)} \hat{\mathbf{i}} + \mathbf{E}_{0y} e^{i(\omega t - kz + \phi_2)} \hat{\mathbf{j}} \quad (4)$$

We will use only the real part of $\mathbf{E}$ for manipulation, to prevent errors. We divide each component of the electric field by its maximum value, so that the problem is reduced to one of the following two sinusoidal varying unit vectors:

$$\begin{aligned} \frac{E_x}{E_{0x}} &= \cos(\omega t - kz + \phi_1) \\ &= \cos(\omega t - kz) \cos \phi_1 - \sin(\omega t - kz) \sin \phi_1 \\ \frac{E_y}{E_{0y}} &= \cos(\omega t - kz) \cos \phi_2 - \sin(\omega t - kz) \sin \phi_2 \end{aligned} \quad (5)$$

When these unit vectors are added together, the resulting equation will describe the path taken by the tip of the resultant vector. The path will create a Lissajous figure.

To obtain the equation for the Lissajous figure, we eliminate the dependence of the unit vectors on $(\omega t - kz)$. First, multiply the equations by $\sin \phi$ and $\sin \phi_1$, respectively and then subtract the resulting equations. Second, multiply the two equations by $\cos \phi_2$ and $\cos \phi_1$ respectively and then subtract the new equations. These two operations yield the following pair of equations:

$$\begin{aligned} \frac{E_x}{E_{0x}} \sin \phi_2 - \frac{E_y}{E_{0y}} \sin \phi_1 \\ = \cos(\omega t - kz) [\cos \phi_1 \sin \phi_2 - \sin \phi_1 \cos \phi_2] \end{aligned} \quad (6)$$

$$\begin{aligned} \frac{E_x}{E_{0x}} \cos \phi_2 - \frac{E_y}{E_{0y}} \cos \phi_1 \\ = \sin(\omega t - kz) [\cos \phi_1 \sin \phi_2 - \sin \phi_1 \cos \phi_2] \end{aligned} \quad (7)$$

The term in brackets can be simplified using the trig identity

$$\begin{aligned} \sin \delta &= \sin(\phi_2 - \phi_1) \\ &= \cos \phi_1 \sin \phi_2 - \sin \phi_1 \cos \phi_2 \end{aligned} \quad (8)$$

After replacing the term in brackets by $\sin \delta$, the two equations are squared and added yielding the equation for the Lissajous figure:

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - \left(\frac{2E_x E_y}{E_{0x} E_{0y}}\right) \cos \delta = \sin^2 \delta \quad (9)$$

The trig identity

$$\cos \delta = \cos(\phi_2 - \phi_1) = \cos \phi_1 \cos \phi_2 + \sin \phi_1 \sin \phi_2$$

was also used to further simplify Eq. (9).

Eq. (9) has the same form as the equation of a conic

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

Geometry defines the conic as an ellipse because, from Eq. (9)

$$B^2 - 4AC = \frac{4}{E_{0x}^2 E_{0y}^2} (\cos^2 \delta - 1) < 0$$

This ellipse is called the polarization ellipse. The orientation of the ellipse, with respect to the x-axis, is

$$\tan 2\theta = \frac{B}{A - C} = \frac{2E_{0x} E_{0y} \cos \delta}{E_{0x}^2 - E_{0y}^2} \quad (10)$$

The tip of the resultant electric field vector obtained from Eq. (4) traces out the polarization ellipse in the plane normal to $\mathbf{k}$, as predicted by Eq. (9). The ratio of the length of the minor to the major axis of the ellipse is equal to the ellipticity, φ , i.e., the amount of deviation of the ellipse from a circle

$$\begin{aligned} \tan \varphi &= \pm \left(\frac{E_m}{E_M} \right) \\ &= \frac{E_{0x} \sin \phi_1 \sin \theta - E_{0y} \sin \phi_2 \cos \theta}{E_{0x} \cos \phi_1 \cos \theta + E_{0y} \cos \phi_2 \sin \theta} \end{aligned} \quad (11)$$

In particle physics, the light would be said to have a negative helicity if it rotated in a clockwise direction. If we look at the source, the electric vector seems to follow the threads of a left-handed screw, agreeing with the nomenclature that left-handed quantities are negative. However, in optics the light that rotates clockwise, as we view it traveling toward us from the source, is said to be right-circularly polarized. The counterclockwise rotating light is left-circularly polarized. The association of right-circularly polarized light, with 'right-handedness' in optics, came about by looking at the path of the electric vector in space at a fixed time, then $\tan \psi = \tan(\phi - kz)$ (Fig. 1).

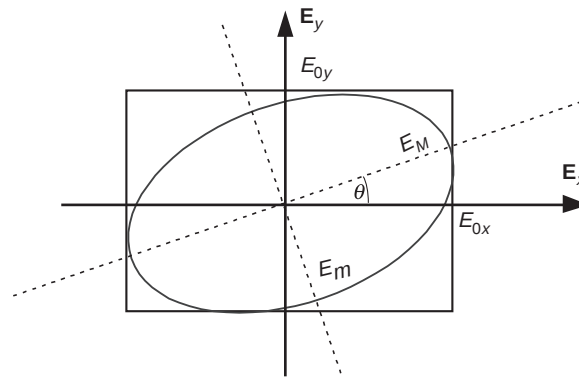


Fig. 1 General form of the ellipse described by Eq. (9). Reproduced with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

Table 1 Typical Stokes vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$
+45° polarization	$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	−45° polarization	$\begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$
Right circular polarization	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$	Left circular polarization	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$

Stokes Parameters

In the formalism associated with Eq. (2), the expansion coefficients, a_i , can be used to form a 2–2 matrix which, in statistical mechanics, is called the density matrix and in optics, the coherency matrix (Table 1). The elements of the matrix are formed by the rule

$$\rho_{ij} = \mathbf{a}_i \mathbf{a}_j^* \quad (12)$$

The matrix is Hermitian, so that $\rho_{ij} = \rho_{ji}^*$. We will not develop the theory of polarization using the coherency matrix, but simply use the coherency matrix to justify the need for four independent measurements to characterize polarization. There is no unique set of measurements required by the theory, but normally measurements made are of the Stokes's parameters which are directly related to the parameters of the polarization ellipse of Eq. (9).

The Stokes parameters of a light wave are measurable quantities, defined as:

$s_0 \Rightarrow$ Total flux density

$s_1 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented parallel to the x -axis and one oriented parallel to the y -axis. The x and y -axes are usually selected to be parallel to the horizontal and vertical directions in the laboratory.

$s_2 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented at 45° to the x -axis and one oriented at 135°.

$s_3 \Rightarrow$ Difference between flux density transmitted by a right-circular polarizer and a left-circular polarizer.

If the Stokes parameters are to characterize the polarization of a wave, they must be related to the parameters of the polarization ellipse. It is therefore important to establish that the Stokes parameters are variables of the polarization ellipse (Eq. (9)).

In its current form, Eq. (9) contains no measurable quantities and thus must be modified if it is to be associated with the Stokes parameters. The time average of the Poynting vector is the quantity observed when measurements are made of light waves. We must, therefore, find the time average of Eq. (9) if we wish to relate its parameters to observable quantities.

The time average Eq. (9) can now be written as

$$\frac{\langle E_x^2 \rangle}{E_{0x}^2} + \frac{\langle E_y^2 \rangle}{E_{0y}^2} - 2 \frac{\langle E_x E_y \rangle}{E_{0x} E_{0y}} \cos \delta = \sin^2 \delta \quad (13)$$

where time average is denoted by $\langle \rangle$

$$\langle E_x^2 \rangle = \frac{1}{T} \int_{t_0}^{t_0+T} E_{0x}^2 [\cos(\omega t - kz) \cos \phi_1 - \sin(\omega t - kz) \sin \phi_1]^2 dt \quad (14)$$

With the time averages, Eq. (13) can be written as

$$\begin{aligned} (E_{0x}^2 + E_{0y}^2)^2 - (E_{0x}^2 - E_{0y}^2)^2 - (2E_{0x}E_{0y} \cos \delta)^2 \\ = (2E_{0x}E_{0y} \sin \delta)^2 \end{aligned} \quad (15)$$

Each term in this equation can be identified with a Stokes parameter:

$$\begin{aligned} s_0 &= \langle E_{0x}^2 \rangle + \langle E_{0y}^2 \rangle & s_1 &= \langle E_{0x}^2 \rangle - \langle E_{0y}^2 \rangle \\ s_2 &= \langle 2E_{0x}E_{0y} \cos \delta \rangle & s_3 &= \langle 2E_{0x}E_{0y} \sin \delta \rangle \end{aligned} \quad (16)$$

Eq. (15) can now be written as

$$s_0^2 - s_1^2 - s_2^2 = s_3^2 \quad (17)$$

For a polarized wave, only three of the Stokes parameters are independent. This agrees with the requirement placed upon elements of the Hermitian, coherency matrix, introduced above.

With this demonstration of the connection between the Stokes parameters and the polarization ellipse, the Stokes parameters can be written in terms of the parameters of the polarization ellipse.

$$\begin{aligned} s_1 &= s_0 \cos 2\varphi \cos 2\theta \\ s_2 &= s_0 \cos 2\varphi \sin 2\theta \\ s_3 &= s_0 \sin 2\varphi \end{aligned} \quad (18)$$

Mueller Calculus

Mueller pointed out that the Stokes parameters can be thought of as elements of a column matrix or a 4-vector (Table 2)

$$\mathcal{S} = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} \quad (19)$$

Table 2 Mueller matrices for polarizers

Polarizer	Transmission axis	Mueller matrix
Linear	Horizontal	$\frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	Vertical	$\frac{1}{2} \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	+45°	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	−45°	$\frac{1}{2} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
Circular	Right	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$
	Left	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$

Each element of the Stokes vector represents a measurable intensity. The vector can represent not only polarized light, but unpolarized or partially polarized light.

In 1929, Soleillet discovered that an optical device performed a linear transformation on the input wave and, in 1942, Perrin put this fact into a matrix formalism involving the Stokes vectors:

$$\mathcal{S}_1 = \mathcal{M} \cdot \mathcal{S}_0 \quad (20)$$

Mueller used experimentally derived 4×4 matrices, the $\mathcal{M}$ in Eq. (20), to describe the effect of an optical device on a light wave's polarization. The $\mathcal{S}$'s in Eq. (20) are column matrices whose elements are the Stokes parameters (Eq. (16)). The matrices are based upon an assumed linear relationship between the incident and transmitted beams (Table 3).

The analysis of the effect of a number of polarizers and retarders is made easier by the use of the Mueller–Stokes matrix calculus, coupled with the use of the Stokes vectors. To determine the Mueller matrices, one must measure the effect a device has on unpolarized, horizontally polarized, linearly polarized at $+45^\circ$, and right-circularly polarized light. It is then an algebraic exercise to generate the elements of the matrix. A few optical devices and their Mueller matrix are listed in Tables 2 and 3.

The Mueller matrix contains 16 parameters but there are only seven independent parameters. The matrix contains no information about absolute phase but it handles partially polarized and unpolarized light without modification (Table 4).

Table 3 Mueller matrices for retarders

Retarder	Transmission axis	Mueller matrix
Quarter-wave plate	Horizontal	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
	$+45^\circ$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$
	-45°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
	$\pm 45^\circ$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

Table 4 Jones vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
+45° polarization	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	−45° polarization	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Right circular polarization	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}$	Left circular polarization	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}$

Table 5 Jones matrices for retarders

Retarder	Transmission axis	Jones matrix
	Horizontal	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
	Vertical	$\begin{bmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
	+45°	$\frac{1}{2} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$
Quarter-wave plate	−45°	$\frac{1}{2} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
	±45°	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Jones Vector

There is one other representation of polarized light, complementary to the Stokes vector, developed by Clark Jones in 1941 and called the Jones vector. It is superior to the Stokes vector in that it handles light of a known phase and amplitude with a reduced number of parameters. However, it is inferior to the Stokes vector in that, unlike the Stokes representation, which is experimentally determined, the Jones representation cannot handle unpolarized or partially polarized light. The Jones vector is a theoretical construct that can only describe light with a well-defined phase and frequency. The density matrix formalism can be used to correct the shortcomings of the Jones vector, but then the simplicity of the Jones representation is lost.

Assuming the coordinate system is such that the electromagnetic wave is propagating along the z -axis, it was shown earlier that any polarization could be decomposed into two orthogonal E vectors, i.e., parallel to the x and y directions. The Jones vector is defined as a two-row, column matrix consisting of the complex components in the x and y direction:

$$\mathbf{E} = \begin{bmatrix} E_{0x} e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi_1)} \\ E_{0y} e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi_2)} \end{bmatrix} \quad (21)$$

If absolute phase is not an issue, then we may normalize the vector by dividing by that number (real or complex) that simplifies the components but keeps the sum of the square of the components equal to one. For example, assume that $E_{0x} = E_{0y}$, then

$$\mathbf{E} = E_{0x} e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi_1)} \begin{bmatrix} 1 \\ e^{i\delta} \end{bmatrix} \quad (22)$$

The normalized vector would be the terms contained within the bracket, each divided by $\frac{1}{\sqrt{2}}$.

Table 6 Jones matrices for polarizers

Polarizer	Transmission axis	Jones matrix
Linear	Horizontal	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
	+45°	$\frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
	−45°	$\frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$
Circular	Right	$\frac{1}{2} \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$
	Left	$\frac{1}{2} \begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

The general form of the Jones vector is

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \quad \mathbf{E}^* = [\mathbf{A}^* \mathbf{B}^*] \quad (23)$$

Some examples of Jones vectors are shown in [Table 4](#).

Jones Calculus

The Jones calculus is complementary to the Mueller calculus and operates on the Jones vector, ([Eq. \(23\)](#)), similar to the way the Mueller matrix operates on the Stokes vector ([Table 5](#)).

The Jones matrix contains eight independent parameters with no redundancy, making it simpler than the Mueller calculus. However, the Jones calculus only applies to polarized light. The Jones calculus can be extended, using the density matrix formalism to allow manipulation of unpolarized light, but with a loss of simplicity. The matrix equation for Jones calculus is

$$\mathbf{E}_{\text{out}} = \mathcal{W} \mathbf{E}_{\text{in}} \quad (24)$$

A few optical devices and their Jones matrices, $\mathcal{W}$, are listed in [Tables 5](#) and [6](#).

For every matrix in Jones calculus, there is a matrix in Mueller calculus, but the converse is not true. For example, it is possible to construct a depolarizer, by using a thick piece of opal glass in the visible or by using gold covered sandpaper in the infrared. Such a device can be described in Mueller calculus, but there is no matrix for such a device in Jones calculus ([Tables 5](#) and [6](#)).

See also: Polarization Introduction

Further Reading

- Born, M., Wolf, E., 1970. Principles of Optics. New York: Pergamon Press.
 Clark Jones, R., 1941. A new calculus for the treatment of optical systems. Journal of the Optical Society of America 32, 488–503.
 Clark Jones, R., 1942. A new calculus for the treatment of optical systems. Journal of the Optical Society of America 31, 486–493.
 Guenther, R.D., 1990. Modern Optics. New York: Wiley.
 Hecht, E., 1998. Optics, 3rd edn. Reading, MA: Addison-Wesley.
 Klein, M.V., 1970. Optics. New York: Wiley.
 Kliger, D.S., Lewis, J.W., Randall, C.E., 1990. Polarized Light in Optics and Spectroscopy. Boston: Academic Press.
 McMaster, W.H., 1954. Polarization and the Stokes parameters. American Journal of Physics 22, 351–362.
 Mueller, H., 1948. The foundation of optics. Journal of the Optical Society of America 38, 661.

Electromagnetic Theory

SG Johnson and JD Joannopoulos, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Photonic crystals are periodically structured electromagnetic media, generally possessing photonic bandgaps: ranges of frequency in which light cannot propagate through the structure. This periodicity, whose lengthscale is proportional to the wavelength of light in the bandgap, is the electromagnetic analog of a crystalline atomic lattice, where the latter acts on the electron wavefunction to produce the familiar band gaps, semiconductors, etc., of solid-state physics. The study of photonic crystals is likewise governed by the Bloch–Floquet theorem, and intentionally introduced defects in the crystal (analogous to electronic dopants) give rise to localized electromagnetic states: linear waveguides and point-like cavities. The crystal can thus form a kind of perfect optical ‘insulator’, which can confine light around sharp bends, in lower-index media, and within wavelength-scale cavities, among other novel possibilities for control of electromagnetic phenomena. Below is introduced the basic theoretical background of photonic crystals in one, two, and three dimensions (schematically depicted in Fig. 1), as well as hybrid structures that combine photonic-crystal effects in some directions with more-conventional index guiding in other directions. (Line and point defects in photonic crystals are discussed in another article.)

Electromagnetic wave propagation in periodic media was first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic ‘twinning’ planes (across which the dielectric tensor undergoes a mirror flip). These correspond to one-dimensional photonic crystals, and he identified the fact that they have a narrow bandgap prohibiting light propagation through the planes. This bandgap is angle-dependent, due to the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. (A similar effect is responsible for many other iridescent colors in nature, such as those of butterfly wings and abalone shells.) Although multilayer films received intensive study over the following century, it was not until 100 years later, when Yablonovitch and John, in 1987, joined the tools of classical electromagnetism and solid-state physics, that the concepts of omnidirectional photonic bandgaps in two and three dimensions was introduced. This generalization, which inspired the name ‘photonic crystal’, led to many subsequent developments in their fabrication, theory, and application, from integrated optics to negative refraction to optical fibers that guide light in air.

Maxwell’s Equations in Periodic Media

The study of wave propagation in three-dimensionally periodic media was pioneered by Felix Bloch in 1928, unknowingly extending an 1883 theorem in one dimension by Gaston Floquet. Bloch proved that waves in such a medium can propagate without scattering, their behavior described by a periodic envelope function multiplied by a planewave. Although Bloch studied quantum mechanics, leading to the surprising result that electrons in a conductor scatter only from imperfections and not from the periodic ions, the same techniques can be applied to electromagnetism by casting Maxwell’s equations as an eigenproblem in analog with Schrödinger’s equation. By combining the source-free Faraday’s and Ampere’s laws at a fixed (angular) frequency ω , i.e., time dependence $e^{-i\omega t}$ one can obtain an equation in only the magnetic field $\vec{H}$:

$$\vec{\nabla} \times \frac{1}{\epsilon} \vec{\nabla} \times \vec{H} = \left(\frac{\omega}{c}\right)^2 \vec{H} \quad (1)$$

where ϵ is the dielectric function $\epsilon(x,y,z)$ and c is the speed of light. This is an eigenvalue equation, with eigenvalue (ω/c^2) and an eigen-operator $\vec{\nabla} \times (1/\epsilon) \vec{\nabla} \times$ that is Hermitian (acts the same to the left and right) under the inner product $\int \vec{H}^* \cdot \vec{H}$ between two

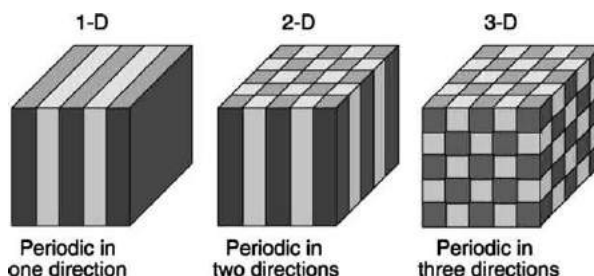


Fig. 1 Schematic depiction of photonic crystals periodic in one, two, and three directions, where the periodicity is in the material (typically dielectric) structure of the crystal. Only a 3d periodicity, with a more complex topology than is shown here, can support an omnidirectional photonic bandgap.

fields $\vec{H}$ and $\vec{E}$. (The two curls correspond roughly to the 'kinetic energy' and $1/\varepsilon$ to the 'potential' compared to the Schrödinger Hamiltonian $\nabla^2 + V$.) It is sometimes more convenient to write a generalized Hermitian eigenproblem in the electric field $\vec{E}$, $\vec{\nabla} \times \vec{\nabla} \times \vec{E} = (\omega/c)^2 \varepsilon \vec{E}$ which separates the kinetic and potential terms. Electric fields that lie in higher ε i.e., lower potential, will have lower ω this is discussed further in the context of the variational theorem of Eq. (3).

Thus, the same linear-algebraic theorems as those in quantum mechanics can be applied to the electromagnetic wave solutions. The fact that the eigen-operator is Hermitian and positive-definite (for real $\varepsilon > 0$) implies that the eigenfrequencies ω are real, for example, and also leads to orthogonality, variational formulations, and perturbation-theory relations that are discussed further below. An important difference compared to quantum mechanics is that there is a transversality constraint: one typically excludes $\vec{\nabla} \cdot \vec{H} \neq 0$ (or $\vec{\nabla} \cdot \varepsilon \vec{E} \neq 0$) eigensolutions, which lie at $\omega = 0$; i.e., static-field solutions with free magnetic (or electric) charge are forbidden.

Bloch Waves and Brillouin Zones

A photonic crystal corresponds to a periodic dielectric function $\varepsilon(\vec{x}) = \varepsilon(\vec{x} + \vec{R}_i)$ for some primitive lattice vectors $\vec{R}_i$ ($i = 1, 2, 3$, for a crystal periodic in all three dimensions). In this case, the Bloch–Floquet theorem for periodic eigenproblems states that the solutions to Eq. (1) can be chosen of the form $\vec{H}(\vec{x}) = e^{i\vec{k} \cdot \vec{x}} \vec{H}_{n,\vec{k}}(\vec{x})$ with eigenvalues $\omega_n(\vec{k})$, where $\vec{H}_{n,\vec{k}}$ is a periodic envelope function satisfying

$$(\vec{\nabla} + i\vec{k}) \times \frac{1}{\varepsilon} (\vec{\nabla} + i\vec{k}) \times \vec{H}_{n,\vec{k}} = \frac{\omega_n(\vec{k})^2}{c} \vec{H}_{n,\vec{k}} \quad (2)$$

yielding a different Hermitian eigenproblem over the primitive cell of the lattice at each Bloch wavevector $\vec{k}$. This primitive cell is a finite domain if the structure is periodic in all directions, leading to discrete eigenvalues labeled by $n = 1, 2, \dots$. These eigenvalues $\omega_n(\vec{k})$, are continuous functions of $\vec{k}$, forming discrete 'bands' when plotted versus the latter, in a 'band structure' or dispersion diagram – both ω and $\vec{k}$ are conserved quantities, meaning that a band diagram maps out all possible interactions in the system. (Note also that $\vec{k}$ is not required to be real; complex $\vec{k}$ gives evanescent modes that can exponentially decay from the boundaries of a finite crystal, but which cannot exist in the bulk.)

Moreover, the eigensolutions are periodic functions of $\vec{k}$ as well: the solution at $\vec{k}$ is the same as the solution at $\vec{k} + \vec{G}_j$ where $\vec{G}_j$ is a primitive reciprocal lattice vector defined by $\vec{R}_i \cdot \vec{G}_j = 2\pi\delta_{ij}$. Thanks to this periodicity, one need only compute the eigensolutions for $\vec{k}$ within the primitive cell of this reciprocal lattice – or, more conventionally, one considers the set of inequivalent wavevectors closest to the $\vec{k} = 0$ origin, a region called the first Brillouin zone. For example, in a one-dimensional system, where $R_1 = a$ for some periodicity a and $G = 2\pi/a$, the first Brillouin zone is the region $k = -\pi/\dots\pi/a$; all other wavevectors are equivalent to some point in this zone under translation by a multiple of G_1 . Furthermore, the first Brillouin zone may itself be redundant if the crystal possesses additional symmetries such as mirror planes; by eliminating these redundant regions, one obtains the irreducible Brillouin zone, a convex polyhedron that can be found tabulated for most crystalline structures. In the preceding one-dimensional example, since most systems will have time-reversal symmetry ($k \rightarrow -k$), the irreducible Brillouin zone would be $k = 0 \dots \pi/a$.

The familiar dispersion relations of uniform waveguides arise as a special case of the Bloch formalism: such translational symmetry corresponds to a period $a \rightarrow 0$. In this case, the Brillouin zone of the wavevector k (also called β) is unbounded, and the envelope function $\vec{H}_{n,\vec{k}}$ is a function only of the transverse coordinates.

The Origin of the Photonic Bandgap

A complete photonic bandgap is a range of ω in which there are no propagating (real $\vec{k}$) solutions of Maxwell's Eq. (2) for any $\vec{k}$ surrounded by propagating states above and below the gap. There are also incomplete gaps, which only exist over a subset of all possible wavevectors, polarizations, and/or symmetries. Both sorts of gaps are discussed in the subsequent sections, but in either case, their origins are the same and can be understood by examining the consequences of periodicity for a simple one-dimensional system.

Consider a one-dimensional system with uniform $\varepsilon = \bar{\varepsilon}$ which has planewave eigensolutions $\omega(k) = ck$ as depicted in Fig. 2(left). This ε has trivial periodicity a for any $a \geq 0$, with $a = 0$ giving the usual unbounded dispersion relation. One is free, however, to label the states in terms of Bloch envelope functions and wavevectors for some $a \neq 0$, in which case the bands for $|k| > \pi/a$ are translated (folded) into the first Brillouin zone, as shown by the dashed lines in Fig. 2(left). In particular, the $k = -\pi/a$ mode in this description now lies at an equivalent wavevector to the $k = \pi/a$ mode, and at the same frequency; this accidental degeneracy is an artifact of the 'artificial' period that has been chosen. Instead of writing these wave solutions with electric fields $\vec{E}(x) \sim e^{\pm i\pi x/a}$, one can equivalently write linear combinations $e(x) = \cos(\pi x/a)$ and $o(x) = \sin(\pi x/a)$ as shown in Fig. 3, both at $\omega = c\pi/a\sqrt{\bar{\varepsilon}}$. Now, however, suppose that one perturbs ε so that it is nontrivially periodic with period a for example, a sinusoid $\varepsilon(x) = \bar{\varepsilon} \cdot [1 + \Delta \cos(2\pi x/a)]$, or a square wave as in the inset of Fig. 2(right). In the presence of such an oscillating 'potential', the accidental degeneracy between $e(x)$ and $o(x)$ is broken: supposing $\Delta > 0$, then the field $e(x)$ is more concentrated in the higher- ε regions than $o(x)$ and so lies at a lower frequency. This opposite shifting of the bands away from the mid-gap frequency $\omega \cong c\pi/a\sqrt{\bar{\varepsilon}}$ creates a bandgap, as depicted in Fig. 2(right). (In fact, from the perturbation theory described subsequently, one can show that for $\Delta \ll 1$ the bandgap, as a fraction of mid-gap frequency, is $\Delta\omega/\omega \cong \Delta/2$.) By the same arguments, it follows that any

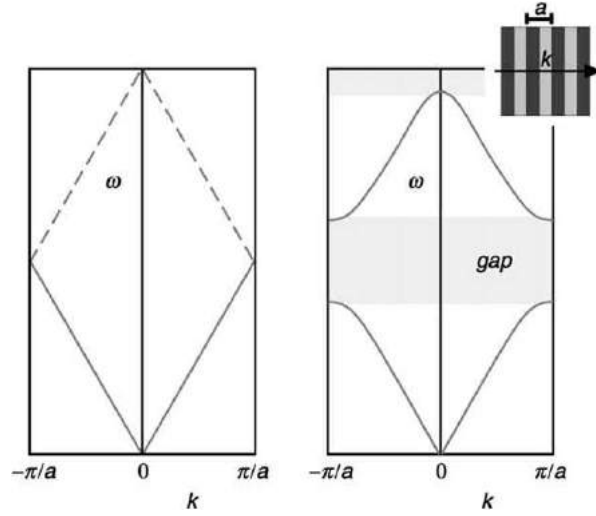


Fig. 2 Left: Dispersion relation (band diagram), frequency ω versus wavenumber k of a uniform one-dimensional medium, where the dashed lines show the 'folding' effect of applying Bloch's theorem with an artificial periodicity a . Right: Schematic effect on the bands of a physical periodic dielectric variation (inset), where a gap has been opened by splitting the degeneracy at the $k = \pm \pi/a$ Brillouin-zone boundaries (as well as a higher-order gap at $k=0$).

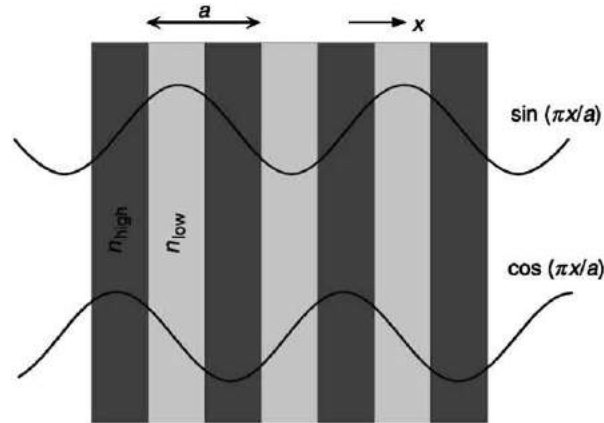


Fig. 3 Schematic origin of the band gap in one dimension. The degenerate $ka = \pi/a$ planewaves of a uniform medium are split into $\cos(\pi x/a)$ and $\sin(\pi x/a)$ standing waves by a dielectric periodicity, forming the lower and upper edges of the bandgap, respectively – the former has electric-field peaks in the high dielectric (n_{high}) and so will lie at a lower frequency than the latter (which peaks in the low dielectric).

periodic dielectric variation in one dimension will lead to a bandgap, albeit a small gap for a small variation; a similar result was identified by Lord Rayleigh in 1887.

More generally, it follows immediately from the properties of Hermitian eigensystems that the eigenvalues minimize a variational problem:

$$\omega_{n,\vec{k}}^2 = \min_{\vec{E}_{n,\vec{k}}} \frac{\int |(\vec{\nabla} + i\vec{k}) \times \vec{E}_{n,\vec{k}}|^2}{\int \epsilon |\vec{E}_{n,\vec{k}}|^2} c^2 \quad (3)$$

in terms of the periodic electric field envelope $\vec{E}_{n,\vec{k}}$ where the numerator minimizes the 'kinetic energy' and the denominator minimizes the 'potential energy'. Here, the $n > 1$ bands are additionally constrained to be orthogonal to the lower bands:

$$\int \vec{H}_{m,\vec{k}}^* \cdot \vec{H}_{n,\vec{k}} = \int \epsilon \vec{E}_{m,\vec{k}}^* \cdot \vec{E}_{n,\vec{k}} = 0 \quad (4)$$

for $m < n$. Thus, at each $\vec{k}$ there will be a gap between the lower 'dielectric' bands concentrated in the high dielectric (low potential) and the upper 'air' bands that are less concentrated in the high dielectric: the air bands are forced out by the orthogonality condition, or otherwise must have fast oscillations that increase their kinetic energy. (The dielectric/air bands are analogous to the valence/conduction bands in a semiconductor.)

In order for a complete bandgap to arise in two or three dimensions, two additional hurdles must be overcome. First, although in each symmetry direction of the crystal (and each $\vec{k}$ point) there will be a bandgap by the one-dimensional argument, these bandgaps will not necessarily overlap in frequency (or even lie between the same bands). In order that they overlap, the gaps must be sufficiently large, which implies a minimum ϵ contrast (typically at least 4/1 in 3d). Since the 1d mid-gap frequency $\sim c\pi/a\sqrt{\epsilon}$ varies inversely with the period a it is also helpful if the periodicity is nearly the same in different directions – thus, the largest gaps typically arise for hexagonal lattices in 2d and fcc lattices in 3d, which have the most nearly circular/spherical Brillouin zones. Second, one must take into account the vectorial boundary conditions on the electric field: moving across a dielectric boundary from ϵ to some $\epsilon' < \epsilon$, the inverse ‘potential’ $\epsilon|\vec{E}|^2$ will decrease discontinuously if $\vec{E}$ is parallel to the interface ($\vec{E}_{\parallel}$ is continuous) and will increase discontinuously if $\vec{E}$ is perpendicular to the interface ($\epsilon\vec{E}_{\perp}$ is continuous). This means that, whenever the electric field lines cross a dielectric boundary, it is much harder to strongly contain the field energy within the high dielectric, and the converse is true when the field lines are parallel to a boundary. Thus, in order to obtain a large bandgap, a dielectric structure should consist of thin, continuous veins/membranes along which the electric field lines can run – this way, the lowest band(s) can be strongly confined, while the upper bands are forced to a much higher frequency because the thin veins cannot support multiple modes (except for two orthogonal polarizations). The veins must also run in all directions, so that this confinement can occur for all $\vec{k}$ and polarizations, necessitating a complex topology in the crystal.

Ultimately, however, in two or three dimensions there are only rules of thumb for the existence of a bandgap in a periodic structure, since no rigorous criteria have yet been determined. This made the design of 3d photonic crystals a trial and error process, with the first example by Ho *et al.* of a complete 3d gap coming three years after the initial 1987 concept. As is discussed by the final section below, a small number of families of 3d photonic crystals have since been identified, with many variations thereof explored for fabrication.

Computational Techniques

Because photonic crystals are generally complex, high index-contrast, two- and three-dimensional vectorial systems, numerical computations are a crucial part of most theoretical analyses. Such computations typically fall into three categories: time-domain ‘numerical experiments’ that model the time-evolution of the fields with arbitrary starting conditions in a discretized system (e.g., finite-difference); definite-frequency transfer matrices wherein the scattering matrices are computed in some basis to extract transmission/reflection through the structure; and frequency-domain methods to directly extract the Bloch fields and frequencies by diagonalizing the eigenoperator. The first two categories intuitively correspond to directly measurable quantities such as transmission (although they can also be used to compute e.g., eigenvalues), whereas the third is more abstract, yielding the band diagrams that provide a guide to interpretation of measurements as well as a starting-point for device design and semi-analytical methods. Moreover, several band diagrams are included in the following sections, and so the frequency-domain method used to compute them is briefly outlined here.

Any frequency-domain method begins by expanding the fields in some complete basis, $\vec{H}_{\vec{k}}(\vec{x}) = \sum_n h_n \vec{b}_n(\vec{x})$, transforming the partial differential Eq. (2) into a discrete matrix eigenvalue problem for the coefficients h_n . Truncating the basis to N elements leads to $N \times N$ matrices, which could be diagonalized in $O(N^3)$ time by standard methods. This is impractical for large 3d systems, however, and is also unnecessary – typically, one only wants the few lowest eigenfrequencies, in which case one can use iterative eigensolver methods requiring only $\sim O(N)$ time. Perhaps the simplest such method is based directly on the variational theorem Eq. (3): given some starting coefficients h_n one iteratively minimizes the variational ‘Rayleigh’ quotient using e.g., preconditioned conjugate-gradient descent. This yields the lowest band’s eigenvalue and field, and upper bands are found by the same minimization while orthogonalizing against the lower bands (‘deflation’). There is one additional difficulty, however, and that is that one must at the same time enforce the $(\vec{\nabla} + i\vec{k}) \cdot \vec{H}_{\vec{k}} = 0$ transversality constraint, which is nontrivial in three dimensions. The simplest way to maintain this constraint is to employ a basis that is already transverse, for example planewaves $\vec{h}_{\vec{G}} e^{i\vec{G} \cdot \vec{x}}$ with transverse amplitudes $\vec{h}_{\vec{G}} \cdot (\vec{G} + \vec{k}) = 0$. (In such a planewave basis, the action of the eigen-operator can be computed via a fast Fourier transform in $O(N \log N)$ time.)

Semi-analytical Methods: Perturbation Theory

As in quantum mechanics, the eigenstates can be the starting point for many analytical and semi-analytical studies. One common technique is perturbation theory, applied to small deviations from an ideal system – closely related to the variational Eq. (3), perturbation theory can be exploited to consider effects such as nonlinearities, material absorption, fabrication disorder, and external tunability. Not only is perturbation theory useful in its own right, but it also illustrates both old and peculiarly new features that arise in such analyses of electromagnetism compared to scalar problems such as quantum mechanics.

Given an unperturbed eigenfield $\vec{E}_{n,\vec{k}}$ for a structure ϵ the lowest-order correction $\Delta\omega_n^{(1)}$ to the eigenfrequency from a small perturbation $\Delta\epsilon$ is given by

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int \Delta\epsilon |\vec{E}_{n,\vec{k}}|^2}{\int \epsilon |\vec{E}_{n,\vec{k}}|^2} \quad (5)$$

where the integral is over the primitive cell of the lattice. A Kerr nonlinearity would give $\Delta\epsilon \sim |\vec{E}|^2$, material absorption would produce an imaginary frequency correction (decay coefficient) from a small imaginary $\Delta\epsilon$ and so on. Similarly, one can compute the shift in frequency from a small $\Delta\vec{k}$ in order to determine the group velocity $d\omega/dk$; this variation of perturbation theory is also called $k \cdot p$ theory. All such first-order perturbation corrections are well known from quantum mechanics, and in the limit of infinitesimal perturbations give the exact Hellman–Feynman expression for the derivative of the eigenvalue. However, in the limit where $\Delta\epsilon$ is a small shift Δh of a dielectric boundary between some ϵ_1 and ϵ_2 an important class of geometric perturbation, Eq. (5) gives a surface integral of $|\vec{E}|^2$ on the interface, but this is ill-defined because the field there is discontinuous. The proper derivation of perturbation theory in the face of such discontinuity requires a more careful limiting process from an anisotropically smoothed system, yielding the surface integral:

$$\Delta\omega_{(1)}^n = -\frac{\omega_n(\vec{k})}{2} \frac{\int \int \Delta h \left(\Delta\epsilon_{12} |\vec{E}_{||}|^2 - \Delta\epsilon_{-1}^2 |D_{\perp}|^2 \right)}{\int \epsilon |\vec{E}_{n,\vec{k}}|^2}$$

where $\Delta\epsilon_{12} = \epsilon_1 - \epsilon_2$, $\Delta\epsilon_{-1}^2 = \epsilon_1^{-1} - \epsilon_2^{-1}$, and $\vec{E}_{||}/D_{\perp}$ denotes the (continuous) interface parallel/perpendicular components of the unperturbed electric/displacement eigenfield. A similar expression is required in high index-contrast systems to employ, e.g., coupled-mode theory for slowly-varying waveguides or Green’s functions for interface roughness.

Standard perturbation-theory techniques also provide expressions for higher-order corrections to the eigenvalue and eigenfield, based on an expansion in the basis of the unperturbed eigenfields. This approach, however, runs into immediate difficulty because the eigenfields are also subject to the transversality constraint, $(\vec{\nabla} + i\vec{k}) \cdot \epsilon \vec{E}_{\vec{k}} = 0$ and this constraint varies with ϵ and $\vec{k}$ – the eigenfields $\vec{E}_{\vec{k}}$ are not a complete basis for the eigenfields constrained at a different ϵ or $\vec{k}$. For ϵ perturbations, this problem can be eliminated by using the $\vec{H}$ or $\vec{D}$ eigenproblems, whose constraints are independent of ϵ . For $\vec{k}$ perturbations, one can employ a transformation by Sipe to derive a corrected higher-order perturbation theory (for e.g., the group-velocity dispersion), based on the fact that all of the non-transverse fields lie at $\omega=0$. Such completeness issues also arise applying the variational Eq. (3), as was noted in the previous section: in order for a useful variational bound to apply, one must operate in the constrained (transverse) subspace.

Two-Dimensional Photonic Crystals

After the identification of one-dimensional bandgaps, it took a full century to add a second dimension, and three years to add the third. It should therefore come as no surprise that 2d systems exhibit most of the important characteristics of photonic crystals, from nontrivial Brillouin zones to topological sensitivity to a minimum index contrast, and can also be used to demonstrate most proposed photonic-crystal devices. The key to understanding photonic crystals in two dimensions is to realize that the fields in 2d can be divided into two polarizations by symmetry: *TM* (transverse magnetic), in which the magnetic field is in the (xy) plane and the electric field is perpendicular (z); and *TE* (transverse electric), in which the electric field is in the plane and the magnetic field is perpendicular.

Corresponding to the polarizations, there are two basic topologies for 2d photonic crystals, as depicted in Fig. 4(top): high index rods surrounded by low index (top) and low-index holes in high index (bottom). Here, a hexagonal lattice is used because, as noted earlier, it gives the largest gaps. Recall that a photonic band gap requires that the electric field lines run along thin veins: thus, the rods are best suited to *TM* light ($\vec{E}$ parallel to the rods), and the holes are best suited to *TE* light ($\vec{E}$ running around the holes). This preference is reflected in the band diagrams, shown in Fig. 4, in which the rods/holes (top/bottom) have a strong *TM*/*TE* band gap. For these diagrams, the rod/hole radius is chosen to be $0.2a/0.3a$, where a is the lattice constant (the nearest-neighbor periodicity) and the high/low ϵ is $12/1$. The *TM*/*TE* bandgaps are then 47%/28% as a fraction of mid-gap frequency, but these bandgaps require a minimum ϵ contrast of $1.7/1$ and $1.9/1$, respectively. Moreover, it is conventional to give the frequencies ω in units of $2\pi c/a$, which is equivalent to a/λ (λ being the vacuum wavelength) – Maxwell’s equations are scale-invariant, and the same solutions can be applied to any wavelength simply by choosing the appropriate a . For example, the *TM* mid-gap ω in these units is 0.36, so if one wanted this to correspond to $\lambda=1.55 \mu\text{m}$ one would use $a=0.36 \cdot 1.55 \mu\text{m}=0.56 \mu\text{m}$.

The Brillouin zone (a hexagon) is shown at left-center, with the irreducible Brillouin zone shaded (following the sixfold symmetry of the crystal); the corners (high symmetry points) of this zone are given canonical names, where Γ always denotes the origin $\vec{k}=0$, K is the nearest-neighbor direction, and M is the next-nearest-neighbor direction. The Brillouin zone is a two-dimensional region of wavevectors, so the bands $\omega_n(\vec{k})$ are actually surfaces, but in practice the band extrema almost always occur along the boundaries of the irreducible zone (i.e., the high-symmetry directions). So, it is conventional to plot the bands only along these zone boundaries in order to identify the bandgap, as is done in Fig. 4.

Actually, the hole lattice can display not only a *TE* gap, but a complete photonic bandgap (for both polarizations) if the holes are sufficiently large (nearly touching). In this case, the thin veins between nearest-neighbor holes induce a *TE* gap, while the interstices between triplets of holes form ‘rod-like’ regions that support a *TM* gap overlapping the *TE* gap.

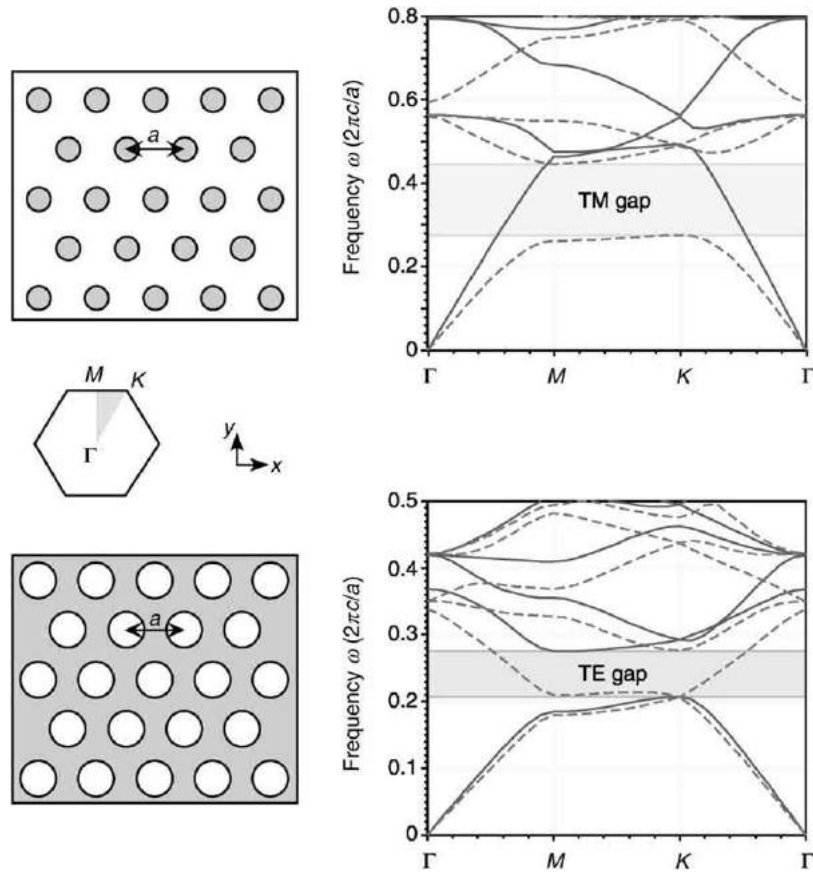


Fig. 4 Band diagrams and photonic band gaps for hexagonal lattices of high dielectric rods ($\epsilon = 12$, $r = 0.2a$) in air (top), and air holes ($r = 0.3a$) in dielectric (bottom), where a is the center-center periodicity. The frequencies are plotted around the boundary of the irreducible Brillouin zone (shaded triangle, left center), with solid/dashed lines denoting TE/TM polarization (electric field parallel/perpendicular to plane of periodicity). The rods/holes have a gap in the TM/TE bands.

Photonic-Crystal Slabs

In order to realize 2d photonic-crystal phenomena in three dimensions, the most straightforward design is to simply fabricate a 2d-periodic crystal with a finite height: a photonic-crystal slab, as depicted in Fig. 5. Such a structure can confine light vertically within the slab via index guiding, a generalization of total internal reflection – this mechanism is the source of several new tradeoffs and behaviors of slab systems compared to their 2d analogs.

The key to index guiding is the fact that the 2d periodicity implies that the 2d Bloch wavevector $\vec{k}_{\parallel}$ is a conserved quantity, so the projected band structure – all states in the bulk substrate/superstrate (the uniform regions far below/above the slab) versus their in-plane wavevector component (projected wavevector) – creates a map of which states can radiate vertically. If the slab is suspended in air, for example, then the eigensolutions of the bulk air are $\omega = c\sqrt{|\vec{k}_{\parallel}|^2 + k_{\perp}^2}$ which when plotted versus $\vec{k}_{\parallel}$ forms the continuous light cone $\omega \geq c|\vec{k}_{\parallel}|$ shown as a shaded region in Fig. 5. Because the slab has a higher $\epsilon(12)$ than the air (1), and frequency goes as $1/\sqrt{\epsilon}$ discrete guided bands are ‘pulled down’ in frequency from this continuum – these bands, lying beneath the light cone, cannot couple to any vertically radiating mode by the conservation law and so are confined to the slab (exponentially decaying away from it). If the horizontal mid-plane of the slab is a mirror symmetry plane, then just as there were TM and TE states in 2d, here there are two categories of modes: even (TE-like) and odd (TM-like) modes under reflections through the mirror plane (which are purely TE/TM in the mirror plane itself). Because the slab here is based on the 2d hole crystal, which had a TE gap, here there is a 26% ‘bandgap’ in the even modes: a range of frequencies in which there are no guided modes. It is not a complete photonic bandgap, not only because of the odd modes, but also because there are radiating (light cone) modes at every ω . The presence of these radiating modes means that if all in-plane translational symmetry is broken by a localized change in the structure, say a waveguide bend or a resonant cavity, then vertical radiation losses are inevitable; there are various strategies to minimize the losses to tolerable levels, however. On the other hand, if only one direction of translational symmetry is broken, as in a linear-defect waveguide, ideally lossless guiding can be maintained.

Photonic-crystal slabs have two new critical parameters that influence the existence of a gap. First, it must have vertical mirror symmetry in order that the gaps in the even and odd modes be treatable separately – such symmetry is broken by the presence of a

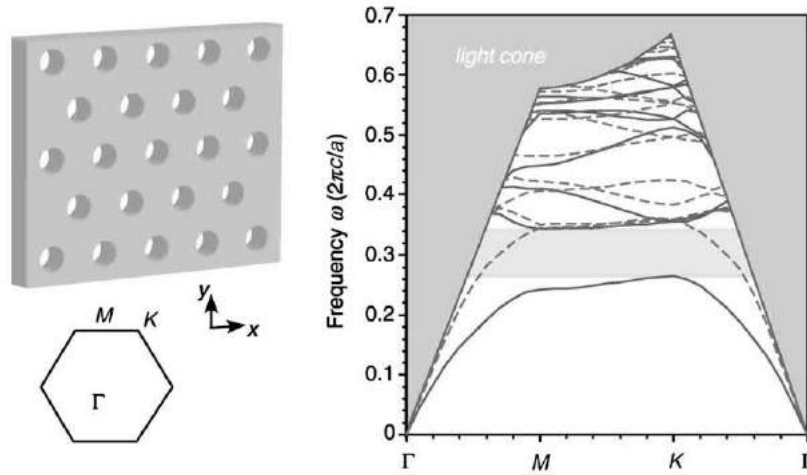


Fig. 5 Projected band diagram for a finite-thickness ($0.5a$) slab of air holes in dielectric (cross section as in Fig. 4 bottom), with the irreducible Brillouin zone at lower left. The shaded region is the light cone: the projection of all states that can radiate in the air. Solid/dashed lines denote guided modes (confined to the slab) that are even/odd with respect to the horizontal mirror plane of the slab, whose polarization is TE-like/TM-like, respectively. There is a ‘bandgap’ (region without guided modes) in the TE-like guided modes only.

substrate that is not the mirror image of the superstrate, but in practice the symmetry breaking can be weak if the index contrast is sufficiently high (so that the modes are strongly confined in the slab). Second, the height of the slab must not be too small (or the modes will be weakly confined) or too large (or higher-order modes will fill the gap); the optimum height is around half a wavelength λ/n_{eff} (relative to an average/effective index n_{eff} that depends on the polarization). In Fig. 5, a height of $0.5a$ is used, which is near the optimum (with holes of radius $0.3a$ and $\varepsilon=12$ as in the previous section).

Three-Dimensional Photonic Crystals

Photonic-crystal slabs are one way of realizing 2d photonic-crystal effects in three dimensions; an example of another way, lifting the sacrifices imposed by the light cone, is depicted in Fig. 6. This is a 3d-periodic crystal, formed by an alternating hole-slab/rod-slab sequence in an ABCABC stacking of bilayers – equivalently, it is an fcc (face-centered cubic) lattice of air cylinders in dielectric, stacked and oriented in the 111 direction, where each overlapping layer of cylinders forms a rod/hole bilayer simultaneously. Its band diagram is shown in Fig. 6 along the boundaries of its irreducible Brillouin zone (from a truncated octahedron, inset), and this structure has a $>20\%$ complete gap ($\Delta\omega$ as a fraction of mid-gap frequency) for $\varepsilon=12/1$, forbidding light propagation for all wavevectors (directions) and all polarizations. Not only can this crystal confine light perfectly in 3d, but because its layers resemble 2d rod/hole crystals, it turns out that the confined modes created by defects in these layers strongly resemble the TM/TE states created by corresponding defects in two dimensions. One can therefore use this crystal to directly transfer designs from two to three dimensions while retaining omnidirectional confinement. Its fabrication, of course, is more complex than that of photonic-crystal slabs (with a minimum ε contrast of $4/1$), but this and other 3d photonic crystal structures have been constructed even at micron (infrared) lengthscales, as described below.

There are three general dielectric topologies that have been identified to support complete 3d gaps for $\varepsilon=12/1$ (e.g., Si:air) contrast: diamond-like arrangements of high dielectric ‘atoms’ surrounded by low dielectric, which can lead to $>20\%$ gaps between the 2nd and 3rd bands; fcc ‘inverse opal’ lattices of nearly-touching low dielectric spheres (or similar) surrounded by high dielectric, giving gaps around 10% between the 9th and 10th bands; and cubic ‘scaffold’ lattices of rods along the cube edges, giving $\sim 7\%$ gaps between the 2nd and 3rd bands. It is notable that the first two topologies correspond to fcc lattices, which have the most nearly spherical Brillouin zones in accordance with the rules of thumb given above. Many variations on these topologies continue to be proposed – for example, the structure of Fig. 6 is diamond/graphite-like – mainly in conjunction with different fabrication strategies, such as the following three successful approaches. First, layer-by-layer fabrication, in which individual crystal layers (typically of constant cross-section) are deposited one-by-one and etched with a 2d pattern via standard lithographic methods (giving fine control over placement of defects, etc.); Fig. 6 can be constructed in this fashion (as well as other diamond-like structures with large gaps). Second, colloidal self-assembly, in which small dielectric spheres in a fluid automatically arrange themselves into close-packed (fcc) crystals by surface forces – these crystals can be back-filled with a high-index material, out of which the original spheres are dissolved in order to form inverse-opal crystals with a complete gap. Third, holographic lithography, in which a variety of 3d crystals can be formed by an interference pattern of four laser beams to harden a light-sensitive resin (which is then back-filled and dissolved, as with colloids, to achieve the requisite index contrast). The second and third techniques are notable for their ability to construct large-scale 3d crystals (thousands of periods) in a short time.

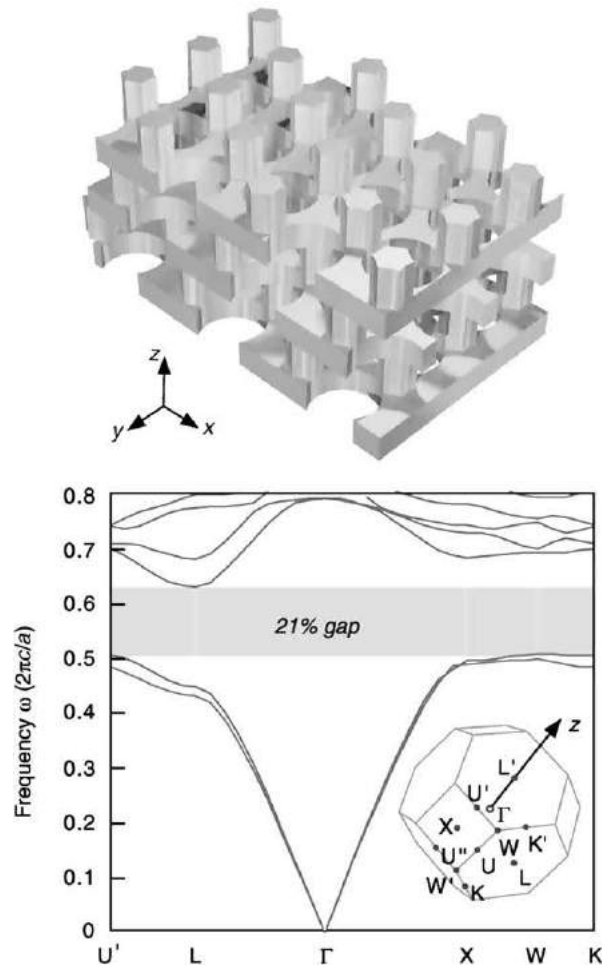


Fig. 6 Band diagram (bottom) for 3d-periodic photonic crystal (top) consisting of an alternating stack of rod and hole 2d-periodic slabs (similar to Fig. 4), with the corners of the irreducible Brillouin zone labeled in the inset. This structure exhibits a $\Delta\omega/\omega_{\text{midgap}} = 21\%$ omnidirectional bandgap.

Further Reading

- Ashcroft, N.W., Mermin, N.D., 1976. *Solid State Physics*. Philadelphia: Holt Saunders.
- Cohen-Tannoudji, C., Diu, B., Laloë, F., 1977. *Quantum Mechanics*. Paris: Hermann.
- Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.
- Johnson, S.G., Joannopoulos, J.D., . *Photonic Crystals: The Road from Theory to Practice*. Boston: Kluwer.
- Johnson, S.G., Ibanescu, M., Skorobogatiy, M., Weisberg, O., Joannopoulos, J.D., Fink, Y., 2002. Perturbation theory for Maxwell's equations with shifting material boundaries. *Physical Review E* 65, 066611.
- Sakoda, K., 2001. *Optical Properties of Photonic Crystals*. Berlin: Springer.
- Sipe, J.E., 2000. Vector $k \cdot p$ approach for photonic band structures. *Physical Review E* 62, 5672–5677.

Nonlinear Optics in Photonic Crystal Fibers

JE Sharping, Cornell University, Ithaca, NY, USA

P Kumar, Northwestern University, Evanston, IL, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Photonic crystal fibers (PCFs) are very similar to normal optical fibers in that they consist of a core surrounded by cladding, such that light is guided within the core of the fiber. The primary difference between PCF and standard optical fibers is that PCFs feature an air-silica cross-section, whereas standard optical fibers have an all-glass cross-section. An electron micrograph of a typical PCF is shown in Fig. 1. The air holes extend along the axis of the fiber for its entire length and the core of the fiber is formed by a defect, or missing hole, in the periodic structure. The core is formed of solid glass, whose refractive index is that of pure silica (or whatever other glass is chosen), and the cladding is formed by the air-glass mixture, whose effective refractive index depends on the ratio of air-to-glass, also known as the air-fill fraction, that comprises the structure. The resulting effective-index of the cladding will be lower compared with that of the core and, as such, will provide the refractive index variation necessary to support total internal reflection at the core-cladding boundary, and guide light in a manner similar to that of standard optical fibers. The fiber design (i.e., size, shape, and the air-fill fraction) dictates solutions to Maxwell's equations for light propagating within the fiber. Valid solutions are referred to as 'modes' which propagate along the fiber in a known manner, and have a well-defined shape in the transverse direction (i.e., they have a well-defined transverse mode structure).

Nonlinear-optical effects in fibers result from the interaction of optical fields with the glass via the $\chi^{(3)}$, or Kerr nonlinearity. The phenomenon of nonlinear refractive index is a manifestation of a light-material interaction mediated by $\chi^{(3)}$. The magnitudes of the components of the third-order susceptibility tensor in glass, $\chi^{(3)}$, are generally quite small compared with the analogous second-order $\chi^{(2)}$ terms for materials exhibiting such nonlinearities (e.g., lithium niobate, beta-barium borate (BBO), etc.). The relatively small $\chi^{(3)}$ nonlinearity in optical fibers makes them ideal for wavelength-division multiplexed optical communication where light propagation subject to a minimum of nonlinear effects is critical. Nonlinearity does, however, eventually become an issue in wavelength-division multiplexed systems as the launched optical power increases and as the channel spacing decreases. On the other hand, one can utilize nonlinear-optical effects in soliton communication systems and to build useful photonic devices. Despite the weak $\chi^{(3)}$ nonlinearity, the net nonlinear-optical effect in fibers can be large due to the ability to tightly confine intense fields within the core of an optical fiber and maintain the interaction over a long distance as the guided fields propagate through the fiber.

The study of nonlinear-fiber optics has benefited from dramatic improvements in optical fiber and fiber-optic device fabrication. The importance of understanding nonlinear-fiber optics is driven by the need to develop fiber-integrated devices, and also by the need to understand and mitigate the problems that these nonlinearities cause in optical communication systems.

This section introduces the unique linear- and nonlinear-optical properties of PCFs in order to understand the reasons why nonlinear-optical effects are often enhanced in such fibers. These discussions pertain to PCFs which are 'highly nonlinear'.

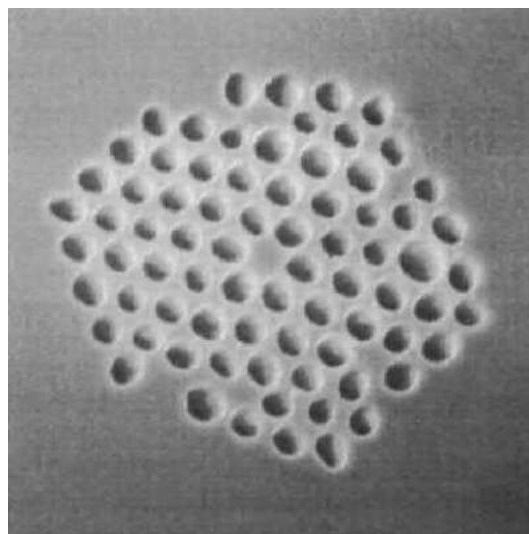


Fig. 1 An electron micrograph showing the periodic microstructure of a typical PCF. The core is formed by the 'missing hole' in the center of the microstructure. Reproduced with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America, courtesy of OFS.

It is essential to clarify that ‘highly nonlinear’ in this context does not mean that the $\chi^{(3)}$ is any larger than that of standard telecommunication fibers, rather that the effect of this nonlinearity is enhanced due to the fiber’s very small core.

PCF Properties

Photonic crystal fibers feature a variety of interesting properties. From the standpoint of nonlinear-fiber optics there are four very useful fundamental properties of PCFs:

- a mechanically robust optical fiber can be fabricated with an extremely small core (a few μm^2);
- a fiber can be made to guide in a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- there are new degrees of freedom that allow one to manipulate the fiber’s group-velocity dispersion (GVD) properties; and
- many, but not all, PCFs are polarization maintaining as a result of form birefringence present in the core.

The fact that small-core PCFs can be fabricated is clear from Fig. 1 by taking note of the fact that the center defect region which comprises the core is about 1.7 μm in diameter. Photonic crystal fibers with even smaller cores have been fabricated.

Transverse Mode Structure

A widely accepted model used to describe the transverse modal behavior of PCFs is called the effective-index model. The effective-index model can be used to understand why some PCFs are ‘endlessly single mode’, meaning that the fiber guides in a single transverse mode over an exceptionally wide wavelength range (370 nm–1600 nm). In the effective-index model, the refractive index of the core, $n_{\text{co}}(\lambda)$, is that of glass, and the refractive index of the cladding, $n_{\text{eff}}(\lambda)$, assumes a value in between that of glass and air. In the context of PCFs, one makes a modification to the standard expression describing single-mode behavior in step-index fibers:

$$V_{\text{eff}} = \frac{2\pi\Lambda}{\lambda} \sqrt{n_{\text{co}}(\lambda)^2 - n_{\text{eff}}(\lambda)^2} < V_{\text{cutoff}} \quad (1)$$

where Λ is the spacing between air holes, λ is the wavelength of light, and V_{cutoff} is the cutoff condition for the PCF. A similar expression for the V parameter is commonly used to understand the modal behavior of standard fibers where the larger the V is, the more transverse modes are supported within the fiber. In standard fibers the cutoff condition below which only a single mode can propagate within the core of a fiber is given by $V_{\text{cutoff}} < 2.405$. In the case of PCFs, a numerical method should be used to determine V_{cutoff} . Mechanically robust PCFs can be fabricated where the dispersion in $n_{\text{eff}}(\lambda)$ (i.e., the variation of n_{eff} with λ) offsets dispersion in $n_{\text{co}}(\lambda)$ and compensates for the $2\pi\Lambda/\lambda$ coefficient in Eq. (1). Therefore, the light within the fiber propagates in a single, Gaussian-like mode because for all wavelengths $V_{\text{eff}} < V_{\text{cutoff}}$. A graph of the variation of V_{eff} with Λ/λ is shown in Fig. 2, where d represents the size of an air hole.

Conceptually, the effective index model can be understood by noting that at short wavelengths the mode field is confined well within the all-silica core, but as λ increases the mode field extends further into the air–glass cladding and V_{eff} and $n_{\text{eff}}(\lambda)$ both decrease.

Dispersion in PCF

Other critical differences between PCFs and standard optical fibers lie in the dispersion properties. When light propagates through a fiber its behavior depends on the light’s optical frequency:

$$E(t, z) = A(t, z) e^{i[\omega t - \beta(\omega)z]} \quad (2)$$

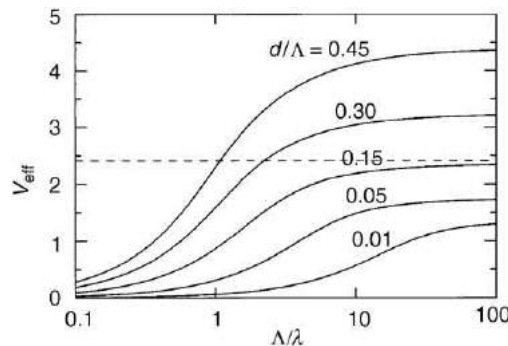


Fig. 2 Variation of V_{eff} for different relative hole diameters d/Λ . The calculation assumes a fiber with an air–glass cross-section where the refractive index of air and glass was taken to be 1 and 1.45, respectively. The dashed line marks $V_{\text{eff}} = 2.405$, the cutoff value for a step-index fiber. Reproduced with permission from Birks TA, Knight JC and Russell PSTJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963. ©1997 Optical Society of America.

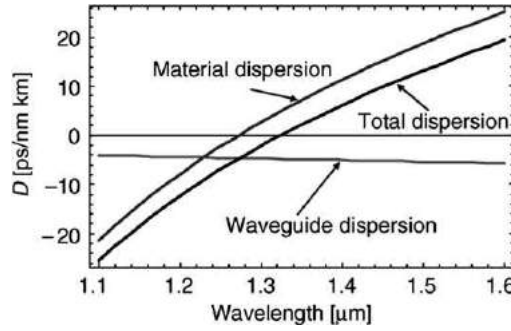


Fig. 3 Plots of the theoretical dispersion coefficient, D as a function of wavelength for a standard optical communication fiber.

Eq. (2) describes the mode as it propagates through the fiber. It is decomposed into a slowly varying envelope, $A(t, z)$, and a rapidly varying exponential component where ω is the frequency of the mode, t is time, z is the position along the length of the fiber, and $\beta(\omega)$ is called the mode-propagation constant. The general term used in describing the frequency dependence of β is chromatic dispersion, which includes contributions from the material as well as the waveguide. Other types of dispersion present in optical fibers include multi-modal (arising from multiple guided transverse modes) and polarization-mode dispersion.

One way to understand the chromatic dispersion of a mode propagating through an optical fiber is to study the Taylor series expansion of the mode-propagation constant, β , about the center frequency of the field, ω_0 :

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + \dots \quad (3)$$

where $\beta_i = d^i\beta/d\omega^i$. The physical significance of the various β_i in Eq. (3) are as follows: the phase-fronts of the electric field move at a speed given by $\omega/\beta_0 = v_p$, and the envelope, $A(t, z)$, moves at a group velocity given by $1/\beta_1$. A GVD term, which governs temporal spreading of the envelope, is given by β_2 . Higher-order β terms are usually negligible for propagation of pulses of ≥ 1 ps duration in optical fibers and are lumped into the category of 'higher-order chromatic dispersion'.

The notation ' β_2 ', as defined above, is often used in the literature with dimensions of ps^2/km . However, another expression is frequently used because of its direct relationship to measured quantities. It is straightforward to measure the relative delay, T , between two pulses having different center wavelengths. Choosing a particular wavelength as a reference, one can then measure relative delay as a function of an injected pulse's center wavelength. The first derivative with respect to λ of the relative delay curve gives the GVD according to

$$D = \frac{\partial\left(\frac{1}{v_g}\right)}{\partial\lambda} = \frac{1}{L} \frac{dT}{d\lambda} = -\frac{2\pi c}{\lambda^2} \beta_2 \quad (4)$$

where v_g is the group velocity, and L is the length of the fiber under test. The dimension commonly used for D is $\text{ps}/(\text{nm km})$.

Chromatic dispersion in single-mode optical fibers results from two different wavelength-dependent fiber parameters. The medium itself, glass in this case, has a wavelength-dependent refractive index. This 'material' contribution has the same magnitude regardless of the various parameters associated with the waveguide. A second contribution has to do with the design of the optical fiber. This 'waveguide' contribution to dispersion arises from the fact that the wavelength-dependent mode depends on the properties of the waveguide (i.e., the core size and refractive index contrast between the core and cladding). Empirical models can be used to describe the material, waveguide, and total GVD for standard communication fibers. Such a set of curves is given in Fig. 3 where it can be seen that it is possible to have positive, negative, or zero values for D . For historical reasons, the regions where D is negative (β_2 is positive) exhibit 'normal GVD', while those where D is positive (β_2 is negative), exhibit 'anomalous GVD'. The wavelength corresponding to $D=0$ is referred to as the zero-dispersion wavelength (λ_0), which for most silica glass fibers is about 1,300 nm.

In contrast with standard optical fibers, where the waveguide contribution to D is always less than zero, small-core PCFs can be fabricated where the waveguide contribution to GVD is positive and quite large. As such, in PCFs, λ_0 can be shifted to wavelengths shorter than the intrinsic dispersion zero of glass. Control over the GVD is essential for phase matching certain nonlinear-optical interactions involving light of different colors co-propagating within a fiber. Indeed, several exciting applications of nonlinear optics in PCF require a fiber with a $\lambda_0 \sim 800$ nm. The GVD is also of great importance when working with pulsed light in PCF, because GVD results in temporal pulse broadening. It also governs pulse temporal walkoff effects, limiting the effective interaction length between pulses of different colors. This new flexibility to manipulate the GVD curve, by varying waveguide design parameters, is a key advantage associated with using PCFs for nonlinear optics.

Birefringence in PCF

The core of an optical fiber often exhibits some amount of anisotropy. The core may be elliptical in form (shape) which leads to a phenomenon referred to as form birefringence. Since mode propagation depends on the fiber structure, a fiber with an elliptical

core will exhibit mode propagation that depends on the electric field's polarization with respect to the axes of the elliptical core. As a result of birefringence, the polarization of the mode varies as it propagates through the fiber (unless care is taken to align the polarization of the injected light with respect to a principal axis of the birefringence).

Polarization-maintaining (PM) fibers are designed to include birefringence in a particular axis of the fiber. By including a well-defined birefringence throughout the length of a fiber that is larger than that induced by external perturbations, fast and slow axes of the optical fiber are created for all guided wavelengths, giving two orthogonal 'polarization modes'. If light is injected into one of the polarization modes (i.e., with its linear polarization along one of the axes) it remains linearly polarized along that axis as it propagates along the fiber. The two polarization modes generally have different group velocities, so pulsed light in each mode will take a different amount of time to propagate through a given segment of fiber. Most PCFs exhibit strong birefringence due to a slightly elliptical core combined with a large core-cladding index difference, and so they behave similarly to PM fibers. Special care must be taken when working with PCF to be sure that the polarization of the light launched into the fiber is aligned with one of the birefringent axes.

In practice, there are a few other features of PCF that are of importance when discussing nonlinear-optic interactions:

- propagation losses are generally larger in PCFs than in standard optical fibers; and
- free-space coupling and splicing are difficult and can result in large coupling losses.

Nonlinear Phenomena

The basic principles determining nonlinear effects in PCFs are the same as those for standard optical fibers. It is the new flexibility in PCFs to obtain transverse-modal and GVD behavior different from that of standard optical fibers that makes PCFs truly interesting for nonlinear optics. The relevant nonlinear-optical effects are: self-phase modulation (SPM); cross-phase modulation (CPM); third-harmonic generation (3HG); four-wave mixing (FWM); Raman scattering; and Brillouin scattering.

Self-phase modulation (also known as the optical Kerr effect) refers to the self-induced phase shift experienced by an optical field as it propagates through a fiber. It becomes particularly important for the case of pulses of light propagating through optical fibers. In small core PCFs, SPM is enhanced due to the high-intensity light propagating within the core. Self-phase modulation can lead to substantial spectral broadening of pulsed light propagating along an optical fiber.

When a pulse of light experiences normal GVD (i.e., $D < 0$) as it propagates, the longer-wavelength components travel faster than the shorter-wavelength components. Anomalous GVD (i.e., $D > 0$) leads to the opposite, short-wavelength components traveling faster than the long-wavelength components. Group-velocity dispersion generally leads to temporal broadening of pulses as they propagate along a fiber. Under ideal conditions, however, SPM in combination with anomalous GVD, leads to pulses which propagate without any temporal or spectral broadening. These self-sustaining pulses are called 'optical solitons'.

When waves of light having different wavelengths co-propagate along a fiber, CPM can occur. It can be understood as a phase shift induced on one wave, due to the presence of the other wave. Cross-phase modulation also leads to spectral broadening and solitonic pulse propagation.

In 3HG and FWM, one or more photons are destroyed and others are created. In 3HG, three 'fundamental' photons are destroyed to create one with three times the energy of the fundamental photons. In FWM, two fundamental photons are destroyed while two others are created. While it is straightforward to conserve energy in 3HG and FWM, these interactions must be 'phase-matched', meaning that the interacting waves must be made to propagate in-phase over a meaningful length. Such phase-matching conditions need to be carefully considered when studying 3HG and FWM. Nevertheless, 3HG and FWM can be used to obtain frequency shifts and all-optical amplification. In comparison with other types of optical fibers, PCFs are particularly useful for 3HG and FWM applications. The small core of the PCF allows interactions to occur at much lower input powers, and the new flexibility associated with the GVD properties permits phase matching in cases which are not possible using standard optical fibers. Finally, the fiber's endlessly single-mode behavior permits very good transverse mode overlap between interacting waves having widely different center wavelengths.

Self-phase modulation, CPM, 3HG, and FWM are photon-photon interactions wherein no energy is exchanged with the medium itself. In contrast, Raman scattering and Brillouin scattering result from photon-phonon interactions. The differences between Raman and Brillouin scatterings lie in the energy of the phonons involved and the direction in which the interactions occur. Raman scattering is an interaction between a photon and an optical-phonon mode of the molecules making up the material. In the case of Raman scattering in glass, the energy shift, associated with molecular vibrational (Raman) modes, corresponds to frequencies of 1–12 THz. Raman scattering occurs in the forward and backward directions. With pulsed light, stimulated Raman scattering can occur when the lower-frequency spectrum of the pulse overlaps with the spectrum of the Raman resonances excited by the higher-frequency spectrum. When this happens, energy can be efficiently shifted in spectrum towards the peak of the Raman resonance. In the forward direction, this 'Raman self-frequency shift' builds up. Additionally, if the Raman self-frequency shift occurs in the presence of anomalous dispersion, a 'Raman soliton self-frequency shift' can result. As the injected power is increased, the spectral shift between the injected pulse and the resulting Raman soliton increases. The principal advantage associated with PCF is the ability to generate Raman solitons for a broader range of wavelengths than was previously possible.

For Brillouin scattering, interaction with the acoustical phonons results in frequency shifts of about 10 GHz and the interaction only occurs in the backward direction. Brillouin scattering is generally a nuisance in fiber-based devices, leading to intensity noise

and other problems. The interesting feature of Brillouin scattering in PCFs is that the threshold intensity where problems begin to occur is higher for PCFs than for standard optical fibers. The higher threshold permits further optimization of fiber-based devices wherein Brillouin scattering limits the performance.

Experiment Examples

In the following subsections, a selection of experiments, demonstrating a few of the relevant nonlinear-optical phenomena, are briefly described.

Supercontinuum Generation

One of the most exciting demonstrations of nonlinear optics in PCF is that of supercontinuum generation. In a typical experiment, 100 femtosecond pulses from a mode-locked Ti:Sapphire laser operating at a wavelength of 800 nm were injected into the PCF. As the injected power was increased, a broad continuum of spectrum was generated from wavelengths of 400 nm up to 1,600 nm. Typical data showing the input and output optical spectra are given in Fig. 4, where it can be seen that well over an octave of frequency spectrum is generated from an input pulse whose spectral width is only about 10 nm.

It is widely accepted that supercontinuum generation results from a combination of linear and nonlinear optical effects conspiring to generate the broad spectrum. As the pulse propagates through the PCF, SPM, FWM, and Raman scattering are all likely to occur with relative efficiencies depending heavily on the input pulses spectral and temporal characteristics, as well as the fibers properties. Indeed, efficient supercontinuum generation can occur in PCFs for pump wavelengths lying near the zero-dispersion point in the normal or anomalous dispersion regime of the PCF, and for fibers as short as a few millimeters in length.

Spectral broadening, due to SPM, is common in standard optical fibers, but what makes this particular experiment interesting is the remarkable width of spectrum generated with a short piece of fiber (~ 1 m) and with comparatively small optical powers.

Optical Switching in PCF

Cross-phase modulation can be used to build an all-optical switch. Such a switch can be implemented as a three-port device where the output port for a given optical bit (the signal pulse), is determined by the presence of an optical control (the pump pulse). Switching can then be achieved by dividing the signal pulse equally on two arms of an interferometer and injecting the strong pump pulse only on one arm. Because the pump pulse co-propagates with only one of the two signal pulses, there exists a CPM-induced phase difference $\phi_{NL} = 2\gamma P_p L$ at the output of the interferometer, where P_p is the peak power of the pump pulse and L is the interaction length. By varying the intensity of the pump pulses one can vary the magnitude of this phase difference. If a π -phase shift is achieved, one can switch the interference from destructive to constructive, or vice-versa, thus realizing an all-optical switch.

Fig. 5 shows an experimental setup used to observe switching near 1,550 nm (a similar apparatus using bulk-optic rather than fiber-optic components can be used to conduct experiments near 780 nm). The pump and signal are synchronous few-ps-duration pulses with a tunable wavelength separation of 5–15 nm. The switching characteristics for this implementation are shown in Fig. 6. The apparatus has the advantage of requiring short fiber lengths, low switching powers, and allows switching of weak pulses. It demonstrates the feasibility of using nonlinear optics in PCF to perform essential functions in high-speed all-optical processing.

Parametric (Mixing) Processes

The first set of experiments with controlled FWM in a PCF achieved nondegenerate parametric gains over a 30 nm range of pump wavelengths near the λ_0 of the PCF. The experiments also confirmed the wavelength dependence of the GVD coefficient of the PCF

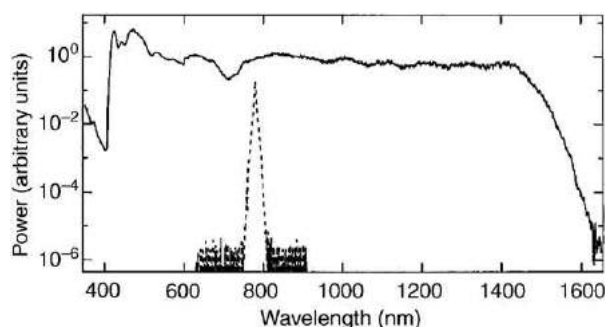


Fig. 4 In supercontinuum generation one observes a broad continuum generated after short pulses of light from a Ti:Sapphire laser propagate through a 75 cm section of PCF. The spectrum of the input pulse is shown as a dashed curve, while the output is a solid curve. Reproduced with permission from Ranka JK, Windeler RS and Strentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America.

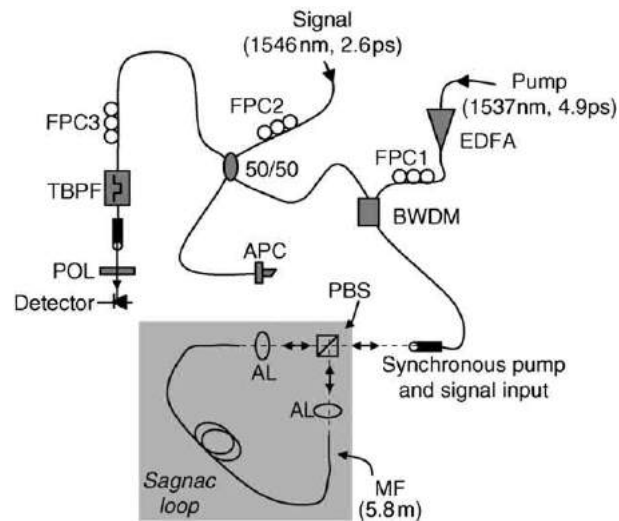


Fig. 5 Experimental setup used to demonstrate all-optical switching near 1,550 nm. (EDFA, erbium-doped fiber amplifier; BWDM, bandpass wavelength-division multiplexor; PBS, polarization beamsplitter; FPC, fiber polarization controller). Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.

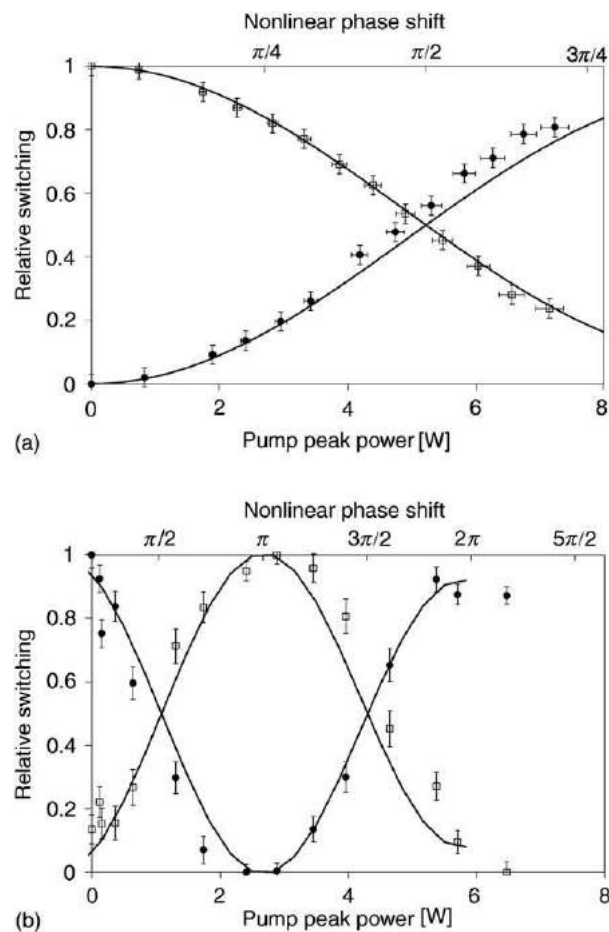


Fig. 6 Switching curves (open boxes and filled circles), showing the relative power measured in each port of the switch vs. the pump peak power, for experiments conducted near (a) 1,550 nm and (b) 780 nm. The curves accompanying the data are generated from numerical solutions of coupled wave equations for CPM. Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.

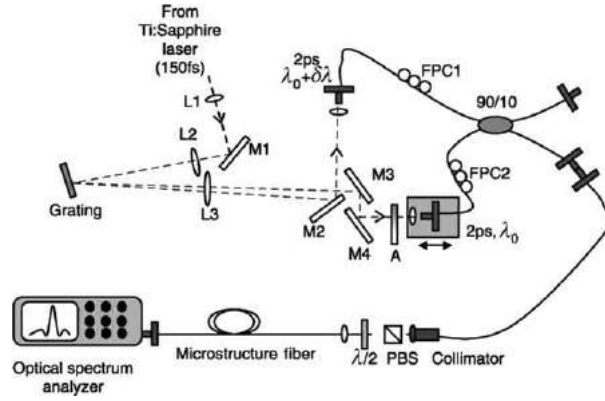


Fig. 7 A schematic of the experimental setup used to investigate FWM in PCFs. Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.

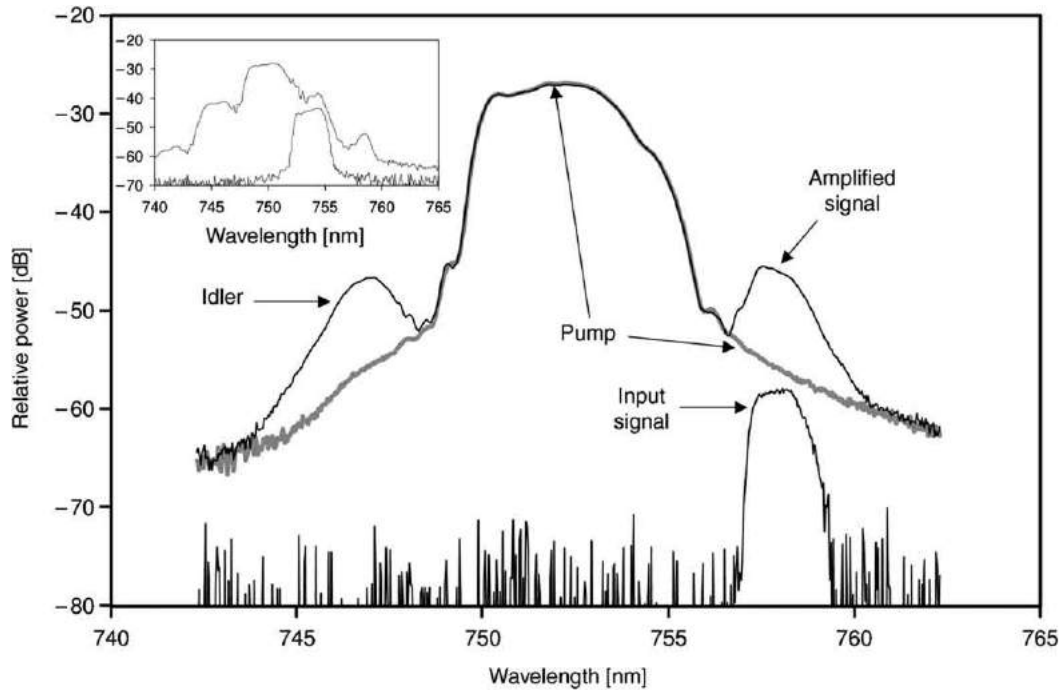


Fig. 8 A typical FWM spectrum observed at the output of the microstructure fiber. The inset shows a spectrum where higher-order cascaded mixing is evident. Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.

near λ_0 . Since the dispersion characteristics of these fibers can be adjusted during the fabrication process, the experiments demonstrate the potential for the use of PCFs in broadband parametric amplifiers, wavelength shifters, and other optical communication devices.

The experimental setup used to demonstrate phase-matched FWM in PCF is shown in Fig. 7. The pump and the input signal are two synchronous pulsed beams having 3–5 nm wavelength separation with the center wavelength tunable over a 720–850 nm range. The maximum peak power of the pump pulses is ≈ 12 W. The two synchronous beams are then combined and injected into the PCF. The pump and signal's optical paths are adjusted to obtain temporal overlap in the PCF and their polarizations are aligned by fiber polarization controllers.

Fig. 8 shows a typical FWM optical spectrum at the output of a 6.1 m long PCF. Here the strong pump beam and the weak signal beam have wavelengths of 753 nm and 758 nm, respectively. The spectrum shows the undepleted pump, the amplified signal, and the generated idler at 747 nm. The spectra in Fig. 8 show that large gain is achievable for a pump-to-signal spacing of 5 nm. Gain values of more than 20 (13 dB) were obtained.

Conclusion

In summary, the advantages of using photonic crystal fibers for demonstrating nonlinear-fiber optical effects arise from four novel properties:

- the nonlinear coefficient is enhanced in small-core PCFs (core area of a few μm^2);
- PCFs can support a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- PCF design parameters allow one to manipulate the fiber's GVD properties; and
- nonlinear interactions are enhanced due to the polarization maintaining properties of PCFs which result from form birefringence present in the core.

These four properties combine to allow efficient interactions to occur in PCFs which are either inefficient or not possible at all in standard optical fibers.

See also: Microstructure Fibers

Further Reading

- Agrawal, G.P., 2000. Nonlinear Fiber Optics, 3rd edn. San Diego, CA: Academic Press.
- Birks, T.A., Knight, J.C., Russell, P.S.J., 1997. Endlessly single-mode photonic crystal fiber. *Optical Letters* 22, 961–963.
- Hansryd, J., Andrekson, P.K., Westlund, M., Li, J., Hedekvist, P., 2002. Fiber-based optical parametric amplifiers and their applications. *IEEE Journal of Selected Topics in Quantum Electronics* 8, 506–520.
- Monro, T.M., Richardson, D.J., Broderick, N.G.R., Bennett, P.J., 2000. Modeling large air fraction holey optical fibers. *Journal of Lightwave Technology* 18, 50–56.
- Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25, 25–27.
- Russell, P.St.J., 2003. Photonic crystal fibers. *Science* 299, 358–362.
- Sharping, J.E., Fiorentino, M., Coker, A., Kumar, P., Windeler, R.S., 2001. Four-wave mixing in microstructure fiber. *Optical Letters* 26, 1048–1050.
- Sharping, J.E., Fiorentino, M., Kumar, P., Windeler, R.S., 2002. All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14, 77–79.

Photonic Crystal Lasers, Cavities and Waveguides

J O'Brien and W Kuang, University of Southern California, Los Angeles, CA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Photonic bandstructure engineering, in which the electromagnetic dispersion relations are intentionally modified, can be accomplished by creating one-, two-, and three-dimensionally periodic dielectric materials. These periodic materials, called photonic crystals, can be used to confine and guide light. This article describes the formation of optical resonant cavities and waveguides in these materials. Most of the discussion will focus on two-dimensional photonic crystals. One-dimensionally periodic systems are treated elsewhere in this volume. This article will begin with a short review of the relevant electromagnetic properties of photonic crystals. The formation and properties of resonant cavities will follow this, and the article will conclude with waveguides formed by linear defects in two-dimensional photonic crystals.

Electromagnetics of Photonic Crystals

A two-dimensional photonic crystal consists of a dielectric material in which the dielectric constant is periodic in two dimensions. The period of these materials is on the order of a half-wavelength of the operating optical wavelength. In this section, we briefly review the electromagnetic properties of two-dimensional photonic crystals. For more details, see the article entitled **Spectroscopy: Raman Spectroscopy** in this volume.

There are five Bravais lattices in two dimensions. Most of the research on photonic crystal resonant cavities and waveguides has focused on square and triangular lattices, and in this article we will consider examples based on the triangular lattice. Since these structures are usually defined lithographically however, there is, in principle, no need to confine the investigation to naturally occurring Bravais lattices. Due to the limitations of current nanofabrication technology, much of the research has focused on photonic crystals that have a finite thickness. The important feature of photonic crystals is that the bandgaps in the electromagnetic spectrum opened up by all of the Bragg planes overlap spatially and spectrally, so that a frequency region exists in which no electromagnetic propagation is allowed. These electromagnetic bandgaps can be used to confine optical modes in very small-volume resonant cavities and in waveguides. In high-contrast dielectric systems, such as a semiconductor/air periodic system, only a few lattice periods may be necessary to confine an electromagnetic mode. In the case of finite thickness photonic crystals, this periodicity has the effect of limiting the bandgap formed to the in-plane directions and in the case of a two-dimensional photonic crystal in a high-index dielectric slab to the formation of a bandgap in the guided modes of the slab. In this case there are still radiation modes of the slab.

Consider a two-dimensional photonic crystal that is periodic in the $x - y$ plane. To start, we can consider materials that are uniform and infinitely extended in z . Electromagnetic fields propagating in the $x - y$ plane are Bloch states. These can be written as:

$$\vec{H}_{\vec{k}_z, \vec{k}_\parallel}(\vec{\rho}, z, t) = e^{i(\omega t - (\vec{k}_\parallel \cdot \vec{\rho} + k_z z))} \vec{u}_{\vec{k}_z, \vec{k}_\parallel}(\vec{\rho}) \quad (1)$$

where $\vec{\rho}$ labels the in-plane coordinates and $\vec{k}_\parallel$ labels the in-plane wavevectors. $\vec{u}_{\vec{k}_z, \vec{k}_\parallel}$ is a periodic function in the $x - y$ plane. Modes propagating in the $x - y$ plane in such a system can be classified as transverse electric (TE) waves or transverse magnetic (TM) waves. TE waves have nonzero E_x , E_y , and H_z field components, and TM waves have nonzero E_z , H_x , and H_y field components. A triangular lattice of air holes patterned into a high refractive index dielectric can be used to create a bandgap for both the guided TE and TM modes of the membrane for a range of hole radii. The TE and TM bandgaps that are formed can be over different spectral ranges, however. For laser applications this is unimportant since in practice the emission occurs from electron-hole recombination in a semiconductor quantum well, and for unstrained or compressively strained quantum well materials this emission is TE polarized. In most cases, waveguides have also been designed to work for a single optical polarization. It is generally true that a photonic lattice that consists of a connected high dielectric region is likely to exhibit a TE bandgap, while a lattice formed by disconnected high dielectric regions is more likely to exhibit a TM bandgap. The bandgap in the TE modes is formed between the first and second bands. The first band is called the dielectric band because the field at the Brillouin zone boundary is a standing wave with its intensity concentrated in the high dielectric regions. The second band is called the air band because at the Brillouin zone boundary this field is a standing wave with its intensity localized in the low dielectric regions. Over a reasonable range of lattice parameters, the bandwidth of the TE bandgap can be changed by changing r/a where r is the radius of the holes and a is the lattice constant of the triangular lattice. As r/a gets larger the dielectric band moves up in frequency. This can be thought of as being due to the fact that this mode has a decreasing effective index as r/a increases. The air band frequency increases as r/a increases. Since more of the electric field of the air band is located in the low dielectric regions than the field of the dielectric band, the air band increases in frequency with increasing r/a faster than the dielectric band and the bandgap therefore increases with increasing r/a . This trend holds in photonic crystals of finite thickness as well.

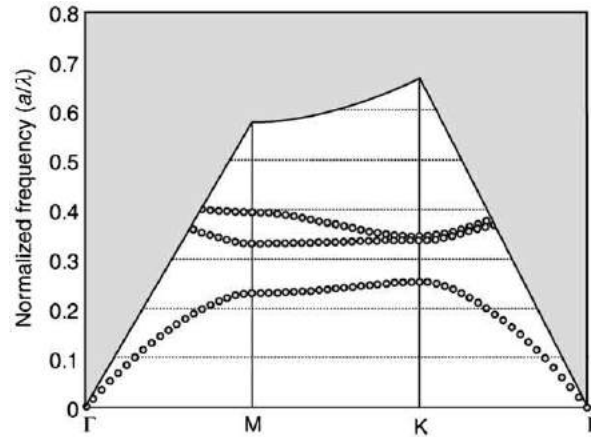


Fig. 1 The photonic band diagram for a photonic crystal slab with a triangular lattice of air holes perforating a high-index dielectric slab. The slab, which assumes a refractive index $n=3.4$, is suspended in the air with a thickness of $d/a=0.6$. The air holes have radii of $r/a=0.3$, where a is the lattice constant. Only the lowest three eigenmodes are shown in the plot.

For photonic crystals of a finite thickness, the ability to simply classify modes as either TE or TM is lost. In the most simple finite thickness photonic crystals, that of a high index dielectric slab in which a two-dimensional photonic crystal has been patterned surrounded above and below by the same low index material which may be air or sapphire, or silicon dioxide, the modes can be classified as being either even or odd modes with respect to the mid-plane of the high index slab. In this mid-plane, however, modes can be classified as being either TE or TM, with each containing the same nonzero vector field components as in the infinite case. Away from this mid-plane, however, modes have in general six nonzero electromagnetic field components. The photonic bandgap that is created in this case is in the guided mode spectrum of the high index dielectric slab. There is no gap formed in the radiation modes of the slab. The radiation modes consist of modes with a ray vector that is not totally internally reflected at the high index slab/low index cladding interfaces. Fig. 1 shows the dispersion relation, plotted over the irreducible Brillouin zone, of the even modes for dielectric slab with a refractive index of 3.4 in which a two-dimensional triangular lattice of holes has been patterned. The surrounding material is air. The figure shows an electromagnetic bandgap formed between the first and the second guided modes of the slab. Radiation modes of the slab have real valued propagation vectors in the z -direction. These modes lie above the in-plane dispersion relation for air and exist inside the shaded region in the figure. This dividing line between the guided modes and the radiation modes is referred to as the light line.

If the upper and lower cladding layers of the high index slab are different and therefore have different indices of refraction, then the ability to classify modes as being either even or odd about the mid-plane of the slab disappears and it is possible to lose the bandgap in the guided modes altogether. In cases where the refractive index of the bottom cladding is not significantly different from that of the top cladding, such as a case where air serves as the top cladding layer and a material such as sapphire or silicon dioxide forms the bottom cladding, it is found that the even and odd modes describe the field reasonably accurately. To emphasize the approximation involved in the model, these modes are most often referred to in the literature as even-like or odd-like. Overall, there are three primary effects of having asymmetric cladding layers such as an air top cladding and sapphire bottom cladding case. These are a reduction in the effective bandgap width, an increase in the area of the radiation modes on the dispersion relation, and the loss of a rigorous bandgap in the guided modes of the slab. The first two of these effects are the most serious for device designers.

Photonic Crystal Resonant Cavities and Lasers

Photonic crystal resonant cavities can be formed at frequencies inside the bandgap or at the bandedges. Modes in the bandgap are formed as a result of a defect in the lattice. Modes at the bandedge are also sometimes used because the fields with wavevectors at Brillouin zone boundary are standing waves. These standing waves are analogous to the resonant modes that are formed in distributed feedback (DFB) laser structures. Here we will focus on modes confined in the bandgap. Allowed modes in electromagnetic bandgaps in photonic crystal slabs can be introduced by introducing one or more defects into the lattice. By perturbing a single lattice site, we can permit a localized mode or modes that have a frequency in the gap. This phenomenon is similar to the formation of deep levels in an electronic bandgap of a solid due to impurities and identical to the formation of resonant modes in vertical cavity surface emitting lasers (VCSELs) and phase-shifted DFB lasers. In each of these last cases, a defect in the periodicity of a precise thickness occurs between two one-dimensional distributed Bragg reflectors. Defect modes in photonic crystal resonant cavities can be engineered to radiate in a particular direction by trading-off the in-plane versus out-of-plane losses, and the resonant frequency is determined by the parameters of the lattice. Very small mode optical mode volumes can also be obtained in these resonant cavities leading to the possibility of modifying spontaneous emission.

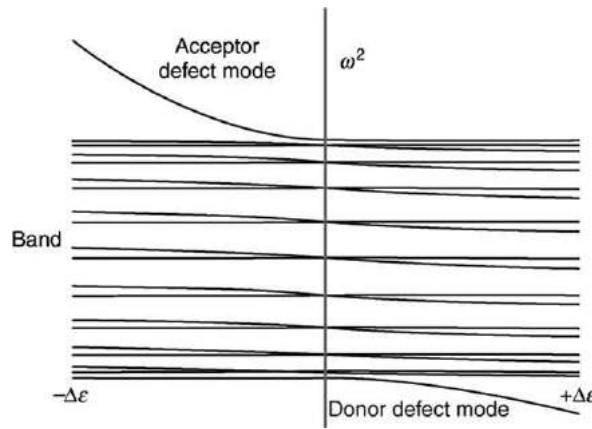


Fig. 2 A one-dimensional Slater-Koster model for a single defect in a 10-unit-cell chain, showing the allowed states as a function of the defect perturbation.

Simple models exist which illustrate the formation of localized modes inside the bandgap as a result of defects in the otherwise periodic lattice. One of these was applied to the formation of deep levels in solids. This model can also be applied to the electromagnetics of defects in periodic structure. For a simple one-dimensional case, assuming that the defect exists at a single unit cell and that the dispersion of the band is dominated by nearest neighbor interactions, this can be solved exactly. **Fig. 2** shows the allowed frequencies of a ten unit cell chain with a single defect located at the center unit cell. Notice that one mode, called a donor mode, drops out of the band for positive values of $\Delta\epsilon$ and one mode, called an acceptor mode, rises out of the band for negative values of $\Delta\epsilon$. This is true in general. A perturbation of the lattice in which the defect has a higher index than the unperturbed lattice will cause a mode to drop out of the air band while a perturbation that has a lower dielectric constant than the background lattice will cause a mode to rise into the bandgap out of the dielectric band. In practice, resonant modes and frequencies in photonic crystals are modeled numerically.

A great deal of variety exists in resonant modes formed by defects in two-dimensional photonic crystal slabs. The simplest examples consist of a single missing hole in a two-dimensional square or triangular lattice patterned into a high-index dielectric slab. However, the resonant mode size, shape, frequency, and polarization can be engineered by tailoring the local dielectric in the region of the resonant mode. Losses in these cavities can be conceptually separated into losses in-plane through the photonic crystal and radiation losses out-of-plane. The in-plane losses are a result of having only a finite number of lattice periods surrounding the localized mode. A finite number of periods results in a finite loss due to tunneling of the fields through the lattice. This loss can in principle be made arbitrarily small by increasing the number of lattice periods surrounding the resonant mode. For a fixed number of lattice periods, the in-plane radiation loss will be reduced as the bandwidth of the photonic bandgap is increased because the magnitude of the in-plane decay constant increases with increasing bandgap width. A resonant mode formed by a defect in a truly three-dimensional photonic crystal will have losses limited by the number of achievable lattice periods and the bandwidth of the photonic crystal bandgap.

Photonic crystal cavities formed by two-dimensional photonic crystals of finite thickness suffer from an out-of-plane radiation loss. Optical confinement in the direction perpendicular to the two-dimensional photonic crystal is due to total internal reflection of the mode that occurs in the high index photonic crystal slab surrounded by lower index materials. Out-of-plane losses are due to the presence of wavevector components in the resonant mode that lie above the light line inside the radiation cone of the cladding. **Fig. 3** shows the Fourier transform of the magnetic field at the mid-plane of photonic crystal slab in which the resonant mode is formed by a single missing hole in a two-dimensional triangular lattice. In this cavity geometry, a doubly degenerate pair of modes is introduced into the photonic bandgap. The figure shows the Fourier transform of the field for one mode of this double degenerate pair. The boundary of the radiation cone in the figure is marked by the solid circle. Components of the mode inside this circle are not confined to the slab and contribute to out-of-plane radiation losses. These losses can be reduced by engineering the mode so that it has a smaller overlap with the radiation cone. This can be accomplished by allowing the mode to expand in real space in directions in which the wavevectors of the mode extend into the radiation cone. This reduces the spread of the Fourier transform of the mode and reduces the overlap of the mode with the radiation cone.

Symmetry can also be used to reduce the coupling to the radiation fields of the slab. Reducing the resonant frequency of the mode can also reduce out-of-plane radiation losses. This has the effect of reducing the size of the radiation cone in k -space. One way to accomplish this is to reduce r/a in the photonic crystal lattice. This will generally reduce frequencies of all of the photonic crystal modes. It will also, however, generally be accompanied by a reduction in the bandwidth of the photonic bandgap which will increase the in-plane losses. As a result, care must be taken in designing photonic crystal resonant cavities. Nevertheless, in the smallest mode volume photonic crystal resonant cavities, the optical losses will generally be dominated by out-of-plane radiation loss.

Quality factors have been predicted to exceed 100 000 in carefully designed photonic crystal resonant cavities in which the optical mode volume is on the order of a few cubic half-wavelengths in the material. Larger resonant cavities formed by

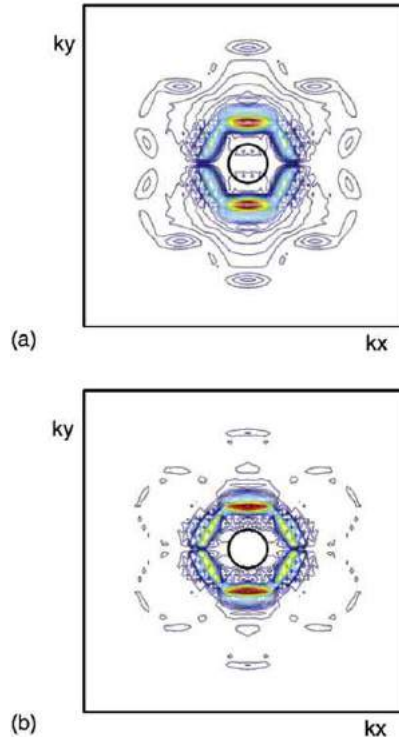


Fig. 3 (a) The spatial Fourier transform of the magnetic field component, H_z , of a defect cavity mode at the mid-plane of a triangular-lattice suspended membrane single-defect photonic crystal microcavity. The inner circle indicates the light cone inside which mode components radiate vertically. (b) The spatial Fourier transform of a modified defect cavity mode. The overlap of the mode with light cone has been reduced by modifying the defect cavity.

introducing multiple defects in the lattice are often less well confined in the real space lattice leading to a reduction in the components overlapping the radiation cone in k -space and a reduction in out-of-plane radiation loss. These cavities also are much more likely to support multiple resonant modes for a given index perturbation.

In analyzing resonance modes of photonic crystal resonant cavities, one of the most important parameters is the quality factor Q of the cold cavity. Practically, a theoretical prediction of the quality factor requires the use of numerical methods and finite-difference time-domain and finite element techniques are commonly used. There are basically three numerical approaches that can be used to calculate the quality factor of the cavity modes. Two of these methods calculate the quality factor from the time domain fields while the third is a frequency domain approach. The first method is to calculate the slope of the exponential decay of the energy of a given cavity mode with time. This decay of energy from the cavity is described by $\exp(-t/t_{ph})$ where t_{ph} is the photon lifetime which is related to the quality factor Q , by $Q = \omega t_{ph}$. This method is most useful for relatively low Q modes where the slope of energy decay is visibly greater than zero.

Another method is to calculate the ratio of the full width at half magnitude (FWHM) of the cavity resonance in the frequency domain, $\Delta\omega$, to the center frequency, ω_0 . However, distortion to the spectrum is often introduced because the numerical simulation terminates before the impulse response is fully evolved. This has the effect of viewing the true time-domain response through a rectangular window, which translates mathematically into the convolution of the true spectrum with a sinc function. This must be accounted for in the determination of the quality factor.

A third method calculates the ratio of cycle-averaged power absorbed in the boundary to the total energy in the cavity mode. This last method has the advantage of being able to separate the radiation losses into different directions:

$$\frac{1}{Q} = \frac{\omega P}{U} = \frac{\omega(P_{\perp} + P_{\parallel})}{U} = \frac{1}{Q_{\perp}} + \frac{1}{Q_{\parallel}} \quad (2)$$

in which, the effective vertical quality factor $Q_{\perp}$ is given by the ratio of power lost to the absorber at the top and bottom $P_{\perp}$ to the total cavity energy $U(t)$, and the effective in-plane quality factor $Q_{\parallel}$ is similarly given by the ratio of in-plane power loss $P_{\parallel}$ to the total cavity energy.

Photonic crystal lasers and resonant cavities have been demonstrated experimentally. The first demonstration of lasing in a two-dimensional photonic crystal laser occurred in a cavity formed by a single defect in a triangular lattice that was patterned into an InGaAsP membrane. The laser operated under optically pumped conditions at 143K. Room temperature pulsed operation of photonic crystal defect lasers was reported shortly thereafter. Fig. 4 shows an electron micrograph of such a resonant cavity. In the figure, the resonant cavity is formed in a suspended membrane in which a triangular lattice photonic crystal with a defect has been

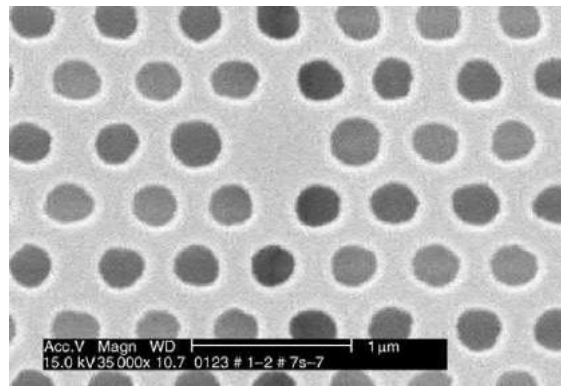


Fig. 4 The scanning electron micrograph of a triangular-lattice suspended membrane photonic crystal single defect cavity.

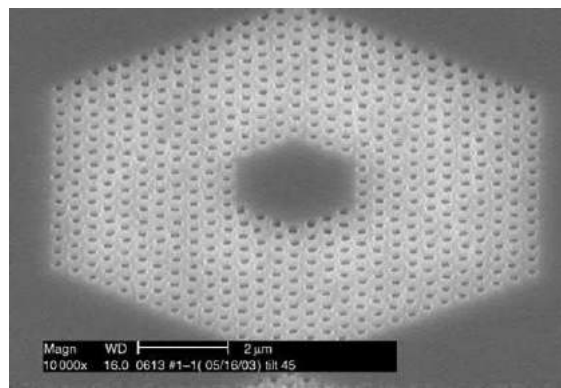


Fig. 5 The scanning electron micrograph of a triangular-lattice, 37-missing-hole, photonic crystal defect cavity, bonded on sapphire which serves as lower cladding to improve heat dissipation. CW operation of the laser cavity was demonstrated.

patterned. The fabrication of these devices involves pattern definition by electron beam lithography and pattern transfer by a series of wet and dry etching steps. The observed quality factors are, of course, sensitive to fabrication imperfections. Many other cavity designs and demonstrations have also been reported.

To operate a photonic crystal laser continuously at room temperature, it is necessary to design a high-Q resonant cavity that dissipates heat well. Poor heat dissipation is a major drawback of suspended membrane resonant cavities. One strategy for improving the heat dissipation in photonic crystal laser cavities is to form the two-dimensional photonic crystal membrane cavity on top of a low-index high-thermal-conductivity substrate. **Fig. 5** shows an electron micrograph of a photonic crystal cavity formed by leaving out 37 holes from a triangular lattice patterned into an InGaAsP membrane. This membrane is bonded to a sapphire substrate, which facilitates heat dissipation. This cavity is capable of room temperature continuous wave operation.

It is believed that these lasers are capable of electrically-pumped room temperature operation. This is based on the fact that large quality factors have been demonstrated in optically-pumped photonic crystal lasers and some quality factor reduction due to free carrier absorption and absorption at metal contacts can be tolerated and still lead to lasing. A reasonable assumption would be that the performance of these electrically-pumped photonic crystal lasers would have much in common with the performance of VCSELs. Both VCSELs and photonic crystal cavities have small mode volumes with the cavity formed in some directions by distributed Bragg reflection. The emission direction of photonic crystal lasers can be engineered to be vertical or in-plane, but the basic resonant cavities do have important similarities. The free carrier absorption loss in photonic crystal laser cavities may be expected to be larger than in VCSELs since a photonic crystal laser mode lacks the standing wave behavior in the vertical direction that allows doping at the nodes of the standing wave to reduce the absorption loss. If free carrier absorption loss leads to larger internal losses for photonic crystal lasers than for VCSELs, then some reduction in the achievable slope efficiencies will also occur for photonic crystal lasers.

Photonic Crystal Waveguides

Another important optical element for integrated photonic circuitry is a linear waveguide. Conventional dielectric waveguides confine propagating beams by an index of refraction difference between the waveguide core and the waveguide cladding. Photonic crystal waveguides are most often formed by a linear defect, which consists of a row of modified unit cells of the lattice, patterned

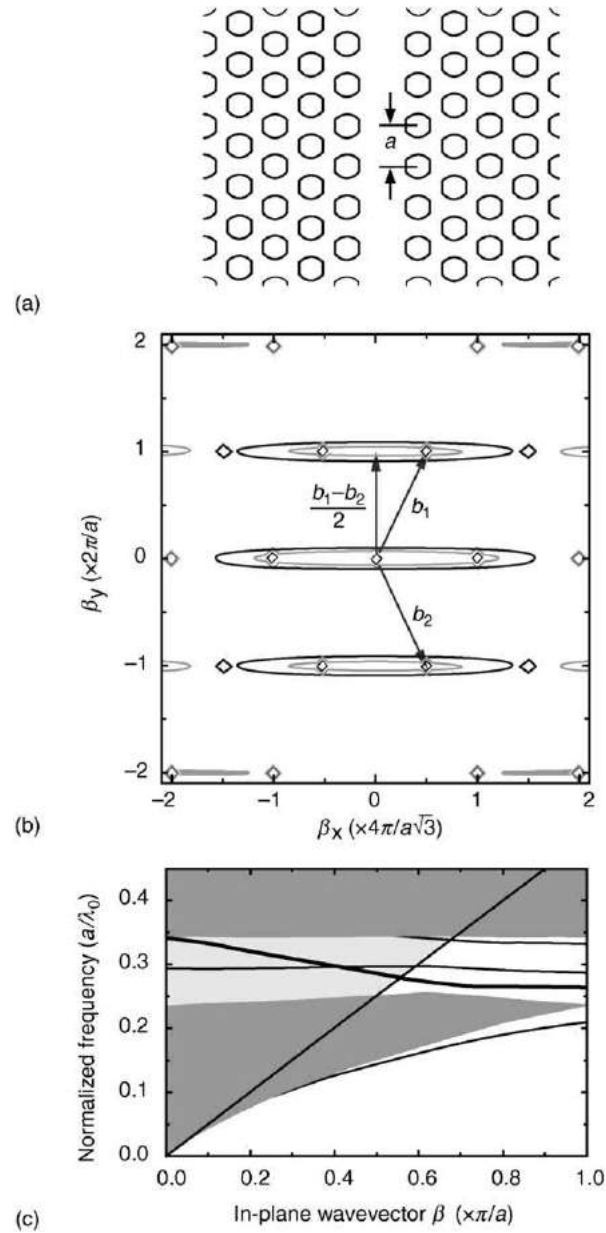


Fig. 6 (a) The top view of a triangular lattice photonic crystal single-line defect waveguide with air hole radius $r/a=0.3$. (b) The 2D spatial Fourier transform of the waveguide dielectric distribution. (c) The photonic band diagram for a suspended membrane single-line photonic crystal waveguide. The membrane has a refractive index of $n=3.4$ and its thickness is $d/a=0.6$. The surrounding material is air. The light gray and dark gray area indicate the regimes where field radiates through the air and photonic crystals, respectively.

into a high index dielectric membrane. The guiding in this case is due to a mixture of total internal reflection at the high index membrane/low index cladding interfaces and distributed reflections from the photonic crystal in-plane. Because this confinement does not exclusively depend on total internal reflection, the transmission loss through small bending radius waveguide branches and bends can be made very small. Predictions and experimental demonstrations exist which show that photonic crystal waveguides are capable of supporting small bending radius waveguide bends with almost total transmission. This makes photonic crystal waveguides a candidate for waveguides in densely integrated photonic circuits.

Fig. 6(a) shows an illustration of the top view of a two-dimensional photonic crystal waveguide. This is a triangular lattice in which one row of unit cells has been modified. Shown in **Fig. 6(b)** is the Fourier transform of index of refraction for the situation in which the photonic crystal consists of low index holes in a high index semiconductor. The presence of the linear defects modifies and broadens the reciprocal lattice from a simple reciprocal triangular lattice, before the creation of linear defect, to the distribution shown in the figure. This Fourier transform has components along the direction of propagation at $(\vec{b}_1 - \vec{b}_2)/2$ as shown in the figure where $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice unit vectors. These reciprocal lattice components at $(\vec{b}_1 - \vec{b}_2)/2$ couple a

planewave with wavevector β to other planewaves with wavevectors $\beta \pm (\vec{b}_1 - \vec{b}_2)/2$. The result of this coupling is that the electric field of the waveguide mode is a Bloch wave. In a photonic crystal waveguide in which the waveguide axis is the z -direction, the Bloch wavefield in the waveguide can be written in the form:

$$\vec{E}(x, y, z, t) = \sum_n c_n \vec{f}_\beta(x, z) \times \exp \left[i \left(\omega t - \left[\beta + n \left(\frac{\vec{b}_1 - \vec{b}_2}{2} \right) \cdot \gamma \right] \right) \right] \quad (3)$$

In this expression the periodic function of the Bloch wave has been explicitly expanded as a Fourier series and each term in the Fourier series is called a spatial harmonic. Note that these fields differ from that of a photonic crystal fiber because the cladding of a photonic crystal waveguide has a periodicity along the direction of a waveguide. A photonic crystal fiber does not.

Most of the realizations of photonic crystal waveguides occur in two-dimensional photonic crystals formed in high index dielectric slabs. The membrane is usually located a few microns above a high index substrate. Because the fields of the guided modes decay exponentially outside the membrane, the effect of a substrate several microns away is very nearly negligible. The photonic band structure for $r/a=0.3$, normalized membrane thickness $d/a=0.6$ and refractive index of $n=3.4$ is shown in Fig. 6 (c). Only the modes with even symmetry along the mid-plane of the membrane are shown. A bandgap opens for modes with odd symmetry only for large low index filling factors in the triangular lattice. Filling factors large enough to create a bandgap for the odd modes are often impractical so that the odd modes are rarely considered. The even and odd modes are orthogonal, so that plotting the dispersion relation for the even modes only is justified. In reality, even and odd modes may be weakly coupled as a result of fabrication imperfections. The horizontal axis of the dispersion relation covers the in-plane wavevector β along the propagation direction of the irreducible Brillouin zone. The vertical radiation light cone and transverse radiating region of the photonic crystal are mapped as light gray and dark gray areas, respectively. They correspond to the regimes in which light radiates through the air cladding and photonic crystal cladding, respectively. Waveguide dispersion relations can be calculated using two-dimensional approaches in which case the guiding due to the high index slab is accounted for by using an effective index of refraction for the mode. The effective index is the ratio of the propagation coefficient to the free space wavevector of the guided

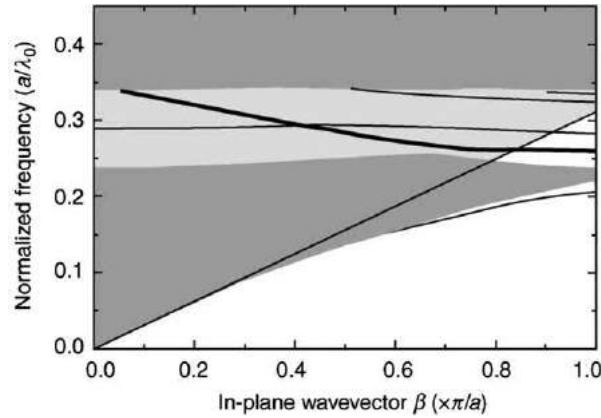


Fig. 7 The photonic band diagram for a triangular-lattice sapphire-bonded single-line photonic crystal defect waveguide. The dielectric membrane has a refractive index of $n=3.4$ and thickness of $d/a=1.0$. The index of the sapphire bottom cladding is assumed to be 1.6.

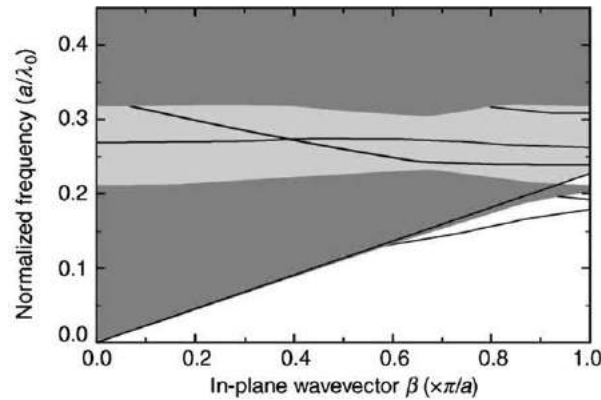


Fig. 8 The photonic band diagram for a triangular-lattice deeply etched single-line photonic crystal defect waveguide. The top cladding, guiding membrane, and bottom cladding have an index of refraction of 3.0, 3.4, and 3.0, respectively. Their normalized thicknesses d/a are 1.0, 1.0, and 6.0, respectively.

mode of the slab. The two-dimensional calculation then uses the effective index of the mode as the high index of material in the photonic crystal. To obtain accurate results, the effective index for each mode calculated separately and the two-dimensional calculation of the photonic crystal waveguide dispersion must be repeated for each waveguide mode.

Examples of photonic crystal waveguide dispersion relations are shown in Figs. 7 and 8 for waveguides formed by photonic crystals in dielectric slabs on high index lower cladding layers. Fig. 7 is the dispersion relation for an oxide lower cladding layer. The area of the light gray region, which is the projection of the oxide cladding light cone onto waveguide propagation direction, increases as a result of higher index of refraction. Fig. 8 shows the dispersion relation for a waveguide with an even larger lower cladding index. In this case the air holes of the two-dimensional photonic crystal are patterned all the way through the lower cladding layer, with an index of 3.0, on a high index substrate. The defect modes lying within the in-plane photonic bandgap are all above the cladding light line in this case. Propagation losses are predicted to increase by over two orders of magnitude from the structure with very low index cladding layers shown in Fig. 6 to the high index lower cladding structure of Fig. 8.

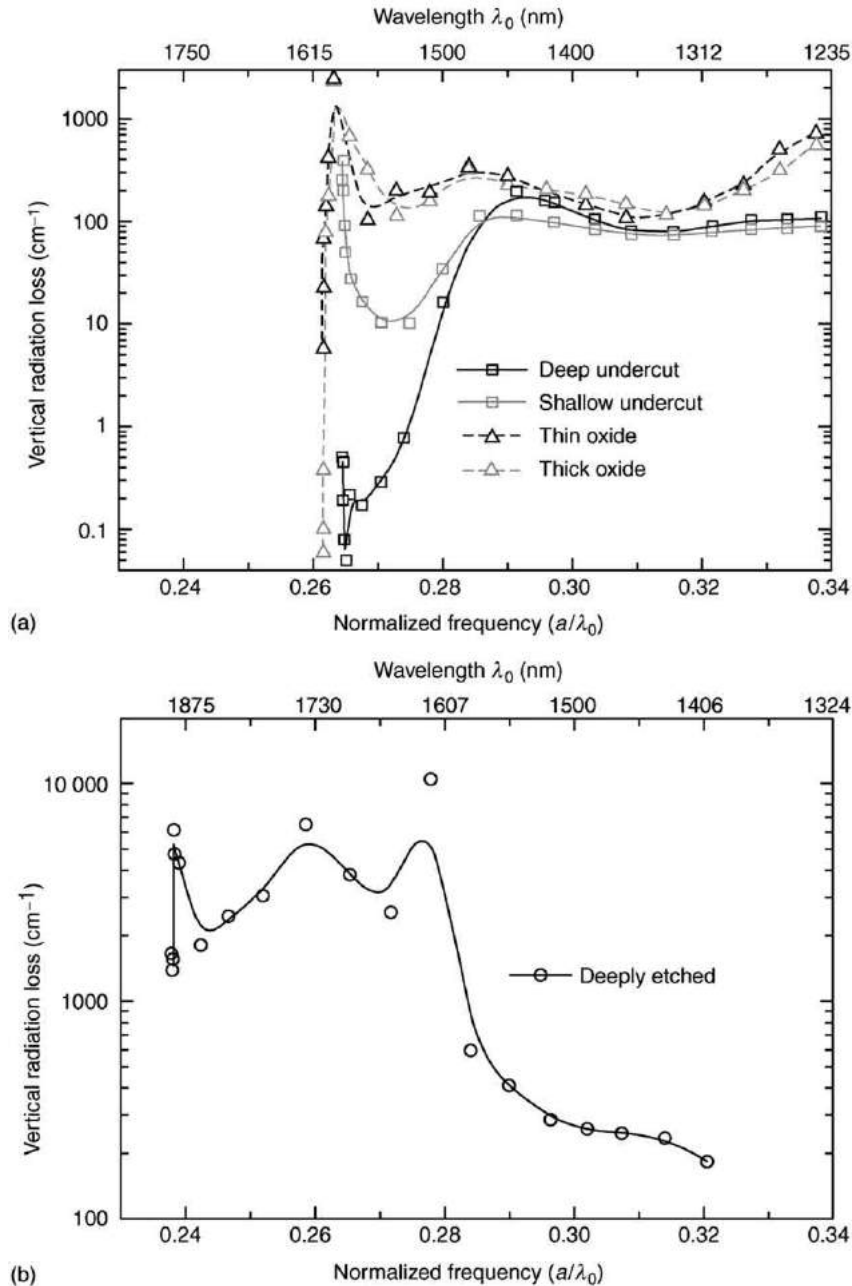


Fig. 9 The out-of-plane radiation loss as a function of normalized frequency for the photonic crystal defect slab waveguides illustrated in Figs. 6–8. The lattice constant $a=450$ nm for deeply etched waveguide and 420 nm for the rest of the cases. Points are calculated values and lines are B-spline curve fits.

Table 1 Photonic crystal defect slab waveguides considered in the calculations

Waveguide structures	Suspended membrane		Oxidized lower cladding		Deeply etched
	Deep undercut	Shallow undercut	Thin oxide	Thick oxide	
Layer description: Material (normalized thickness d/a)	—	—	Air (>3.0)	—	AlGaAs (1.0)
	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (1.0)
	Air (3.0)	Air (1.0)	Al _x O _y (2.0)	Al _x O _y (5.0)	AlGaAs (6.0)
			GaAs substrate (>3.0)		

Refractive index of 3.4 is assumed for GaAs, 3.0 for AlGaAs, 1.6 for Al_xO_y, and 1.0 for air.

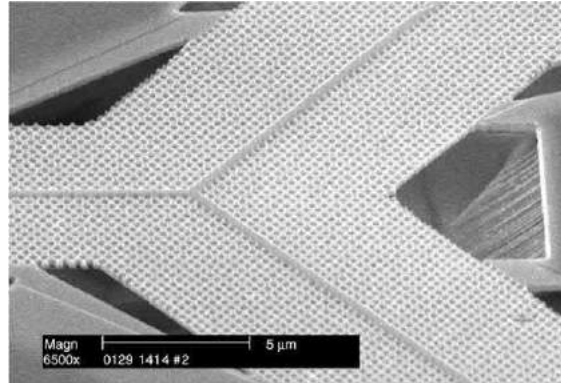

Fig. 10 The scanning electron micrograph of a photonic crystal Y-branch fabricated on InP material system.

Fig. 9(a) and **(b)** shows the results of a fully three-dimensional calculation for the propagation loss of the three structure shown in **Figs. 6–8**. The details of these structures are shown in **Table 1**. This calculated loss is due to coupling between the guided mode and the radiation modes of the high index substrate. A wavelength scale is also included using $a=420$ nm in **Fig. 9(a)** and 450 nm in **Fig. 9(b)**. The suspended membrane suffers the least vertical radiation loss among all of the waveguides considered and has a minimum loss around 0.2 cm^{-1} . Low-loss waveguides formed on high index cladding layers may be possible, but care should be taken to eliminate coupling between the guided modes and the radiation modes of the cladding. In this case, the design strategy for reducing the optical loss is to reduce the magnitude of the Fourier component of the electromagnetic field inside the light cone. It should also be noted that deeply-etched waveguides formed by removing multiple rows of holes have been demonstrated with significantly lower radiation loss than is predicted for the single line defect waveguides. However, these waveguides are multi-moded at all frequencies. It is also important to remember that the nanofabricated waveguides will suffer additional losses due to fabrication imperfections that were not included in the numerical model. **Fig. 10** shows a nanofabricated photonic crystal waveguide. The image is an electron micrograph of a y-branch component of a Mach–Zehnder interferometer formed in a two-dimensional photonic crystal waveguide patterned into a suspended InGaAsP membrane.

Finally, it is worth noting that the group velocity of photonic crystal waveguide modes, where the group velocity is given by the gradient of the dispersion surface:

$$v_g = \nabla_{\vec{k}} \omega \quad (4)$$

is very small near the zone boundary. This can be seen from the dispersion relations shown in **Figs. 6–8**. In fact, the slope of the dispersion relations is likely to be a parameter which can be engineered by careful waveguide design. This property may allow photonic crystal waveguides to find applications in optical processing functions such as delay lines.

See also: Electromagnetic Theory

Further Reading

- Baba, T., 1997. Photonic crystals and microdisk cavities based on GaInAsP–InP system. *IEEE Journal of Selected Topics in Quantum Electronics* 3, 808–830.
- Joannopoulos, J.D., Meade, R.D., Winn, J.D., 1995. *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.
- Mekis, A., Chen, J.C., Kurland, I., Fan, S., Villeneuve, P.R., Joannopoulos, J.D., 1996. High transmission through sharp bends in photonic crystal waveguides. *Physics Review Letters* 77, 3787–3790.

- Painter, O., Lee, R.K., Yariv, A., *et al.*, 1999. Two-dimensional photonic crystal defect laser. *Science* 284, 1819–1821.
- Sakoda, K., 2001. *Optical Properties of Photonic Crystals*. New York: Springer.
- Slater, J.C., 1967. *Insulators, Semiconductors, and Metals*, vol. 3. New York: McGraw-Hill Book Company.
- Srinivasan, K., Painter, O., 2002. Momentum space design of high-Q photonic crystal optical cavities. *Optics Express* 10, 670–684.
- Villeneuve, P.R., Fan, S., Joannopoulos, J.D., 1996. Microcavities in photonic crystals: mode symmetry, tunability, and coupling efficiency. *Physics Review B* 54, 7837–7842.
- Yariv, A., Yeh, P., 1984. *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Microstructure Fibers

RS Windeler, OFS Laboratories, Murray Hill, NJ, USA

© 2018 Elsevier Inc. All rights reserved.

Introduction

Microstructure fibers have unique properties that cannot be obtained from traditional fibers (i.e. all glass, doped silica fibers) and can deliver functionalities superior to many of today's best transmission and specialty fibers. Their unique properties are obtained from an intricate cross-section of high and low index regions that traverse the length of the fiber. The vast majority of fibers consist of silica for the high-index region and air for the low-index region. These fibers are known by several different names including microstructure fiber, holey fiber, and photonic crystal fiber. The term microstructure fiber is used in this chapter because it is the most generic, encompassing all the fiber types.

Microstructure fiber properties vary greatly and are determined by the size and arrangement of air holes in the fiber. For example, fibers have been fabricated such that the numerical aperture of the core or cladding approaches one. They have been fabricated with unique dispersion profiles, such as a zero-group velocity dispersion in the near visible regime or an ultraflat dispersion over hundreds of nanometers wide. They have been fabricated to generate a super continuum over two octaves wide. Some guide light with an air core via bandgap guidance. In addition, the air holes in these microstructure fibers can be filled with highly tunable materials, giving one the capability of controlling the fiber properties for use in energy efficient devices.

Microstructure fibers can have doped regions like traditional fibers. These hybrid fibers combine the benefits of traditional and microstructure fibers. The doped core typically guides most of the light, but its guidance properties can be strongly influenced by the air regions. Applications for hybrid fibers include dispersion compensation, polarization maintaining, and bend insensitive fibers.

Loss is always an important factor when determining whether a microstructure fiber will compete with or replace a traditional optical fiber. The index profiles that make these fibers so special can also lead to fibers that have an inherently high loss at connections or along the length of the fiber. Loss at connections typically occurs due to mode mismatch between the two fibers, undesired hole collapse, or poor alignment. Loss along the fiber length can occur due to impurities, hole surface roughness, or poor confinement.

The dispersion profile (zero dispersion wavelength, dispersion slope, normal or anomalous) of microstructure fibers can be optimized more than a traditional fiber, because the hole size and placement can be arranged to make the cladding index strongly wavelength dependent. Dispersion causes a pulse to spread because the phase velocity is wavelength dependent. Normal dispersion is when longer wavelengths travel faster than shorter wavelengths. Anomalous dispersion is just the opposite—shorter wavelengths travel faster than longer wavelengths. Dispersion determines the amount of interaction between different wavelengths. For applications such as transmission fiber, one wants little interaction and for high nonlinear applications, one wants significant interaction.

This chapter examines the methods for fabricating microstructure fibers and reviews the different types of microstructure fibers and their properties. Fibers that guide by total internal reflection are reviewed separately from bandgap guided fibers because their properties are significantly different. Lastly, a description and example of an active device is presented in which the air regions of a microstructure fiber are filled with controllable material.

Microstructure Fiber Fabrication

At first glance, fabrication of microstructure fibers looks similar to traditional fibers in that the fibers are fabricated in a two-step process. First a preform is fabricated and then it is drawn (stretched) into a fiber. However, when the process is examined in more detail, both the preform fabrication and draw depart significantly from traditional methods. First, a novel preform fabrication method is used to incorporate air holes that run the length of the preform. Second, novel drawing procedures are used to keep the air holes open as the preform diameter is reduced by several orders of magnitude during draw.

Preform Fabrication

Microstructure preforms are cylinders of amorphous material usually less than a few centimeters in diameter that have an index profile running the length of the preform similar to the desired fiber. The most common preform consists of air and a single material such as silica, polymer, or high nonlinear glass. These air-containing preforms are relatively easy to fabricate but are difficult to draw because the structures can deform.

The basic parameters of the fiber are usually determined by the geometry of the holes in the preform. These include hole diameter (d), hole position, pitch (center-to-center hole spacing, Λ), core diameter (a), and number of layers. However, changes in

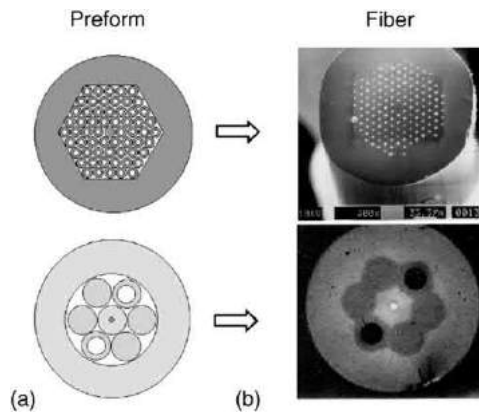


Fig. 1 (a) Tubes, rods, and core rods are stacked together in a close packed arrangement and held together with an overclad tube. (b) During draw the desired air region stay open forming a microstructure fiber.

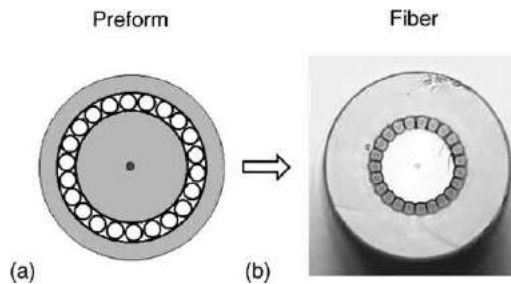


Fig. 2 (a) An air-clad preform is fabricated by stacking a ring of small, thin walled tubes around a large core rod. (b) Photograph of the resulting fiber.

the size and shape of the holes in the fiber can be made purposely or accidentally during draw, which will cause deviations between the fiber and preform profiles.

Several methods are used to fabricate microstructure preforms including stacking, casting, extrusion, and drilling. Each process has its advantages and disadvantages. By far, the most common method is stacking in which tubes, rods, and core rods (rods containing doped cores fabricated by traditional optical fiber techniques) are stacked in a close packed pattern such as a triangular or hexagonal lattice (Fig. 1). The bundle of tubes and rods is then bound together with a large tube, called an overclad tube. Fibers can be made with complicated or asymmetric index profiles by strategically placing tubes with the same outer diameter (for good stacking) but a different inner diameter to change the index in that location. For example, a fiber with smaller holes on opposite sides of the core produces high birefringence.

The main advantages of stacking are that no special equipment is needed for fabricating microstructure preforms and that doped cores are easily added. However, there are several disadvantages. First, the stacking method is limited to simple geometries of air holes because the tubes are stacked in a close pack arrangement. Second, unless hexagonal tubes are used, interstitial areas are created between the tubes that may not be desired in the final fiber. Third, the method is labor intensive and requires significant glass handling.

A variation of the stacking method is used for fabricating air-clad fibers. Air-clad preforms are created by placing a layer of small tubes around the perimeter of a large core rod. The assembly is then overcladed for strength (Fig. 2).

Extrusion and casting of glass powder, polymer, or sol-gel slurry are also used for making single material microstructure preforms. The main advantage of extrusion and casting is that complicated structures can be fabricated in which the position, size, and shape of the air regions are independent of one another. The disadvantage of this method is that preform fabrication is more difficult compared to the other fabricate methods. This method may become more common as complex air structures are needed to make advanced microstructure fibers.

The last method consists of drilling holes in a preform or rod. Drilling is well understood and used for other specialty fibers. The advantage of this method is that it is easy to drill holes of various sizes in any position in a preform, including doped regions. The disadvantages of drilling are that the holes cannot be drilled very deep compared to a traditional preform length; the distance between holes may be limited due to cracking; and the fiber may experience high loss due to surface roughness of the holes and impurities incorporated during drilling.

Fiber Draw

It is significantly more difficult to draw a microstructure preform with air holes than a traditional preform. This is because the air holes tend to close due to surface tension. This force can be overcome or minimized using two different techniques. The first is to draw the preform under very high tension (low temperature). Minimal hole collapse occurs when the draw tension is significantly larger than the surface tension. The problem with this method is that the break rate increases substantially at higher tension.

The second method is to apply an external pressure to the holes to counteract the surface tension. If a single pressure source is used, the holes can be made larger or smaller during draw by changing the pressure as needed. However, the process can become unstable because a large hole needs a lower pressure to maintain its size than a small hole. If holes of different sizes exist, a larger hole will grow at the expense of a smaller one regardless of the pressure used. To minimize the instability, this method is typically performed in conjunction with high-tension drawing.

To alleviate the draw problem, a second solid material can be used in the preform to obtain the desired index profile. With the air regions removed from the preform, the drawing process becomes significantly easier. In addition, when a preform consists of two solid materials, one of the materials does not have to be continuous, as in single material preforms. This allows for simpler designs such as concentric rings, which may be easier to fabricate and model (Fig. 13(c)). The disadvantage of this approach is finding two materials that have compatible physical and chemical properties, have low loss, and have a large index of refraction difference.

Microstructure Fiber Types

Microstructure fibers are categorized by their method of guidance, properties, and function. Index guided microstructure fibers most closely resemble traditional fibers and are described first. These fibers guide light by total internal reflection and have a core index of refraction greater than the cladding. Bandgap fibers are examined next. The unique guidance of bandgap fibers allows the core index to be lower than the cladding index. Tunable microstructure fibers are described last. These microstructure fibers are filled with tunable materials so that the fiber's properties can be actively manipulated.

Index Guided Fibers

Most microstructure fibers guide light by total internal reflection. The size and spacing of air holes determine the guiding properties. For example, a solid silica core surrounded by a cladding consisting of small air holes that are spread apart creates a small core-to-cladding index difference similar to what can be achieved in traditional fibers. At the other extreme, a solid silica core surrounded by a cladding of large holes, which are closely spaced together, forms a large core-to-cladding index difference. This structure creates a large numerical aperture in the core and optically isolated regions within the fiber. The ability to create regions of significantly different indices in various parts of the fiber distinguishes the index guided microstructure fibers from traditional fiber. Different types of microstructure fiber are formed by varying the size, spacing and pattern of the air holes.

Different techniques are used to model these fibers, depending on spacing and size of the air holes relative to the wavelength of the guided light. If the hole diameter and spacing is much smaller than the wavelength of the guided light, the cladding index can be modeled as an average index weighted by the volume fraction of the two materials (typically air and silica). As the hole diameter or spacing approaches the wavelength of the guided light, complicated models are used such as finite difference method, finite element method, beam propagation method, or multipole method.

Endlessly single-mode fiber

The endlessly single-mode fiber consists of a solid core surrounded by a two dimensional array of small air holes in the cladding (Fig. 3). The size and position of the air holes are arranged such that only the fundamental mode exists in the core regardless of the wavelength of the guided light.

The prevention of higher order modes can be understood by examining the V number. The V number is a dimensionless parameter that describes the guiding nature of the waveguide. In traditional fiber, when the V number is less than 2.405, the fiber will guide only the fundamental mode. The V number is given by

$$V = \frac{2\pi a}{\lambda} (n_{\text{core}}^2 - n_{\text{clad}}^2)^{1/2} \quad (1)$$

where a is the core diameter, λ is the wavelength of the guided light, and n_{core} and n_{clad} are the index of refractions of the core and cladding, respectively. In traditional fiber, the core and cladding index (n_{core} and n_{clad}) are for the most part constant, so the V number increases as the wavelength decreases.

However, for the endlessly single mode microstructure fiber, the V number does not increase indefinitely with decreasing wavelength, but approaches a constant value. Obtaining a V number independent of wavelength is understood by examining the distance the light propagates into the cladding as a function of wavelength. As the wavelength becomes shorter, light penetrates a shorter distance into the cladding, interacting with less of the air regions. This causes the effective cladding index (n_{clad}) to increase with decreasing wavelength, such that V approaches a constant value.

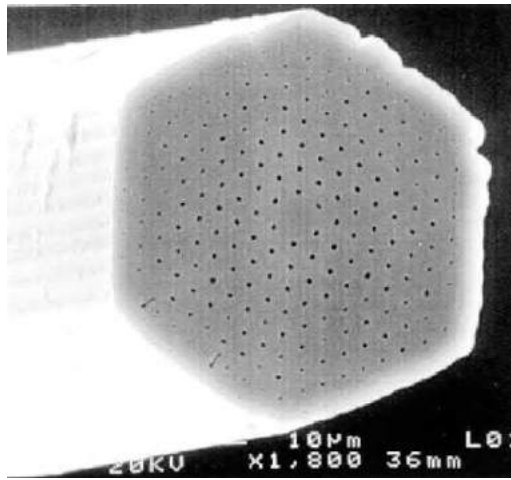


Fig. 3 Photograph of an endlessly single mode fiber. Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963.

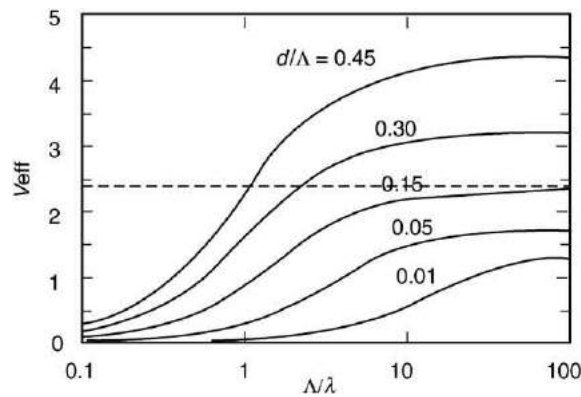


Fig. 4 The effective V number as a function of Λ/λ for various d/Λ . Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963.

Microstructure fibers are made endlessly single-mode over all wavelengths by properly designing the hole diameter-to-pitch ratio (d/Λ). When the d/Λ is low enough, the microstructure fiber guides light only in the single mode regardless of wavelength and core size (Fig. 4). However, these large core diameter fibers are limited by long wavelength bend loss in the same way as large core single mode traditional fibers.

High delta core fiber

High delta core microstructure fibers (HDCMF) consist of a small (typically less than 2 microns) solid silica core surrounded by one or more layers of large air holes closely spaced together (Fig. 5). This creates a very large index difference between the core and cladding. A core-to-cladding index difference of 0.4 is easily achieved in a HDCMF, which is an order of magnitude larger than can be achieved with traditional fiber.

Even though the core is very small, these fibers are almost always multimoded due to the large index difference between the core and cladding. However, because of its small core, it is difficult to launch anything other than the fundamental mode into the fiber. Once a mode is guided in the fiber it remains in that mode due to the large difference in effective indices between modes (Fig. 6). When higher-order modes are purposely generated in the fiber, they also guide over long lengths without coupling to other modes (Fig. 7). This strong guidance makes the fiber extremely bend insensitive.

HDCMF can have a unique dispersion profile because it can guide the fundamental mode exclusively in the multimode regime. This allows greater flexibility in controlling the shape of the dispersion profile, especially the ability to move the zero dispersion to shorter wavelengths. The HDCMF is the first silica fiber with anomalous dispersion of single-mode light below 1270 nm (the wavelength at which the material dispersion of silica is zero). These fibers are often made to have a zero-group velocity dispersions around 770 nm for use with Ti-sapphire lasers.

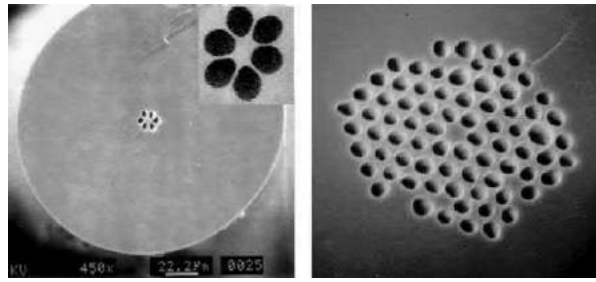


Fig. 5 High delta core fibers consist of a small solid core surrounded by at least one layer of large air holes closely spaced together. Right-hand panel reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

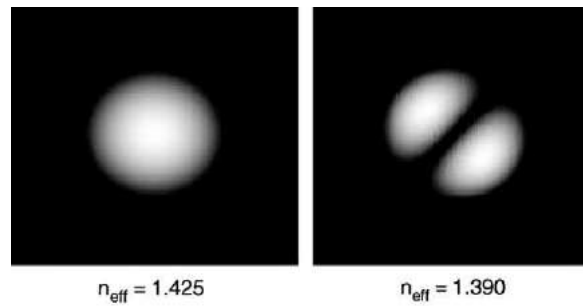


Fig. 6 Individual modes do not couple well due to large differences in effective indices between modes. The figure shows the mode profile and effective index of the fundamental and the next higher mode for a fiber with a two-micron diameter core with a core-to-cladding index difference of 0.45. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

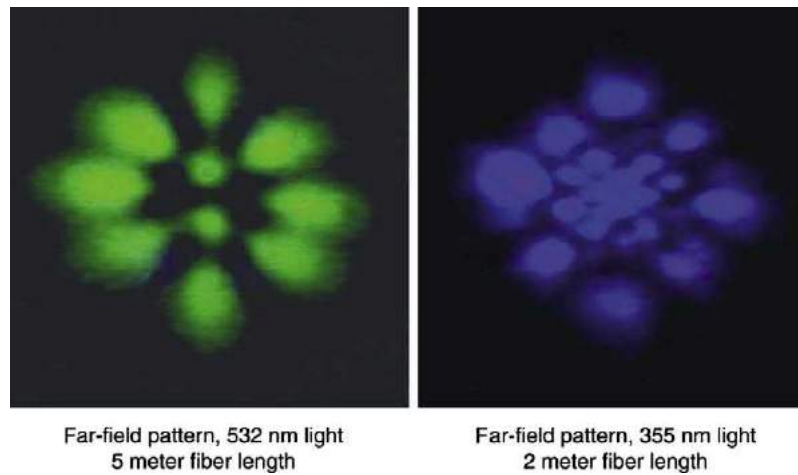


Fig. 7 Far-field patterns show that higher order modes propagate long distances in the HDCMF without coupling to other modes. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

HDCMF is ideal for performing nonlinear experiments in the near-visible regime because the fiber's small, high-index core creates high intensity light, and the fiber's low dispersion creates long interaction times. For example, a broadband continuum can be generated over two octaves using less than a meter of HDCMF fiber (Figs. 8 and 9).

Tapered microstructure fiber (TMF)

Coupling light in and out of HDCMF is difficult and results in a large loss due to its small core. When free space optics are used to couple the light, frequent realignment of the beam is required to minimize loss due to drifts in the equipment. Splicing HDCMF

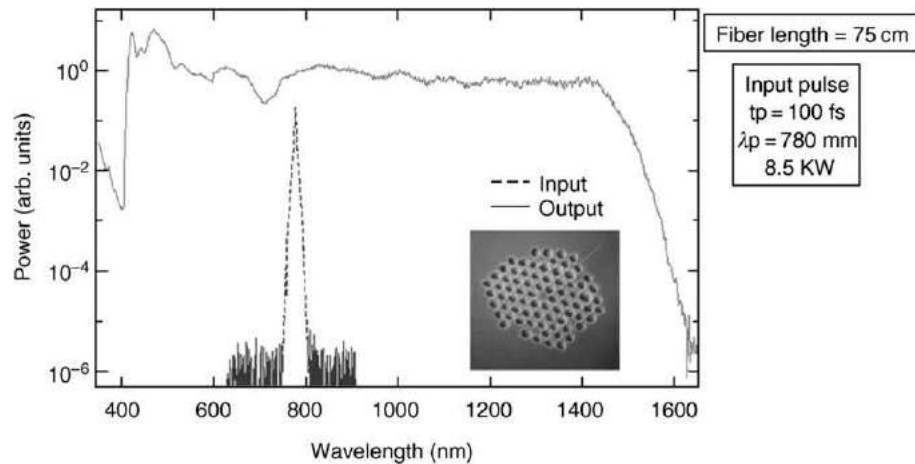


Fig. 8 Plot of a supercontinuum spectrum over two octaves wide generated from a 75 cm piece of HDCMF. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25: 25–27.

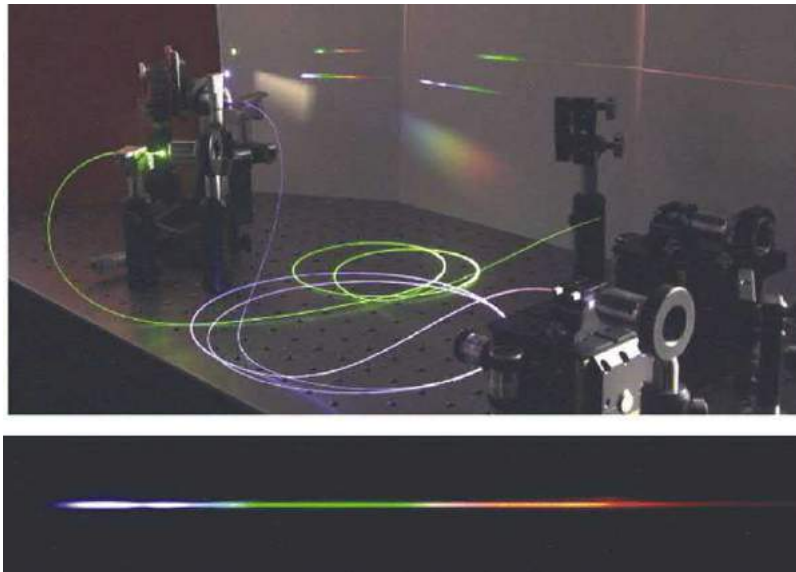


Fig. 9 Photograph of fiber generating the supercontinuum. The bottom photo shows the continuum after passing through a prism.

fiber to traditional fiber also results in high loss due to a large modal mismatch. These coupling problems are solved with tapered microstructure fibers.

Tapered microstructure fibers (TMF) consist of a traditional germanium single-mode core surrounded by a small silica cladding (Fig. 10). The inner cladding is surrounded by a layer of air holes and then a protective outer silica cladding for strength. Since the core of the TMF is very similar to a traditional single-mode fiber core, the splice loss is low. Splice losses below 0.075 dB are typical.

Properties similar to the HDCMF are obtained by adiabatically tapering the TMF while maintaining the cross-sectional profile of the fiber. In the taper region, the fundamental mode is no longer guided by the germanium core and evolves into the fundamental mode of the silica region, where it is confined by the ring of air holes. The taper waist is typically ten to twenty centimeters long and has properties identical to the HDCMF. When the fiber is adiabatically expanded back to its original size, the light is guided back into the standard diameter germanium core, making it easy to splice to traditional fiber with low loss.

The same effect can be obtained by stretching a traditional fiber from 125 microns down to 2 microns. However, there are several disadvantages compared to the tapered microstructure fiber. The taper region is much longer in the traditional fiber; the optimal diameter of the silica core is harder to maintain; the taper region must remain clean and uncoated; and the taper region is much weaker.

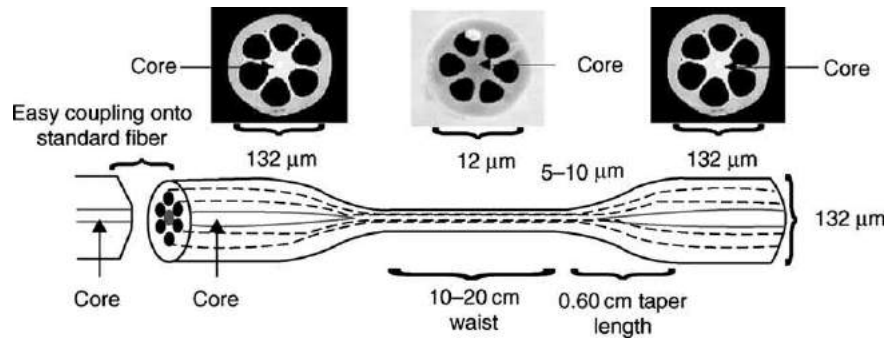


Fig. 10 Schematic drawing of the tapered microstructure fiber. Reprinted with permission from Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1) (©2001 IEEE).

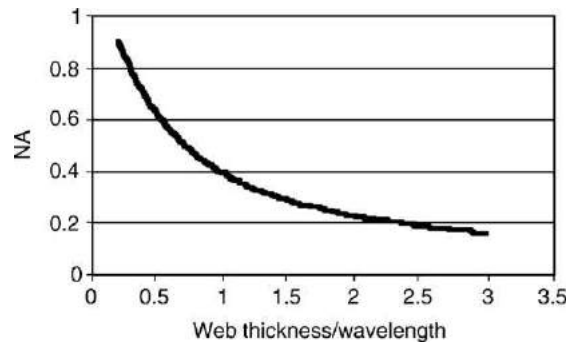


Fig. 11 Predicted NA as a function of the web thickness divided by the wavelength using an infinite slab model.

Air-clad fiber

Air-clad fibers consist of a doped core surrounded by a silica inner cladding. The inner cladding is surrounded by a ring of air, which in turn is surrounded by an outer silica cladding for strength (Fig. 2). Thin silica webs connect the inner and outer claddings to hold the inner fiber region in place. The region consisting of air and thin silica webs creates an optical barrier, preventing most of the light from escaping the inner cladding. Such fibers are referred to as double clad fiber.

In traditional double-clad fibers, a material of lower index (typically another layer of glass or a layer of polymer) immediately surrounds the inner silica cladding. A key optical parameter describing the light guiding properties of the inner cladding is the numerical aperture (NA). The NA is defined as the sine of the largest (acceptance) angle of light that will be guided in the inner cladding. The upper NA values of traditional double-clad fibers (~ 0.22 and ~ 0.45 for glass and polymer outer claddings, respectively) are limited by the refractive indices of the available cladding materials.

For the air-clad fiber, the NA is determined by the ability of light to escape from the inner cladding through the silica webs. The NA of the air-clad fiber can be calculated by modeling the web as an infinitely long slab waveguide. As seen in Fig. 11, the NA increases as the webs become thinner or the wavelength of light becomes longer. To obtain NA similar to what is achieved by coating a traditional fiber with a low index polymer ($NA \sim 0.45$), the web thickness must be about equal to the wavelength of the light. The NA of the air-clad fiber increases dramatically as the web thickness becomes less than half the wavelength of the guided light. Numerical apertures above 0.90 have been experimentally demonstrated.

The large NA values achievable with air-clad fibers are an important advantage for advanced double-clad amplifiers and lasers. A double-clad amplifier or laser (also known as cladding pumped amplifier or laser) contains a double-clad fiber with a rare earth doped core. High-power pump light, launched into and guided within the inner cladding, excites the rare earth ions as the pump light crosses through the core. The excited ions, which emit light with the same wavelength and phase as the signal, amplify the signal. The fiber's efficiency and the rare earth inversion level increase with the amount of pump light absorbed by the core. The light absorbed by the core, in turn, increases with the core-to-inner cladding area ratio and the pump light intensity in the inner cladding.

Traditional fibers have a practical limit to their core-to-cladding ratio. The maximum core diameter is limited by bend loss. And, while ultimately the minimum fiber diameter is limited by draw capabilities, in practice it is limited by the ability to couple pump light into a fiber having a limited inner-cladding NA. For a pump source of given brightness, the product of beam diameter and NA cannot be increased. Such light can be successfully coupled into a larger inner cladding with smaller NA or a smaller inner cladding with larger NA.

Air-clad fibers have other advantages over traditional double-clad fibers. The outer glass cladding in air-clad fibers is optically inactive. This allows the inner cladding diameter to be decoupled from the physical fiber diameter, eliminating draw-related

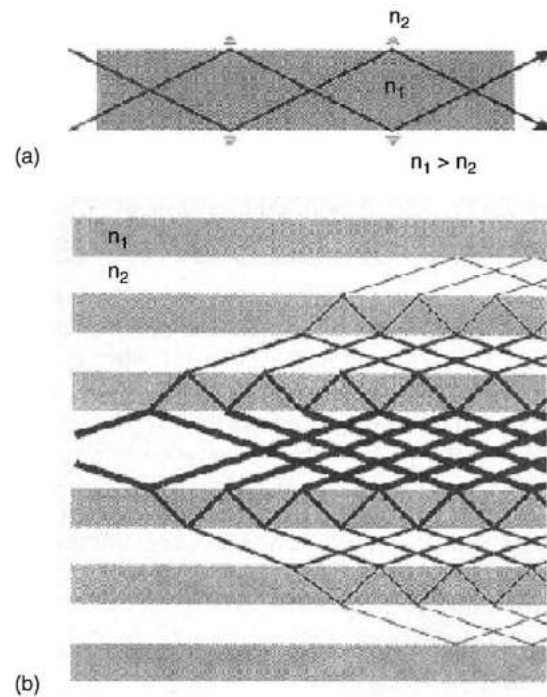


Fig. 12 Schematic of the mechanism of (a) index guided light and (b) band gap guided light. Reprinted with permission from Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285. Copyright 1999 American Association for the Advancement of Science.

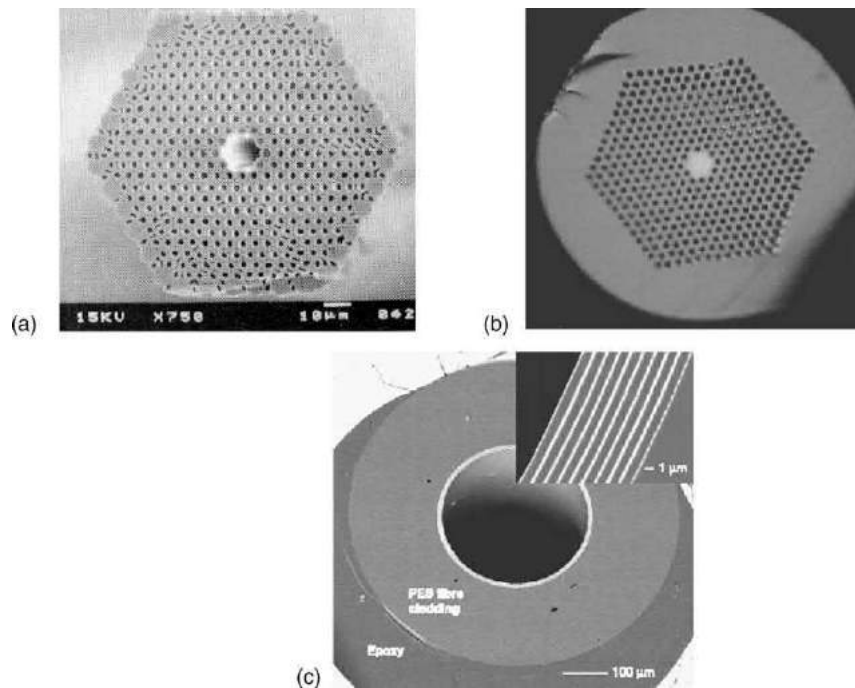


Fig. 13 Photographs and drawing of three types of bandgap fibers: (a) triangular array cladding with an air core (Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* vol. 285), (b) tunable with a silica core, (c) concentric rings with an air core. Reprinted with permission from Temelkuran *et al.* *Nature* 420. Copyright 1999 American Association for the Advancement of Science.

constraints on minimum inner cladding size. Since only the inner cladding is optically active, its size can be optimized (made smaller) while keeping the total fiber diameter at the standard 125 microns. In addition, the pump light does not interact with the polymer coating, which can be important if the polymer properties are affected by the surroundings.

Bandgap Guided Fibers

Bandgap fibers guide light in the core (also referred to as a defect) by confining the light through constructive interference due to Bragg scattering (Fig. 12). Unlike traditional fibers, this mechanism allows light to propagate in a core that has an index lower than the cladding. Bandgap fibers with an air core could theoretically be very low loss, propagate high powers, have a large effective area core, and exhibit very low nonlinearities. In addition, the fibers can be used in novel devices by filling the core with special gasses or liquids for nonlinear processes.

Bandgap fibers can be generalized into three types (Fig. 13). The first type of fiber has an air core surrounded by a cladding that consists of a periodic array of air holes in silica. Fibers of this type have been designed to guide a single mode and hold the greatest promise of guiding light over long distances with very low loss. The second type of bandgap fiber consists of a silica core and a silica cladding with a periodic array of holes which are filled with a high-index, tunable liquid. This design allows the bandgap properties to be tuned and is of interest for use in devices. The third fiber type consists of rings of alternating high- and low-index dielectric material with a center air core. They typically have a large, multimode core which is ideal for sending high powers. In addition, very little light is in the dielectric materials so the fiber can be designed to guide any wavelength with relatively little loss.

The spectrum of a bandgap fiber is quite different than an index guided fiber (Fig. 14). Frequencies that lie within the bandgap cannot propagate in the cladding and are confined to the defect in the lattice (the core). Bandgap fibers with ideal geometry are predicted to have losses over an order of magnitude lower than can be obtained in an ideal traditional fiber. Although the loss of bandgap fibers has not been demonstrated to be less than traditional fiber at telecommunication wavelengths, bandgap fibers have been shown to have much lower loss than traditional fibers at other wavelengths.

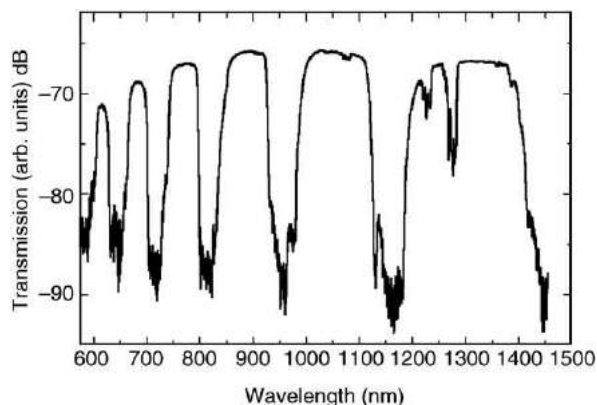


Fig. 14 Spectrum of a typical bandgap fiber.

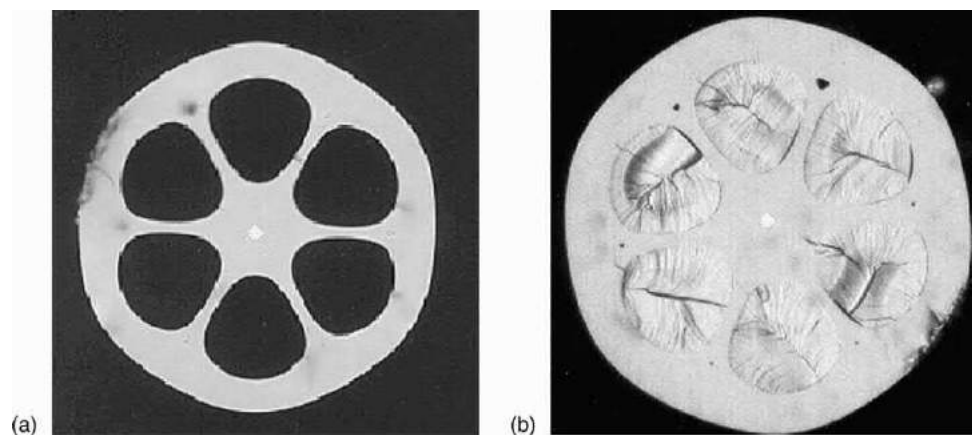


Fig. 15 Holes in the microstructure fibers are filled with controllable materials to affect the fiber properties actively. Photos show (a) original fiber and (b) fiber after polymer is inserted and cured in the microstructure fiber.

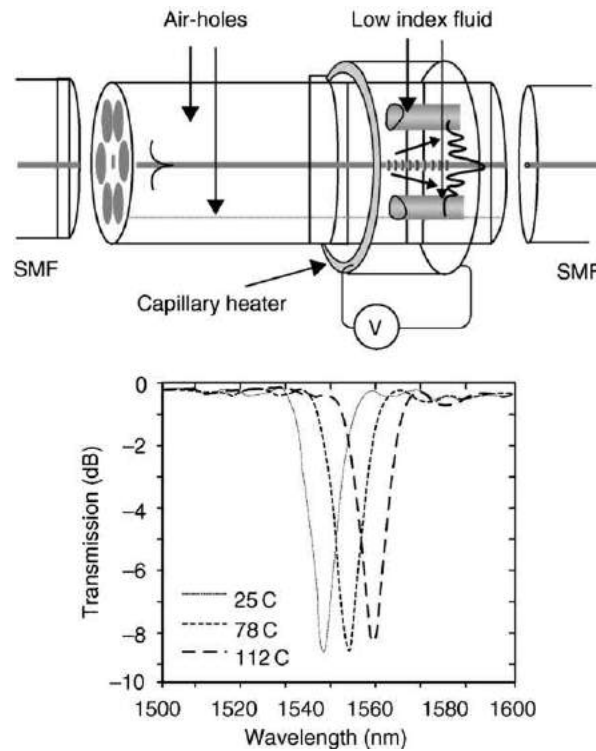


Fig. 16 Schematic of a tunable device in which low index liquid is positioned over the grating. The plot shows the shift in the spectrum as a function of temperature. Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier.

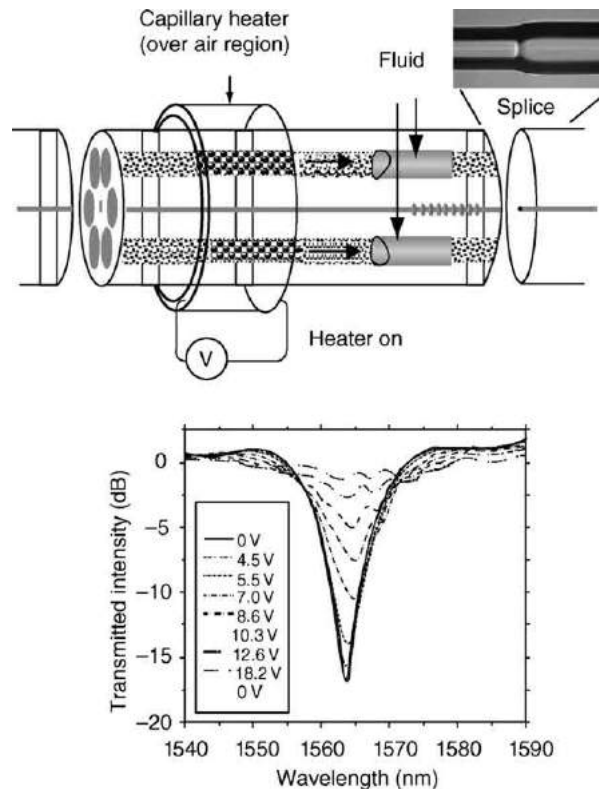


Fig. 17 Schematic of a tunable device in which a high index liquid plug is moved over the grating. The plot shows the intensity of the spectrum as a function of temperature. Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier.

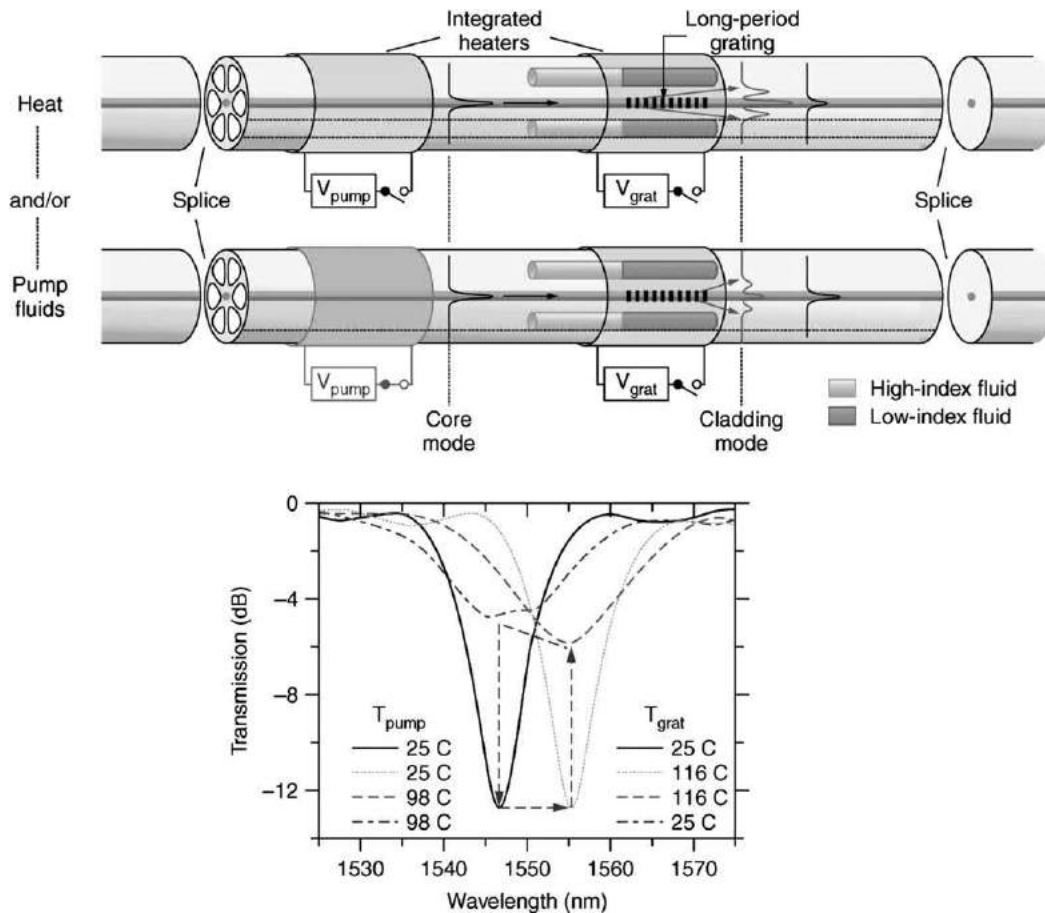


Fig. 18 Schematic of a tunable device using both high and low index liquids with two heaters. The plot shows that the position and strength of the spectrum are adjusted independently. Reproduced with permission from Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* 80(23). Copyright (2002) by the American Physical Society.

Tunable Microstructure Fiber Devices

Unique tunable devices are made by filling the air holes of microstructure fibers with materials whose properties can be controlled actively (Fig. 15). The materials can be positioned in the core or cladding to affect the corresponding modes. Most commonly, the index of refraction is changed using temperature sensitive liquids or polymers, although materials that respond to electric or magnetic fields enable a quicker response time. These material-filled active fibers can be used to fabricate devices like tunable filters, switches, broadband attenuators, and fibers with tunable birefringence. Below, a few examples of tunable filters are presented in which the device properties depend on the type and position of the liquid inserted in the fiber holes.

A microstructure tunable filter works by controlling the spectral position, shape, or strength of a longitudinal long-period grating written in the fiber. A grating is written in the fiber before filling the holes with tunable liquid. The position of the peak absorption wavelength is controlled by placing a liquid, with an index close to silica and strongly temperature dependent, in air holes in the cladding region over the grating (Fig. 16). The cladding index is tuned by varying the temperature of the liquid, using a small thin film heater located over the liquid. The position of the peak absorption wavelength moves as the cladding index is changed.

The strength or depth of the filter is controlled by inserting a moveable plug of high-index fluid in a sealed air section of a microstructure fiber (Fig. 17). The strength of the grating is determined by the fraction of the grating surrounded by high index fluid. The position of the liquid plug relative to the grating can be finely adjusted with a thin film heater located over the air region adjacent to the material plug. As heat is applied to the air, the air expands and pushes the liquid plug over the grating, decreasing the coupling of the fundamental mode to higher order cladding modes. Fig. 18 shows an example in which the position and strength of a filter are varied independently using two heaters and two adjacent materials.

See also: Fabrication of Optical Fiber

Further Reading

- Birks, T.A., Knight, J.C., Russell, P., 1997. Endlessly single mode photonic crystal fiber. *Optics Letters* 22 (13).
- Bjarklev A, Broeng J and Bjarklev AS, *Photonic Crystal Fibers*, Boston: Kluwer Academic Publishers.
- Broeng, J., Mogilevstev, D., Bjarklev, A., Barkou, S.E., 1999. Photonic crystal fibers: A new class of optical waveguides. *Optical Fiber Technology* vol. 5.
- Chandalia, J.K., Eggleton, B.J., Windeler, R.S., Kosinski, S.G., Liu, X., Xu, C., 2001. Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13 (1).
- Cregan, R.F., Mangan, B.J., Knight, J.C., Birks, T.A., Russell, P.J., Roberts, P.J., Allan, D.C., 1999. Single-mode photonic band gap guidance of light in air. *Science* 285.
- Eggleton, B.J., Westbrook, P.S., White, C.A., Kerbage, C., Windeler, R.S., Burdge, G.L., 2000. Cladding-mode resonance in air-silica microstructure optical fiber. *Journal Lightwave Technology* 18 (8).
- Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. *Photonic Crystals*. New Jersey: Princeton University Press.
- Johnson, S.G., Ibanesc, M., Skorobogatiy, M., Weisberg, M., Engeness, O., Soljacic, M., Jacobs, S.A., Joannopoulos, J.D., Fink, Y., 2001. Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9 (13).
- Kerbage, C., Hale, A., Yablon, A., Windeler, R.S., Eggleton, B.J., 2001. Integrated all fiber variable attenuator based on hybrid microstructure fiber. *Applied Physics Letters* 79 (19).
- Kerbage, C., Windeler, R.S., Eggleton, B.J., Dolinskoi, M., Mach, P., Rogers, J.A., 2002. Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204, 179–184.
- Knight, J.C., Birks, T.A., Russell, P.J., Atkin, D.M., 1996. All-silica single-mode optical fiber with photonic crystal cladding. *Optics Letters* 21 (19).
- Knight, J.C., Birks, T.A., Russell, P.J., 2001. Hole' silica fibers. In: Markel, V.A., George, T.F. (Eds.), *Optics of nanostructured materials*. New York: John Wiley & Sons.
- Mach, P., Dolinski, M., Baldwin, K.W., Rogers, J.A., Kerbage, C., Windeler, R.S., Eggleton, B.J., 2002. Tunable microfluidic optical fiber. *Applied Physics Letters* vol. 80.
- Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25 (1).
- Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25, 796–798.

Omnidirectional Surfaces and Fibers

S Hart, G Benoit, and Y Fink, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Elsevier Ltd. All rights reserved.

Omnidirectional Reflecting Mirrors

Mirrors are probably the most prevalent of optical devices. Known to the ancients and used by them as objects of worship and beauty, mirrors are currently employed for imaging, solar energy collection, and also in laser cavities. Their intriguing optical properties have captured the imagination of scientists as well as artists and writers.

One can distinguish between two types of mirrors, the age-old metallic, and the more recent multilayer dielectric. Metallic mirrors reflect light over a broad range of frequencies incident from arbitrary angles (i.e., omnidirectional reflectance). However, at infrared and optical frequencies a few percent of the incident power is typically lost due to absorption.

Multilayer dielectric mirrors are used primarily to reflect a narrow range of frequencies incident from a particular angle or particular angular range. Unlike their metallic counterparts, dielectric reflectors can be extremely low loss. The ability to reflect light of arbitrary angle of incidence for all-dielectric structures has been associated with the existence of a complete photonic bandgap, which can exist only in a system with a dielectric function that is periodic along three orthogonal directions. In fact, a sufficient condition for the achievement of omnidirectional reflection in a periodic system with an interface, is the existence of an overlapping bandgap regime in phase space above the light cone of the ambient media.

Consider a system that is constructed of alternating dielectric layers coupled to a homogeneous medium – characterized by n_0 (such as air with $n_0=1$), at the interfaces. Electromagnetic waves are incident upon the multilayer film from the homogeneous medium. While such a system was analyzed extensively in the literature the possibility of omnidirectional reflectivity was not recognized until recently.

The generic system is described by the index of refraction profile in Fig. 1, where h_1 and h_2 are the layer thickness, and n_1 and n_2 are the indices of refraction of the respective layers. The incident wave has a wavevector $\vec{k} = k_x \hat{e}_x + k_y \hat{e}_y$, and frequency of $\omega = c|k|$. The wavevector together with the normal to the periodic structure defines a mirror plane of symmetry which allows us to distinguish between two independent electromagnetic modes: transverse electric (TE) modes and transverse magnetic (TM) modes. For the TE mode the electric field is perpendicular to the plane, as is the magnetic field for the TM mode.

General features of the transport properties of the finite structure can be understood when the properties of the infinite structure are elucidated. In a structure with an infinite number of layers, translational symmetry along the direction perpendicular to the layers leads to Bloch wave solutions of the form:

$$u_K(x, y) = E_K(y) e^{iKx} e^{ik_x x} \quad (1)$$

where $E_K(y)$ is periodic, with a period of length a , and K is the Bloch wave number. These waves represent solutions to an eigenvalue problem and are completely and uniquely defined by the specification of K , k_x and ω .

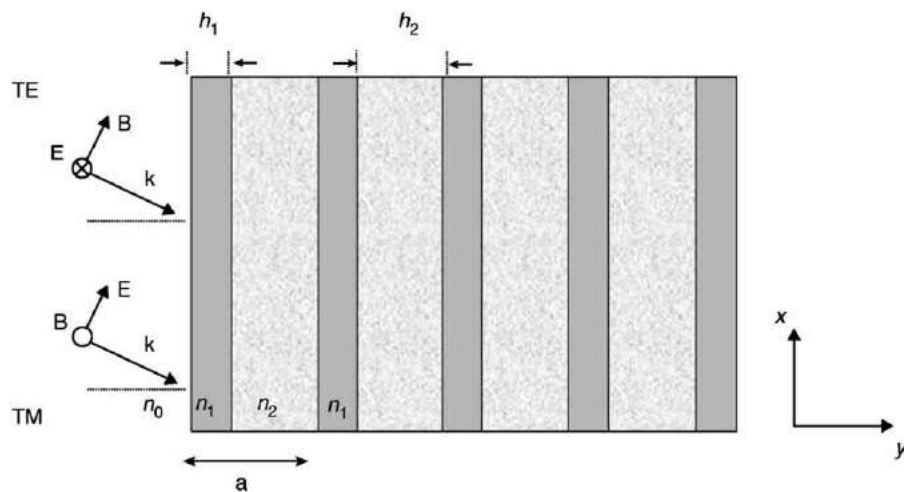


Fig. 1 Schematic of the multilayer system showing the layer parameters (n_x , h_x – index of refraction and thickness of layer x), the incident wavevector $\vec{k}$ and the electromagnetic mode convention.

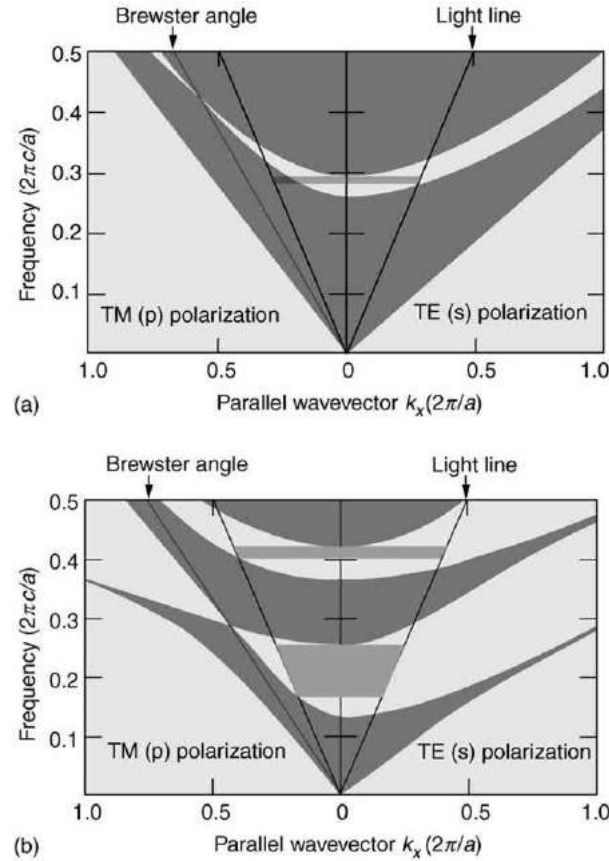


Fig. 2 (a) Projected bandstructure of a multilayer film with the light line and Brewster line, exhibiting a reflectivity range of limited angular acceptance with ($n_1=2.2$ and $n_2=1.7$ and a thickness ratio of $h_2/h_1=2.2/1.7$). (b) Projected bandstructure of a multilayer film together with the light line and Brewster line, showing an omnidirectional reflectance range at the first and second harmonic (propagating states – dark gray, evanescent states – white, omnidirectional reflectance range – light grey). The film parameters are $n_1=4.6$ and $n_2=1.6$ with a thickness ratio of $h_2/h_1=1.6/0.8$. These parameters are similar to the actual polymer–tellurium film parameters measured in the experiment. Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

Solutions can be propagating or evanescent, corresponding to real or imaginary Bloch wave numbers, respectively. It is convenient to display the solutions of the infinite structure by projecting the $\omega(K, k_x)$ function onto the $\omega - k_x$ plane; Fig. 2(a) and (b) are examples of such projected structures.

The gray background areas highlight phase space where K is strictly real, that is, regions of propagating states, while the white areas represent regions containing evanescent states. The shape of the projected bandstructures for the multilayer film can be understood intuitively. At $k_x=0$, the bandgap for waves traveling normal to the layers is recovered. For $k_x>0$, the bands curve upward in frequency. As $k_x \rightarrow \infty$, the modes become largely confined to the slabs with the high index of refraction and do not couple between layers (and are therefore independent of k_x).

In a finite structure, the translational symmetry in the directions parallel to the layers is preserved, hence k_x remains a conserved quantity and can be used to label solutions even though these solutions will no longer be of the Bloch form. The relevance of the band diagram to finite structures is that it allows for the prediction of regions of phase space where waves are evanescently decaying in the multilayer structure.

Since we are primarily interested in waves originating from the homogeneous medium external to the periodic structure, we will focus only on the portion of phase space lying above the light line. Waves originating from the homogeneous medium satisfy the condition $\omega \geq ck_x/n_0$, where n_0 is the refractive index of the homogeneous medium, and therefore they must reside above the light line. States of the homogeneous medium with $k_x=0$ are normal incident, and those lying on the $\omega=ck_x/n_0$ line with $k_y=0$ are incident at an angle of 90° .

The states in Fig. 2(a), that are lying in the restricted phase space defined by the light line and that have a (ω, k_x) corresponding to the propagating solutions (gray areas) of the crystal can propagate in both the homogeneous medium and in the structure. These waves will partially or entirely transmit through the film. Those with (ω, k_x) in the evanescent regions (white areas) can propagate in the homogeneous medium but will decay in the crystal – waves corresponding to this portion of phase space will be reflected off the structure.

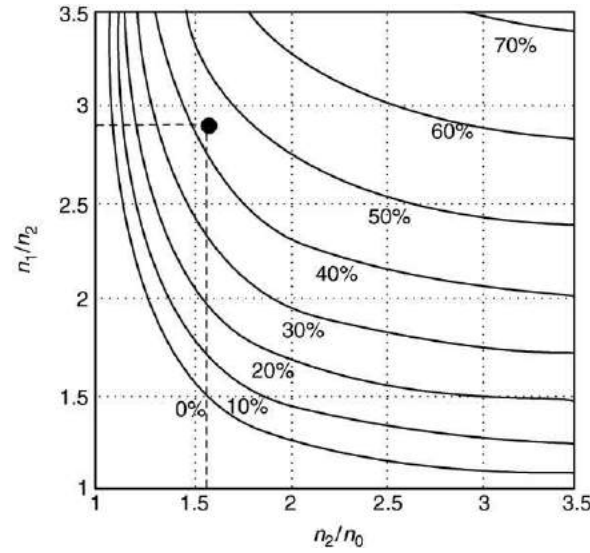


Fig. 3 The range to mid-range ratio $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$, for the fundamental frequency range of omnidirectional reflection, plotted as contours. Here, the layers were set to quarter wave thickness and $n_1 > n_2$. The ratio for our materials is approximately 45%. ($n_1/n_2 = 2.875$, $n_2/n_0 = 1.6$) is located at the intersection of the dashed lines (black dot). Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

The multilayer system leading to **Fig. 2(a)** represents a structure with a limited reflectivity cone, since for any frequency one can always find a k_x vector for which a wave at that frequency can propagate in the crystal – and hence transmit through the film. The necessary and sufficient criterion for omnidirectional reflectivity at a given frequency is that there exists no transmitting state of the structure inside the light cone – this criterion is satisfied by frequency ranges marked in light gray in **Fig. 2(b)**. In fact, the system leading to **Fig. 2(b)** exhibits two omnidirectional reflectivity ranges.

The omnidirectional range is defined from above by the normal incidence bandedge ω_h ($k_y = \pi/a$, $k_x = 0$) (**Fig. 2(b)**) and below by the intersection of the top of the TM allowed bandedge with the light line ω_l ($k_y = \pi/a$, $k_x = \omega_l/c$) (**Fig. 2(b)**).

A dimensionless parameter used to quantify the extent of the omnidirectional range is the range to mid-range ratio defined as $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$.

Fig. 3 is a plot of this ratio as a function of n_2/n_1 and n_1/n_0 , where ω_h and ω_l are determined by solutions of **Eq. (2)** with quarter wave layer thickness. The contours in this figure represent various equi-omnidirectional ranges for different material index parameters and could be useful for design purposes.

At normal incidence, there is no distinction between TM and TE modes. At increasingly oblique angles, the gap of the TE mode increases, whereas the gap of the TM mode decreases. In addition, the center of the gap shifts to higher frequencies. Therefore, the criterion for the existence of omnidirectional reflectivity can be restated as the occurrence of a frequency overlap between the gap at normal incidence and the gap of the TM mode at 90° . Analytical expressions for the range to mid-range ratio can be obtained by setting:

$$\omega_h = \frac{2c}{h_2 n_2 + h_1 n_1} \cos^{-1} \left(-\frac{|n_1 - n_2|}{n_1 + n_2} \right) \quad \omega_l = \frac{2c}{h_2 \sqrt{n_2^2 - 1} + h_1 \sqrt{n_1^2 - 1}} \times \cos^{-1} \left(\left| \frac{n_1^2 \sqrt{n_2^2 - 1} - n_2^2 \sqrt{n_1^2 - 1}}{n_1^2 \sqrt{n_2^2 - 1} + n_2^2 \sqrt{n_1^2 - 1}} \right| \right) \quad (2)$$

Moreover, the maximum range width is attained for thickness values that are not equal to the quarter wave stack though the increase in bandwidth gained by deviating from the quarter wave stack is typically only a few percent.

In general, the TM mode defines the lower frequency edge of the omnidirectional range; an example can be seen in **Fig. 2(b)** for a particular choice of the indices of refraction. This can be proven by showing that:

$$\left. \frac{\partial \omega}{\partial k_y} \right|_{\text{TM}} \geq \left. \frac{\partial \omega}{\partial k_y} \right|_{\text{TE}} \quad (3)$$

in the region that resides inside the light line. The physical reason for **Eq. (3)** lies in the vectorial nature of the electric field. In the upper portion of the first band the electric field concentrates its energy in the high dielectric regions. Away from normal incidence, the electric field in the TM mode has a component in the direction of periodicity, this component forces a larger portion of the electric field into the low dielectric regions. The group velocity of the TM mode is therefore enhanced. In contrast, the electric field of the TE mode is always perpendicular to the direction of periodicity and can concentrate its energy primarily in the high dielectric region. While the omnidirectional reflection criteria can be used to confine light in various geometries, it is the cylindrical fiber configuration that appears to present significant application opportunities.

Mesostructured Fibers for External Reflection Applications

Polymer fibers are ubiquitous in applications such as textile fabrics, due to their excellent mechanical properties and the availability of low-cost, high-volume processing techniques; however, the control over their optical properties has so far remained relatively limited. Conversely, dielectric mirrors are used to precisely control and manipulate light in high performance optical applications, but the fabrication of these typically fragile mirrors has been mostly restricted to planar geometries and typically involves multiplicity of deposition sequences in a high vacuum thin film deposition system.

Fabrication Approach

The fabrication of extended lengths of omnidirectionally reflecting fibers with large bandgaps and high layer counts poses considerable challenges. To illustrate the nature of this formidable task, one needs to consider the necessity of maintaining the uniformity of sub-100 nm layer thicknesses over kilometer-length scales; creating continuous layers with an aspect ratio of $\sim 10^{10}$ in a single process! To meet this and other challenges associated with the fabrication of mesostructured fibers, we have developed a preform-based fabrication approach (Fig. 4). A scaled-up version of the final fiber, called a preform, is fabricated which shares the geometry and materials of the final fiber but exhibits macroscopic lateral features. The preform is heated up and drawn under tension into the fiber using a simple cylindrical furnace. The macroscopic layers are reduced in the process to microscopic dimensions while maintaining the overall geometry and symmetry of the original preform. Conservation of mass determines the final length of the fiber – typically this length is equal to the lateral reduction factor squared multiplied by the length of the preform. The nature of the process and requirements on the fiber's optical properties lead to the definition of materials selection rules. First, in order to achieve omnidirectional reflectivity, one needs to identify two solid materials exhibiting an index contrast given by the plot in Fig. 3. Second, to enable the codrawing of the two dissimilar materials both will need to have viscosities that are lower than $\sim 10^8$ poise at the drawing temperature. In order to maintain high draw speeds, the majority component needs to be amorphous and the adhesion between these two materials needs to be sufficient to prevent delamination. Finally, their thermal expansion coefficients need to be close or alternatively at least one of the materials needs to be capable of relieving the stress due to CTE mismatch.

Materials Selection Criteria

Pairs of materials that are compatible with the process and fiber property requirements have been identified, including: high glass transition temperature (T_g), thermoplastic polymers such as poly(ether-sulfone) (PES), poly (ether-imide) (PEI) and members of the chalcogenide glass family, such as arsenic triselenide (As_2Se_3) or arsenic trisulfide. Pairs of these materials will have substantially different refractive indices, as shown in Fig. 5, which is a measurement of the real and imaginary indices of refraction of PES and As_2Se_3 , obtained using a broadband spectroscopic ellipsometer (SOPRA GES5). They nevertheless exhibit similar thermo-mechanical properties within a certain thermal processing window.

Adhesion and extensional viscosity in the fluid state are difficult to measure in general, and the measurement of high-temperature surface tension is quite involved. Thus, limited data on these properties are available and it was necessary to empirically identify materials that could be used to draw out mirror-fibers. Various high-index chalcogenide (S, Se, and Te containing) glasses and low-index polymers were identified as potential candidates based on their optical properties and overlapping thermal softening regimes. Adhesion and viscosity matching were tested by thermal evaporation of a chalcogenide glass layer on top of a polymer film or rod and elongation of the coated substrate at elevated temperatures. The choice of a high-temperature polymer, PES, and a simple chalcogenide glass, As_2Se_3 , resulted in excellent thermal co-deformation without film cracking or delamination. Approximate matching of extensional viscosity in this manner was also demonstrated using As_2Se_3 and PEI. The properties, processing, and applications of chalcogenide glasses have been explored extensively elsewhere. One advantage in choosing As_2Se_3 for this application is that not only is it a stable glass, but it is a stoichiometric compound that can be readily deposited in thin films through thermal evaporation or sputtering without dissociation. Additionally, As_2Se_3 is transparent to IR

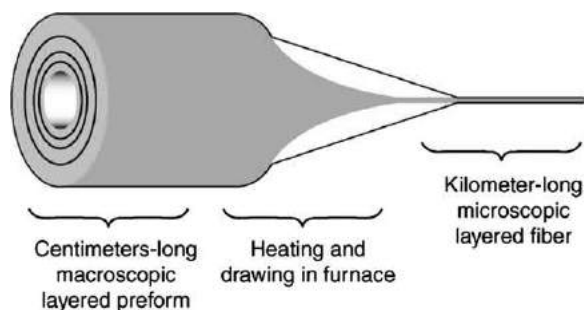


Fig. 4 Conceptual preform based fabrication process for meso-structured fibers.

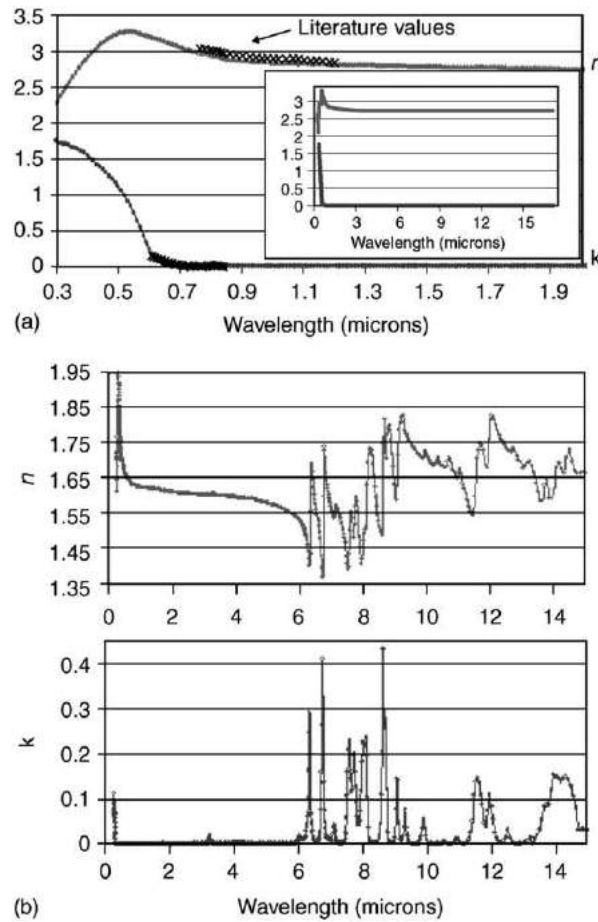


Fig. 5 (a) Real part (n – light gray) and imaginary part (k – dark gray) of the refractive index of annealed As_2Se_3 . The black crosses correspond to literature values. (b) Real part (n – upper) and imaginary part (k – lower) of the refractive index of PES.

radiation from approximately 0.8 to 17 μm as shown in the ellipsometric data (Fig. 5(a)), and has a refractive index of ~ 2.8 in the mid-IR. PES is a high-performance, dimensionally stable thermoplastic with a refractive index of ~ 1.6 and good transparency to EM waves in a range extending from the visible regime into the mid-IR ($\sim 5 \mu\text{m}$), as shown in Fig. 5(b).

Preform Construction Process and Fiber Draw

The selected materials were used to construct a multilayer preform rod, which essentially is a macroscale version of the final fiber. In order to fabricate the dielectric mirror fiber preform, an As_2Se_3 film was deposited through thermal evaporation on either side of a free-standing PES film which was then rolled on top of a PES tube substrate, forming a structure having 21 alternating layers of PES and As_2Se_3 , as shown in Fig. 6.

The resulting multilayer fiber preform was subsequently drawn down using an optical fiber draw tower into hundreds of meters of multilayer fiber with a precisely controlled submicron layer thickness, creating a photonic bandgap in the mid-IR. Fibers of outer diameters varying from 175–500 μm with a typical standard deviation of 10 μm from target, were drawn from the same preform to demonstrate adjustment of the reflectivity spectra through thermal deformation. The spectral position of the photonic bandgap was controlled by the optical monitoring of the outer diameter (OD) of the fiber during draw, which was later verified by reflectivity measurements on single and multiple fibers of different diameters. Scanning electron micrographs (SEMs) of the cross-section of these fibers are depicted in Fig. 7.

Bandstructure for Multilayer Fibers for External Reflection Applications

In theoretically predicting the spectral response of these fibers, it is helpful to calculate the photonic bandstructure that corresponds to an infinite one-dimensional photonic crystal. This allows for the analysis of propagating and evanescent modes in the structure, corresponding to real or imaginary Bloch wave number solutions.

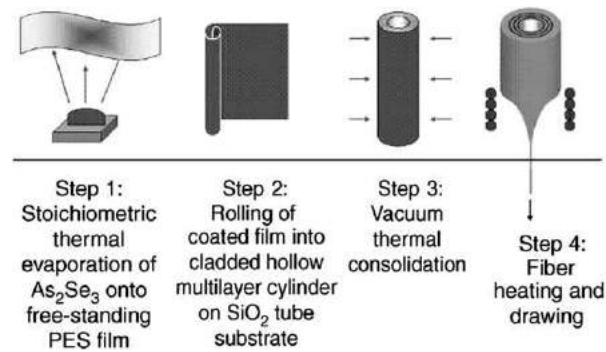


Fig. 6 Multilayer preform fabrication sequence.

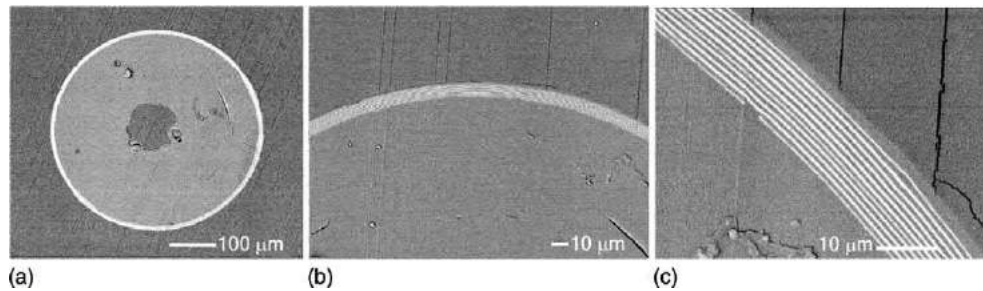


Fig. 7 SEM micrographs of 400 μm OD fiber cross-section. The entire fiber is embedded in epoxy. (a) shows the entire fiber cross-section, with mirror structure surrounding the PES core; (b) demonstrates that the majority of the fiber exterior is free of significant defects and that the mirror structure adheres well to the fiber substrate; and (c) reveals the ordering and adhesion within the alternating layers of As_2Se_3 (bright layers) and PES. Stresses developed during sectioning caused some cracks in the mounting epoxy that are deflected at the fiber interface. Fibers from this batch were used in the reflectivity measurements recorded below in **Fig. 9(a)**. Reproduced with permission from Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513. Copyright 2002 American Association for the Advancement of Science.

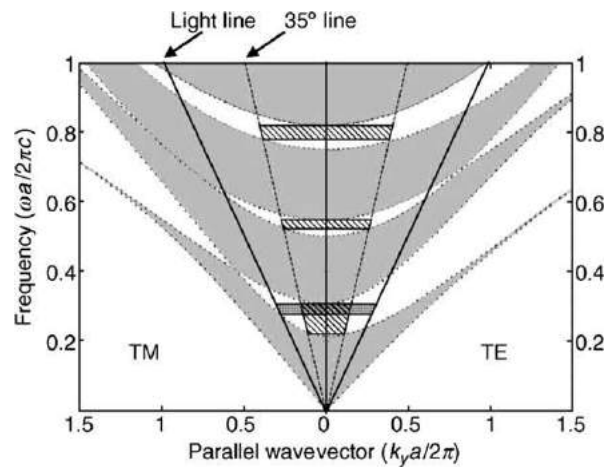


Fig. 8 Photonic band diagram for a one-dimensional photonic crystal having a periodic refractive index alternating between 2.8 and 1.55. Gray regions represent propagating modes within the structure, while white regions represent evanescent modes. Hatched regions represent photonic bandgaps where high reflectivity can be expected for external EM waves over an angular range extending from normal to 35° incidence. The lower dark shaded trapezoid represents a region of external omnidirectional reflection.

The electric or magnetic field vector is parallel to the mirror layer interfaces for the TE and TM polarized modes, respectively. The parallel wavevector (k_y) is the component of the incident electromagnetic (EM) wavevector that is parallel to the layer interfaces. The phase space accessible from an external ambient medium is contained between the light lines (defined by the glancing-angle condition $\omega = ck_y/n_0$), and the modes between the 35° lines correspond to those sampled experimentally. Axes are normalized to the thickness a of one mirror bilayer (a period consisting of one high- and one low-index layer). **Fig. 8** depicts

the photonic band diagram for an infinite structure having similar periodicity and refractive indices to the mirror structures fabricated here. Three photonic bandgaps are present where high reflectivity is expected within the 0–35° angular range, and the fundamental gap contains a region of external omnidirectional reflectivity.

Optical Characterization of ‘Mirror Fibers’

Mirror fiber reflectivity was measured from both single fibers and parallel fiber arrays using a Nicolet/SpectraTech NicPlan Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860). The microscope objective (SpectraTech 15×, Reflachromat) used to focus on the fibers had a numerical aperture (NA) of 0.58. This results in a detected cone where the angle of reflection with respect to the surface normal of the structure could vary from normal incidence to ~35°, which is determined by the NA of the microscope objective. As a background reference for the reflection measurements, we used gold-coated PES fibers of matching diameters. Dielectric mirror fibers, drawn to 400 μm OD, exhibited a very strong reflection band centered at 3.4 μm wavelength (Fig. 9(a)). Measured reflectivity spectra agree well with planar-mirror transfer matrix method (TMM) simulations, where the reflectivity was averaged across the aforementioned angular range for both polarization modes. Fibers drawn down to 200 μm OD show a similar strong fundamental reflection band centered near 1.7 μm (Fig. 9(b)). This shifting of the primary photonic bandgap clearly illustrates the precise tuning of the reflectivity spectra over wide frequency ranges through thermal deformation processing. Strong optical signatures are measurable from single fibers as small as 200 μm OD. Fiber array measurements, simultaneously sampling reflected light from multiple fibers, agree quite well with single-fiber data (Fig. 9(b)).

These reflectivity results are strongly indicative of uniform layer thickness control, good interlayer adhesion, and low interdiffusion through multiple thermal treatments. This was confirmed by SEM inspection of fiber cross-sections (Fig. 7). The layer thicknesses observed ($a=0.90$ μm for the 400 μm fibers; $a=0.45$ μm for the 200 μm fibers) correspond well to the measured reflectivity spectra. The fibers have a hole in the center, due to the choice of a hollow rod as the preform substrate, which experienced some nonuniform deformation during draw. The rolled-up mirror structure included a double outer layer of PES for mechanical protection, creating a noticeable absorption peak in the reflectivity spectrum at ~3.2 μm (Fig. 9(a)).

A combination of spectral and direct imaging data demonstrates excellent agreement with the photonic band diagram. The measured gap width (range to mid-range ratio) of the fundamental gap for the 400 μm OD fiber is 27%, compared to 29% in the photonic band diagram.

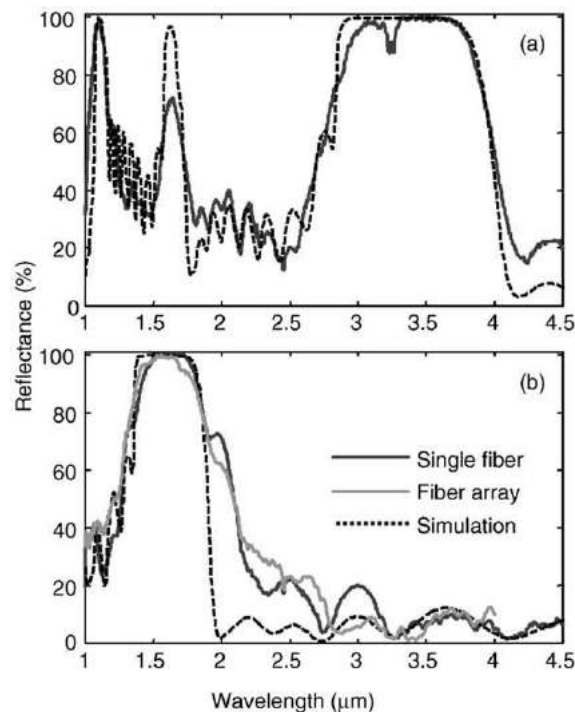


Fig. 9 Measured reflectance spectra for 400 μm OD (a) and 200 μm OD (b) dielectric mirror fibers relative to gold-coated fibers of the same diameter. (a) shows a single-fiber reflectivity measurement, while (b) compares single-fiber reflectivity to that measured from a multifiber array. Simulations were performed using the transfer matrix method.

Tunable 'Fabry-Perot' Fibers

The fabrication of fibers surrounded by or lined with alternating layers of materials with a large disparity in their refractive indices, presents interesting opportunities for passive and active optical devices. While a periodic multilayer structure, such as the one reported above, leads to the formation of photonic bandgaps and an associated range of high reflectivity, it is the incorporation of intentional deviations from periodicity, also called 'defects', which allows for the creation of localized electromagnetic modes in the vicinity of the defect. These structures, sometimes called optical cavities, in turn can provide the basis for a large number of interesting passive and active optical devices such as vertical cavity surface emitting lasers (VCSELs), bi-stable switches, tuneable dispersion compensators, tuneable drop filters, etc. Here we report on the fabrication of a fiber surrounded by a Fabry-Perot cavity structure and demonstrate that by application of axial mechanical stress, the spectral position of the resonant Fabry-Perot mode can be reversibly tuned.

Structure and Optical Properties of the Fabry-Perot Fibers

The fabrication technique described above allows for the accurate placement of optical cavities that can encompass the entire or partial fiber circumference. The fibers discussed above are made of As_2Se_3 and PES and have a low index Fabry-Perot cavity (Fig. 10).

This structure was achieved by introducing an extra polymer layer in the middle of the periodic multilayer structure of the preform, thus generating a defect mode in the photonic bandgaps of the drawn fibers. The position of the bandgap center is linearly related to the optical thickness of the layers by the Bragg condition. Through accurate outer diameter control afforded by the laser micrometer mounted on the draw tower we have been able to place gaps (Fig. 11) at wavelengths ranging from

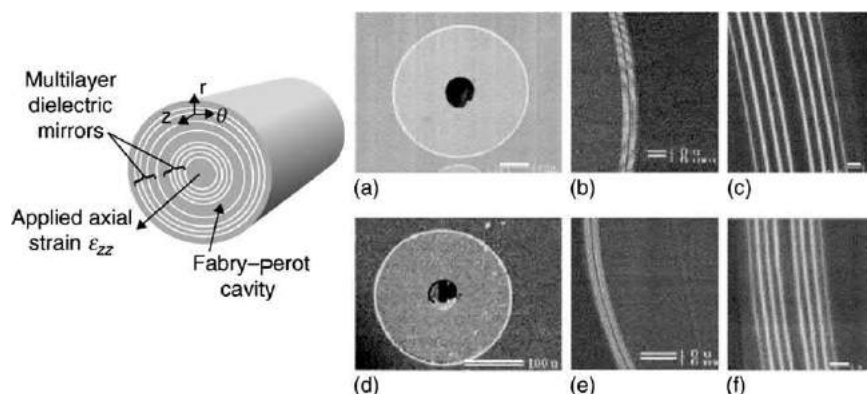


Fig. 10 Schematic of the structure of dielectric mirror fibers made of As_2Se_3 (light gray) and PES (dark gray) with low index Fabry-Perot cavity. The local cylindrical coordinate system is represented as well as the applied axial strain ϵ_{zz} . Typical radii are $\sim 150 \mu\text{m} \pm 100 \mu\text{m}$. Backscattered SEM micrographs of the cross-sections of a 460 micron diameter fiber (a, b, c) and of a 240 micron diameter fiber (d, e, f) embedded in epoxy and microtomed. (a) and (d) show the entire cross-section of the fibers, (b) and (e) demonstrate long-range layer uniformity, and (c) and (f) reveal the ordering and adhesion of the Fabry-Perot cavity structure. Bar scales have been redrawn for clarity. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

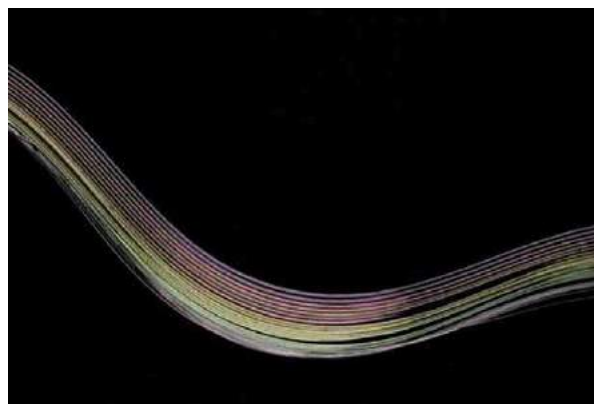


Fig. 11 Array of parallel fibers with outer diameter ranging from $\sim 420 \mu\text{m}$ (bottom) to $\sim 100 \mu\text{m}$ (top). The colors are due to the narrow 4th order photonic bandgap in the visible.

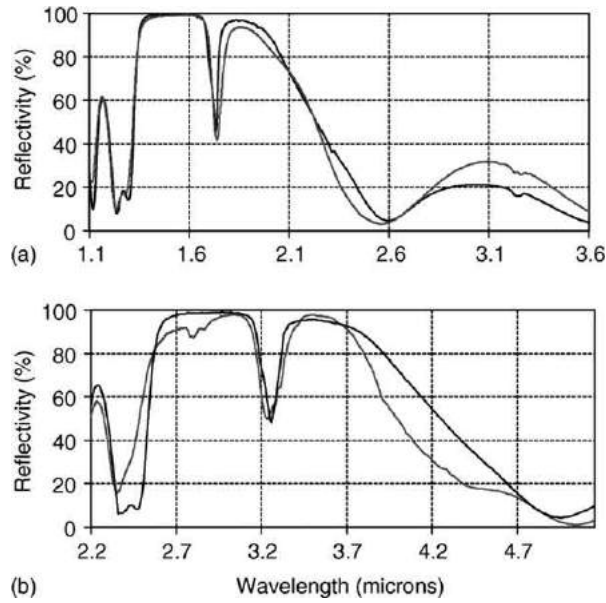


Fig. 12 Computed (black lines) and measured (grey lines) reflectivity spectra for the 240 micron (a) and of the 460 micron diameter fibers (b) with Fabry-Perot resonant modes at 1.74 and 3.2 μm , respectively. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

11 microns to below 1 micron. Such cost-effective tuneable optical filters could lead to applications such as optical switches for wavelength-division-multiplexing (WDM) systems and sensors.

The cross-sectional structures of a 240 micron and a 460 micron diameter fiber were observed by an SEM using a backscattered electron detector (Fig. 10). The structure is composed of a hollow core polymer rod surrounded by six bilayers of As_2Se_3 and PES, separated in the middle by an extra polymer layer, which forms the Fabry-Perot cavity. An extra polymer layer protects the fiber surface. For the 240 micron (460 micron) diameter fiber, the glass layers are ~ 135 nm (250 nm) thick, except for the first and the last ones which are half as thick due to the fabrication technique; the polymer layers are ~ 270 nm (540 nm) thick and the defect layer is ~ 610 nm (1170 nm) thick.

Reflectivity spectra measurements were performed under a microscope using a (Nicolet SpectraTech NicPlan) Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860) with a lens numerical aperture (NA) corresponding to 30 degrees of angular spread. They exhibited a single-mode Fabry-Perot resonant mode at 1.74 and 3.2 μm for the 240 and 460 micron diameter fiber, respectively (Fig. 12).

Because of the range of incident angles, the measured quality factor (Q-factor, defined as its spectral position divided by the full width half maximum (FWHM)) was equal to 31. Using the different thicknesses and the real and imaginary part of the refractive index (n & k) of As_2Se_3 and PES carefully measured with a broadband (300 nm to 16 microns) spectroscopic ellipsometer (Sopra GES-5), the reflectivity spectra of these fibers were computed with the TMM by approximating them as a one-dimensional planar stacking. By averaging the calculated spectra over the accessible incident angles and polarizations, we obtained a very good agreement between the simulation and the measurements.

Simulation of the Opto-Mechanical Behavior of the Fabry-Perot Fibers

We focused our analysis on the elastic regime, which ultimately limits the operational range of these fibers. The fiber's Young's modulus (E) can be approximated by modeling this multilayer structure as independent parallel springs under axial (along the z -axis of the fiber) strain assumed to be equal for all the layers, leading to a Young's modulus of 2.64 GPa (using $E_{\text{PES}} = 2.4$ GPa for bulk PES and $E_{\text{As}_2\text{Se}_3} = 15$ GPa reported for 1.5 μm thick films).

Neglecting the possible strain-induced refractive index variation of the materials, the normalized shift of the Fabry-Perot resonant mode can be related to the applied axial strain by calculating the radial stresses σ_r (resulting from the difference between the Poisson ratios of the materials – $\nu_{\text{PES}} = 0.45$ and $\nu_{\text{As}_2\text{Se}_3} = 0.289$ – and the adhesion condition between the layers) and displacements u_r in each layer under axial strain. Starting from the equilibrium equations:

$$\frac{\partial \sigma_{zz}}{\partial z} = 0 \quad (4)$$

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r} = 0 \Leftrightarrow \frac{d^2 u_r}{dr^2} + \frac{1}{r} \frac{du_r}{dr} - \frac{u_r}{r^2} = 0 \quad (5)$$

and using Hooke's law and Lamé's equations, we can derive a general expression for u_r and σ_{rr} with two unknowns A and B per layer:

$$u_r = \frac{A}{r} + Br \quad (6)$$

$$\sigma_{rr} = -\frac{2GA}{r^2} + 2(\eta + G)B + \lambda \varepsilon_{zz} \quad (7)$$

where $G = E/[2(1 + \nu)]$ is the shear modulus and $\eta = E\nu/[(1 + \nu)(1 - 2\nu)]$ the Lamé modulus. The continuity of the displacement and the radial stress at each interface, plus the boundary conditions (σ_{rr} vanishes at free surfaces) can be expressed as a linear system whose unique solution allows us to relate the radial strain in each layer to the applied axial strain as a linear relation $\varepsilon_{rr} = C\varepsilon_{zz}$, where C can be interpreted as an effective Poisson ratio. Finally, taking into account that the Fabry–Perot resonant mode itself is linearly shifted within the bandgap because of the different effective Poisson ratios of the glass and polymer layers, we obtain a linear relation between the normalized shift of the Fabry–Perot resonant mode and the applied axial strain:

$$\frac{\Delta\lambda}{\lambda} = -0.373\varepsilon_{zz} \quad (8)$$

All the layers (except the outer protective polymer layer) are under tensile radial stress whose maximum (0.22 MPa under 1% axial strain) is located at the interface between the layers and the polymer core where delamination is most likely to occur.

Mechanical Tuning Experiment and Discussion

Measurements were performed on fibers ~ 30 cm long, which were fixed at one end with epoxy to a load cell (Transducer Techniques MDB-2.5) while the other end was attached with strong tape to a pole mounted on a stepper rotational stage (Newport PR50) (Fig. 13). This end of the fiber was also screwed to the pole to further secure it in place.

The diameter of the pole was equal to 2.1 cm, leading to a normalized shift precision below 0.005%. The uncertainty on the Young's modulus, due to the precision of the load cell, was lower than 30 MPa. All the reflectivity spectra were normalized to a background taken with a flat gold mirror. The measurements were realized as far as possible from the fixed ends of the fiber where edge effects are likely to occur. These edge effects result in a reduction of the length of the fiber that deforms uniformly and consequently increase the real strain far from the edges by a factor of 1.14 and 1.15 for the 240 and the 460 micron diameter fiber, respectively (determined experimentally by measuring the position of two reference points on the fiber) compared to the strain calculated from the rotation of the pole. Moreover, to avoid measuring slight variations in the spectral position of the bandgap resulting from outer diameter variations, typically of the order of 4–5 microns over meters of fibers, the measurements were realized at a fixed reference position on the fiber.

By focusing on the fundamental bandgap of the 240 and the 460 micron diameter fiber, we demonstrated the tuning of the Fabry–Perot resonant mode under increasing axial strain (Fig. 14).

A drop in the reflectivity of 13% was observed at $1.71 \mu\text{m}$ (dash line) for the 240 micron diameter fiber when increasing the applied axial strain from 0.23% (light gray) to 1.07% (dark gray). The normalized-shift versus strain curves (Fig. 15) appeared to be linear for both fibers up to approximately 0.9% (dash line) with a slope equal to -0.3859 and -0.3843 for the 240 and the 460 micron diameter fiber, respectively, close to the predicted value (-0.373 , Eq. (8)).

The normalized shift was equal to -0.347% for 0.9% applied axial strain, which seems to be the limit of the elastic regime and corresponds to small applied loads: 76 and 300 g for the 240 and 460 micron diameter fiber, respectively. Under higher strains, the normalized-shift versus strain curves were no longer linear and the measured loads would start decreasing slightly with time, which could be a sign of delamination between the layers and the polymer core, of plastic deformation and/or of relaxation in the layers. The stress–strain curves also exhibit an elastic regime up to approximately 1% strain with a corresponding Young's modulus equal to 2.59 GPa for the 240 μm OD fiber and to 2.39 GPa for the 460 μm OD fiber. These values are slightly lower than the predicted value, possibly due to relaxation effects in the polymer and the glass layers (for As_2Se_3 and PES, $T_g = 175$ and 220°C , respectively).

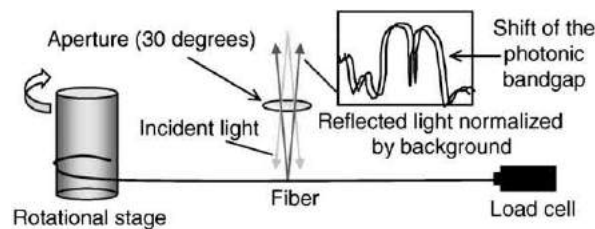


Fig. 13 Experimental setup used for mechanical tuning demonstration. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

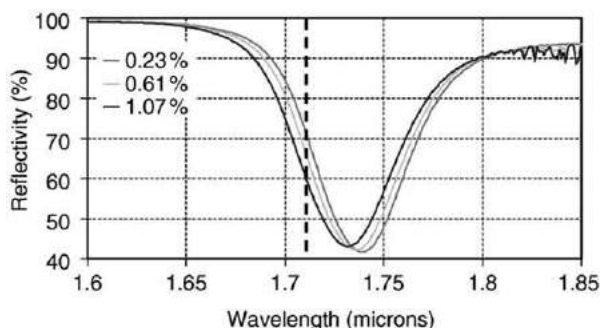


Fig. 14 Reflectivity versus wavelength plot showing the shift of the Fabry–Perot resonant mode of the 240 micron diameter fiber for three increasing values of the applied axial strain. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

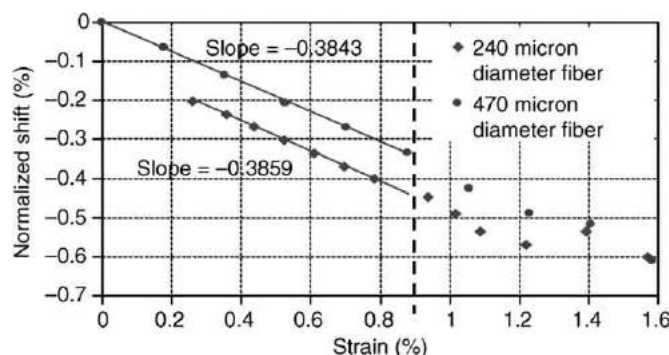


Fig. 15 Normalized shift of the Fabry–Perot resonant mode (defined as $\Delta\lambda/\lambda$) versus applied axial strain for the 240 micron (diamonds) and the 460 micron diameter fiber (dots). The curve obtained for the 240 micron diameter fiber is lower because the experimental 0% strain is likely to correspond to a nonzero positive strain, necessary to keep this thinner fiber straight for the measurement (equivalent to a load less than 10 g). Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

Wavelength-Scalable Hollow Optical Fibers with Large Photonic Bandgaps for CO₂ Laser Transmission

Hollow optical transmission fibers offer the potential to circumvent fundamental limitations associated with conventional index guided fibers and thus have been the subject of active research in recent years. Here we report on the materials selection, design, fabrication, and characterization of extended lengths of hollow optical fiber lined with an interior omnidirectional dielectric mirror. These fibers consist of a hollow air core surrounded by multiple alternating submicron-thick layers of a high-refractive-index glass and a low-index polymer, resulting in large infrared photonic bandgaps. These gaps provide strong confinement of optical energy in the hollow fiber core and lead to light guidance in the fundamental and up to fourth-order gaps. We show that the fiber transmission windows can be scaled over a large wavelength range covering at least 0.75 to 10.6 microns. The utility of our approach is further demonstrated by the design and fabrication of tens of meters of hollow photonic bandgap fibers for 10.6 micron radiation transmission. We demonstrate transmission of carbon dioxide (CO₂) laser light with high power-density through more than 4 meters of hollow fiber and measure the losses to be less than 1.0 dB/m at 10.6 microns. This establishes suppression of fiber waveguide losses by orders of magnitude compared to the intrinsic fiber material losses.

Silica optical fibers have been extremely successful in telecommunications applications, and other types of solid-core fibers have been explored at wavelengths where silica is not transparent. However, all fibers that rely on light propagation principally through a solid material have certain fundamental limitations stemming from nonlinear effects, light absorption by electrons or phonons, material dispersion, and Rayleigh scattering that limit maximum optical transmission power and increase attenuation losses. These limitations have, in turn, motivated the study of a fundamentally different light-guiding methodology: the use of hollow waveguides having highly reflecting walls. Light propagation through air in a hollow fiber eliminates or greatly reduces the problems of nonlinearities, thermal lensing, and end-reflections, facilitating high-power laser guidance and other applications which may be impossible using conventional fibers. Hollow metallic or metallo-dielectric waveguides have been studied fairly extensively and found useful practical application, but their performance has been bounded by the notable losses occurring in metallic reflections at visible and infrared (IR) wavelengths, as well as by the limited length and mechanical flexibility of the fabricated waveguides. Hollow all-dielectric fibers, relying on specular or attenuated total reflection, have also been explored, but

high transmission losses have prevented their broad application. More recently, all-dielectric fibers, consisting of a periodic array of air holes in silica, have been used to guide light through air using narrow photonic bandgaps. Solid-core, index-guiding versions of these silica photonic crystal fibers have also been explored for interesting and important applications, such as very large core single-mode fibers, nonlinear enhancement and broadband supercontinuum generation, polarization maintenance, and dispersion management. However, the air-guiding capabilities of such waveguides thus far remain inferior to transmission through solid silica, due to various factors such as the difficulties in fabricating long, uniform fibers which must have a high volume fraction of air and many air-hole periods, as well as by the large electromagnetic (EM) penetration depths associated with the small photonic bandgaps achievable in these air-silica structures. In our fiber, the hollow core is surrounded by a solid high-refractive-index-contrast multilayer structure leading to large photonic bandgaps and omnidirectional reflectivity. The pertinent theoretical background and recent analyses indicate that such fibers may be able to achieve ultralow losses and other unique transmission properties. The large photonic bandgaps result in very short EM penetration depths within the layer structure, significantly reducing radiation and absorption losses while increasing robustness. Omnidirectional reflectivity is expected to reduce intermode coupling losses.

To achieve high index contrast in the layered portion of the fiber, we combined a chalcogenide glass with a refractive index of ~ 2.8 As_2Se_3 , and a high-performance polymer with a refractive index of ~ 1.55 PES. We recently demonstrated that these materials could be thermally co-drawn into precisely layered structures without cracking or delamination, even under large temperature excursions. The same polymer was used as a cladding material, resulting in fibers composed of $\sim 98\%$ polymer by volume (not including the hollow core) and thus combine high optical performance with polymeric processability and mechanical flexibility. We fabricated a variety of fibers by depositing a 5–10 micron thick As_2Se_3 layer through thermal evaporation onto a 25–50 micron thick PES film and the subsequent ‘rolling’ of that coated film into a hollow multilayer tube called a fiber preform. This hollow macroscopic preform was consolidated by heating under vacuum and clad with a thick outer layer of PES; the layered preform was then placed in an optical fiber draw tower and drawn down into tens or hundreds of meters of fiber having well-controlled submicron layer thicknesses. The nominal positions of the photonic bandgaps were determined by laser monitoring of the fiber OD during the draw process. Typical standard deviations in the fiber OD were $\sim 1\%$ of the OD. The resulting fibers were designed to have large hollow cores, useful in high-energy transmission.

SEM analysis (Fig. 16) reveals that the drawn fibers maintain proportionate layer thickness ratios and that the PES and As_2Se_3 films adhere well during rigorous thermal cycling and elongation. Within the multilayer structure shown in Fig. 1, the PES layers (gray) have a thickness of 900 nm, and the As_2Se_3 layers (bright) are 270 nm thick (except for the first and last As_2Se_3 layers, which are 135 nm). Broadband fiber transmission spectra were measured with a Fourier transform infrared (FTIR) spectrometer (Nicolet Magna 860), using a parabolic mirror to couple light into the fiber and an external detector. The results of these measurements are shown in the lower panel of Fig. 2 for fibers having two different layer structures. For each spectrum, light is guided at the fundamental and high-order photonic bandgaps. Also shown in the upper panel of Fig. 2, is the corresponding photonic band diagram for an infinite periodic multilayer structure calculated using the experimental parameters of our fiber (layer thicknesses and indices). Good agreement is found between the positions of the measured transmission peaks and the calculated bandgaps, corroborated with the SEM-measured layer thicknesses, verifying that transmission is dominated by the photonic bandgap mechanism. In order to demonstrate ‘wavelength scalability’ (i.e., the control of transmission through the fiber’s structural parameters) another fiber was produced, having the same cross-section but with thinner layers. We compared the transmission spectra for the original 3.55 micron bandgap fibers to the fiber with the scaled-down layer thicknesses.

Fig. 17 shows the shifting of the transmission bands, corresponding to fundamental and high-order photonic bandgaps, from one fiber to the next. The two fibers analyzed in Fig. 17 were fabricated from the same fiber preform using different draw-down

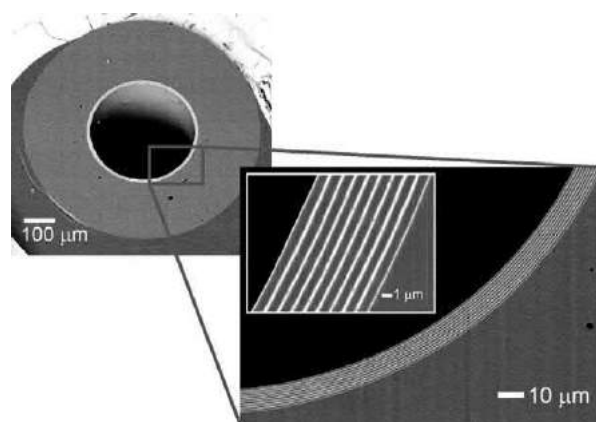


Fig. 16 Cross-sectional SEM micrographs at various magnifications of hollow cylindrical multilayer fiber mounted in epoxy. The hollow core appears black, the PES layers and cladding gray, and the As_2Se_3 layers bright white. This fiber has a fundamental photonic bandgap at a wavelength of ~ 3.55 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

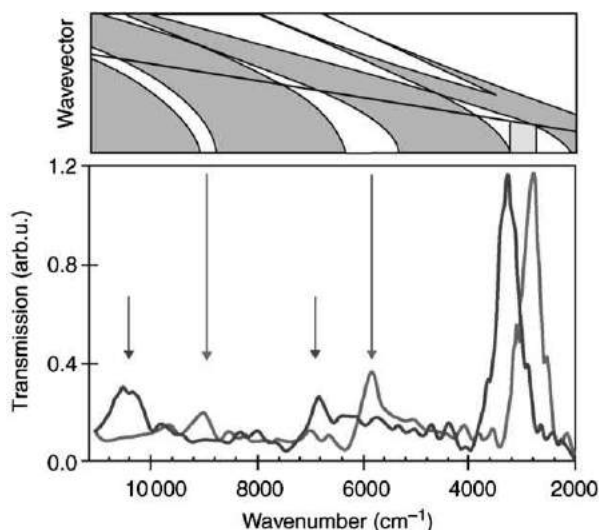


Fig. 17 Upper panel: Calculated photonic bandstructure associated with the dielectric mirror lining of the hollow fiber. Modes propagating through air and reflected by the fiber walls lie in the bandgaps (white) and within the light cone defined by the glancing-angle condition (black line). The gray regions represent modes radiating through the mirror. The fundamental bandgap has the widest range-to-mid-range ratio, with a region of omnidirectional reflectivity highlighted in black. Lower panel: Comparison of transmission spectra for two different hollow fibers of ~ 30 cm length having similar structure but scaled photonic crystal (multilayer) period dimensions. The spectrum in gray is from a fiber with a fundamental photonic bandgap at 3.55 microns; the black spectrum is from a fiber where the corresponding bandgap is at 3.1 microns. High-order bandgaps (indicated by arrows) are periodically spaced in frequency. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO₂ laser transmission. *Nature* 420: 650–653.

ratios (fibers with a fundamental bandgap centered near 3.55 microns have an OD of 670 microns; those with a gap at 3.1 microns have an OD of 600 microns). The high-order bandgaps are periodically spaced in frequency, as expected for such a photonic crystal structure.

The wavelength scalability of our fibers was further demonstrated in the fabrication of hollow fibers designed for the transmission of 10.6 micron EM radiation. This not only shows that these structures can be made to guide light at extremely disparate wavelengths, but that specific useful bandgap wavelengths can be accurately targeted during fabrication and fiber drawing. Powerful and efficient CO₂ lasers are available that emit at 10.6 microns and are used in such applications as laser surgery and materials processing, but waveguides operating at this wavelength have remained limited in length or loss levels. Using the fabrication techniques outlined above, we produced fibers having hollow core diameters of 700–750 microns and ODs of 1300–1400 microns with a fundamental photonic bandgap spanning the 10–11 micron wavelength regime, centered near 10.6 microns. **Fig. 5** depicts a typical FTIR transmission spectrum for these fibers, measured using ~ 30 cm long straight fibers.

In order to quantify the transmission losses in these 10.6 micron bandgap hollow fibers, fiber cutback measurements were performed. This involved the comparison of transmitted intensity through ~ 4 meters of straight fiber with the intensity of transmission through the same section of fiber cut to shorter lengths (**Fig. 18** inset). This test was performed on multiple sections of fiber, and the results found to be nearly identical for the different sections tested. The measurements were performed using a 25 watt CO₂ laser (GEM-25, Coherent-DEOS) and high power detectors (Newport 818T-10). The fiber was held straight, fixed at both ends as well as at multiple points in the middle to prevent variations in the input coupling and propagation conditions during fiber cutting. The laser beam was sent through focusing lenses as well as 500 micron diameter pinhole apertures and the input end face of the fiber was coated with a metal film to prevent accidental laser damage from misalignment. The transmission losses in the fundamental bandgap at 10.6 microns were measured to be 0.95 dB/m, as shown in the inset of **Fig. 18**, with an estimated measurement uncertainty of 0.15 dB/m. These loss measurements are comparable to some of the best reported loss values for other types of waveguides operating at 10.6 microns. A bending analysis for fibers with a bandgap centered at 10.6 microns revealed bending losses below 1.5 dB for 90 degree bends with bending radii from 4–10 cm. We expect that these loss levels could be lowered even further by increasing the number of layers, through optimization of the layer thickness ratios and by creating a cylindrically symmetric multilayer fiber with no inner seam (present here because of the ‘rolling’ fabrication method). In addition, using a polymer with lower intrinsic losses should greatly improve the transmission characteristics.

One reasonable figure of merit for optical transmission losses through hollow all-dielectric photonic bandgap fibers is to compare the hollow fiber losses to the intrinsic losses of the materials used to make the fiber. As₂Se₃ has been explored as an IR-transmitting material, yet the losses at 10.6 microns reported in the literature are ~ 10 dB/m for highly purified material, and more typically are greater than 10 dB/m for commercially available materials such as those used in our fabrication. Based on FTIR transmission and spectroscopic ellipsometer measurements that we have performed on PES, the optical losses associated with propagation through solid PES should be greater than 40 000 dB/m at 10.6 microns. This demonstrates that guiding light through

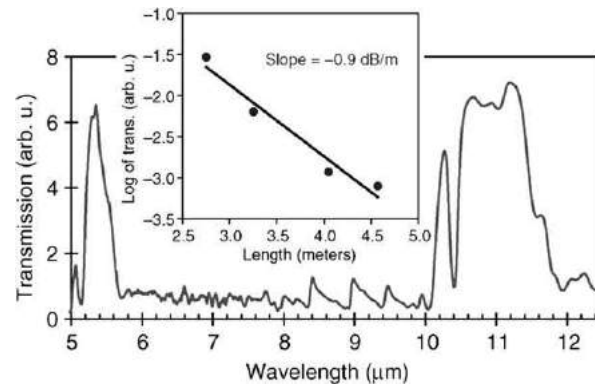


Fig. 18 Typical transmission spectrum of hollow fibers designed to transmit CO₂ laser light. The fundamental photonic bandgap is centered near a wavelength of 10.6 microns and the second-order gap is at ~5 microns. Inset: Log of transmitted power (arbitrary units) versus length of fiber (meters). The slope of this graph is the loss in dB/m. The measured fiber has a hollow core diameter of 700 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO₂ laser transmission. *Nature* 420: 650–653.

air in our hollow bandgap fibers leads to waveguide losses that are orders of magnitude lower than the intrinsic fiber material losses, which has been one of the primary goals of hollow photonic bandgap fiber research. These comparatively low losses are made possible by the very short penetration depths of EM waves in the high refractive index contrast photonic crystal structure, allowing these materials to be used at wavelengths that may have been thought improbable. Another long-standing motivation of infrared fiber research has been the transmission of high-power laser light. As a qualitative demonstration of the potential of these fibers for such applications, both straight and smoothly bent fibers of lengths varying from 0.3–2.5 meters were used to transmit enough CO₂ laser energy to burn holes through paper. The maximum laser power density coupled into our fibers in these trials was approximately 300 W/cm², more than sufficient to burn a homogeneous polymer material, including PES. No damage to the fibers was observed when the laser beam was properly coupled into the hollow fiber core. These results indicate the feasibility of using hollow multilayer photonic bandgap fibers as a low-loss wavelength-scalable transmission medium for high-power laser light.

Further Reading

- Abeles, F., 1950. Investigations on the propagation of sinusoidal electromagnetic waves in stratified media. Application to thin films. *Annales de Physique* 5, 706.
- Benoit G and Fink Y, Spectroscopic Ellipsometry Database, <http://mit-pbg.mit.edu/Pages/Ellipsometry.html>.
- Born, M., Wolf, E. (Eds.), 1980. *Principles of Optics*, 6th edn. New York: Pergamon Press, p. 67.
- Engeness, T.D., Ibanescu, M., Johnson, S.G., *et al.*, 2003. Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. *Optics Express* 11, 1175–1198, <http://www.opticsexpress.org/abstract.cfm?URI=OPEX-11-10-1175>.
- Fink, Y., Winn, J.N., Fan, S., *et al.*, 1998. A dielectric omnidirectional reflector. *Science* 282, 1679–1682.
- Fink, Y., Ripin, D.J., Fan, S., Chen, C., Joannopoulos, J.D., Thomas, E.L., 1999. Guiding optical light in air using an all-dielectric structure. *Journal of Lightwave Technology* 17, 2039–2041.
- Hart, S.D., Maskaly, G.R., Temelkuran, B., Prideaux, P.H., Joannopoulos, J.D., Fink, Y., 2002. External reflection from omnidirectional dielectric mirror fibers. *Science* 296, 510–513.
- Ibanescu, M., Johnson, S.G., Soljacic, M., *et al.*, 2003. Analysis of mode structure in hollow dielectric waveguide fibers. *Physics Review E* 67, 1–8. 046608.
- Johnson, S.G., Ibanescu, M., Skorobogatiy, M., *et al.*, 2001. Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9, 748–779, <http://www.opticsexpress.org/abstract.cfm?URI=OPEX-9-13-748>.
- Kuriki, K., Shapira, O., Hart, S.D., *et al.*, 2004. Hollow multilayer photonic bandgap fibers for NIR applications. *Optics Express* 12, 8.
- Soljacic, M., Ibanescu, M., Johnson, S.G., Joannopoulos, J.D., Fink, Y., 2003. Optical bistability in axially modulated OmniGuide fibers. *Optics Letters* 28, 516–518.
- Temelkuran, B., Hart, S.D., Benoit, G., Joannopoulos, J.D., Fink, Y., 2002. Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO₂ laser transmission. *Nature* 420, 650–653.
- Yeh, P., Yariv, A., Hong, C.-H., 1977. Electromagnetic propagation in periodic stratified media. 1. General theory. *Journal of the Optical Society of America* 67 (4), 423.
- Yeh, P., Yariv, A., Marom, E., 1976. Theory of Bragg fiber. *Journal of the Optical Society* 68, 1196–1201.

Guided Wave Optics

Alan Mickelson, University of Colorado at Boulder, Boulder, CO, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Already in 1842, Colladan had published his observation of light being guided within a curving jet of water (Colladan, 1842a,b). The demonstration led to some scientific study as well as numerous artistic and carnival displays (Colladan, 1842b; Hecht, 2004). From this 19th century curiosity, guided wave optics has grown to be the physical level technology that transfers most of the world's information. The waveguide of the worldwide telecommunications network is the optical fiber. The optical fiber itself is but a passive waveguide and guided wave optics is the technology which includes all of the passive and active components which are necessary to generate light, impress (electrical) information on light to produce information bearing optical signals, regenerate these optical signals during transmission, and convert optical signals back to (electrical) information streams at the transmission system output. In this article, some introduction to this rather encompassing topic of guided wave optics will be given.

This article will be separated into six sections. In the next section, discussion will be given to fiber optics, that is, the properties of these optical waveguides which allow light to travel in distinctly non-rectilinear paths over terrestrial distances. In Section Active Fiber Compatible Components of the present article will then turn to the components which can be used along with the fiber optical waveguides to create transmission systems. These components include light sources, modulators and detectors as well as optical amplifiers. In Section Telecommunications Technology of the article, we will discuss the telecommunications network which has arisen due to the availability of fiber optics and fiber optic compatible components. The article would hardly be complete without some discussion of the ever growing field of optical data communications that is Section Data Communications of the exposition. The article closes with a Summary section.

Fiber Optics

Fiber optics refers to a technology in which light (actually infrared, visible or ultraviolet radiation) is transmitted through the transparent core of a small (250 μm diameter – a human hair is circa 75 μm diameter) thread of composite material. The composite material consists of a core concentric with a cladding of lower optical density (index of refraction) than the core and a coating, generally a polymer, that is applied during manufacture for environmental protection (see Fig. 1). The composite material is referred to as an optical fiber. Typical dimensions for a telecommunications fiber (which most fiber is) are core diameter of 10 μm for a single mode fiber and 50 μm for a multimode fiber, cladding diameter 125 μm and coating of 250 μm .

The fiber coating here is included as a part of the fiber, unlike in most textbooks interested in mathematical analysis alone, because uncoated fiber has a short lifetime. Rarely does one see an uncoated fiber unless one has just mechanically or chemically stripped off the coating in anticipation of further processing of the bare fiber, that is, splicing or other attachment to a device. Fiber that is clear and coiled on a spool is coated fiber. When fiber is jacketed it appears much as conventional insulated wire although the resulting wire is lighter and more flexible than copper (or other metal core) wire. A fiber cable contains many jacketed fibers

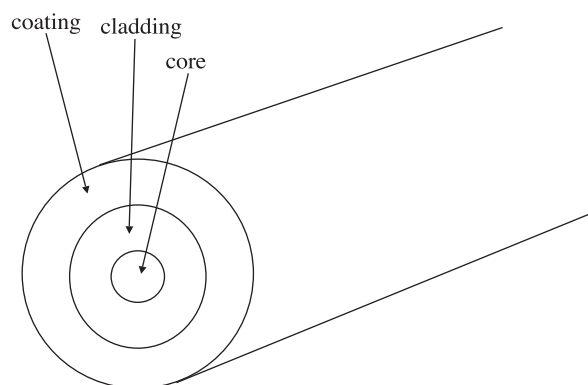


Fig. 1 Schematic depiction of the concentric layers that make up a fiber optic thread. The fiber consists of a core in which the light is guided, a cladding that confines the light to the core as well as adding mechanical support and a coating that protects the fiber from the environment. Silica fiber, the choice of material for telecommunications fiber, is especially susceptible to contamination by humidity. Typical dimension for a telecommunications fiber are a coating diameter of 250 μm , a cladding diameter of 125 μm and a core diameter of 10 μm for single mode fiber and 50 μm for multimode fiber. At the wavelength of 1.55 μm , the index of refraction of a pure silica cladding is 1.45 and of the germano-silicate core circa 1.46 for a typical numerical aperture of 0.024.

(840 is a usual maximum number for the largest of sea cables) with some support material for rigidity. Optical fiber for telecommunications is fabricated by a process of gas phase chemical deposition of fused silica doped with various trace chemicals. Because of the distances involved in the world telephone network, it is safe to say that practically all optical fiber in existence is telecommunications fiber. Fiber can be made from a number of different material systems and in a number of different configurations for use with various types of light sources in specialty applications. In what follows, we will limit discussion to the basic properties of the light guided by the fiber and leave fabrication and manufacture for the those readers who will to research on their own from the numerous sources available on these topics.

There are two complimentary mathematical descriptions of the propagation of light in an optical waveguide. Ray theory is an approximate description that is quite accurate for properties of multimode fibers. Wave theory is a more rigorous description that is necessary to elucidate the properties of single mode propagation but is needlessly complex for the description of multimode fiber where there are typically 1000 modes. We will briefly consider both descriptions of the propagating light in order to understand coupling of light to and from fiber, attenuation of light in fiber and the dispersion of light pulses when propagating along a fiber axis.

In the ray description, light is considered to be made up of a bundle of rays. In a uniform homogeneous medium, each ray is an arrow that exhibits rectilinear propagation from its source to its next interface with a dissimilar material. These rays satisfy the law of reflection (incident angle equals reflected angle) and Snell's law of refraction (bending is proportional to the optical density of the material) at interfaces between materials with dissimilar optical properties. That is, at an interface, a fraction of the light is reflected backwards at an angle equal to the incident angle and a portion of the light is transmitted in a direction which is directed more toward the normal to a plane interface when the index of refraction increases across the boundary and is directed more away from the normal when the index decreases. Using these laws at the input enface to the fiber, a portion of the energy guided by each ray is reflected back into space due to the change in refractive index at the guide surface. The refracted ray enters the fiber to exhibit a continued path.

In a step index optical fiber where the index of refraction is uniformly higher in the fiber core than in a surrounding cladding, a ray will propagate along a straight path until encountering the core cladding interface. Unguided rays (see Fig. 2) will be only partially reflected at the core cladding interface. The quantity of light remaining in the fiber will rapidly die out while propagating along the axis of the fiber due to this loss of intensity (energy) at each successive surface reflection. Rapidly is obviously a relative term. Typical distances between successive encounters with the boundary are 1 mm and total propagation distances may be many kilometers. The loss of even a tenth of 1% per reflection will result in the wave being decreased in intensity by about a factor of ten in 1 m (1000 reflections).

Because of the law of refraction, the unguided rays refracted at the interface are directed ever more closely to the direction of the core cladding boundary as the for the rays that are ever more directed along the axis. There is then an incident angle at which the direction of the refracted ray is along the interface. If the incident angle is more axially directed than this cut-off angle, there is no refracted ray in the cladding. These guided rays (see Fig. 3) are totally internally reflected back into the fiber core to again be totally internally reflected at the next core cladding interface. Although none of the guided rays suffer loss at the interface, for each

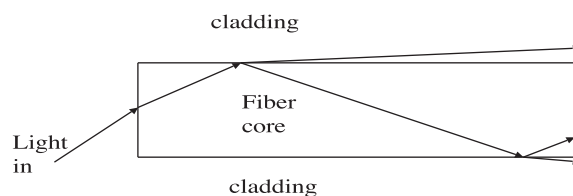


Fig. 2 A graphic indicating a ray path for an unguided (unbound) light ray. In the ray picture of light, a source is represented as an emitted bundle of rays. A consequence of Snell's law of refraction is that there is an incident angle for which the transmitted ray travels along the interface when the ray is incident from a higher to a lower optical density (index of refraction). When the ray is incident at an angle closer to the normal than this critical angle, a ray is transmitted with an amplitude given by a Fresnel coefficient. As light is shared with the cladding at each reflection point, the light ray is unbound.

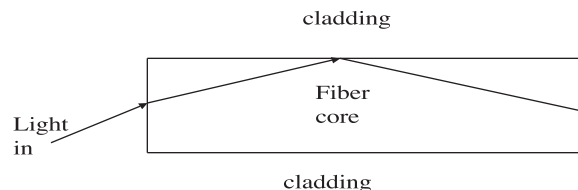


Fig. 3 A graphic indicating a ray path for a guided (bound) light ray. In the ray picture of light, a source is represented as an emitted bundle of rays. A consequence of Snell's law of refraction is that there is an incident angle for which the transmitted ray travels along the interface when the ray is incident from a higher to a lower optical density (index of refraction). For a ray incident more nearly tangential to the interface than the critical angle, there is no transmitted ray. This is to say that rays incident in this range of angles are bound to the core of the fiber.

incident angle, the guided ray will travel a different distance along its path in traversing a given distance along the fiber axis. This effect is referred to as modal dispersion. A bit of information is a single pulse of light. A single pulse of light if launched into a number of modes will then spread as it propagates down a step index, multimode fiber. As the quantity of information transmitted is proportional to the number of pulses transmitted per unit time, dispersion limits the quantity of information that can be transmitted over a given distance along the fiber axis.

In a graded index fiber, the refractive index within the fiber core varies continuously from a maximum somewhere within the core to a minimum at the core cladding interface. The guided ray paths within such fibers are curved whereas the unguided simply exit the fiber when they reach the cladding. That is, the guided rays in graded index medium are characterized by the fact that once in the fiber they never encounter the cladding. Unlike in step index fiber, there is no clear relation between the angle at which ray enters the fiber and the distance that a guided ray actually traverses along its path in order to traverse a distance along the fiber axis. The dispersion of a graded index multimode fiber is dependent on the grading of the index. Modal dispersion is generally expressed as the number of cycles of modulation that can be transmitted over a given length with only a 50% decrease in cycle amplitude. The unit of a cycle is a Hz (hertz). A 1 Hz bandwidth is roughly equivalent an information rate of 1 bps (bit per second). As bps is a small unit, thousands (Kbps), millions (Mbps), billions (Gbps) and trillions (Tbps) are generally used. The bandwidth of a step index telecommunications fiber is roughly 100 MHz km and a graded index fiber 1 GHz km.

Guided rays suffer no excess loss per propagation cycle, that is, a pair of reflections in a step index or complete curvilinear path in a graded index fiber. That is not to say that the propagation is loss free. There is some residual absorption and scattering of the light that are dependent on the wavelength of the light. Telecommunications wavelengths are clustered into three wavelength windows, one about 0.85 μm , the second at 1.3 μm and the third at 1.55 μm . These intrinsic losses are result in the wave being attenuated by a factor of 10 in about 3 km at 0.85 μm , 20 km at 1.3 μm and 50 km at 1.55 μm . These losses are much lower than the leakage losses associated non guidance discussed above.

The ray description of fiber propagation is quite simple. Rays incident on the input surface at less than a critical angle (defined by the fiber numerical aperture (NA) of the fiber) are guided. Other rays do not couple. At the fiber output, the converse is true. Rays exit from the enface at all angles up to the cut-off angle defined by the NA (which happens to be the sine of the cut-off angle) of the fiber. What the ray picture does not take into account is the wave nature of light. In the wave picture, each ray is carrying a clock that remembers how long it has been propagating along its ray path. When two rays come together, they can either add or subtract depending on the reading on their respective clocks and the wavelength of the source. This is the process known as interference. The ray picture gives a poor description of the lowest order mode on an optical fiber, the one that follows the axis. When the fiber has a small enough core and NA (small enough acceptance angle), only this fundamental mode can propagate. This defines a single mode fiber. As there is only one ray path, all light coupled into the fiber propagates at the same velocity. Hence, we say that there is no modal dispersion. There is still dispersion as difference wavelengths will travel at different velocities.

Multimode fibers have larger cores and higher numerical apertures than single mode fibers. They are easier to couple to. As their description (the ray description) did not require interference to be taken into account, coupling does not require the source to radiate at only one wavelength, that is, to be monochromatic. Monochromatic, single wavelength, is the lower end limit of what we call the spectral (band) width of a source. One can couple light from even a white light source (broad spectral band) into a sufficiently multimode fiber. As each more, though, propagates at a different velocity, the ease of coupling and alignment comes at the cost of information rate. The information content of signal is inversely proportional to the temporal width of the pulses (how many pulses can be jammed into a given time interval) in the signal. The bandwidth of the signal we can send over a multimode fiber is limited by the number of modes. In contrast, a signal mode fiber requires small dimensions and tight tolerances to couple. Single mode coupling is also phase dependent so a narrow band source such as a single mode laser is necessary in order to couple light efficiently. Pulses, though, disperse much more slowly with propagation distance in the single mode fiber than in the multimode fiber. Much more information can be transmitted over the same distance although at a cost of higher alignment tolerances and use of more coherent sources. As stated above, a graded index multimode fiber can support an information bandwidth of roughly 1 GHz-km, that is, one can transmit 1 GHz 1 km, half a GHz over 2 km, etc. The bandwidth of a single mode fiber is defined by the source wavelength and spectral width as well as the length of propagation. The dispersion can also be managed by alternating the type of fiber in which the signal propagates along the propagation path. At present (2017), single mode fibers can transmit bandwidths of 25 GHz over 40 km spans between repeaters. Further improvement is possible.

Active Fiber Compatible Components

The previous discussion of propagation in fiber optic waveguides would be rather sterile were there not a number of available fiber compatible components that can generate and detect light streams. In order to construct fiber optic systems, components to impress information streams on light and to read out the impressed information at a receiver are also necessary. In this section, we will discuss some of the theory of operation of such active optical components with a goal of understanding possibilities and limitations of guided wave technology.

In a passive component such as an optical fiber, the optical signal is considered as given. The power carried by the signal can only be attenuated. The attenuation may be through imperfect coupling or loss mechanisms. Signal fidelity can be altered through dispersion but the impressed information is considered as given. New information cannot be created. Active components are different. Active components require that the optical fields interact with a medium such that energy and/or information can be

transferred from the active medium to the light field or from the light field to the active medium. Light can be amplified. In an optical amplifier, energy is transferred from an atomic medium to a propagating wave. Light can be generated where there was no light. That is, amplification (more correctly said, oscillation) can take place even without an input signal. Information can be impressed on a light stream. That is, in a modulator, a signal is applied to the atomic medium while a constant optical traverses it. The result is the impression of an electrical information stream on an optical carrier. This operation of mixing signals is inherently nonlinear. There is no direct inverse operation. However, a light stream can be converted back to an electrical stream. A detector can work as a source but in reverse. Fortunately, optical detectors with short response times can be used to follow the variations of a modulated optical signal and thereby recover a modulating electrical information stream while converting the optical to electrical power.

Sources compatible with fiber optic systems are quite generally non-thermal sources. Fiber compatible sources operate on quantum mechanical principles. The electrons in an atom can only exist in certain energy states, that is, in certain atomic configurations. An isolated atom is then characterized by a set of energy levels, that is, energies that are absorbed or emitted when the atom changes configuration. Energy must be applied to reconfigure the atom to a higher energy state. Energy is released when an atom relaxes to a lower energy state. The energy can take the form of a single particle of light of a particular wavelength (frequency). Such a particle of light is called a photon. It is usual to consider a single pair of levels of an atom. A pair of levels is referred to as a transition. For example, the rare earth ions of a rare earth doped optical amplifier act independently of one another. The medium acts as a sum of atomic media. A single transition of the rare earth ion interacts with a stream of light, that is, of photons of light with a given wavelength. By preparing the atoms such that they are in the upper state of their transition, one can extract energy from the medium to the light.

In a semiconductor, atoms act collectively. That is, inner electrons are bound to specific nuclei whereas the outer electrons are shared between all of the nuclei. The shared electrons can either all be weakly bound to the nuclei, or free to roam in such a manner that charge neutrality is preserved. The level spacing of interest in the semiconductor is that of the band gap. The band gap is the spacing between the least energetic freely propagating (conduction) electron and the most energetic bound (valence) electron. In the semiconductor, then, the transition of interest is the valence to conduction band transition. The wavelength of operation of such an active medium is determined by the energy spacing between the upper and lower levels of this quantum mechanical transition. As illustrated in Fig. 4, injecting current into a semiconductor diode introduces electrons into the conduction band from the negative electrode. Valence band electrons in the junction that recombine with these conduction band electrons to give off photons at the wavelength of the transition. A source, be it light emitting diode (LED) or laser diode (see Fig. 4) will emit light at a wavelength which corresponds to an energy slightly above the minimum gap energy. Detectors are in essence the inverse of sources (see Fig. 5). However, there are important differences.

A detector can detect almost any energy greater than the band edge. That is, a laser or LED will emit light in a band of wavelengths around a center that corresponds to an energy value (photon energy is equal to a constant (that happens to be Planck's constant times the speed of light) divided by the wavelength of the photon) slightly greater than the band gap. A detector is not limited on the high energy end. If the detector sees red light, it will also see blue light. A more important difference, though, has to do with the direction of the transition. In a detector, the final electron state is a conduction electron. Conduction electrons are meant to be swept out of a junction as current. In a source, the final state is a photon that is due to recombination of a conduction and valence electron. It is not so easy to radiatively combine electrons and holes as the valence electrons are called. To make a long story short (the concerned reader will be able to find numerous sources on semiconductor lasers), only direct band gap materials can be used for sources. Silicon and germanium, the materials of most electronics, are indirect gap. GaAs and InGaAsP are direct gap and are the materials of choice for semiconductor lasers.

The historical development of the fiber system sources serves to explain what was previously only stated. Telecommunications employs essentially three windows of wavelengths, one centered about $0.85\ \mu\text{m}$, one about $1.3\ \mu\text{m}$ and a third about $1.55\ \mu\text{m}$. Why the first window is referred to as first can be explained by the development by the ternary source material GaAlAs before the

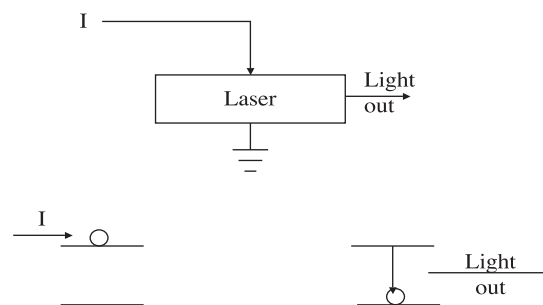


Fig. 4 Schematic depiction of the quantum mechanical action of an electrically pumped, non-thermal light source, that is, a light emitting diode (LED) or laser. Current is injected into a semiconductor diode structure that results in injected electrons entering the conduction band of the light emitting region. The conduction band is an excited band at an energy level higher than that of the ground state valence band. To relax to the valence band, the electron must give up energy in the form of a photon, a particle of light whose frequency corresponds to the transition energy that is related to the transition frequency by energy equals Planck's constant times the frequency.

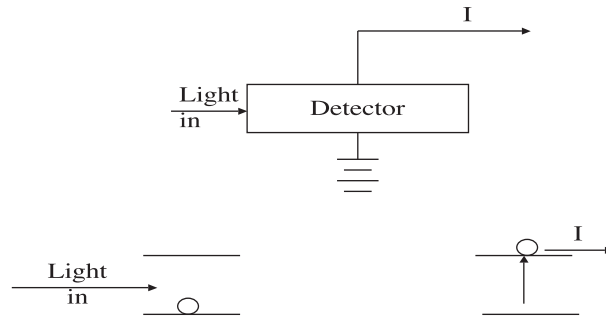


Fig. 5 Here, the detection process in a semiconductor medium is illustrated. The process is inverse to the emission process in some respects but not others. A photon is incident on the detecting region of a semiconductor. If the photon energy is higher in energy than the band gap then there is a probability that it will be absorbed by a valence band electron that will be elevated in energy to the conduction band. Electrical bias of the active region of the detector will then sweep the conduction band carrier out of the junction area. Whereas in a source, recombination requires that both an electron and hole are present before emission, any valence band electron can be promoted to the almost empty conduction band and then swept out of the junction by a bias field.

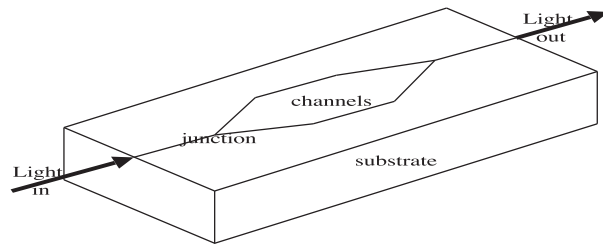


Fig. 6 Schematic depiction of the arrangement of waveguide regions in a Mach-Zehnder modulator. Light that enters the modulator through the single mode waveguide to the left is split, and then propagated through straight parallel channels before being recombined. If the light in the two arms is in phase on recombination, the output power will equal the input power. If the light is out of phase, the light will radiate from the down coupling region. If the channels are electro optic and a time-varying signal is applied between them, the relative phase induced by the signal will be transduced into a time varying intensity at the output junction.

quaternary materials of the InGaAsP family. Pure GaAs has a gap that corresponds to $0.85\ \mu\text{m}$ whereas increasing the Al fraction shifts the operation to shorter wavelengths. Doped fused silica is the material used for telecommunications grade optical fiber. The loss of fused silica fibers decreases with increasing wavelength until $1.6\ \mu\text{m}$ with a single exception of a broad absorption peak around $1.1\ \mu\text{m}$. The loss abruptly increases orders of magnitude above $1.6\ \mu\text{m}$. The window around $0.85\ \mu\text{m}$ is where the wavelength of GaAlAs lasers overlap a minimum loss window of fused silica. InGaAsP is a family of materials whose resonant wavelength can be tuned to lie from circa $1.0\ \mu\text{m}$ to $1.6\ \mu\text{m}$. The dispersion minimum of fused silica lies about $1.3\ \mu\text{m}$ and the minimum loss of fused silica lies at $1.55\ \mu\text{m}$.

The last piece of the active device puzzle concerns modulation. Semiconductor sources that operate by current injection offer the option of direct modulation, that is, one can add an information stream in the form of an input current stream. A problem with such modulation is that it disturbs the spectral characteristics of the source. Much has been done to minimize this distortion. There are a myriad of different semiconductor sources with a myriad of acronyms to describe that were mainly introduced to minimize spectral distortion under modulation. An important one is DFB or distributed feedback laser. In the most exacting of applications, one uses an external modulator such as the Mach-Zehnder structure (MZ) seen in Fig. 6. No attempt has been made here to specify the technology or even the electrode structure. The material systems and drive systems are too numerous to describe in this article. The operating principle, though, is simple. It is also the principle shared by most free space interferometers. In an external modulator system, a stabilized laser output is input into the left hand input of the MZ. The light is split as equally as possible into two streams. In the channel region, these streams propagate in parallel but at a distance to each other such that they do not interact. If nothing happens in this region and the path lengths are equal, the light will recombine. If an information source is used to affect the relative phases of the waves in the two arms, though, the light will not recombine. The phase modulation will be converted to intensity modulation.

How to impress phase modulation of the arms of the MZ? The highest speed method is to use the electro optic effect in an electro optic medium. By placing electrodes over the two channels and applying the information stream between these channels, one can alter the index of refraction and hence optical path in one arm relative to the others. The electro optic effect in a number of materials (lithium niobate, silicon, GaAs, InGaAsP, etc.) is electronic in origin and, therefore, can respond to an electrical signal at rates as high as optical frequencies. In fact, light can be used to modulate light. More commonly, the modulation frequency, though, is limited by the frequencies that can be electrically generated and applied to electrodes. We leave this topic to the

concerned reader and leave the optic to say that, for example, silicon photonics MZ modulators that modulate to 50 Gbps are presently available at a number of foundries.

Telecommunications Technology

The physical transmission layer of the world wide telecommunications network has come to be dominated by fiber optic technology. The long haul network that transmits the majority of the world's bits is comprised of wavelength division multiplexed (WDM) single mode fiber optic links operating in the 1.55 μm telecommunications window. In the following, we will first discuss how this transformation to optical communications took place and then go on to discuss some of the specifics of the technology. The highest capacity channels are present (2017) transmit 100 Gbps per channel by using four level phase modulation at 25 Gbps on each of the polarization states of the fiber at one of the International Telecommunications Union (ITU) defined wavelengths within the so-called C- band of optical communications spectrum. The C-band spans 35 nm from 1530 to 1565 nm in the infrared. At 1550 nm, a 1 nm of wavelength bandwidth is equivalent to 125 GHz of frequency bandwidth (using that wavelength equals the speed of light divided by frequency). It is, therefore, possible to place 40 of these 100 Gbps channels (each one occupying circa 20 GHz on 50 GHz spacing) in the C-band in order to transmit a composite rate of 8 Tbps per optical fiber. Bandwidths transmitted by transoceanic cables are now approaching composite bandwidths of 1 Pbps (10^{15} bits/s).

Already by the middle of the 1960s, it was clear that changes were going to have to take place in order that the exponential growth of the telephone system in the United States as well as in Europe could continue. The telephone system then, pretty much as now, consisted of a hierarchy of tree structures connected by progressively longer lines. A local office is used to connect a number of lines emanating in a tree structure to local users. The local offices are connected by trunk lines. The lines from there on up the hierarchy are long distance ones which are termed long lines and can be regional and longer. The most pressing problem in the late 1960s was congestion in the so-called trunk lines which was most severe in urban areas. In many cities, these trunk lines were two kilometers in length, a length that was fixed by the underpayment ducts in which the cables were placed. The problem was that there was no more space in the ducts that housed these lines in any number of cities including New York City. The solution to duct congestion had traditionally been to use progressively more time division multiplexing (TDM) to increase the traffic that could be carried by each of the lines already buried in the conduit.

The human voice produces sounds in the frequency range of 20 Hz to as high as 20 kHz. The upper range (above circa 10 kHz), though, is necessary only for opera, dog whistles and a some other more esoteric purposes. Usual speech is still understandable if frequencies only as high as 4 kHz are included. By digitally sampling a 4 kHz band limited conversation twice per period and quantizing each amplitude into 8 bits (where $2^8 = 64$ levels), a digital copy of that conversation can be stored as 64 Kbps. If one then TDM's, 24 of these conversations, one obtains a composite rate of circa 1.53 Mbps, a standard known as T1. T1 TDM was already becoming common in the 1960s with the dawning of the digital communications era. If 28 of these T1 channels are then TDM'ed together, one obtains a T3 channel of roughly 45 Mbps. By 1975, the plan for the next generation of trunk lines was to use T3 (also called DS1). The digital equipment was ready but the twisted pair lines that were already installed in the ducts were problematic. The twisted pair lines with a bandwidth of circa 10 MHz km would not even carry the T2 rate of 4 T1 (6 MHz) over two kilometers. That is, dispersion of 10 MHz km over two kilometers (bandwidth of 5 MHz) would smear out the edges of the time varying bit streams reducing the modulation power transmitted. In 1975, the only viable lasers were GaAlAs operating in the first (0.85 μm window) with silicon detectors. The best step index, multimode fiber (evidently for use at 0.85 μm) in 1975 was rated at 100 MHz km and would transmit the edges of a 45 Mbps (T3) signal. The problem was thought to lie in the sources and detectors, but not so.

A 1975 demonstration in Atlanta (Hecht, 2004) of a multimode fiber optic system operating in the 0.85 μm window was successful. Subsequent progress was meteoric. Already by 1980, advances in single mode laser and single mode fiber technology operating at the 1.3 μm wavelength had made the inclusion of fiber into the long lines viable as well. For roughly the decade from 1985 onward, single mode fiber systems dominated long line replacement for terrestrial as well as transoceanic links. The erbium doped fiber amplifier which operated in the 1.55 μm wavelength third telecommunication window had proven itself to be viable for extending repeater periods already by around 1990. Development of the InGaAsP quaternary semiconductor system allowed for reliable lasers to be manufactured for this window as development of strained lattice laser technology in the InGaAs system allowed for efficient pump lasers for the fiber amplifiers. As the optical amplifiers could amplify across the wavelength band that could be occupied by many aggregated channels of TDM signals, the move of the long line systems to the third telecommunications window was accompanied by the adoption of wavelength division multiplexing or WDM.

WDM, unlike TDM, can be carried out using only optics. That is, dispersive optical elements (e.g., diffraction gratings) can split channels into channels of spectral widths as small as 0.2 nm (25 GHz). The International Telecommunications Union defined a standardized wavelength grid down to 25 GHz spacing in the earl 1990s. Soon after, 10 Gbps channels on 50 GHz (0.4 nm) spacing were allowing single fibers to carry up to 32 channels. The 10 GHz proved to be a challenge that electronics was not able to surmount until the 2010s. At present, 25 Gbps modulation rates and ethernet cards are becoming ubiquitous along with the 100 Gbps rates per channel with which the discussion of this section began. But internet demand continues to rise at a rate of 22% per year so the story of the telecommunications network is nowhere near complete.

Data Communications

As is evident from the exposition of the last section on guided wave optics in telecommunications, fiber optics dominates all applications that require a bandwidth length product greater than some limit. The limit is somewhat application dependent (due to cost) but lies in the range of a 100 MHz km for telecommunications. Although fiber to the home (FTTH) has been discussed for many years and even deployed in a number of trials in a number of countries, digital telephony with a requirement of 64 kbps per conversation does not lead to rates requiring a fiber solution for single homes. The rates required for data communication, however, continue to increase. Clock rates of computer chips have stagnated near 4 GHz since circa 2007, but stream serialization has resulted most recently in 25 Gbps ethernet cards for individual servers within data networks.

Fig. 7 illustrates the structure of a present day warehouse scale computer (WSC). A WSC is a type of data center, that is, an arrangement of processors and memory that can be used to both store data as well as carry out operations on that day bank. WSC's are the elements of what is commonly called the cloud. Various companies use worldwide networks of WSC's to ply their trade. Google uses WSC's for inquiry response, Microsoft rents WSC capacity as a source of income, Facebook uses WSC's to store and retrieve member data whereas Amazon uses WSC's as both a cloud provider and as a sales house.

At the lowest level in the diagram of **Fig. 7** are the individual servers, each server consisting of processing and storage. The servers, each of which possessing an ethernet card output, are arranged into 6 foot high racks of perhaps 50 servers per rack. The racks are arranged into clusters and the clusters are arranged to fill a warehouse. The racks are limited to contain roughly 1000 cores (20 cores per server is common). The limit is economic. Cores dissipate circa 10 W and cooling a rack that dissipates more than 10 kW is expensive. The WSC is generally limited to roughly 10,000 racks or 100 MW of dissipation. The limit here the requirement of 250 MW of power (100 MW of dissipation requires 100 MW of cooling and 50 MW of input servers) as power stations larger than 250 MW become exponentially more expensive.

The racks of a data center require interconnection lengths of circa 10 m. A cluster may require lengths up to 100 m. Gigabit ethernet became available already in circa 2000. 1 Gbps ethernet cards already were requiring 100 MHz km bandwidth length products and, indeed, it was at this time that the rack top to rack top interconnections of WSC's first used went optical. As multimode grade index fiber have achieved bandwidths of 1 GHz km, the original data center links consisted of multimode graded index fiber with vertical cavity surface emitting laser (VCSEL) transmitters in the 0.85 μm windows and silicon detectors. In the time since 2000, ethernet cards for servers first increased in maximum rate to 10 Gbs and most recently to 25 Gbps. The 10 Gbps rates are still primarily multimode VCSEL driven links. With the 25 Gbps, links are being to migrate to single mode fiber (SMF) in the second and third windows with InGaAsP sources and detectors. The move to SMF is also seeding efforts to increase rates to 100 Gbps though WDM.

At present, about 70% of all internet traffic passes through a data center. This percentage is increasing as more businesses are moving to the cloud to eliminate the need to maintain enterprise data centers. The WSC's as of 2015 consume roughly 1.8% of all US electrical power (Shehabi *et al.*, 2016) and internet traffic is increasing at a rate of circa 22% per year. The energy use is so pervasive that new efforts to increase throughput are now inevitably coupled with efforts to improve energy efficiency. Such efforts are bound to include increasing levels of TDM and WDM that in turn will make optics viable at shorter interconnection length scales. The optical revolution continues.

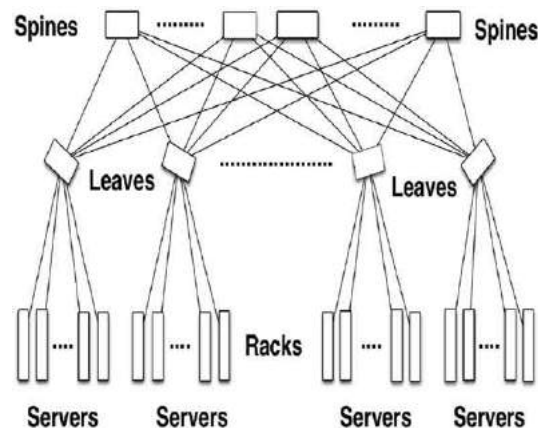


Fig. 7 A sketch that illustrates the interconnection structure of a warehouse scale computer (WSC). In the WSC, racks of servers are interconnected with each other through rack top switches. Each server has an ethernet card that is connected with a copper cable to the rack top switch. The rack top switches are then connected to higher level switches with optical fibers that have transceivers on either end of the fiber connection. The transceivers contain optical sources, modulators that are driven by the information stream from the computer to generate the optical signal that is to be transmitted to the distant location.

Summary

Although optical guidance was both understood and applied in the 19th century, the development of the low loss (less than 100 dB/km) optical fiber (Kapron *et al.*, 1970) and the room temperature semiconductor laser in 1970 (Alferov *et al.*, 1970; Hayashi and Panish, 1970) nucleated a new field of transmission technology, that of fiber optics. The first system demonstration of the technology already occurred in 1975 (Hecht, 2004). Optical fibers come in two primary varieties, multimode and single mode. Multimode fiber is easier to attach to other components than single mode but suffers from much higher dispersion. That is, the information bandwidth that a multimode fiber can propagate over any length is limited and can not easily be extended. Single mode fiber (SMF) not only has much larger bandwidth length product but exhibits dispersion of a type that can be compensated. Hence, SMF is the choice for telecommunications. And telecommunications is the field that employs more fiber than any there field because of the distances involved.

Fiver compatible components include light emitting diodes, semiconductor lasers, optical amplifiers, optical detectors and external modulators, primarily Mach-Zhender (MZ) modulators. Again, to the present, telecommunications has been the most high volume of optical component application areas. Telecommunication components operate in three windows, one around 0.85 μm , 1.3 μm and 1.55 μm (Agrawal, 2010). In the first window, the lasers are GaAs, detectors are silicon and most of the fiber is multimode. In the second and third windows, the laser and detectors are primarily InGaAsP and much of the fiber is single mode. Optical amplifiers operate in the third window and are generally rare earth doped and optically pumped. Data rates in this third channel can now (2017) be as high as 100 Gbps per channel with 32 channels of WDM in the C-band. These numbers are likely to continue to increase.

Data communications, especially within warehouse scale computers (WSC's), is switching to optics. The original data link employed data rates that could be supported over the required lengths by multimode fiber in the first window. The required rates for rack to rack interconnections have already increased to the point where second and third window components with WDM and SMF are necessary. It is expected that data rates will continue to increase and that optics will therefore become viable for shorter and shorter length links.

See also: Fabrication of Optical Fiber

References

- Agrawal, G.P., 2010. Fiber Optic Communication Systems, fourth ed. Wiley.
- Alferov, Z.I., Andreev, M.V., Portnoi, E.L., Trukan, M.K., 1970. Alas-gaas heterojunction injection lasers with a low room-temperature threshold. *Soviet Physics Semiconductors* 3, 1107–1110.
- Colladan, J.-D., 1842a. Sur les reflexions d'un rayon de lumiere a l'interieur d'une veine liquide parabolique. *Comptes Rendus* 15, 800–882.
- Colladan, J.-D., 1842b. La fontaine colladon. *La Nature* 15, 525–526.
- Hayashi, I., Panish, M.B., 1970. Gaas-gaalas heterostructure injection lasers which exhibit low thresholds at room temperature. *Journal of Applied Physics* 41, 150–163.
- Hecht, J., 2004. City of Light: The Story of Fiber Optics, Expanded, revised ed. Oxford University Press.
- Kapron, F.P., Keck, D.B., Maurer, R.D., 1970. Radiation losses in glass optical waveguides. *Applied Physics Letters* 17, 423–425.
- Shehabi, A., Smith, S., Dale, S., *et al.*, 2016. United states data center energy usage report. Technical Report LBNL-1005775, Ernest Orlando Lawrence Berkeley National Laboratory, June.

Optical Fiber Gratings

Paul S Westbrook and Tristan Kremp, OFS Fitel, LLC, Somerset, NJ, United States

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

ASE Amplified spontaneous emission
CMT Coupled mode theory
CPA Chirped pulse amplification
CW Continuous wave
DFB Distributed feedback
FBG Fiber Bragg grating
HR High reflector

IR Infrared
LPG Long period grating
LP Linearly polarized
MPI Multipath interference
OC Output coupler
TAP Turn around point
UV Ultraviolet
WDM Wavelength division multiplexing

Fiber gratings are periodic index variations inscribed into the core of an optical fiber and are important devices for manipulating fiber guided light. Since their first fabrication in the late 1970s, fiber grating use and manufacture has increased significantly and they are now employed commercially as in-fiber optical filters and reflectors in telecommunications systems, fiber lasers, and sensors. We review optical fibers and the fundamentals of fiber grating characteristics as well as fabrication. Finally we consider several of the major industrial and scientific applications of fiber gratings. Fiber gratings have been reviewed in two monographs (Kashyap, 2010; Othonos and Kalli, 1999) and various review articles (Hill *et al.*, 1997).

Fiber Modes

A fiber grating couples light between the modes of an optical fiber, and therefore a knowledge of these modes is essential to understanding the characteristics of fiber gratings (Kogelnik, 1990; Marcuse, 1974; Adams, 1981). In what follows, we emphasize single mode fibers used near 1550 nm because these are very important in most practical applications that use fiber gratings. A single mode fiber guides light in a raised index core region, typically germanium doped silica, surrounded by a silica cladding layer and protective polymer coating. The core diameter of a few microns is sufficiently small that only a single transverse mode is guided through total internal reflection at the core-cladding boundary. If the index is only slightly higher in the core than in the cladding, then the electric field (E-field) of the fundamental mode propagating in the core may be described with a characteristic transverse E-field profile and a propagation constant k :

$$E = E(\rho, \theta)e^{-i(\omega t - kz)}, \quad k = 2\pi n_{\text{eff}}/\lambda \quad (1)$$

Here, ρ and θ are the radial and azimuthal positions, z is the longitudinal position along the fiber, ω is the radial frequency, λ is the vacuum wavelength, and we assume that the E-field is linearly polarized (LP), i.e., it points in the same direction everywhere. As the second equation indicates, the fiber mode propagates in the fiber as though it has an effective refractive index n_{eff} .

The silica cladding acts to isolate the core mode from the outside, hence the core mode penetrates very little into the cladding. However, if the outer cladding surface allows total internal reflection, the cladding may guide light as well. Such “cladding modes” are also important in understanding fiber gratings, since the grating couples the core mode with the cladding modes. As shown in Fig. 1, the cladding modes extend outside the core into the cladding region, and their effective index is therefore always lower than that of the core mode. We also note that in the case when the outer coating of the fiber does not support total internal reflection, “leaky” cladding modes result. In the limit that there is no reflection at the cladding boundary, a continuum of “radiation modes” results. A grating couples power from the core mode to the cladding/radiation modes, which exist regardless of the presence of the grating. The grating is typically only a small quasi-periodic perturbation on the index profile of the waveguide and the corresponding mode fields.

Fiber Grating Theory

Fiber gratings are longitudinally periodic variations in the refractive index (or, more generally, the electric permittivity) of the core and/or cladding of an optical fiber. The scattering from any grating may be understood simply by considering the scattering off of each successive period and adding these contributions while taking the phase of the E-field into account. Fig. 2 illustrates this analysis in the case of a fiber Bragg reflector. When the reflections from each period of the grating add constructively, a strong back reflection results. This is known as Bragg reflection. The condition for Bragg reflection is given by:

$$\lambda_{\text{Bragg}} = 2n_{\text{eff}}\Lambda_{\text{grating}} \quad (2)$$

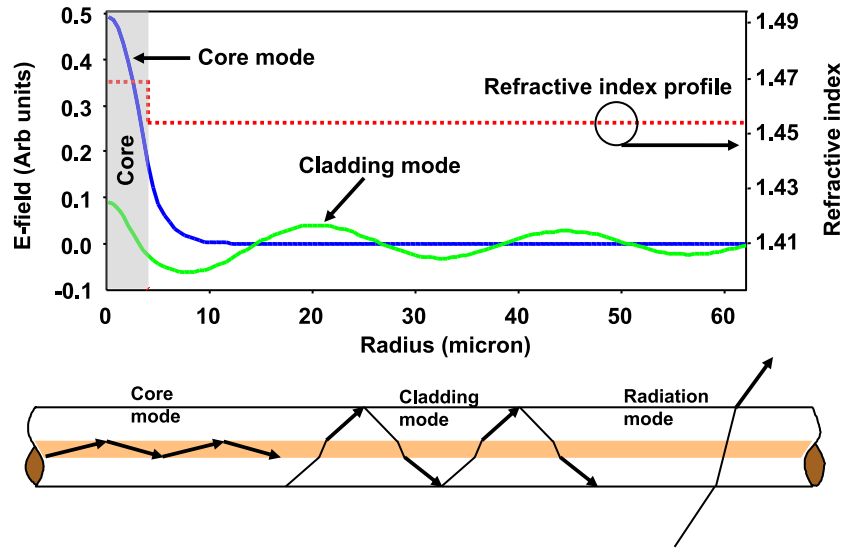


Fig. 1 Fiber modes. Refractive index profile of a step index, single mode fiber (dashed line). Radial profiles of mode E-fields of the fundamental (LP₀₁) core mode, guided within the raised index region defined by the core, and one of the many higher order cladding modes guided by the outer, air-silica interface, assuming an uncoated fiber. Also shown is a fiber indicating the ray paths for core, cladding and radiation modes.

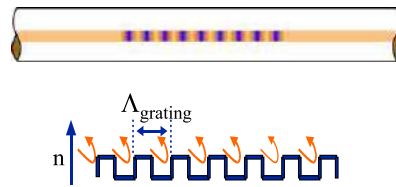


Fig. 2 Bragg reflection. Each period of the grating reflects a small fraction of the incident light. When the period of the grating is adjusted so that these reflections add coherently, a strong Bragg reflection results and a narrow dip is observed in the transmission spectrum of the fiber (see resonance near 1557 nm in Fig. 3(a)). Bragg reflection can also occur into cladding modes as shown in the resonances below 1548 nm in Fig. 3(a).

Here, λ_{Bragg} is the vacuum wavelength of the light (corresponding to a radial frequency $\omega_{\text{Bragg}} = 2\pi c_0 / \lambda_{\text{Bragg}}$), c_0 is the speed of light in vacuum, n_{eff} is the effective index of the mode, and Λ_{grating} is the period of the grating refractive index modulation. This Bragg reflection is evident in a narrow dip in the transmission spectrum of the grating, see Fig. 3(a). The ability of a fiber grating to selectively reflect the core light at a particular wavelength is one of its most important practical properties since it can be used as a narrow-band filter or reflector. As a fiber component, the fiber grating has the advantage of being compact and providing in-line wavelength filtering.

The condition for Bragg reflection may be expressed more generally as a phase matching condition. Phase matching refers to the condition when the phases of all the grating-scattered waves are the same and there is constructive interference. More specifically, this condition expresses the relationship between the spatial frequencies of the incident light, the scattered light, and the grating modulation, necessary for constructive interference in the scattered light. For a grating, the phase matching condition may be expressed in terms of the grating wave vector, $K_{\text{grating}} = 2\pi / \Lambda_{\text{grating}}$ and of the two mode propagation constants, k_{incident} and $k_{\text{scattered}}$ (defined in Eq. (1))

$$k_{\text{incident}} - K_{\text{grating}} = k_{\text{scattered}} \quad (3)$$

Using phase matching, we can understand the various transmission spectra that result from fiber gratings. Fig. 3 shows the transmission characteristic observed for core guided light in a single mode fiber containing three different types of gratings: Bragg reflector, radiation mode coupler, and co-propagating mode coupler. Each grating has a different type of phase matching condition. Fig. 3 also depicts the phase matching condition in Eq. (3) graphically. The vertical line contains a mark representing the propagation constants, k , of modes of a single mode fiber (both positive and negative for the two propagation directions) and a solid black region representing the radiation mode continuum. Bragg reflection into backward propagating core and cladding modes occurs for short grating periods ($K_{\text{grating}} \sim 2k_{\text{core}}$). For a Bragg resonance near 1.5 μm , this period is roughly 500 nm. Such gratings are known as Fiber Bragg Gratings (FBGs). A typical transmission spectrum with a strong Bragg reflection and several cladding mode resonances is shown in Fig. 3(a). Since the effective indices of cladding modes are lower than the effective index of the core mode, the Bragg wavelength $\lambda_{\text{Bragg}} = (n_{\text{eff,incident}} + n_{\text{eff,scattered}}) \Lambda_{\text{grating}}$, which we obtain from Eq. (3) as a generalization of Eq. (2), is shorter for core-cladding mode reflection in comparison to core-core Bragg reflection for any given grating period

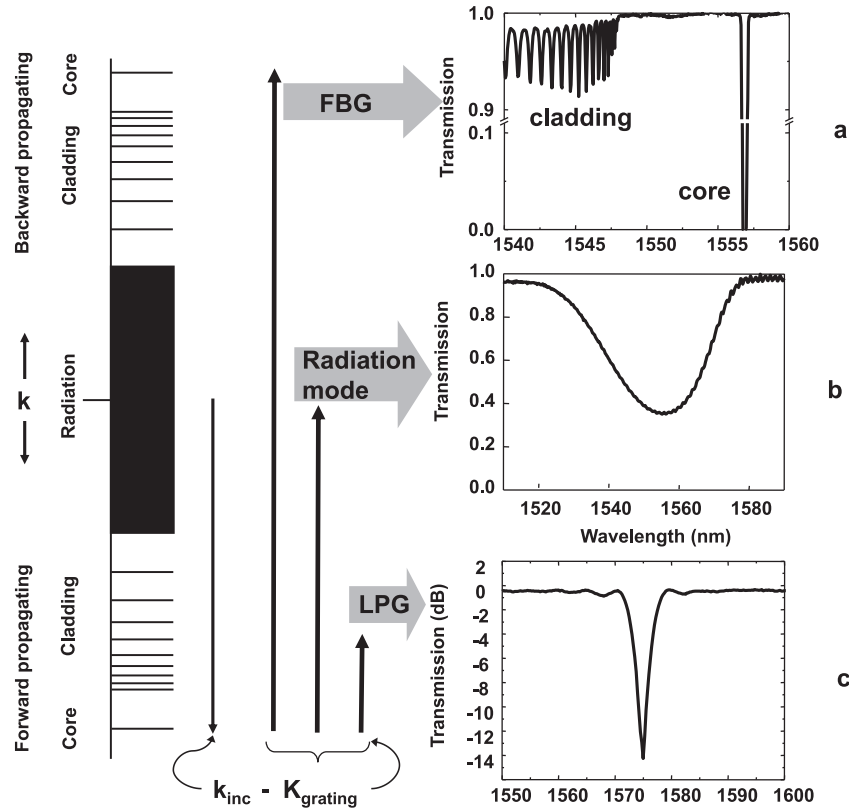


Fig. 3 Phase matching. Fiber grating resonances occur when the phase matching condition is satisfied: $k_{\text{inc}} - K_{\text{grating}} = k_{\text{scat}}$. Propagation constants ($k = 2\pi n_{\text{eff}}/\lambda$) of the fiber modes are indicated as horizontal lines on the vertical line. The black region indicates the regime of unguided radiation modes. The fiber grating is represented by its longitudinal spatial frequency (wave vector) $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}}$, where Λ_{grating} is the period of the grating. Assuming an incident forward going core mode, the grating wave vector then phase matches to (a) counter propagating core and cladding modes (Fiber Bragg Grating, FBG), (b) continuum of radiation modes, or (c) a co-propagating cladding mode (Long Period Grating, LPG).

Λ_{grating} . When the grating period is very long ($\Lambda_{\text{grating}} \sim 1000$ times that of an FBG, typically hundreds of microns), phase matching to co-propagating cladding modes occurs. Such gratings are known as Long Period Gratings (LPGs). A typical LPG spectrum is shown in Fig. 3(c). The LPG resonance corresponds to coupling between the core mode and a single co-propagating cladding mode. For wave vectors in between these two regimes, coupling to the continuum of radiation modes is possible. An example of such a grating spectrum is shown in Fig. 3(b). We note that fiber cladding modes can be either well guided (resulting in the sharp grating resonances of Fig. 3(a)) or poorly guided (resulting in a broad continuum as in Fig. 3(b)), depending on the fiber surface. The radiation mode approximation is valid in the limit that the cladding mode is very poorly guided (i.e., has high loss) within the grating region.

A fundamental difference between FBGs and LPGs is the direction of the scattered mode, i.e., the sign of $k_{\text{scattered}}$ in Eq. (3). In the case of LPGs, $k_{\text{scattered}}$ has the same sign as k_{incident} , corresponding to a co-propagating mode. Thus, the grating wave vector $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}} = k_{\text{incident}} - k_{\text{scattered}}$ is orders of magnitude smaller than in the case of FBGs, where the counterpropagating $k_{\text{scattered}}$ has the opposite sign of k_{incident} . As we discuss below, this effect also makes LPGs more sensitive to modal dispersion.

While phase matching yields the *resonance wavelengths* of a fiber grating using only the effective indices, the *amplitudes* of these resonances depend on the length and profile of the gratings as well. These amplitudes must be derived from Maxwell's equations or a sufficiently accurate approximation. The most common approximation used in grating calculations is known as Coupled Mode Theory (CMT) (Kashyap, 2010; Othonos and Kalli, 1999; Kogelnik, 1990). In CMT, the longitudinal refractive index variation is assumed to be a small perturbation to the transverse refractive index distribution. Hence, in the framework of a first-order perturbation analysis, only the propagation constants of the modes change, but not the transverse structure of the E-fields of the modes (Eq. (1)). CMT then assumes that the E-field amplitude of each mode varies slowly (compared to a wavelength) along the grating. A pair of coupled equations relating the two mode amplitudes as a function of position along the grating can then be derived, and these allow for computation of the grating reflection and transmission. The coupling between two modes is determined by an overlap integral κ_{12} between the product of the E-fields of the two modes (E_1 and E_2) and the periodic refractive index variation that defines the grating:

$$\kappa_{12} \sim \iint_{\text{Area}} E_1(\rho, \theta) E_2^*(\rho, \theta) \delta n(\rho, \theta) dA, \quad (4)$$

where $\delta n(\rho, \theta)$ is the transverse part of the uniform grating refractive index modulation

$$\delta n(\rho, \theta, z) = \delta n(\rho, \theta) \cos(K_{\text{grating}} z) \quad (5)$$

For sufficiently weak gratings, this overlap integral allows for a comparison of the strength of resonances for different modes without the need to obtain a full CMT solution. In the simple case of Bragg reflection of the core mode in a standard single mode fiber (Fig. 1), the mode coupling coefficient κ_{12} becomes:

$$\kappa = \frac{\pi \eta \delta n}{\lambda} \quad (6)$$

where $\eta \approx 0.8$ is the overlap integral of the core mode E-field with itself within the core region that has the grating index modulation.

To understand fiber grating spectra, we first discuss a uniform grating of length L and period Λ . By definition, such a grating has uniform index modulation along its length, which is defined in Eq. (5) and illustrated in Fig. 4(a) and (b). Depending on the length and refractive index modulation of the grating, it may be classified as either weak or strong. The division between these two regimes may be understood by considering the light scattered by each successive period of the grating. The fraction of light

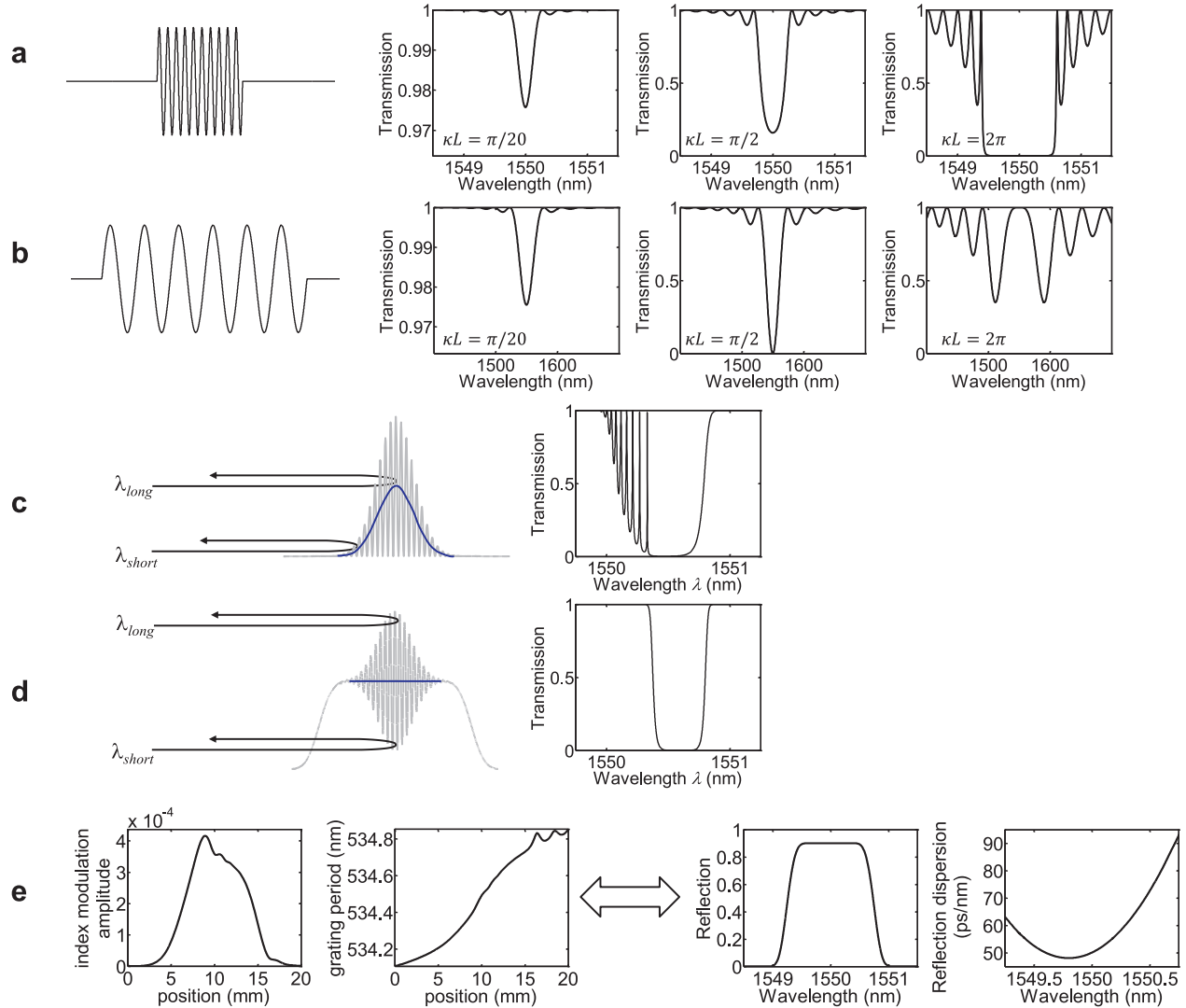


Fig. 4 Fiber grating spectra. (a) Uniform Bragg grating of length $L=3$ mm with period $\Lambda=534.5$ nm and (b) uniform long period grating of length $L=3$ cm with period $\Lambda=534.5$ μm , both with increasing levels of refractive index modulation, given by κL . (c) An “AC apodized” Gaussian grating profile of length $L=3$ cm and strength $\kappa L=8\pi$ showing Fabry-Perot like structure within the grating resonance. (d) A “100% DC apodized” Gaussian grating profile of length $L=3$ cm and strength $\kappa L=8\pi$ showing a symmetric transmission spectrum. (e) Complex grating profile and spectrum designed using inverse scattering. The two plots on the right show the desired flat reflection amplitude and nonuniform chromatic dispersion of reflected light. The two plots on the left show the computed profile of the grating index modulation and grating period required to obtain the desired spectrum.

scattered from each period can be approximated by $\delta n/2n$ using the Snell's Law reflectance formula. The boundary between a strong and a weak grating occurs when the grating is either long enough or strong enough that the sum of the scattering from all periods (L/Λ) is of the order 1: $(L/\Lambda)(\delta n/2n) \approx 1$. This division may also be expressed in terms of κ and L : $\kappa L \approx 1$ (here we are neglecting a factor $\pi/2 \approx 1$).

In the weak limit, the incident light propagates through the grating with very little scattering. In this case, both the LPG and FBG resonances have transmission spectra that approximate the Fourier transform of the grating index profile (See discussion below on inverse scattering. Also, for LPGs, this may not hold if modal dispersion is large; see the discussion below.). In the case of a uniform grating profile, this results in a 1-sinc²-shaped transmission spectrum with a spectral width $\delta\lambda/\lambda \sim \Lambda/L = 1/N_{\text{periods}}$ as shown in Fig. 4(a) for a weak FBG and in Fig. 4(b) for a similarly weak LPG. The characteristic side lobes of such a perfectly uniform grating result from the sudden "turn on" of the grating modulation. In the FBG, they can be understood as resonances of the effective Fabry-Perot cavity of length L that is defined by the sharp grating boundaries. Asymptotically, the spectral distance between adjacent side lobes is therefore $c_0/(2n_{\text{eff}}L)$ in frequency and $\lambda_{\text{Bragg}}^2/(2n_{\text{eff}}L)$ in wavelength. In the case of an LPG, the term $2n_{\text{eff}}$, which is the sum of the effective indices of the forward and backward propagating core mode, needs to be replaced by the difference $|n_{\text{eff,incident}} - n_{\text{eff,scattered}}|$ of the effective indices of the co-propagating modes.

In the strong limit, a substantial fraction of light is scattered out of the incident mode, and the FBG and LPG transmission spectra become different. As shown in Fig. 4(a), the FBG spectrum increases to a width determined only by δn : $\delta\lambda/\lambda = \delta n/n$. The incident light penetrates only a short distance ($\sim \lambda/\delta n$) into the grating before being completely reflected. The strong LPG in Fig. 4(b), on the other hand, becomes "overcoupled" as the second mode is reconverted into the incident mode at the center of the spectrum. The result is a higher transmission for the incident core mode (provided that the scattered mode propagates without loss). If we continuously increase κL , the transmission at the exact center of the spectrum of a uniform LPG becomes periodically stronger and weaker, with the period π : At $\kappa L = \pi/2, 3\pi/2, 5\pi/2$ etc., the transmission is zero, and at $\kappa L = 0, \pi, 2\pi$, etc., the transmission is maximized, see also Fig. 4(b). In contrast, the transmission in the center of a uniform FBG spectrum decreases monotonically with respect to κL , see Fig. 4(a).

In the case of a uniform Bragg grating of length L , the peak reflectivity may be computed analytically (Kashyap, 2010; Othonos and Kalli, 1999): $R_{\text{peak}} = \tanh^2 \kappa L$, where κ is the coupling constant defined in Eq. (6). This formula is approximately correct for many non-uniform Bragg gratings (e.g., the Gaussian apodized gratings of Fig. 4(c) and (d)) as well, and, with an appropriate choice of the effective length L , provides a useful estimate of the value of κ and the index modulation of the grating.

While uniform gratings are easy to model and manufacture, most practical fiber gratings have a nonuniform profile. In a nonuniform grating profile, the refractive index modulation amplitude, $\delta n_{\text{AC}}(z)$, and the average effective index value, $\delta n_{\text{DC}}(z)$, as well as the local grating wave vector $K_{\text{grating}}(z)$ and phase $\varphi(z)$ may vary along the fiber grating length:

$$\delta n(z) = \delta n_{\text{DC}}(z) + \delta n_{\text{AC}}(z) \cos(K_{\text{grating}}(z)z + \varphi(z)) \quad (7)$$

By controlling these parameters, a greatly improved filter may be fabricated. Fig. 4(c) shows the index profile and the grating transmission spectrum of a strong fiber Bragg grating with an "AC apodized" Gaussian profile. Such a profile results when a grating is inscribed with a spot exposure using a Gaussian beam. Note that due to the smooth increase in modulation amplitude, the sidebands of the uniform grating have been eliminated. However, because the DC index (and hence Bragg condition) is not constant within the grating, a sharp resonant structure is observed within the grating resonance. This structure can be understood as Fabry-Perot resonances in the cavity formed by the nonuniform Bragg condition within the grating. In contrast, Fig. 4(d) shows a grating with a "100% DC apodized" Gaussian profile. This grating has a constant DC refractive index in the region where the AC refractive index δn_{AC} is nonzero, and therefore the resonance spectrum is smooth and symmetric. Most useful bandpass filters require a profile close to 100% apodization. Although not shown in Fig. 4, a smoothly varying index profile also eliminates the sidelobes in LPG resonances. In practice, LPGs are more often fabricated with a uniform profile because of their greater sensitivity to variations in the phase matching condition within the grating.

Other important types of nonuniform fiber gratings include chirped gratings, phase shifted and superstructure fiber gratings. In a chirped fiber grating, the grating period $K_{\text{grating}}(z)$ slowly changes along the length of the grating. This spatial chirp of the grating period induces a proportional spatial chirp of the local grating wavelength and is analogous to the temporal chirp of an increasing or decreasing audio signal. As discussed below, chirped fiber gratings may be used as pulse compressors and chromatic dispersion compensators. In a phase shifted grating, a single discrete phase shift, for example, $\Delta\varphi = \pi$, is introduced in a uniform profile. Such a phase shifted grating can be used as a distributed feedback laser cavity (DFB). Superstructure gratings have a spatially modulated profile and can exhibit the same spectrum at regular frequency intervals determined by the period of the spatial modulation. The properties of nonuniform gratings may be computed using CMT and can also be understood intuitively using "band-diagrams" and an effective index model of the fiber grating response (Poladian, 1993).

While uniform and chirped fiber gratings are useful for most applications, it is also possible to design gratings with complex filter responses using a process called "inverse scattering", see, for example, Buryak et al. (2009). This process is usually applied to the design of FBGs, while having only limited use in designing LPGs. In such an algorithm, a given desired phase and amplitude response is used in one of several design methods to obtain the desired filter response. For instance, for some pulse compression applications, it is necessary to have both first and second order chromatic dispersion, while still maintaining a uniform reflection over the spectral bandwidth of the pulse. These parameters can be used to design a specific grating profile that will give the desired chromatic dispersion. Fig. 4(e) shows the grating profile and spectrum for such a grating.

Inverse scattering can be computationally intensive due the effect of multipath interference (MPI) within the grating. However, for sufficiently weak or short gratings, MPI effects can be neglected and the grating spatial profile and optical spectrum are more easily related. In this single scattering approximation, the coupled mode equations result in a simple Fourier transform relation between the (in general complex-valued) mode coupling coefficient κ and the dimensionless complex E-field reflection spectrum

$$r(\lambda) = \int_0^L \kappa(z) e^{i4\pi n_{\text{group}} z / \lambda} dz \quad (8)$$

where n_{group} is the group index of the propagating mode. This Fourier transform relationship between grating profile and spectrum can be seen in the weak ($\kappa L = \pi/20$) LPG and FBG spectra of Fig. 4(a) and (b).

Grating spectra showing radiation mode coupling are unlike FBGs or LPGs because there is little or no conversion of the radiation modes back to the incident core mode. Since the radiation mode spectrum is continuous, the spectrum of a grating that couples to radiation modes also smooths out into a broad continuum. The grating transmission spectrum depends less on the refractive index profile of the grating than on the overlap of the core mode and grating with the radiation modes (Kashyap, 2010). Radiation mode coupling is greatly affected by tilting the grating planes with respect to the axis of the fiber. Such tilted gratings may be designed as loss filters and optical fiber taps. Grating tilt has also been applied to decrease core mode back reflection (Kashyap, 2010; Othonos and Kalli, 1999).

While the resonance of an FBG is determined by the sum of the effective indices of two modes, an LPG resonance depends on the relatively tiny difference of two effective indices according to Eq. (3). This makes the resonance condition of an LPG very sensitive to any sort of perturbation along the waveguide. For instance, LPG resonances may exhibit a strong dependence on the modal dispersion (or variation of the propagation constant, k , with wavelength). This wavelength dependence can change the phase matching condition enough to dominate the wavelength dependence arising from CMT, making the resonance more or less narrow than predicted by a similar CMT with a linear phase matching curve (Kashyap 2010; Othonos and Kalli, 1999). In an extreme case, very broadband LPG resonances (> 100 nm at 1550 nm) may be produced between the core mode and a specially designed higher order mode whose propagation constant varies such that it is resonant with the core mode over a large bandwidth. Fig. 5 shows the spectrum and phase matching curves for both a standard LPG with a near linear phase matching curve, and a broadband LPG whose phase matching curve exhibits a turn around point (TAP). An LPG with a period and resonance wavelength near the TAP will exhibit a very broadband resonance.

Fiber Grating Fabrication

Photosensitivity

Fiber gratings can be fabricated with various high power light sources, including pulsed and continuous wave (CW) ultra violet (UV) sources, femtosecond infrared (IR) lasers and CO₂ lasers. The key early discovery leading to the study and practical application of fiber gratings was the phenomenon of UV-induced photosensitivity in the conventional germanosilicate optical fibers used in telecommunications systems (Hill *et al.*, 1978). UV irradiation of the germanosilicate core region can produce refractive index changes in the range 0.01 to 0.0001. Such refractive index changes are large enough to produce useful filters and reflectors. The refractive index change is most efficiently produced by irradiation close to 242 nm or beyond 200 nm, where

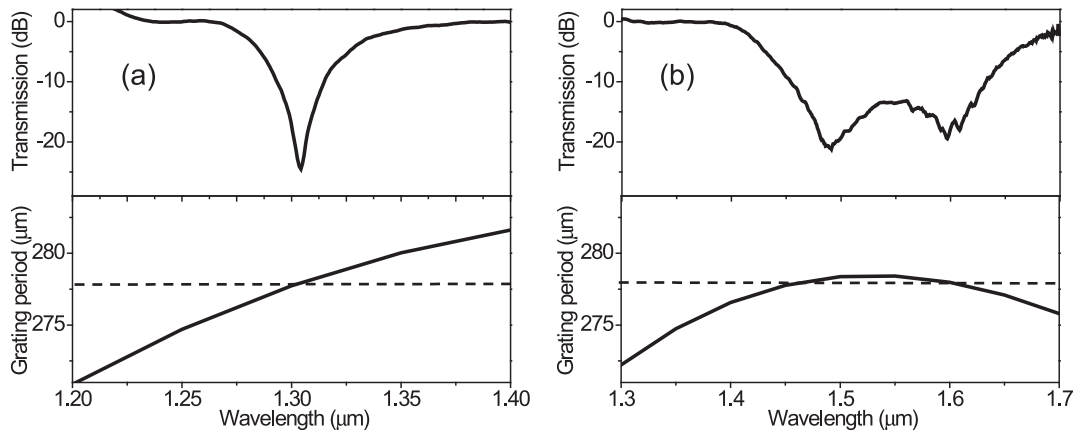


Fig. 5 Long Period Grating (LPG) phase matching curves. LPG fundamental mode transmission spectrum (top) and phase matching curve (bottom) for (a) standard, nearly linear phase matching, and (b) “turn around point” phase matching, resulting in a very broad transmission resonance. The solid line on the phase matching curves gives the resonance wavelength (x -axis) for a grating with a period indicated on the y axis. Dashed line indicates the grating period.

germanosilicate glass has strong absorptions. Numerous other dopants have been shown to be photosensitive at wavelengths ranging from 484 nm to 155 nm, including P, Ce, Pb, Sn, N, Sn-Ge, Sb, and Ge-B. Other dopants, such as Al and F, show very little photosensitivity to UV radiation. With other laser sources such as IR femtosecond and CO₂ lasers, the index modification results from thermal and multiphoton processes which can be used to inscribe gratings in a much larger range of materials, including pure silica. A drawback of this type of photosensitivity, though, is increased loss and more stringent alignment tolerances to yield the required intensity and placement of the grating index perturbations. For CO₂ lasers operating near 10.6 μm , there is a limit to the minimum grating period that can be inscribed. As a result, such lasers are used only for long period gratings.

A second major discovery enabling practical manufacture of fiber gratings was the effect of “hydrogen loading” (Lemaire *et al.*, 1993). Using this procedure, the fiber is placed in a pressurized hydrogen atmosphere until molecular hydrogen has diffused into the core region to a level of a few mole percent (a level similar to that of the photosensitive dopants). UV irradiation of the resulting fiber produces index changes an order of magnitude larger than in the unloaded fiber. Hydrogen loading allows gratings to be easily imprinted in conventional single mode fibers, whose Ge content is too low to allow for strong gratings. The increase in photosensitivity is true for most photosensitive dopants. Index changes in excess of 10^{-2} have been produced with hydrogen loading. In order to avoid increased OH absorption in the telecommunications bands (1520 nm–1630 nm), deuterium (D₂) is normally used in place of hydrogen.

At least two mechanisms are generally agreed to contribute to the UV induced index change. Firstly, as is evident from Fig. 6, the UV absorption spectrum is modified during UV exposure. This bleaching corresponds to a change in refractive index in the infrared which may be computed using the Kramers-Kronig relations. In particular, the absorption band at 242 nm resulting from germanium oxygen deficient centers is bleached, indicating that these defects are modified during exposure. The infrared index change results both from this bleaching and from the change in the UV absorption edge below 200 nm. Secondly, absorption of UV causes the germanosilicate matrix to contract. This compaction and the resulting stress distribution also change the refractive index.

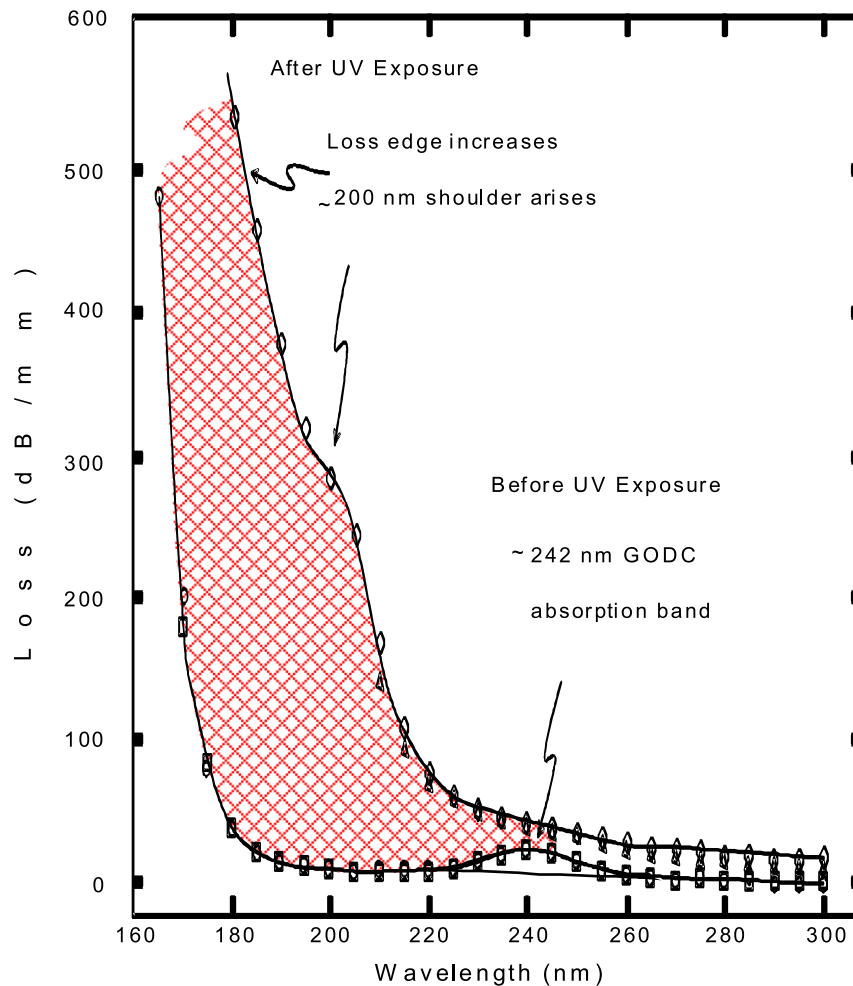


Fig. 6 UV absorption spectrum of Ge doped silica before and after irradiation with 242 nm light, showing bleaching of the Ge-oxygen deficient defect feature at 242 nm as well as a shift in the UV absorption edge.

Both mechanisms are enhanced by the presence of hydrogen in the silica lattice. The hydrogen reacts with oxygen, thus increasing the number of UV absorbing germanium oxygen deficient centers. As a result, in hydrogen loaded fibers, more of the Ge atoms participate in the index change, explaining the large increase in photosensitivity. The number of oxygen deficient defects may also be increased during fiber manufacture by collapsing the fiber preform in a low oxygen atmosphere. Such fibers show increased photosensitivity without the need for hydrogen loading. Co-doping a Ge doped fiber with boron has also been shown to increase the level of photosensitivity of a fiber for a given level of index change in the core.

An important aspect of UV induced index changes is that they require thermal annealing to ensure long term stability (Hill *et al.*, 1997). Sufficient exposure to elevated temperatures can completely erase the refractive index change in some Ge-doped gratings, for example, by several hours of heating to greater than 500°C. Thermal stability has been described by an activation energy model which assumes a distribution of energies for the defects giving rise to the refractive index change. Therefore, at a given anneal temperature, a fixed fraction of the index change will quickly decay, leaving only stable defects that easily survive at lower temperatures. By annealing at several temperatures for a fixed time, a relationship between temperature and time required for a given refractive index decay may then be derived and used in accelerated aging tests of fiber gratings. Annealing conditions vary depending on the application. One typical annealing condition for 20 year reliability is 120°C for 24 h (Kashyap, 2010; Othonos and Kalli, 1999). Gratings written using IR femtosecond radiation can show greater stability than UV written gratings due to the more significant modification that the IR multiphoton process produces in the fiber.

Grating Inscription

In general, fiber gratings are fabricated by exposure of the fiber through the side of the fiber. In the most common method, a UV interference pattern is projected through the side of the fiber onto the core region. The first demonstration of this technique employed a free space interferometer to produce the UV fringes (Meltz *et al.*, 1989). A significant improvement on this technique employs a zero order nulled phase mask to generate the interference pattern. The phase mask is typically a quartz plate with grooves patterned and etched into it using standard semiconductor industry techniques. Such a phase mask splits an incident beam into multiple orders, and these form the interference pattern that is projected onto the core of the fiber. The method is shown in Fig. 7. The advantage of the phase mask technique is that the fringes are very stable, since the UV beams propagate only a short distance. Another important technique for writing fiber gratings is the point-by-point or direct writing method. In a point-by-point technique, the UV interference pattern is modulated and translated along a fiber in such a way that a long grating may be formed. Although complex and sensitive, such methods can yield gratings as long as a few meters and also allow for the fabrication of complex grating profiles. Point-by-point methods have been used to fabricate broadband chirped dispersion compensating gratings and Bragg reflectors with very low dispersion.

Both phase mask and point-by-point methods can be used with IR femtosecond writing beams as well (Mihailov *et al.*, 2004). In addition, femtosecond lasers can be used to inscribe the grating line by line with a focused beam that creates a single index modification. Precision translation is required to ensure precise periodicity and a well defined optical spectrum.

LPGs may also be fabricated with UV or IR radiation. An amplitude mask is typically used, though point-by-point methods are also possible. Requirements for beam coherence is greatly reduced because of the long period (typically a few hundred microns). Other techniques may also be used to fabricate LPGs. The fiber may be periodically deformed by heating or with periodic micro-bends. Dynamic LPGs may also be produced by propagating acoustic waves along the fiber (Kashyap, 2010; Othonos and Kalli, 1999).

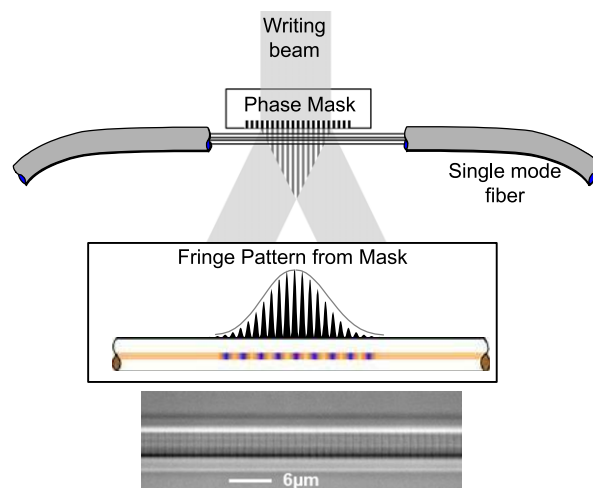


Fig. 7 Phase mask fabrication of a fiber grating. Diffraction by the phase mask grooves yields two beams that form an interference pattern projected into the fiber to write the grating. Bottom: High resolution image of the grating refractive index modulation.

Applications

Fiber gratings are key enabling technologies in a number of important applications. In this section we review these, highlighting the unique functionalities that can be provided by fiber gratings.

Fiber Sensors

Fiber gratings are sensitive to both strain and temperature and may therefore be used as sensors. Both LPG and FBG resonances change with strain and temperature. Since LPGs show greater but more complicated sensitivity, FBGs are typically used in such applications. Strain sensitivity arises from the fact that gratings are extended in length. Because fibers are very thin (diameter $\sim 125 \mu\text{m}$), significant extension is possible. A smaller strain sensitivity arises from the stress-optic effect. Typical sensitivity at 1550 nm is $\sim 1.2 \text{ pm}/1 \mu\epsilon$ in FBGs. Temperature sensitivity arises from temperature dependence of the refractive index and gives rise to a change of order $0.01 \text{ nm}/^\circ\text{C}$ of temperature change of the FBG. FBG strain sensors have been applied to map stress distributions and can be employed in a distributed manner with many Bragg reflectors in one fiber being examined with one detector. Fig. 8 shows a typical distributed Bragg grating sensor application that uses wavelength division multiplexing (WDM). Every position along the array has a sensor with a different Bragg wavelength, allowing for many sensing points that are mapped onto a spectrum of reflection peaks. Shifts in these peaks correspond to variations in strain at the corresponding sensing point.

These sensitivities may also be used to make fiber grating filters tunable. Temperature changes typically give 1–2 nm of tunability in an FBG resonance at 1550 nm. Strain tuning is more difficult to package commercially but can give up 4 nm in extension and more than 10 nm in compression. Thermal tuning is typically slower than strain tuning. Passive thermal stabilization of grating resonances may also be achieved by designing packages in which strain and temperature variations cancel each other.

It is also possible to inscribe gratings in fiber with more than one core. Such multicore fiber gratings can be used in applications such as shape sensing (Westbrook *et al.*, 2017). Light scattering from the center core and several offset cores is collected in an interrogator and used to reconstruct the shape of the fiber. Twisted offset cores provide sensitivity to the bend and twist direction of the fiber. Fig. 9 shows a twisted multicore fiber grating. Gratings may be inscribed in such fibers using phase mask inscription and UV irradiation similar to that used in single core gratings.

Fiber Grating Stabilized Diode Lasers and Fiber Lasers

One of the first important commercial applications of fiber gratings was in stabilization of fiber coupled diode lasers. Normally, these lasers operate on several cavity modes making the output unstable. If feedback is provided in the form of a narrow band reflection, then the laser output will be only in this narrow bandwidth. The fiber grating is implemented in the fiber pigtail of the diode laser and typically provides a few percent reflectivity over a bandwidth of $\sim 1 \text{ nm}$. The application is depicted schematically in Fig. 10(a).

More generally, fiber gratings have been used to realize fiber laser cavities. Typically, a laser cavity may be defined by inscribing a high reflector (HR) and an output coupler (OC) at matched wavelengths. The resulting cavity may then be pumped at a wavelength outside the reflection spectra of the narrow HR/OC pair. Fig. 10(b) illustrates an example of a fiber laser formed by an HR/OC pair. Fig. 10(c) shows an example spectrum of the broadband HR grating and the much weaker OC grating. Another important application is the cascaded Raman resonator fiber laser, in which several FBGs recycle Raman pump light to produce high power, fiber-coupled light at any design wavelength. As described earlier, it is also possible to inscribe distributed feedback (DFB) cavities using fiber Bragg gratings. Such gratings typically have a uniform profile and a single phase shift at the center of the

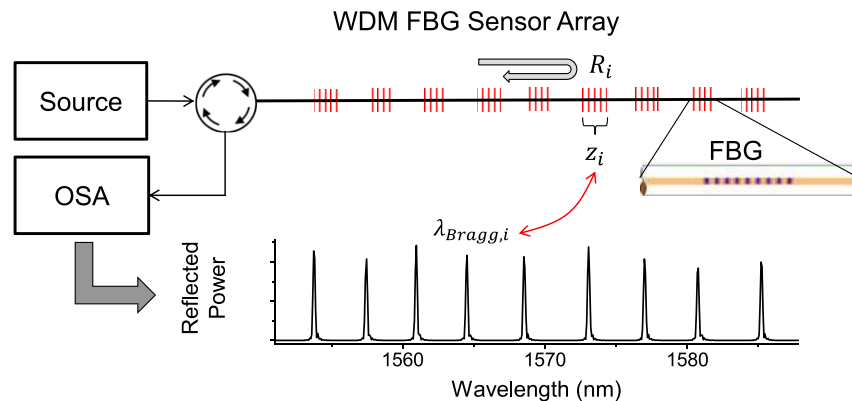


Fig. 8 Wavelength division multiplexed (WDM) fiber Bragg grating (FBG) sensor array. Signals reflected from sensor grating at z_i are measured with the Optical Spectrum Analyzer (OSA) and mapped to Bragg wavelength $\lambda_{\text{Bragg},i}$ whose resonance wavelength shift yields strain and temperature information from position z_i .

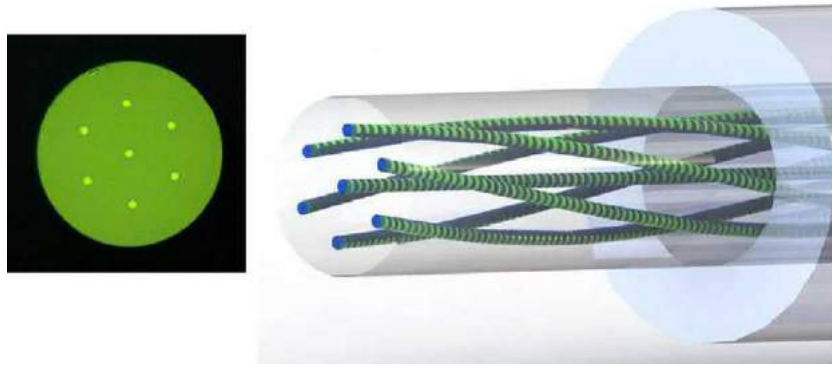


Fig. 9 Twisted multicore fiber with continuous grating sensor array used for fiber bend and shape sensing.

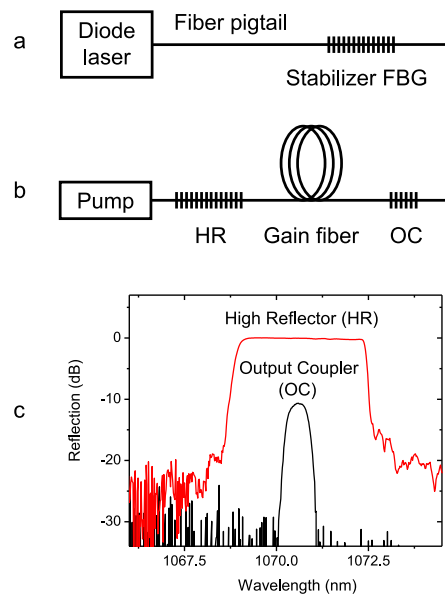


Fig. 10 FBGs in laser applications. (a) Fiber Bragg Grating stabilized diode laser. Fiber grating reflector in the laser pigtail provides small level of feedback to stabilize laser output power and wavelength. (b) Fiber laser cavity defined by a high reflector (HR) and output coupler (OC) grating. (c) Typical spectra for the HR and OC.

grating to define an efficient, low threshold cavity. Fiber DFB lasers can provide excellent performance at extremely narrow linewidths below 1 kHz.

Optical Filtering

Fiber gratings offer unprecedented in-fiber, low loss, filtering of guided wave signals. Filter bandwidths from 10 pm to more than 100 nm are possible. A primary strength of fiber gratings is the diversity of filter shapes that is achievable through modification of the grating profile. One example is a gain flattening filter. Such filters have seen important application in telecommunications systems in order to smooth the wavelength dependence of fiber amplifiers (such as Er or Raman amplifiers). Both LPGs and FBGs may be employed as gain flattening filters. The desired loss spectrum is translated into a corresponding grating profile. LPGs have the advantage of low return loss and are useful for filters with bandwidths larger than a few nm. FBGs may also be used in both tilted and untilted form and are often accompanied by an optical isolator to reduce back reflections. Phase shifted FBGs can be configured with a circulator to provide a very narrow band filter for some applications. **Fig. 11** shows an example of a gain equalizing loss filter using LPGs.

Chromatic Dispersion Control

Chromatic dispersion arises from the variation in propagation velocity with wavelength. A temporally sharp pulse is thus dispersed and broadens out in the time domain. Fiber Bragg gratings can provide strong chromatic dispersion when operated in reflection

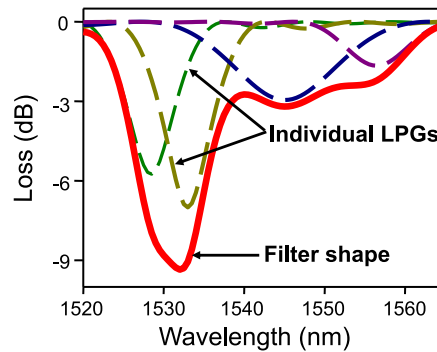


Fig. 11 Long Period Grating gain equalizing filter. Several individual gratings (dashed line) are combined to produce the overall shape (solid line).

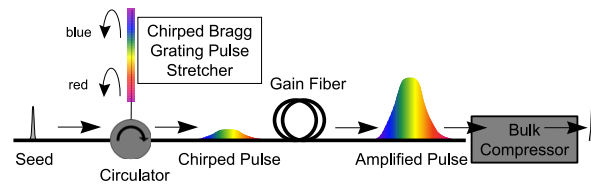


Fig. 12 Chirped pulse amplifier (CPA) using a chirped fiber Bragg grating to stretch the input pulse before amplification.

with a 3-port circulator. In this case, the period is made to vary along the length of the grating. This means that different wavelengths in a pulse reflect at different positions in the grating. This gives rise to wavelength dependent group delay and therefore chromatic dispersion.

Such chirped gratings can be used to compress and disperse pulses. Pulse compression can improve pulsed telecommunications signals that have been broadened by the dispersion of an optical fiber. On the other hand, in high power pulse amplification, such gratings can be used to greatly broaden a pulse, thereby reducing its peak power and allowing for amplification without excess nonlinearities. Such an amplifier is called a chirped pulse amplification system (CPA). **Fig. 12** shows the use of a chirped fiber grating pulse stretcher in a CPA system.

Nonlinear Applications of Fiber Gratings

While most applications of gratings involve linear propagation of light, when the optical power in the fiber is high enough, nonlinear propagation effects are also observed. While such nonlinearities can limit the performance of fiber grating devices, they can also lead to potentially useful effects, some of which we discuss here. Optical nonlinearities arise from the intrinsic nonlinearity of the material system in which the grating is inscribed (Agrawal, 2001). In the case of optical fibers, the dominant nonlinearity is the Kerr nonlinearity, which is a third order effect that manifests as a change in the refractive index of the material that depends on the peak intensity of the optical field: $n = n_0 + n_2 I$. This nonlinearity has been studied in a number of effects, including all-optical switching, bistability, multistability, soliton generation, pulse compression and wavelength conversion (Slusher and Eggleton, 2013). Critical to these phenomena is the shift in the grating reflectivity as the input power is increased. Sufficiently high power can result in increased transmission through a strong Bragg grating due to the shift of the Bragg resonance. For certain high power pulses, these propagation effects may result in the formation of a Bragg soliton, i.e., a pulse that maintains its shape during propagation, within the photonic bandgap formed by the fiber grating. The enormous dispersion provided by the fiber Bragg grating is balanced by the fiber nonlinearity in a manner similar to solitons observed in standard single mode fiber where the dispersion is determined by material properties and waveguide design. Since the dispersion of a Bragg grating can be orders of magnitude larger than that of the bare fiber, Bragg solitons can be observed in gratings of only a few centimeters in length. Such Bragg solitons can show a considerable decrease in group velocity, allowing the possibility of an optical delay line.

Another set of nonlinear effects can occur if continuum generation occurs within a fiber Bragg grating. Continuum generation of very intense pulses can occur in a fiber due to the interplay of the full third order nonlinear interaction with the effect of fiber dispersion. For certain input pulsed sources, the resulting continuum is comprised of a comb of individual frequency lines, and such optical frequency combs can be used in precision spectroscopy. When such a process occurs in the presence of a fiber Bragg grating, the formation of the continuum is greatly modified in the spectral region near the grating (Westbrook *et al.*, 2004). The spectral density near the Bragg resonance can increase by more than an order of magnitude due to the interplay of the grating dispersion and the continuum formation, in effect producing a more efficient phase matching of the nonlinear processes in that

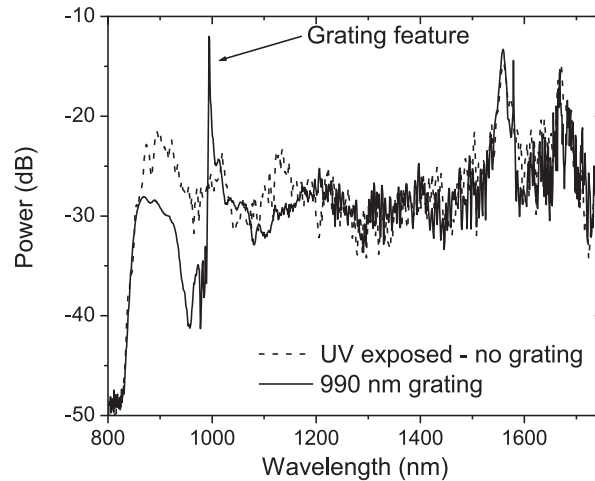


Fig. 13 Grating enhanced continuum. Continuum generation in a fiber with and without a fiber Bragg grating at 990 nm, showing a large enhancement peak in the continuum near the fiber grating resonance.

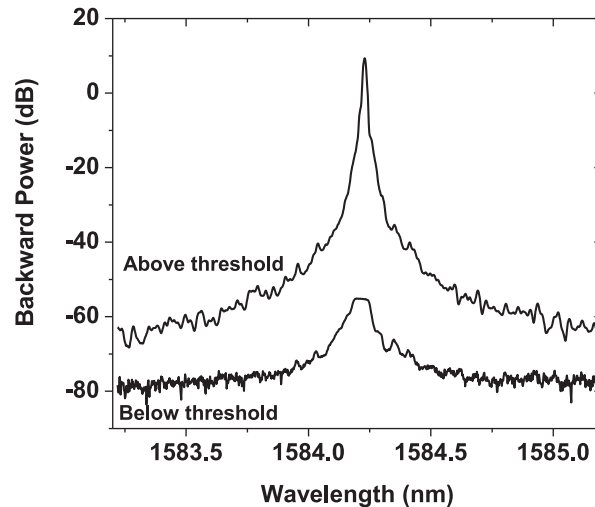


Fig. 14 Raman fiber distributed feedback (DFB) laser. Optical spectrum back reflected from a phase shifted fiber grating at 1584.2 nm pumped by a Raman pump near 1480 nm, showing a Bragg reflection spectrum at low pump power and a narrow linewidth lasing line when the Raman gain exceeds the lasing threshold within the Bragg grating DFB cavity.

spectral region. **Fig. 13** shows an optical continuum with and without a grating present in the fiber. On the long wavelength side of the grating, the spectrum is enhanced by more than a factor of 10. Such grating enhancement peaks have been shown to increase the signal-to-noise ratio by more than an order of magnitude in applications that use optical frequency combs for precision measurements.

It is also well known that fibers exhibit Raman gain, another manifestation of the third order nonlinearity of glass materials. An optical pump at any wavelength will result in Raman gain for a range of wavelengths longer than the pump wavelength. Such gain can be used to produce Raman fiber amplifiers and Raman fiber lasers. Raman lasers are very noisy and exhibit large linewidths, however, it is possible to exploit Raman gain within a fiber distributed feedback (DFB) cavity to achieve narrow linewidth lasing. The DFB cavity is formed by a FBG with a phase shift in the center. **Fig. 14** shows an example of such a Raman DFB fiber laser (Westbrook *et al.*, 2011). The lower trace on the plot shows the light back reflected from the DFB fiber grating at low pump power. Raman amplified spontaneous emission (ASE) generated in the fiber back reflects off of the grating spectrum and reveals a typical flat-topped reflection spectrum. The upper trace shows lasing from the cavity when the Raman pump power is increased beyond the lasing threshold. Raman gain and feedback from the FBG DFB cavity cause light intensity to build up, resulting in a strong and very narrow back reflected signal centered at the Bragg wavelength. Such lasers can generate light over a range of frequencies defined by the pump and DFB Bragg wavelengths and exhibit linewidths orders of magnitude narrower than a typical Raman fiber laser.

See also: Diffraction Gratings. Nonlinear Optics

References

- Adams, M.J., 1981. An Introduction to Optical Waveguides, vol. 14. New York, NY: Wiley.
- Agrawal, G., 2001. Applications of Nonlinear Fiber Optics. Academic press.
- Buryak, A., Bland-Hawthorn, J., Steblina, V., 2009. Comparison of inverse scattering algorithms for designing ultrabroadband Bragg gratings. *Optics Express* 17 (3), 1995–2004.
- Hill, K.O., Fujii, Y., Johnson, D.C., Kawasaki, B.S., 1978. Photosensitivity in optical fiber waveguides: Application to reflection filter fabrication. *Applied Physics Letters* 32 (10), 647–649.
- Hill, K.O., Russell, P.S.J., Meltz, G., Vengsarkar, A.M. (Eds.), 1997. *IEEE Journal of Lightwave Technology: Special Issue on Fiber Gratings, Photosensitivity, and Poling* 15 (8), 1261–1511.
- Kashyap, R., 2010. Fiber Bragg Gratings, second ed. New York, NY: Academic.
- Kogelnik, H., 1990. Theory of optical waveguides. In: Tamir, T. (Ed.), *Guided-Wave Optoelectronics*. Berlin: Springer-Verlag.
- Lemaire, P.J., Atkins, R.M., Mizrahi, V., Reed, W.A., 1993. High pressure H/sub 2/loading as a technique for achieving ultrahigh UV photosensitivity and thermal sensitivity in GeO/sub 2/doped optical fibres. *Electronics Letters* 29 (13), 1191–1193.
- Marcuse, D., 1974. *Theory of Dielectric Optical Waveguides*. New York, NY: Academic.
- Meltz, G., Morey, W., Glenn, W.H., 1989. Formation of Bragg gratings in optical fibers by a transverse holographic method. *Optics Letters* 14 (15), 823–825.
- Mihailov, S.J., Smelser, C.W., Grobnc, D., *et al.*, 2004. Bragg gratings written in all-SiO₂ and Ge-doped core fibers with 800-nm femtosecond radiation and a phase mask. *Journal of Lightwave Technology* 22 (1), 94.
- Othonos, A., Kalli, K., 1999. *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.
- Poladian, L., 1993. Graphical and WKB analysis of nonuniform Bragg gratings. *Physical Review E* 48 (6), 4758.
- Slusher, R.E., Eggleton, B.J. (Eds.), 2013. *Nonlinear Photonic Crystals*, vol. 10. Springer Science & Business Media.
- Westbrook, P.S., Abedin, K.S., Nicholson, J.W., Kremp, T., Porque, J., 2011. Raman fiber distributed feedback lasers. *Optics Letters* 36 (15), 2895–2897.
- Westbrook, P.S., Kremp, T., Feder, K.S., *et al.*, 2017. Continuous multicore optical fiber grating arrays for distributed sensing applications. *Journal of Lightwave Technology* 35 (6), 1248–1252.
- Westbrook, P.S., Nicholson, J.W., Feder, K.S., Li, Y., Brown, T., 2004. Supercontinuum generation in a fiber grating. *Applied Physics Letters* 85 (20), 4600–4602.

Optical Amplifiers: SOAs

Michael J Connelly, University of Limerick, Limerick, Ireland

© 2018 Elsevier Inc. All rights reserved.

Introduction

The rapid growth in the development of optical communication systems requires small, inexpensive and easy to integrate Semiconductor Optical Amplifiers (SOAs) for basic applications such as power boosters, in-line amplifiers and receiver pre-amplifiers. SOAs can also be used to carry out optical signal processing functions such as wavelength conversion, modulation, demultiplexing, switching, logic and regeneration (Connelly, 2002; Dutta and Wang, 2013).

Optical Fiber Amplifiers (OFAs), which cannot be integrated, are often the choice for power, in-line and preamplifier applications where small-size is usually not a critical factor. Their advantages include wide bandwidth (10s–100s of nm), high gain, high saturation output power $P_{o,sat}$, defined as the output power at which the amplifier gain is half the unsaturated gain, low Noise Figure (NF) and low polarization sensitivity. Because OFA gain dynamics are very slow, they do not impart significant distortion to amplified optical data signals irrespective of the modulation format; however this precludes their use in optical signal processing applications.

The gain mechanism in SOAs is based on achieving a population inversion between the conduction and valence bands of the amplifier active region material driven by an electrical current. Net amplification results when stimulated emission induced by the amplified lightwave exceeds losses due to stimulated absorption and other material or structural losses. By appropriate choice of the active material SOAs can be designed to operate in the wavelength region of choice, usually one of the optical communications bands in the 1.3 μm and 1.55 μm regions. SOAs have gain, NF, $P_{o,sat}$ and bandwidth values comparable to OFAs, as well as low polarization sensitivity. Compared to OFAs, the principle advantages of SOAs are their small size and compatibility with Photonic Integrated Circuits (PICs). SOAs have much faster dynamics than OFAs, which can result in signal distortion; but in association with nonlinearities can be exploited to realize all-optical signal processing functions. Advanced SOAs based on Quantum-Dot (QD) structures have demonstrated processing speeds in the THz range.

This article reviews SOA principles and structures, some simple SOA models that provide insights into SOAs physics and performance, basic network applications and the important all-optical signal processing applications of wavelength conversion, optical logic, data regeneration, optical remodulation and microwave phase shifting.

SOA Principles

The principle of operation of an SOA is shown in Fig. 1, where Antireflection Coatings (ARs) are used to suppress end facet reflections that would result in the SOA acting as a Fabry-Perot cavity leading to ripples in the gain spectrum and the possibility of oscillation at high gains. Such SOAs are often referred to as traveling-wave SOAs. The input lightwave is amplified as it propagates through the electrically pumped active waveguide. Although SOA waveguides are designed to be single-mode they support two orthogonal polarization modes, the Transverse Electric (TE) and Transverse Magnetic (TM) modes. Confinement of the signal lightwave to the active region is quantified by the optical confinement factor Γ , defined as the fraction of the transverse (to the propagation direction) optical intensity overlapping with the active region. Γ is usually polarization dependent, so the TE and TM confinement factors Γ_{TE} and Γ_{TM} respectively are unequal except in SOAs having a square cross-section active waveguide; however such SOAs are difficult to fabricate. SOAs usually have rectangular cross-section waveguides so $\Gamma_{TE} \neq \Gamma_{TM}$; a consequence of which is that if the active material gain is polarization independent the SOA gain will be polarization dependent.

A consequence of optical amplification is the addition of broadband Amplified Spontaneous Emission (ASE) noise due to the spontaneous recombination of the active material charge carriers – conduction band electrons and valence band holes. In the linear operating region (low input signal power), the gain and ASE do not depend on the input signal power. In most practical applications a narrowband optical bandpass filter of bandwidth B_o , which passes the amplified signal, is placed after the SOA to reduce the ASE. If the filter is assumed to be an ideal rectangular filter centered at the signal photon energy E_s , the noise power

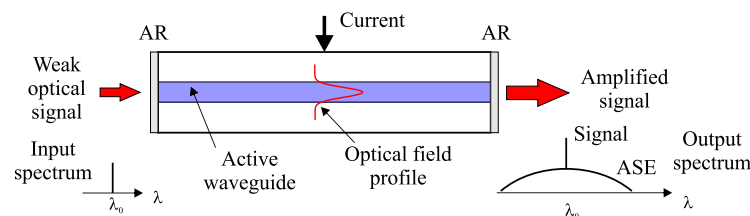


Fig. 1 Basic concept of an SOA.

spectral density, in a single-polarization, is given by

$$\rho_{ASE} = n_{sp} E_s (G - 1) \quad (1)$$

where G is the gain and n_{sp} is the population inversion factor. A common metric for quantifying the noise properties of an amplifier is the NF , defined as the ratio in dB of the input to output Signal-to-Noise Ratios (SNRs) as measured in terms of the signal and noise levels of the current from an ideal photodetector having responsivity R (A/W) and electrical bandwidth B_e ; in linear terms the ratio is called the noise factor F . If the input signal power is P_s , its associated electrical current is RP_s . Assuming the input signal is quantum noise (often referred to as shot noise) limited, its current variance is $\sigma_{in}^2 = 2eRP_s B_e$ (A^2) and so the input SNR $SNR_{in} = P_s / (2E_s B_e)$. The output noise current consists of several uncorrelated components: the quantum noise of the amplified signal, the quantum noise of the ASE, the signal-spontaneous beat noise and the spontaneous-spontaneous beat noise, having respective current variances (Olsson, 1989).

$$\sigma_s^2 = 2eRGP_s B_e \quad (2)$$

$$\sigma_{sp}^2 = 2eR\rho_{ASE} B_o B_e \quad (3)$$

$$\sigma_{s-sp}^2 = 4R^2 GP_s \rho_{ASE} B_e \quad (4)$$

$$\sigma_{sp-sp}^2 = 2R^2 \rho_{ASE}^2 B_e (2B_o - B_e) \quad (5)$$

The total current variance $\sigma_{out}^2 = \sigma_s^2 + \sigma_{sp}^2 + \sigma_{s-sp}^2 + \sigma_{sp-sp}^2$. The output SNR $SNR_{out} = (RGP_s)^2 / \sigma_{out}^2$. If B_o is small enough the signal dependent noise components dominate, in which case

$$NF = 10 \log_{10} \left(\frac{2\rho_{ASE}}{GE_s} + \frac{1}{G} \right) \quad (6)$$

When, $G \gg 1$, the signal-spontaneous beat noise dominates, so

$$NF = 10 \log_{10} (2n_{sp}) \quad (7)$$

The maximum possible value of n_{sp} is 1, so the minimum achievable NF is 3 dB.

SOA Structures

The key parameters required for practical SOAs are:

- Low end reflectivities (typically $< 10^{-4}$).
- Low polarization sensitivity (< 0.5 dB)
- Wide optical bandwidth (10s nm).
- High gain at low currents.
- High $P_{o,sat}$.
- Low fiber-to-chip coupling losses.
- Fast gain recovery time to reduce amplified signal distortion.

The aim of most SOA structural designs, choice of active region material and packaging is to realize some or all of the above objectives. SOAs can also be incorporated into complex configurations using PICs or discrete optics to perform optical signal processing functions, for which it may be desirable for the SOA to possess enhanced nonlinearities. In this section a bulk material SOA is described, which illustrates practical SOA operation, as well as SOAs based on semiconductor nanostructures that have several important advantages over bulk devices.

Bulk SOA

A typical SOA, fabricated from bulk material, operating in the 1550 nm region, is shown in Fig. 2. The active region is sandwiched between two Separate Confinement Heterostructure (SCH) layers having refractive indices less than the active region refractive index, which provides waveguiding, i.e., helps prevent the propagating lightwave from spreading into the surrounding lossy regions. The p-n junctions formed by the p- and n-type InP layers act as current blocks and provide good confinement of the injected carriers in the active region. Tensile strain is introduced between the active region and SCH layers via lattice mismatching. This increases the ratio of the TM to TE material gain coefficients to compensate for the higher Γ_{TE} caused by the waveguide asymmetry thereby reducing polarization sensitivity. The tapered regions act as mode-expanders that couple light from the active waveguide to an underlying passive waveguide to simplify coupling of the amplified light to lensed optical fibers. Low end reflectivities ($< 10^{-5}$) are obtained by combining buried windows and a 7° waveguide tilt angle with respect to the AR coated end facets.

Typical tensile-strained SOA static characteristics are shown in Fig. 3. The small-signal gain spectrum and ASE spectrum (Fig. 3 (a) and (b)) at a given current have similar shapes near the peak wavelength but are significantly different as the wavelength deviation from the peak increases. At high bias currents the 3 dB gain bandwidth is approximately 65 nm. At high currents the gain

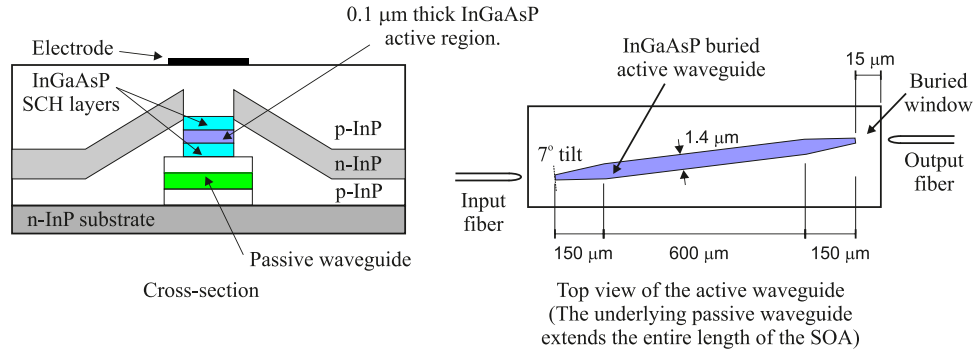


Fig. 2 Tensile-strained bulk SOA with some typical dimensions.

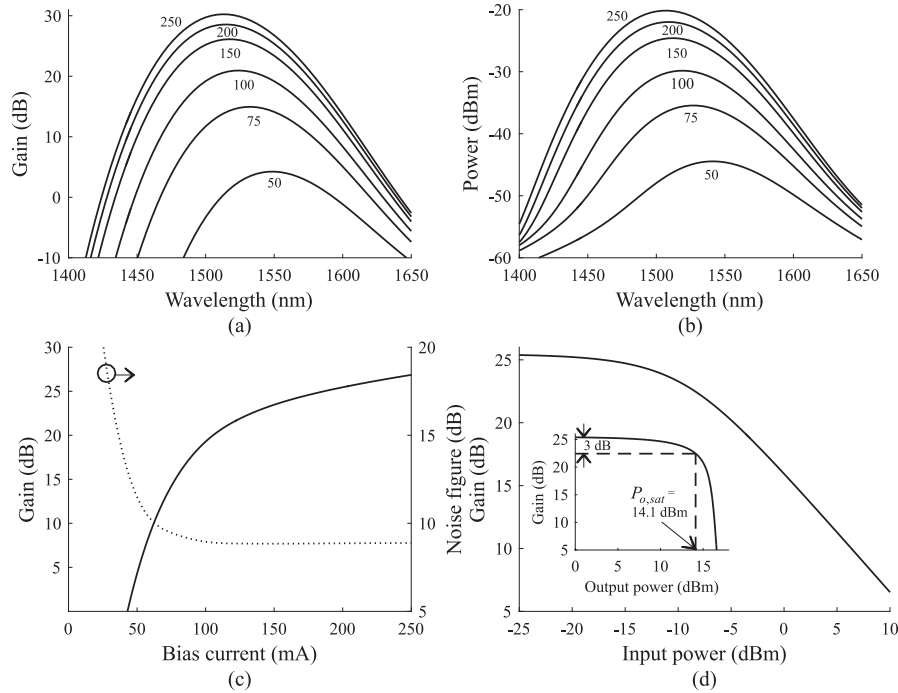


Fig. 3 Typical SOA experimental static characteristics. (a) Small-signal gain vs. wavelength and (b) output ASE spectrum with 0.07 nm resolution. The parameter is the bias current in mA. (c) Small-signal gain and NF vs. bias current at 1550 nm. (d) Gain vs. input signal power at 1550 nm and a bias current of 200 mA. The inset shows the gain vs. output signal power and the value of $P_{o,sat}$.

saturates because the ASE depletes the conduction band electrons, which are then not available to impart gain to the signal. Saturation is also evident in the small-signal gain versus bias current characteristic shown in Fig. 3(c). The NF decreases as the gain increases as shown in Fig. 3(c), which includes a degradation of 2 dB due to the output fiber coupling loss. The SOA gain dependency on input power is shown in Fig. 3(d); the inset shows the gain as a function of output power from which $P_{o,sat}$ can be determined.

Semiconductor Nanostructure SOAs

In a bulk SOA, the motion of the carriers is not restricted. Placing semiconductor layers or structures with a reduced dimensionality in the active region results in a restriction of the carrier motion and has a very significant impact on SOA behavior. The motion of a carrier is restricted if the material dimension is of the order of the carrier de Broglie wavelength – of the order of 7–70 nm in typical semiconductors. Reducing the material dimension in one (Quantum-Well, QW), two (Quantum-Wire) and three dimensions (quantum-dot) results in carrier confinement in 1D, 2D and 3D respectively. Reduced dimensionality semiconductors are referred to as semiconductor nanostructures and are usually embedded in a bulk semiconductor of larger bandgap energy. A Quantum Dash (QDash), or short wire, is an elongated nanostructure, having a cross-section similar to a QD but with a length of typically 100 s of nm (Lelarge *et al.*, 2007).

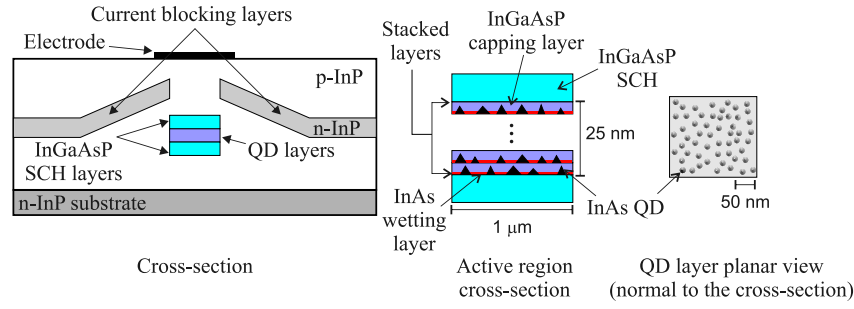


Fig. 4 QD SOA structure with some typical dimensions (not to scale). Typical dot densities are $(2 - 10) \times 10^{24} \text{ mm}^{-2}$.

The active region of a single QW SOA is very similar to that for a bulk SOA except that the thickness is much lower. This permits many QWs to be stacked to form a Multiple-QW (MQW) active region, to mitigate the effect of the reduced single QW confinement factor and thereby increase the gain coefficient. MQW SOAs have a number of advantages compared to bulk SOAs. Because of the small active region volume compared to bulk SOAs, a reduced injection current is sufficient to create a large carrier density resulting in a broader gain spectrum and shorter carrier lifetime τ_c (time for conduction band electrons to recombine with valence band holes). Lower τ_c values result in reduced gain recovery times, which is of particular importance in reducing pattern effects when the SOA is used to amplify high-speed data signals. The loss coefficient in MQW active regions is significantly smaller than that in bulk devices, which leads to an improvement in noise performance. The intrinsic polarization sensitivity of QW structures can be reduced by using strained QWs. Strain-induced band structure modifications also result in a reduction in loss mechanisms such as Auger recombination and intervalence band absorption (Chuang, 2009). The Linewidth Enhancement Factor (LEF), a proportionality factor relating phase changes to gain changes is also reduced, which results in amplified pulses experiencing less spectral broadening, leading to superior high-speed performance compared to bulk and unstrained QW SOAs. Bulk and QW SOAs have typical gain recovery times in the range of 100 s picoseconds so, when used to amplify data signals, significant pattern effects can be present at symbol rates above 10s Gbaud/s. Pattern effects also induce Pattern Dependent Phase Distortion (PDPD) due to Self-Phase Modulation (SPM). SPM in an SOA arises when a high intensity signal depletes the carrier density, which leads to a modification in the refractive index and thereby the instantaneous phase of the propagating lightwave. PDPD can be a more serious performance degradation factor than pattern effects in coherent optical communication systems employing advanced multi-level modulation formats such as Phase Shift Keying (PSK) and especially Quadrature Amplitude Modulation (QAM). The amount of SPM is directly related to the LEF.

A Quantum-Dot (QD) is a semiconductor nanostructure having dimensions that typically range from 2 to 10 nm, which confines the motion of injected carriers to all three spatial directions. In QD SOAs (Akiyama *et al.*, 2007) the device structure for implementing electrical pumping and waveguiding, surrounding the QD material active region is very similar to that for bulk and QW SOAs as shown in Fig. 4. The active region consists of stacked layers of QD material fabricated using the Stranski-Krastanov (SK) growth technique that results in self-assembled QDs. Initially, a lattice matched strained InP 2D film (wetting layer) is formed. The accumulated strain is relieved by the random formation of 3D InP islands. The growth of an InGaAsP capping layer completes the QD formation. If the capping layer has a wider bandgap than the substrate a QW structure around the dots is formed, which extends the emission wavelength of the dots thereby leading to a larger gain bandwidth. Self-assembled SK grown QDs are inherently polarization dependent having only TE mode gain. Low polarization sensitivity QD SOAs have been realized by using tensile-strained capping layers or through the use of optimized QD shapes. Compared to bulk and MQW SOAs, QD SOAs have shown to have significantly improved gain bandwidth (100s nm), $P_{o,sat}$ ($> 25 \text{ dBm}$) and NF ($\sim 5 \text{ dB}$). The former two properties are especially important when the SOA is used to simultaneously amplify many channels as is the case in Wavelength Division Multiplexed (WDM) systems. Of particular significance is the short τ_c typically a few picoseconds, which is a factor of ten lower than that for bulk and QW SOAs, so QD SOAs are capable of amplifying ultrafast data signals with little or no pattern effects. The physical mechanism leading to short τ_c is that when the QD electrons are depleted by an incoming lightwave, they can be refilled very quickly by the electron population in the material surrounding the dots. Because of these advantages and also enhanced nonlinear effects, QD SOAs are very useful for efficient high-speed all-optical signal processing applications. QD SOA LEFs can be much less than for bulk and MQW SOAs and so are particularly suited for amplifying very high baud rate advanced modulation format data signals.

QDash SOAs are of interest as an alternative to QD SOAs, since they possess some dot-like properties, especially wide gain bandwidth and ultrafast gain recovery time (similar to QD SOAs), and can more easily be made to operate in the $1.55 \mu\text{m}$ region.

SOA Models

Mathematical models of SOAs are useful in understanding static, dynamic and nonlinear behavior of SOA; in particular the dependency on device materials and structure (Connelly, 2002; Piprek, 2017). In this section simple steady-state, time domain and pulse amplification models are used to describe fundamental SOA physics and characteristics.

Steady-State Model

To determine the factors that influence gain, a simple traveling-wave based model, omitting ASE, can be used (Connelly, 2002). The active material gain g_m per unit length at the signal photon energy E_s is taken to be a linear function $g_m = a(n - n_t)$ of the carrier density n at distance z (from the SOA input) and time t , where the differential gain coefficient $a = \partial g_m / \partial n$ at E_s and n_t is the transparency carrier density. n is determined from the rate equation

$$\frac{dn}{dt} = \frac{\eta I}{eV} - \frac{n}{\tau_c} - \frac{\Gamma g_m P}{AE_s} \quad (8)$$

where I is the bias current, the current injection efficiency η is the fraction of the injected current entering the active region. The active region cross-section area and volume are $A = dW$ and $V = AL$ respectively, where L , d , W and are the SOA length and the active waveguide thickness and width respectively. τ_c models interband effects such as radiative and non-radiative spontaneous recombination processes. In practice τ_c reduces as n increases. The signal power P in the SOA is determined from the traveling-wave equation

$$\frac{dP}{dz} = (\Gamma g_m - \alpha)P \quad (9)$$

where α is the loss coefficient. In the steady-state $dn/dt = 0$, so from Eq. (8),

$$n = \frac{P_{sat}}{P + P_{sat}} \left(\frac{\tau_c \eta I}{eV} + n_t \frac{P}{P_{sat}} \right) \quad (10)$$

The saturation power P_{sat} is defined as

$$P_{sat} = \frac{AE_s}{\Gamma a \tau_c} \quad (11)$$

Defining the normalized power $p(z) = P(z)/P_{sat}$, unsaturated carrier density $n_0 = \tau_c \eta I / (eV)$ and unsaturated gain coefficient $g_0 = \Gamma a (n_0 - n_t)$, Eq. (10) can be written as

$$n(z) = \frac{(n_0 + n_t p)}{1 + p} \quad (12)$$

Inserting Eq. (12) into Eq. (9) gives

$$\frac{dp}{dz} = \left(\frac{g_0}{1 + p} - \alpha \right) p \quad (13)$$

The unsaturated gain $G_0 = \exp[(g_0 - \alpha)L]$, which corresponds to a particular value of bias current. Eq. (13) can be solved numerically, using, for example, the Runge-Kutta method. The boundary condition is $p(0) = P_{in}/P_{sat}$, where P_{in} is the input signal power. The amplifier gain G is the ratio of the output power $P_{out} = p(L)P_{sat}$ to the input power. The calculated gain versus output power is shown in Fig. 5(a) for various values of G_0 with $\exp(\alpha L) = 10$. The carrier density is calculated using Eq. (12). As the input signal power is increased the gain reduces as the electrically pumped active material conduction band electrons are depleted by stimulated recombination with valence band holes. The normalized saturation output power $p_{o,sat}$ is defined as the normalized output power at which the amplifier gain is half the unsaturated gain. The saturation output power $P_{o,sat} = p_{o,sat}P_{sat}$. From Fig. 5(a) it can be seen that $p_{o,sat} \approx -3.8$ dB for 20 dB unsaturated gain and is almost independent of G_0 .

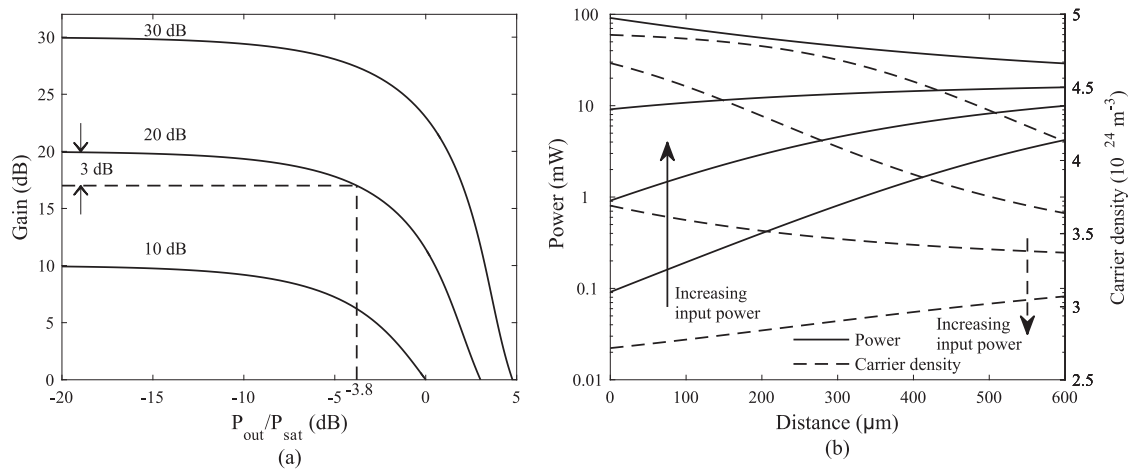


Fig. 5 (a) Gain vs. output normalized power. The parameter is the unsaturated gain. (b) Carrier density and signal distributions for normalized input powers of 0.01, 0.1, 1 and 10. The unsaturated gain is 20 dB.

Consider a 600 μm long square cross-section bulk SOA with w and d equal to 0.4 μm , $\Gamma=0.45$, $\tau_c=0.5$ ns, $a=1 \times 10^{-20}$ m^2 , $n_i=2.5 \times 10^{24}$ m^{-3} and $\exp(\alpha L)=10$. If the unsaturated gain is 20 dB at a current of 150 mA and $\eta=1$, then $n_o=4.9 \times 10^{24}$ m^{-3} , $g_0=1.1 \times 10^4$ m^{-1} and if the signal wavelength is 1550 nm, $P_{\text{sat}}=9.6$ dBm and $P_{o,\text{sat}}=5.8$ dBm. Fig. 5(b) shows the distributions of $P(z)$ and $n(z)$ for various values of $p(0)$. When $P_{\text{in}} \ll P_{\text{sat}}$, n is uniform throughout the amplifier. Models that include ASE show that when the amplifier is unsaturated n has a non-uniform symmetrical distribution with minima at the SOA ends and a maximum at the center. As the saturation level increases the n distribution becomes less symmetric; at high levels of saturation n approaches n_i . A further disadvantage of the model is that the spontaneous recombination term in Eq. (8) is taken to be a linear function of n . In practice spontaneous recombination is more accurately modelled as a polynomial function of carrier density so τ_c is actually n dependent. In power booster applications the most important parameter is $P_{o,\text{sat}}$. The model shows that $P_{o,\text{sat}}$ can be increased by reducing a , τ_c or Γ . One method for increasing $P_{o,\text{sat}}$ is to use a wide dilute waveguide, which has a low Γ ; however disadvantages include increased device length and current. Compared to bulk materials QD materials have significantly higher values of differential gain, allowing even higher values of $P_{o,\text{sat}}$ to be achieved. Modeling of semiconductor nanostructure based SOAs is significantly more complex than for bulk and QW SOAs.

Time-Domain Model and Pattern Effects

SOAs are usually used for modulated signal amplification, so it is of interest to model time domain behavior. Eqs. (8) and (9) can be used to model SOA dynamics having timescales as short as 100 s of picoseconds (Connelly, 2002). First the carrier density is initialized to some suitable value. The carrier density $n(z, t + \Delta t)$ at the next time step is determined from the finite time difference solution of Eq. (8). Assuming that the propagation time of a lightwave through the SOA, which is typically of the order of 1 – 10 ps, is much less than the time variation of the optical input, $P(z, t + \Delta t)$ can be determined from Eq. (9) as

$$P(z, t + \Delta t) = P_{\text{in}}(t + \Delta t) \exp\left(\int_0^z \Gamma g_m(z, t + \Delta t) dz - \alpha z\right) \quad (14)$$

The calculated $P(z, t + \Delta t)$ is then used in Eq. (8) to determine the carrier density at the next time step until the time span of interest is completed. Fig. 6 shows the simulated normalized output power, spatially averaged carrier density and instantaneous gain for an amplified 5 Gb/s Non-Return-to-Zero (NRZ) data stream for an SOA having the same parameters as in the steady-state model above. When the input data is high (logical '1'), the carrier density is depleted and so the gain is reduced. If the gain response time (related to τ_c and the saturation level) is comparable to the inverse of the bit rate then signal distortion will arise within a bit period. If a long string of '1's occurs then the gain will saturate at a constant value. If a '1' is followed by a '0' the gain begins to recover; however the gain experienced by the next '1' may not have recovered to its unsaturated value. Hence the signal distortion is bit pattern dependent. Such pattern effects lead to intersymbol interference and can have a severe impact on system performance.

Because bulk and QD SOAs have relatively large τ_c PDPD will also be severe, which severely impacts on coherent system performance. Because PDPD is a nonlinear effect, compensation is very difficult to achieve using current digital signal processing

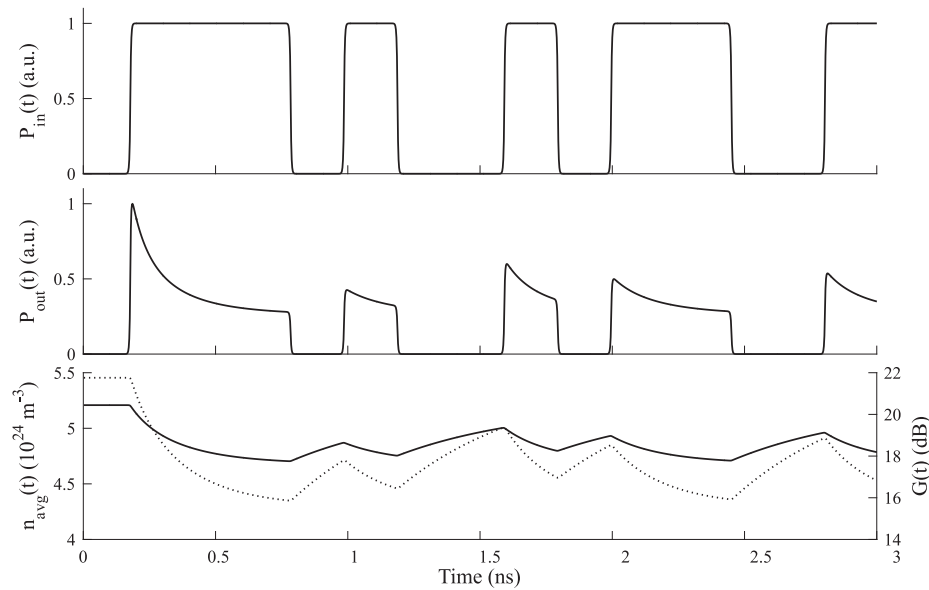


Fig. 6 Input bit stream and simulated SOA output power, spatially averaged carrier density and instantaneous gain for an input NRZ 5 Gb/s data sequence. The input data has high and low level powers of 0.2 mW and 0.2 μW respectively and a rise/fall time of 50 ps. The bias current, unsaturated gain and signal wavelength are 160 mA, 22 dB and 1550 nm respectively.

technology used in digital coherent receivers. QD SOAs have τ_c values significantly smaller than for bulk and QW SOAs so gain recovery is faster, which greatly reduces pattern effects. Pattern effect free amplification with QD SOAs has been demonstrated in Intensity Modulation (IM) systems at bit rates exceeding 160 Gb/s. PDPD is also much lower in QD SOAs, which allows phase transparent amplification of high-order modulation formats at extremely high data rates (Lange *et al.*, 2013).

Two time constants are necessary to describe QDash SOA dynamics: a slow component (~ 100 ps) and a fast component (~ 1 –2 ps). The former is related to carrier transport between individual dashes; the latter to carrier dynamics within a single QDash. At high bit rates (> 10 Gb/s), the carrier transport effect cannot follow signal intensity changes, so the gain response dynamics are similar to QD SOAs.

Pulse Amplification Model

Because SOAs have wide optical bandwidths and fast gain recovery lifetimes, they can be used to amplify optical pulses with Full-Width at Half Maximum (FWHM) pulsewidths in the picosecond range. By exploiting ultrafast SOA nonlinearities in materials such as QDs, it is possible to amplify femtosecond pulses (Sugawara *et al.*, 2004). We use a simple model (Agrawal and Olsson, 1989) applicable to pulsewidths much larger than the interband relaxation time (typically < 0.1 ps in bulk and QW SOAs). Intraband processes, which are not considered in this model, such as carrier heating, spectral hole burning, two-photon absorption and interband refractive index dynamics can be exploited to implement ultrafast optical signal processing functions. If it is assumed that $\alpha \ll g$, where $g = \Gamma g_m$, the output pulse instantaneous power $P_{out}(\tau)$ and phase $\phi_{out}(\tau)$ are given by

$$P_{out}(\tau) = G(\tau)P_{in}(\tau) \quad (15)$$

$$\phi_{out}(\tau) = \phi_{in}(\tau) - \frac{\alpha_H}{2}h(\tau) \quad (16)$$

where τ is the local time with respect to a time frame moving with the pulse, $P_{in}(\tau)$ and $\phi_{in}(\tau)$ the input pulse power and phase respectively. α_H is the LEF, which in general is carrier density and wavelength dependent. The instantaneous SOA gain $G(\tau) = \exp[h(\tau)]$. If the input FWHM pulsewidth $\tau_p \ll \tau_c$, then $h(\tau)$ is given by

$$h(\tau) = -\ln \left\{ 1 - \left(1 - \frac{1}{G_0} \right) \exp \left[-\frac{U_{in}(\tau)}{E_{sat}} \right] \right\} \quad (17)$$

where G_0 is the unsaturated gain, the saturation energy $E_{sat} = E_s W d / (\Gamma a)$ and

$$U_{in}(\tau) = \int_{-\infty}^{\tau} P_{in}(\tau') d\tau' \quad (18)$$

The temporal dependence of $G(\tau)$ induced by the amplified pulse power leads to dynamic changes in the pulse phase, i.e., SPM. The frequency chirp $\Delta v_{out} = (1/2\pi) d\phi_{out}/d\tau$ of the output pulse is given by

$$\Delta v_{out}(\tau) = \Delta v_{in}(\tau) + \frac{\alpha(G_0 - 1)}{4\pi G_0} \frac{P_{out}(\tau)}{E_{sat}} \exp \left[-\frac{U_{in}(\tau)}{E_{sat}} \right] \quad (19)$$

where Δv_{in} is the input pulse chirp. The above closed form equations are applicable to pulsewidths of the order of picoseconds to 10s of picoseconds. To illustrate the effects of an SOA on an amplified pulse consider an unchirped Gaussian input pulse of zero phase and power $P_{in}(\tau) = E_{in}/(\tau_0\sqrt{\pi})\exp[-(\tau/\tau_0)^2]$, where E_{in} is the pulse energy. τ_0 is related to τ_p by $\tau_p \approx 1.665 \tau_0$. In this case $U_{in}(\tau) = (E_{in}/2)[1 + \text{erf}(\tau/\tau_0)]$, where erf is the error function. Simulated amplified pulse power, chirp and power spectrum are shown in Fig. 7. The amplified pulse is asymmetric because the leading edge of the pulse experiences a larger gain than the trailing edge. The amplified pulse is chirped, resulting in a broadening of the pulse spectrum. At high gains the amplified pulse spectrum exhibits a multipeak structure caused by SPM induced frequency chirp. The complex pulse structure and spectral broadening can lead to a significant increase in chromatic dispersion when the pulse is transmitted in an optical fiber and a consequent degradation in system performance. It is possible under appropriate operating conditions for an SOA to act as a pulse narrower or, by inducing an additive chirp whose sign is the opposite of the input pulse chirp, as a dispersion compensator.

Basic Applications

The three basic applications of SOAs are: booster amplifier, in-line amplifier and preamplifier. Systems using relatively inexpensive low-power lasers often need the signal power to be increased prior to transmission. Boosting laser power in an optical transmitter can also be used to overcome modulator losses, compensate for splitting and tap losses in optical distribution networks and to increase the power budget of optical links. The most critical requirement of a booster amplifier is a high $P_{o,sat}$. The NF and polarization sensitivity are of less importance but it is advantageous to keep them as low as possible. Post amplification optical filtering is also not critical because operating the amplifier in saturation greatly reduces the ASE.

Conventional optical transmission systems employ IM of a carrier lightwave. At the receiver end the signal is reconverted to the electrical domain by a photodetector, which is a phase and polarization insensitive process referred to as Direct Detection (DD). If an optical preamplifier is not used, the dominant receiver noise is thermal noise originating from the receiver load resistance and electronic amplifiers. If an optical preamplifier, followed by a narrowband optical filter, is placed in front of the photodetector the

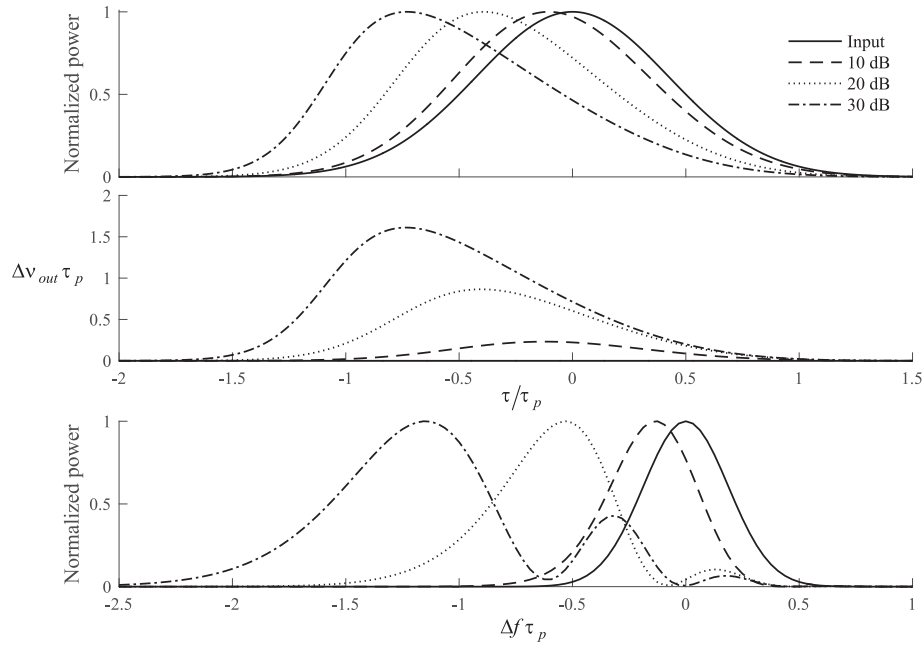


Fig. 7 SOA input and amplified unchirped Gaussian pulse normalized power, chirp and normalized power spectrum. $E_{in}/E_{sat}=0.1$. The parameter is the unsaturated gain and $\alpha_H=5$. Δf is the frequency deviation from the input lightwave center frequency.

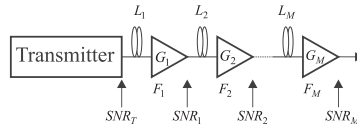


Fig. 8 Optical amplifier chain. The last amplifier in the chain, if immediately followed by a receiver, acts as an optical preamplifier.

dominant noise process is signal-spontaneous beat noise. The power required at the detector is larger than that required if no preamplifier were used; however the power required at the SOA input is lower so the overall receiver sensitivity is improved. It is important the sufficient gain is provided by the SOA such that signal-spontaneous noise dominates; low NF and polarization sensitivity are also important. In practical DD receivers typical sensitivities for a 10^{-9} Bit-Error-Rate (BER) are of the order of a few thousand photons per bit. The use of an SOA preamplifier can result in sensitivity improvements typically in the range of 10–20 dB.

In long-haul optical transmission systems, in-line SOAs can be used to compensate for link losses as shown in **Fig. 8**, where each amplifier compensates for the loss of the preceding fiber link. When several SOAs are cascaded the accumulated ASE will reduce the SNR. If the ASE is high the amplifiers will become saturated, which limits the signal amplification. To avoid excessive noise accumulation it may be necessary to follow each SOA by a narrowband optical filter; however this can limit the use of WDM to increase system capacity. The total noise factor F_{total} of a chain of M amplifiers, defined as the ratio of the transmitter SNR SNR_T , to the output SNR SNR_M of the last amplifier in the chain, is given by

$$F_{total} = \frac{F_1}{L_1} + \frac{F_2}{L_1 G_1 L_2} + \dots + \frac{F_M}{L_M \prod_{k=1}^{M-1} L_k G_k} \quad (20)$$

where L_k is the loss of the k -th fiber span and F_k and G_k are the noise factor and gain of the k -th in-line amplifier respectively. If the amplifiers are identical ($F_k=F, G_k=G$) and exactly compensate for their preceding fiber spans ($L_i=G^{-1}$) then

$$F_{total} = MGF \quad (21)$$

If SNR_T is known then the SNR at the receiver input can be determined using **Eqs. (20)** or **(21)**. In addition to high gain, $P_{o,sat}$ and NF are important in-line amplifier parameters.

Signal distortion can occur even when an SOA is slightly saturated and can significantly increase in severity as the number of amplifiers in the chain increases. A gain-clamped SOA can be used to greatly reduce signal distortion and Cross-Gain Modulation (XGM) induced interchannel crosstalk that occurs when IM channels dynamically saturate the amplifier gain, and is of importance in WDM systems. SOA gain-clamping can be achieved by producing lasing action, at a wavelength remote from the signal wavelength, by introducing wavelength dependent feedback such as placing integrated distributed Bragg reflectors at opposite ends

of the SOA. In the lasing state the SOA carrier density is clamped at a fixed value. Changes in the input signal power lead to opposing changes in the lasing mode power, the effect of which is to keep the carrier density and resulting gain fixed. Gain-clamping can also be achieved by the use of an externally injected Continuous Wave (CW) lightwave: using this technique almost pattern free transmission of a 112 Gb/s four level pulse amplitude modulation data over a 40 km fiber link has been demonstrated (Chan and Way, 2015).

Large increases in system capacity and performance can be achieved by employing multi-level modulation formats such as PSK and QAM. The reception of multi-level optical signals requires coherent detection, whereby the signal is combined with a Local Oscillator (LO) laser and simultaneously detected; the resulting electrical signals are then processed to recover the received lightwave phase and amplitude. Coherent systems are now common in backbone networks. A major advantage of coherent detection is that the LO power can be increased to a level such that receiver thermal noise is negligible, in which case receiver sensitivity is determined by the SNR γ_s of the received symbols. In contrast to DD receivers, coherent receivers are linear so post-detection signal processing can be used to compensate for local and transmission link impairments. In the quantum noise limited regime, encountered in links not employing optical amplifiers, $\gamma_s = \eta n_s$ where η is the quantum efficiency of the receiver photodetectors and n_s is the average number of photons per received symbol. When in-line optical amplifiers are used to periodically compensate for the transmission fiber loss and where the last amplifier acts as a preamplifier, as shown in Fig. 8, the signal-spontaneous beat noise dominates in which case $\gamma_s = 2n_s/MF$. If NF is equal to the minimum possible value of 3 dB then $\gamma_s = n_s/M$.

Functional Applications

SOAs can be used to implement many optical signal processing functions especially important at ultra-fast data rates where real-time electronic processing is not possible. In this section the important SOA ultra-fast signal applications of wavelength conversion, optical logic and regeneration are described but also the low speed applications of optical remodulation and microwave phase shifting, all of which exploit various SOA nonlinearities.

SOA FWM Wavelength Converter

Wavelength conversion whereby data is transferred from one carrier wavelength to another is an important function required in WDM networks. Non-degenerate Four-Wave Mixing (FWM) in SOAs is one of the most promising wavelength conversion techniques as it has the advantages of ultrafast operation and preserves the input data signal amplitude and phase modulation and as such can be used in IM and coherent communication systems. FWM is a coherent third-order nonlinear effect that can occur between two or more co-propagating co-polarized optical fields (Agrawal, 1988). If there are two fields, a high power pump at frequency ν_0 and a lower power probe at frequency $\nu_0 - \Delta\nu$, where $\Delta\nu$ is the pump-probe detuning, a refractive index modulation at $\Delta\nu$ will result that leads to the creation of a conjugate field at frequency $\nu_0 + \Delta\nu$. In the case of a CW pump and modulated signal the conjugate is identical to the signal, albeit with a different power level, as shown in Fig. 9 and can be selected using a bandpass filter at the SOA output. FWM in SOAs arises from different physical phenomena. At low $|\Delta\nu| (< 1/\tau_c)$ the refractive index modulation is caused by modulation of the carrier density and is only significant for $|\Delta\nu|$ less than a few 10s GHz. At higher frequencies, two other effects dominate: Spectral Hole Burning (SHB) and Carrier Heating (CH). SHB is caused by the pump creating a hole in the intraband carrier distribution, which modulates the occupation probability of carriers within a band leading to fast gain and refractive index modulation. CH is caused by stimulated emission and free carrier absorption. Stimulated emission subtracts carriers that are cooler than average while free carrier absorption moves carriers to higher energy levels in the band. The resulting increase in temperature decreases the gain. Both SHB and CH have characteristic times in the hundreds of femtosecond range, so FWM based wavelength converters can operate with $|\Delta\nu|$ in the 100s GHz range and at ultrahigh bit rates (100s Gb/s). FWM also allows the simultaneous conversion of multiple input wavelengths to another set of multiple output wavelengths which is useful in multicasting applications. The conversion efficiency η , defined as the ratio of the SOA output converted signal power to

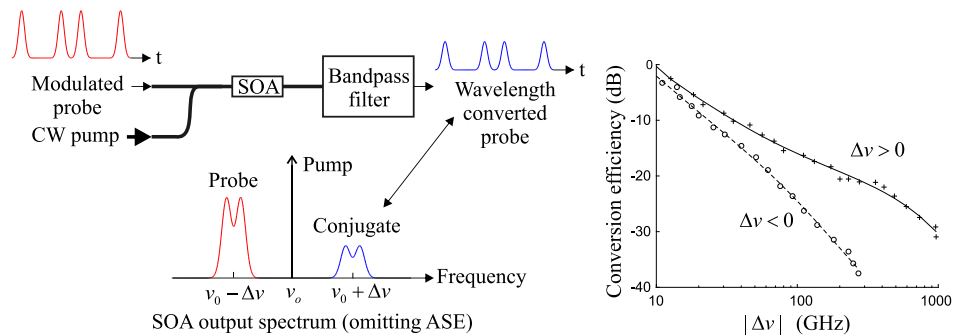


Fig. 9 SOA FWM wavelength converter and typical conversion efficiency vs. detuning frequency characteristic.

the input signal power, depends on a number of factors including the SOA gain, input pump and probe powers, $\Delta\nu$ and the time constants associated with the physical processes described above. A typical experimental plot of η versus $\Delta\nu$ is shown in Fig. 9.

Optical Logic

Optical logic gates have important applications such as addressing, header recognition, data encoding and encryption. Many different schemes for implementing different logic operations have been investigated, including several based on the exploitation of SOA nonlinearities such as XGM and FWM (Stubkjaer, 2000). Interferometric configurations containing SOAs are useful for implementing fast optical switching and logic functions. The most commonly used structure is the SOA Mach-Zehnder Interferometer (MZI), which for stable operation must be integrated. An SOA MZI XOR gate and associated truth table are shown in Fig. 10. The XOR gate is of special interest since it is the basic building block for a wide range of more complex logic functions. The data signals (A and B) input into the two SOAs modulates the SOA carrier densities leading to refractive index modulations and thereby XGM and Cross-Phase Modulation (XPM) of the co-propagating input CW light, having a different wavelength to the data signals. The input data powers and SOA operating current are chosen such that the CW light experiences the same phase shift in each SOA when A and B have the same power level (identical logic levels) leading to destructive interference (logic '0') at the interferometer output. Conversely when A and B have different logic levels the relative phase shift is π , which results in the CW light experiencing constructive interference (logic '1'). SOA MZIs often have tunable phase shifting elements placed in one or both of the interferometer branches to allow fine tuning and optimization of the switching process. Other logic operations can be implemented by choosing suitable input data power levels and SOA-MZI operating conditions. SOA interferometric logic gates have the advantage of having very steep transfer functions so the speed of operation is not limited by the slow carrier recovery times. Only low input signal power levels are needed to introduce a phase difference between the interferometer arms, so that efficient logic operation conversion is obtained almost independently of wavelength. Also because the carrier density modulation is small, the frequency chirp imposed on the output signal is not large. Bulk or QW SOA MZI logic gates can operate at speeds in the 10s Gb/s range. Much higher speeds (> 250 Gb/s) can be achieved using QD SOAs. SOA-MZIs can also be used as wavelength converters but only for IM signals.

All-Optical Data Regeneration

Regeneration of a data signal is needed if the signal quality has been adversely impacted by noise or other impairments and involves one or more of three functions: amplification, reshaping (of the signal amplitude and/or phase) and retiming. Conventional SOA based regenerators for IM signals use bulk or QW SOA MZI switching structures. However, due to its faster gain dynamics a single QD SOA can be used to perform some regeneration functions such as noise compression at ultrafast bit rates (> 40 Gb/s) as shown in Fig. 11 (Akiyama *et al.*, 2007). A noisy input NRZ data stream has an average binary '1' power level that induces gain saturation in the SOA. If the gain dynamics are much faster than the signal transitions the SOA effectively responds instantaneously to the input data thereby compressing the '1' level noise. This scheme does not reduce the '0' level noise; however it can be reduced by following the SOA with a saturable absorber.

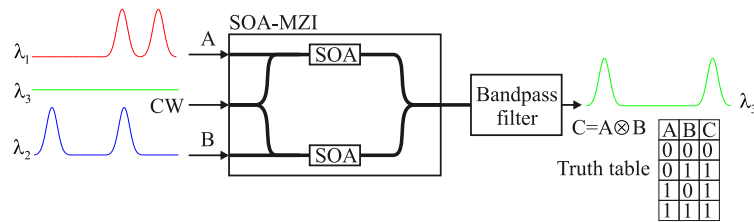


Fig. 10 SOA MZI XOR logic and truth table.

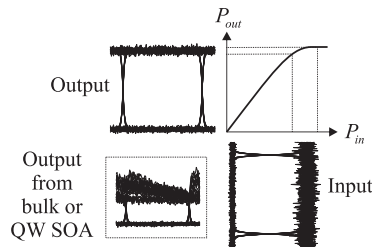


Fig. 11 Eye diagrams illustrating QD SOA noise compression at high bit rates. If a bulk or QW SOA is used, pattern effects will be present due to the much slower gain recovery time.

Regeneration of amplitude/phase modulated signals is significantly more challenging and requires the use of a nonlinear optical loop mirror or Phase Sensitive Amplifier (PSA), which precludes the direct use of phase insensitive SOAs. A PSA is capable of amplifying only one of the two quadrature components of a lightwave signal and attenuating the other. In theory a PSA can realize noise free amplification ($NF=0$); however most implementations are based on complex bulk crystal or fiber systems. Recently a monolithically integrated PSA based on a dual-pump degenerate FWM process that uses an SOA as the nonlinear element has been demonstrated (Li *et al.*, 2016), which shows great promise for use in coherent optical communications.

Reflective SOA based Passive Optical Network

There is an increasing effort to develop cost-effective Fiber-To-The-Home (FTTH) networks. WDM Passive Optical Networks (PONs) are one of the most promising solutions for realizing next generation FTTH. For successful deployment of WDM-PONs the Optical Network Unit (ONU) located in the subscriber premises needs to be wavelength independent as well as being cost effective. One such WDM-PON which uses a carrier remodulation scheme based on the saturation properties of a Reflective SOA (RSOA) is shown in Fig. 12 (Lee *et al.*, 2005). An RSOA is a TW SOA with a highly-reflective coating applied to one of the end facets. RSOAs have similar optical bandwidths to TW-SOAs but have a steeper gain vs. input optical power characteristic and are therefore easier to saturate. In the WDM-PON Central Office (CO), low extinction ratio (ER) NRZ downstream signals are generated by direct modulation of lasers and transmitted through an Arrayed Waveguide Grating (AWG) wavelength multiplexer, the fiber link and at the Remote Node (RN) separated by an AWG wavelength demultiplexer. At the ONU the downstream channel is split by a coupler; one output of which is sent to a receiver for conventional detection and processing, the other output is sent to an RSOA. The power level of the RSOA input is chosen such that the amplifier operates in saturation, thereby suppressing the downstream modulation component. The upstream channel (RSOA output) is generated by modulating the RSOA current with NRZ data. The downstream channel data rate is higher than that of the upstream. At the upstream receiver located in the CO, the unsuppressed component of the downstream signal, which would lead to a power penalty, is filtered out by an electrical lowpass filter. Error-free bidirectional transmission with 1.25 Gb/s downstream data rates was demonstrated for 20 km transmission distance.

SOA Microwave Phase Shifter

Microwave Photonics (MWP) is concerned with the generation, transport and processing of radio frequency, microwave and millimeter wave signals in the optical domain (Xue *et al.*, 2010). Wideband tunable microwave delays and phase shifters are a key element required in MWP, with applications in microwave filtering and optically fed phased array antennas, however their implementation in the microwave domain is very difficult. One of the most promising photonic based techniques uses Coherent Population Oscillations (CPOs) in SOAs, which allows tunability based on the SOA current or input optical power. CPOs are induced by injecting a sinusoidally modulated lightwave at frequency f_m as shown in Fig. 13. Therefore the input optical signal is a

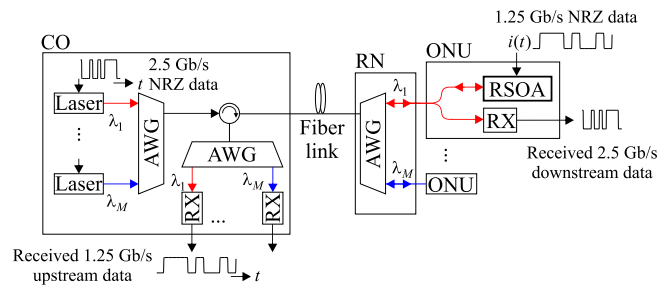


Fig. 12 WDM-PON topology. RX, receiver. Modified from Lee, W., Park, M.Y., Cho, S.H., *et al.*, 2005. Bidirectional WDM-PON based on gain-saturated reflective semiconductor optical amplifiers. IEEE Photonics Technology Letters 17, 2460–2462.

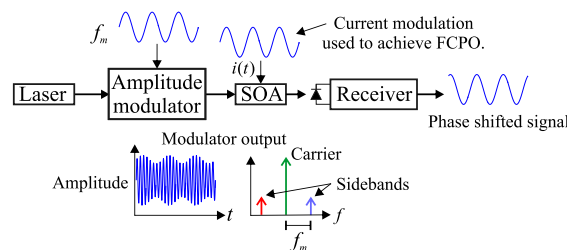


Fig. 13 SOA MWP phase shifter.

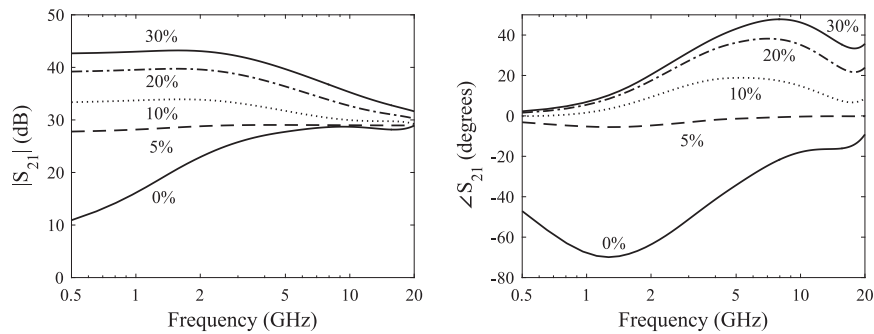


Fig. 14 Typical bulk SOA phase shifter transfer functions for different values of the current modulation index. The unsaturated gain at the lightwave wavelength of 1550 nm is 25 dB at the bias current of 175 mA. The average input optical power is -5 dBm, at which the SOA gain is 14 dB, with an optical modulation index of 10%.

double-sideband unsuppressed carrier amplitude modulated signal having a strong carrier and two relatively weak sidebands. This leads to the carrier density being modulated at the beat (modulation) frequency between the sidebands and carrier, which induces a frequency dependent change in the group index and therefore the group velocities of the sidebands. After propagation through the SOA the envelope of the optical signal will be delayed or advanced (slow or fast light). Detection of the output optical signal results in an output microwave signal that is phase shifted with respect to the modulator input. At modulation frequencies less than the inverse of the carrier lifetime, the beat-signal gain is low leading to a poor signal-to-noise ratio. The low frequency response can be enhanced by using Forced CPO (FCPO), which involves simultaneously modulating the input optical power and SOA current as shown in Fig. 13. The SOA phase shifter transfer function $S_{21}(f)$ is defined as the ratio of the SOA optical output to input detected beat signal currents. $|S_{21}(f)|$ and $\angle S_{21}(f)$ are the phase shifter signal gain and phase shift respectively; typical frequency plots of which for CPO (no current modulation) and FCPO operation are shown in Fig. 14 for various values of the current modulation index. Adjustment of the current modulation index allows control of the level of advancement and to switch from fast to slow light. Further tunability can be obtained by adjusting the bias current and optical power.

Conclusions

SOAs have many applications in optical networks especially as optical signal processing devices. Although SOAs have been investigated for many years and are commercially available, usually as single devices fabricated from bulk or QW material, the development of new SOA structures, advanced materials such as QDs, PICs and exploitation of ultrafast non-linearities make SOA research and development an exciting field of study. The commercial deployment of ultrafast coherent optical communication systems requires new optical signal processing functions such as full regeneration of amplitude/phase modulated signals, and doubtless new SOA based technologies will be developed to meet this need.

See also: Erbium Doped Fiber Amplifiers for Lightwave Systems

References

- Agrawal, A.P., 1988. Population pulsations and nondegenerate four-wave mixing in semiconductor lasers and amplifiers. *Journal of the Optical Society of America B* 5, 147–159.
- Agrawal, G.P., Olsson, N.A., 1989. Self phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. *IEEE Journal of Quantum Electronics* 25, 2297–2306.
- Akiyama, T., Sugawara, M., Arakawa, Y., 2007. Quantum-dot semiconductor optical amplifiers. *Proceedings of the IEEE* 95, 1757–1766.
- Chan, T.K., Way, W.I., 2015. 112 Gb/s PAM4 transmission over 40 km SSMF using 1.3 μm gain-clamped semiconductor optical amplifier. In: 2015 Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, pp. 1–3.
- Chuang, S.L., 2009. *Physics of Optoelectronic Devices*. New York, NY: Wiley.
- Connelly, M.J., 2002. *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic.
- Dutta, N.K., Wang, Q., 2013. *Semiconductor Optical Amplifiers*. Singapore: World scientific.
- Lange, S., Contestabile, G., Yoshida, Y., Kitayama, K.I., 2013. Phase-transparent amplification of 16 QAM signals in a QD-SOA. *IEEE Photonics Technology Letters* 25, 2486–2489.
- Lee, W., Park, M.Y., Cho, S.H., *et al.*, 2005. Bidirectional WDM-PON based on gain-saturated reflective semiconductor optical amplifiers. *IEEE Photonics Technology Letters* 17, 2460–2462.
- Lelarge, F., Dagens, B., Renaudier, J., *et al.*, 2007. Recent advances on InAs/InP quantum dash based semiconductor lasers and optical amplifiers operating at 1.55 μm . *IEEE Journal of Selected Topics in Quantum Electronics* 13, 111–124.

- Li, W., Lu, M., Mecozzi, A., *et al.*, 2016. First monolithically integrated dual-pumped phase-sensitive amplifier chip based on a saturated semiconductor optical amplifier. *IEEE Journal of Quantum Electronics* 52, 1–12.
- Olsson, N.A., 1989. Lightwave systems with optical amplifiers. *Journal of Lightwave Technol* 7, 1071–1082.
- Piprek, J., 2017. *Handbook of Optoelectronic Device Modeling and Simulation: Fundamentals, Materials, Nanostructures, LEDs, and Amplifiers*. vol. 1. CRC Press.
- Stubkjaer, K.E., 2000. Semiconductor optical amplifier-based all-optical gates for high-speed optical processing. *IEEE Journal of Selected topics in Quantum Electronics* 6, 1428–1435.
- Sugawara, M., Ebe, H., Hatori, N., *et al.*, 2004. Theory of optical signal amplification and processing by quantum-dot semiconductor optical amplifiers. *Physical Review B* 69, 235332.
- Xue, W., Sales, S., Capmany, J., Mørk, J., 2010. Wideband 360° microwave photonic phase shifter based on slow light in semiconductor optical amplifiers. *Optics Express* 18, 6156–6163.

Wavelength Division Multiplexing

Klaus Grobe, ADVA Optical Networking SE, Martinsried, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction to WDM

Wavelength Division Multiplexing (WDM) is a multiplexing and transmission scheme in fiber-optical telecommunications where different wavelengths, emitted by several lasers, each carry dedicated information. These wavelengths are multiplexed by means of WDM filters. Likewise, they are separated, or demultiplexed, in the receive end by means of similar filters, or coherent detection using tunable local oscillators.

As a multiple-access mechanism, WDM allows separating different customers' traffic in the wavelength domain. This is called Wavelength Division Multiple Access (WDMA). WDM can be combined with any other multiplexing or multiple-access schemes, namely electrical Time Division Multiplexing (TDM, TDMA), Sub-Carrier Multiplexing (SCM, SCMA), and Orthogonal Frequency Division Multiplexing (OFDM, OFDMA, including the real-valued variant Discrete Multi-Tone).

WDM splits into high-performance Dense WDM (DWDM) with high channel count (> 100) and channel spacing down to 12.5 GHz, and lower-cost Coarse WDM (CWDM) with up to 18 channels spaced at 20 nm. CWDM is used in access and backhaul, DWDM is used in all high-capacity and long-haul transport applications. Driven by improvements of components and modulation and equalization techniques, the total DWDM capacity has largely increased since the first experiments in the late 1980s, see Fig. 1.

DWDM systems have approached an area of slowed-down capacity improvement. Over the next years, DWDM on Standard Single-Mode Fibers (SSMF) will finally reach the Nonlinear Shannon Limit (Essiambre *et al.*, 2010). Further progress beyond this limit will possibly require new fiber types.

Overviews on WDM can be found, e.g., in Grobe and Eiselt (2014), Agrawal (1992).

Transmission Effects in Optical Fibers

WDM transmission depends on the fiber type that is used. SSMF show frequency dependence, time variance, and nonlinear behavior. The resulting transmission impairments are.

- Linear effects
 - a. Attenuation
 - b. Chromatic Dispersion (CD)
 - c. Polarization-Mode Dispersion (PMD)
 - d. Polarization-Dependent Loss (PDL)
- Nonlinear effects
 - a. Kerr effects
 - b. Scattering effects (nonlinear, stimulated)

Detailed discussions of these effects can be found in the literature, e.g., Agrawal (1992), (1995). For WDM long-haul transmission, all these effects and their interactions have to be considered.

Linear Effects

Attenuation

Attenuation in Silica single-mode fibers is mainly caused by intrinsic Silica-glass loss (Rayleigh scattering, Infrared absorption), extrinsic loss due to impurities and bending loss.

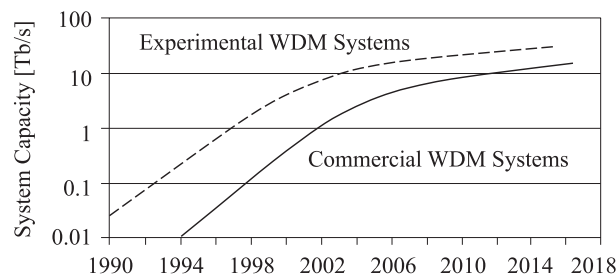


Fig. 1 Development of WDM Systems Transport Capacity over time.

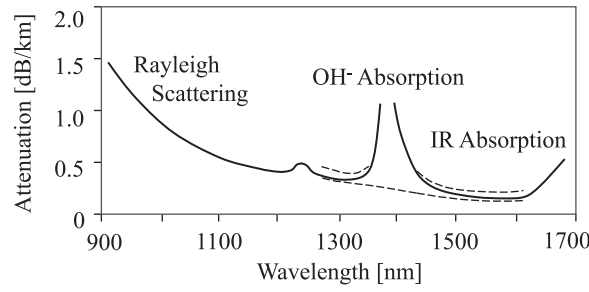


Fig. 2 Spectral loss in single-mode fibers.

The spectral loss caused by **intrinsic effects** and **impurity** is shown in **Fig. 2**. The attenuation peak around 1400 nm is caused by OH^- ions. In fibers vintages up into the 1980s, this peak was very pronounced, in new fibers it has almost been eliminated. The dashed lines in **Fig. 2** indicate the loss tolerance as specified in ITU-T Recommendation G.695 (ITU-T, 2015) for SSMF.

Fibers exhibit loss caused by perturbations of the ideal waveguide geometry. Such perturbations can result from fiber bends, where bending radii with $R \gg \lambda$ (macro-bending) and $R \approx \lambda$ (micro-bending) with λ the optical wavelength have to be differentiated.

Macro-bending loss occurs when fiber-optic cables are bent, e.g., in patch frames etc. It increases exponentially with increasing wavelength and decreasing bending radii. For wavelengths < 1300 nm and bending radii > 20 mm, macro-bending loss can be neglected. WDM wavelengths around 1550 nm can already be affected significantly. The region > 1600 nm is sensitive to macro bending.

Micro-bending loss is caused when the fiber is subject to radial pressure. This happens when a fiber (within a cable) is subject to mechanical pressure. This forces deformations of the fiber's core-cladding boundary which have similar dimension as the wavelengths. This causes destructive light interference and consequently loss.

Chromatic dispersion

Chromatic Dispersion (CD) causes signal distortions via the dependence of the velocity of propagation on the frequency of the respective spectral components. Since modulation always leads to spectral broadening of carrier waves, every information transmission in fibers is subject to CD. The resulting temporal spread of the signal leads to Inter-Symbol Interference (ISI).

In the absence of nonlinearity and birefringence, and with constant attenuation α , propagation of a light wave with envelope E is given by:

$$E(z, t) = E_0 \exp(-\alpha z) \exp(j\omega t - j\beta(\omega)z) \quad (1)$$

The phase constant $\beta(\omega)$ is usually developed into a Taylor series around a mean carrier frequency ω_0 (corresponding to, e.g., 1550 nm). From the third Taylor coefficient, β_2 , the so-called Dispersion Parameter is derived (Agrawal, 1992, 1995):

$$D = -\omega^2 \beta_2 / 2\pi c_0 \quad (2)$$

D has the dimension (ps/(nm · km)). It describes the temporal spread (in ps) of signal pulses with a certain optical bandwidth (in nm) over a certain transmission distance (in km). The other coefficients are related to *phase velocity* ($v_p = \omega_0 / \beta_0 = c_0 / n$), *group velocity* ($v_g = 1 / \beta_1$), and *dispersion slope* (β_3). More coefficients are typically not considered.

Fig. 3 shows D parameters for single-mode fibers according to ITU-T Recommendations G.652 (ITU-T, 2016d), G.653 (ITU-T, 2010b) and certain G.655 (ITU-T, 2009b) fiber brands (OFS, 2017; Corning, 2014; Ohsono *et al.*, 2003; Prysmian Group, 2010; Corning, 1998).

The D parameter can be used to define a transfer function which considers CD:

$$H_{CD}(z, j\omega) = \exp \left[j \left(\beta_0 + \beta_1(\omega - \omega_0) - \frac{1}{2} \frac{\lambda_0^2}{2\pi c_0} D(\omega - \omega_0)^2 \right) z \right] \quad (3)$$

The transfer function $H_{CD}(z, j\omega)$, complemented by a simple attenuation term, can be used for calculations or simulations of a linear fiber model (neglecting nonlinearity).

Polarization mode dispersion

Polarization Mode Dispersion (PMD) is caused by differences of the effective refractive index of the transmission fiber between (any) two orthogonal polarizations. This leads to birefringence and is caused by fiber geometry perturbation or lateral stress on the fiber. These cannot be perfectly avoided. While in a perfectly circular fiber no particular pair of orthogonal polarization modes is distinguished, birefringence leads to two particular modes developing, the *Principal States of Polarization* (PSP). They depend on wavelength. The PSPs are defined at the input to the fiber as those states of polarization, for which, when slightly varying the signal frequency, the output polarization remains constant. The time delay between both PSPs is called *Differential Group Delay* (DGD).

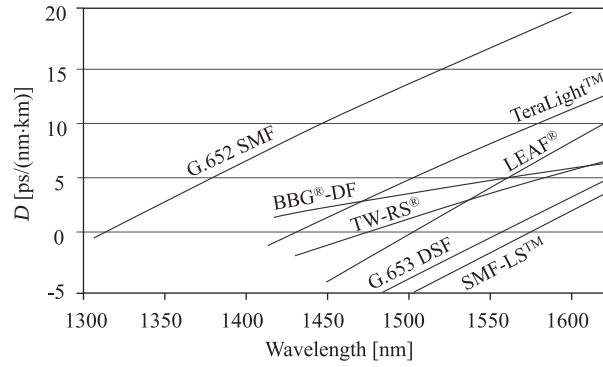


Fig. 3 D parameters of various single-mode fibers.

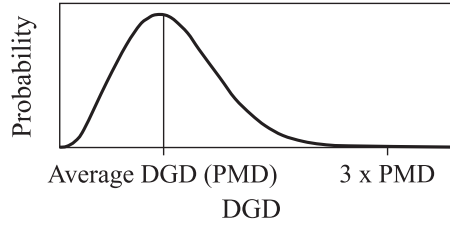


Fig. 4 Distribution function for DGD.

PMD results in pulse spread ΔT^{PMD} , depending on fiber length L and PMD Parameter D_{PMD} :

$$\Delta T^{\text{PMD}} = D_{\text{PMD}} \sqrt{L} \quad (4)$$

Long fibers can be modeled as cascades of k birefringent sections with random polarization orientations each. The resulting total PMD then yields:

$$\Delta T_{\text{tot}}^{\text{PMD}} = \sqrt{\sum_k (\Delta T_k^{\text{PMD}})^2} \quad (5)$$

The PMD parameter D_{PMD} has the dimension (ps/ $\sqrt{\text{km}}$). Typical values are listed in [Table 3](#).

The PMD parameter is an average over time or wavelength. The actual pulse broadening is related to the instantaneous DGD and depends on the actual State of Polarization of the signal relative to the PSPs. The pulse broadening is therefore described by a stochastic process. The DGD has a Maxwellian distribution, as shown in [Fig. 4](#).

The *average* DGD is referred to as PMD. As most transmission systems are impacted by the instantaneous DGD, a DGD tolerance is specified. Since $\text{DGD} > 3 \cdot \text{PMD}$ only occurs with a probability of $4 \cdot 10^{-5}$ (~ 21 min/year), the specified PMD tolerance is usually three times smaller than the DGD tolerance to guarantee low outage probability ([Bülow, 1998](#)).

Nonlinear Fiber Effects

Silica fibers exhibit weak nonlinearity. This effect becomes apparent at high power levels and long-distance transport without regeneration. The latter is achieved in all WDM long-haul transport, which is ultimately limited by this nonlinearity ([Essiambre et al., 2010](#)).

Nonlinear fiber effects split into two classes, instantaneous Kerr effects and (stimulated, i.e., laser-like) scattering effects with specific gain spectra.

The **Kerr effects** can be classified into inter- (WDM-) channel and intra-channel effects, and into signal-signal or signal-noise interactions, see [Fig. 5](#) ([Winzer and Essiambre, 2006](#)).

The excitation of the individual Kerr effects, i.e., their relative relevance, depends on power level, fiber type (especially CD) and per-channel bit rate, as shown in [Fig. 6](#) ([Winzer and Essiambre, 2006](#)). For low CD, FWM (i.e., generation of new spectral components via parametric mixing) is dominant, and for very high Baud rates, the intra-channel effects dominate.

The propagation constant β is proportional to the effective refractive index n_{eff} seen by the electrical field, $\beta = 2\pi \cdot n_{\text{eff}}/\lambda$. The effective refractive index becomes dependent on the instantaneous power, causing nonlinearity:

$$n_{\text{eff}} = n_0(\lambda) + n_2 P / A_{\text{eff}} \quad (6)$$

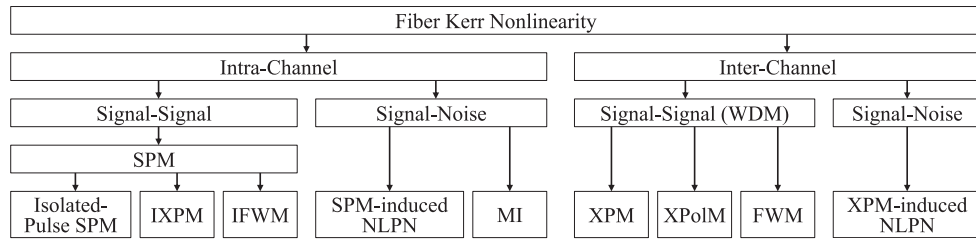


Fig. 5 Kerr-effect classification. MI, Modulation Instability; NLPN, Nonlinear Phase Noise; SPM, Self-Phase Modulation. XPoIM: Cross-Polarization Modulation. (I)XPM: (Intra-channel) Cross-Phase Modulation. (I)FWM: (Intra-channel) Four-Wave Mixing.

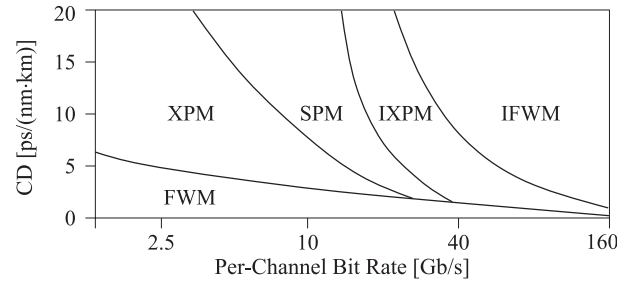


Fig. 6 Dominance of nonlinear effects.

Table 1 Values of nonlinearity for various types of transmission fiber

Fiber type	A_{eff} typ.	γ typ. at 1550 nm
SSMF G.652	80 μm^2	1.3/(W · km)
TrueWave [®] G.655 (OFS, 2017), DSF G.653	55 μm^2	1.8/(W · km)
LEAF [®] G.655 (Corning, 2014)	72 μm^2	1.4/(W · km)

Here, n_2 is the nonlinear fiber coefficient, A_{eff} is the effective area of the fiber, which is the average area occupied by the field in the fiber, and P is the instantaneous signal power. A nonlinearity coefficient γ is often used to describe the nonlinear properties of the fiber. It is given by $\gamma = 2\pi n_2 / \lambda A_{\text{eff}}$. Values for different fiber types are listed in **Table 1**.

The Kerr effects can be calculated numerically using a nonlinear propagation equation, the so-called *Nonlinear Schrödinger Equation* (NLSE) (Agrawal, 1995):

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \frac{j}{2}\beta_2 \frac{\partial^2 A}{\partial t^2} + j\gamma|A|^2A \quad (7)$$

A is the complex envelope of the propagating field. The NLSE (7) does not consider scattering effects, and is limited in accuracy for ultra-broadband signals. A more general description has to be based on Wiener-Volterra series (Schetzen, 1980). It is possible to extend Eq. (7) by higher-order CD and a Raman shock term (Agrawal, 1995), resulting in the *Generalized Nonlinear Schrödinger Equation* (here, without attenuation term):

$$\frac{\partial A}{\partial z} = \sum_{i=1}^3 \beta_i \frac{(-j)^{i+1}}{i!} \frac{\partial^i A}{\partial t^i} - \frac{j\beta_0 n_2}{n_0} \left(1 + \frac{j}{\omega_0} \frac{\partial}{\partial t}\right) \left(A \int_0^\infty R(\tau) |A(z, t - \tau)|^2 d\tau\right) \quad (8)$$

Eq. (7) can efficiently be solved numerically with the Split-Step Fourier Algorithm (Agrawal, 1995). The GNLSE (8) has to be solved with more CPU-demanding pseudo-spectral methods.

Raman Scattering is scattering of photons at glass molecules with certain nuclear and vibrational states. These states become dependent on the incoming intensity, causing nonlinearity. The molecular states cover a certain spectrum. Due to the amorphous structure of Silica glass, *harmonic broadening* occurs, and the scattered photons are down-shifted in frequency by continuous scattering spectra. During scattering, the system molecule-plus-photon is in a virtual, i.e., forbidden, energy state. Hence, scattering takes place quickly, within less than 1 ps (Agrawal, 1995).

For strong Stimulated Raman Scattering (SRS), amplification of signals of incident power $P_S(0)$ over length L is given by:

$$P_S(L) = P_S(0) \times \exp(G_R(\Delta f) I_0 L) \quad (9)$$

I_0 is the intensity of the optical pump, and $G_R(\Delta f)$ is the Raman gain as shown in **Fig. 7** (Agrawal, 1995; Chraplyvy, 1990).

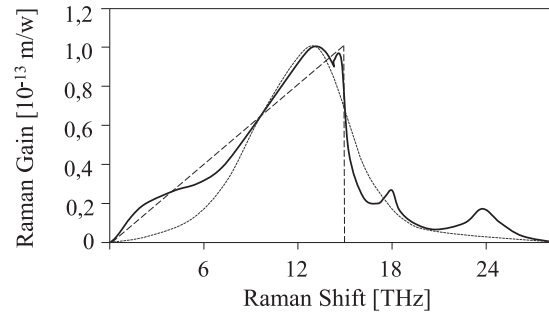


Fig. 7 Raman gain spectrum. Triangle (dashed) and single-Lorentzian (dotted) approximations were sometimes used for numerical approximations.

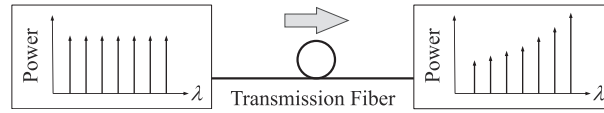


Fig. 8 Gain tilt in WDM channels due to SRS.

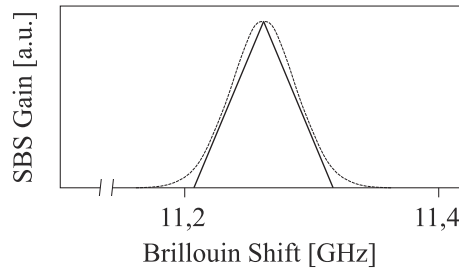


Fig. 9 Brillouin gain of a G.653 fiber with almost triangular shape according to Agrawal (1995), and approximation with a single Lorentzian (dotted).

In WDM systems, higher-frequency channels pump lower-frequency channels through SRS. This leads to power loss of higher-frequency channels and gain tilt, as shown in Fig. 8. In addition, SRS leads to crosstalk in the lower-frequency channels following the (intensity) bit patterns in the pump channels.

If the Raman-induced loss for the pump channels has to be < 1 dB, the product of total bandwidth and total power of all channels has to be < 50 GHz · W (Chraplyvy, 1990).

Brillouin Scattering is inelastic scattering caused by spatial refractive-index fluctuations which are generated by density fluctuations. The latter are caused by hypersonic sound waves or by electrical fields (electrostriction). Similar to SRS, the effect can be stimulated, leading to Stimulated Brillouin Scattering (SBS).

SBS has the lowest power threshold of all nonlinear fiber effects and can limit WDM systems in their maximum per-channel power. The maximum SBS gain is $G_B(\Delta f_B) \approx 4 \cdot 10^{-11}$ m/W and is achieved at a frequency shift of $\Delta f_B \approx 11$ GHz. The SRS gain bandwidth for optical pumps around 1550 nm is limited to 20–100 MHz.

Similar to SRS, amplification of incident power $P_S(0)$ by SBS is given by

$$P_S(L) = P_S(0) \times \exp(G_B(\Delta f) I_0 L) \quad (10)$$

I_0 is the intensity of the optical pump, L is the fiber length, and $G_B(\Delta f)$ is the frequency-dependent SBS gain. An SBS gain example is shown in Fig. 9.

SRS causes *backscattering* which can be blocked by optical isolators. However, due to the low SBS threshold, per-channel launch power must be < 10 dBm in order to avoid significant reflection and power loss (Chraplyvy, 1990). Due to the narrow-band gain characteristics, SBS is also reduced by widening the laser line widths of WDM channels. This can be achieved through additional slow amplitude modulation (so-called dither (Fishman and Nagel, 1993)).

Components and Sub-Systems

End-to-end WDM transmission consists of transmitters, transmission channel, and receivers. The WDM transmission channel always comprises SMFs. WDM transmitters and receivers consist of several functions, see Fig. 10.

Necessary functions in the transmit/receive end include channel coding/decoding (e.g., the application of Forward Error Correction), modulation/demodulation, and WDM and optional polarization multiplexing/demultiplexing. Optional functions include amplification, encryption/decryption, and time-domain multiplexing/demultiplexing.

WDM Transmitters

Laser diodes

WDM systems use (tunable) Laser Diodes (LD) as optical transmitters. LD output spectra depend on the optical-cavity geometry and the gain-medium wavelength dependence. Semiconductor materials for wavelengths of 0.5 μm to 4 μm are shown in Fig. 11 (Agrawal and Dutta, 1986).

LD direct current modulation causes intensity and parasitic frequency modulation. The latter is called (frequency) chirp. To avoid chirp, *external modulators* have to be used.

Tunable lasers are required for network-capacity optimization and remote reconfiguration, together with Reconfigurable Optical Add/Drop Multiplexers (ROADM). They also allow protection schemes based on wavelength switching. Tunable lasers further reduce the number of different transmitter variants, thus reducing device and inventory cost.

Mode number m and wavelength λ of a LD depend on the effective refractive index n and effective cavity length L , $m\lambda/2 = nL$. Consequently, LD tunability can be achieved by tuning any of these parameters

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta n}{n} + \frac{\Delta L}{L} - \frac{\Delta m}{m} \quad (11)$$

On the first two terms on the right-hand side, electronic, thermal or mechanical tuning can be applied (Coldren *et al.*, 2004). Electronic tuning refers to the application of an electric field which tunes the laser frequency. Several implementations exist, which are basically derivatives of the three-section Distributed-Bragg-Reflector (DBR) LD.

Thermal tuning can be done with certain Distributed-Feedback (DFB) LD. Mechanical tuning can be applied to Micro-Electromechanical Vertical Cavity Surface Emitting Lasers (MEM-VCSEL) or External Cavity Lasers. Electronic and mechanical implementations enable fast tuning speeds, thermal tuning is somewhat slower.

External modulators

External modulators are used for high-speed WDM transmission. They allow broad modulation bandwidth, (partly) avoid the chirp resulting from direct laser modulation, and can enable complex (I + Q) modulation.

Electro-Absorption Modulators (EAM) are similar to reverse-biased p-i-n diodes with a bulk active region or multiple quantum-wells (MQWs) as absorption layer. They can have low power consumption and low cost. However, they are mainly restricted to intensity modulation.

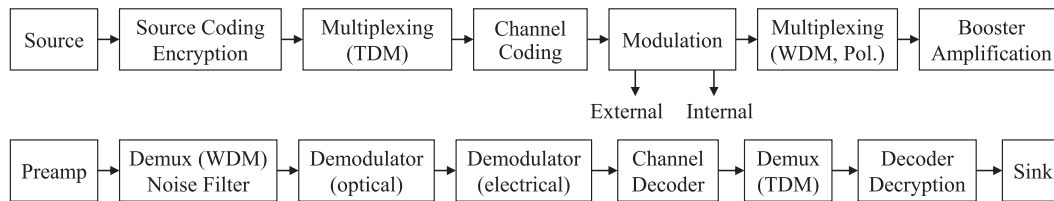


Fig. 10 Transmit-end (top) and receive-end (bottom) of WDM transmission system.

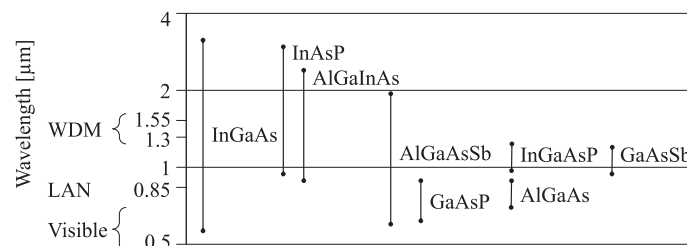


Fig. 11 Semiconductor materials for WDM laser diodes.

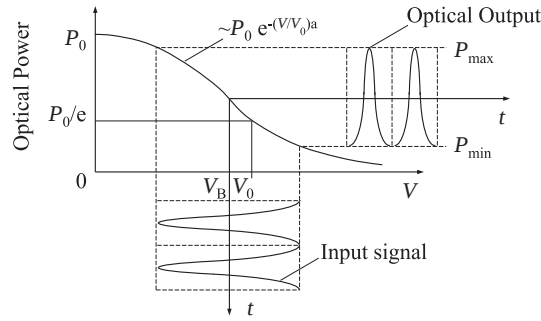


Fig. 12 Modulation of an EAM.

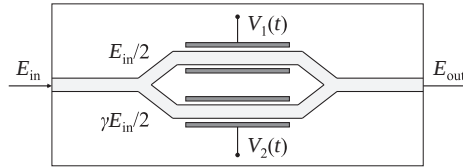


Fig. 13 MZM structure.

The output intensity of an EAM is a nonlinear function of the modulating voltage, see Fig. 12. For intensity modulation at high extinction ratio (ER) and low insertion loss, small bias voltage has to be used. For high bias voltage, ER decreases and insertion loss increases.

EAMs can be monolithically integrated with laser diodes, e.g., DFB lasers. This avoids adiabatic laser chirp. Since the same semiconductor system is used, transient chirp remains. EAMs are mainly used for 10-Gb/s intensity-modulated WDM transmission. Reach > 100 km on SSMF without dispersion compensation can be achieved.

In **electro-optic modulators**, the distribution of electrons within the medium is distorted if an electrical field is applied so that the refractive index changes anisotropically. This leads to phase changes of traversing optical signals. With the change of refractive index Δn , the resulting phase change of a signal propagating through an electro-optic medium is given by $\Delta(\varphi) = \Delta\beta_0 L = k_0 \Delta n L$. A relevant material for electro-optic modulators is LiNbO₃ (Lithium Niobate).

Electro-optic modulators are used as phase modulators or as nearly chirp-free amplitude modulators. For amplitude modulation, a Mach-Zehnder Interferometer (MZI) structure according to Fig. 13 is used. This structure is also called Mach-Zehnder Modulator (MZM).

The transfer function of the MZM is given as a function of the modulator voltages V_1 and V_2 :

$$H(V_1, V_2) = E_0 \cos[\pi(V_1 - V_2)/2V_\pi] \cdot \exp[j\pi(V_1 + V_2)/2V_\pi] \quad (12)$$

The (nonlinear!) cosine term accounts for amplitude modulation, and the exponential term for phase modulation or chirp. MZM can achieve bandwidths of > 100 GHz and high extinction ratio of 15–20 dB.

Single-Mode Fibers

Fibers have major impact on long-haul WDM transmission. Neglecting specialty and multi-mode fibers (the latter are relevant for LAN and inside data centers), relevant fibers are:

- Standard Single-Mode Fibers (ITU-T Recommendation G.652A-D (ITU-T, 2016d))
- Dispersion-Shifted single-mode Fibers (DSF, ITU-T G.653 (ITU-T, 2010b))
- Cut-off-Shifted single-mode Fibers (CSF, ITU-T G.654. A-C (ITU-T, 2016b))
- Non-Zero Dispersion-Shifted single-mode Fibers (NZ-DSF, ITU-T G.655A-E (ITU-T, 2009b))
- Ultra-broadband, non-zero dispersion-shifted single-mode fibers (ITU-T G.656 (ITU-T, 2010a))
- Low bending-loss single-mode fibers (ITU-T G.657A/B (ITU-T, 2016a))

SSMF are the oldest single-mode fibers. The first low-loss SSMF were built in 1978. Over time, several Recommendations (G.652A to G.652D) were produced which mainly improved water-absorption loss around 1400 nm and the PMD parameter D_{PMD} .

DSF combine lowest loss and lowest CD around 1550 nm. They were optimized of single-channel – *non-WDM* – transmission. Their close-to-zero CD is critical for WDM long-haul transmission since it allows phase matching between WDM channels which can lead to disastrous FWM. DSF have never been deployed on broad scale.

CSF have their cut-off wavelength shifted into the region of 1530 nm, and ultra-low loss. For wavelengths higher than cut-off, a fiber becomes single-mode. Cut-off shifting can be achieved by increasing the fiber core area, which also decreases nonlinearity. Fibers according to G.654B and G.654C have low PMD, and the G.654B type has CD *increased* as compared to SSMF. This makes G.654B CSF suitable for ultra-long-haul applications.

NZ-DSF are an improvement of DSF with regard to higher CD around 1550 nm. They aimed at a CD compromise between CD-limited reach and CD-supported nonlinearity suppression. Over time, CD was increased and PMD decreased. NZ-DSF also include large-area fibers.

G.656 fibers are the extension of G.655 fibers towards even higher CD. Recommendation G.656 further defines higher CD over increased bandwidth, low PMD and very low loss.

G.657 low bending-loss fibers were developed to decrease macro-bending loss resulting from cabling. G.657A fibers are splice-compatible with SSMF and have somewhat decreased macro bending loss. G.657B fibers are incompatible with SSMF and have even better bending loss.

Relevant dispersion parameters (D , dispersion slope S , zero-dispersion wavelength λ_0 , PMD parameter D_{PMD}) are listed for relevant fibers in [Table 2](#).

Some G.655 fibers have smaller effective core areas, compared to G.652 fibers, see [Table 1](#). This results from refractive-index profile manipulations which often leads to multiple claddings, and also decreased core diameter, as shown in [Fig. 14](#). Other G.655 fibers have increased effective areas (Large-Area (LA) or Large-Effective-Area fibers (LEAF) ([Corning, 2014](#))).

The technically achievable maximum transport capacity of SMF is approaching ([Essiambre et al., 2010](#)). *New fiber types* which possibly lead to WDM capacity increase are Few-Mode Fibers (FMF), Multi-Core Fibers (MCF), and Photonic Crystal Fibers (PCF), respectively.

In **Few-Mode Fibers**, several guided modes are modulated and demodulated independently. More guided modes are allowed by slightly increasing the core area, thus shifting cut-off such that the modes become allowed. FMFs are easily splicable, similar to SMF or CSF. All modes in an FMF can be amplified simultaneously by a single EDFA, given they are amplified equally. Due to increased core area, FMFs also have reduced nonlinear effects.

One challenge of FMF is the need for high-performance Multiple-Inputs Multiple Outputs (MIMO) receivers. These are required since mode de-/multiplexers, fiber distortion and other lumped devices will cause unavoidable mode coupling. For an FMF supporting the LP_{01} and the two degenerated LP_{11} modes, a 6×6 MIMO is required (each mode has two polarization modes). Further challenges relate to low-loss mode multiplexers and demultiplexers. The early devices used so far did not fully solve the low-loss requirement.

In **Multi-Core Fibers**, multiple single-mode cores are integrated in one fiber. Ideally, these cores are amplified by single fiber amplifiers (which can also be based on multiple cores), however challenges may occur with proper splicing.

Different MCF have been proposed and tested, originally based on three, seven, and 19 cores. Such MCF can have strong crosstalk between the cores. Then, MIMO receivers are required, similar to FMF. Crosstalk can also be suppressed by proper fiber design, e.g., by separating the cores with additional holey barriers. A MIMO may nonetheless be required if crosstalk is caused in fiber bends and lumped components.

Table 2 Dispersion characteristics of relevant single-mode fibers

	G.652A/C	G.652B/D	G.654B	G.654C	G.655A
D @ 1550 nm (ps/(nm · km))	16–21	16–21	22	20	0.1–6.0
λ_0 [nm] (Zero-Dispersion)	1311.5 ± 10	1311.5 ± 10	(~1300)	(~1300)	1480 ± 30
S @ 1550 nm (ps/(nm ² · km))	≤ 0.092	≤ 0.092	0.07	0.07	≤ 0.05
D_{PMD} (ps/√km)	0.5	0.2	0.2	0.2	0.5
	G.655B	G.655C	G.655E	G.656	G.657A
D @ 1550 nm (ps/(nm · km))	1.0–6.0	1.0–6.0	5.5–10	2.0–14	16–21
λ_0 (nm) (Zero-Dispersion)	1480 ± 30	1480 ± 30	1440	1480 ± 30	1312 ± 12
S @ 1550 nm (ps/(nm ² · km))	≤ 0.05	≤ 0.05	≤ 0.052	≤ 0.05	≤ 0.092
D_{PMD} (ps/√km)	0.5	0.2	0.2	0.2	0.2

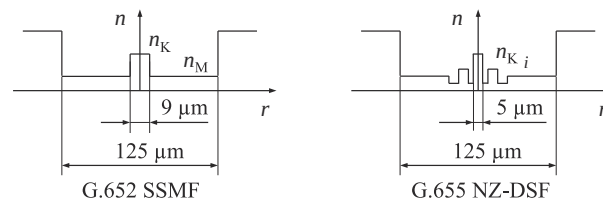


Fig. 14 Refractive index profile $n(r)$ for G.652 SSMF (left) and typical G.655 NZ-DSF (right).

In addition, new devices for separating the cores are required. Such devices can be based on extensions of fiber tapers. Further problems include the necessary splicing to amplifiers.

Photonic Crystal Fibers (PCF) have micro-structured cross sections, often comprising axial holes running along the fiber. They can either increase the fiber's confinement, or allow air guidance. This allows improving bending loss, PMD or nonlinear effects (which can be either decreased or increased, according to the needs). PCF split into two categories: index-guided and Photonic Bandgap- (PBG-) guided fibers (Russell, 2006). Index-guided PCF typically have a solid core and surrounding holes, the aim of PBG-guided fibers is hollow-core air guidance.

Solid-core, index-guided fibers with holey cladding have lower *effective* refractive index in the cladding and higher contrast than Silica fibers. They produce stronger confinement which allows stronger nonlinearities, birefringence, or bending sensitivity. The best attenuation reported so far is 0.28 dB/m at 1550 nm. Today, it is not clear if these fibers will play a more important role in WDM transmissions. One relevant application may be the crosstalk suppression provided by holey barriers in MCF.

PBG-guided fibers aim at air-guidance, allowing very low effective index (i.e., increased propagation velocity), and almost zero nonlinearity and PMD. Therefore, it is likely that they will play an increasingly important role in WDM transmissions. PBG-guided fibers split into Bragg fibers and hollow-core PCFs, see Fig. 15. Bragg fibers utilize the fact that 1D crystals can reflect light from all angles and polarizations. They form waveguides similar to hollow metal waveguides. They hence support the TE_{01} as the lowest-loss mode, which is immune to PMD and does not split into polarization modes.

Hollow-core PCFs consist of 2D-periodic photonic crystals and a hollow core. Here, the PBG is distorted, allowing guided fields to be confined into the hollow core region. This allows low nonlinearity, PMD and effective refractive index. The latter makes PCF suitable for "fast" fibers where the group velocity exceeds the one known from Silica SSF. Best results reported so far are 1.2 dB/km at 1620 nm over limited fiber length (Roberts et al., 2005).

Optical Amplifiers

In optical amplifiers, the WDM Optical Multiplex Section is amplified in the analog optical domain. Optical amplifiers are transparent with regard to payload protocols, Baud rates, and modulation schemes of the individual WDM channels. They do not provide equalization, pulse re-shaping or re-timing. Instead, they *add noise*. This noise is called Amplified Spontaneous Emission (ASE).

The optical amplifiers most relevant to WDM transmission are Erbium-Doped Fiber Amplifier (EDFA), Fiber Raman Amplifiers (FRA), and Semiconductor Optical Amplifiers (SOA). Parametric optical amplifiers (using the fiber's Kerr nonlinearity) so far have not proven relevance, and Brillouin fiber amplifiers cannot be used for WDM signals due to lack of bandwidth (see Fig. 9). WDM amplifier technologies are shown in Fig. 16 (Hirano, 2002).

EDFAs are the most relevant amplifiers for WDM transmission (Desurvire, 1994). Originally developed for DWDM C-band transmission, similar fiber amplifiers exist for other wavelength bands as well. Examples include (co-doped) L-band EDFAs, and Thulium-doped and Praseodymium-doped amplifiers for S-band and O-band amplification, respectively.

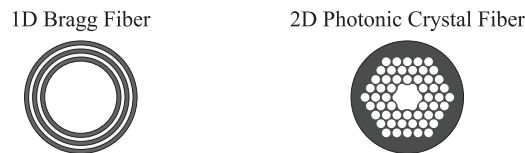


Fig. 15 Air-guiding Photonic Bandgap (PBG) fibers.

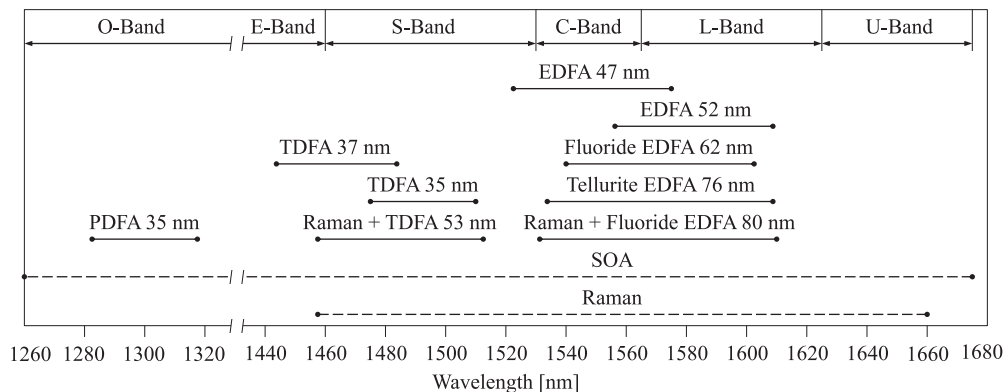


Fig. 16 WDM amplifier technologies.

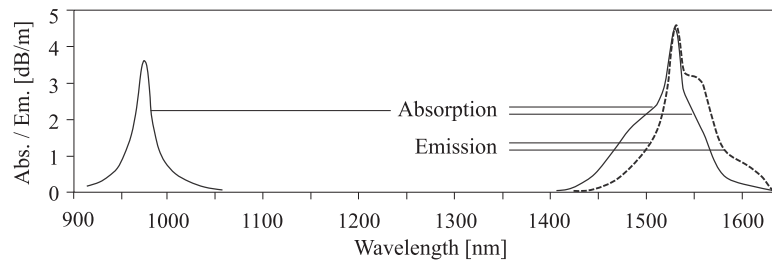


Fig. 17 Absorption and emission bands of an EDFA.

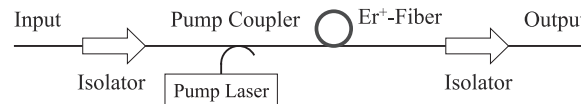


Fig. 18 Principle EDFA configuration with isolators.

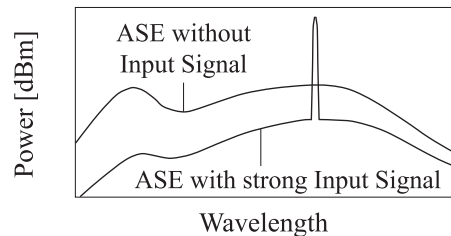


Fig. 19 EDFA ASE spectrum.

Erbium can produce gain at 1520–1610 nm, has absorption bands elsewhere which are accessible with low-cost pump lasers, and achieves high population inversion at reasonable pump-power levels. The absorption and emission bands of an EDFA are shown in [Fig. 17](#). Two absorption bands are shown, one centered around 980 nm, the other overlapping with the emission band. Here, pumping is possible at 1480 nm, which is no relevant WDM payload wavelength.

Since an EDFA produces ASE, optical isolators are necessary to prevent back-propagation of noise. The resulting simplified block diagram of an EDFA is shown in [Fig. 18](#).

Both, the ASE spectrum and the gain characteristics of an EDFA strongly depend on the (WDM) input signals, as seen from [Fig. 19](#).

The gain dependence on input power necessitates power-level control specially for cascaded amplifiers. Power levels can be controlled at the transmit side and in in-line sites which perform equalization, e.g., with variable optical attenuators. This is often implemented in ROADMs.

The gain spectrum of **Fiber Raman Amplifiers** depends on the spectral location of the Raman pumps. Given the availability of cost-effective high-power pumps, Raman amplification is possible throughout the entire low-loss region of the optical fiber (it is difficult at best at strong water absorption around 1400 nm).

The SRS gain spectrum, relative to its pump wavelength, is shown in [Fig. 7](#). With two pumps, gain bandwidths exceeding 90 nm can be achieved, as indicated in the top insert of [Fig. 20](#). FRA use the transmission fiber itself as gain medium. Pumps can propagate co- and/or counter-directionally with the WDM signals to be amplified. Backward pumping is preferred since forward pumping requires higher pump-power levels. It is often used as pre-amplification for EDFAs. Both operation modes are shown in [Fig. 20](#).

Distributed Raman amplification depends on the transmission-fiber type. On G.655 NZ-DSF, C-band amplification becomes less effective because the pumps fall into the zero-CD region and become subject to FWM.

Due to the ultra-short time constant of SRS, sophisticated transient and power-level control is required. Transient control can be based, e.g., on gain clamping.

Semiconductor Optical Amplifiers are similar in structure to index-guided Fabry-Pérot laser diodes ([Agrawal, 1992](#)). SOAs have broad gain bandwidth, and with different devices, the complete wavelength region of 1280–1650 nm can be covered. For CWDM (ranging from 1271 to 1611 nm), SOAs are the only cost-efficient amplification technology.

For multi-channel WDM operation, gain-control mechanisms are required. At system level, a saturating reservoir channel or small-signal operation (linear regime) can be implemented. At device level, laser gain clamping can be used. In a gain-clamped

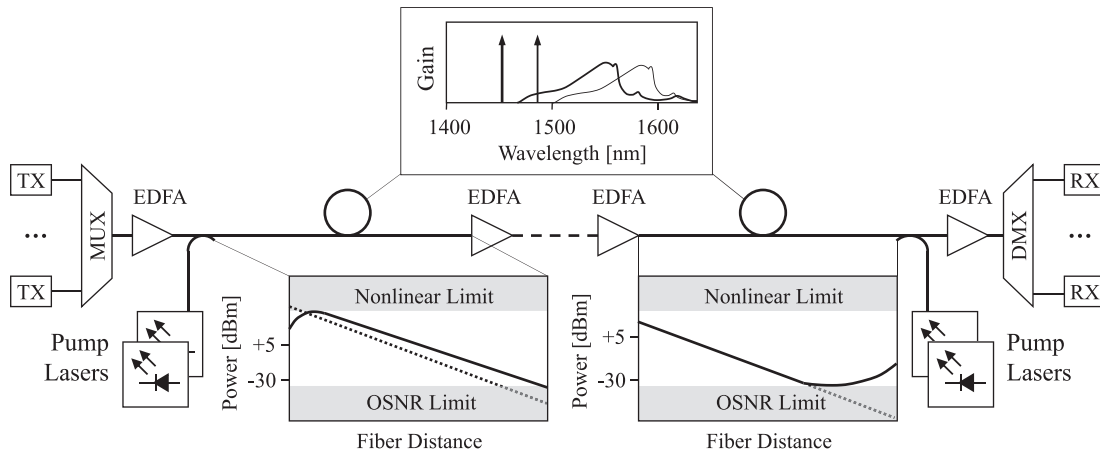


Fig. 20 Raman amplification with co- and counter-propagating pumps.

Table 3 Comparison of relevant SOA and fiber-amplifier parameters

	EDFA	FRA	SOA
Max. Gain (dB)	> 40	> 25	> 30
Wavelength Region (nm)	1530–1610	1280–1650	1280–1650
3-dB Bandwidth (nm)	30–60	Pump-dependent	60
P_{Sat} (dBm)	22	$0.75 \times \text{Pump}$	18
Polarization Dependence (dB)	0	0	< 0.5
Noise Figure, typ. (dB)	4–6	3–5	6–8
Pump Power	25 dBm	> 30 dBm	< 400 mW
Time Constant	10 ms	fs–ps	0.2 ns
Switchable	No	No	Yes

SOA, the gain medium is shared between SOA and a laser. Hence, the device is operated above threshold and produces constant gain. Lasing occurs at wavelengths different from the amplification band.

Table 3 lists relevant SOA parameters in comparison with those of fiber amplifiers.

Chromatic-Dispersion Compensation

Compensation of CD became necessary in the 1990s with WDM per-channel bit rates increasing to 10 Gb/s. Due to its linear, deterministic nature, CD can be compensated quasi-statically in the optical domain. The most important techniques were based on (Agrawal, 1992, 1995):

- Dispersion Compensation Fibers (DCF)
- Dispersion-compensating gratings, e.g., Fiber Bragg Grating (FBG)

In addition, fiber nonlinearity can be used for (partial) CD compensation. This is achieved by mid-span spectral inversion or Self-Phase Modulation (Soliton effect).

CD design must provide correct amounts of net dispersion at each add/drop node in a network. Performance penalties caused by nonlinearity and CD for each *potential* connection must be as low as possible. Optimum CD design depends on per-channel bit rate and modulation, number of WDM channels and fiber type. For 10-Gb/s On-Off Keying (OOK), a dispersion map with Distributed Undercompensation (DUCS) in each span is preferred. In contrast, 40-Gb/s (NRZ-DPSK) signals prefer slight overcompensation in each span, and zero residual dispersion at the receiver. A DUCS dispersion map is shown in Fig. 21.

In addition to quasi-static compensation, various Tunable Optical Dispersion Compensators (TODC) have been developed, especially for the directly-detected 40-Gb/s WDM systems. TODCs became necessary to compensate the remaining receive-end residual CD. They may become relevant for direct-detect 400-Gb/s (based on Pulse-Amplitude Modulation with four levels (PAM-4)) again.

After Y2000, digital signal processing (DSP) became powerful enough to cope with real-time dispersion equalization in high-speed fiber-optic transmission. DSP is mostly applied in coherent receivers, but pre-distortion at the transmitter is also possible. The most commonly used real-time equalizers are Feed-Forward Equalizers (FFE).

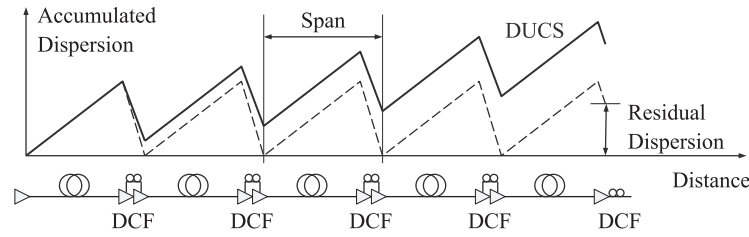


Fig. 21 Compensated multi-span long-haul link with DUCS.

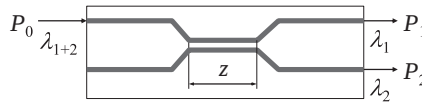


Fig. 22 Wavelength-selective fused fiber coupler.

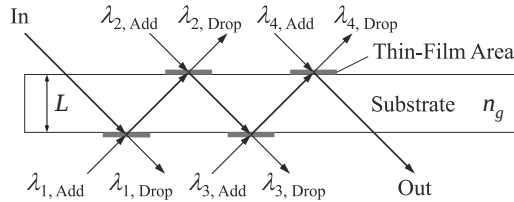


Fig. 23 Thin-film WDM filter.

Passive WDM Filters and Couplers

Passive static optical filters are used as de-/multiplex or add/drop filters and as noise-bandwidth limiters. They are now partially replaced by Re-configurable Optical Add/Drop Multiplexers (ROADMs).

Wavelength-selective directional couplers can be made from two fibers with different core diameters and refractive indices. These fibers are matched at only one wavelength each, the device can hence be used as *diplexer* (Senio and Jamro, 2009). The principle is shown in Fig. 22.

For the diplexer, the *isolation loss* is given as the power of the desired wavelength at a given output port divided by the power of the other wavelength at the same port: $\alpha_{ISO} = 10 \log_{10}(P_{\lambda_1, \text{port1}}/P_{\lambda_2, \text{port1}})$. In multi-port filters, this is called *adjacent-channel isolation* and is given as ratio of desired wavelength at one port, divided by the power of the *two adjacent* wavelengths at the same port. The *total isolation* in a multi-port device is given by the power of the desired wavelength divided by the accumulated power of *all* other wavelength.

Dielectric Thin-Films Filters (TFF) are relevant for WDM de-/multiplexing and add/drop. TFF are based on Fabry-Pérot filters (FPF), the principle is shown in Fig. 23.

A FPF is essentially a resonator, bound on either end by a partial mirror. Light of frequencies other than resonant frequencies is mostly reflected internally. Light at resonant frequency enters the cavity and exits through the opposite mirror. In the TFF shown in Fig. 23, layers of dielectric thin films are used as mirrors. Thin-film reflection or transmission is wavelength-dependent. Hence, the add/drop locations depend on wavelength, and add/drop wavelengths are assigned individual filters.

Relevant parameters of an FPF are the finesse and the Free Spectral Range (FSR). The finesse is a measure for the filter quality. Higher reflectivity of the thin-film mirrors leads to higher finesse. Resonant wavelengths are given by $\lambda = 2Ln_g/m$, where n_g is the refractive index of the substrate, and $m = 1, 2, 3, \dots$. The FSR is the distance between peaks in the filter response. It is given by $FSR = c/n_gL$. The finesse is given as the ratio of FSR and Full-Width at Half-Maximum (FWHM), $FSR/FWHM = \pi\sqrt{R}/(1-R)$. R is the reflectance of the mirrors.

Arrayed Waveguide Gratings (AWGs) can multiplex/demultiplex high WDM channel numbers (up to ~ 100) with relatively low loss (Dragone, 2005). Fig. 24 shows an $N:M$ AWG. It consists of two couplers (free-propagation regions, FPR) which are connected by a waveguide array with equal length difference ΔL between adjacent array waveguides.

Light coming from an input waveguide is diffracted and coupled into the arrayed waveguides by the first FPR. The optical path-length difference ΔL between adjacent array waveguides equals an integer multiple of the central wavelength λ_0 of the demultiplexer. As a consequence, the field distribution at the input aperture will be reproduced at the output aperture of the second FPR. Therefore, at the center wavelength, the light focuses in the center of the image plane if the input waveguide is centered in the input

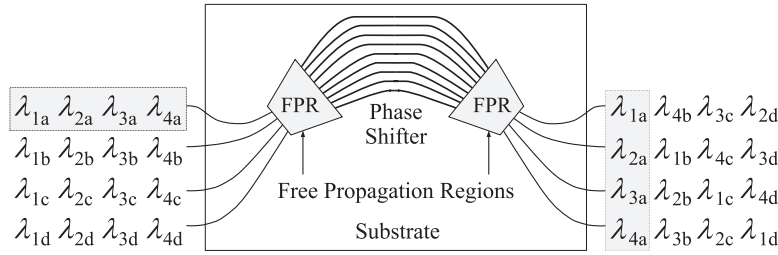


Fig. 24 $N:M$ AWG as column-row converter.

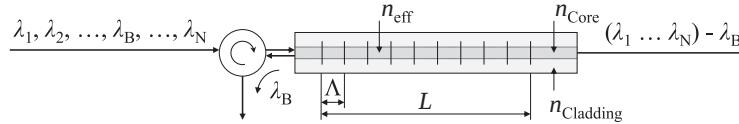


Fig. 25 Fiber Bragg Grating (FBG).

plane. If the input wavelength deviates from the central wavelength, phase changes occur in the array branches which increase linearly from the inner to outer array waveguides. This causes the wavefront to be tilted at the output aperture. Therefore, spatial separation of output wavelengths is given.

The FSR of an AWG is given by $FSR = c/n_2\Delta L$ (with n_2 the refractive index in the waveguides between the FPRs). AWGs can route one wavelength each in several wavelength bands, which are separated by the FSR. The respective devices are also called *cyclic* AWGs. Cyclicity can be used for capacity upgrades of WDM systems or in systems based on single-fiber working, utilizing, e.g., the C-band and the L-band. In ITU Recommendation G.698.3, cyclic AWGs with up to 48 ports (100 GHz in the C-band) are standardized.

As integrated waveguide devices, AWGs show strong temperature dependence. The temperature-induced wavelength shift can be expressed as (Grave de Peralta *et al.*, 2004):

$$\frac{\partial \lambda}{\partial T} = \frac{\lambda}{nL} \left(\frac{\partial(nL)}{\partial T} \right) = \frac{\lambda}{nL} \left(n \frac{\partial L}{\partial T} + L \frac{\partial n}{\partial T} \right) = \lambda \left(\frac{1}{L} \frac{\partial L}{\partial T} + \frac{1}{n} \frac{\partial n}{\partial T} \right) = \lambda \left(\alpha + \frac{1}{n} \frac{\partial n}{\partial T} \right) \quad (13)$$

For silica, $\partial n/\partial T$ is in the range of $7.5 \cdot 10^{-6}/^\circ$, and for silicon, $\alpha = 2.63$ ppm/ $^\circ$, respectively. A silica-on-silicon device has a wavelength drift (red shift) of $d\lambda/dT = 12$ nm/ $^\circ$. A 50-GHz DWDM device (~ 400 pm channel spacing) is completely transparent every 400 pm, and opaque in between. The device becomes a wavelength stop if the temperature changes by $\sim 17^\circ$. Hence, AWGs must be temperature-stabilized or athermalized. This can be achieved with mechanical and/or material compensation. Typical total insertion loss of athermalized 50-GHz AWGs is in the range of 6 dB, which is almost independent on the WDM port count.

Fiber Bragg Gratings (FBGs) are periodic perturbations of the refractive index in the propagating medium. An FBG in an add/drop configuration is shown in Fig. 25. Λ is the period of the grating (or perturbation). The Bragg phase-matching condition is satisfied for two waves with β_0 and β_1 if $|\beta_0 - \beta_1| = 2\pi/\Lambda$. If a wave with β_0 propagates through the grating, its energy is coupled onto a scattered wave traveling in the counter direction at the same wavelength if $|\beta_0 - (-\beta_0)| = 2\beta_0 = 2\pi/\Lambda$.

Fiber Bragg Gratings can be produced from photosensitive (Ge-doped) single-mode fibers. The grating is written into the fiber with UV lasers, where regions with higher UV intensity produce higher refractive index. The required refractive-index change is in the range of $\Delta n \approx 10^{-4}$.

Advantages of FBGs include low loss (down to 0.1 dB), ease of coupling to transmission fibers, polarization insensitivity, and high crosstalk suppression. The typical temperature coefficient is in the range of $\sim 1.2 \cdot 10^{-2}$ nm/ $^\circ$ C (Lima *et al.*, 2005). Therefore, FBGs must be operated in temperature-controlled environment, or be temperature-compensated.

FBG with short period ($\Lambda \approx 0.5$ μ m) or long period ($\Lambda \approx 100$ – 1000 μ m) exist. *Short-period* FBGs are well-suited as filters or channelized CD compensation devices in WDM systems. *Long-period* FBGs can be used, e.g., for broadband EDFA gain equalization.

Interleavers are periodic filters used for combining or separating (de-interleaving) two WDM multiplex sections with the same channel spacing and a grid offset of half the spacing. For example, a 50-GHz DWDM signal can be composed of two 100-GHz signals. Due to the required periodicity, Mach-Zehnder Interferometers are suitable devices for interleavers. Similar to MZIs used as modulators (see Fig. 13), MZIs used as interleavers introduce a delay ΔL to one of the branches, as shown in Fig. 26.

The signal coming from Input 1 and going to Output 1 through the upper arm acts as reference. Then, the signals coming from Input 1 and going through the lower arm to Output 1 and Output 2 experience phase lags of $\pi/2 + \beta\Delta L + \pi/2 = \pi + \beta\Delta L$ and $\pi/2 + \beta\Delta L - \pi/2 = \beta\Delta L$, respectively. For $\beta\Delta L = k\pi$, $k = (2n + 1)$, the signals in the upper output add in phase. At Output 2, the phase difference is $(2n + 1)\pi$ which is out of phase, hence there is no signal at the output. Opposite behavior is achieved for

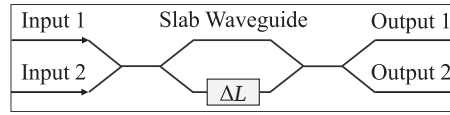


Fig. 26 MZI as (de-) interleaver.

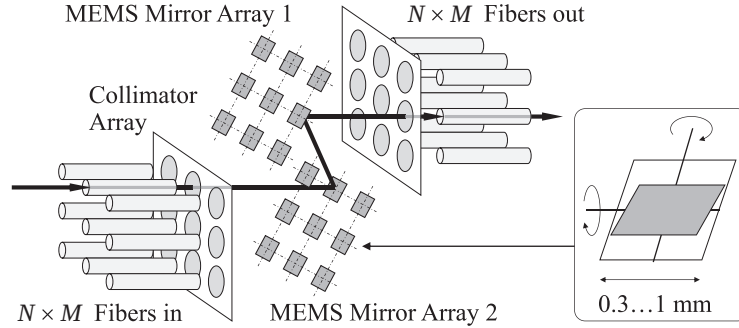


Fig. 27 3D-MEMS used for multi-dimensional fiber cross connect.

$\beta\Delta L = 2n\pi$. The MZI transfer function is given by:

$$\begin{pmatrix} T_{11}(f) \\ T_{12}(f) \end{pmatrix} = \begin{pmatrix} \sin^2(\beta\Delta L/2) \\ \cos^2(\beta\Delta L/2) \end{pmatrix} \quad (14)$$

For use as a de-interleaver, λ_1 and λ_2 (both at Input 1) are chosen to coincide with the minima and maxima of the transfer function. With $\beta = 2\pi n_{\text{eff}}/\lambda$, the delay ΔL and the corresponding phase lag $\beta\Delta L$ can be expressed as $\Delta L = m_i \lambda_i / 2n_{\text{eff}}$, $\beta\Delta L = m_i \pi$. For $\lambda_1 = 2\Delta L n_{\text{eff}} / m_i$, m_i odd, and Output 1 has a signal whereas Output 2 has no signal. For $\lambda_2 = 2\Delta L n_{\text{eff}} / m_i$, m_i even, the output assignment is vice versa. Hence, the device acts as a de-interleaver. Operation as interleaver is similar.

Tunable WDM Filters

Tunable filters have been investigated since the early WDM days. Advantages of tunable filters include remote reconfigurability, and reduction of system items and spare parts. Tunability can be based on several technologies:

- TFF, thermally tuned.
- Liquid-Crystal filter, electrically tuned.
- MEM-tunable devices.
- FBG, temperature- or strain-tuned.
- Acousto-optic filter, tuned via electrostriction.
- Electro-optical filter, electrically tuned.
- AWG, thermally tuned.
- MZI, cascaded, with heaters.
- Ring-Resonator filters, thermally tuned.

So far, not all of these technologies have been commercialized.

Wavelength Switching Devices

Optical switching matrices in WDM systems are used for cross connects or Reconfigurable Optical Add/Drop Multiplexers (ROADMs). They must have low insertion loss, high isolation, and low back reflection. In addition, they need to be highly reliable, fast (for certain application), compact, low power-consuming, and scalable. Relevant techniques include Micro Electro-Mechanical Systems (MEMS) and Liquid Crystal Technologies (LCT).

MEMS refer to arrays of tiny tilting mirrors sculpted from semiconductor materials such as silicon. In 3D-MEMS arrays, the mirrors can be tilted in any direction. 3D-MEMS arrays can be used for cross connects on wavelength, wavelength-band or fiber level. A 3D-MEMS assembly is shown in Fig. 27.

As MEMS is based on mechanical movement, it offers the best isolation performance available. For the same reason MEMS cannot be used for drop-and-continue applications.

In LCT, liquid crystals alter their transmission characteristics when applying an electric field. LCT devices can throttle the amount of light that passes through them. They can hence be used as variable attenuators or power splitters with configurable splitting ratio. This enables drop-and-continue applications. LCT has low power consumption (as known from displays). It can be combined with electro-holographic technology for increasing switching speeds.

Liquid-Crystal-on-Silicon (LCoS) is the latest addition to LCT (Sakurai *et al.*, 2012). It allows WDM-channel switching based on many sub-elements. Variable numbers of these sub-elements can be bonded for varying channel bandwidth without attenuation dips, as shown in Fig. 28. LCoS can be used for *flexible-grid* ROADMs (with internal frequency granularity of 6.25 GHz).

ROADMs

Reconfigurable Optical Add/Drop Multiplexers are used in flexible WDM networks. Early technologies led to Degree-2 ROADMs for WDM rings or linear add/drop links. In meshed networks, Multi-Degree (MD-) ROADMs are required. Most devices are based on *Wavelength-Selective Switches* (WSS, based on MEMS or LCT/LCoS). This evolved from Degree-4 ROADMs (4 network ports plus local add/drop port) via Degree-8 to degrees exceeding 30.

The functionality of MD-ROADMs evolved from colored and directed add/drop via directionless and colorless ROADMs to contentionless devices, see Fig. 29(A–D). Meanwhile, these variants can be implemented with *flexible-grid* technology.

In a **colored/directed ROADM**, each terminated network fiber (degree) requires dedicated add/drop filters. A client device hence must be connected to the correct add/drop filters and filter ports. This problem is solved in **directionless** ROADMs. An additional WSS (Fig. 29(B)) connects the add/drop ports of the line-terminating WSSs. An add/drop filter is connected to this additional WSS. Drop signals can now be selected from any of the terminated lines, and add signals be directed respectively. Add/drop filter ports are still colored, i.e., specific wavelengths terminate at specific ports. Due to the use of a single filter in the add/drop path, each wavelength can only be terminated once, leading to potential wavelength blocking.

In **colorless, directionless** ROADMs, a further WSS provides uncolored add/drop ports, see Fig. 29(C). Different wavelengths can be terminated at any WSS port. This is one application area for WSS with (add/drop) port counts ≥ 20 . Wavelength blocking can still occur. Colorless, directionless ROADMs allow *restoration* switching.

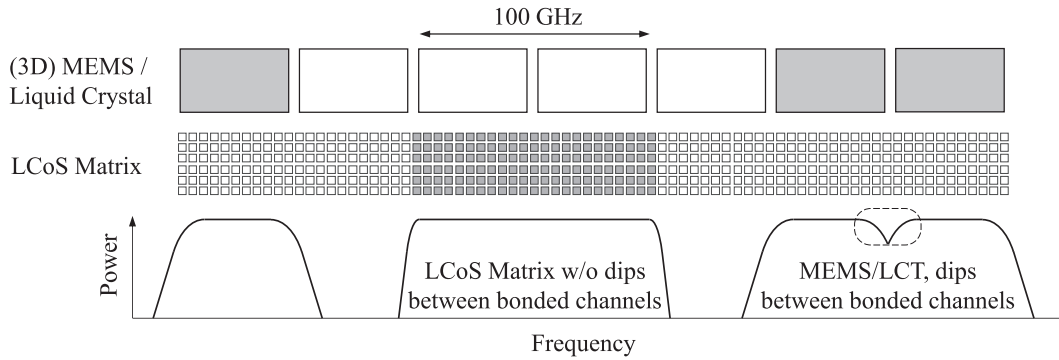


Fig. 28 LCoS flexible-grid matrix WSS.

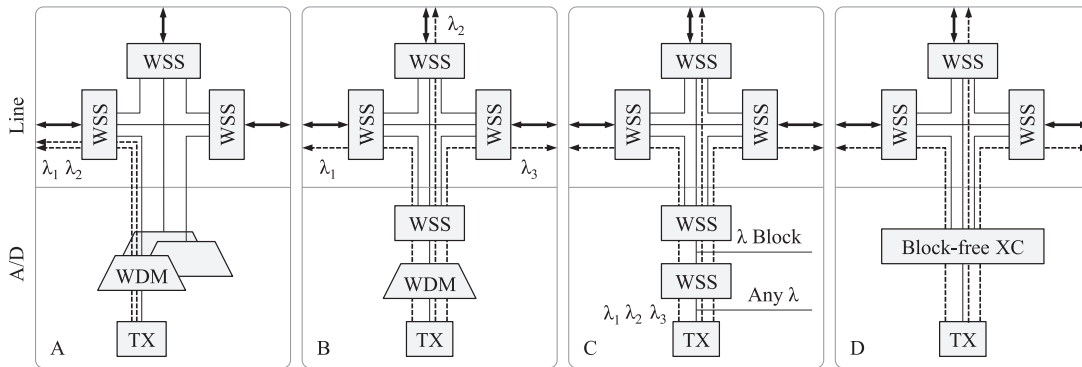


Fig. 29 Basic ROADM functionality. (A): colored per-degree add/drop. (B): directionless add/drop. (C): directionless and colorless add/drop. (D): directionless, colorless, and contentionless add/drop.

Wavelength-blocking is solved in **contentionless** ROADMs with a blocking-free cross connect (XC in Fig. 29(D)) in the add/drop. One implementation is shown in Fig. 30. It is based on WSSs in the outgoing (line) directions and power splitters for the incoming directions. The add/drop consists of WSS/power-splitter combinations, $N \times M$ switches, and a passive shuffle (i.e., patches) connecting all WSS/power-splitter combinations to all switches.

ROADMs, like other filters, exhibit bandwidth narrowing when being cascaded. This must be considered in transmission at high Baud rates. Depending on the filter and signal characteristics and number of cascaded filters, penalties on the receive-end OSNR must be applied. The effect can be counter-acted by proper spectral shaping, or with flexible-grid ROADMs.

Receivers

In WDM systems, two types of photo diodes are used as photo detectors. In a **PIN photo diode**, a p-n junction of suitable semiconductor material is used as high-speed photo detector. To improve responsivity, a lightly-doped intrinsic semiconductor is put between the p- and n-type semiconductors. To allow absorption in the intrinsic region, the p- and n-regions ideally are kept transparent.

The primary photocurrent resulting from absorption is given by $I_p = RP_0$. The responsivity R of the PIN diode results as $R = \eta q / h\nu [A/W]$. The external quantum efficiency η is the ratio of photo-generated electron-hole pairs and incident photons, $\eta = (I_p/q)/(P_0/h\nu)$. P_0 is the incident optical power, q is the electron charge, and $h\nu$ is the photon energy.

Fig. 31 shows the responsivity of InGaAs and Ge which are the relevant materials covering the relevant WDM wavelengths (Azadeh, 2009). For wavelengths $\lambda < 900$ nm, Si can be used.

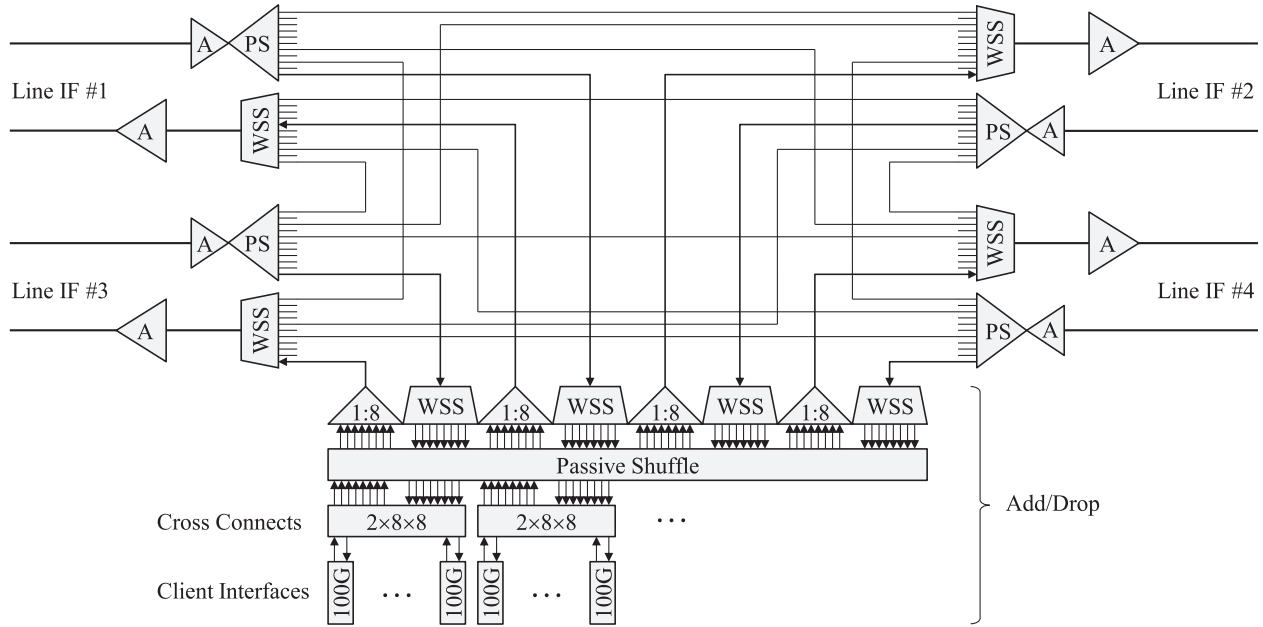


Fig. 30 Block diagram of Degree-N directionless, colorless, contentionless ROADM. PS: power splitter. A: amplifier (EDFA).

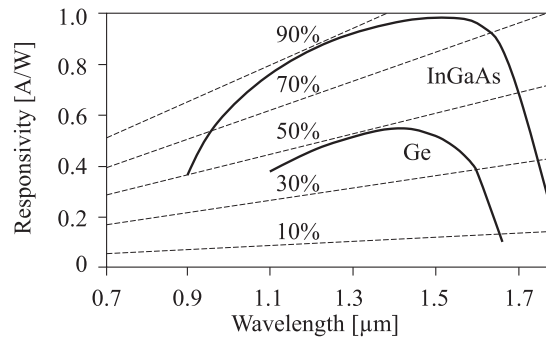


Fig. 31 Spectral responsivity of Ge and InGaAs photo diodes.

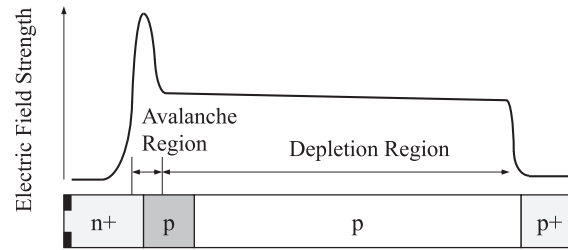


Fig. 32 Structure of an APD, and electrical field strength along its structure.

Table 4 Comparison of photo diodes

		<i>Ge</i>	<i>InGaAs</i>
Wavelength range, λ (nm)	PIN, APD	900–1650	1000–1700
Responsivity, R (A/W)	PIN	0.4–0.5	0.7–0.95
Quantum efficiency, η (%)	PIN	50–55	60–70
Avalanche gain, M	APD	50–200	10–40
k -Factor, k_A	APD	0.7–1.0	0.5–0.7
Bandwidth, B (GHz)	PIN	0.5–3	1–40
	APD	1.5 max.	1–5
Bias Voltage, V_B (V)	PIN	5–10	5–6
	APD	20–40	20–30

Reproduced from Azadeh, M. Fiber Optics Engineering. Dordrecht Heidelberg: Springer; 2009. ISBN 978-1-4419-0304-4.

Avalanche Photo Diodes (APDs) provide inherent gain. They are doped such that a region with very high field strength results when the diode is reverse-biased with high voltage. This region produces internal current gain due to impact ionization (avalanche effect). Higher reverse bias leads to higher gain, including stronger noise generation. The structure of an APD is shown in Fig. 32, together with the electrical field strength along the device (Azadeh, 2009).

APDs for WDM applications use the same semiconductor materials used for PIN diodes (see Fig. 31). InGaAs has less multiplication noise than Ge.

The responsivity of an APD can be expressed by the gain M , $R_{APD} = R_{PIN}M$ (M is set to 1 for PIN diodes). M is given by $M = I_M/I_P$, with I_M the average value of the total multiplied output current. Since the noise figure $F(M)$ increases with M , there is an optimum value of M that maximizes signal/noise.

A comparison of relevant photo-diode parameters is given in Table 4 (Azadeh, 2009).

Non-Fiber-Related Effects

In Section Transmission Effects in Optical Fibers, effects related to the transmission fibers were discussed. Here, an overview is given on crosstalk (from demultiplexing) and noise (from amplification).

Linear Crosstalk

Linear crosstalk originates in WDM receivers from imperfect demultiplexing or faulty routing of certain channels. According to ITU Recommendation G.Sup39 (ITU-T, 2012b), two mechanisms exist:

- *Inter-channel crosstalk*. Ratio of total power in the disturbing channels to that in the wanted channel. Wanted and disturbing channels have different wavelengths.
- *Interferometric or intra-channel crosstalk*. Ratio of disturbing power *not* including ASE noise to the wanted power *within the same* wavelength.

Inter-channel crosstalk occurs when WDM channels are demultiplexed in a filter with non-perfect isolation. At the demultiplexer output port under consideration, the disturbing (neighbouring) channels are attenuated with respect to the wanted channel by the (adjacent) filter isolation I (dB). The worst case is given if the power of the channel under consideration is at its guaranteed minimum, and all other channels are at their allowed maximum. This maximum difference is denoted d (dB) in Fig. 33. IL denotes the filter insertion loss.

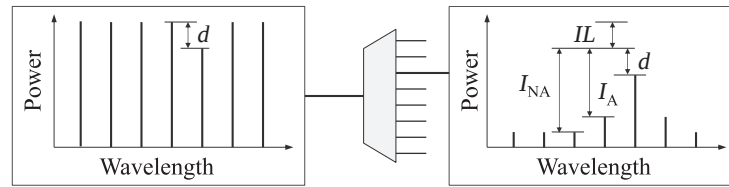


Fig. 33 WDM demultiplexer isolation.

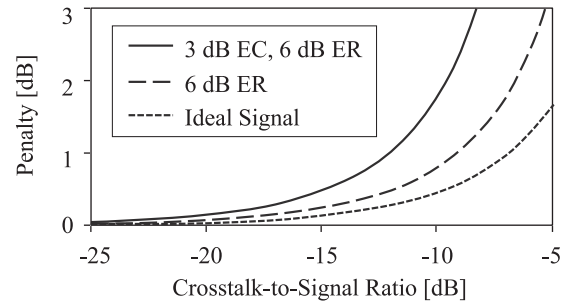


Fig. 34 Optical penalty vs. inter-channel crosstalk from single interferer.

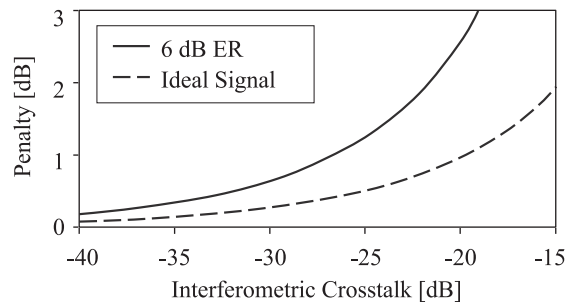


Fig. 35 Optical penalty vs. interferometric crosstalk (single interferer, optimum decision threshold).

In practical demultiplexers, the isolation I_A for the channels adjacent to the wanted channel is smaller than the isolation I_{NA} of the non-adjacent channels, $I_A < I_{NA}$.

The single-interferer inter-channel crosstalk penalty for various values of the eye closure (EC) and extinction ratio (ER) of the wanted signal is shown in Fig. 34. Penalties for real systems are somewhere below the highest curve. For multiple interferers, the penalty becomes smaller than the single-interferer ideal-signal penalty for small crosstalk, and higher for large crosstalk.

Interferometric crosstalk occurs when the disturbing channel and the wanted channel are at the same nominal wavelengths. Obviously, the disturbing signals cannot be suppressed with WDM filters. Interferometric crosstalk can occur:

- In a node where a wavelength is incompletely dropped before the new signal is added
- In a multiplexer where a transmitter emits at the wavelength of another channel. This can occur if channels are combined with power splitters rather than filters. Interference can then result from wrong laser tuning or insufficiently suppressed *laser side modes*.
- In a cross connect or MD-ROADM with poor isolation, where light from more than one network fiber (degree) can reach the respective receive port
- In any component, subsystem or network element with more than one light path that connects to the receiver. This effect is also called *Multi-Path Interference*.
- *Intentionally*. This is also referred to as crosstalk attack.

Interferometric crosstalk results from *field beating*. Therefore, lower crosstalk levels lead to the same penalties, compared to inter-channel crosstalk, see Fig. 35.

Interferometric multi-interferer crosstalk may have to be reduced by 2–4 dB as compared to single-interferer crosstalk for penalties in the range of 0.5–1 dB.

Noise in Optical Transmission Systems

Fiber and other components' loss leads to decreasing signal power levels. On long transmission links, this has to be compensated by optical amplifiers which amplify several WDM channels simultaneously. Like all amplifiers, optical amplifiers add noise (ASE). Together with the unavoidable noise sources in receivers, this decreases the Signal/Noise Ratio (SNR) which is a relevant measure for the achievable Bit-Error Rate (BER).

In order to derive the SNR which a photo diode generates, the primary noise sources must be considered. These are quantum (shot) noise and dark-current noise. Quantum noise arises from the statistical nature of photo-generated current. Dark-current noise results from the current that continues to flow through the bias circuit in the absence of light. It usually is not significant for PIN photo diodes, but can become significant in APDs. Further relevant noise sources include thermal noise from electronic receivers (Alexander, 1997) and ASE from optional optical pre-amplifiers. Consequently, PIN-based, APD-based and optically pre-amplified receivers must be considered separately.

Following, the achievable SNR and sensitivity of optical receivers (PIN, APD, optical pre-amplified) is given, followed by the noise figures of (chains of) optical amplifiers.

Achievable SNR (or, in the case of a pre-amplifier, noise figure) and sensitivities can be derived considering the relevant noise sources (Agrawal, 1992; Alexander, 1997). For weak and large PIN photo-diode input, thermal noise and shot noise dominate, respectively, leading to different SNR:

$$SNR_{PIN,therm} = \frac{R_L \sigma_p^2}{4k_B T B}, \quad SNR_{PIN,Shot} = \frac{\sigma_p^2}{2qI_p B} \quad (15)$$

R_L is the matched load resistance of the amplifier, $k_B = 1.38 \times 10^{-23}$ is Boltzmann's constant, T is the temperature, and B is the electronic (effective noise) receiver bandwidth. I_p is the photocurrent, q is the electron charge, and σ_p^2 is the input signal variance. These SNR values have to be decreased by the noise figure F_N of the electronic amplifier which follows the load resistor.

For APDs, SNR is dominated by shot noise, and avalanche gain M and excess noise factor $F(M)$ must be considered. The total shot noise (to replace $2qI_p B$ in Eq. (15)) is given by $\sigma_s^2 = 2q(I_p + I_D)BM^2F(M) + 2qI_L B$. I_D and I_L are the bulk dark current and surface current of the APD. The excess noise factor $F(M)$ is given by $F(M) = kM + (1-k)(2-1/M)$. The range for the factor k is $0 < k < 1$, see Table 4. For the avalanche gain M , also refer to Table 4.

Typical PIN receivers are dominated by thermal noise. They can be combined with *Optical Pre-Amplifiers* (OPA) in order to increase sensitivity. OPAs generate signal-ASE and ASE-ASE beat noise, respectively. For large OPA gain G , signal-ASE beat noise becomes the sole dominant noise. Then, the OPA noise figure F_N can be approximated as $F_N \approx 2n_{sp}$. Here, n_{sp} is the population-inversion factor. Large gain requires high population inversion, hence ideally $n_{sp} \approx 1$ and consequently, the optimum OPA noise figure is $F_{N,opt} \approx 2$. An ideal low-noise OPA has a noise figure of 3 dB and helps avoiding the thermal-noise limit of PIN diodes.

Sensitivity of an optical receiver is defined as the minimum average power which is required to achieve a BER of 10^{-12} . For On-Off Keying with direct detection, sensitivity of approximately -26 dBm, -36 dBm and -50 dBm results for receivers with PIN diodes, APDs, and OPAs, respectively. For the APD, $M=10$ and $F(M)=1.3$ have been assumed.

The achievable sensitivities heavily depend on the modulation and detection scheme. Sensitivity increases from direct detection via heterodyne detection to homodyne detection. In Tonguz and Wagner (1991), the equivalence between optically pre-amplified direct and asynchronous heterodyne detection has been shown. Quantum-limited sensitivity of 9 photons per bit is reached for coherent homodyne PSK reception.

Noise in cascaded optical amplifiers is most often derived in a semiclassical description using field-beating theory (Desurvire, 1994), considering signal-ASE beating.

A co- and counter-pumped dual-stage EDFA is shown in Fig. 36.

Here, Friis' formula for the total noise figure F_{tot} of cascaded amplifiers applies, $F_{tot} = F_1 + (F_2 - 1)/G_1$ (Friis, 1944). For sufficiently large gain G_1 , the total noise figure is dominated by the first stage. The first EDFA stage should be pumped with 980 nm, which achieves better inversion as compared to 1480-nm pumping. This leads to noise figures of 980-nm-pumped EDFAs which are typically better by 1 dB, compared to 1480-nm pumping (Kaminow and Koch, 1997).

Cascaded amplifiers in long-haul links are shown in Fig. 37. Amplifiers have total noise figures F_{Ni} and gains G_i , and the intermediate spans have attenuations A_i (given by αL_i , with α the loss coefficient and L_i the respective span length).

The noise figure of a chain of amplifiers along the link is then given by:

$$F_{total} \approx \frac{F_1}{A_1} + \frac{F_2}{A_1 G_1 A_2} + \dots + \frac{F_n}{A_n G_1 A_1 \dots G_{n-1} A_{n-1}} \quad (16)$$

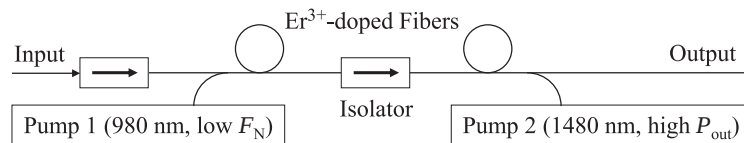


Fig. 36 Dual-stage EDFA.

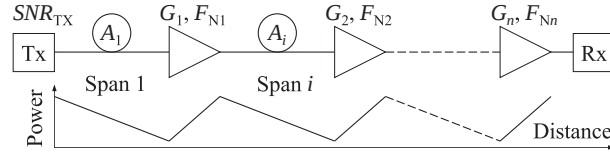


Fig. 37 Long-haul link with cascaded amplifiers.

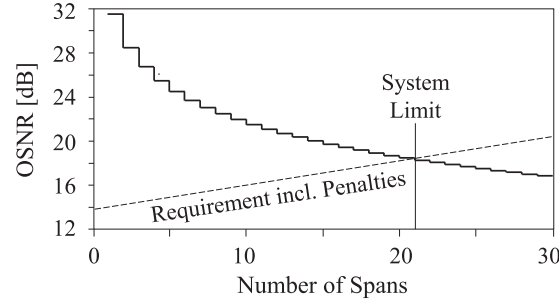


Fig. 38 Development of OSNR in long chains of (equidistantly spaced) amplifiers.

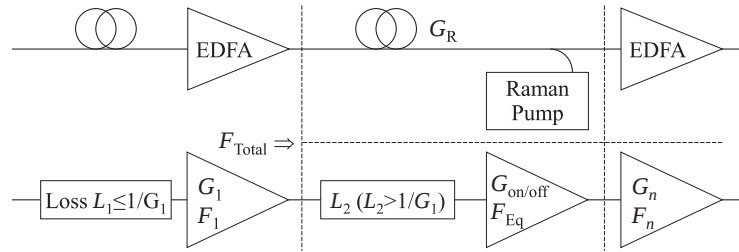


Fig. 39 Multi-span link with EDFAs and Raman pre-amplifiers, and equivalent model.

If $G_i \gg 1/A_i$, only the first amplifier dominates the receive-end noise behavior. If $G_i \leq 1/A_i$, then F_{total} is increased significantly by the respective spans (Loudon, 1985). If all amplifiers have equal gain G_i , and all links have equal loss A_i with $G_i = 1/A_i$, one gets $F_{\text{total}} = nF_N/A_i$. The resulting OSNR is often expressed in dB:

$$\text{OSNR}_{0.1\text{nm}} = P_{\text{out,Ch}} - A - F_N - 10\log(n) + 58 \quad (17)$$

This definition relates to a reference bandwidth of $B_{\text{o,ref}} = 125$ GHz (or 0.1 nm). A is the mean span loss (now in dB), n is the number of spans (or amplifiers). The OSNR in long-haul systems, assuming $G = 1/A$, drops logarithmically with increasing span number, see Fig. 38. Here, noise figures of $F_N = 6.5$ have been assumed. The system limit is explained in Section Long-Haul WDM Systems.

In ultra-long-haul systems with low-noise design, several 100 amplifiers can be cascaded (Loudon, 1985). In regional systems with irregular span lengths (and spans with $A_i > 1/G_i$), the number of amplifiers is much smaller, often it is < 20 .

The OSNR resulting from very long single spans or very long multi-span links can be improved with FRA. This is often done with backward-pumped Raman pre-amplifiers, acting as *distributed* pre-amplifiers to subsequent EDFAs, see Fig. 39. The EDFA input is prevented from getting a certain critical OSNR, as shown in Fig. 20. This results in improved noise figure of the combined amplifiers.

Often, distributed FRA are described via an *Equivalent* Noise Figure F_{Eq} and the Raman on/off gain $G_{\text{on/off}}$ of a *lumped* Raman amplifier (Bristiel et al., 2006). These are given by $F_{\text{Eq}} = F_{N,\text{FRA}} \exp(-\alpha_s L)$ and $G_{\text{on/off}} = G_R(L) \exp(\alpha_s L) = \exp(g_R P_{\text{P0}} L_{\text{eff}})$. $F_{N,\text{FRA}}$ is the noise figure of the FRA, and $G_R(z)$ and g_R are the length-dependent Raman gain and the Raman-gain coefficient (see Fig. 7), respectively. P_{P0} is the pump power launched at $z = L$, and L_{eff} is the effective fiber length at the pump wavelength. α_s is the fiber loss coefficient at signal wavelengths. The on/off gain is explained in Fig. 40 (also see Fig. 20).

Depending on fiber characteristics (loss, mode-field diameter, CD) and patch and splice losses between fiber segments, typical value of $F_{\text{Eq}} \approx -2.5$ dB and $G_{\text{on/off}} = 15$ –20 dB result. The total noise figure F_{tot} of combined EDFA and Raman pre-amplification is given by:

$$F_{\text{tot}} = F_{\text{Eq}} + [(F_{\text{EDFA}}/\alpha_c) - 1]/G_{\text{on/off}} \quad (18)$$

Here, α_c is the (coupling) loss between the fiber output (at the end of L_2) and the EDFA.

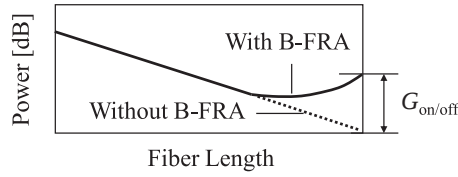


Fig. 40 On/off Raman gain. B-FRA: backward-pumped FRA.

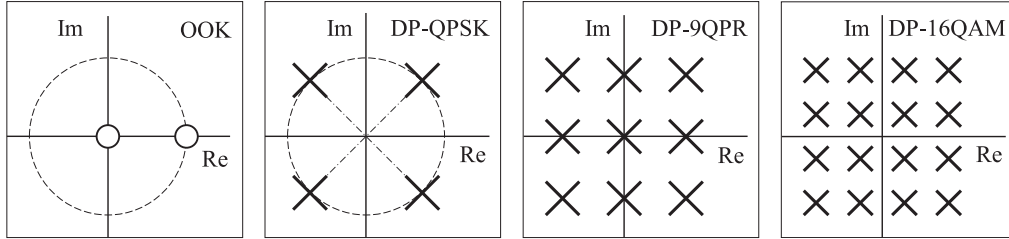


Fig. 41 Some relevant fiber-optic modulation schemes' constellation diagrams.

The receive-end OSNR in very long multi-span transmission can be improved by Raman-only amplification, using forward- and backward-pumped distributed FRAs. Noise figures for forward-pumped FRA can achieve $F_{N,FRA} = 4\text{--}5$ dB (Premaratne, 2004).

Modulation of WDM Signals

Modulation Formats

Typically, WDM channels up to 100 Gb/s per channel are modulated individually, without considering other channels. Starting with 100 Gb/s, the respective capacity has also been split onto several *sub-carriers* which together form so-called WDM *super-channels*. Sub-carriers are modulated via a common entity.

Per WDM channel or sub-carrier, all parameters of the respective laser can be used for modulation. The simplest (and oldest) variant is *intensity modulation*. It can be detected non-coherently in direct-detection receivers, leading to Intensity Modulation with Direct Detection (IM/DD). For binary modulation, this is also called On-Off-Keying (OOK). Currently, a variant with four amplitude levels, Pulse-Amplitude Modulation-4 (PAM-4), is under consideration for shorter-reach 400-Gb/s WDM super-channel transport with direct detection.

If the instantaneous frequency of the laser is modulated, this is called *Frequency-Shift Keying*. It does not play an important role in WDM transmission.

If the carrier phase is chosen to carry the information, this is called *Phase-Shift Keying* (PSK). PSK has been used in commercial WDM system in the binary and quaternary, differentially modulated versions (DBPSK, DQPSK). DBPSK is still relevant for ultra-long-haul WDM. DQPSK is identical to differentially encoded 4QAM.

The combination of (multi-level) amplitude or intensity modulation and PSK leads to *Quadrature Amplitude Modulation* (QAM). Today, coherently detected QAM dominates ultra-high-speed WDM transmission.

In coherent detection, receivers must detect signals in orthogonal polarization planes, because the PSPs of the receive signal are generally not known. Then, two orthogonal polarizations can also be used for independent modulation and associated capacity increase. This is often called *Dual-Polarization* (DP) modulation.

Any modulation can use *full-response* or *partial-response* pulses. For Partial Response (PR), the pulse spreads over several symbols, causing intentional Intersymbol Interference (ISI). PR leads to smaller signal bandwidth and better CD tolerance. The added ISI can be eliminated with receive-end Maximum-Likelihood Sequence Estimation. The simplest form of PR in WDM transmission is Optical Duobinary (ODB), the PR variant of OOK. PR can be extended to complex symbol constellations, leading to *M*-ary Quadrature Partial Response (QPR).

Often, (de-) modulation is characterized graphically in constellation diagrams where the respective Inphase (I) and Quadrature (Q) components of each symbol state are shown. For DP modulation, this can be done for both polarizations. In Fig. 41, some examples are shown. Crosses (×) indicate DP modulation.

Independent from full vs. partial response, pulse or spectral shaping can be added. In some long-haul systems, this was done by Return-to-Zero (RZ) pulse shaping to increase tolerance against nonlinearity. Different RZ duty cycles were implemented, e.g., 33%, 50%, and 67% (called RZ33, RZ50 and RZ67, respectively). All RZ variants increase bandwidth over Non-Return-to-Zero (NRZ). Some resulting optical power spectra, normalized to the bit rate, are shown in Fig. 42.

Pulse (or spectral) shaping can also be done in order to decrease the bandwidth of (sub-) carriers. The intention here is to allow denser (sub-) carrier spacing, in order to increase spectral efficiency. This approach is sometimes referred to as Nyquist WDM (Bosco et al., 2010). Spectral shaping for a super-channel with four sub-carriers is visualized in Fig. 43.

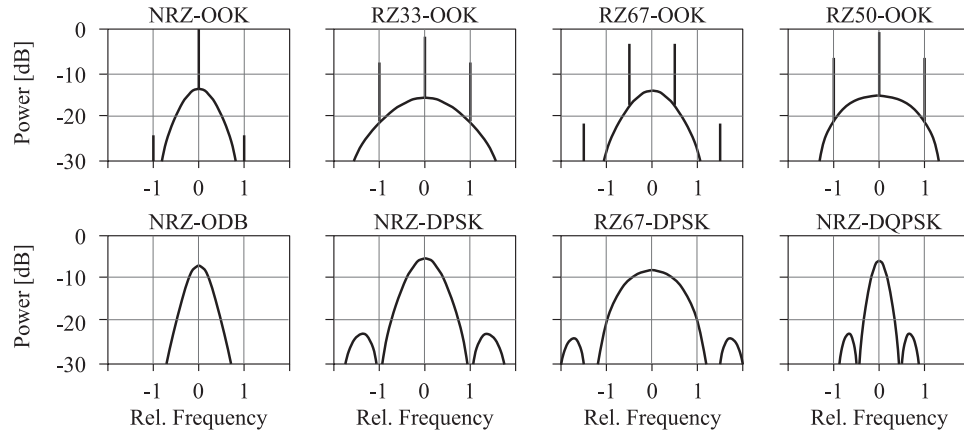


Fig. 42 Effect of modulation, pulse shaping and duty cycle on optical bandwidth.

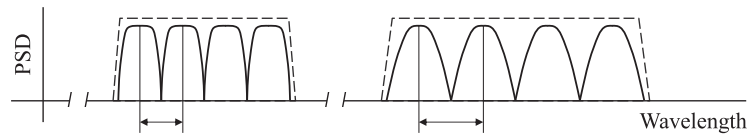


Fig. 43 Spectral shaping of sub-carriers to allow denser spacing.

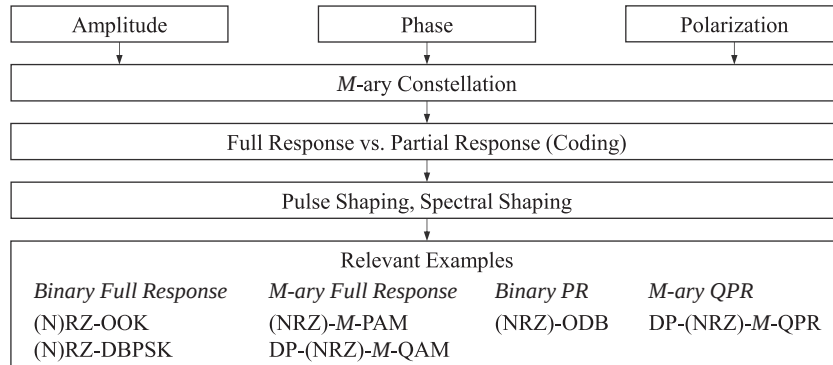


Fig. 44 Overview on fiber-optic modulation and pulse-shaping schemes.

The classification of modulation schemes suitable and relevant for WDM is given in [Fig. 44](#), together with (commercially) relevant past and present examples.

Up to and including 10-Gb/s WDM transmission, NRZ-OOK (IM/DD) has been most important. For 40 Gb/s, RZ-OOK, ODB and DPSK have been used, and first DP-DQPSK systems been introduced. For 100 Gb/s per channel and beyond, long-haul systems are based on DP-QAM with intradyne detection. Shorter reach, e.g., for data-center interconnects, can also be based on ODB or PAM-4 with multiple sub-carriers.

Achievable Bit-Error Rates

Closed analytical calculation of the receive-end Bit Error Rate (BER) P_B of a WDM transmission channel including fiber non-linearity, PMD, CD, crosstalk and noise is not possible. Hence, a simplified AWGN (Added White Gaussian Noise) system model must be used, e.g., [Agrawal \(1992\)](#), [Linke and Gnauck \(1988\)](#). All distorting effects must be considered via penalties on the receive-end OSNR.

For coherent detection, the analytical analysis leads to the Q-function which can be derived from the Complementary Error Function, erfc :

$$\text{erfc}(x) = 2Q(\sqrt{2}x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} dy \quad (19)$$

Following, the AWGN-channel BER is derived for relevant (de-) modulation techniques.

Intensity modulation

Intensity modulation can be split into *binary modulation* (OOK), and multi-level modulation. The latter is called Pulse-Amplitude Modulation (PAM), with PAM-4 currently being the most relevant variant for WDM. OOK is mostly detected directly. Then, the BER results as:

$$P_B = \frac{1}{2} \exp\left(-\frac{E_b}{4N_0}\right) \quad (20)$$

E_b is the bit energy (in J or Ws), N_0 is the noise-power density (in W/Hz). The simplified block diagram of an NRZ-OOK system is shown Fig. 45. The transmitter consists of a continuous-wave Laser Diode (LD) and an external modulator. The receiver consists of pre-amplifier, Optical Band-Pass Filter (OBPF) and direct-detection receiver. The OBPF is necessary for WDM channel separation and as a noise-bandwidth limiter. The direct-detection receiver consists of a Sample-and-Hold (S + H) unit which approximates a matched filter, followed by a decision unit.

PAM-4 can basically use the same optical frontend as OOK. Added effort results from the 4-level signal generation (including serial-parallel (S/P) and parallel-serial (P/S) conversion), and the necessity to equalize the received signal, refer to Fig. 46.

PAM-4 is currently under investigation for WDM applications with ~ 100 km reach, e.g., for data-center interconnects. For 400-Gb/s transport, super-channels with 8×56 Gb/s can be used.

Phase-shift keying

PSK has first been considered for ultra-long-haul WDM transmission in the 1990s because of its superior OSNR. For homodyne detection, the BER of BPSK is given by:

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \quad (21)$$

This is the best BER achievable by the modulation formats discussed in Section Modulation Formats.

Early commercial BPSK systems were based on non-coherent detection with differential precoding and delay demodulation (DBPSK). Its BER is:

$$P_B = \frac{1}{2} \exp\left(-\frac{E_b}{2N_0}\right) \quad (22)$$

This BER is worse as compared to homodyne BPSK by ~ 3.5 dB.

The most relevant implementation of *multi-level PSK* is (Differential) QPSK. It has been used in early non-coherent systems, and today is the basis of the quasi-standardized intradyne 100 Gb/s transmission. The BERs for non-coherent and homodyne detection are 3 dB worse compared to BPSK, for the same symbol energies E_s .

Quadrature amplitude modulation

The term QAM is often applied to rectangular symbol constellations (see Fig. 41), however, other constellations based on multiple PSK stars or even non-symmetric constellations exist.

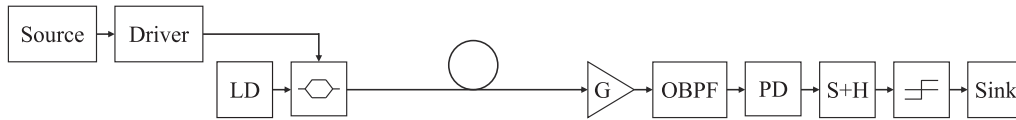


Fig. 45 NRZ-OOK with external modulation and Direct Detection (DD).

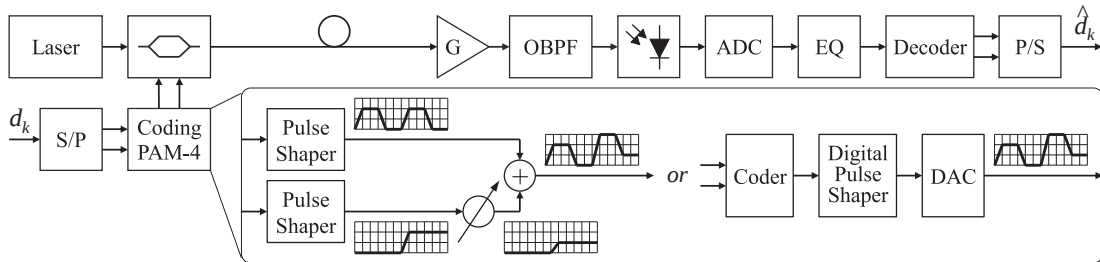


Fig. 46 PAM-4 system with two alternatives for 4-level signal generation.

For rectangular constellations and homodyne detection, the *symbol error rate* is given by:

$$P_S(M) = 4 \left(1 - \frac{1}{\sqrt{M}} \right) Q \left(\sqrt{\frac{3 \log_2 M}{M-1} \frac{E_b}{N_0}} \right) \quad (23)$$

The BER is given by $P_B \approx P_S / \log_2 M$ for $P_S \ll 1$.

Today, long-haul ultra-high-speed WDM relies on Dual-Polarization QAM with digital **intradyn**e detection. For 100-Gb/s transport, this leads to DP-4QAM (DP-QPSK). For bit rates ≥ 200 Gb/s, DP-16QAM can be used, eventually with a multi-carrier approach. For tight sub-carrier spacing, this may be complemented by Partial-Response Signaling (QPR) or other spectral shaping. Typically, differential encoding is used to avoid error propagation.

Intradyn detection is based on tuning the local laser such that the resulting intermediate frequency is close to zero. Then, analog-to-digital conversion is performed, and exact phase tracking is done in the digital domain. This allows implementing digital dispersion-compensating filters which lead to almost zero penalty with regard to polarization and CD effects.

Coherent systems require polarization-diverse receivers. Therefore, they make use of polarization multiplexing. The resulting configuration for a Dual-Carrier system is shown in [Fig. 47](#). For single carriers, a similar configuration was reported in [Savory et al. \(2007\)](#).

Each transmit laser is split into orthogonal polarizations by means of a polarization beam splitter (PBS). Both polarization signals are independently modulated and re-combined by a polarization beam combiner (PBC). At the receiver, the input signal is split into orthogonally polarized signals by another PBS. The local laser signal is also split and then combined with the input signal by means of two 90° hybrids. Each 90° hybrid has dual output ports for the respective inphase and quadrature components. The output signals are detected by four balanced receivers. These are followed by fast Analog-to-Digital Converters (ADC) which feed the digital signal processing (DSP). [Fig. 48](#) shows the DSP in more detail.

After detection and sampling at approximate Nyquist rate, the I and Q components have orthogonal but arbitrary polarization planes each. The four components are then processed in several filter stages. In a first stage, bulk CD compensation is performed. This stage can be complemented by Nonlinear Compensation (NLC). CD compensation is based on an n -tap Feed-Forward Equalizer (FFE). The upper bound for the tap number at a Baud rate of B GBd is given by $0.032 \cdot B^2$ per 1000 ps/nm of chromatic dispersion ([Savory et al., 2007](#)). At 28 GBd (112-Gb/s DP-QPSK), this translates to 25 taps per 1000 ps/nm.

NLC primarily tackles intra-channel effects (SPM). Inter-channel effects (XPM, FWM) require equalizers which are currently too complex. Linear and nonlinear intra-channel compensation can be done iteratively, switching between time and frequency domain by means of FFT and IFFT, respectively, see [Fig. 49](#). This approach is called Backward Propagation. The improvement in launch power (and therefore, OSNR) of NLC typically is limited to 1–1.5 dB.

In the next filter stage, Clock Data Recovery (CDR) and re-sampling to 2 samples/symbol are performed. CDR is based on a digital filter-and-square timing recovery. It also performs re-sampling to 2 samples/symbol.

PMD and residual-CD equalization, and polarization recovery (demultiplexing) is performed in a Multiple-Input Multiple-Output (MIMO) filter which follows re-timing. The filter tap weights are usually optimized using blind adaptation. This can be done with the Constant-Modulus Algorithm (CMA) ([Johnson et al., 1998](#)). The CMA can equalize M -ary PSK signals with $M \geq 4$ only. Derivatives for BPSK, QAM and QPR exist.

The last filter stage performs Carrier Phase Estimation (CPE), i.e., the digital phase tracking. CPE for QPSK can be done with 4th-power Viterbi-and-Viterbi recovery. For BPSK, 2nd-power estimation can be used, whereas for (square) QAM, modifications are required. These can consist of considering symbols of equal magnitude separately.

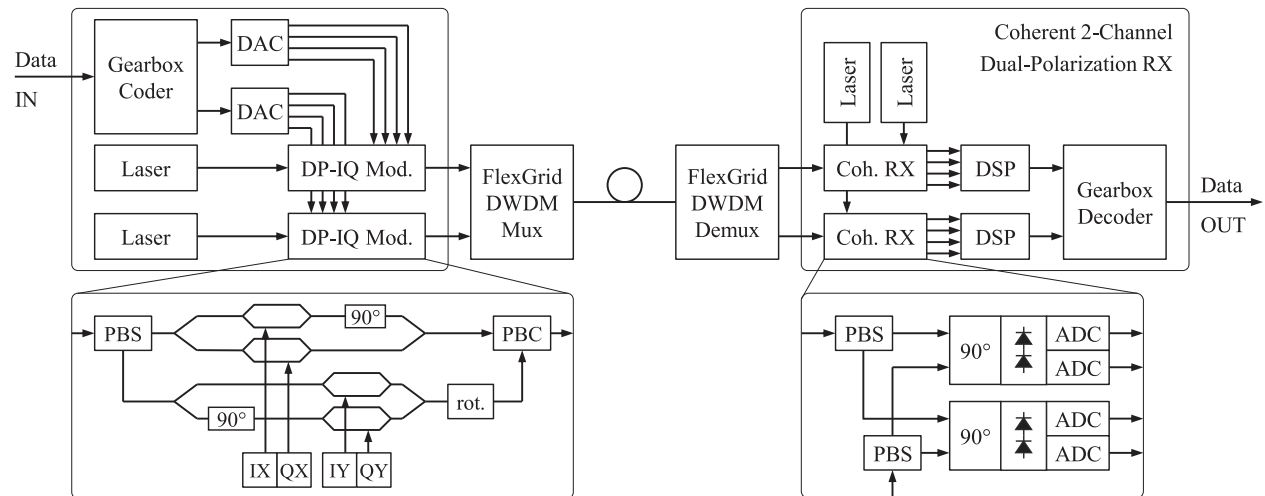


Fig. 47 Coherent intradyne Dual-Carrier DP-QAM (DP-QPSK) transmission.

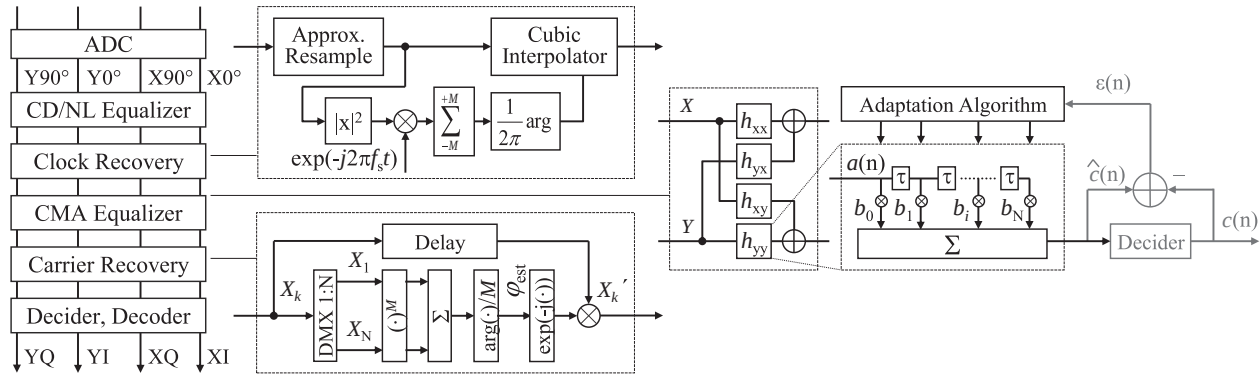


Fig. 48 Coherent intradyne receiver: digital realization.

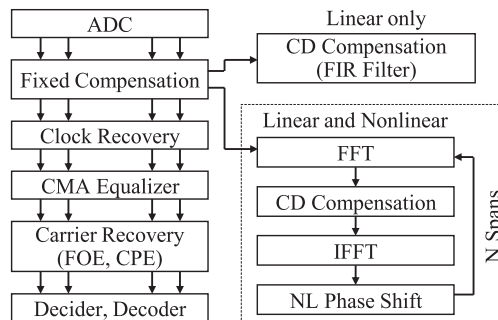


Fig. 49 Linear-only or linear plus nonlinear compensation (backward propagation).

Partial response signaling

Optical Duobinary, ODB, is the simplest form of *partial response* signaling in WDM transmission. It intentionally produces *known* ISI in order to achieve higher spectral efficiency. The ODB spectrum is compared to OOK and PSK in Fig. 42. PR signaling has been studied since the 1960s (Kobayashi and Tang, 1971).

ODB encoding can be achieved by a delay-and-add filter, followed by a Low-Pass Filter (LPF). The LPF bandwidth is $0.5 - 0.75 \cdot R_B$ (R_B the bit rate). This encoder can be replaced by a sole LPF with bandwidth in the range of $\sim 0.28 \cdot R_B$ (so-called *filtered* Optical Duobinary). The ISI introduced at the transmitter can be unraveled at the receiver by differential decoding. Since this can produce error propagation, differential *precoding* at the transmitter is preferred, allowing symbol-by-symbol detection. The resulting transmitter is shown in Fig. 50.

With direct detection, the BER of ODBis in the range of the BER of OOK, depending on receiver implementation. ODB can also be detected coherently (Lyubomirsky, 2006). The BER is then given by:

$$P_B \approx \frac{3}{2} Q\left(\frac{\pi}{2} \sqrt{N_P}\right) \quad (24)$$

$N_p = E_b/hf_0$ is the average per-bit photon number. With optimized threshold, homodyne ODB achieves sensitivity (at BER = 10^{-9}) of 15 photons/bit (compare to 9 photons/bit for BPSK).

ODB can be used for cost-effective medium-reach WDM transport of 100-Gb/s signals. Then, four DWDM sub-carriers carrying ~ 28 Gb/s each are used, see Fig. 51.

Reach according to [Fig. 51](#) is OSNR-limited and in the range of ~ 500 km. The sub-carrier spacing is 50 or 25 GHz. As compared to 50 GHz, 25 GHz leads to a reach penalty which is caused by higher sub-carrier crosstalk and filter effects. CD tolerance is in the range of ± 300 ps/nm, mandating optical CD compensation (Section Chromatic-Dispersion Compensation) ([ADVA, 2016](#)).

Partial response coding can be applied to complex signals, leading to *Quadrature PR* (QPR). QPR is under frequent study for spectrally efficient WDM transmission (Lyubomirsky, 2010). Here, 9QPR has gained relevance so far. 9QPR has a symmetric 3×3 constellation, see Fig. 41.

The BER for coherent homodyne M -ary QPR ($M=9, 25, 49, \dots$) is given by (Smith, 2003):

$$P_{B,M-QPR} = \frac{4}{\log_2 L} \left(1 - \frac{1}{L^2}\right) Q \left[\frac{\pi}{2} \sqrt{\frac{3E_b(\log_2 L)}{(L^2 - 1)N_0}} \right] \quad (25)$$

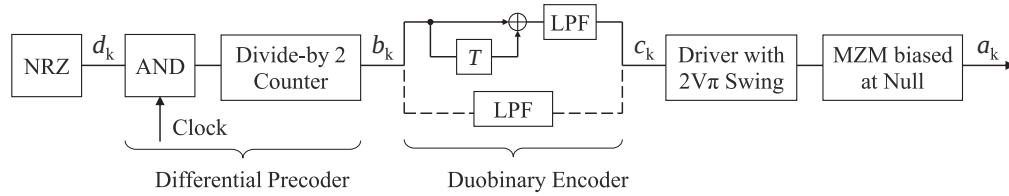


Fig. 50 Optical Duobinary coding.

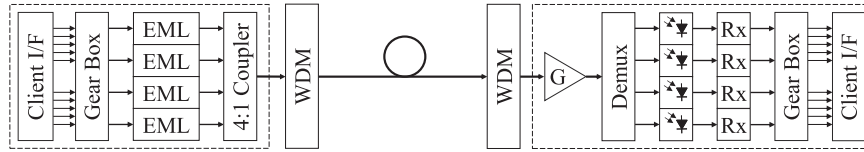


Fig. 51 High-speed 4-subcarrier Duobinary system with Externally Modulated Lasers (EML) and direct detection. Gear box is the synchronization circuit needed for 100-Gb/s client interfaces.

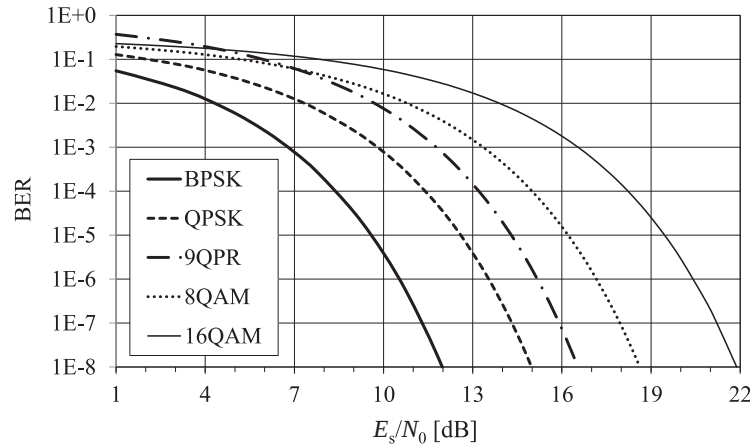


Fig. 52 BER of relevant modulation schemes (homodyne detection) for same symbol energy.

here, $L = \sqrt{M}$. 9QPR leads to ~ 2 dB penalty compared to QPSK, and a gain of ~ 2 dB over 16QAM, for coherent detection and same bit energy. This does not yet take an optimum receiver into consideration. For QPR, this is an MLSE which ideally can eliminate all penalty caused by PR coding. Simplified MLSEs have been shown for ODB and 9QPR. They are called Ambiguity Zone Detectors (AZD) (Kobayashi and Tang, 1971). AZD can improve the penalty for 9QPR against QPSK to ~ 1 dB at little implementation effort.

Comparison of modulation formats

In Fig. 52, bit-error rates are shown for the coherent modulation schemes that are most relevant for WDM transmission today. All schemes are normalized to the same symbol energy. The incoherent schemes DBPSK, OOK and ODB are not shown here. The resulting DBPSK BER curve is located very near to (~ 0.5 dB on the right-hand side of) the QPSK curve. OOK BER is 3 dB worse compared to DBPSK, and ODB BER is in the range of OOK BER.

These results imply ideal reception. Real systems (specially the receivers) lead to penalties over the numbers given in this chapter. These penalties are in the range of ≥ 3 dB, depending on implementation details.

System Realization

WDM in Metro and Access Networks

Large telecommunications networks are hierarchically organized. Residential and business access is based on wireless (2G, 3G, 4G, 5G upcoming) and wireline (fiber point-to-point, Passive Optical Networks (PON), Hybrid Fiber-Coaxial (HFC), copper twisted-pair) technologies. Access concentrators (base stations, PON Optical Line Terminations, Multi-Service Access Nodes, etc.) are

backhauled to aggregation sites of a first level. These sites are often referred to as Local eXchanges (LX) or Central Offices (CO). This backhaul is based on CWDM or DWDM. There is also a strong trend towards mobile *fronthaul* in LTE-Advanced (4G, 5G) networks. Here, several Baseband Units (BBU) are concentrated in BBU Hotels (BBUH) and connected to their respective antennas via digital high-speed links. Since these links must not be statistically multiplexed, and have tight latency and jitter requirements, they must run via point-to-point fiber or WDM channels.

Several LX can be connected, mostly via DWDM, to metropolitan-area core Points-of-Presence (PoPs). PoPs can accommodate aggregation switches of a secondary level, and they are connected via protected DWDM rings or mesh networks. Typically, the metro core network is connected to national and international backbones via two redundant PoPs (which accommodate core routers, Broadband Remote Access Servers (BRAS), etc.). The backbones use high-capacity, long-haul DWDM. An example network is shown in Fig. 53.

Real networks can deviate from Fig. 53. They may have less aggregation levels. They can be based on differing combinations of ring, mesh or point-to-point structures, and they can also use legacy technologies like SONET/SDH which are not shown here. However, most networks have in common that they use some sort of MPLS-over-WDM in the core, and some sort of Ethernet-over-WDM in the backhaul.

Metropolitan-area networks have been relevant applications for DWDM from the 1990s. Metro and regional networks are based on multi-span WDM systems. Many of these networks rely on *ring* architectures. One reason for (WDM) rings is that they require comparatively few fibers for redundant (i.e., protected) connections between a given number of sites. Rings are also relatively easy to manage (Maier et al., 2002). A metro DWDM ring system is depicted in Fig. 54.

WDM nodes consist of add/drop filters or ROADMs, amplifiers, filters and termination for an Optical Supervisory Channel (OSC, which is part of the Data Communications Network), and optional CD compensation. The static filter structure shown in Fig. 54 is increasingly being replaced by ROADMs. Main advantages of ROADMs include the capabilities of single-channel add/drop, which increases the number of logical connections without wavelength conversions (regenerations), and remote reconfiguration.

So far, most metro WDM systems make use of CD compensation. Main reason is the mixture of services, including transparent carriers' carrier services, which does not allow the exclusive use of coherent systems with in-built equalization. Due to distance restrictions, effects of accumulated nonlinearity are smaller than in long-haul.

Parameters for metro WDM link design are summarized in Table 5 (Grobe and Eiselt, 2014).

A difficulty in metro WDM link design is the potential mixture of services and bit rates, and the related transceiver technologies and parameters. On certain links, it may be impossible to perfectly align receiver sensitivities, power levels or CD compensation requirements, which leads to additional penalties.

WDM in Data Center Interconnects

Relevant WDM applications are reach extensions for Storage Area Networks (SAN), and data center interconnects. Often, these are point-to-point applications which require very high WDM capacity over distances of <100 km.

Disk mirroring is primarily related to disaster recovery (DeCusatis, 2008). It provides redundancy for those data centers where failure/unavailability will cause significant costs. Disk mirroring can be performed *synchronously* or *asynchronously*. Synchronous disk mirroring means that the remote server can take on action immediately *without any interruption*. It has strict latency requirements which limit maximum distance between the data centers. Asynchronous mirroring allows some *limited* data loss or application interruption. It has relaxed latency requirements. In Table 6, typical latencies of WDM system components are listed.

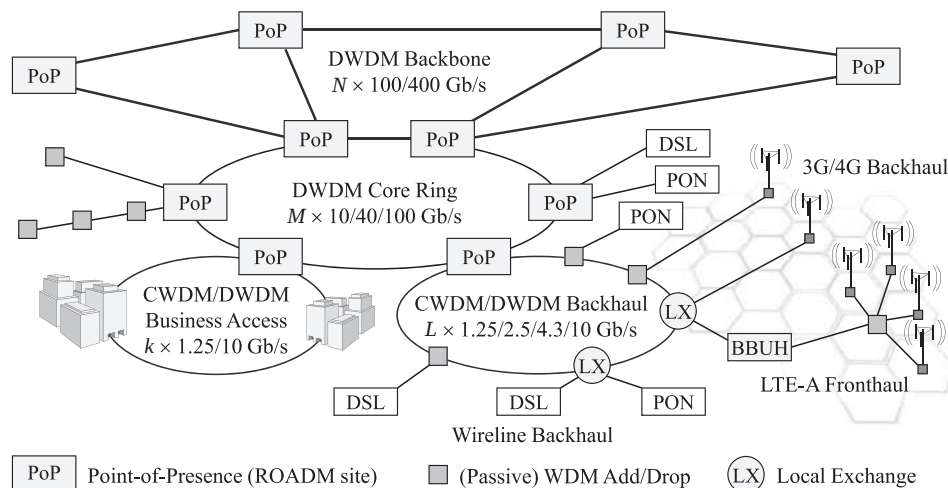


Fig. 53 Hierarchical fiber-optic telecommunications network. BBUH: Baseband Unit Hotel.

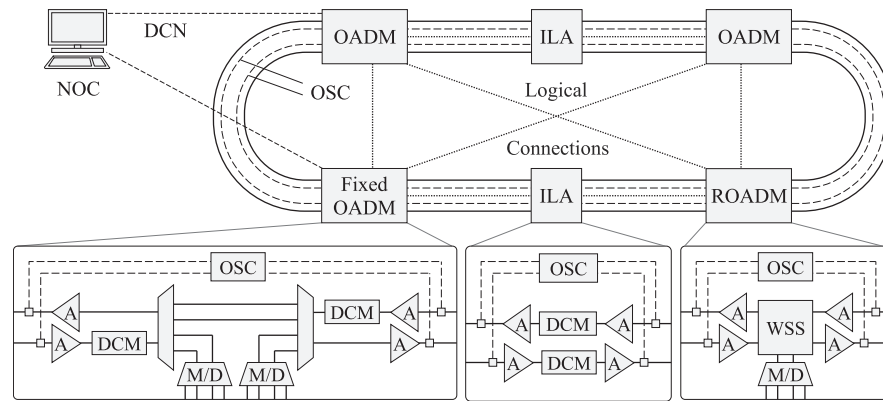


Fig. 54 Metro WDM ring. A: amplifier, DCM: Dispersion Compensating Module, DCN: Data Communications Network, ILA: In-Line Amplifier, M/D: Mux/Demux, NOC: Network Operations Center, OSC: Optical Supervisory Channel, WSS: Wavelength-Selective Switch.

Table 5 Parameters of metro and regional WDM link design

Parameter	Value
End-of-Life Penalty (Patches etc.) (dB)	2–3
Average Fiber Loss (dB/km)	0.25 (regio)–0.35 (metro)
Optical Path Penalty for CD, PMD, Nonlinearity (dB)	1 (2G5)–2.5 (10G, 40G)
Insertion Loss of WDM Filter plus OSC, east or west (dB)	7–10
Filtering Penalty for serial 40G Services for 50-GHz System	0.5 dB/Node
Typ. OSNR Requirement 2G5/10G/40G/coh. 100G (dB)	18 (no FEC) / 14/17/15 (FEC)
Typ. Sensitivity 2G5/10G/serial 40G/coh. 100G (dBm)	–28/–22/–20/–20
Typ. guaranteed Launch-Power Window (dB)	0–4
Typ. Amplifier Gain/Noise Figure (dB)	20–25/5.5–6.5

Reproduced from Grobe, K., Eiselt, M.H., Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, NJ: John Wiley & Sons; 2014. ISBN 978-0-470-62302-2.

Table 6 Latencies of different transmission systems

Transmission system	Latency/(group) delay
Transparent WDM transponder w/o framing	<10 ns
TDM multiplexing with OTN framing	5–20 μ s
FEC (10 Gb/s, 100 Gb/s)	5–10 μ s
EDFA (single-stage)	~200 ns
100 km SSMF round-trip	~1 ms

Reproduced from Grobe, K., Eiselt, M.H., Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, NJ: John Wiley & Sons; 2014. ISBN 978-0-470-62302-2.

In data centers, often Ethernet is used for the LAN, and Fiber Channel (FC) is used for the SAN. In addition, InfiniBand may be used for high-performance compute clusters (DeCusatis, 2008). In highly available data centers, key components are duplicated per site, and sites are connected for even higher availability, and to create virtual servers. Data-center interconnection then uses two separate WDM links, see Fig. 55.

Some applications require very high availability. In such mission-critical applications, all major components are duplicated, including the transmission paths. If necessary, more than two fiber connections (ducts) can be used. Options for redundant WDM transport are shown in Fig. 56.

The options shown here can be combined for highest path availability in the range of 99.9999%.

Long-Haul WDM Systems

In long-haul WDM transport, all system impairments (linear and nonlinear effects, noise) have to be considered. The influence of these effects varies with data rate, modulation format, fiber type, dispersion management, and the type of optical amplification.

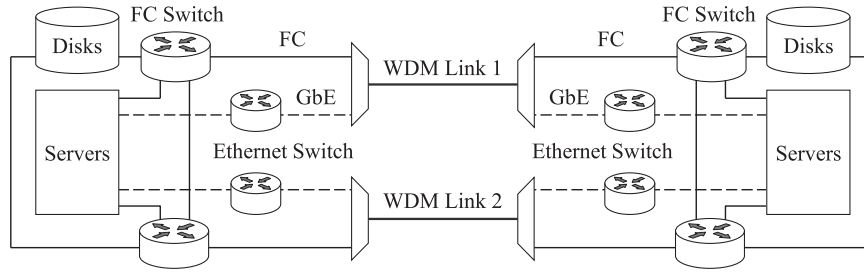


Fig. 55 Data-center interconnect via Ethernet (LAN) and FC (SAN) links.

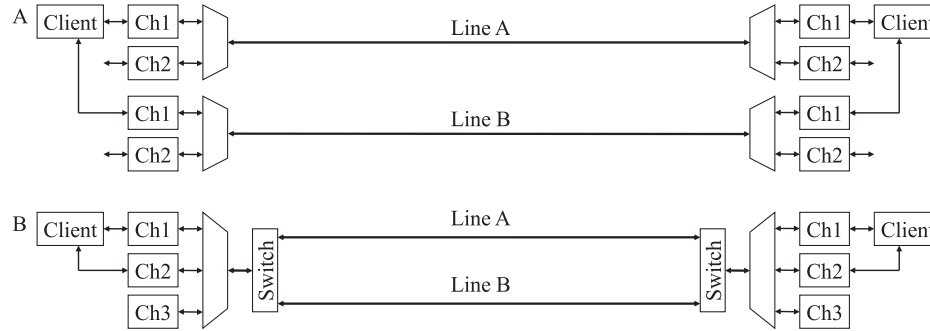


Fig. 56 Redundancy in data-center interconnects. A: Path protection, B: line protection.

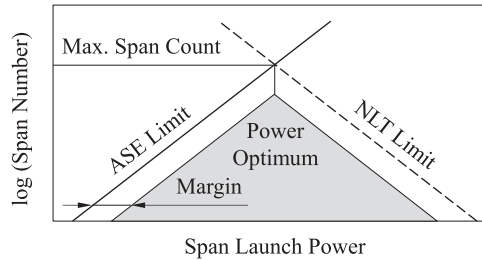


Fig. 57 The “working zone” of optical-layer design.

Since simulation of nonlinear fiber transmission causes immense calculation effort, there is the need for engineering rules enabling quick first-order performance estimates and identification of limiting effects in multi-path meshed networks. Each potential channel has to be in the power range above noise limitation and below severe nonlinearity accumulation. This limits the valid power-level range, see [Fig. 57](#).

One relevant system limitation is given by the *Nonlinear Tolerance* (NLT). The NLT law $\log(N_{\text{span}}) + P_{\text{Launch}} \text{ (dBm)} = \text{NLT (dBm)} = \text{const}$ holds for different fiber types, data rates, modulation formats, and channel spacings ([Mohs et al., 2000](#)). To achieve high transmission distances, the per-span launch power has to be kept low. This corresponds to shorter amplifier spans, leading to better noise figures at the cost of more amplifiers. Typical resulting reach is shown for (mixed) 10G/100G signals in [Fig. 58](#).

For [Fig. 58](#), fiber loss of 0.25 dB/km and a patch-panel and repair margin of 1 dB each were assumed for every span. Shorter spans improve total reach to a certain point only. Beyond this point, accumulated amplifier tilt and ripple lead to negative effects.

The loss of any components along a given WDM link is not constant. Spectral fluctuations induced by random loss non-uniformities cannot be fully compensated by optical amplification and accumulate as power ripples. These ripples arise from filter-loss and amplifier-gain non-uniformities, combined with transmitter launch-power variations. This leads to typical channel-power spread as shown in [Fig. 59](#).

High-power channels are subject to stronger signal distortion by nonlinear effects, whereas the BER of the lowest-power channels is impacted by the low OSNR and/or low power at the receiver. *Power management* aims at controlling the WDM channel power levels at particular points in the networks to ensure optimum reach for all channels. It can be performed in

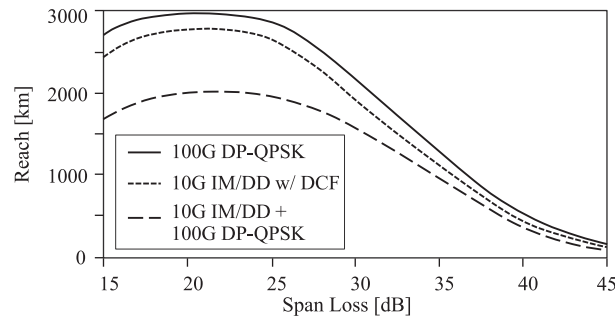


Fig. 58 Transparent reach for different span losses.

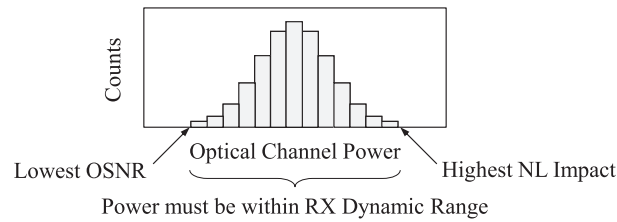


Fig. 59 Histogram of WDM channel power.

ROADMs which are located after every N spans, with $N=5-10$ for terrestrial networks. This is typically supported by a distributed control plane.

The minimum OSNR calculated for each link including all OSNR penalties must be compared with the OSNR required by any of the receivers used. The system limit is reached if the OSNR of the worst link, including all penalties, is lower than the OSNR required at the receiver. After that, the optical signal must be electrically regenerated. This system limit was already shown in [Fig. 38](#), where OSNR penalties (in dB) are assumed to grow linearly with the span count.

Mixed 10 Gb/s/100 Gb/s Design

In certain WDM networks, combined 10-Gb/s IM/DD and coherent 100 Gb/s transport is required. This causes potential problems:

- Coherent 100 Gb/s do not require optical CD compensation, whereas most 10-Gb/s IM/DD systems do
- Coherent 100 Gb/s systems can be subject to reach penalties when DCF are used. This results from increased XPM inside the DCF, and the fact that DCF compensate the channel walk-off (leading to stronger XPM in the transmission fibers).
- Phase-modulated signals suffer penalties caused by IM/DD signals. These are generated by XPM-induced Nonlinear Phase Noise (X-NLPN).

The DCF-related penalty can partly be reduced by carefully adjusting the DCF launch power. Further methods for reducing the DCF-induced XPM penalty are:

- Separate 10-Gb/s and 100 Gb/s transport on dedicated photonic layers (coherent overlay)
- Use CD compensators other than DCF, i.e., channelized FBGs

Reduction mechanisms for IM/DD-induced X-NLPN include:

- Reducing power of 10 Gb/s channels compared to 100 Gb/s channels.
- Using channelized FBGs instead of DCFs to allow walk-off between 10 Gb/s and 100 Gb/s channels.
- Increasing the spectral separation between 10 Gb/s and 100 Gb/s channels (guard band).

The last two methods decorrelate the 10 Gb/s from the 100 Gb/s signals along their paths.

Another method of partial nonlinearity suppression is based on using IM/DD with inbuilt pre-equalization. It avoids DCF and the related XPM, but does not avoid X-NLPN.

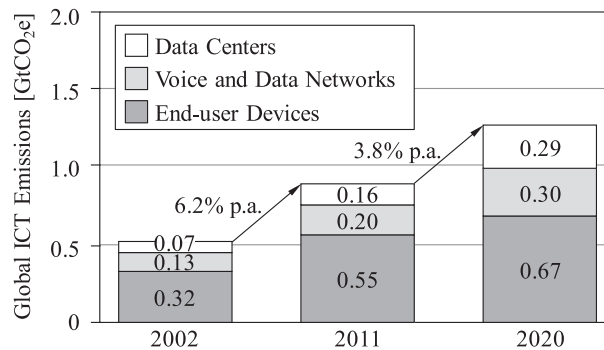
In addition, the Carrier Phase Estimation (CPE) in coherent receivers can be configured to achieve optimum results in the presence of nonlinearity.

For all methods, certain limitations or costs result. This is summarized in [Table 7](#).

Beyond 100 Gb/s, further future functionality should be considered, e.g., super-channels for coherent 400-Gb/s and 1-Tb/s systems. For optimum performance, these systems require coherent-only, uncompensated photonic layers. This excludes both, coexistence scenarios based on optical CD compensation, and the general use of any IM (DD) technology.

Table 7 Cost of 10-Gb/s and 100-Gb/s coexistence

Method	Generalized cost	Potential advantages
Coherent overlay	Almost doubles cost of photonic layers, especially fibers.	Only safe solution for 400+ Gb/s super-channels
IM/DD with inbuilt EDC	10-Gb/s EDC does not reduce X-NLPN, and hence decreases 100-Gb/s reach.	Avoids compensated photonic layer and DCF-incurred penalties (XPM)
Use of FBGs instead of DCFs	Phase-ripple penalty to IM/DD channels (~1 dB penalty after 8–10 devices). <i>Channelized filters contradict FlexGrid</i>	Significantly reduces X-NLPN from IM/DD. Eliminates XPM
Guard bands between 10-Gb/s and 100-Gb/s channels	Decrease of system and network capacity, added network-planning complexity.	Can achieve almost optimum performance for IM/DD and QPSK. After gradual growth to all-100-Gb/s, the guard band can be eliminated.
Power-level reduction for IM/DD	Reduces IM/DD performance. Requires CD compensation for IM/DD (unless IM/DD uses inbuilt EDC).	Can lead to less or negligible IM/DD-induced penalty for QPSK.

**Fig. 60** Global ICT carbon footprint. Reproduced from Mohs, G., *et al.*, 2000. Maximum link length versus data rate for SPM limited systems. In: Proceedings European Conference on Optical Communication (ECOC) 2000, Munich.

WDM Sustainability

In the recent years, consideration of the environmental impact of the global Information and Communications Technologies (ICT) sector has become a major requirement by large network operators and legislation.

Global ICT produces a significant amount of the global CO₂ footprint. In 2007, the ICT portion of global CO₂ production was in the range of 2% (GeSI, 2020, 2012). This amount is projected to significantly increase over the next years, see Fig. 60.

From the global ICT CO₂ footprint, telecommunications infrastructure is in the range of ~25%. WDM transport is responsible for ~7% of network power consumption. Nonetheless, increased energy efficiency is relevant due to the exponentially increasing bandwidths.

The analysis of various environmental impact parameters of any systems is known as Life-Cycle Assessment (LCA). LCA according to ISO14040/14044 considers all relevant phases of systems or products, from extraction of raw materials via manufacture, distribution, the use phase to end of life (the latter preferably meaning recycling or partial reuse). Various environmental-impact parameters like Global Warming Potential (which is basically coupled to CO₂ footprint) can be derived. This requires detailed knowledge of the respective parameters of any components used, plus knowledge of the contributions from any logistics etc.

Fig. 61 shows an LCA of a WDM system in a simple point-to-point pre-amplified configuration. The respective system supports up to 80 channels and uses 10-Gb/s channel cards. Other WDM configurations (rings, 100-Gb/s cards,...) show similar results.

It can be seen that various environmental-impact parameters, in particular GWP and ODP, are dominated by the use phase. This is driven by energy consumption, assuming electricity mix of the year 2016. In this LCA, a use phase of 8 years has been assumed. The dominance of the use phase in LCA underpins the necessity of highest WDM energy efficiency.

Energy efficiency is ranked in so-called Telecommunications Energy-Efficiency Ratings (TEER). Definitions for WDM can be found in the ANSI Standard ATIS-0600015.02.2009 (ANSI/SCTE, 2009) and the Ecology Guidelines for the ICT Industry (ECO ICT 2017). In ECO ICT (2017), the relevant metric (Figure of Merit, FoM) is given as maximum throughput (in Gb/s) divided by average power consumption (in Watts). Average power consumption is defined as linear average of power consumption of a

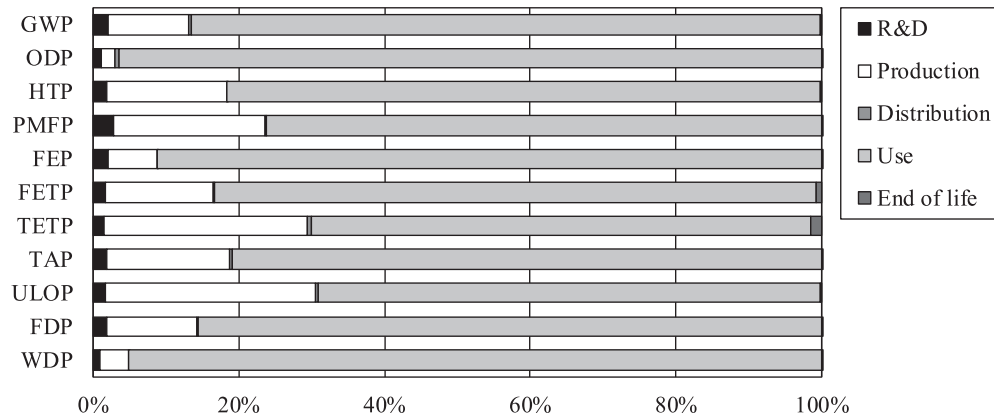


Fig. 61 LCA of a WDM system with 8 years of use phase. The diagram shows the potentials for Global Warming (GWP), Ozone Depletion (ODP), Human Toxicity (HTP), Particulate Matter Formation (PMFP), Freshwater Eutrophication (FEP), Freshwater EcoToxicity (FETP), Terrestrial EcoToxicity (TETP), Terrestrial Acidification (TAP), Urban Land Occupation (ULOP), Fossil Depletion (FDP) and Water Depletion (WDP), respectively.

Table 8 TEEER for WDM system FSP 3000 according to ECO ICT (2017).

ECO ICT Council TEER		Ref. –Config. 2 DWDM w/ROADM, 80 × 100G					
		100G old (2013)		100G new (2015)		400G (2016)	
PC @ 1 Wavelength (W)		1296		725		597	
PC @ full Wavelengths (W)		13,863		8118		6719	
Max. Throughput (Gbps)		8000		8000		8000	
N.R. (Gbps/W)	PC @ N.R. (W)	0.86	9302.32	0.86	9302.32	0.86	9302.32
★★		0.96	8372.09	0.96	8372.09	0.96	8372.09
★★★		1.08	7441.86	1.08	7441.86	1.08	7441.86
★★★★		1.23	6511.62	1.23	6511.62	1.23	6511.62
★★★★★		1.43	5581.39	1.43	5581.39	1.43	5581.39
FoM (Gbps/W)	Avg. PC (W)	1.06	7579	1.80	4435	2.19	3658
Result		★★		★★★★★		★★★★★	

Source: Reproduced from ECO ICT, 2017. Ecology guideline for the ICT industry (Version 7.1), ICT Ecology Guideline Council. Available at: tca.or.jp/information/pdf/ecoguideline/guideline_eng_7_1.pdf.

system equipped with one wavelength and power consumption of a fully loaded system. Similar definitions are given in ANSI/SCTE (2009).

For relevant WDM configurations, Normative References (N.R.) have been defined which have to be matched by the respective WDM systems. Systems perform the better the clearer they exceed the N.R.

Table 8 lists the Figures of Merit (FoM) for a specific WDM transport system, the Fiber Service Platform 3000 (ADVA, 2016). The configuration analyzed here (Reference Configuration 2 of ECO ITC (2017)) is an 80-channel DWDM system with coherent 100-Gb/s OTN transponders and Degree-4 colorless, directionless ROADMs. Three generations of transport cards are analyzed, showing a strong increase of energy efficiency over time.

In order to further increase WDM energy efficiency, every layer of WDM transport systems and networks must be considered and optimized. This includes:

- Components (power supplies, ASICs, ADC/DAC,...).
- Thermal design (fans, also: Heating, Ventilation, Air Conditioning (HVAC)).
- Low-energy modes (partial deactivation, sleep modes,...).
- Networking aspects like grooming, flexibility, reconfigurability.

Even combination of these aspects may not be sufficient to keep WDM power consumption stable with exponentially increasing bit rates. Fig. 62 displays power consumption of different WDM generations with channel bit rates of 2.5, 10, 40, 100 and 400 Gb/s, respectively. It can be seen that efficiency massively increased over time, now approaching 0.1 W/(Gb/s). However, absolute consumption also slightly increased.

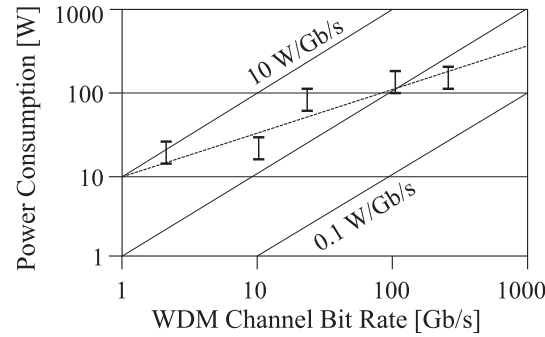


Fig. 62 Development of WDM power consumption over time, with increasing bit rates.

High savings in CO₂ footprint can be enabled by ICT, and especially telecommunications, outside the ICT sector. Examples include reduction of travel and truck rolls, dematerialization, control of energy grids, and many more. The telecommunications-related CO₂ abatement for 2020 is estimated to be five times higher as the ICT contribution to CO₂ production (GeSI, 2020, 2012).

Overview on the Optical Transport Network

OTN Layers

The Optical Transport Network (OTN) is specified in several ITU-T Recommendations (G.872 on architecture (ITU-T, 2017), G.709 on frames and formats (ITU-T, 2009a), G.798 on atomic functions and processes (ITU-T, 2012a)). It is sometimes also called Optical Transport Hierarchy (OTH). It combines TDM and WDM into a common transport system.

The TDM part is hierarchically structured, with Optical Channels (OCh) forming its basis. They are structured in hierarchical levels called Optical Channel Payload Unit (OPU), Optical Channel Data Unit (ODU) and Optical Channel Transport Unit (OTU).

The ODU signals are the end-to-end networking entities. Eight ODU signals have been defined in G.709. They address different bit rates and mapping schemes. The nominal ODU_k rates are approximately 1.25 Gb/s (ODU0), 2.50 Gb/s (ODU1), 10.04 Gb/s (ODU2), 40.32 Gb/s (ODU3), and 104.79 Gb/s (ODU4). For 10 GbE LAN PHY services, an overclocked ODU2e with 10.40 Gb/s has been defined. In addition, two ODUFlex signals for Constant Bit-Rate (CBR) clients and clients which are mapped via the Generic Framing Procedure (GFP, according to ITU-T Recommendation G.7041 (ITU-T, 2016c)) have been specified. Their nominal bit rates are $239/238 \times \text{CBR client bit rate}$ and $n \times 1.25 \text{ Gb/s}$ (with integer n), respectively. Jitter tolerances are ± 20 ppm for most ODUs, and ± 100 ppm for ODU2e and ODUFlex for CBR services.

The layered structure of OTN is shown in Fig. 63.

The OTN point-to-point single-wavelength frame structure is called Optical channel Transport Unit (OTU). The nominal OTU_k rates are approximately 2.67 Gb/s (OTU1), 10.71 Gb/s (OTU2), 43.02 Gb/s (OTU3), and 111.81 Gb/s (OTU4). The OTU_k frame structure always contains 4×4080 bytes. Therefore, the frame duration is not constant across hierarchy levels.

Multiplexing of several OCh by means of WDM creates the Optical Multiplex Section (OMS) and the Optical Transport Section (OTS). The OMS and OTS can be accompanied by an Optical Supervisory Channel (OSC, i.e., a dedicated wavelength carrying supervision and management information only), as indicated in Fig. 63.

The OMS refers to sections between optical multiplexer and demultiplexer, and the OTS to sections between optical amplifiers, respectively (Gorshe, 2010). This is shown in Fig. 64. Here, two OTN interfaces, Inter-Division Interface (IrDI) and Intra-Division Interface (IaDI), are shown. These interfaces are defined in ITU-T Recommendation G.872. They allow interworking between (IrDI) and within (IaDI) network domains. The IrDI interfaces are defined with full regeneration at each end of the interface.

OTN Mapping and Multiplexing

All relevant client signals can be mapped efficiently into OTN. Via OTN multiplexing, they can also be (time-domain) multiplexed onto high-bit-rate wavelengths. The resulting mapping options are summarized in Fig. 65.

SONET/SDH client signals can directly be mapped into OTN OPUs (which perform the client adaptation). Although OTN does not require synchronization, it can support the synchronization requirements of SONET/SDH and of synchronous Ethernet.

For the different Layer-2 signals, different mapping mechanisms can be used. One option is GFP. It allows mapping of various Layer-2 signals into SONET/SDH or ODU frames. The client signals can be Protocol Data Unit-oriented (like IP/PPP or Ethernet MAC) or block-code-oriented CBR streams such as Fiber Channel (FC).

For 10 GbE, two physical implementations have been defined in IEEE 802.3ae (IEEE, 2002). The WAN PHY operates at 9.954 Gb/s and is bit-rate-compatible with SONET OC-192 and SDH STM-64. It can be mapped into ODU2. The LAN PHY, which

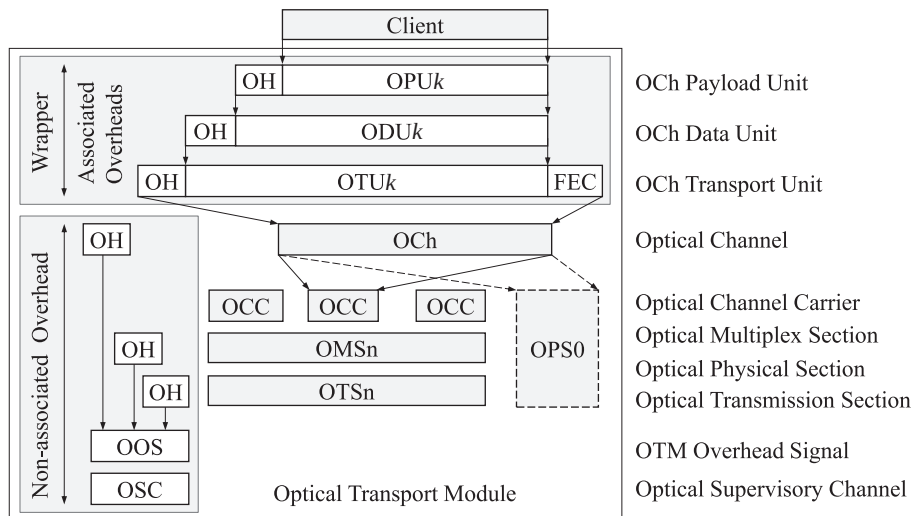


Fig. 63 OTN layering and monitoring. OH: Overhead.

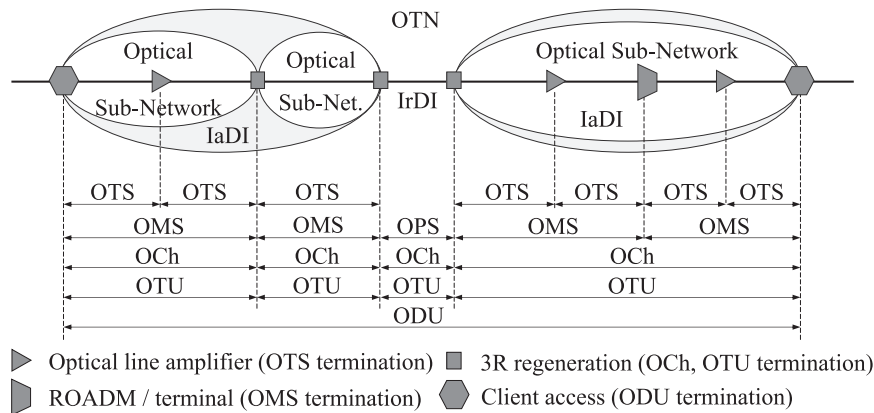


Fig. 64 OTN network layers and interfaces.

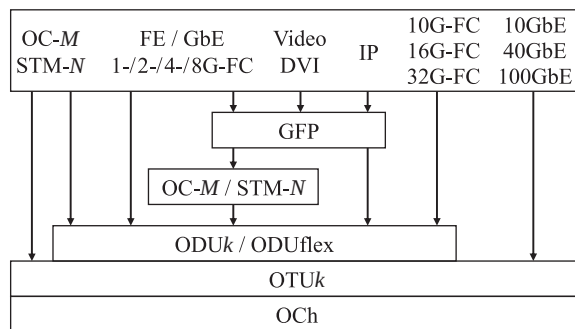


Fig. 65 OTN mapping options.

represents the majority of 10 GbE interfaces, operates at 10.3125 Gb/s. Direct mapping is possible into the overlocked ODU2e. ODU2e can be transported via ODU3 and ODU4, or directly on a wavelength with overlocked OTU2.

40 GbE and 100 GbE can directly be mapped into OTN. For Fiber Channel and other CBR services, the Generic Mapping Procedure (GMP) as defined in ITU-T G.709 can be used for mapping into ODUFlex (CBR type), as an alternative to GFP.

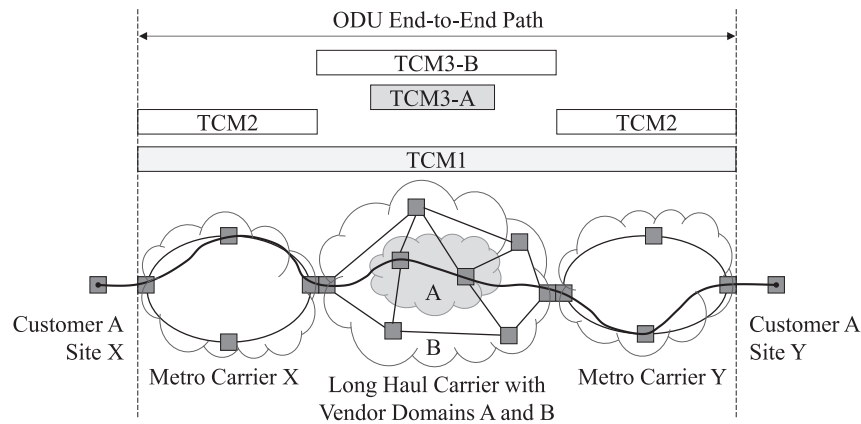


Fig. 66 OTN Tandem Connection Monitoring example.

OTN Operations and Monitoring Aspects

An important OTN feature is *Tandem Connection Monitoring* (TCM). TCM allows end-to-end monitoring across different administrative domains (i.e., network, operator or vendor domains). 6 independent TCM levels have been defined, allowing monitoring for nested and cascaded domains. Fig. 66 shows a transmission example.

Two metro carriers and one long haul-carrier with two nested vendor domains provide the end-to-end connection. Every network domain and also the end-to-end connection (TCM1) can be monitored simultaneously.

Conclusion

Wavelength Division Multiplexing is a multiplexing and multiple-access technology, used in fiber-optic transmission in order to maximize transmitted bit rates. Its earliest beginnings, in the form of groundbreaking work on optical fibers and semiconductor lasers, are roughly 50 years old. Nonetheless, without WDM, the Internet as it is known today would not exist. Practically all long-haul transmission of data, video and voice is based on WDM. According to the Cisco Virtual Network Index (Cisco, 2017), global Internet traffic is approaching some 100 Tb/s around 2020. Today, latest commercial WDM systems have transport capacity, over up to 1000 km reach, of 20 Tb/s. Since this capacity exceeds the one of any other transmission channel by orders of magnitude, there are no alternatives to long-haul WDM.

In addition, backhaul of any end-customer access technology – including mobile – is increasingly based on WDM, and even in direct fiber-optic access, WDM has started to be used. With copper-based wireline access diminishing, the remaining Internet areas without WDM are dedicated-fiber, non-WDM (“gray”) applications, the last mile in mobile networks and some microwave and satellite communications. Here, it is worth noting that even in free-space optical communications, WDM started to be used.

In the field of optics, WDM is strongly related to *modern optics* and *photonics*. Relevant areas are quantum optics and optoelectronics, but WDM also covers aspects of optomechanics and electro-optics. These areas also cover semiclassical approaches, including wave optics. The latter is obviously used for describing field propagation in fibers. Different semiclassical approaches are also used for describing optical-amplifier noise figures. As such, WDM is also one of the relevant applications – and success stories – of modern optics.

References

- ADVA, 2016. ADVA optical networking SE: Fiber service platform 3000R7 module and system specification, Release 16.1, Issue A (5/3/2016).
- ANSI/SCTE, 2009. Energy efficiency for telecommunication equipment: Methodology for measurement and reporting transport requirements, American National Standard for Telecommunications, ATIS-060015.02.2009.
- Agrawal, G.P., 1992. Fiber-Optic Communication Systems. New York: John Wiley & Sons.
- Agrawal, G.P., 1995. Nonlinear Fiber Optics, second ed. San Diego: Academic Press.
- Agrawal, G.P., Dutta, N.K., 1986. Long-Wavelength Semiconductor Lasers. New York: Van Nostrand Reinhold.
- Alexander, S.B., 1997. Optical Communication Receiver Design, vol. TT22. Bellingham: SPIE, (ISBN 0-8194-2023-9).
- Azadeh, M., 2009. Fiber Optics Engineering. Dordrecht Heidelberg: Springer, (ISBN 978-1-4419-0304-4).
- Bosco, G., Carena, A., Curri, V., *et al.*, 2010. Performance limits of nyquist-WDM and CO-OFDM in high-speed PM-QPSK systems. IEEE Photonics Technology Letters 22 (15), 1129–1131.
- Bristiel, B., Jiang, S., Gallion, P., *et al.*, 2006. New model of noise figure and RIN transfer in fiber raman amplifiers. IEEE Photonics Technology Letters 18 (8), 980–982.
- Bülöw, H., 1998. System outage probability due to first- and second-order PMD. IEEE Photonics Technology Letters 10 (5), 1548–1557.
- Chraplyvy, A.R., 1990. Limitations on lightwave communications imposed by optical-fiber nonlinearities, invited. IEEE Journal of Lightwave Technology 8 (10), 1548–1557.
- Cisco, 2017. The zettabyte era: Trends and analysis, Cisco White Paper. Available at: www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html.

- Coldren, L.A., Fish, G.A., Akulova, Y., *et al.*, 2004. Tunable semiconductor lasers: A tutorial. *IEEE Journal of Lightwave Technology* 22 (1), 193–202.
- Corning, 1998. Corning® SMF-LSTM CPC6 Single-mode non-zero dispersion-shifted optical fiber, corning incorporated, product information P11050, Issued 9/1998.
- Corning, 2014. LEAF® Optical Fiber, Data Sheet, Corning Incorporated. Available at: www.corning.com/media/worldwide/coc/documents/Fiber/LEAF%20Optical%20fiber.pdf.
- DeCusatis, C., 2008. Handbook of Fiber Optic Data Communication, third ed. Burlington, MA: Elsevier Academic Press, (ISBN 978-0-12-374216-2).
- Desurvire, E., 1994. Erbium-Doped Fiber Amplifiers. New York: John Wiley & Sons.
- Dragone, C., 2005. Low-loss wavelength routers for WDM optical networks and high-capacity IP routers. *IEEE Journal of Lightwave Technology* 23 (1), 66–79.
- ECO ICT, 2017. Ecology guideline for the ICT industry (Version 7.1), ICT Ecology Guideline Council. Available at: tca.or.jp/information/pdf/ecoguideline/guideline_eng_7_1.pdf.
- Essiambre, R.-J., Kramer, G., Winzer, P.J., *et al.*, 2010. Capacity limits of optical fiber networks. *IEEE Journal of Lightwave Technology* 28 (4), 662–701.
- Fishman, D.A., Nagel, J.A., 1993. Degradations due to stimulated brillouin scattering in multigigabit intensity-modulated fiber-optic systems. *IEEE Journal of Lightwave Technology* 11 (11), 1721–1728.
- Friis, H.T., 1944. Noise figures of radio receivers. *Proceedings of the Institute of Radio Engineers* 32 (12), 419–422.
- GeSI, 2012. GeSI SMARTer 2020: The role of ICT in driving a sustainable future, GeSI and BCG. Available at: gesi.org/assets/js/lib/tinymce/jscripts/tiny_mce/plugins/ajaxfilemanager/uploaded/SMARTer%202020%20The%20Role%20of%20ICT%20in%20Driving%20a%20Sustainable%20Future%20-%20December%202012.pdf.
- Gorshe, S., 2010. A tutorial on ITU-T G.709 optical transport networks (OTN), technology White Paper, PMC-Sierra, Inc., PMC-2081250, 1.
- Grave de Peralta, L., Bernussi, A.A., Gorbounov, V., *et al.*, 2004. Temperature-insensitive reflective arrayed-waveguide grating multiplexers. *IEEE Photonics Technology Letters* 16 (3), 831–833.
- Grobe, K., Eiselt, M.H., 2014. Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, New Jersey: John Wiley & Sons, (ISBN 978-0-470-62302-2).
- Hirano, A., 2002. Optical amplifiers and their standardization in ITU-T & IEC, ITU-T Workshop on IP/Optical. Available at: www.itu.int/itudoc/itu-t/workshop/optical/s8-p02r.html.
- IEEE, 2002. IEEE Std 802.3ae™-2002 (Amendment to IEEE Std 802.3-2002), Media access control (MAC) parameters, physical layers, and management parameters for 10 Gb/s operation, August.
- ITU-T, 2009a. Interfaces for the optical transport network, Recommendation ITU-T G.709.
- ITU-T, 2009b. Characteristics of a non-zero dispersion-shifted single-mode optical fibre and cable, Recommendation ITU-T G.655, 11/2009.
- ITU-T, 2010a. Characteristics of a fibre and cable with non-zero dispersion for wideband optical transport, Recommendation ITU-T G.656, 07/2010.
- ITU-T, 2010b. Characteristics of a dispersion-shifted, single-mode optical fibre and cable, Recommendation ITU-T G.653, 07/2010.
- ITU-T, 2012a. Characteristics of optical transport network hierarchy equipment functional blocks, Recommendation ITU-T G.798, 12/2012.
- ITU-T, 2012b. Optical system design and engineering considerations, Recommendation ITU-T Series G, Supplement 39, 09/2012.
- ITU-T, 2015. Optical interfaces for coarse wavelength division multiplexing applications, Recommendation ITU-T G.695.
- ITU-T, 2016a. Characteristics of a bending-loss insensitive single-mode optical fibre and cable, Recommendation ITU-T G.657, 11/2016.
- ITU-T, 2016b. Characteristics of a cut-off shifted single-mode optical fibre and cable, Recommendation ITU-T G.654, 11/2016.
- ITU-T, 2016c. Generic framing procedure, Recommendation ITU-T G.7041/Y.1303, 08/2016.
- ITU-T, 2016d. Characteristics of a single-mode optical fibre and cable, Recommendation ITU-T G.652, 11/2016.
- ITU-T, 2017. Architecture of optical transport networks, Recommendation ITU-T G.872, 01/2017.
- Johnson, C.R., Schniter, P., Endres, T.J., *et al.*, 1998. Blind equalization using the constant modulus criterion: A review, invited paper. *Proceedings of the IEEE* 86 (10), 1927–1950.
- Kaminow, I.P., Koch, T.L., 1997. Optical fiber telecommunications III. San Diego: Academic Press, (ISBN 0-12-395170-4).
- Kobayashi, H., Tang, D.T., 1971. On decoding of correlative level coding systems with ambiguity zone detection. *IEEE Transactions on Communication Technology COM-19* (4), 467–477.
- Lima, M.J.N., Nogueira, R.N., Silva, J.C.C., *et al.*, 2005. Comparison of the temperature dependence of different types of bragg gratings. *Microwave and Optical Technology Letters* 45 (4), 305–307.
- Linke, R.A., Gnauck, A.H., 1988. High-capacity coherent lightwave systems. *IEEE Journal of Lightwave Technology* 6 (11), 1750–1769.
- Loudon, R., 1985. Theory of noise accumulation in linear optical-amplifier chains. *IEEE Journal of Quantum Electronics* QE-21 (7), 766–773.
- Lyubomirsky, I., 2006. Coherent detection for optical duobinary communication systems. *Photonics Technology Letters* 18 (7), 868–870.
- Lyubomirsky, I., 2010. Quadrature duobinary for high-spectral efficiency 100G transmission. *IEEE Journal of Lightwave Technology* 28 (1), 91–96.
- Maier, G., Pattavina, A., De Patre, S., *et al.*, 2002. Optical network survivability: Protection techniques in the WDM layer. *Photonic Network Communications* 4 (3/4), 251–269.
- Mohs, G., *et al.*, 2000. Maximum link length versus data rate for SPM limited systems. In: *Proceedings European Conference on Optical Communication (ECOC) 2000*, Munich.
- Ohsono, K., *et al.*, 2003. High performance optical fibers for next generation transmission systems. *Hitachi Cable Review*, 22).
- OFS, 2017. TrueWave® RS Optical Fiber, Data Sheet, Doc ID: fiber-120, OFS Fitel, LLC, 06/2017. Available at: fiber-optic-catalog.ofsoptics.com/Asset/TrueWaveRSLWP-120-web.pdf.
- Premaratne, M., 2004. Analytical characterization of optical power and noise figure of forward pumped Raman amplifiers. *Optics Express* 12 (18), 4235–4245.
- Prysmian Group, 2010. TeraLight™ ultra optical fiber, draka communications, data sheet product type G.655.E, Issue Date 08/2010. Available at: www.prysmiangroup.com/sites/default/files/business_markets/markets/downloads/datasheets/SMF—TeraLight-Ultra-Optical-Fiber.pdf.
- Roberts, P.J., Couny, F., Sabert, H., *et al.*, 2005. Ultimate low loss of hollow-core photonic crystal fibers. *Optics Express* 13 (1), 236–244.
- Russell, P.S.J., 2006. Photonic crystal fibers. *IEEE Journal of Lightwave Technology* 24 (12), 4729–4749.
- Sakurai, Y., Khan, S., Takamura, H., *et al.*, 2012. LCOS-based gridless wavelength blocker array for broadband signals at 100Gbps and beyond, OFC2012, L.A., paper OTh3D.
- Savory, S.J., Mikhailov, V., Killey, R.I., *et al.*, 2007. Digital coherent receivers for uncompensated 42.8Gb/s transmission over high PMD fibre, ECOC2007, Berlin, September.
- Schetzen, M., 1980. The Volterra and Wiener Theories of Nonlinear Systems. New York: John Wiley & Sons.
- Senio, J.M., Jamro, M.Y., 2009. Optical Fiber Communications: principles and Practice, 3rd ed. Harlow: Pearson Education Limited.
- Smith, D.R., 2003. Digital Transmission Systems, 3rd ed. Norwell, MA, USA: Kluwer Academic Publishers, (ISBN 1-4020-7587-1).
- Tonguz, O.K., Wagner, R.E., 1991. Equivalence between preamplified direct detection and heterodyne receivers. *IEEE Photonics Technology Letters* 3 (9), 835–837.
- Winzer, P.J., Essiambre, R.-J., 2006. Advanced modulation formats for high-capacity optical transport networks, invited paper. *IEEE Journal of Lightwave Technology* 24 (12), 4711–4728.

Coherent Lightwave Systems

Michael J Connelly, University of Limerick, Limerick, Ireland

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Most optical communication systems (typically operating in the 1300 nm or 1550 nm optical fiber communication bands) are based on intensity modulation (IM) of a carrier lightwave by an electrical data signal and Direct Detection (DD) of the received light. The simplest IM-DD schemes employ On/Off Keying (OOK) whereby turning on or off a carrier lightwave transmits a binary '0' or '1'. The lightwave is transmitted via an optical fiber and at the receiver detected by a photodetector. The resulting photocurrent is then processed to determine if a '0' or '1' was received. Photodetection is insensitive to the phase and polarization state of the received lightwave. In the ideal case, assuming a monochromatic carrier and noiseless receiver, the Bit Error Rate (BER) is related to the quantum noise of the received light; the number of received photons/bit required (sensitivity) to achieve a given BER is termed the quantum-limit; for example for a BER of 10^{-9} the quantum limit is 10 photons/bit. In practice the sensitivity of DD receivers is 10–30 dB less than the quantum limit, because of the relatively much higher receiver noise. The use of a high-gain optical preamplifier can greatly improve the sensitivity.

The capacity of IM-DD links can be greatly increased by the use of wavelength division multiplexing (WDM). Typical commercial WDM-IM-DD systems utilize an optical channel spacing of 50 GHz for a 40 Gb/s channel data rate. Erbium Doped Fiber Amplifiers (EDFAs) are used to compensate for transmission link losses (fiber attenuation and other losses such as splitting losses). The usable bandwidth of an EDFA is typically 30 nm, which is equivalent to 3.77 THz if centered at 1550 nm. The number of 50 GHz WDM channels that can fit into the EDFA bandwidth is 75 leading to a gross data rate of approximately 3 Tb/s. Higher gross data rates are achievable by using more complicated amplifier configurations to extend the usable bandwidth. The spectral efficiency of OOK modulation is 1 bit/s/Hz/channel.

Increased gross data rates can be achieved by employing spectrally efficient high-order modulation formats employing phase and amplitude modulation of the carrier lightwave (Nakazawa *et al.*, 2010; Seimetz, 2010; Winzer, 2012). A further advantage of high-order formats is that the symbol rate (R_s) corresponding to a specific data rate (R_b) is reduced thereby relaxing the bandwidth requirements of the transmitter and receiver components. System capacity can be further increased by Dual Polarization (DP) transmission, in which the two orthogonal polarizations of the carrier lightwave are separately modulated and recombined. Low-order phase modulated signals can be detected using differential detection receivers based on delay interferometers and DD. Such receivers have been employed in commercial systems but are not suitable for detection of more complex modulation formats. Detection of a phase/amplitude modulated lightwave necessitates the use of linear receivers that preserve the lightwave amplitude and phase. Linear receivers use coherent detection, whereby the received modulated lightwave is combined with a Local Oscillator (LO) laser and simultaneously detected; the resulting electrical signals are then processed to recover the received lightwave phase and amplitude. Because coherent receivers are linear, post-reception Digital Signal Processing (DSP) can be used to compensate for channel and receiver impairments and implement polarization demultiplexing, decoding and error correction. The combination of coherent optical reception and DSP is called a Digital Coherent Receiver (DCR) and forms the basis for all coherent optical receivers. This article gives an overview of optical modulation, high-order modulation formats, differential and coherent detection, DCRs and current developments in the field.

Optical Modulation

A monochromatic lightwave field can be expressed as

$$e_{CW}(t) = \sqrt{2P_s} \cos(\omega_s t) \quad (1)$$

where P_s is the power (averaged over an optical period) and ω_s is the angular optical frequency. An arbitrary modulation of the lightwave can be expressed as

$$e_s(t) = \sqrt{2P_s} a(t) \cos[\omega_s t + \theta_s(t)] \quad (2)$$

with time dependent amplitude modulation $a(t)$ and phase modulation $\theta_s(t)$. Dividing (2) by $\sqrt{2P_s}$ gives the normalized lightwave field

$$\frac{e_s(t)}{\sqrt{2P_s}} = a(t) \cos[\omega_s t + \theta_s(t)] \quad (3)$$

which can be expanded as

$$a(t) \cos[\omega_s t + \theta_s(t)] = a(t) \{ \cos(\omega_s t) \cos[\theta_s(t)] - \sin(\omega_s t) \sin[\theta_s(t)] \} \quad (4)$$

Taking $\cos(\omega_s t)$ as the phase reference, the In-phase (I) and Quadrature (Q) components of (4) are $a(t)\cos[\theta_s(t)]$ and $a(t)\sin[\theta_s(t)]$ respectively. The phasor representation of (4) $A_s(t) = a(t)\exp[j\theta_s(t)]$, as shown in Fig. 1, is called the complex amplitude. Coherent lightwave communication systems always use the I and Q components of the carrier.

Modulators

Optical phase modulation can be implemented using an integrated Phase Modulator (PM) or Mach-Zehnder Modulator (MZM) as shown in Fig. 2. The PM comprises an optical waveguide embedded in an electro-optic material, usually Lithium Niobate (LiNbO_3). Phase modulation $\theta(t)$ of the input lightwave is achieved by applying a drive voltage $v(t)$ between the electrodes. Assuming a lossless device the relationship between the input $E_{in}(t)$ and output $E_{out}(t)$ complex fields is

$$E_{out}(t) = E_{in}(t)\exp[j\theta(t)] = E_{in}(t)\exp\left[j\frac{v(t)}{V_\pi}\right] \quad (5)$$

where V_π is the voltage required to obtain a π phase shift.

A MZM comprises two PMs embedded in a Mach-Zehnder Interferometer. In a dual-drive MZM, shown in Fig. 2, the PMs are independently driven. The input lightwave is split into two lightwaves of equal power. The phase shifted lightwaves in each arm of the interferometer are recombined resulting in interference, which depending on the drive voltages, can be varied between destructive and constructive. The modulator transfer function is given by

$$\frac{E_{out}(t)}{E_{in}(t)} = \frac{\exp[j\theta_1(t)] + \exp[j\theta_2(t)]}{2} = \exp\left\{j\frac{[\theta_1(t) + \theta_2(t)]}{2}\right\} \cos\left[\frac{\theta_1(t) - \theta_2(t)}{2}\right] \quad (6)$$

Each PM induces a phase shift $\theta_k(t) = v_k(t)\pi/V_{\pi,k}$, where $k=1,2$ denotes the upper or lower PM respectively. If the PMs are identical then $V_{\pi,k} = V_\pi$. If $v_1(t) = v_2(t)$ (push-push mode) then $\theta_1(t) = \theta_2(t)$ resulting in phase modulation $\theta(t) = \theta_1(t)$, without any accompanying amplitude modulation. If $v_2(t) = -v_1(t)$ (push-pull mode), the phase term in (6) is equal to unity so the modulator acts as an amplitude modulator. The power transfer function (magnitude squared of (6)), is

$$\frac{P_{out}(t)}{P_{in}(t)} = \frac{1}{2} \left\{ 1 + \cos\left[\frac{v(t)}{V_\pi}\pi\right] \right\} \quad (7)$$

where $v(t) = 2v_1(t)$ is the MZM drive voltage. Plots of (6) and (7) are shown in Fig. 3 for push-pull mode operation. The MZM can be operated as an intensity modulator by biasing at the quadrature point $-V_\pi/2$ and driving with a V_π peak-to-peak modulation. By biasing at minimum transmission point $-V_\pi$ and driving with a $2V_\pi$ peak-to-peak modulation the MZM can be operated as a Binary Phase Shift Key (BPSK) modulator since there is a phase jump of π every time the minimum transmission point is crossed.

A major disadvantage of a single MZM is that independent modulation of the I and Q components is not possible, which complicates the drive voltage waveforms required to generate high-order modulation formats. An alternative modulator is based on two push-pull mode MZMs in parallel, one of which is in series with a $\pi/2$ phase shifter. Such an IQ Modulator (IQM), shown in Fig. 2, has a transfer function, normalized such that its maximum magnitude is equal to unity, given by

$$\frac{E_{out}(t)}{E_{in}(t)} = \frac{1}{\sqrt{2}} \cos\left[\frac{v_I(t)}{2V_\pi}\pi\right] + j\frac{1}{\sqrt{2}} \cos\left[\frac{v_Q(t)}{2V_\pi}\pi\right] \quad (8)$$

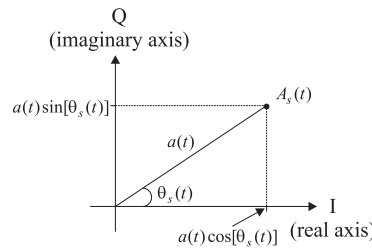


Fig. 1 I-Q representation of the complex amplitude of a modulated lightwave.

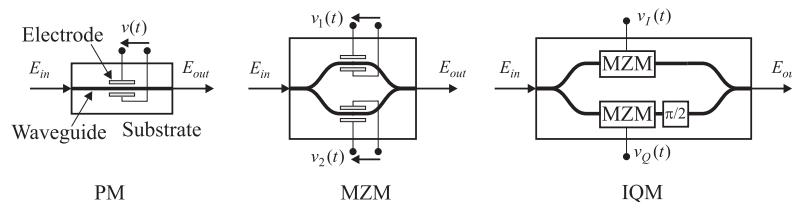


Fig. 2 Integrated PM, dual-drive MZM and IQM.

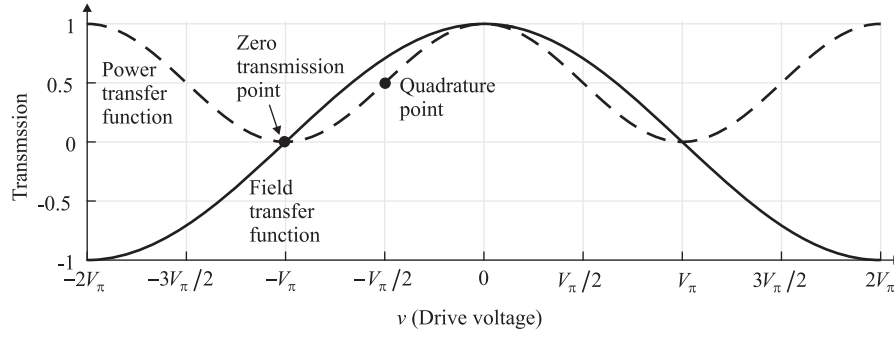


Fig. 3 Push-pull mode MZM transfer functions.

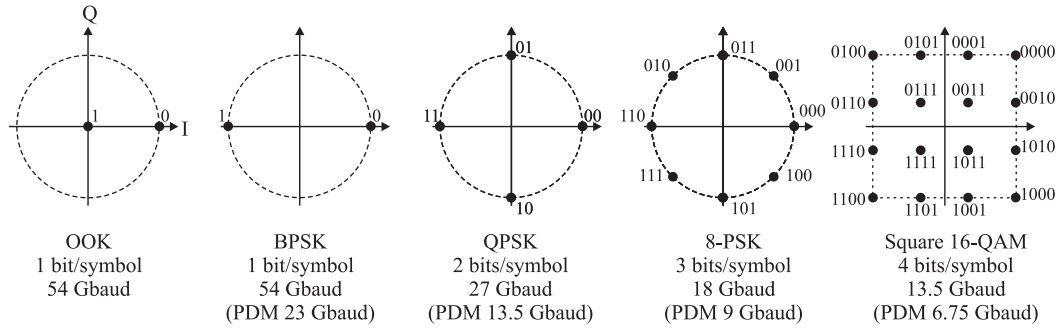


Fig. 4 Constellation diagrams with Gray coded symbols for common modulation formats assuming SP transmission. The symbol rates for SP and PDM transmission corresponding to 54 Gb/s data rate are also shown.

where $v_I(t)$ and $v_Q(t)$ are the I and Q branch MZM drive voltages respectively. The polar form of (8) is

$$\frac{E_{out}(t)}{E_{in}(t)} = a(t) \exp[j\theta(t)] \quad (9)$$

with amplitude modulation

$$a(t) = \frac{1}{\sqrt{2}} \left\{ \cos^2 \left[\frac{v_I(t)}{2V_\pi} \pi \right] + \cos^2 \left[\frac{v_Q(t)}{2V_\pi} \pi \right] \right\}^{1/2} \quad (10)$$

and phase modulation

$$\theta(t) = \arg \left\{ \cos \left[\frac{v_I(t)}{2V_\pi} \pi \right] + j \cos \left[\frac{v_Q(t)}{2V_\pi} \pi \right] \right\} \quad (11)$$

Any point within the normalized I-Q space (unit circle) can be reached by choosing suitable values of v_I and v_Q .

High-Order Modulation

In digital optical communications, the complex amplitude of the carrier lightwave can only take on discrete values, whereby b data bits are mapped onto M discrete symbol points located in the I-Q plane (constellation diagram). M is a power of 2 so $b = \log_2 M$. Bit-to-symbol mapping can be chosen arbitrarily; in most systems the Gray code is used, whereby only one bit per symbol differs from a neighboring symbol leading to optimum receiver BER performance. In M -ary PSK (m -PSK), symbols are mapped to M possible values of the carrier phase between 0 and 2π with a step size of $2\pi/M$; the phase values associated with BPSK are $[0, \pi]$; with 4-PSK (Quadrature PSK, QPSK), $[0, \pi/2, \pi, 3\pi/2]$ and with 8-PSK $[0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 7\pi/4, 9\pi/4]$ as shown in Fig. 4 (Gnauck and Winzer, 2005). In M -ary Quadrature Amplitude Modulation (m -QAM), symbols are mapped to one of M combinations of the carrier amplitude and phase. The constellation points can be arranged in a square (square QAM), as shown for 16-QAM in Fig. 4, or on multiple concentric circles (star QAM). 2-QAM and 4-QAM are the same as BPSK and QPSK respectively.

Demodulation of PSK signals using differential detection requires a phase reference, which is not available in DD based receivers. However, demodulation can be achieved if the data is pre-coded such that the information is carried in the phase changes between successive bits; such modulation is called Differential PSK (DPSK). The term DPSK is taken to mean differential binary DPSK. Differential Quadrature DPSK (DQPSK) is the most common differential PSK format used in commercial systems.

Single Polarization (SP) transmission is a 2D format since two degrees of freedom are exploited (I and Q). Two more degrees of freedom can be exploited by using Dual Polarization (DP) transmission, in which the two orthogonal polarizations of the carrier

lightwave are modulated by I-Q data (4D format). The two most common forms of DP transmission are Polarization Division Multiplexed (PDM) and Polarization Switched (PS) systems (Krongold *et al.*, 2012). In PDM, each polarization is independently modulated with half of the symbol bits, thereby halving the symbol rate compared with single-polarization transmission. For example in PDM-QPSK systems, each polarization is modulated with QPSK data so the number of possible symbols $M=16$ (Roberts *et al.*, 2009). The principle of PS can be illustrated by considering the most power efficient format of all 4D modulation formats, PS-QPSK which has 8 possible symbols (3 bits/symbol) (Karlsson and Agrell, 2009). Two of the bits define a QPSK symbol, while the remaining bit, corresponding to whether the particular polarization component is turned on or off, indicates which polarization the QPSK symbol is transmitted on as shown in Fig. 5. The eight levels of the PS-QPSK format are not possible to Gray code, since each point in the signal constellation has 6 nearest neighbors.

High-order modulation of a carrier lightwave is generally implemented in a number of stages. First the data bits to be transmitted are encoded, typically using Forward Error Correction (FEC), and mapped onto the desired signal constellation (Djordjevic *et al.*, 2009). If required the symbols can be differentially encoded by a pre-coder. The resulting digitally generated I and Q signals are then converted to analog electrical signals by Digital-to-Analog converters (DACs), electrically filtered if necessary, and sent to the IQM modulator I and Q ports inputs via drive amplifiers. Transmitter DSP can also include functions such as Nyquist spectral shaping (to minimize the modulated lightwave bandwidth) and signal pre-distortion (to mitigate transmission impairments).

A specific example of high-order modulation is the Differential QPSK (DQPSK) transmitter, shown in Fig. 6. The input data stream is first encoded using FEC and then mapped to QPSK symbols, resulting in I and Q data bit streams u_k and v_k respectively, where the subscript denotes the k -th bit. u_k and v_k are differentially encoded, using a serial pre-coder. Serial pre-coder can be realized for speeds of up to 40 Gb/s; higher speeds require more complex parallel architectures. The pre-coder output bit streams I_k and Q_k are converted to analog Non-Return-to-Zero (NRZ) voltage signals by the DACs. These signals drive the I and Q ports of the IQM, such that its MZMs are operated as BPSK modulators, which results in QPSK modulation of the input lightwave. The constant amplitude output lightwave has a phase $\theta_k = \tan^{-1}[\cos(Q_k\pi)/\cos(I_k\pi)]$ relative to the input lightwave. The modulator output constellation diagram is the same as the QPSK constellation shown in Fig. 4 but rotated 45° anticlockwise; however constant phase shifts are of no consequence in the demodulation process.

All modulators suffer from additive chirp (time dependent instantaneous frequency of the lightwave carrier frequency), especially at symbol transitions. When an MZM is operated in push-pull mode, the chirp is greatly reduced because the drive voltages are opposite in sign and so the induced chirp caused by each PM are opposite in sign and cancel. When the modulated lightwave is transmitted down an optical fiber, the chirp can lead to an increase in Chromatic Dispersion (CD) and a consequent degradation in system performance. In the MZM modulator, there are also undesirable power dips at the modulation signal transitions. Both of these effects can be greatly reduced by using Return-to-Zero (RZ) transmission. The required RZ pulses can be generated using a pulse carver, either prior to or after modulation. The most common RZ pulse carver uses a biased low-chirp MZM driven by a full or half symbol rate sinusoid. Commonly used RZ pulse streams have full-width at half maximum pulsewidths of 33%, 50% and 67% of the symbol period. The former two pulse streams have spectrums that contain a significant

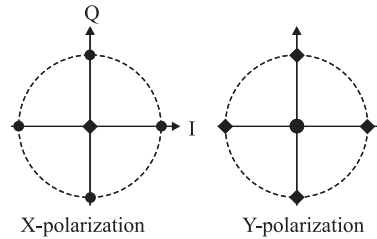


Fig. 5 Constellation diagrams for the two orthogonal polarizations (X and Y) of PS-QPSK. The circles indicate a QPSK symbol is transmitted on the X polarization, while a zero is transmitted on the Y polarization. The opposite is the case with the diamond constellation points. The symbol rate corresponding to 54 Gb/s data rate is 18 Gb/s.

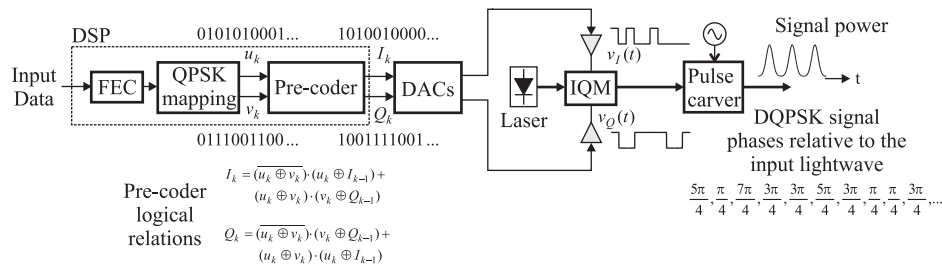


Fig. 6 DQPSK transmitter structure. The serial pre-coder logical relationships $\oplus$, $\cdot$ and $+$ are the XOR, AND and OR operators respectively. Lowpass Filters (LPFs) are often employed after the DACs.

carrier component, which is useful in retrieving the symbol clock at the receiver. The latter pulse stream has an alternating phase and is called Carrier Suppressed RZ (CSRZ), which compared to standard RZ-OOK is more tolerant to filtering and CD because of its narrower spectrum.

PDM-QPSK can be implemented by passing the orthogonal polarization components of the input lightwave through separate IQMs and recombining their outputs using a polarization beam combiner. PS-QPSK can be achieved in a similar way to PDM-QPSK except that the orthogonal outputs from each IQM are passed through on/off intensity modulators, to set the third bit of the symbol to be transmitted, before recombination.

Differential Detection

A differentially encoded PSK lightwave can be detected using differential detection and is especially applicable to DQPSK formats. The DQPSK demodulator shown in Fig. 7 consists of two parallel Delay Interferometers (DIs). The incoming signal is split into two arms each with two sub-paths. In each arm the upper sub-path is delayed by the symbol period and the lower sub-path phase is shifted by $\pi/4$ or $-\pi/4$. The Balanced Detector (BD) output photocurrents corresponding to the k -th received symbol are given by

$$i_{r,k} = -\frac{RP_s}{2\sqrt{2}} [\cos(\Delta\phi_k) - \sin(\Delta\phi_k)] \quad (12)$$

$$i_{s,k} = -\frac{RP_s}{2\sqrt{2}} [\cos(\Delta\phi_k) + \sin(\Delta\phi_k)] \quad (13)$$

where R and P_s are the photodiode responsivity and the demodulator optical input power respectively and $\Delta\phi_k = \phi_k - \phi_{k-1}$ is the phase difference between the k -th and $(k-1)$ -th symbols. The logical values corresponding to (12) and (13) are

$$r_k = \frac{-[\cos(\Delta\phi_k) - \sin(\Delta\phi_k)] + 1}{2} \quad (14)$$

$$s_k = \frac{-[\cos(\Delta\phi_k) + \sin(\Delta\phi_k)] + 1}{2} \quad (15)$$

Each of the four possible values of $\Delta\phi_k$, $[0, \pi/2, \pi, 3\pi/2]$ results in the symbols $\{r_k, s_k\} = [\{0,0\}, \{1,0\}, \{1,1\}, \{0,1\}]$. If received signal is distortion and noise free, r_k and s_k correspond to the transmitted data channels u_k and v_k (as in Fig. 6) respectively.

PDM-DQPSK transmission can be achieved by separately modulating the orthogonal polarizations of the carrier and recombining using a polarization beam combiner. When PDM signals are transmitted in an optical fiber they experience, Polarization Mode Dispersion (PMD) and Polarization Dependent Loss (PDL); which can be very severe in long-haul links and as such must be compensated for at the receiver. Demultiplexing of the two polarization streams is also necessary. In a DD based receiver polarization demultiplexing is accomplished optically using polarization tracking, which is also used for PMD and PDL compensation. Optically based polarization tracking configurations involve multiple optical and optoelectronic components (such as beam splitters and photodetectors) and electronic control and can be very challenging. The requirement for optically based tracking can be removed by using coherent detection and post-processing of the resulting electrical signals. The practical implementation of DD based receiver structures becomes impractical for higher order modulation formats. The alternative is to employ coherent detection, which allows polarization demultiplexing, demodulation and fiber channel compensation to be realized in the electrical domain thereby significantly reducing the receiver optical front-end complexity especially in terms of interferometric demodulation structures.

Coherent Detection

In the context of coherent lightwave systems the term coherent refers to techniques employing mixing between two optical waves on a photodetector (Ip *et al.*, 2008, Kikuchi, 2016). Optical receivers employing coherent detection are linear and thereby allow access, in the electrical domain, to all of the information – amplitude, frequency, phase and polarization – present in the detected

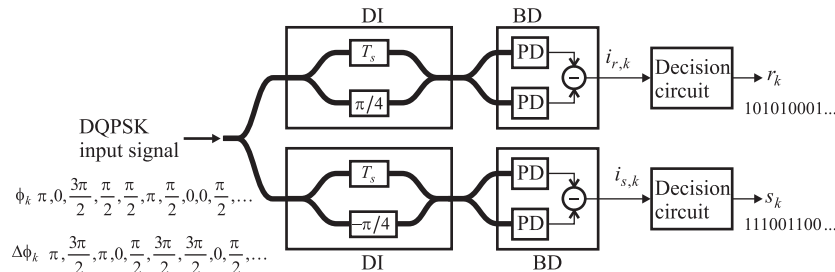


Fig. 7 DQPSK demodulator.

lightwave. The development of high speed Analog to Digital Converters (ADCs) and Application Specific Integrated Circuits (ASICs) has made it possible to carry out real-time DSP on multi-Gb/s data rate signals, which can be employed to equalize fiber impairments, carry out polarization demultiplexing and demodulate high-order modulation formats. DSP also allows the implementation of and flexible tunable WDM receivers. Photonic Integrated Circuit (PIC) technology has experienced rapid development to the point where most of the optical and electronic components required in coherent systems are commercially available or at an advanced state of research and development. This makes it possible to realize reliable and cost-effective solutions for meeting the demand for ultra-fast data rates.

Basic Principles

The fundamental concept in coherent detection is to multiply a modulated optical signal field $E_s(t)$ by a LO field $E_l(t)$ that acts as a phase reference, which enables full recovery of the amplitude and phase information of the signal. The complex optical fields of the signal and LO at the receiver input are

$$E_s(t) = \sqrt{2P_s}A_s(t)\exp\{j[\omega_s t + \theta_{sn}(t) + \theta_0]\} \quad (16)$$

$$E_l(t) = \sqrt{2P_l}\exp\{j[\omega_l t + \theta_l(t)]\} \quad (17)$$

where P_l is the LO power, ω_s and ω_l the received signal and LO optical angular frequencies respectively, $\theta_{sn}(t)$ and $\theta_l(t)$ the signal and LO phase noises respectively and θ_0 is the static phase shift between the signal and LO. The basic form of a balanced coherent receiver is shown in Fig. 8, in which the identical photodiodes have responsivity R . Assuming co-polarized signal and LO fields, the fields prior to detection by the upper and lower photodiodes as given by

$$E_1 = \frac{1}{\sqrt{2}}(E_s + E_l) \quad (18)$$

$$E_2 = \frac{1}{\sqrt{2}}(E_s - E_l) \quad (19)$$

with corresponding photocurrents

$$i_1(t) = \frac{R}{2} \left\{ P_s a(t)^2 + P_l + 2\sqrt{P_s P_l} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_{sn}(t) - \theta_l(t) + \theta_0] \right\} \quad (20)$$

$$i_2(t) = \frac{R}{2} \left\{ P_s(t) a(t)^2 + P_l - 2\sqrt{P_s P_l} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_{sn}(t) - \theta_l(t) + \theta_0] \right\} \quad (21)$$

Subtracting $i_2(t)$ from $i_1(t)$, gives the BD output

$$i(t) = 2R\sqrt{P_s P_l} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_{sn}(t) + \theta_0] \quad (22)$$

having a total phase noise $\theta_n(t) = \theta_{sn}(t) - \theta_l(t)$. $\Delta\omega = \omega_s - \omega_l$ is the Intermediate Frequency (IF).

Heterodyne Detection

In heterodyne detection $\Delta\omega$ is non-zero. In order to avoid signal distortion caused by spectral folding, $|\Delta\omega|$ must be larger than the modulation bandwidth ($\approx 2\pi R_s \text{ rad/s}$) of the optical carrier, i.e., $|\Delta\omega| > 2\pi R_b$, as shown in Fig. 9. It is possible to recover $a(t)$ using envelope detection, which involves electronically mixing $i(t)$ with itself and low-pass filtering to eliminate the component centered at $2\omega_{IF}$ and pass the baseband term proportional to $a^2(t)$. However, phase information cannot be recovered by this technique. Constant envelope formats such as M-PSK can be demodulated by using electrical differential detection which, in a manner similar to optical differential detection, determines the phase difference between consecutive symbols. However, pre-coding of the transmitted data is required.

Synchronous electrical detection can be employed to fully recover $A_s(t)$ such as in the receiver shown in Fig. 10. $i(t)$ is split into two and electrical LOs are used to translate the IF signals to baseband. LPFs pass the baseband signal and eliminate the terms centered at $2|\Delta\omega|$. An electrical Phase Locked Loop (PLL) provides carrier phase recovery by compensating for the IF signal phase noise. If the phase noise is totally eliminated then the output photocurrents $i_I(t)$ and $i_Q(t)$ are given by

$$i_I(t) = R\sqrt{P_s P_l} a(t) \cos[\theta_s(t)] \quad (23)$$

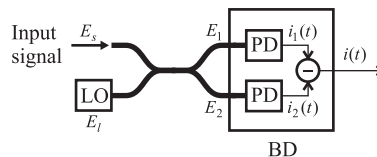


Fig. 8 Coherent receiver using balanced detection.

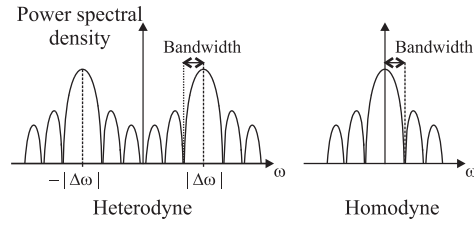


Fig. 9 Power spectrums of heterodyne and homodyne down-converted signals.

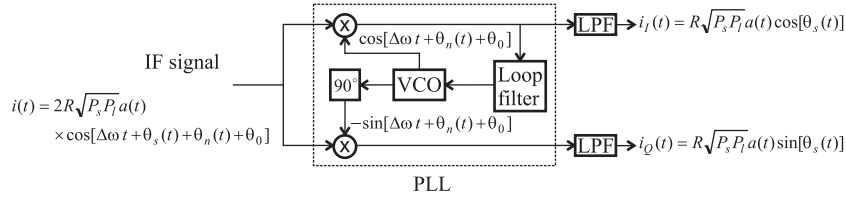


Fig. 10 Synchronous heterodyne receiver utilizing an electrical PLL. In the locked state the PLL tracks the IF signal phase noise to produce a Voltage Controlled Oscillator (VCO) output that acts as a phase reference for the I-channel.

$$i_Q(t) = R\sqrt{P_s P_l} a(t) \sin[\theta_s(t)] \quad (24)$$

which, when divided by $R\sqrt{P_s P_l}$, are equal to the I and Q-components of $A_s(t)$. A major disadvantage of the heterodyne receiver is the requirement that $|\Delta\omega| > 2\pi R_b$, so the minimum bandwidth requirement of the photodiodes is $2\Delta\omega$, which for high symbol rates can be very large. It is also difficult to design wideband low-distortion electrical mixers. An alternative to using an electrical PLL is to use an Optical PLL (OPLL) to match the LO phase to the input Signal Carrier (SC) phase; however the practical implementation of an OPLL is difficult mainly due to limitations in the achievable loop bandwidth. An additional difficulty is the relatively high phase noise of Distributed Feedback (DFB) semiconductor lasers, which are the preferred choice for the signal carrier and LO (an alternative is to use external cavity lasers, which however are of high cost compared to DFB lasers). Because of advances in high-speed DSP circuits, the practical implementation of wideband electrical PLLs is feasible.

Homodyne and Intradyne Detection

In homodyne detection $\Delta\omega=0$, in which case (22) is given by

$$i(t) = 2R\sqrt{P_s P_l} a(t) \cos[\theta_s(t) + \theta_n(t) + \theta_0] \quad (25)$$

To achieve $\Delta\omega=0$ requires synchronization of the LO frequency with the signal carrier frequency. In order to decode the transmitted symbols, the LO phase must track the SC phase noise. This can be achieved by an OPLL. When SC-LO phase synchronization is achieved, (25) gives the real part of $A_s(t)$; however the imaginary part of $A_s(t)$ cannot be detected. The latter can be detected by the use of a phase diversity receiver, discussed later in the context of intradyne reception.

In commercial coherent systems intradyne reception is employed, whereby $|\Delta\omega|$ is less than the symbol rate; but not necessarily equal to zero. This means that the SC and the LO do not need to be phase locked to each other thereby avoiding the need for an OPLL. In intradyne systems $\Delta\omega$ is referred to as the Frequency Offset (FO). The FO results in a steady angular rotation $\Delta\omega t$ experienced by the received signal constellation. With the use of commercial lasers, the magnitude of the FO may be up to several GHz. If the angular rotation over a symbol period is significant, FO estimation and compensation is necessary.

Phase Diversity Receiver

If the signal and LO are co-polarized, full recovery of $A_s(t)$ is possible by using phase-diversity detection as shown in Fig. 11. The receiver consists of a 90° Optical Hybrid (OH) and two BDs. The receiver has the major advantage of allowing carrier recovery to be carried out at baseband, which is now feasible with the availability of high-speed DSP circuits thereby negating the need for an OPLL.

Assuming non-zero $\Delta\omega$, the photocurrents from the BDs are given by

$$i_I(t) = R\sqrt{P_s P_l} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \quad (26)$$

$$i_Q(t) = R\sqrt{P_s P_l} a(t) \sin[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \quad (27)$$

Because $1/(2\pi|\Delta\omega|)$ is typically much smaller than the symbol period, (26) and (27) are effectively baseband signals, having bandwidths approximately equal to the modulation bandwidth. This means that the receiver bandwidth requirements are greatly relaxed compared to heterodyne detection, which negates the requirement for wideband electrical mixers, prior to any necessary

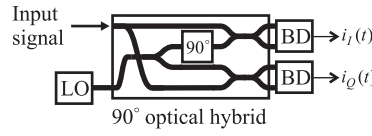


Fig. 11 Phase-diversity receiver.

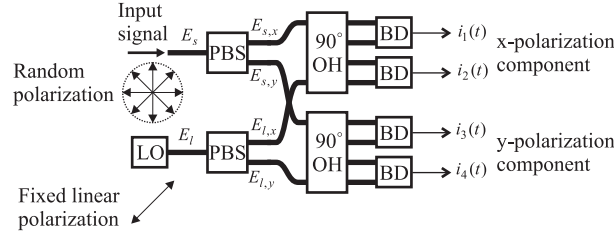


Fig. 12 Phase and polarization diversity receiver.

DSP. (26) and (27) can be combined to form a complex amplitude $A_c(t)$, given by

$$A_c(t) = \frac{1}{R\sqrt{P_l}} [i_I(t) + i_Q(t)] = \sqrt{P_s} a(t) \exp\{j[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0]\} \quad (28)$$

$A_s(t)$ can be recovered if the FO, phase noise and θ_0 are removed from (28) by applying suitable DSP.

Phase and Polarization Diversity Receiver

In practical systems the SC and LO polarizations are different. In the worst case scenario, when the SC and LO polarizations are orthogonal, the phase diversity receiver photocurrents will be zero. This problem can be overcome by the use of the phase and polarization diversity receiver shown in Fig. 12. The LO is polarized at 45° with respect to the Polarization Beam Splitter (PBS) polarization axes (x and y) to give equal LO powers at the PBS outputs. The split signal and LO components are then input to two 90° OHs. If it is assumed that the signal x- and y-polarization components are in-phase then the output photocurrents are given by

$$i_1(t) = R\sqrt{\frac{\alpha P_s P_l}{2}} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \quad (29)$$

$$i_2(t) = R\sqrt{\frac{\alpha P_s P_l}{2}} a(t) \sin[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \quad (30)$$

$$i_3(t) = R\sqrt{\frac{(1-\alpha) P_s P_l}{2}} a(t) \cos[\Delta\omega t + \theta_s(t) + \theta_n(t)] \quad (31)$$

$$i_4(t) = R\sqrt{\frac{(1-\alpha) P_s P_l}{2}} a(t) \sin[\Delta\omega t + \theta_s(t) + \theta_n(t)] \quad (32)$$

where α is the ratio of the x-polarization signal power to the total signal power. The complex amplitudes of the x- and y-polarizations corresponding to (29–32) can be written as

$$E_x(t) = \frac{1}{R} \sqrt{\frac{2}{P_l}} [i_1(t) + j i_2(t)] = \sqrt{\alpha P_s} a(t) \exp\{j[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0]\} \quad (33)$$

$$E_y(t) = \frac{1}{R} \sqrt{\frac{2}{P_l}} [i_3(t) + j i_4(t)] = \sqrt{(1-\alpha) P_s} a(t) \exp\{j[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0]\} \quad (34)$$

When the FO, phase noise and θ_0 are removed from (33,34), then $A_s(t)$ can be recovered in a polarization independent manner.

Signal-to-Noise Ratio

A major advantage of coherent detection is that the LO power can be increased to a level such that the receiver noise is negligible, in which case the receiver noise performance is determined by the Signal-to-Noise Ratio (SNR) of the received symbols. There is no difference in the performance of the phase diversity homodyne receiver and heterodyne receiver if in the latter an optical filter is used to reject optical frequency bands centered at $\omega_l \pm \Delta\omega$, which otherwise are down-converted to the same IF as the signal of interest and introduce WDM crosstalk. In a heterodyne system employing optical amplifiers, optical filtering avoids excess Amplified Spontaneous Emission (ASE) noise. The phase diversity receiver I and Q outputs can be combined to form a complex

valued signal, accompanied by additive complex valued white noise. The value of the SNR per symbol γ_s depends on the dominant noise source. In the quantum noise limited regime, encountered in links not employing optical amplifiers, γ_s is equal to the average number of photons per symbol n_s . In long-haul links, in-line optical amplifiers are used to periodically compensate for the transmission fiber loss as shown in Fig. 13, where the number of fiber spans N_a equals the number of (identical) polarization independent amplifiers. The detected noise (signal-spontaneous beat noise) is much larger than the quantum noise of the signal itself. Assuming that the noise figure of each amplifier is the minimum possible value of 3 dB, then γ_s is equal to n_s/N_a .

Spectral Efficiency and BER

The Spectral Efficiency (SE) (bit/s/Hz) of a communications link is the data rate that can be transmitted over a given bandwidth, excluding the use of error-detection codes. It is a measure of how efficiently the limited available spectrum is utilized. The maximum achievable SE (bit/s/Hz/polarization/channel) of a linear communication link in the presence of Additive White Gaussian Noise (AWGN) is given by the Shannon limit

$$SE_{\max} = \log_2(1 + SE_{\max}\gamma_b) \quad (35)$$

where γ_b is the ratio of the energy per bit to noise power spectral density, which is equal to the Signal-to-Noise Ratio (SNR) per bit. The symbol SNR $\gamma_s = b\gamma_b$. (35) can be rewritten as

$$\gamma_b = \frac{2^{SE_{\max}} - 1}{SE_{\max}} \quad (36)$$

which enables $SE_{\max}$ to be determined as a function of γ_b . The SE of a specific modulation format is given by

$$SE = \frac{\log_2(M)}{D/2} \quad (37)$$

where D is the dimension of the modulation format. In coherent systems, both the I and Q components of the carrier lightwave are used so $D=2$ for SP transmission and $D=4$ for DP transmission.

The principle metric for evaluating the performance of a communication link is the BER. In the absence of any transmission link impairments the BER only depends on γ_b . The symbol error rate and BER are approximately equal when Gray coding is used. Gray coding is not possible for PS formats; the best that can be done is to encode the levels so that constellation point pairs that are furthest away from each other have inverted binary code words, in which case the BER is approximately half the symbol error rate. Exact and approximate BER expressions for common modulation formats are given in Table 1, from which γ_b required to achieve a target BER can be calculated. A target BER of 10^{-3} is typical in systems employing FEC. Fig. 14 shows the SE as a function of γ_b for different modulation formats and also the Shannon limit, which shows that the use of high-order modulation formats increases SE at the expense of a larger SNR. It is also evident PS has the highest power efficiency among 4D formats.

Digital Coherent Receiver

In order to recover the transmitted symbols, the x- and y-polarization signals from the phase and polarization diversity optical receiver must be processed. The required processing can be carried out using a DCR, the main functional blocks of which are shown in Fig. 15. First the real and imaginary parts of the complex amplitudes of the x- and y-polarizations from the optical front-end are passed through Anti-Aliasing Filters (AAFs), digitized by a 4-channel ADC and combined to obtain the sampled complex amplitudes $E_x(n)$ and $E_y(n)$, where n is the sample number. The next stage is the removal, using a fixed equalizer, of Intersymbol Interference (ISI) caused by the fiber fixed Group Velocity Dispersion (GVD). Clock recovery, to determine the symbol rate and optimum sampling times, is usually performed after GVD compensation (Zhou, 2014). Linear polarization dependent fiber impairments are removed using adaptive equalization, and if necessary the DP signals (X and Y) can be demultiplexed. Next FO estimation and compensation are carried out followed by carrier phase detection and compensation after which the symbols are decoded and FEC applied.

Sampling

In order to minimize the DSP load, it is desirable to band limit the received signal to the minimum possible value that allows successful data recovery. The theoretical minimum bandwidth of the optical signal, termed the Nyquist bandwidth, is equal to the

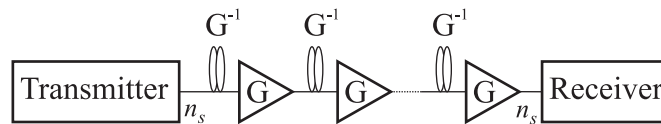
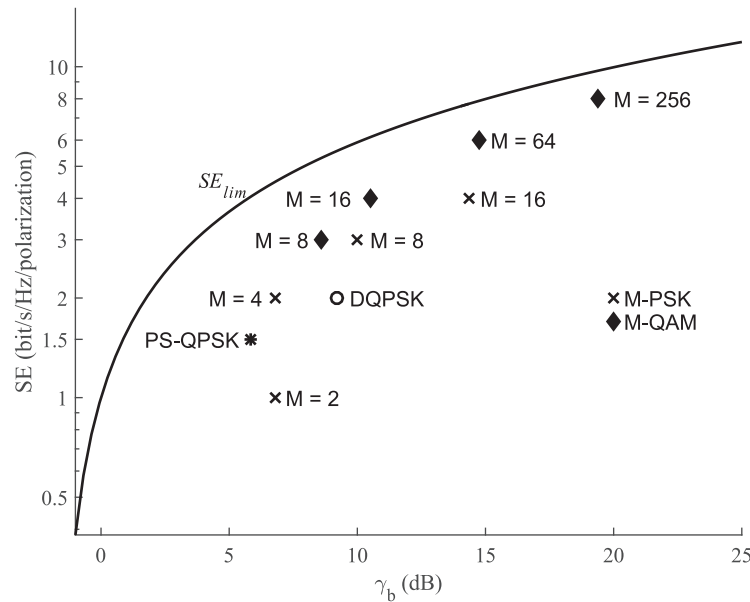


Fig. 13 Optical amplifier chain. The amplifier gain G compensates for the loss G^{-1} of the preceding fiber span.

Table 1 Exact and approximate BER expressions for various receivers assuming Gray coding, except in the case of PS-QPSK

Detection scheme	BER
Differential detection: DQPSK	$= Q(\alpha, \beta) - \frac{1}{2} I_0(\alpha\beta) \exp\left[-\frac{1}{2}(\alpha^2 + \beta^2)\right]$,
Coherent detection: BPSK and QPSK	where $\alpha = \sqrt{2\gamma_b(1 - \sqrt{1/2})}$ and $\beta = \sqrt{2\gamma_b(1 + \sqrt{1/2})}$ $= \frac{1}{2} \text{erfc}(\sqrt{\gamma_b})$
Coherent detection: M-PSK ($M > 4$)	$\approx \frac{1}{b} \text{erfc}\left[\sqrt{b\gamma_b} \sin\left(\frac{\pi}{M}\right)\right]$
Coherent detection: M-QAM	$\approx \frac{2}{b} \left(1 - \frac{1}{\sqrt{M}}\right) \text{erfc}\left[\sqrt{\frac{3b\gamma_b}{2(M-1)}}\right]$
Coherent detection: PS-QPSK	$\approx \frac{1}{2} \left\{1 - \frac{1}{\sqrt{\pi}} \int_0^\infty [1 - \text{erfc}(x)]^3 \exp\left[-(x - \sqrt{3\gamma_b})^2\right] dx\right\}$

Q : Marcum Q-function; I_0 : zero-order modified Bessel function of the first kind; erfc : complementary error function.

**Fig. 14** Spectral efficiency vs. SNR per bit for different modulation formats for a BER = 10^{-3} . The Shannon limit is also shown. 4-QAM is equivalent to 4-PSK. PS based formats can only operate in DP mode.

symbol rate (per polarization), so the minimum bandwidth of the baseband I and Q signals is $R_s/2$, which means that the theoretical minimum required sampling rate is equal to R_s . It is not possible to construct an ideal AAF that band limits the polarization diversity receiver outputs to a bandwidth equal to $R_s/2$; practical AAFs have an amplitude response that extends beyond $R_s/2$. Symbol-rate sampling is also susceptible to sampling time errors. This limitation can be overcome by using a sampling rate pR_s , where $1 < p < 2$ is a rational oversampling ratio. This also avoids aliasing up to a signal bandwidth of $pR_s/2$. Clock recovery and adaptive equalization require two-times oversampling, so if fractional sampling is used, the sampled signals must be resampled at a rate of $2R_s$, which increases the processing complexity.

Group Velocity Dispersion Equalization

GVD is caused by CD, which is a combination of waveguide and material dispersion. The fiber refractive index $n(\omega)$ is frequency dependent, which leads to a frequency dependent group velocity $v_g(\omega) = c/n(\omega)$. Each optical pulse contains a continuum of components having frequencies within the pulse bandwidth, which travel at different group velocities. Thereby each of these components have different arrival times at the receiver, resulting in pulse broadening and ISI, the severity of which depends on the

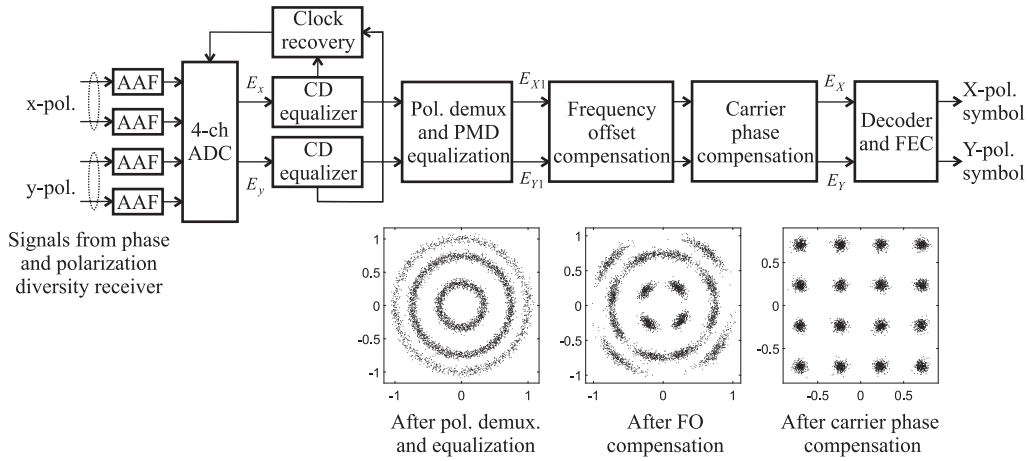


Fig. 15 DCR configuration showing a square 16-QAM signal constellation at various stages in the signal recovery process.

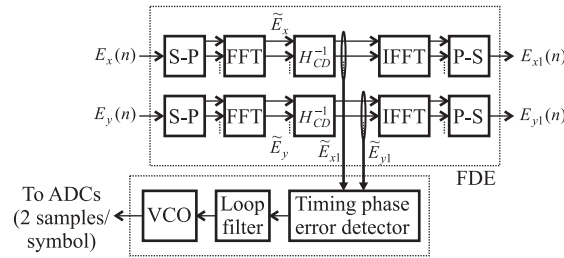


Fig. 16 General configuration of a FDE and frequency domain based clock extraction.

pulse bandwidth, transmission length and the fiber dispersion at the operating wavelength. In the frequency domain the fiber CD has the scalar transfer function

$$H_{CD}(\omega) = \exp \left[-jL \left(\frac{\beta_2 \omega^2}{2} + \frac{\beta_3 \omega^3}{6} \right) \right] \quad (38)$$

where ω is the angular frequency deviation from ω_s and L , β_2 and β_3 are the fiber length, dispersion parameter and dispersion slope respectively at ω_s . Because the GVD is time invariant it can be removed from each polarization tributary by the use of a digital filter, operating at two samples per symbol, having a transfer function H_{CD}^{-1} (equal to the complex conjugate H_{CD}^*). There are two ways of implementing the filter; a Finite Impulse Response (FIR) filter or a Frequency Domain Equalizer (FDE). FIR filters operate on serial data but are difficult to implement if the number of required taps is large. FDE equalization greatly reduces the computational load by using block-by-block processing; with the additional advantage that it is especially suited for use with some common clock recovery schemes. The general structure of a FDE is shown in Fig. 16. The digitized x- and y-polarization complex amplitudes are divided into blocks (of suitable size N) using a Serial-to-Parallel (s-P) converter and transformed to the frequency domain by calculating their Fourier transforms $\tilde{E}_x$ and $\tilde{E}_y$, using the Fast Fourier Transform (FFT) algorithm. If two-times oversampling is used, the discrete frequency k components of the FFTs range from $-1/T_s$ to $1/T_s$, with a resolution equal to $2/NT_s$. The spectral components are then multiplied by H_{CD}^{-1} , reconverted to the time domain by applying an Inverse FFT (IFFT), and serialized by a Parallel-to-Serial (p-S) converter. In practical systems, the GVD value is monitored by a pilot tone, added at the transmitter, and the required filter tap coefficients determined from the monitored GVD value. A major advantage of using electrical GVD compensation is that it eliminates the need for dispersion-compensation fiber.

Clock Recovery

After GVD compensation, clock recovery is carried out to determine the symbol clock frequency and optimum sampling times. Clock recovery can be carried out using well known techniques such as the Gardner method, which is particularly useful if the received signal contains some energy at the clock frequency. Two samples per symbol and knowledge of the previous symbol timing are required in order to estimate the timing error for the current symbol. If the x-polarization signal is considered, the timing error τ_e is calculated, after CD equalization, as

$$\tau_e = \sum_{n=0}^{N/2-1} [E_{x1}(2n-1) - E_{x1}(2n+1)] E_{x1}^*(2n) \quad (39)$$

which is computed in the frequency domain as

$$\tau_e = \sum_{k=0}^{N/2-1} \text{Im} [\tilde{E}_{x1}(k) \tilde{E}_{x1}^*(k + N/2)] \quad (40)$$

where E_{x1} and $\tilde{E}_{x1}$ are the CD compensated x-polarization signal and its FFT respectively. The estimated timing error is then passed through a loop-filter, whose output controls the frequency and phase of the clock VCO, in such a way as to optimize the sampling times. This method works well when the received optical signals do not have significant PMD. In order to tolerate large PMD, both x- and y-polarizations need to be considered. A modified Gardner method that uses both polarizations, calculates the frequency domain timing phase error as

$$\tau_e = \sum_{k=0}^{N/2-1} \text{Im} \left\{ \left[\tilde{E}_{x1}(k) + \tilde{E}_{y1}(k) e^{j(\phi)_U} \right] \left[\tilde{E}_{x1}^*(k + N/2) + \tilde{E}_{y1}^*(k + N/2) e^{j(\phi)_L} \right] \right\} \quad (41)$$

which includes rotational phase angles ϕ_U and ϕ_L in the positive and negative frequency components respectively. Appropriate values of ϕ_U and ϕ_L are chosen to mitigate the effects of first-order PMD by introducing a relative phase between the x- and y-polarizations.

PMD and PDL Equalization and Polarization Demultiplexing

PMD is caused by the differential arrival time of the different polarization components of an input light pulse at the fiber output (Galtarossa and Menyuk, 2006). In single-mode fiber PMD is due to the fiber birefringence (polarization dependent effective refractive index) caused by noncircularity of the core and inhomogeneity of the fiber as well as external stress-induced material birefringence, such as bends and twists. The birefringence changes randomly along the fiber length and is wavelength and time dependent. Modeling the propagation of a pulse through a long length of fiber is complicated because of random polarization mode coupling and pulse splitting at every change in the local birefringence.

PMD is commonly described using the principal states model. For a given length of fiber and in the absence of PDL there exist two orthogonal polarization modes (Principal States of Polarization, PSPs) at the input for which, to a first-order approximation, the output polarization is independent of the frequency content of the lightwave. The PSPs propagate at different speeds according to a slow and fast axis induced by the birefringence of the fiber. For each pair of input PSPs there is a corresponding pair of orthogonal PSPs at the fiber output. When a light pulse is launched into any polarization state other than a PSP, the two polarization components slowly separate so the resulting composite pulse is broadened and distorted as shown in Fig. 17. The time difference between the fast and slow polarization states after a given propagation distance is called the instantaneous Differential Group Delay (DGD) $\Delta\tau$. Together, the frequency independent DGD and the PSPs are the fundamental manifestations of first-order PMD. The instantaneous value of $\Delta\tau$ (at a particular wavelength) at the fiber output is a random variable. From experiment and theory it has been shown that $\Delta\tau$ has a Maxwellian probability distribution entirely specified by a single parameter, the mean DGD $\langle\Delta\tau\rangle$, called the PMD of the fiber. For fiber lengths usually encountered in optical transmission systems, $\langle\Delta\tau\rangle$ is proportional to the square root of the fiber length. The constant of proportionality is called the PMD coefficient $\text{PMD}_{\text{coeff}}$ which for new types of fiber is typically of the order of $0.1 \text{ ps}/\sqrt{\text{km}}$. Older installed fiber has $\text{PMD}_{\text{coeff}}$ values which are usually at least a magnitude higher. The average value of the frequency interval over which the PSPs are frequency independent is called the PSP bandwidth $\Delta\nu_{\text{PSP}}$, which is inversely proportional to $\langle\Delta\tau\rangle$. A reasonably accurate value of $\Delta\nu_{\text{PSP}}$ in the 1550 nm region is 125/

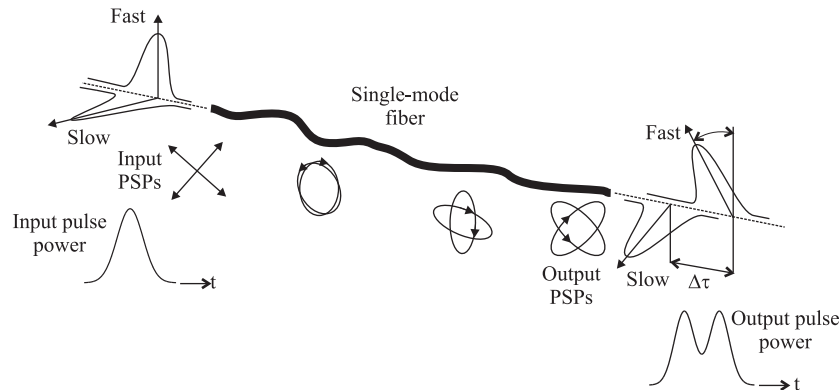


Fig. 17 First-order PMD and pulse distortion in an optical fiber. It is assumed that the input pulse power is equally distributed into the fiber input slow and fast PSPs. In the absence of PDL, the output PSPs are orthogonal and rotated with respect to the input PSPs. Slowly varying birefringence along the fiber leads to rotation of the PSPs. The differential group delay between the PSPs leads to pulse distortion.

$\langle\Delta\tau\rangle$ GHz, when $\langle\Delta\tau\rangle$ is expressed in ps. Considering a fiber with $PMD_{coeff}=0.1ps/km$ and a length of 1000 km, $\langle\Delta\tau\rangle=1ps$ and $\Delta\nu_{PSP}=125$ GHz, so in this case it can be assumed that the first order PMD approximation is valid for data having symbol rates of less than 125 Gbaud.

When the spectral width of the signal exceeds the PSP bandwidth, both the DGD and the PSPs vary across the signal spectrum, which leads to higher order PMD effects. Second order PMD has two components. The first component is polarization dependent CD, which adds or subtracts from the fiber CD depending on the input polarization state to the fiber. The second and dominant component is rotation of the PSPs across the pulse spectrum, which results in pulse distortion such as overshoots and generation of satellite pulses.

For a given modulation format and data rate, the impact of first-order PMD depends on the DGD and relative intensities of the light in the PSPs. The impairment is most severe when equal amounts of the transmitted light pulse are coupled into the input PSPs, and the impairment is negligible when all of the light is coupled to a single PSP. The transmitted pulses experience distortion and broadening, with the possibility of ISI, the consequence of which is a limitation in the acceptable transmission distance. While some simple practical rules for the maximum tolerable mean DGD in IM-DD schemes are available; in general evaluation of the effects of PMD on signal transmission requires complex simulations. The effects of PMD can be reduced by the use of low PMD_{coeff} fiber. The use of high-order modulation formats, for a given data rate, leads to a reduction in the optical signal bandwidth, thereby reducing the impact of PMD. Optical or electrical compensation offers the potential of reducing PMD impairment to acceptable levels, but because PMD can fluctuate on time scales of the order of a millisecond the compensators need to be rapidly adjustable. The availability of fast DSP circuits has made electrical compensation the preferred option. In the DCR, GVD compensation is followed by PMD equalization and polarization demultiplexing.

The complex amplitude of the optical signal at the input of the transmission system can be written as a vector $[E_{x,in}(t), E_{y,in}(t)]^T$, where T denotes the transpose and $E_{x,in}(t)$ and $E_{y,in}(t)$ are the x- and y-polarization complex amplitudes respectively. The Fourier transform of the received complex amplitude is given by

$$\begin{bmatrix} \tilde{E}_x(\omega) \\ \tilde{E}_y(\omega) \end{bmatrix} = \mathbf{H}_o(\omega) \begin{bmatrix} \tilde{E}_{x,in}(\omega) \\ \tilde{E}_{y,in}(\omega) \end{bmatrix} \quad (42)$$

where $[\tilde{E}_{x,in}(\omega), \tilde{E}_{y,in}(\omega)]^T$ is the Fourier transform of $[E_{x,in}(t), E_{y,in}(t)]^T$, ω is the angular frequency deviation from ω_s and the 2×2 matrix $\mathbf{H}_o(\omega)$ is the link transfer function. In the linear region $\mathbf{H}_o(\omega)$ can be modeled as $\mathbf{H}_o(\omega) = H_{CD}(\omega)\mathbf{H}_p(\omega)$, where the link polarization properties are taken into account by the matrix $\mathbf{H}_p(\omega) = \mathbf{U}(\omega)\mathbf{T}\mathbf{K}$. The frequency dependent PMD matrix $\mathbf{U}(\omega)$ is given by

$$\mathbf{U}(\omega) = \mathbf{R}_1^{-1} \begin{bmatrix} \exp(j\omega\Delta\tau/2) & 0 \\ 0 & \exp(-j\omega\Delta\tau/2) \end{bmatrix} \mathbf{R}_1 \quad (43)$$

where $\mathbf{R}_1$ is a unitary matrix converting each PSP into the x- or y-polarization component. The frequency independent PDL matrix $\mathbf{T}$ is given by

$$\mathbf{T} = \mathbf{R}_2^{-1} \begin{bmatrix} \sqrt{\alpha_{max}} & 0 \\ 0 & \sqrt{\alpha_{min}} \end{bmatrix} \mathbf{R}_2 \quad (44)$$

where $\mathbf{R}_2$ is a unitary matrix converting each PDL eigenmode (polarization states for which the PDL is either a maximum or minimum) into an x- or y-polarization component. α_{max} and α_{min} are the power transmission coefficients of the PDL eigenmodes. $\mathbf{K}$ is a frequency independent 2×2 unitary matrix which mixes the two polarizations at the receiving end of the fiber.

Because all of the above impairments are linear, they can be removed by passing $[\tilde{E}_x(\omega), \tilde{E}_y(\omega)]^T$ through an equalizer having a transfer function equal to $\mathbf{H}_o^{-1}(\omega)$. Because the fiber CD is fixed or varies very slowly it is first removed by using FDE equalization as described above. Polarization demultiplexing and polarization impairment equalization can be implemented by filtering the sampled CD compensated signal $[E_{x1}(n), E_{y1}(n)]^T$ with an equalizer having a transfer function $\mathbf{H}_{eq}(\omega) = \mathbf{H}_p^{-1}(\omega)$. Because $\mathbf{H}_p(\omega)$ is time dependent, adaptive equalization is required. $\mathbf{H}_{eq}(\omega)$ can be written as

$$\mathbf{H}_{eq}(\omega) = \begin{bmatrix} h_{xx}(\omega) & h_{xy}(\omega) \\ h_{yx}(\omega) & h_{yy}(\omega) \end{bmatrix} \quad (45)$$

$\mathbf{H}_{eq}(\omega)$ can be implemented as a 2×2 butterfly structured filter, each element of which is realized by FIR filters with tap coefficient vectors $\vec{h}_{xx}$, $\vec{h}_{xy}$, $\vec{h}_{yx}$ and $\vec{h}_{yy}$, of equal lengths k , as shown in Fig. 18. The most common method for adaptively finding the optimum tap coefficients is the Constant Modulus Algorithm (CMA). The input signal is normalized such that $|E_{x1}(n)|=|E_{y1}(n)|=1$, as is the case for NRZ M-PSK type signals. The tap coefficient vectors are updated on a symbol by symbol basis as follows:

$$\vec{h}_{xx}(n+1) = \vec{h}_{xx}(n) + \mu e_x(n) \vec{E}_{x1}^*(n) \quad (46)$$

$$\vec{h}_{xy}(n+1) = \vec{h}_{xy}(n) + \mu e_x(n) \vec{E}_{y1}^*(n) \quad (47)$$

$$\vec{h}_{yx}(n+1) = \vec{h}_{yx}(n) + \mu e_y(n) \vec{E}_{x1}^*(n) \quad (48)$$

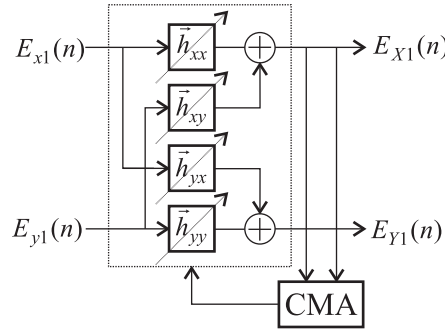


Fig. 18 Configuration of a 2×2 butterfly structured adaptive filter.

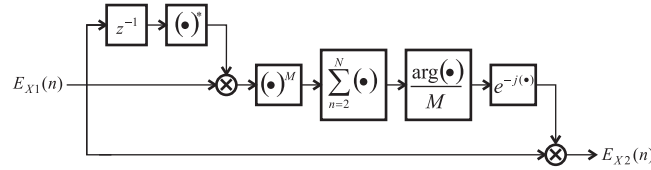


Fig. 19 DSP implementation of the time-domain differential phase M-th power FO estimation and compensation.

$$\vec{h}_{yy}(n+1) = \vec{h}_{yy}(n) + \mu e_y(n) \vec{E}_{y1}^*(n) \quad (49)$$

where $e_x(n) = [1 - |E_{X1}(n)|^2]E_{X1}(n)$ and $e_y(n) = [1 - |E_{Y1}(n)|^2]E_{Y1}(n)$ are error signals, μ a step-size parameter and $\vec{E}_{x1}(n)$ and $\vec{E}_{y1}(n)$ are input vectors comprising the current input sample and the previous k samples. When the algorithm converges (such that $e_x(n) \rightarrow 0$ and $e_y(n) \rightarrow 0$) the polarization dependent impairments are compensated and polarization demultiplexed X- and Y-polarization components $E_{X1}(n)$ and $E_{Y1}(n)$ respectively of the fiber input PDM signals will be present at the equalizer output ports. If SP transmission is used, the signal of interest will be present at only one of the ports. CMA based equalization works well for M-PSK signals but is difficult to apply to M-QAM signals, whose amplitude has multiple levels.

FO Compensation

The adaptive equalizer does not remove the FO or the phase noise from the received signal. A non-zero FO results in the signal experiencing a phase rotation $\Delta\omega T_s$ over a symbol period. After adaptive equalization, the demultiplexed X- and Y-polarization signals are downsampled to one sample per symbol. If we consider the X-polarization component, the downsampled complex amplitude at the equalizer output has the form

$$E_{X1}(n) = a(n) \exp\{j[\Delta\omega t_n + \theta_s(n) + \theta_n(n) + \theta_0]\} \quad (50)$$

where t_n is the sampling time and $a(n)$, $\theta_s(n)$ and $\theta_n(n)$ are the sampled modulation signal amplitude and phase and phase noise. In order to remove the FO, it must first be estimated and then (50) reverse rotated by the phase offset $\Delta\omega T_s$ accumulated over a symbol period. For M-PSK modulated signals a common technique for estimating the FO is the time-domain differential phase M-th power algorithm. Considering an M-PSK signal constant $a(n)$, the phase modulation is removed by raising (50) to the M-th power to give

$$E_{X1}^M(n) = a^M \exp\{jM[\Delta\omega t_n + \theta_n(n) + \theta_0]\} \quad (51)$$

The complex conjugate of the previous sample raised to the power of M is given by

$$[E_{X1}^*(n-1)]^M = a^M \exp\{-jM[\Delta\omega(t_n - T_s) + \theta_n(n-1) + \theta_0]\} \quad (52)$$

Multiplying (51) by (52) gives

$$E_{X1}^M(n)[E_{X1}^*(n-1)]^M = a^{2M} \exp\{jM[\Delta\omega T_s + \Delta\theta_n(n)]\} \quad (53)$$

where $\Delta\theta_n(n) = \theta_n(n) - \theta_n(n-1)$ is the phase noise induced random phase within a symbol period. The phase noise can be removed by summing (53) over a suitably long number $N-1$ of single estimates and the FO estimate $\Delta\omega_e$ calculated as

$$\Delta\omega_e = \frac{1}{M T_s} \arg \left[\sum_{n=2}^N E_{X1}^M(n)[E_{X1}^*(n-1)]^M \right] \quad (54)$$

The implementation of this algorithm and subsequent removal of the FO is shown in Fig. 19. Because the arg operator has a range of $[-\pi, \pi]$, the maximum resolvable FO is $\pm 1/2MT_s$, which for a 40 Gbaud 8-PSK signal is ± 2.5 GHz. Therefore coarse tuning of the LO may be necessary to ensure that the FO is within the resolvable range. The performance of the M-power method

for high-order QAM is poor because only a small amount of the constellation points having equal phase separation is useable for FO estimation.

The FO can also be estimated by computing the FFT of the data erased signal, which has a peak at $M\Delta\omega$. The accuracy of the estimated FO can be greater than that for the time-domain method; however the computational complexity is much greater. It is not possible to determine the sign of the FO; to do so, necessitates further signal processing. Other FO estimation methods include blind frequency search (no training sequence is required), where the estimated FO is scanned over a particular range and the optimal FO determined using a minimum phase or Euclidian distance method. Another method uses a starting training sequence to obtain an initial FO estimate and then uses the recovered phase angle from the following carrier phase recovery unit to track the FO change. Both of these techniques work for arbitrary modulation formats.

Carrier Phase Compensation and Decoding

After the FO has been compensated, the phase difference between the received signal and LO must be removed before the final decoding stage. The X-polarization complex amplitude after FO compensation is given by

$$E_{X2}(n) = a(n)\exp\{j[\theta_s(n) + \theta_n(n) + \theta_0]\} \quad (55)$$

The linewidth of DFB lasers used in the transmitter and receiver is usually in the range of 100 kHz to 10 MHz, so the phase noise varies much more slowly than the phase modulation. In practice (55) is accompanied by additive noise, originating from the optical signal quantum noise, ASE noise if optical amplifiers are present in the transmission link and receiver electronic noise. It is possible to obtain an accurate estimate of the phase noise and offset by averaging the phase of (55) over a large number N of symbol periods. For M-PSK signals the feedforward M-th power method can be used; as shown in Fig. 20. First, each symbol in the l -th block $E_{X2,1}$ of N symbols is raised to the M-th power to erase the phase modulation. The phase estimate over the l -th block is then calculated as

$$\theta_l = \frac{1}{M} \arg \left\{ \sum_{n=1}^N E_{X2,l}[n + (l-1)N] \right\} \quad (56)$$

Averaging reduces the effect of additive noise but the longer the averaging period the worse the phase estimate; hence the optimal block length is a trade-off between the additive noise and the phase noise; narrower linewidths requiring longer lengths. The symbols are then corrected by reverse rotating the symbols in the block by θ_l . Because the arg operator has a range of $[-\pi, \pi]$, the phase estimate is limited to values between $\pm\pi/M$ and so the resulting symbols will have a phase ambiguity of $2\pi/M$. This can be overcome by the use of differential modulation; however this results in a doubling of the error rate. After phase estimation, $E_{X2,1}$ is reverse rotated by an angle $-\theta_l$ to obtain the phase corrected block $E_{X,1}$, with corresponding complex amplitudes

$$E_X(n) = a(n)\exp[j\theta_s(n)] \quad (57)$$

which represents the correctly aligned X-polarization symbols. In practice (57) will also have an additive noise component due to the intensity noise of the optical signals and receiver noise. The complex amplitudes $E_X(n)$ are then decoded and FEC applied to recover the X-polarization symbols with an acceptable BER.

The M-th power method cannot be directly applied to high-order QAM formats because of the multiplicity of symbol amplitude levels. In the case of square 16-QAM the symbols can be divided into two classes: Class I and Class II as shown in Fig. 21. Class I symbols can be regarded as QPSK ($M=4$) signals having two amplitude levels; the phase can be estimated by using only these symbols in the 4-th power technique and the estimated phase then used to correct the entire QAM constellation. Class II symbols are not used for the phase estimation.

Nonlinear Impairments

The dominant nonlinear impairment in optical fiber is due to the Kerr effect, which causes a refractive index change proportional to the signal intensity, resulting in nonlinear waveform distortion. This can severely limit the maximum transmission distance of high-order QAM signals. FIR based adaptive filters cannot be used to compensate for such nonlinear effects. The propagation of a DP signal though an optical fiber, in the presence of fiber attenuation, CD and Kerr nonlinearity is described by the nonlinear

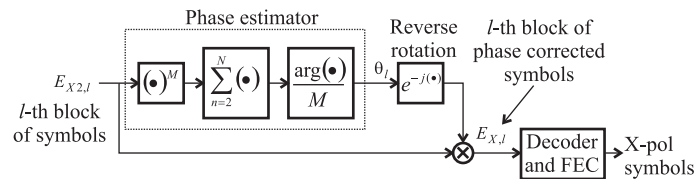


Fig. 20 DSP implementation of feedforward M-th power phase estimation and compensation for M-PSK signals.

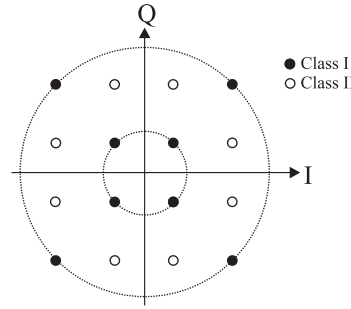


Fig. 21 Square 16-QAM symbol classification.

Schrodinger equation

$$\frac{\partial E_{x,y}}{\partial z} = -\frac{\alpha}{2}E_{x,y} - \frac{j}{2}\beta_2 \frac{\partial^2 E_{x,y}}{\partial t^2} + j\gamma(|E_x|^2 + |E_y|^2)E_{x,y} \quad (58)$$

where α is the attenuation coefficient, γ the nonlinear coefficient and z and t are the propagation distance and time respectively. By reversing the signs of the coefficients in (58), the virtual back-propagation process is described by

$$\frac{\partial E_{x,y}}{\partial z} = \frac{\alpha}{2}E_{x,y} + \frac{j}{2}\beta_2 \frac{\partial^2 E_{x,y}}{\partial t^2} - j\gamma(|E_x|^2 + |E_y|^2)E_{x,y} \quad (59)$$

If the received signal is digitally back-propagated using (59) the effects of deterministic nonlinear impairments can be reversed. If present, optical amplifiers can be included in the back-propagation process as virtual lumped amplifiers having gains equal to the inverse of the link amplifier gains. (59) is usually solved using the split-step Fourier algorithm; however the computational complexity is very high, thereby precluding their use in commercial systems. Therefore there is a critical need for develop improved algorithms that can provide nonlinearity compensation at feasible computational cost. In WDM systems, the signal also suffers deterministic nonlinear effects due to the fields of the neighboring channels, such as cross-phase modulation and four-wave mixing. In amplified systems, Nonlinear Phase Noise (NLPN) results from the interaction between ASE noise and the signal via the Kerr effect. Developing computationally efficient methods for compensation of non-deterministic nonlinear impairments is also of great importance.

Conclusion

Coherent lightwave systems offer the potential to realize ultra-fast communication systems. The combination of coherent detection and the DCR provides new capabilities that are possible without optical signal phase detection. 100 Gb/s DP-QPSK DCR systems are in wide commercial use for core and metropolitan networks. Data rates as high as 400 Gb/s using DP 16-QAM have recently been demonstrated. Higher order format transmission at even higher data rates is an active area of research. Photonic integration technology, including integration with electronic circuits, is undergoing rapid progress enabling more complex optical transmitters and receivers, including LO laser sources, leading to improved performance and reliability. The development of high-speed ASICs, ADCs and DACs underpins the implementation of advanced DCRs that can carry out more complex real-time DSP functions. Concerning the DSP functions, there is a need for flexible and computationally efficient algorithms that can be used to implement adaptive modulation and coding to mitigate transmission link impairments, especially fiber nonlinearities. As coherent lightwave systems technology matures it will also be applied to medium and short reach applications such as data centers and access networks.

See also: Atom Optics. Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

References

- Djordjevic, I.B., Arabaci, M., Minkov, L.L., 2009. Next generation FEC for high-capacity communication in optical transport networks. *Journal of Lightwave Technology* 27, 3518–3530.
- Galtarossa, A., Menyuk, C.R. (Eds.), 2006. *Polarization Mode Dispersion 1*. Berlin Heidelberg: Springer-Verlag.
- Gnauck, A.H., Winzer, P.J., 2005. Optical phase-shift-keyed transmission. *Journal of Lightwave Technology* 23, 115–130.
- Ip, E., Lau, A.P.T., Barros, D.J., Kahn, J.M., 2008. Coherent detection in optical fiber systems. *Optics Express* 16, 753–791.
- Karlsson, M., Agrell, E., 2009. Which is the most power-efficient modulation format in optical links? *Optics Express* 17, 10814–10819.
- Kikuchi, K., 2016. *Fundamentals of coherent optical fiber communications*. *Journal of Lightwave Technology* 34, 157–179.
- Krongold, B., Pfau, T., Kaneda, N., Lee, S.C.J., 2012. Comparison between PS-QPSK and PDM-QPSK with equal rate and bandwidth. *IEEE Photonics Technology Letters* 24, 203–205.

- Nakazawa, M., Kikuchi, K., Miyazaki, T. (Eds.), 2010. High Spectral Density Optical Communication Technologies 6. Berlin Heidelberg: Springer-Verlag.
- Roberts, K., O'Sullivan, M., Wu, K.T., *et al.*, 2009. Performance of dual-polarization QPSK for optical transport systems. *Journal of Lightwave Technology* 27, 3546–3559.
- Seimetz, M., 2010. High-Order Modulation for Optical Fiber Transmission. Berlin Heidelberg: Springer-Verlag.
- Winzer, P.J., 2012. High-spectral-efficiency optical modulation formats. *Journal of Lightwave Technology* 30, 3824–3835.
- Zhou, X., 2014. Efficient Clock and Carrier Recovery Algorithms for Single-Carrier Coherent Optical Systems, 31. *IEEE Signal Processing Magazine*. pp. 35–45.

Measuring Fiber Characteristics

A Girard, EXFO, Quebec, Canada

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The optical fiber is divided in two types: multimode and singlemode. Each type is used in different applications and wavelength ranges and is consequently characterized differently. Furthermore, the corresponding test methods also vary.

The optical fiber characteristics may be divided into four categories:

- The optical characteristics (transmission related);
- The dimensional characteristics;
- The mechanical characteristics; and
- The environmental characteristics.

These categories will be reviewed in the following sections, together with their corresponding test methods.

Fiber Optical Characteristics and Corresponding Tests Methods

The following sections will describe the following optical characteristics:

- attenuation;
- macrobending sensitivity;
- microbending sensitivity;
- cut-off wavelength;
- multimode fiber bandwidth;
- differential mode delay for multimode fibers;
- chromatic dispersion;
- polarization mode dispersion;
- polarization crosstalk; and
- nonlinear effects.

Attenuation

The spectral attenuation of an optical fiber follows exponential power decay from the power level at a cross-section 1 to the power level at cross-section 2, over a fiber length L as follows:

$$P_2(\lambda) = P_1(\lambda) \cdot e^{-\gamma(\lambda)L} \quad (1)$$

$P_1(\lambda)$ is the optical power transmitted through the fiber core cross-section 1, expressed in mW; $P_2(\lambda)$ is the optical power transmitted through the fiber core cross-section 2 away from cross-section 1, expressed in mW; $\gamma(\lambda)$ is the spectral attenuation coefficient in linear units, expressed in km^{-1} ; and L is the fiber length expressed in km.

Attenuation may be characterized at one or more specific wavelengths or as a function of wavelength. In the later case, attenuation is referred to spectral attenuation. Fig. 1 illustrates such power decay.

Eq. (1) may be expressed in relative units as follows:

$$\log_{10} P_2 = (\log_{10} P_1) - \gamma L \cdot \log_{10} e \quad (2)$$

P is expressed in dBm units using the following definition.

The power in dBm is equal to 10 times the logarithm in base 10 of the power in mW; or

$$0 \text{ dBm} = 10 \log_{10}(1 \text{ mW}) \quad (3)$$

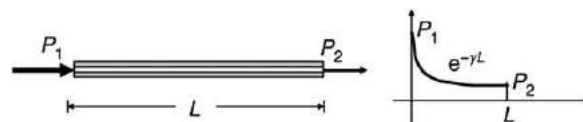


Fig. 1 Power decay in an optical fiber due to attenuation.

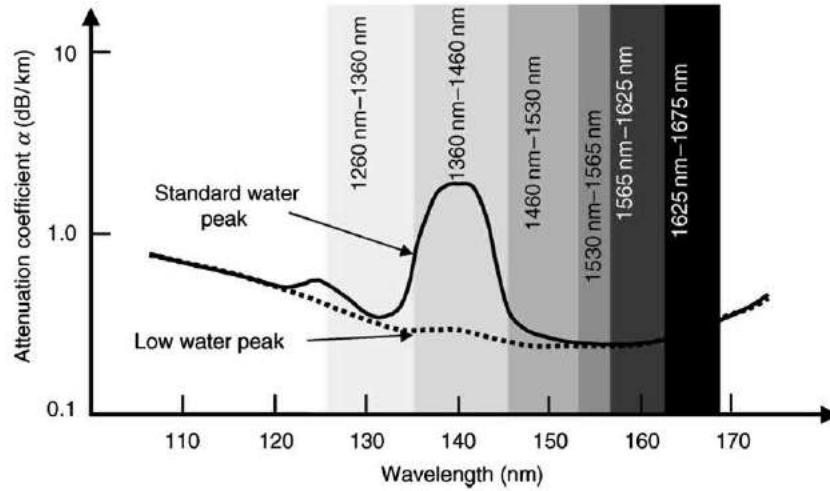


Fig. 2 Typical spectral attenuation of a singlemode fiber.

Then

$$\gamma(\lambda)[\text{km}^{-1}] = \frac{[(10\log_{10}P_1) - (10\log_{10}P_2)]}{L\log_{10}e} \quad (4)$$

$$\alpha(\lambda)[\text{dB/km}] = [P_1(\text{dBm}) - P_2(\text{dBm})]/L \quad (5)$$

A new relative attenuation coefficient $\alpha(\lambda)$ with units of dB/km has been introduced, which is related to the linear coefficient $\gamma(\lambda)$ with units of km^{-1} as follows:

$$\alpha(\lambda) = (\log_{10}e) \cdot \gamma(\lambda) \approx 4,343\gamma(\lambda) \quad (6)$$

$\alpha(\lambda)$ is the spectral attenuation coefficient of a fiber and is illustrated in **Fig. 2**.

Test methods for attenuation

The attenuation may be measured by the following methods:

- cut-back method;
- backscattering method; and
- insertion loss method.

Cut-back method

The cut-back method is a direct application of the definition in which the power levels P_1 and P_2 are measured at two points of the fiber change of input conditions. P_2 is the power emerging from the far end of the fiber and P_1 is the power emerging from a point near the input after cutting the fiber.

The output power P_2 is recorded from the fiber under test (FUT) placed in the measurement setup. Keeping the launching conditions fixed, the FUT is cut to the cut-back length, for example 2 m from the launching point. The FUT attenuation, between the points where P_1 and P_2 have been measured, is calculated using P_1 and P_2 , from the definition equations provided above.

Backscattering method

The attenuation coefficient of a singlemode fiber is characterized using bidirectional backscattering measurements. This method is also used for:

- attenuation uniformity;
- optical continuity;
- physical discontinuities;
- splice losses; and
- fiber length.

An optical time domain reflectometer (OTDR) is used for performing such characterization. Adjustment of laser pulsewidth and power is used to obtain a compromise between resolution (a shorter pulsewidth provides a better resolution but at lower power) and dynamic range/fiber length (higher power provides better dynamic range but with longer pulsewidth). An example of such an instrument is shown in **Fig. 3**.

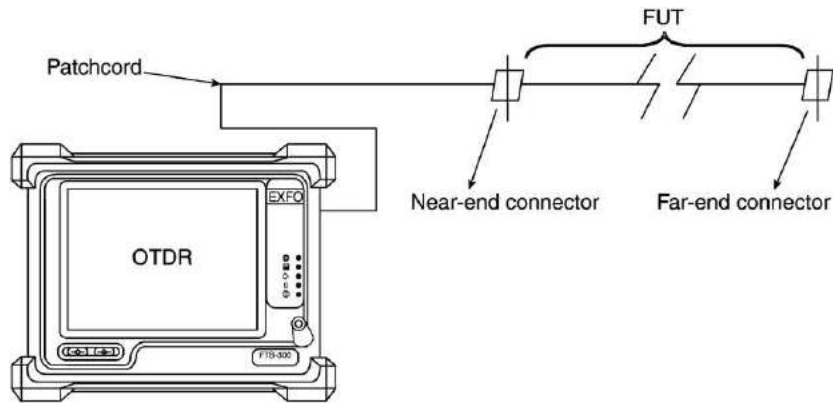


Fig. 3 The backscattering method (OTDR).

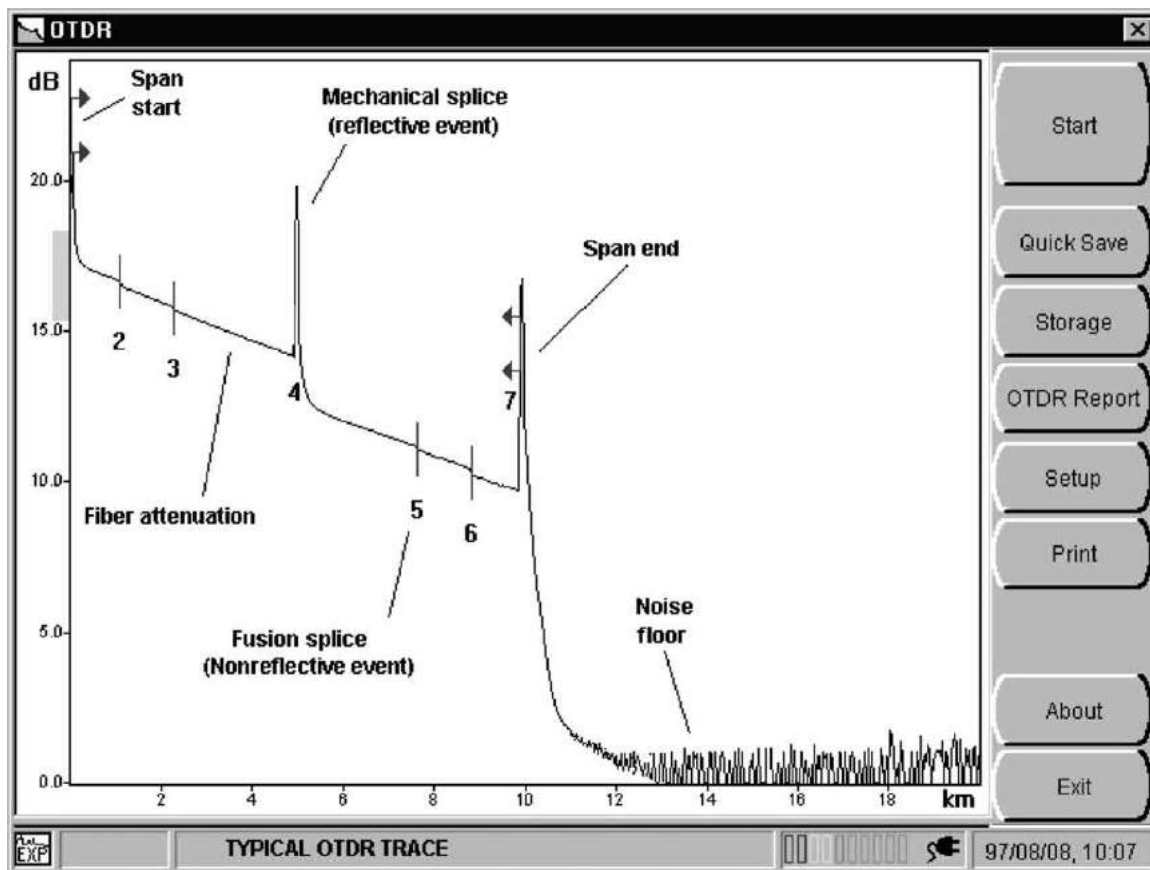


Fig. 4 Unidirectional OTDR backscattering loss measurement.

The measurement is applicable to various FUT configurations (e.g., cabled fiber in production or deployed in the field, fiber on a spool, etc.). Two unidirectional backscattering curves are obtained, one from each end of the fiber (Fig. 4).

Each backscattering curve is recorded on a logarithmic scale, avoiding the parts at the two ends of the curves, due to parasitic reflections.

The FUT length is found from the time interval between the two ends of the backscattering loss curve together with the FUT group index of refraction, n_g , as:

$$L_f = c_{fs} \cdot \frac{\Delta T_f}{n_g} \quad (7)$$

c_{fs} is the velocity of light in free space.

The bidirectional backscattering curve is obtained using two unidirectional backscattering curves and calculating the average loss between them. The end-to-end FUT attenuation coefficient is obtained from the difference between two losses divided by the difference of their corresponding distances.

Insertion loss method

The insertion loss method consists of the measurement of the power loss due to the FUT insertion between a launching and a receiving system, previously interconnected (reference condition). The powers P_1 and P_2 are evaluated in a less straightforward way than in the cut-back method. Therefore, this method is not intended for use in manufacturing.

The insertion loss technique is less accurate than the cut-back, but has the advantage of being non-destructive for the FUT. Therefore, it is particularly suitable in the field.

Macrobending Sensitivity

Macrobending sensitivity is the property by which there is a certain amount of light leaking (loss) into the cladding when the fiber is bent and the bending angle is such that the condition of total internal reflection is no longer met at the core-cladding interface.

Fig. 5 illustrates such a case.

Macrobending sensitivity is a direct function of the wavelength: the longer the wavelength and/or the smaller the bending diameter, the more loss the fiber experiences. It is recognized that 1625 nm is a wavelength that is very sensitive to macrobending (see Fig. 6).

The macrobending loss is measured by the power monitoring method (OTDR, especially for field assessment, see Fig. 6) or the cut-back method.

Microbending Sensitivity

Microbending is a fiber property by which the core-cladding concentricity randomly changed along the fiber length. It causes the core to wobble inside the cladding and along the fiber length.

Four methods are available for characterizing microbending sensitivity in optical fibers:

- expandable drum for singlemode fibers and optical fiber ribbons over a wide range of applied linear pressure or loads;
- fixed diameter drum for step-index multimode, singlemode, and ribbon fibers for a fixed linear pressure;
- wire mesh and applied loads for step-index multimode and singlemode fibers over a wide range of applied linear pressure or loads; and
- 'basketweave' wrap on a fixed diameter drum for singlemode fibers.

The results from the four methods can only be compared qualitatively. The test is nonroutine for general evaluation of optical fiber.

Cut-off Wavelength

The cut-off wavelength is the shortest wavelength at which a single mode can propagate in a singlemode fiber. This parameter can be computed from the fiber refractive index profile (RIP). At wavelengths below the cut-off wavelength, several modes propagate and the fiber is no longer singlemode, but multimode.

In optical fibers, the change from multimode to singlemode behavior does not occur at a specific wavelength, but rather over a smooth transition as a function of wavelengths. Consequently, from a fiber-optic network standpoint, the actual threshold wavelength for singlemode performance is more critical. Thus an effective cut-off wavelength is described below.

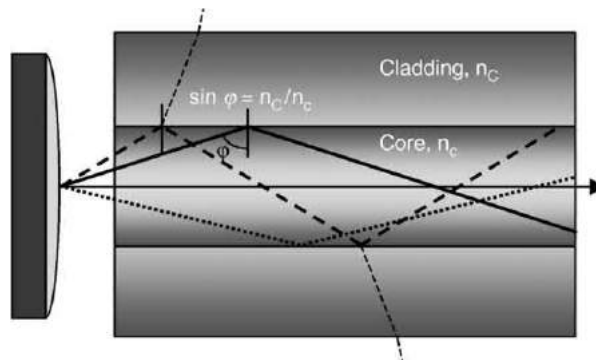


Fig. 5 Total internal reflection and macrobending effect on the light rays.

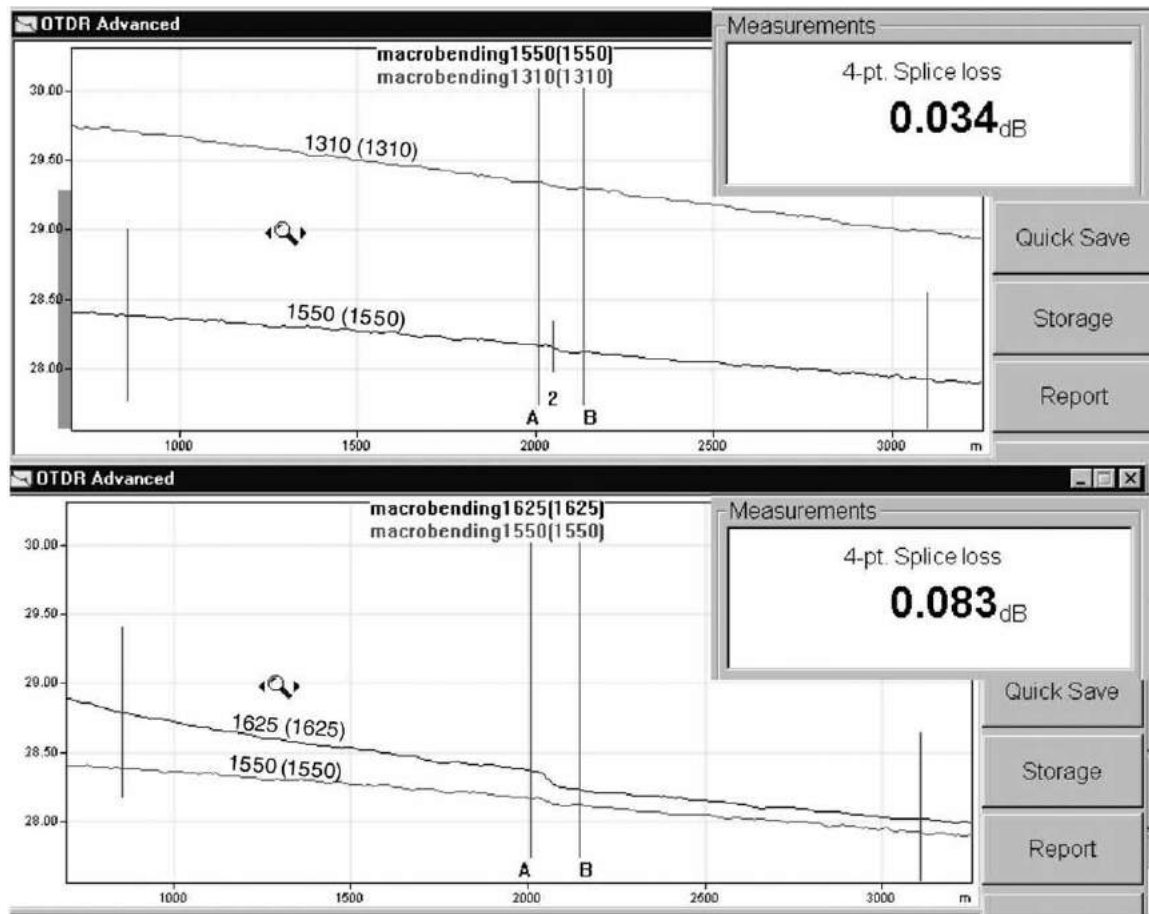


Fig. 6 Macrobending sensitivity as a function of wavelength.

The cut-off wavelength is defined as the wavelength greater than the ratio between the total power, in the higher-order modes and the fundamental mode, which has decreased to less than 0.1 dB. According to this definition, the second-order mode LP_{11} undergoes 19.3 dB more attenuation than the fundamental LP_{01} mode when the modes are equally excited.

Because the cut-off wavelength depends on the fiber length, bends, and strain, it is defined on the basis of the following three cases:

- fiber cut-off wavelength;
- cable cut-off wavelength; and
- jumper cable cut-off wavelength.

Fiber cut-off wavelength: Fiber cut-off wavelength λ_{fc} is defined for uncabled primary-coated fiber and is measured over 2 m with one loop of 140 mm radius loosely constrained, with the rest of the fiber kept essentially straight. The presence of a primary coating on the fiber usually will not affect λ_{fc} . However, the presence of a secondary coating may result in λ_{fc} being significantly shorter than that of the primary coated fiber.

Cable cut-off wavelength: Cable cut-off wavelength is measured prior to installation on a substantially straight 22 m cable length prepared by exposing 1 m of primary-coated fiber at both ends, the exposed ends each incorporating a 40 mm radius loop. Alternatively, this parameter may be measured on 22 m primary-coated uncabled fiber in the same configuration as for the λ_{fc} measurement.

Jumper cable cut-off wavelength: Jumper cable cut-off wavelength is measured over 2 m with one loop of 76 mm radius, or equivalent (e.g., split mandrel), with the rest of the jumper cable kept essentially straight.

Multimode Fiber Bandwidth for Multimode Fibers

The -3 dB bandwidth of a multimode optical fiber (or modal bandwidth) is defined as the lowest frequency where the magnitude of the baseband frequency response in optical power has decreased by 3 dB relative to the power at zero frequency. Modal bandwidth is also called intermodal dispersion as it takes into account the dispersion between the modes of propagation of the transmitted signal into the multimode fiber.

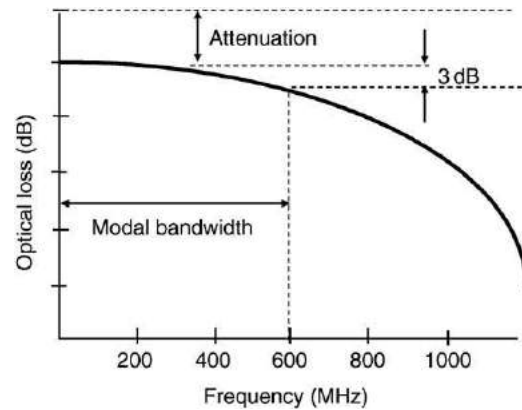


Fig. 7 Determination of modal bandwidth.

Various methods of reporting the results are available, but the results are typically expressed in terms of the -3 dB (optical power) frequency. [Fig. 7](#) illustrates modal bandwidth.

The bandwidth or pulse broadening may be normalized to a unit length, such as GHz·km, or ns/km.

Two methods are available for determining transmission capacity of multimode fibers:

- the frequency domain measurement method in which the baseband frequency response is directly measured in the frequency domain by determining the fiber response to a sinusoidally modulated light source; or
- the optical time domain measurement method (pulse distortion) in which the baseband response is measured by observing the broadening of a narrow pulse of light.

Differential Mode Delay for Multimode Fibers

Differential mode delay (DMD) characterizes the modal structure of a graded-index glass-core multimode fiber. DMD is useful for assessing the bandwidth performance of a fiber when used with short-pulse, narrow spectral-width laser sources.

The output from a singlemode probe fiber excites the multimode FUT at the test wavelength. The probe spot is scanned across the FUT endface, and the optical pulse delay is determined at specified radial offset positions between an inner and an outer limit. DMD is the difference in optical pulse delay time between the FUT fastest and slowest modes excited for all radial offset positions between and including the inner and the outer limits.

The related critical issues influencing DMD are the temporal width of the optical pulse, jitter in the timing, the finite bandwidth of the optical detector, and the mode broadening due to the source spectral width and the FUT chromatic dispersion.

The test method is commonly used in production and research facilities, but is not easily accomplished in the field. DMD may be a good predictor of the source launching conditions. DMD may be normalized to a unit length, such as ps/m.

Chromatic Dispersion

Chromatic dispersion in a singlemode fiber is a combination of material dispersion and waveguide dispersion (see [Fig. 8](#)), and it contributes to pulse broadening and distortion in a digital signal.

Material dispersion is produced by the dopants used in glass and is important in all fiber types. Waveguide dispersion is produced by the wavelength dependence of the index of refraction and is critical in singlemode fibers only.

From the point of view of the transmitter, this is due to two causes:

- The presence of wavelengths in the source optical spectrum. Each wavelength has a different phase delay and group delay (different group velocities) along the fiber, because they travel under different index of refraction (or phase) as the index varies as a function of wavelengths, as shown in [Fig. 9](#).
- The other cause is the modulation of the source, which itself has two effects:
 - As bit-rates increase, the spectral width of the modulated signal increases and can become comparable to or exceed the spectral width of the source.
 - Chirp occurs when the source wavelength spectrum varies during the pulse. By convention, positive chirp at the transmitter occurs when the spectrum during the rise/fall of the pulse shifts towards shorter/longer wavelengths respectively. For a positive fiber dispersion coefficient, longer wavelengths are delayed relative to shorter wavelengths. Hence if the sign of the product of chirp and dispersion is positive, the two processes combine to produce pulse broadening. If the product is negative, pulse compression can occur over an initial fiber length until the pulse reaches a minimum width and then broadens again with increasing dispersion.

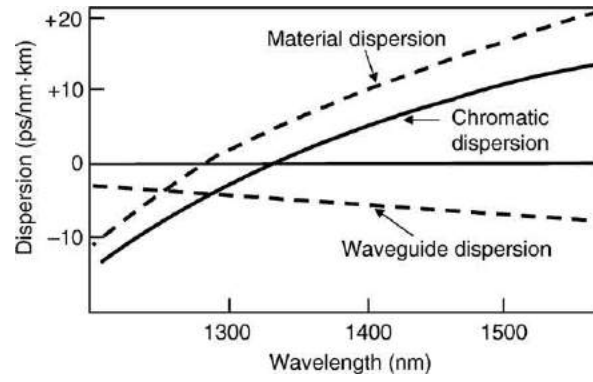


Fig. 8 Contribution of the material and waveguide dispersions to the chromatic dispersion.

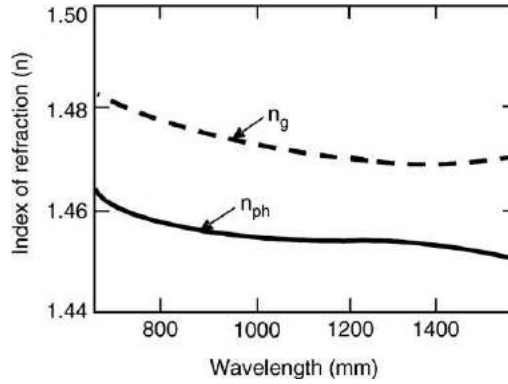


Fig. 9 Difference between the phase and the group index of refraction.

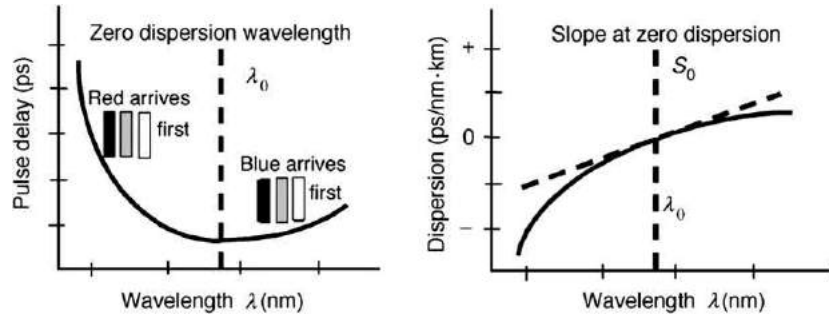


Fig. 10 Relation between the pulse (group) delay and the (chromatic) dispersion.

The electric field propagating into the FUT may be simply described as follows:

$$E(t, z) = E_0 \sin(\omega t - \beta z) \quad (8)$$

$\omega = 2\pi c/\lambda$ [rad/s] is the angular frequency; $\beta = kn = (\omega/c)n$ is the effective index; as β has units of m^{-1} it is often referred to as wavenumber or even sometimes propagation constant; k is the propagation constant.

The group delay τ_g is then given by:

$$d\beta/d\omega = \beta_1 = \tau_g \quad (9)$$

An example of the group delay spectral distribution is shown in **Fig. 10**. Assuming n is not complex; β_1 is the first-order derivative of β .

The group velocity v_g is given by

$$v_g = (d\beta/d\omega)^{-1} = -(\lambda^2/2\pi c)(d\beta/d\lambda)^{-1} \quad (10)$$

The FUT input-output delay is given by

$$\tau_d = L/v_g \quad (11)$$

L is the FUT length.

The group index of refraction n_g is given by

$$n_g = c/v_g = n - \lambda(dn/d\lambda) \quad (12)$$

The dispersion parameter or dispersion coefficient D (ps/nm · km) is given by

$$\begin{aligned} D &= -(\omega/\lambda)(d\tau_g/d\omega) = -(2\pi c/\lambda^2)(d^2\beta/d\omega^2) \\ &= -(\lambda/c)(d^2n/d\lambda^2) \end{aligned} \quad (13)$$

$$d^2\beta/d\omega^2 = \beta_2 \quad (14)$$

An example of the spectral distribution of D obtained from the group delay is shown in Fig. 10.

β_2 (ps²/km) is the group velocity dispersion parameter, so D may be related to β_2 as follows:

$$D = -(\omega/\lambda)\beta_2 \quad (15)$$

When β_2 is positive then D is negative and vice-versa. The region where β_2 is positive is called normal dispersion while the negative- β_2 region is called anomalous dispersion.

At λ_0 , β_1 is minimum, and $\beta_2=0$, then $D=0$.

The dispersion slope S (ps/nm² · km), also called the differential dispersion parameter or second-order dispersion, is given by

$$S = dD/d\lambda = (\omega/\lambda)\beta_3 + (2\omega/\lambda^2)\beta_2 \quad (16)$$

$$\beta_3 = d\beta_2/d\omega = d^3\beta/d\omega^3$$

At λ_0 , β_1 is minimum, $\beta_2=0$, then $D=0$; but S is not zero and depends on β_3 . An example of the spectral distribution of S and S_0 is illustrated in Fig. 10.

Overall, the general expression of β is given by

$$\begin{aligned} \beta(\omega) &= \beta_0 + (\omega - \omega_0)\beta_1 + (1/2)(\omega - \omega_0)^2\beta_2 \\ &\quad + (1/12)(\omega - \omega_0)^3\beta_3 + \dots \end{aligned} \quad (17)$$

Fig. 11 illustrates the difference between dispersion unshifted fiber (ITU-T Rec. G.652), dispersion shifted fiber (ITU-T Rec. G.653) and nonzero dispersion shifted fiber (ITU-T Rec. G.655).

Test methods for the determination of chromatic dispersion

All methods measure the group delay at a specific wavelength over a range and use agreed fitting functions to evaluate λ_0 and S_0 .

In the phase shift method, the group delay is measured in the frequency domain, by detecting, recording, and processing the phase shift of a sinusoidal modulating signal between a reference, a secondary fiber path, and the channel signal.

Setup variances exist and some do not require the secondary reference path. For instance, by using a reference optical filter at the FUT output it is possible to completely decouple the source from the phasemeter. With such an approach, chromatic dispersion may now be measured in the field over very long links using optical amplifiers (see Fig. 12).

In the differential phase shift method, two detection systems are used together with two wavelength sources at the same time. In this case the chromatic dispersion may be determined directly from the two group delays. This technique usually offers faster and more reliable results but costs much more than the phase-shift technique which is usually preferred.

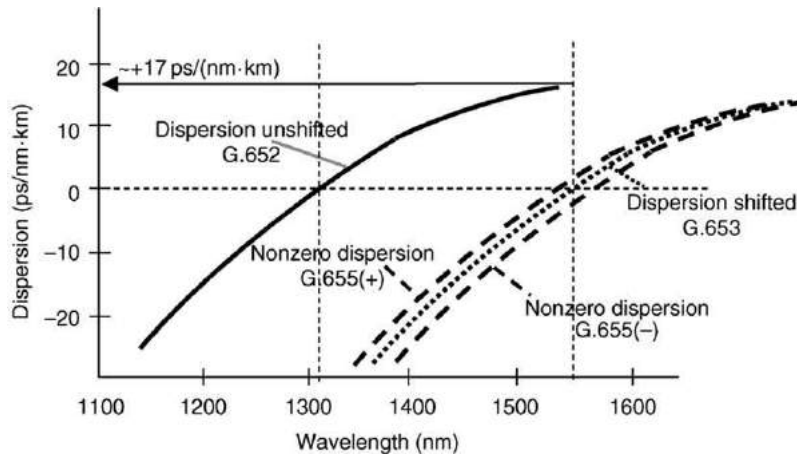


Fig. 11 Chromatic dispersion for various types of fiber.

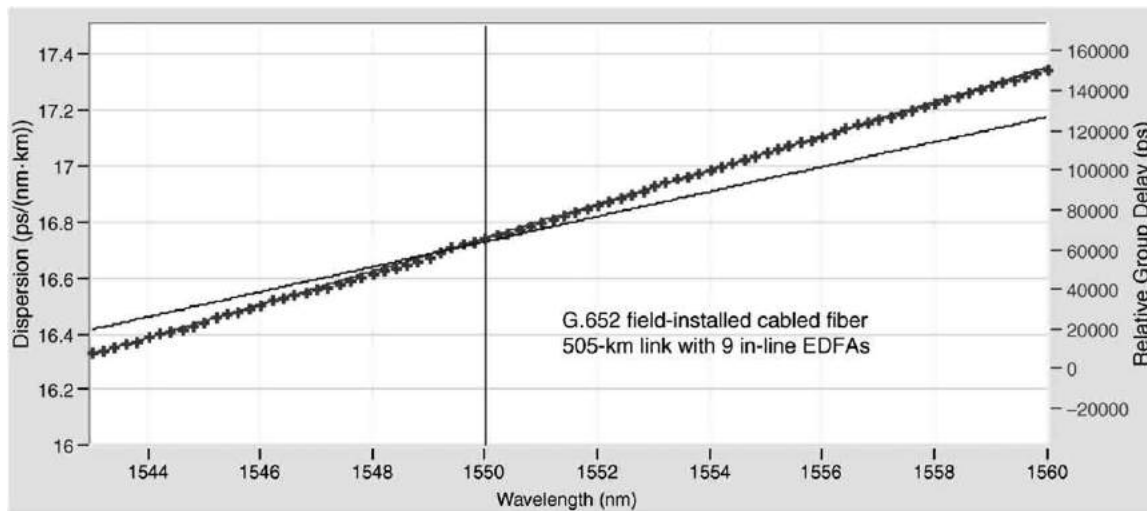


Fig. 12 Test results for field-related chromatic dispersion using the phase-shift technique.

In the interferometric method, the group delay between the FUT and a reference path is measured by a Mach-Zehnder interferometer. The reference delay line may be an air path or a singlemode fiber standard reference material (SRM). The method can be used to determine the following characteristics:

- longitudinal chromatic dispersion homogeneity; and
- effect of overall or local influences, such as temperature changes and macrobending losses.

In the pulse delay method, the group delay is measured in the time domain, by detecting, recording, and processing the delay experienced by pulses at various wavelengths.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) causes an optical pulse to spread in the time domain and may impair the performance of a telecommunications system. The effect can be related to differential phase and group velocities and corresponding arrival time of different polarization components of the pulse signal. For a sufficiently narrowband source, the effect can be related to a differential group delay (DGD), $\Delta\tau$, between a pair of orthogonally polarized principal states of polarization (PSP) at a given wavelength or optical frequency (see Fig. 13(a)). In an ideal circular symmetric fiber, the two PSPs propagate with the same velocity. However:

- a real fiber is not perfectly circular;
- the core is not perfectly concentric with the cladding;
- the core may be subjected to microbending;
- the core may present localized clusters of dopants; and
- the environmental conditions may stress the deployed cable and affect the fiber.

Each time the fiber undergoes local stresses and consequently birefringence. These asymmetry characteristics vary randomly along the fiber and in time, lead to a statistical behavior of PMD.

For a deployed cabled fiber at a given time and optical frequency, there always exist two PSPs such that the pulse spreading due to PMD vanishes, if only one PSP is excited. On the contrary, the maximum pulse spread due to PMD occurs when both PSPs are equally excited, and is related to the difference in their group delays, the DGD associated with the two PSPs. For broadband transmission, the DGD statistically varies as a function of wavelengths or frequencies and result in an output pulse that is spread out in the time domain (see Figs. 13(a)–(c)). In this case, the spreading can be related to the RMS (root mean square) of DGD values $\langle\Delta\tau^2\rangle^{1/2}$. However, if a known distribution such as the Maxwell distribution may be fit to the DGD distribution probability, then a mean (or average) value of the DGD $\langle\Delta\tau\rangle$ may be correlated to the RMS value and used as a system performance predictor in particular with a maximum value of the DGD distribution associated to a low probability of occurrence. This maximum DGD may then be used to define the quality of service that would tolerate values lower than this maximum DGD.

Test methods for polarization mode dispersion

Three methods are generically used for measuring PMD. Other methods or analyses may exist but they are generally not standardized or are limited in their applications.

- Stokes parameter evaluation (SPE)

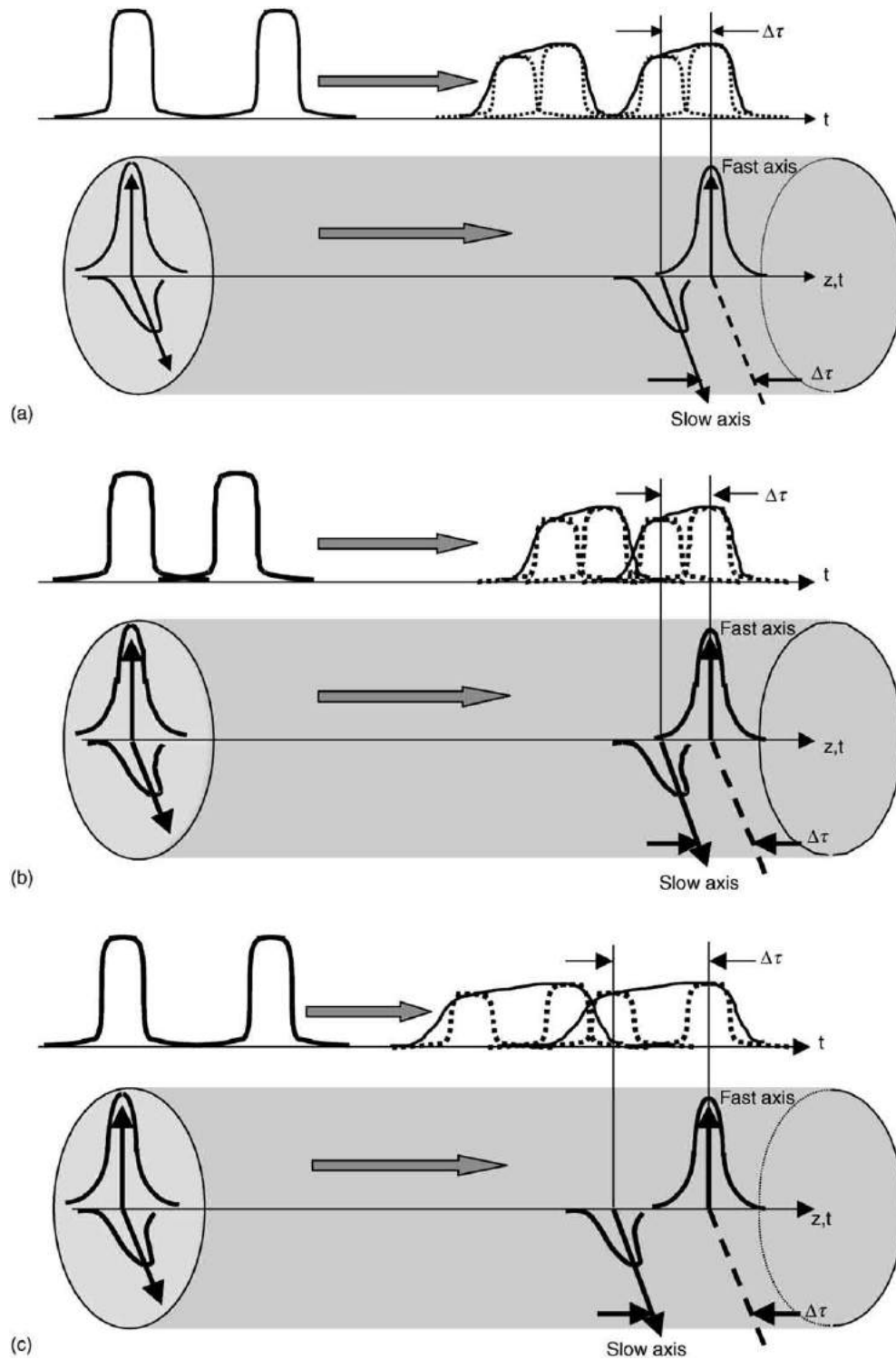


Fig. 13 PMD effect on the pulse broadening and its possible pulse impairment. (a) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is too low to create an impairment. (b) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is high enough to create an impairment. (c) The DGD $\Delta\tau$ is large enough even at low bit rate to make the pulse spreading and creating impairment.

- Jones matrix eigenanalysis (JME)
- Poincaré sphere analysis (PSA)
- Fixed analyzer (FA)

- Extrema counting (EC)
- Fourier transform (FT)
- Interferometry (INTY)
 - Traditional analysis (TINTY)
 - General analysis (GINTY).

All methods use a linearly polarized source at the FUT input and are suitable for laboratory measurements of factory lengths of fiber and cable. However, the interferometric method is the only method appropriate for measurements of cabled fiber that may be moving or vibrating such as is found in the field.

Stokes parameter evaluation

SPE determines PMD by measuring a response to a change of narrowband light (from a tuneable light source with broadband detector – JME, or a broadband source with a filtered detector such as an interferometer – PSA) across a wavelength range. The Stokes vector of the output light is measured for each wavelength. The change of these Stokes vectors with angular optical frequency (wavelength), ω and with the change in input SOP (state of polarization), yields the DGD as a function of wavelength.

For both JME and PSA analyses, three distinct and known linear SOPs (orthogonal on the Poincaré sphere) must be launched for each wavelength. **Fig. 14** illustrates the test setup and examples of test results.

The JME and PSA method are mathematically equivalent.

Fixed analyzer

FA determines PMD by measuring a response to a change of narrowband light across a wavelength range. For each SOP, the change in output power that is filtered through a fixed polarization analyzer, relative to the power detected without the analyzer, is measured as a function of wavelength. **Fig. 15** illustrates a test setup and examples of test results.

The resulting measured function can be analyzed in one of two ways:

- by counting the number of peaks and valleys (EC) of the curve and application of a formula. This analysis is considered as a frequency domain approach; and
- by taking the Fourier transform (FT) of the measured function. This FT is equivalent to the pulse spreading obtained by TINTY.

Interferometry

INTY uses a broadband light source and an interferometer. The fringe pattern containing the source auto-correlation, together with the PMD related cross-correlation of the emerging electromagnetic field, is determined by the interference pattern of the output light, i.e., the interferogram. The PMD determination for the wavelength range associated with the source spectrum is based on the envelope of the fringe pattern interferogram. Two analyses are available to obtain the PMD:

- TINTY uses a set of specific operating conditions for its successful applications and a basic setup; and
- GINTY uses no limiting operating conditions, but in addition to the same basic setup, also using a modified setup compared to TINTY.

Fig. 16 illustrates the test setup for both approaches and examples of test results.

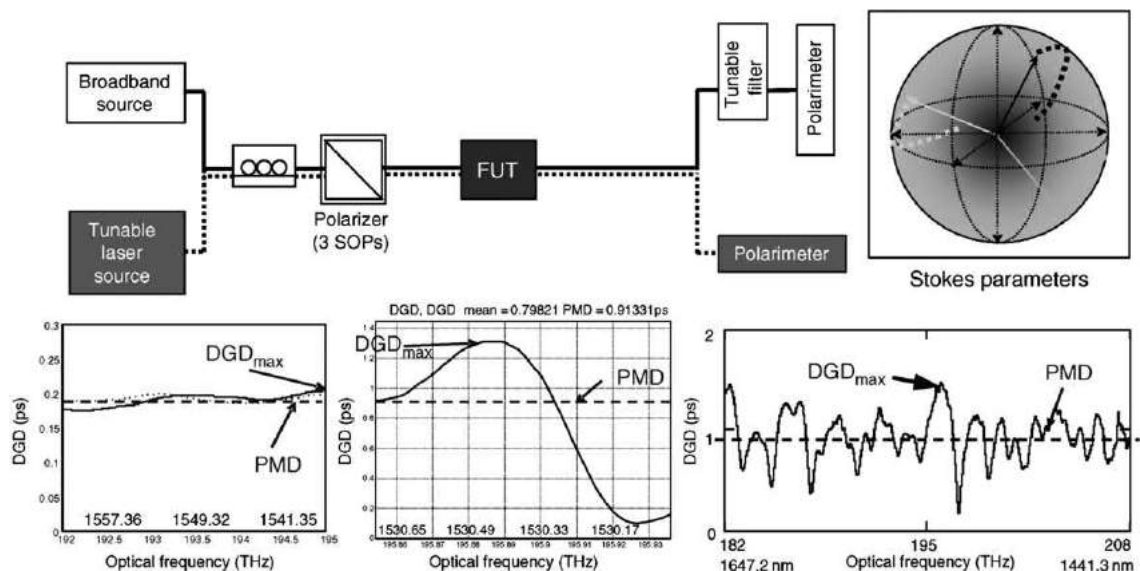


Fig. 14 PMD by Stokes parameter evaluation method.

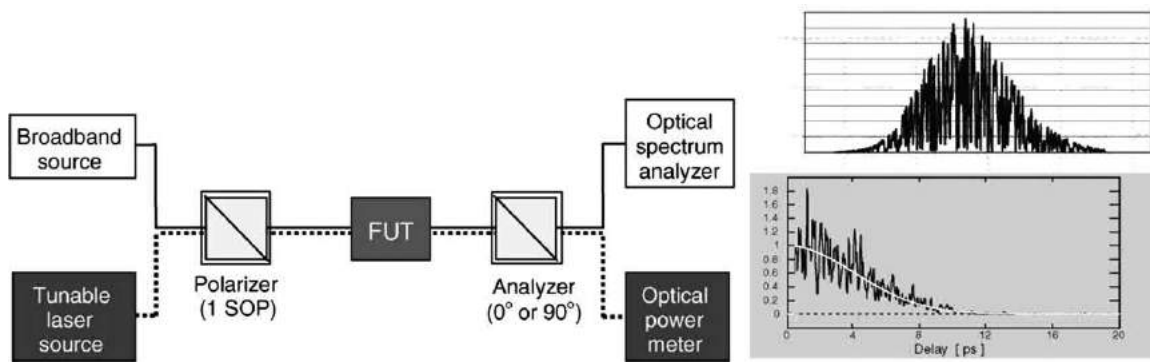
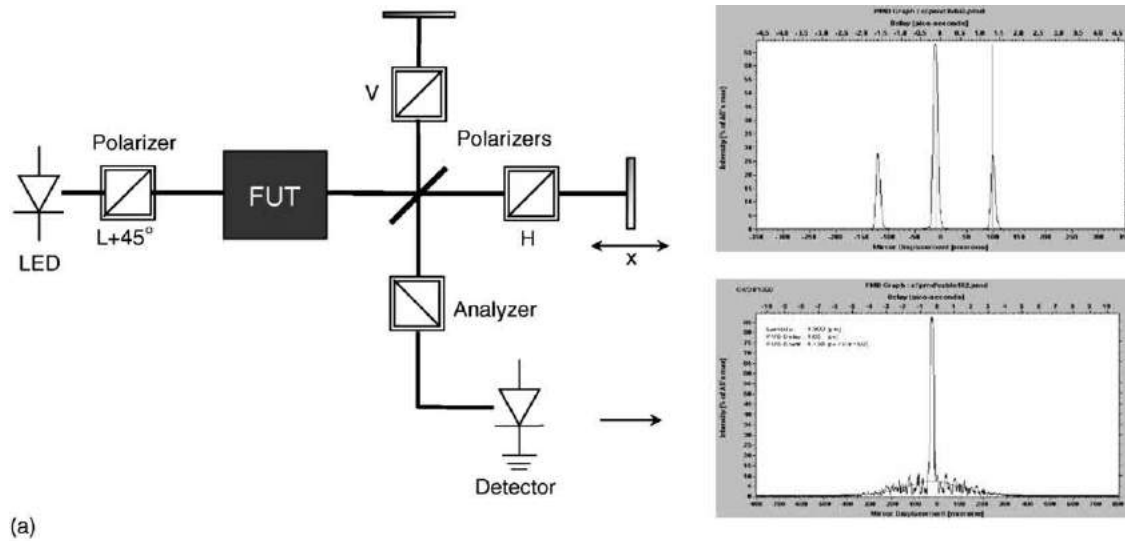
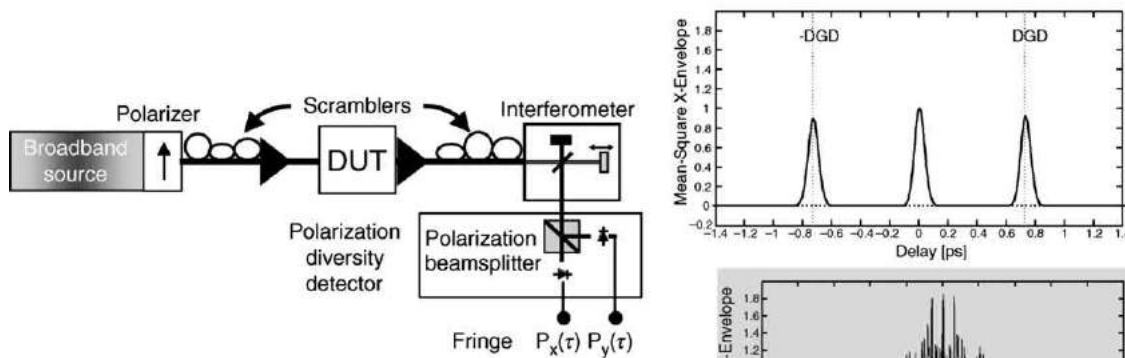


Fig. 15 PMD by fixed analyzer method.



(a)



NOTE: The optional input/output-SOP scramblers are used to improve the measurement uncertainty.

(b)

Fig. 16 PMD by the interferometric method. (a) TINTY, (b) GINTY.

Polarization Crosstalk

Polarization crosstalk is a characteristic of energy mixing/transfer/coupling between the two PSPs in a PMF (polarization maintaining fiber) when their isolation is imperfect. It is the measure of the strength of mode coupling or output power ratio between the PSPs within a PMF.

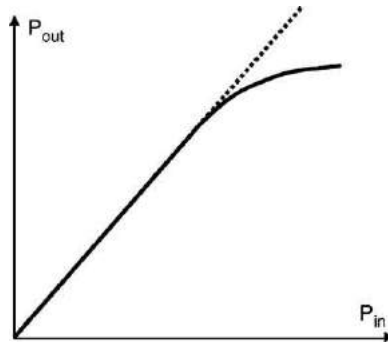


Fig. 17 Power output-to-power input relationship for production of nonlinear effects.

A PMF is an optical fiber capable of transmitting, under external perturbations, such as bending or lateral pressures, both HE_{11}^x and HE_{11}^y polarization modes whose electric field vector directions are orthogonally to each other and which have different propagation constants β_x and β_y .

Two methods are available for measuring the polarization crosstalk of PMF:

- *power ratio method*, which uses the maximum and minimum values of output power at a specified wavelength, and is applicable to fibers and connectors jointed to a PMF, and to two or more PMFs joined in series; and
- *in-line method*, which uses an analysis of the Poincaré sphere, and is applicable to single or cascaded sections of PMF, and to PMF interconnected with optical devices.

Nonlinear Effects

When the power of the transmission signal is increased to achieve longer span lengths at high bit rates the interaction between the signal and the fiber medium creates nonlinearity. The output signal from the fiber does not increase linearly anymore (as shown in Fig. 17) and part of the signal is lost and will appear in another part of the spectrum.

This is a key issue both in high capacity systems and in long unregenerated links. These nonlinear effects can be generally categorized as follows:

- Scattering phenomena:
 - Stimulated Brillouin scattering (SBS); and
 - Stimulated Raman scattering (SRS).
- Kerr-effect related phenomena:
 - Self-phase modulation (SPM);
 - Cross-phase modulation (XPM);
 - Modulation instability;
 - Soliton formation; and
 - Four wave mixing (FWM).

The Kerr effect is related to the intensity dependence of the index of refraction.

The nonlinear effects are influenced by a number of parameters such as:

- Fiber total dispersion;
- Fiber effective area;
- Channel spacing in WDM systems;
- Unregenerated link distance;
- Degree of longitudinal uniformity of fiber characteristics; and
- Signal intensity and source linewidth.

Nonlinear coefficient (n_2/A_{eff})

Starting from a particularly intense field, the fiber index of refraction becomes dependent on optical intensity inside the fiber. The new index can be expressed as follows:

$$n = n_0 + n_2 I \quad (18)$$

n is the nonlinearity dependent index; n_0 is the linear part of the index; n_2 is the nonlinear index, also called the Kerr nonlinear index (2.2 to $3.4 \times 10^{-16} \text{ cm}^2/\text{W}$); and I is the optical intensity inside the fiber.

The field propagation at a distance L into the fiber is described by the following equation:

$$E_{out}(L) = E_{in}(0) \exp[-a/2 + i\beta + \gamma P(L, t)/2]L \quad (19)$$

$a/2$ is the attenuation; $i\beta$ is the phase of the wave; and $\gamma P(L, t)/2$ is the nonlinearity term;

$$\gamma = 2\pi n_2 / \lambda A_{\text{eff}} \quad (20)$$

γ is the nonlinearity coefficient and may be a complex number; A_{eff} is the fiber core effective area; $P(L, t)$ is the total power; and λ is the signal wavelength and t the time variable.

The nonlinear coefficient is defined as n_2/A_{eff} . This coefficient plays a critical role in the fiber and is closely related to system performance degradation due to nonlinearities when very high power is used.

Methods for measuring the nonlinear coefficient

Two methods are available for measuring the nonlinear coefficient:

- Continuous-wave dual-frequency (CWDF); and
- Pulsed single-frequency (PSF).

In the CWDF method, light from two wavelengths is injected into the fiber. At higher power, the light from the two wavelengths beat due to the nonlinearity and produce an output spectrum that is spread. The relationship of the power level to a particular spreading is used to calculate the nonlinear coefficient.

In the PSF method, the pulsed light from a single wavelength is injected into the fiber. Very short pulses (< 1 ns) and their input peak power must be measured and related to the nonlinear spreading of the output spectrum.

Stimulated Brillouin scattering

In an intensity modulated system using a source with a narrow linewidth, significant optical power is transferred from the forward-propagating signal to a backward-propagating signal when the SBS power threshold is exceeded. At that point periodic regions of index of refraction produce a grating traveling at speed of sound away from the source. The grating reflects backward part of the incident light. The reflected sound waves (acoustic phonons) scatter light back to the source. Phase matching (or momentum conservation) dictates that the scattered light preferentially travels in the backward direction. The scattered light will be Doppler-shifted (downshifted or Brillouin-shifted) by approximately 11 GHz (at 1550 nm, for G.652 fiber). The scattered light has a very narrow spectrum (very coherent) and very close to the carrier signal and may be very detrimental.

Stimulated Raman scattering

SRS is an interaction between light and the fiber molecular vibrations as adjacent atoms vibrate in opposite directions (an 'optical phonon'). Some of the energy of the main carrier (optical pump wave) is transferred to the molecules, thereby further increasing the amplitude of their vibrations. If the vibrations become large enough, a threshold is reached at which the local index of refraction changes. These local changes then scatter light in all directions similar to Rayleigh scattering. However, unlike Rayleigh scattering, the wavelength of the Raman scattered light is shifted to longer wavelengths by an amount that corresponds to the molecular vibration frequencies and the Raman signal spreads over a large spectrum.

Self-phase modulation

SPM is the effect that a powerful pulse has on its own phase, considering that in eqn (18), $I(t)$ varies in time:

- $I(t) \rightarrow n(t) = n_0 + n_2 I(t) \rightarrow$ modulates the phase $\beta(t)$ of the pulse; and
- $dl/dt \rightarrow dn/dt \rightarrow d\beta/dt(\text{chirp}) \rightarrow$ broadening in the frequency domain $\rightarrow$ broadening in the time domain.

$I(t)$ peaks at the center of the pulse (peak power) and consequently increases the index of refraction. A higher index causes the wavelengths in the center of the pulse to accumulate phase more quickly than at the wings of the pulse:

- this causes wavelength stretching (shift to longer wavelengths) at pulse leading edge (risetime); and
- this causes wavelength compression (shift to shorter wavelengths) at pulse trailing edge (falltime).

The pulse will then broaden with negative (normal) dispersion and shorten with positive (anomalous) dispersion. SPM may then be used for dispersion compensation considering that self-phase modulation imposes $C > 0$ (positive chirp). It can cancel the dispersion if properly manage as a function of the sign of the dispersion. SPM is one of the most critical nonlinear effects for the propagation of soliton or very short pulses over very long distance.

Cross-phase modulation

XPM is the effect that a powerful pulse has on the phase of an adjacent pulse from another WDM system channel traveling in phase or at slightly the same group velocity. It concerns spectral interference between two WDM channels:

- the increasing $I(t)$ at the leading edge of the interfering pulse shifts the other pulse to longer wavelengths; and
- decreasing $I(t)$ at the trailing edge of the interfering pulse shifts the other pulse to shorter wavelengths.

This produces spectral broadening, which dispersion converts to temporal broadening depending on the sign of the dispersion. XPM effect is similar to SPM except it depends on the channel count.

Table 1 Fiber dimensional characteristics

<i>Attribute</i>	<i>Measured parameter</i>
Fiber geometry	Core/cladding diameter Core/cladding noncircularity Core-cladding concentricity error
Numerical aperture	
Mode field diameter	
Coating geometry	
Length	

Four-wave (four-photon) mixing

FWM is the by-product production effect from two or more WDM channels. For two channels $I(t)$ modulates the phase of each signal (ω_1 and ω_2). An intensity modulation appears at the beat frequency $\omega_1 - \omega_2$.

Two sideband frequencies are created in a similar way as harmonics generation. New wavelengths are created in a number equal to $N^2(N-1)/2$, where N =number of original wavelengths.

Fiber Dimension Characteristics and Corresponding Tests Methods

Table 1 provides a list of the various fiber dimensional characteristics and their corresponding test methods.

Fiber Geometry Characteristics

The fiber geometry is related to the core and cladding characteristics.

Core

The core center is the center of a circle which best fits the points at a constant level in the near-field intensity profile emitted from the central region of the fiber, using wavelengths above and/or below the cut-off wavelength.

The RIP can be measured by refracted near field (RNF) or transverse interferometry techniques and transmitted near field (TNF).

The core concentricity error is the distance between the core center and the cladding center. This definition applies very well for multimode fibers. The distance between the center of the near field profile and the center of the cladding is also used for singlemode fibers.

The mode field diameter (MFD) represents a measure of the transverse electromagnetic field intensity of the mode in a fiber cross-section and it is defined from the far-field intensity distribution.

The MF is the singlemode field distribution of the LP_{01} mode, giving rise to a spatial intensity distribution in the fiber.

The MF concentricity error is the distance between the MF center and the cladding center.

The core noncircularity is a measure of the core ellipticity. This parameter is one of the causes for creating birefringence in the fiber and consequently PMD.

Cladding

The cladding is the outermost region of constant refractive index in the fiber cross-section.

The cladding center is the center of a circle best fitting the outer limit (boundary) of the cladding.

The cladding diameter is the diameter of the circle defining the cladding center.

The cladding noncircularity is a measure of the difference between the diameters of the two circles defined by the cladding tolerance field divided by the nominal cladding diameter.

Coating

The primary coating is one or more layers of protective material applied to the cladding during or after the drawing process to protect the cladding surface (e.g., a 250 μm protective coating). The secondary coating is one or more layers of protective material applied over the primary coating in order to give additional protection or to provide a particular structure.

Measurement of the fiber geometrical attributes

The fiber geometry is measured by the following methods:

- TNF;
- RNF;
- Side-view technique/transverse interference;
- TNF image technique; and
- Mechanical diameter.

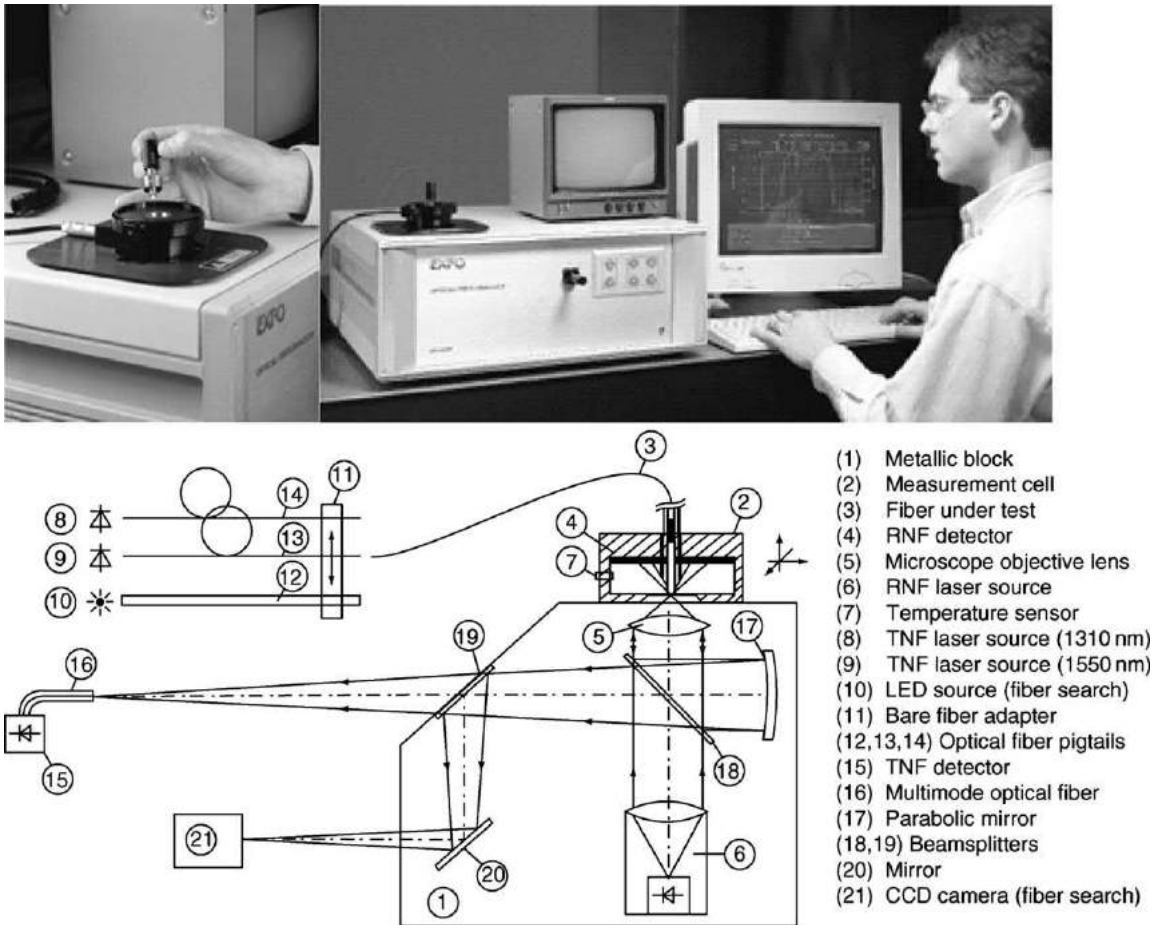


Fig. 18 RNF/TNF combined instrumentation.

Test instrumentation may incorporate two or more methods such as the one shown in [Fig. 18](#).

Transmitted near-field technique

The cladding diameter, core concentricity error, and cladding noncircularity are determined from the near-field intensity distribution. [Fig. 19](#) provides a series of examples of test results from TNF measurements.

Refracted near-field technique

The RIP across the entire fiber (core and cladding) can be directly obtained from the RNF measurement, as shown in [Fig. 20](#).

The geometrical characteristics of the fiber can be obtained from the refractive index distribution using suitable algorithms:

- core/cladding diameter;
- core/cladding concentricity error;
- core/cladding noncircularity;
- maximum numerical aperture (NA); and
- index and relative index of refraction difference.

[Fig. 21](#) illustrates the core geometry.

Side-view technique/transverse interference

The side-view method is applied to singlemode fibers to determine the core concentricity error, cladding diameter and cladding noncircularity by measuring the intensity distribution of light that is refracted inside the fiber. The method is based on an interference microscope focused on the side view of an FUT illuminated perpendicular to the FUT axis. The fringe pattern is used to determine the RIP.

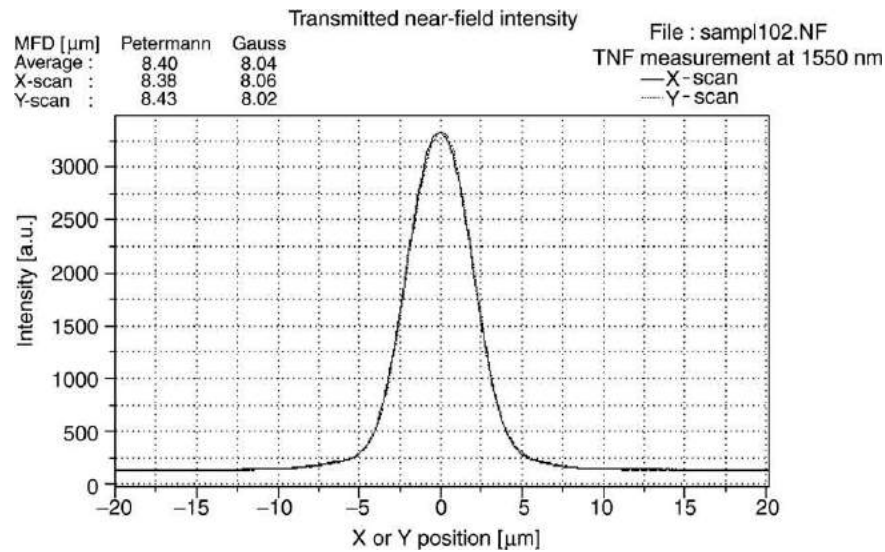


Fig. 19 MFD measurement by TNF.

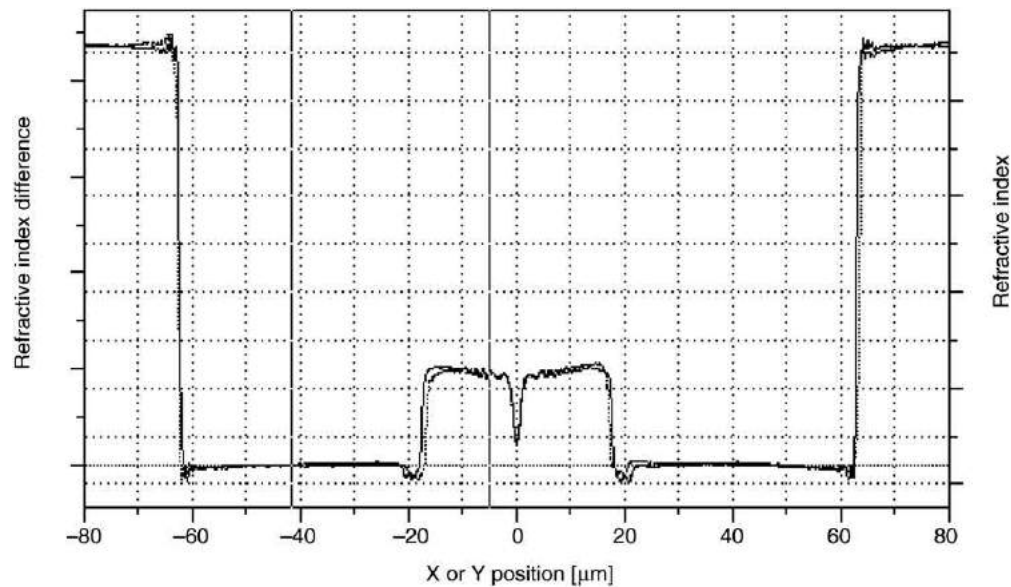


Fig. 20 RIP from RNF measurement.

TNF image technique

The TNF image technique, also called near-field light distribution, is used for the measurement of the geometrical characteristics of singlemode fibers. The measurement is based on analysis of magnified images at the FUT output. Two subsets of the method are available:

- grey-scale technique which performs an x-y near-field scan using a video system; and
- Single near-field scanning technique performing a one-dimensional scan.

Mechanical diameter

This is a precision mechanical diameter measurement technique used to accurately determine the cladding diameter of silica fibers. The technique uses an electronic micrometer such as based on a double-pass Michelson interferometer. The technique is used for providing calibrated fibers to the industry as SRM.

Numerical Aperture

The NA is an important attribute for multimode fibers in order to predict their launching efficiency, joint loss at splices and micro/macrobending characteristics.

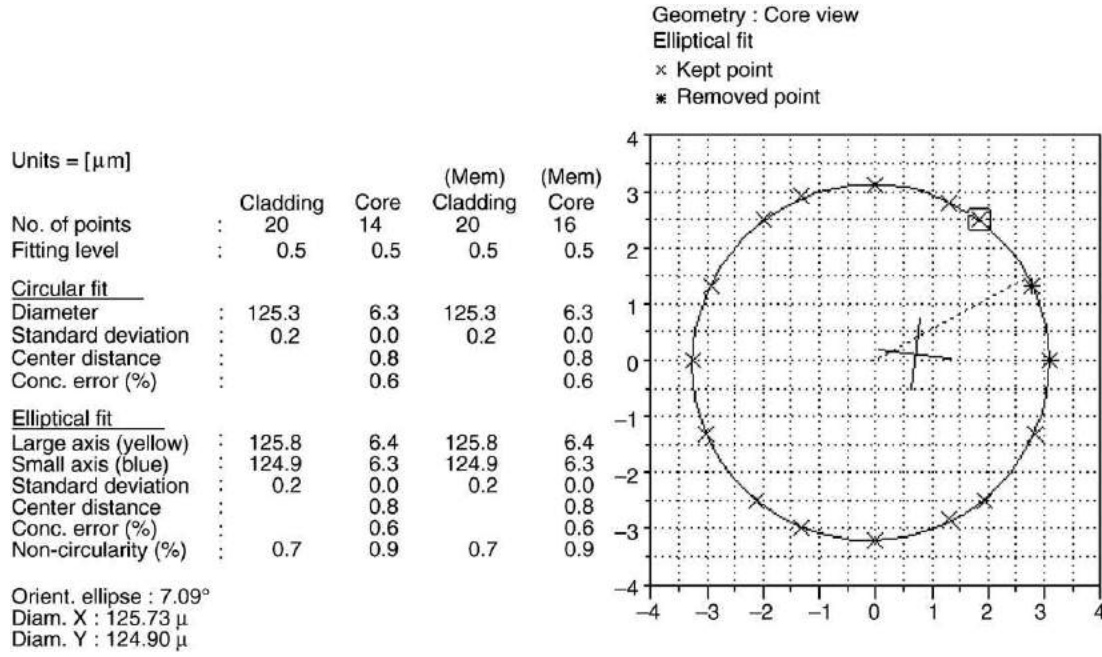


Fig. 21 Core geometry by RNF measurement.

A method is available for the measurement of the angular radiant intensity distribution (far-field) distribution or the RIP at the output of an FUT. NA can be determined from the analysis of the test results.

Mode Field Diameter

The definition of the MFD is given in the section describing the core center above. Four measurement methods are available:

- direct far-field scan determines MFD from the far-field intensity distribution;
- variable aperture in far field determines MFD from the complementary aperture transmission function $a(x)$, $x = d \tan \theta$ being the aperture radius and d the distance between the aperture and the FUT:

$$\text{MFD} = \frac{\lambda}{\pi d} \left[\int_0^\infty a(x) \frac{x}{(x^2 + d^2)^2} dx \right]^{-1/2} \quad (21)$$

Eq. (21) is valid for small- θ approximation.

- near-field scan determines MFD from the near-field intensity distribution I_{NF} , r being the radial coordinate:

$$\text{MFD} = 2 \left[2 \frac{\int_0^\infty r I_{\text{NF}}(r) dr}{\int_0^\infty r \left[\frac{dI^{1/2}(r)}{dr} \right]^2 dr} \right]^{1/2} \quad (22)$$

Eq. (22) is valid for small- θ approximation; and

- bidirectional backscattering uses an OTDR and bidirectional measurements to determine MFD by comparing the FUT results with a reference fiber.

Effective Area

A_{eff} is a critical nonlinearity parameter and is defined as follows:

$$A_{\text{eff}} = \frac{2\pi \left[\int_0^\infty I(r) r dr \right]^2}{\int_0^\infty I(r)^2 r dr} \quad (23)$$

$I(r)$ is the field intensity distribution of the fiber fundamental mode at radius r . The integration in the equation is carried out over the entire fiber cross-section. For a Gaussian approximation:

$$I(r) = \exp - 2 \left(\frac{2r}{\text{MED}} \right)^2 \quad (24)$$

which yields:

$$A_{\text{eff}} = \frac{\pi}{4} \text{MFD}^2 \quad (25)$$

Three methods are available for the measurement of A_{eff} :

- direct far-field;
- variable aperture in far-field; and
- near-field.

Mechanical Measurement and Test Methods

Table 2 describes the various mechanical characteristics and their corresponding test methods.

Proof Stressing

The proof stress level is the value of tensile stress or strain applied to a full fiber length over a short period of time. The method for fiber proof stressing is the longitudinal tension which describes procedures for applying tensile loads to a length of fiber. The fiber stress is calculated from the applied tension. The tensile load is applied over short periods of time but not too short in order for the fiber to experience proof stress.

Residual Stress

Residual stress is the stress built up by the thermal expansion difference between core and cladding during the fiber drawing process or splicing. Methods are available for measuring residual stress based on polarization effects. A light beam produced by a rotating polarizer propagates to the x -axis direction while the beam polarization is in the $y - z$ plane and the fiber longitudinal axis in the z -axis. The light experiences different phase shift in the y - and z -axis due to the FUT birefringence. The photoelastic effect gives the relationship between this phase change and residual stresses.

Stress Corrosion Susceptibility

The stress corrosion susceptibility is related to the dependence of crack growth on applied stress. It depends on the environmental conditions and static and dynamic values may be observed.

Environmental Characteristics

Table 3 lists the fiber characteristics related to the effect of the environment.

Hydrogen Aging for Low-Water-Peak Single-Mode Fiber

Hydrogen aging on low-water peak fibers, such as G.652.C, is based on a test performed at 1.0 atmosphere of hydrogen pressure at room temperature over a period of one month. Other proportional combinations are possible.

Table 2 Fiber mechanical characteristics

Attribute

Proof stress
Residual stress
Stress corrosion susceptibility
Tensile strength
Stripability
Fiber curl

Table 3 Fiber environmental characteristics

Attribute
Hydrogen aging
Nuclear gamma irradiation
Damp heat
Dry heat
Temperature cycling
Water immersion

Nuclear Gamma Irradiation

Nuclear radiation is considered on the basis of a steady state response of optical fibers and cables exposed to gamma radiation and to determine the level of related radiation-induced attenuation produced in singlemode or multimode cabled or uncabled fibers.

The fiber attenuation generally increases when exposed to gamma radiation. This is primarily due to the trapping of radiolytic electrons and holes at defect sites in the glass (i.e., the formation of ‘color centers’). Two regimes are considered:

- the low dose rate suitable for estimating the effect of environmental background radiation; and
- the high dose rate suitable for estimating the effect of adverse nuclear environments.

The effects of environmental background radiation are measured by the attenuation (cut-back method). The effects of adverse nuclear environments are tested by power monitoring before, during and after FUT exposure.

Further Reading

Agrawal, G.P., 1997. Fiber-Optic Communication Systems, 2nd edn. New York: John Wiley & Sons.
Agrawal, G.P., 2001. Nonlinear Fiber Optics, 3rd edn. San Diego, CA: Academic Press.
Girard, A. (Ed.), 2000. Guide to WDM Technology and Testing. Quebec: EXFO, p. 194.
Hecht, J., 1999. Understanding Fiber Optics, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
Masson, B., Girard, A. (Eds.), 2004. FTTx PON Guide, Testing Passive Optical Networks. Quebec: EXFO, p. 56.
Miller, J.L., Friedman, E. (Eds.), 2003. Optical Communications Rules of Thumb. New York: McGraw-Hill, p. 428.
Neumann, E.-G., 1988. Single-Mode Fibers, Fundamentals. Berlin: Springer-Verlag.

Optical Fiber Cables

G Galliano, Telecom Italia Lab, Torino, Italy

© 2005 Elsevier Ltd. All rights reserved.

Introduction

In order to be put to practical use in the telecommunication network, optical fibers must be protected from environmental and mechanical stresses which can change their transmission characteristics and reliability.

The fibers, during manufacturing, are individually protected with a thin plastic layer (extruded during the drawing) to preserve directly their characteristics from damage due to the environment and handling. This coating surrounding the fiber cladding (with an external diameter of 125 μm) is known as the primary coating, which is obtained with a double layer of UV resin for a maximum external diameter of about 250 μm .

Primary coated fiber, at the end of the manufacturing process, is subjected to a dynamic traction test (proof-test) to monitor its mechanical strength. This test, carried out on all fibers manufactured, consists of the application of a well-defined strain (from 0.5% to 1%) for a fixed time (normally 1 s) and thus permits the exclusion of, from successive cabling, those fibers with poor mechanical characteristics.

At the end of the manufacturing process the optical fibers are not directly usable. It is necessary to surround the single fiber (or groups of fibers) with a structure which provides protection from mechanical and chemical hazards and allows for stresses that may be applied during cable manufacture, installation, and service.

Suitable protection is provided using fiber conditioning within the cable (secondary coating or loose cabling), while entire protection is obtained with some cable elements, such as strength members, core assembling, filling, and outer protections. Moreover, the cable structure and its design is heavily influenced by the application (e.g., number of fibers), installation (laying and splicing), and environmental conditions of service (underground, aerial, or submarine).

In this article, the main characteristics of optical cables (with both multimode or single mode fibers) are described and analyzed, together with suitable solutions.

Secondary Fiber Coating

The fiber, before cabling, can be protected with a secondary coating (a tight jacketing) to preserve it during cable stranding. However, fiber with only a primary coating may be cabled with a suitable cable protection.

In the first case the fiber can be directly stranded to form a cable core; in the second case, a single group of fibers must be inserted into a loose structure (tubes or grooves) with a proper extra-length. These two solutions are not always distinguishable; in fact, tight fibers inserted into loose tubes or grooves are often used.

The implementation of the secondary coating and the choice of proper materials requires particular attention, in order to avoid microbending effects on the fiber, which can cause degradation of the transmission characteristics. Microbending is caused by the pressure of the fiber on a microrough surface, or by the fiber buckling, due to the contraction of the structure (e.g., secondary coating) containing it. Microbending causes an increase in fiber attenuation at long wavelengths (1300 and 1550 nm) for both single and multimode fibers, but is particularly emphasized in multimode fibers.

Tight Jacket

In a tight protection, the primary coating fiber is surrounded by a second plastic jacket in contact with the fiber. This jacket mainly protects the fiber from lateral stresses that can cause the microbending.

The fiber can be individually protected with a single or double layer of plastic material or in a ribbon structure. In Fig. 1 four different types of tight jacket are shown. In Fig. 1(a) and (b) the fibers are singularly protected with a single or double layer up to an external diameter of 0.9 mm. The second type is normally preferred because the external hard plastic gives a good fiber protection and the inner soft plastic avoids microbending effects on the fiber.

In a ribbon structure, from 4 to 16 primary coated fibers are laid close and parallel and covered with plastic. Two types of ribbon are normally classified 'encapsulated' (Fig. 1(c)) or 'edge bonded' (Fig. 1(d)). In the first case, the fibers are assembled with a common secondary coating (one or two layers), in the second case, an adhesive is used to stick together the fibers with only the primary coating.

Care should be taken in applying tight protection, in particular for a ribbon structure, to avoid an attenuation increase due to the action of the jacket on the fiber (in particular for the thermal contraction of the coating materials). The ribbon structure is designed to permit the simultaneous junction of multiple fibers; for this reason the ribbon geometry must be precision controlled.

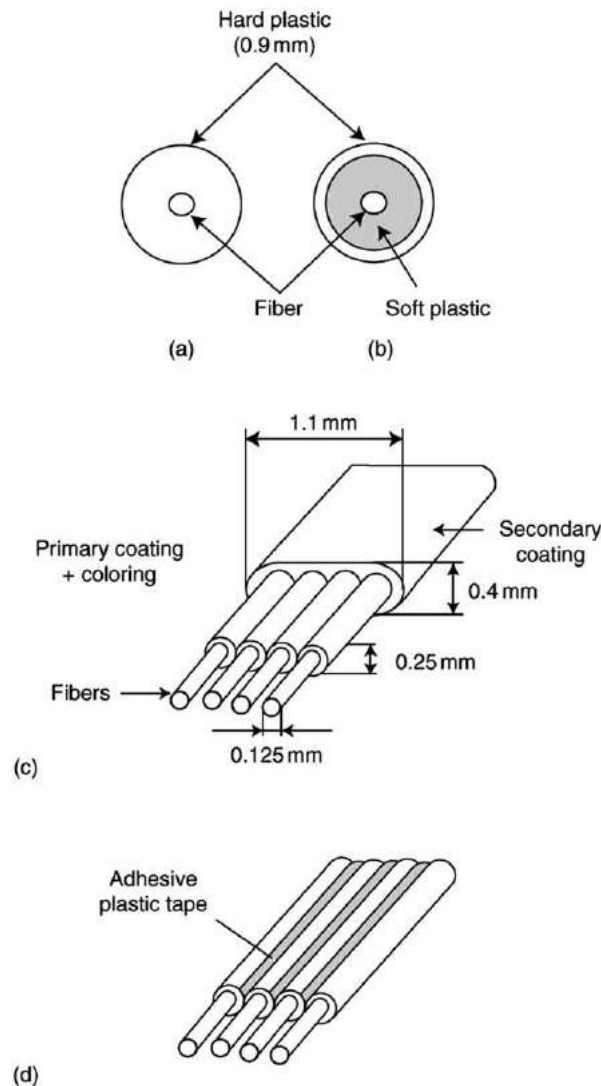


Fig. 1 Different types of tight coatings.

Loose Jacket

In a loose protection the fiber can be inserted into a cable with only the primary coating. A thin layer of colored plastic may be added onto the primary coating to identify the fiber in the cable. The fibers are arranged in a tube or in a groove on a plastic core (slotted core). The diameter of the tube or the dimension of the core must be far larger than the diameter of the fiber so that the fiber is completely free inside the structure. The fibers can be inserted into a tube or into a slotted core, singularly or in groups. The space inside the tube around the fibers can be empty or filled with a suitable jelly. Some examples of loose structures are shown in [Fig. 2](#). In the tube structure, the external diameter of the jacket depends on the number of the fibers: for a single fiber up to 2 mm ([Fig. 2\(a\)](#)), and for a multifiber tube ([Fig. 2\(b\)](#)) up to 3 mm (for 10 fibers).

Examples of slotted core structures are shown in [Fig. 2\(c\)](#) and [\(d\)](#), respectively, for one fiber and for a group of fibers, the dimensions of grooves depending on the number of fibers.

Plastic materials with a high Young modulus are normally used for the tube that is extruded on the single or on the group of fibers. Instead, for the slotted core, plastic materials with good extrusion are used to obtain a precise profile of the grooves.

The loose structure guarantees freedom to the fibers in a defined interval of elongation and compression. Out of this interval, microbending effects can arise which increase fiber attenuation. Axial compression generally occurs when the cable is subjected to low temperatures and is generated from the thermal contraction of the cable materials that is different from that of the fibers. Axial elongation normally occurs when the cable is pulled or by the effects of high temperature.

For a correct design of a cable with a loose structure, the extra length of the fiber with respect to the tube or groove dimension and the stranding pitch must be evaluated, starting from the range of axial compression and elongation. In fact, care must be taken to prevent the fiber from being strained or forced against the tube or groove walls in the designed range.

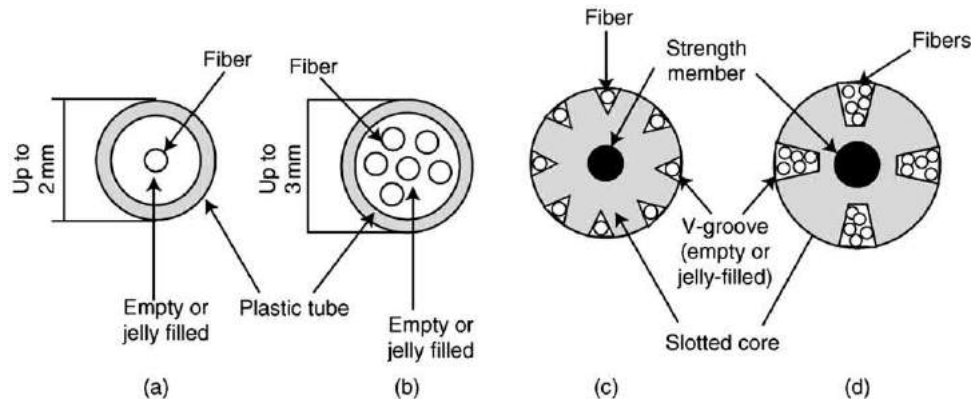


Fig. 2 Different types of loose coatings.

For the tube solution, the extra length of the fibers is controlled during the tube extrusion; for the slotted core solution, the length is controlled during cabling, giving a controlled tension to the slotted core. The loose structure is often used for the cabling of fibers with tight jacketing, joining the advantages of the tight protection and loose cable. This solution is typical for ribbon cables.

Fiber Identification

To identify the fibers in a cable, a proper coloration with defined color codes must be given to the single fibers. In the tight protection, the tight buffer is directly colored during extrusion. In the loose protection, a thin colored layer is added on the primary coating of the fiber; tubes or slotted core may be also colored to identify groups of fibers in high potentiality cables.

Cable Components

The fundamental components of an optical cable are: the cable core with the optical fibers, the strength member (one or more) which provide the mechanical strength of the cable, sheaths which provide the external protection, and filling materials.

Cable Core

The optical fibers with the secondary coating (tight or loose) are rejoined together in a cable core. For tight fibers or loose tubes, the cable core is obtained by stranding the fibers or the tubes around central elements that normally act as strength members too (see below).

The stranding pitch must be long enough to avoid excessive bending of the fibers and short enough to guarantee stress apportionment among the fibers in cable bends and maximum elongation of the cable without fiber strain. Normally, a helical stranding is used, but often SO stranding, that consists in an inversion of stranding direction every three or more turns, is employed. The second stranding type makes the cabling easier but may reduce its performance.

The unit core or the slotted core are made of plastic and normally incorporate a central strength member.

Strength Member

The strength member may be defined as an element that provides the mechanical strength of the cable, to withstand elongation and contraction during the installation and to ensure thermal stability. Fiber elongation, when the cable is pulled or the fiber compressed due to low temperatures, depends greatly on the characteristics of the strength member inserted in the cable.

As the maximum elongation values suggested for the fibers are in the range from 20 to 30% of the proof-test value, the maximum elongation allowed for a tight cable is $\sim 0.1\%$ instead of $\sim 0.3\%$ for a loose cable. To warrant these performances, a material with a high Young modulus is normally used as a strength member. Also, the strength member must be light (to avoid excessive cable weight) and flexible: these properties make it easier when laying the cable in ducts.

The strength member may be metallic or nonmetallic and may be inserted in the core (**Fig. 3(a)**) or in the outer part of the cable (**Fig. 3(b)**), normally under the external sheath. Often two strength members are placed in an optical cable: the first one in the core and the second one under the sheath (**Fig. 3(c)**).

Metallic strength members, generally one or more wires of steel, are inserted in the center or in the outer part of the cable. The steel, for its low thermal expansion coefficient and high Young modulus, behaves as a good strength member, particularly when it is arranged in the central core. Steel is a nonexpensive material, but needs protection against corrosion and electrical-induced voltages, such as by lightning or current flow in the ground.

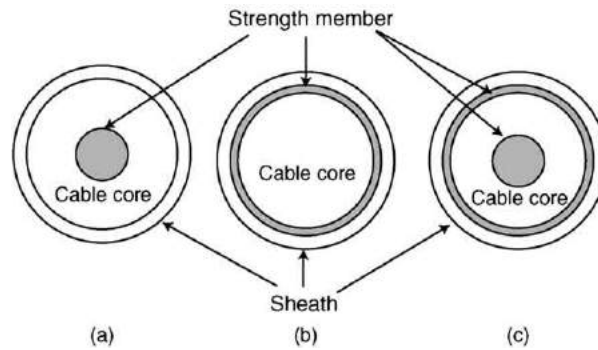


Fig. 3 Different locations of the strength member in an optical cable.

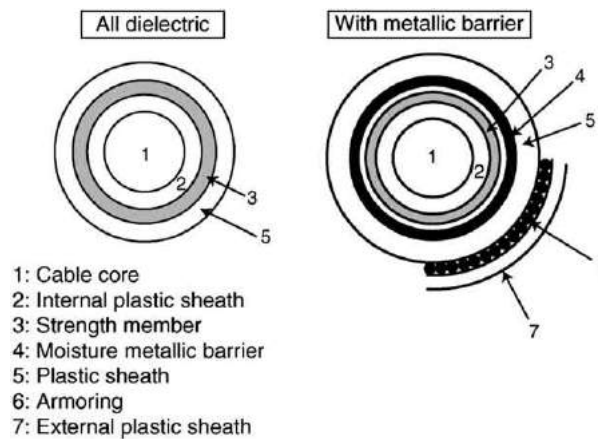


Fig. 4 Examples of optical cable sheaths.

If cables with high flexibility are requested, steel is not suitable and aramid fiber yarn or glass-fiber reinforced plastic (G-FRP) are used, particularly when the strength member is placed in the outer part of the cable. Glass fiber elements are generally used as a compressive strength member in the central cable core to avoid contraction at low temperatures, but they may be used as strength elements when the traction strength is low.

Aramid yarns are not active as resisting compression elements but are normally arranged in the outer part of the optical cable. Aramid yarns have a high Young modulus (as with steel) and low specific weight: in commercial form they consist of a large number of filaments gathered into a layer. Normally two layers stranded in opposite directions are used to avoid cable torsion during the installation.

Metal-free cables are protected by G-FRP and aramid yarns. This solution results in optical cables which are insensitive to electrical interferences.

Sheaths

The functions of the sheaths are to protect an optical cable during its life, in particular from mechanical and environmental stresses. During the installation or when the cable is in service, it may be subjected to mechanical stresses, for example, presence of humidity or water and chemical agents. The sheaths act as a barrier to avoid degradations of the fibers in the internal cable core.

The plastic sheaths (one or more, depending on the cable type and structure) are extruded over the core. This extrusion is a delicate operation in cable manufacturing because the cable core may be subjected to tension, while during the cooling the thermal compression of the plastic produces a stress on all the cable elements. Examples of cable sheaths are shown in [Fig. 4](#).

Different plastic materials, such as polyvinylchloride (PVC), polyethylene, and polyurethane are normally used. PVC has excellent mechanical properties, flexibility, and is flame retardant but is permeable to humidity. On the other hand, high-density polyethylene has a permeability lower than PVC, with good mechanical properties and low coefficient of friction, but it is more flammable than PVC and less flexible. Finally, polyurethane, for its softness and high flexibility, is often used for inner sheaths. When low toxicity is requested (e.g., in indoor cables), halogen-free materials are employed.

Often a metallic barrier is used under the external sheath to prevent moisture entering the cable core. The type most widely used is an aluminum ribbon (of a few tenths of a millimeter) bonded into the external polyethylene sheath, during the sheath extrusion. This structure is normally indicated as a LAP (laminated aluminum polyethylene) sheath.

When the cable is directly buried, a metallic armor (corrugated steel tape or armoring wires) is generally inserted. The metallic armor protects the cable core against radial stresses but it is also used as protection against rodents and insects.

Filling Compounds and Other Components

Fillings are used in optical cables to avoid the presence of moisture and water propagation in the cable core should a failure occur in the cable sheath. Suitable jelly compounds, with constant viscosity at a wide range of temperatures and chemical stabilities, are generally inserted in the cable core and, sometimes, in the outer protection layer. In loose structures, the jelly is in direct contact with the fibers: the compound must be compatible with them and guarantee the fiber freedom. Special compounds may be used to adsorb and permanently fix any hydrogen present in the cable core. The presence of hydrogen gas, that may be generated from the internal materials of the cable (mainly metals), is harmful to the fibers because it causes permanent attenuation increases and thus damage to them.

With the generic terminology 'other components', some other components used in the optical cables are included. They are often similar to those used in conventional cables. The most important are: insulated conductors, plastic fillers, and cushion layers and tapes around the fiber core.

Insulated copper conductors may be incorporated in the cable core, for example, to carry power supply or as service channels, or to detect the presence of moisture in the cable (in this case, the insulation is missing or partial). The conductors may cause problems due to the voltages induced by power lines or lighting.

Plastic fillers are often used to complete the cable core geometry. Cushion layers are used to protect the cable core against radial compression. Normally, they are based on plastic materials wound around the core. Tapes are wound around the cable core to hold the assemblies together and secondly to provide a heat barrier during the sheath extrusion. Mylar tapes are usually employed.

Optical Cable Structures, Types, and Applications

Cable Structures

The structure of an optical cable is strictly dependent on the construction method and the application type that determines the design and the grade of external protection. The main core (or inner) structures of an optical cable can be classified as: stranded structures (tight and loose); slotted core cable; or ribbon cable. In this section, a few examples of cable structures are presented. Below, separate sections are devoted to submarine cables and to special application cables, for their particular structures and features.

In the stranded tight structure, a low number of tight or loose fibers is stranded around a strength member to form a cable unit. Some of this cable unit may be joined to constitute a cable with a medium/high number of fibers (**Fig. 5(a)**). Alternatively the fibers may be arranged in a multilayer structure (**Fig. 5(b)**).

The main advantages of a tight stranded structure are the small dimensions of the cable core, even when a large number of fibers are involved, and also the facility of handling. Care must be taken in the design of the strength member: in fact, the fibers are immediately subjected to tension when the cable is pulled into the ducts or compressed when subjected to low temperatures.

In the stranded loose structures, two categories of cables are defined, one fiber per tube and a group of fibers (up to 20) per tube (**Fig. 5(c)**). As already mentioned, in the loose structures, the fibers are free inside the tube and for this reason the cable is less critical than the tight cable. The external dimensions of loose cables are larger than tight cables, but the solution with multiform tubes allows for cables with a high number of fibers in relatively small dimensions.

Slotted core cables (**Fig. 2(c)** and **(d)**) contain fibers with only the primary coating and may be divided, as in the stranded loose structure, into one fiber per groove or into a group of fibers per groove. The cable design and the choice of a suitable plastic

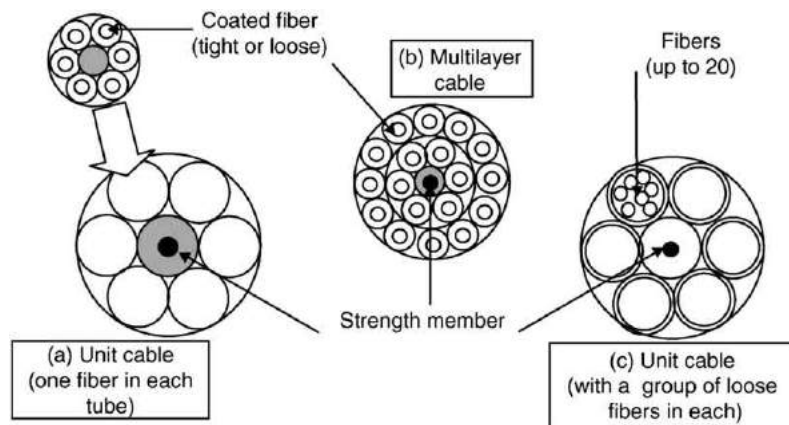


Fig. 5 Structures of stranded core cables.

material for the slotted core joined to a proper strength member, make it possible to obtain cables with fiber stability at a wide range of temperature. As in the loose stranded solutions, the slotted core solution is adopted to obtain cables with a high number of fibers in small dimensions.

In the ribbon solution, the fibers (from 4 to 16) are placed in a linear ribbon array. Many ribbons are stacked together to constitute the optical unit. The ribbon cable normally has a loose structure (loose tube or slotted core). Three structures are shown in Fig. 6. In the first example (classical AT & T cable), 12 ribbons of 12 fibers are directly stacked and stranded in a tube (Fig. 6(a)). In the second and third examples, ribbons are inserted into tubes or grooves (Fig. 6(b) and (c)). Several tubes or slotted cores can be assembled together to give a cable with very high fiber density. The ribbon approach may be efficient when a large number of fibers must be simultaneously handled, spliced, and terminated at connectors.

Types and Applications

The choice of an optical cable is heavily established by many factors, but mainly on the type of installation. On this basis, some different typology of cables may be identified: aerial cables, duct cables, direct buried cables, indoor cables, underwater cables, etc.

Aerial cables can be clasped onto a metal strand or can be self supporting. In the first case, the cable is not subjected to a particular tension but requires good thermal and mechanical performances. This solution is chosen when the cable must withstand strong ice and wind and when long distances between the poles are required. All the structures mentioned above may be adopted. In the second case, the cable is normally exposed to high mechanical stresses during its life and high tensile strength must be guaranteed to maintain the fiber elongation at a safety level. Generally, loose structures are preferred for their greater fiber freedom. Often, in self-supporting cables, the strength member is external to the cable core. An example is shown in Fig. 7.

Cables pulled in ducts must be resistant to pulling and torsion forces, and light and flexible to permit the installation of long sections. The cable must also protect the fibers against water and moisture that may be present in ducts or manholes. Usually the cable is filled with a jelly compound, a metallic barrier is normally used and, often, an armoring is added when protection against rodents is necessary. All the structures described above may be employed in duct cables.

The characteristics of direct buried cables are very similar to those of duct cables, but additional protections, such as metallic armoring, are required to avoid the risk of damages from digging and other earthmoving work.

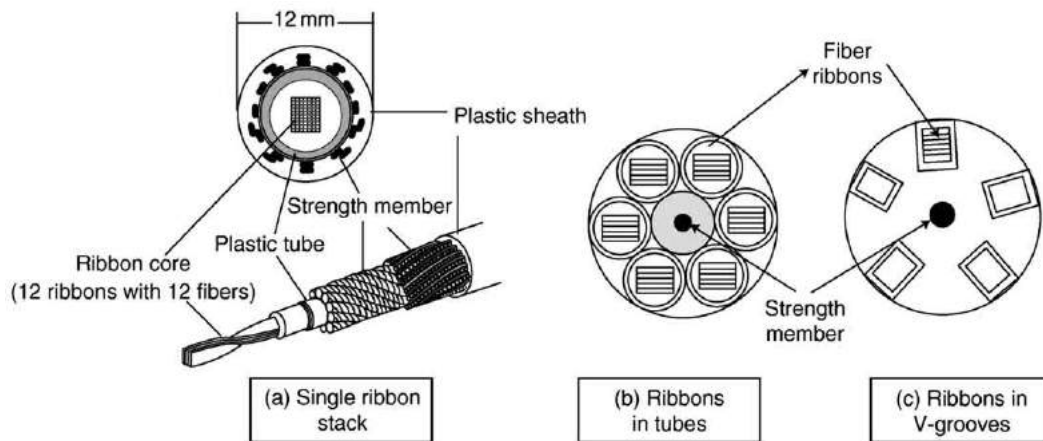


Fig. 6 Structures of ribbon cables.

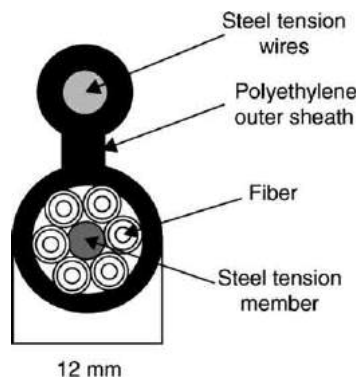


Fig. 7 Self-supporting aerial cable.

Cables for indoor applications normally require a smaller number of fibers. Their main characteristics are: small dimensions, flexibility, small curvature radius, and ease in handling and jointing. For indoor applications, the sheath must not permit flame propagation and must have a low emission of toxic gases and dark smoke. Tight cables are preferably used for indoor applications due to their compactness and small dimensions.

Underwater cables are employed in the crossing of rivers, in lakes and lagoons, and for shallow water applications in general. Such cables must have stringent requirements such as resistance to moisture and water penetration and propagation, pulling resistance (during the installation, and recovery in event of failure), and high resistance to static pressure and core structure compatible with land cables (when it is part of a ground link). In addition, due to the presence of metallic structures, attention should be paid to hydrogen effects.

Submarine Cables

Submarine cable is manufactured for laying in deep-sea conditions. In designing a submarine cable, it is necessary to provide high reliability that the mechanical and transmission characteristics of the optical fibers will be stable over a long period of time. Different cables are used for submarine applications, but generally follow the basic scheme of Fig. 8.

The central core (with a few fibers – up to 12 – in a slotted core or in a tight structure) is surrounded by a double layer of steel wires that act as strength members against tensile action and water pressure. Metal pipes at the inner or outer sides of the strength member, act as a water barrier and as power supply conductors (for the supply of the undersea regenerators).

The outer polyethylene insulating sheath is the ultimate protection for the ordinary deep-sea cable. For cables requiring special protection, steel wire armoring (anti-attack from fishes), and further polyethylene are added.

The cable must have stringent mechanical requirements, such as resistance to traction, torsion, crushing, impact, and to shark attack. The cable must be suitable for installation using standard cable-laying ships. Furthermore the cable materials must have low content of hydrogen and emissions. If necessary, a hydrogen absorber may be included in the cable core.

Special Cables

This category comprises cables not specifically used in ordinary telecommunication networks, but cables used for specific applications and according to specified requirements. Cables for military use are normally employed for temporary plant restoration. They must be light but crush resistant and with good mechanical characteristics. Cables for mobile applications (robots, elevators, aircrafts, submarines, etc.) generally require high flexibility and tensile strength. Special cables may be installed in particular environments, where they must be insensitive, for instance, to nuclear radiation, chemical agents, and high temperatures.

Optical cables can be installed in power lines, together with the power conductors, or into the ground wires of high-voltage overhead power lines. In this last type of cable, the central stranded steel wire, making up the ground wire, is replaced with a metal tube containing the optical unit (normally a small groups of optical fibers inserted in a tube or in a slotted core).

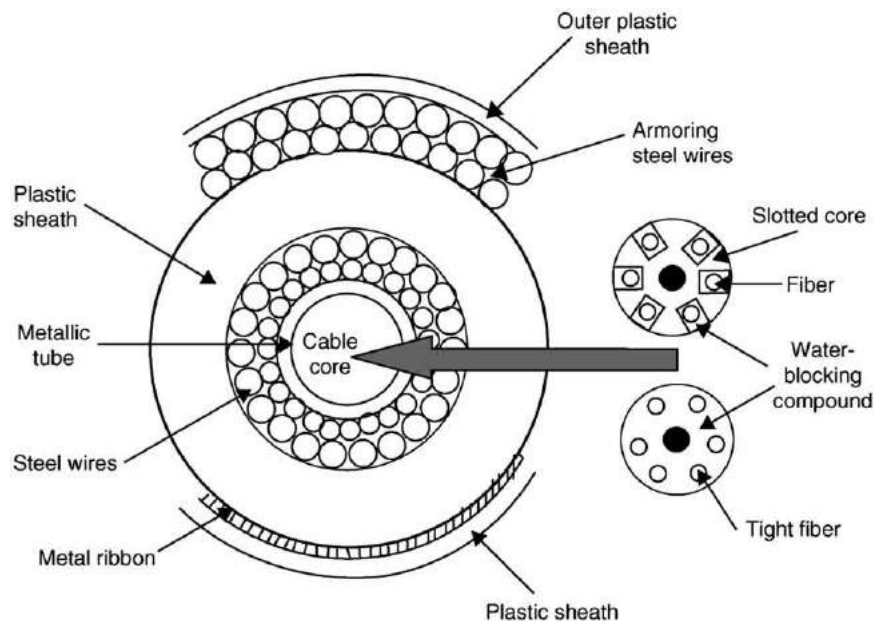


Fig. 8 Basic structure of a submarine cable.

Blown Fiber

This technology consists of the installation of a single fiber or of small fiber bundles (or cables) into pre-installed tubes or ducts using a flux of compressed air. Initially, cables with empty tubes are installed and successively, at the time of need, the tubes can be filled by blowing fibers (singly, or in bundles) or cables. The installation time is normally fast; some compressors can blow tens of meters in a minute. The typical maximum blowing distance is about 1 km for horizontal planes, but it is reduced along winding roads. The advantage of this technique is to produce variable and flexible structures, increasing the number of fibers to fit the network need. Besides, substitutions or change of fibers can be quickly carried out without substitution of the cable.

See also: Fabrication of Optical Fiber

Further Reading

Griffioen, W., 1989. The installation of conventional fiber-optic cables in conduits using the viscous flow of air. *Journal of Lightwave Technology* 7 (2), 297–302.
Hughes, H., 1997. *Telecommunications Cables. Design, Manufacture and Installation*. John Wiley and Sons.
Keiser, G., 1986. *Optical Fiber Communications*. McGraw-Hill.
Murata, H., 1988. *Handbook of Optical Fibers and Cables*. Marcel Dekker.
Technical Staff of CSELT, 1990. *Fiber Optic Communications Handbook*, second ed. TAB Books.

Passive Optical Components

D Suino, Telecom Italia Lab, Torino, Italy

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Modern telecommunications networks are based on fiber optics as signal transporters. In these networks the signals travel in the form of light confined within the fibers. As light power travels through the fibers from the source apparatus to the receiver, a number of components are employed to manage the optical signal. All other components that are not used to manage the signal are called 'passive optical components'. They serve several functions, such as to join two different fibers to distribute the signal onto more optical branches to limit the optical power to select particular wavelengths, and so on.

These components can be located anywhere in the network, depending on their function and on the network architecture. However, their main location is near the apparatus in which it is necessary to manage the signal before it is sent through the network or at the junction in the network where it is sent on to the receiver.

The main types of passive components that are found in a network can be grouped as follows:

- Joint devices:
 - Optical connectors;
 - Mechanical splices.
- Branching devices:
 - Nonwavelength-selective branching devices;
 - WDM.
- Attenuators;
- Filters;
- Isolators;
- Circulators.

Optical fibers are also divided into several groups (multimode 50/125, multimode 62.5/125, singlemode, dispersion shifted, NZD, and so on). There are also components of different typology that match up with each particular fiber. However, the general concept in terms of functionality, technology, and characteristic parameters, are the same among passive optical components of the same type (connectors, branching devices, or attenuators) but of different groups (single-mode, multimode, etc.).

Optical Joint Devices

The function of an optical joint is to perform a junction between two fibers, giving optical continuity to the line. Two different classes of joint can be considered: the connectors and the splices. The joints made by using connectors, are flexible points in the network, which means they can be opened and reconnected several times in order, for example, to reconfigure the network plan. However, a splice is a fixed joint and usually cannot be opened and re-mated. There are two type of splice: mechanical splices, holding joints to fibers by glue or mechanical crimp, and fusion splices fusing the two fiber heads with a suitable fusion splicer.

General Considerations

Considering that fibers can be described as pipes, in the core of which flows light, continuity between two fibers means that the two cores must be aligned in a stable and accurate way. Obviously this alignment must be carried out with minimum power loss in the junction. The main factors that affect the power loss in a joint point between two fibers are:

- x =Lateral offset;
- z =Longitudinal offset;
- θ =Angular misalignment; and
- w_T/w_R =Mode field diameter ratio (in the single mode fiber) or numerical aperture ratio (in the multimode fiber).

These geometrical mismatching between the two fibers is represented in [Fig. 1](#).

The function that describes the joint attenuation is different for single-mode and multimode connectors, but the parameters that contribute to it are the same. The general formulation for these functions are complicated and, in particular, for the multimode connector it involves an integral on the shape distribution of the light in the fiber core. These functions become simpler in particular conditions: for example, if we consider a single-mode joint without a gap between the two fibers (that is the more usual condition in the modern connectors) the z parameter is nil, and in this case the function that describes the attenuation

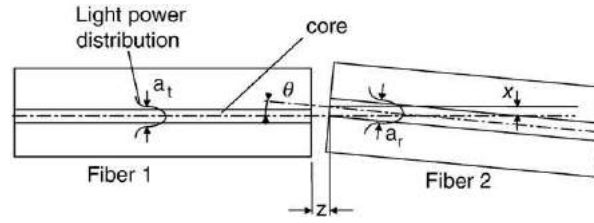


Fig. 1 Main parameter that affects the joint power loss.

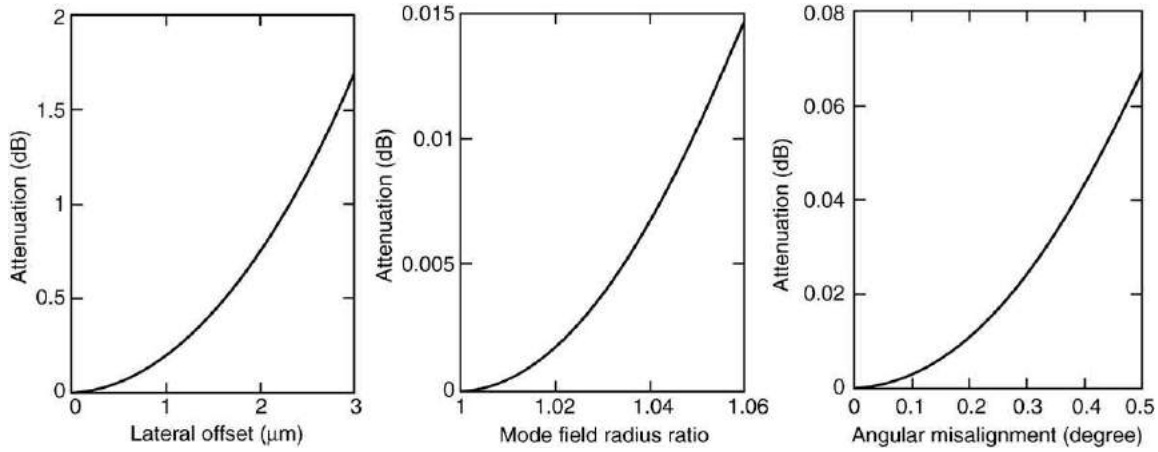


Fig. 2 Single mode connector attenuation as a function of the geometrical parameters. Each curve is plotted keeping the other parameter constant.

of the connector is

$$A(\text{dB}) = -10 \log \left\{ \left(\frac{2w_T w_R}{w_T^2 + w_R^2} \right)^2 \times \exp \left[-2 \frac{(x\lambda)^2 + (\pi n w_T w_R \sin(\theta))^2}{\lambda^2 (w_T^2 + w_R^2)} \right] \right\} \quad (1)$$

where n is the fiber core refractive index and λ is the wavelength of the light.

Fig. 2 shows the trend of **Eq. (1)**, where the parameter ranges in the graphs are typical for the most common products on the market. It is evident that the main contribution to attenuation arises from the lateral offset.

For connectors of good quality, both for single-mode and multimode fibers, the attenuation values are in the range of a few tenths of dB (typically less than 0.3 dB). Another important factor for a joint, in particular for connectors used in networks for high speed or analog signals, is the return loss (RL). This is a parameter that gives a measure of the quantity of light back-reflected onto the optical discontinuity of the connector, due to the Fresnel effect. The RL is defined as the ratio in dB between the incident light power (P_{inc}) and the reflected light power (P_{ref}):

$$\text{RL} = -10 \log \left(\frac{P_{\text{ref}}}{P_{\text{inc}}} \right) \quad (2)$$

Fresnel reflection arises when the light passes between two media with different refractive indices. In the case of connectors in which the two fibers are not in contact, the light passes from the silica of the fiber core to the air in the gap. This causes the reflection of 4% of the incident light that, as RL, is about 14 dB.

To improve the RL performances of an optical joint, it is necessary to reduce the refractive index difference. This can be achieved either by index matching material between the two fibers that reduces the differences between the refractive index of the fiber and the gap, or by performing a physical contact between the fiber heads. The first technique is typically used in mechanical splices but, due to the problem related with the pollution contamination of the index matching, it is not a good solution for connectors that are to be opened and reconnected several times.

The physical contact solution is the most frequently adopted for these connectors. Physical contact is obtained by polishing the ferrule and the fiber end-faces in a convex shape, with the fiber core positioned in the apex (**Fig. 3**). With this technique, the typical values of return loss for a single mode connector are in the range from 35 dB to 55 dB, depending on the polishing grade. The residual back-reflected light is generated in the thin layer of the fiber end surface, because of slight changes in the refractive index due to the polishing process.

Higher return loss values (small quantity of reflected light) are achieved with the angled physical contact (APC) technique. This is obtained by polishing the convex ferrule end face angled with respect to the fiber axis. In this way, the light is not back reflected

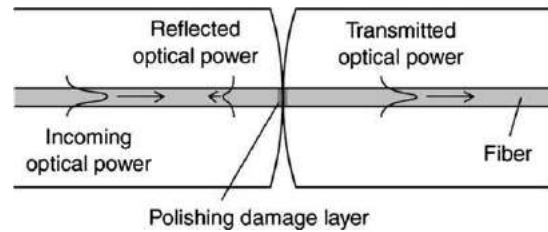


Fig. 3 Back reflection effects in PC connectors.

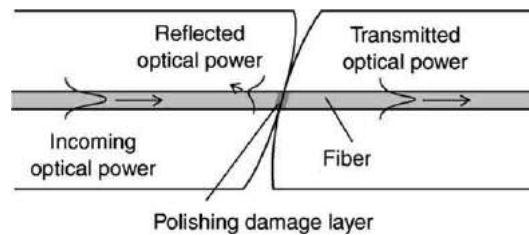


Fig. 4 Back reflection effects in APC connectors.

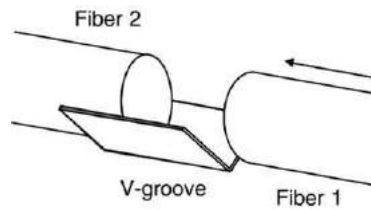


Fig. 5 Scheme of the bare fibers alignment on a V-groove.

into the fiber core, but into the cladding, and is thus eliminated (Fig. 4). In APC connectors, typical return loss values are greater than 60 dB.

Optical Connectors

On the market are several types of optical connectors. It is possible to divide them into three main groups:

- Connectors with direct alignment of bare fiber;
- Connectors with alignment by means of a ferrule; and
- Multifiber connectors.

Connectors with direct alignment of bare fiber

In this type of connector the bare fibers are aligned on an external structure that is usually a V-groove (Fig. 5). Two configurations can be performed: plug and socket or plug–adapter–plug. In the first case, the fiber in the socket is fixed and the one in the plug aligns to it. In the plug–adapter–plug configuration, two identical plugs containing the bare fibers are locked onto an adapter in which the fibers are placed on the aligned structure. These connectors are cheaper than the connectors with a ferrule, (described below), but this is unreliable. In fact, the bare fibers are delicate and brittle and they can be affected by the external mechanical or environmental stresses.

Connectors with alignment by means ferrule

In this type of connector, the fibers are fixed in a secondary alignment structure, usually of cylindrical shape, called a ferrule. The fiber alignment is obtained by inserting the two ferrules containing the fibers into a sleeve. Usually this type of connector is in the plug–adapter–plug configuration (Fig. 6). With this technique, the precision of the alignment is moved from the V-groove precision to the dimensional tolerance of the ferrule. Typical tolerance values on the parameters for a cylindrical ferrule with a diameter of 2.5 mm, are in the range of a tenth of a micrometer, and must be verified on the diameter, eccentricity between the hole (through which the fiber is fixed) and cylindricity of the external surface.

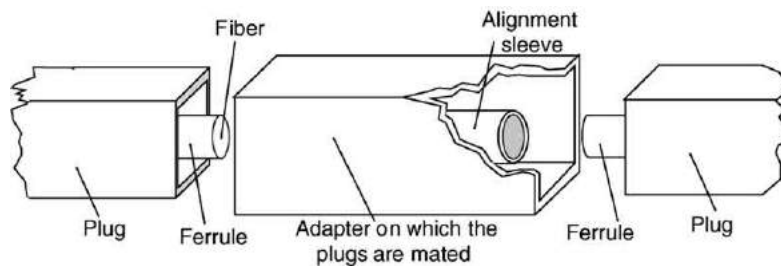


Fig. 6 Scheme of a plug-adaptor-plug connector with alignment by ferrules and sleeve.

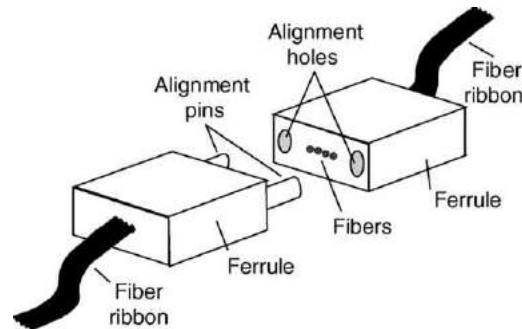


Fig. 7 Multifiber connector.

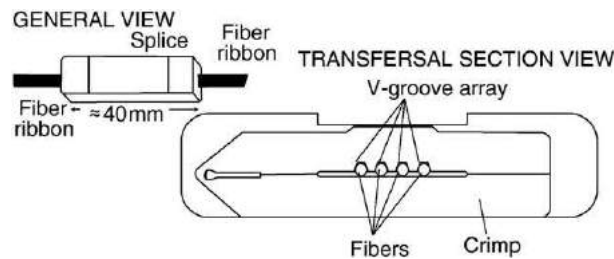


Fig. 8 Example of a multifiber mechanical splice.

Multifiber connectors

In some types of optical cable, the fibers are organized in multiple structures, in which many fibers are glued together to form a ribbon. The ribbon can have 2, 4, 8, 12, or 16 fibers, for example. When these cables are used, it is convenient to maintain the ribbon fiber structure in the connectors (until the point in which it is necessary to use the fibers singularly). Multifiber connectors are typically based on a secondary alignment on a plastic ferrule of rectangular shape in which the fibers maintain the same geometry that they have in the ribbon. The rectangular ferrules of the two plugs are normally aligned by means of two metallic pins placed in two hole in line with the fiber. In Fig. 7, a scheme of a multifiber connector is shown.

Mechanical Splices

Mechanical splices are yet another way to join together two fibers. All the main considerations for the connectors are true for the mechanical splices. The difference is that the mechanical splice is a fixed joint and, normally, it cannot be opened and re-mated. Mechanical splices, both for single fiber and for multifiber structures, are the most usually manufactured. In a mechanical splice, the fibers are aligned directly in a V-groove and fixed by glue or a mechanical holder.

These components are cheaper than fusion splices but their optical, mechanical, and environmental functions are lower than the fusion splices. Fig. 8, shows a mechanical splice for multifibers in which the fibers are crimped on a V-groove array. Usually, to avoid high optical reflection in the junction point, this type of joint is filled with index matching material.

Branching Devices

Branching devices are passive components devoted to distributing the optical power from an input fiber to two or more fibers. These components can be classified in two broad classes, on the basis of the wavelength dependence. The nonwavelength-selective

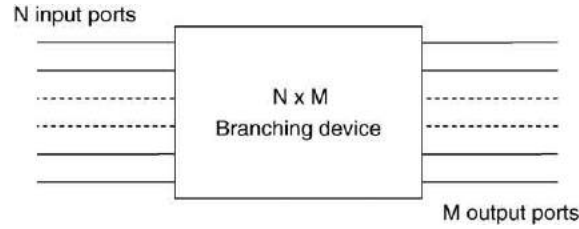


Fig. 9 Scheme of a generic $N \times M$ branching device.

branching devices, called also couplers or splitters, share the input power to two or more fibers independently from the light wavelength. The components that perform the light distribution on more fibers on the basis of the different wavelengths of the input light are called a WDM (wavelength division (de)multiplexer).

Branching devices have three or more ports, bare fiber or connectors, for the input and/or output of optical power, and share optical power among these ports. A generic branching device can be represented as a multiport device having N input ports and M output ports, as shown in the Fig. 9.

The optical characteristics of a branching device can be represented by a square transfer matrix of $n \times n$ coefficients, where n is the total number of the component ports. Each coefficient t_{ij} in the matrix is the fractional optical power transferred among designated ports, that is the ratio of the optical power P_{ij} transferred out of port j with respect to input power P_i into port n :

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & & & \\ \vdots & & t_{ij} & \vdots \\ & & & \ddots \\ t_{n1} & & & t_{nn} \end{bmatrix} \quad (3)$$

where:

$$t_{ij} = \frac{P_{ij}}{P_i} \quad (4)$$

So, according to the relationship between two ports, the coefficients represent several parameters. For example, if the port i is an input port and j is an output port, t_{ij} is the transferred power signal through the component, while if they exist as two isolated ports the coefficients represent the crosstalk between the two ports. The three main techniques to obtain this type of component are:

- (i) All fiber components;
- (ii) Micro-optics; and
- (iii) Integrated optics.

Nonwavelength-Selective Branching Devices

A nonwavelength-selective branching device is a component having three or more ports, which shares an input signal among the output ports in a predetermined fashion. Usually these components are completely reversible and so they also work as a combiner of signals from more input ports onto a single output port. There are several types of couplers for different applications:

- Y-coupler with one input port and two output ports;
- X-coupler with two input ports and two output ports; and
- Star coupler with more than two ports in input and/or in output.

Moreover the couplers can be symmetrical, and in this case they divide the input power light onto n equal output flows, or asymmetrical devices in which the quantity of power light in the output ports is shared in a predefined nonuniform way.

The techniques used to carry out this type of device are mainly the fusion technique or planar technique. The first one is based on the phenomenon of evanescent wave coupling: when the core of two optical guides (as fibers) are located in close proximity, a power exchange from a guide to another is possible. This configuration can be obtained by stretching the fibers under controlled fusion (see Fig. 10).

The planar technique, that is the simpler integrated optic application, is based on the optical guides with appropriate shapes being placed directly onto a silicon wafer by means of lithographic processes (Fig. 11).

The fusion technique has the advantage of being 'all in fiber' so the component is ready to be inserted into the optical networks by means of usual joint techniques. But for a high number of ports, the coupling ratio among the branches cannot be controlled in

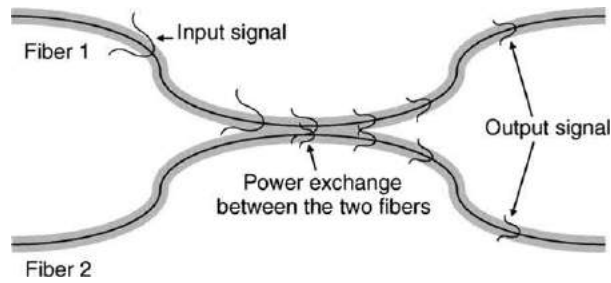


Fig. 10 Scheme of a fusion 2×2 coupler.

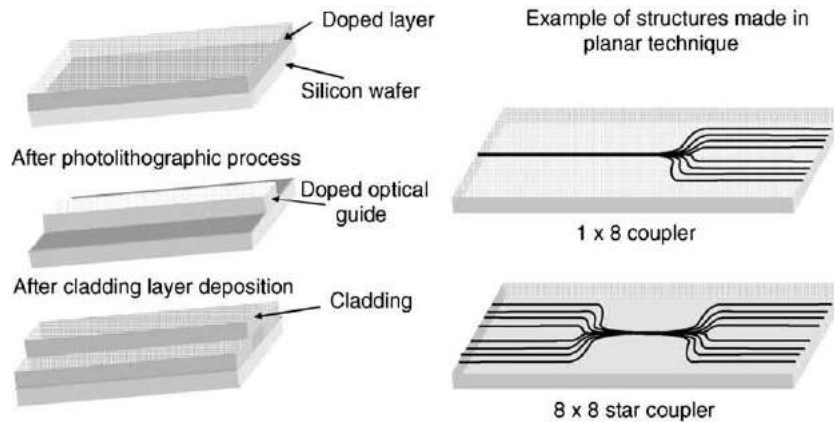


Fig. 11 Scheme of the process to obtain a planar coupler and an example of a device.

a precise way and the cost of the fusion device is proportional to the number of the branches. On the contrary, a device obtained with the planar technology must be coupled with fibers normally used and this operation may be difficult due to the different geometry of the planar guide (rectangular) and the fiber (circular). On the other hand, this technology allows a high precision in the optical characteristics of the component and the cost of the device is independent of the branch number.

WDM

A wavelength-selective branching device, usually called WDM (Wavelength Depending Multiplexer), is a component having three or more ports which shares an input signal among the output ports in a predetermined fashion, depending on the wavelength, so that at least two different wavelength ranges are nominally transferred between two different couples of ports.

WDM devices may be divided in three categories on the basis of the bandwidth of the channels (the spacing between two adjacent wavelength discriminated by the device):

- (i) DWDM (Dense WDM) device operates for channel spacing equal or less than 1000 GHz (about 6–8 nm in the range of typical optical wavelength used in telecommunications systems, 1310 or 1550 nm);
- (ii) CWDM (Coarse WDM) device operates for channel spacing less than 50 nm and greater than 1000 GHz; and
- (iii) WWDM (Wide WDM) device operates for channel spacing equal or greater than 50 nm.

These components are used to combine, at the input side of an optical link in only one fiber, more optical signals with different wavelengths and separate them at the receiving end. This technology allows to send along a fiber, more communication channels, thus increasing the total bit rate transmitted. These types of components can be obtained using filters to select single a wavelength (as a diffractive grating that resolves the light into its monochromatic components) or using the phenomena of evanescent field coupling occurring between two adjacent fiber cores (as described for coupler devices), controlling in a precise way the coupling zone; this is, in fact, dependent on the wavelength (Fig. 12).

Using wavelength filters, as they select a singular wavelength, it is necessary to perform a cascade filter structure for separating more than two wavelengths.

Micro-optic technologies are generally used for multimode fiber due to the critical collimation problem. For the single-mode fiber the evanescent field coupling technique is applied directly to fiber or to planar optical guides.

Another efficient technique to create WDM devices is with a fiber Bragg grating (FBG) wavelength filter directly applied to the fiber. FBGs consist in periodic modulation (see Fig. 13) of the refractive index along the core of the fiber, and they act as reflection filters.

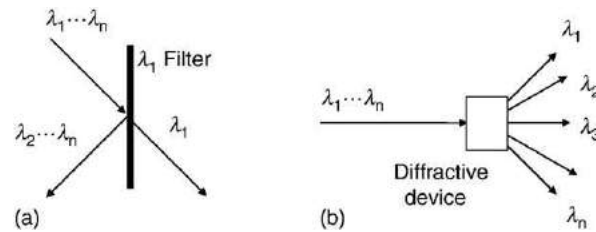


Fig. 12 (a) Scheme of the filter technique and (b) diffractive device technique.

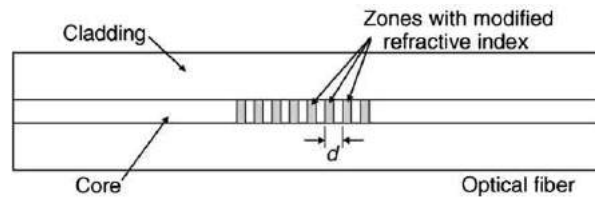


Fig. 13 Scheme of FBG.

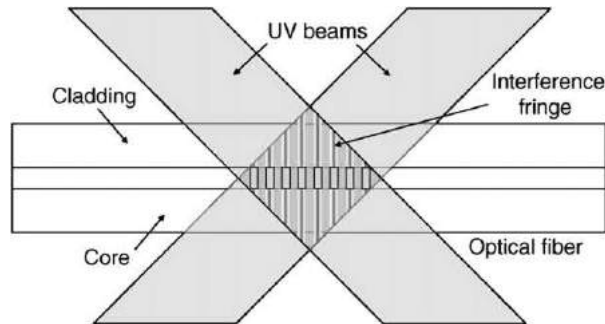


Fig. 14 Scheme of FBG production.

During the fabrication process (see Fig. 14) the core of the optical fiber is exposed to UV radiation, and this exposure increases the refractive index of the core in defined zones, with a periodic shape obtained by the interference fringe created by crossing two coherent beams.

Fig. 15 shows a representative spectral curve of a FBG filter centered at 1625 nm. The central wavelength, the width, and the value of reflection depend on the grating period (d) and on the deposited UV energy in the fiber core.

Other Passive Optical Components

In an optical network, besides the devices described above, there are a number of other passive optical components that perform particular functions on the optical signal; the main ones are:

Attenuators These are devices used to attenuate the optical power, for example, in front of a receiver when the optical line is short and the power of the received signal is still too high for the detector, or when it is necessary to equalize the signals arriving from different optical lines. It exists both in tunable and fixed versions. The first ones allow adjustment of the attenuation value, however, the second ones have a fixed and predefined value. The fixed optical attenuators can be obtained by inserting a filter or stressing a fiber in a controlled and permanent way. The tunable attenuators are based on variable optical parameters as, for example, the distance or the lateral offset in an optical connection. They exist both as connect lengths of fiber (attenuated optical jumpers) or as compact components similar to a double connector (plug stile attenuator).

Filters These are components with two ports employed to select or to filter some fixed wavelengths. They can be divided into the following categories:

- short-wave pass (only wavelengths lower than or equal to a specified value are passed);
- long-wave pass (only wavelengths greater than or equal to a specified value are passed);
- band-pass (only an optical window is allowed); and
- notch (only an optical window is inhibited).

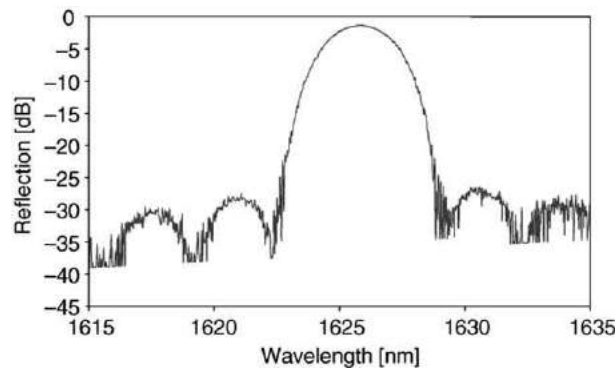


Fig. 15 Transmission characteristic of FBG.

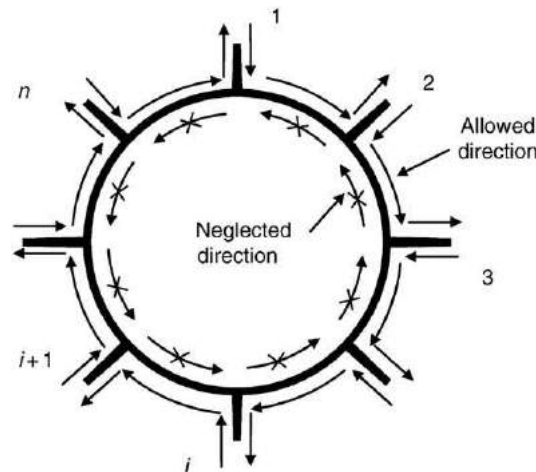


Fig. 16 Principle scheme of an optical circulator.

It is also possible to have a combination of the above categories. The optical filtering can be obtained by inserting one or more filtering devices (as interferential films) into a fiber or creating a filter directly on the fiber by the FBG described above.

Isolators These are nonreciprocal two-port optical device in which the light flows in the forward direction, and not in the reverse one. The isolators are used to suppress backward reflections along an optical fiber transmission line, while having minimum insertion loss in the forward direction. Fiber optic isolators are commonly used to avoid back reflections into laser diodes and optical amplifiers, which can make the laser and amplifiers oscillate unstably, and cause noise in the fiber optic transmission system.

Circulators These are passive components having three or more ports numbered sequentially (1, 2, ..., n) through which optical power is transmitted nonreciprocally from port 1 to port 2, ..., from port (i) to port ($i + 1$), ... and from port n to port 1 (Fig. 16).

See also: Broadband Passive Optical Access Networks

Further Reading

- Adam, M.J., 1981. *An Introduction to Optical Waveguides*. New York: John Wiley & Sons.
- Blonder, G.E., *et al.*, 1989. Glass waveguides on silicon for hybrid optical packaging. *IEEE Journal of Lightwave Technology* 7, 1530.
- Digonnet, M., Shaw, H.J., 1983. Wavelength multiplexing in single-mode fiber couplers. *Applied Optics* 22, 484.
- Jeunhomme, L.B., 1983. *Single-Mode Fiber Optics – Principles and Applications*. New York: Marcel Dekker.
- Kashima N (1995) *Passive Optical Components for Optical Fiber Transmission...*
- Kashima, N., 1993. *Optical Transmission for the Subscriber Loop*. Norwood, MA: Artech House.
- Miller, C.M., 1986. *Optical Fiber Splices and Connectors – Theory and Methods*. New York: Marcel Dekker.
- SC86B IEC, 1999. IEC 60874-1: Connectors for Optical Fibres and Cables – Part 1: Generic Specification, 4th edn. Geneva: International Electrotechnical Commission.
- SC86B IEC, 1999. IEC 61073-1: Optical Fibres and Cables – Mechanical Splices and Fusion Splice Protectors – Part 1: Generic Specification, 3rd edn. Geneva: International Electrotechnical Commission.

- SC86B IEC, 1999. IEC 60869-1: Fibre Optic Attenuators – Part 1: Generic Specification, 3rd edn. Geneva: International Electrotechnical Commission.
- SC86B IEC, 2000. IEC 60875-1: Non-wavelength Selective Fibre Optic Branching Devices – Part 1: Generic Specification, 4th edn. Geneva: International Electrotechnical Commission.
- SC86B IEC, 2000. IEC 62077-1: Fibre Optic Circulators – Part 1: Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
- SC86B IEC, 2000. IEC 61977-1 Ed. 1.0: Fibre Optic Filters – Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
- SC86B IEC, 2000. IEC 61202-1: Fibre Optic Isolators – Part 1: Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
- Technical Staff of CSELT, 1990. Fiber Optic Communications Handbook, 2nd edn. New York: McGraw-Hill.
- Yokohama, I., *et al.*, 1988. Novel mass-fabrication of fiber couplers using arrayed fiber ribbons. *Electronic Letters* 24, 1147.

Nonlinear Effects (Basics)

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

n_2 Nonlinear refractive index [$\text{m}^2 \text{W}^{-1}$]

I Optical intensity [$\text{GW cm}^{-2} = 10^9 \text{W cm}^{-2}$]

CPM cross phase modulation

FWM four wave mixing

MI modulational instability

SPM self-phase modulation

Introduction

Many physical systems in various areas such as condensed matter or plasma physics, biological sciences, or optics, give rise to localized large-amplitude excitations having a relatively long lifetime. Such excitations lead to a host of phenomena referred to as nonlinear phenomena. Of the many disciplines of physics, the optics field is probably the one in which practical applications of nonlinear phenomena have been the most fruitful, in particular, since the discovery of the laser in 1960. This discovery has thus led to the advent of a new branch in optics, referred to as *nonlinear optics*. The applications of nonlinear phenomena in optics include the design of various kinds of laser sources, optical amplifiers, light converters, light-wave communication systems for data transmission purposes, to name a few.

In this article, we present an overview of some basic principles of nonlinear phenomena that result from the interaction of light waves with dielectric waveguides such as optical fibers. These nonlinear phenomena can be broadly divided into two main categories, namely, parametric effects and scattering phenomena. Parametric interactions arise whenever the state of the dielectric matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between energy levels in the medium. More fundamentally, parametric interactions originate from the electron motion under the electric field of a light wave, whereas scattering phenomena originate from the motion of heavy ions (or molecules).

Linear and Nonlinear Signatures

The macroscopic properties of a physical system can be obtained by analyzing the response of the system under an external excitation. For example, consider at time t the response of a system, such as an amplifier, to an input signal $E_1 = A \sin(\omega t)$. In the low-amplitude limit of the output signal, the response R_1 of the system is proportional to the excitation:

$$R_1 = \alpha_1 E_1 \quad (1)$$

where α_1 is a constant. This type of behavior corresponds to the so-called linear response. In general, a physical system executes a linear response when the superposition of two (or more) input signals E_1 and E_2 yields a response which is a superposition of the output signals, as schematically represented in Fig. 1:

$$R = \alpha_1 R_1 + \alpha_2 R_2 \quad (2)$$

Now, in almost all real physical systems, if the amplitude of an excitation E_1 becomes sufficiently large, distortions will occur in the output signals. In other words, the response of the system will no longer be proportional to the excitation, and consequently,

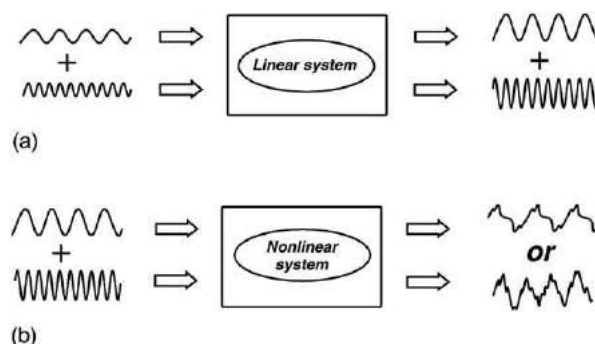


Fig. 1 Schematic representation of linear and nonlinear responses of a system, to two input signals.

the law of superposition of states will no longer be observed. In this case, the response of the system may take the following form:

$$R = \alpha_1 E_1 + \alpha_2 E_1^2 + \alpha_3 E_1^3 + \dots \quad (3)$$

which involves not only a signal at the input frequency ω , but also signals at frequencies 2ω , 3ω , and so on. Thus, harmonics of the input signal are generated. This behavior, called *nonlinear response*, is at the origin of a host of important phenomena in many branches of sciences, such as in condensed matter physics, biological sciences, or in optics.

Physical Origin of Optical Nonlinearity

Optical nonlinearity originates fundamentally from the action of the electric field of a light wave on the charged particles of a dielectric waveguide. In contrast to conductors where charges can move throughout the material, dielectric media consist of *bound* charges (ions, electrons) that can execute only relatively limited displacements around their equilibrium positions. The electric field of an incoming light wave will cause the positive charges to move in the polarization direction of the electric field whereas negative charges will move in the opposite direction. In other words, the electric field will cause the charged particles to become dipoles, as schematically represented in Fig. 2.

Then each dipole will vibrate under the influence of the incoming light, thus becoming a source of radiation. The global light radiated by all the dipoles represents the scattered light. When the charge displacements are proportional to the excitation (incoming light), i.e., in the low-amplitude limit of scattered radiation, the output light vibrates at the same frequency as that of the excitation. This process corresponds to Rayleigh scattering. On the other hand, if the intensity of the excitation is sufficiently large to induce displacements that are not negligible with respect to atomic distances, then the charge displacements will no longer be proportional to the excitation. In other words, the response of the medium, which is no longer proportional to the excitation, becomes nonlinear. In this case, the scattered waves will be generated not only at the excitation frequency (the Kerr effect), say ω , but also at frequencies that differ from ω (e.g., 2ω , 3ω). On the other hand, it is worth noting that when all induced dipoles vibrate coherently (that is, their relative phase does not vary randomly), their individual radiation may, under certain conditions, interfere constructively and lead to a global field of high intensity. The condition of constructive interference is the phase-matching condition.

In practice, the macroscopic response of a dielectric is given by the polarization, which corresponds to the total amount of dipole moment per unit volume of the dielectric. As the mass of an ion is much larger than that of an electron, the amplitude of the ion motion is generally negligible with respect to that of the electrons. As a consequence, the electron motion generally leads to the dominant contribution in the macroscopic properties of the medium. The behavior of an electron under an optical electric field is similar to that of a particle embedded in an anharmonic potential. A very simple model (called the Lorentz model) that provides a deep insight into the dielectric response consists of an electron of mass m and charge $-e$ connected to an ion by an elastic spring (see Fig. 2). Under the electric field $E(t)$, the electron executes a displacement $x(t)$ with respect to its equilibrium position, which is governed by the following equation:

$$\frac{d^2x}{dt^2} + 2\lambda \frac{dx}{dt} + \omega_0^2 x + (a^{(2)}x^2 + a^{(3)}x^3 + \dots) = -\frac{e}{m} E(t) \quad (4)$$

where $a^{(2)}$, $a^{(3)}$, and so on, are constant parameters, ω_0 is the resonance angular frequency of the electron, and λ is the damping coefficient resulting from the dipolar radiation. When the amplitude of the electric field is sufficiently large, then the restoring force on the electrons becomes a nonlinear function of x ; hence the presence of terms such as $a^{(2)}x^2$, $a^{(3)}x^3$, and so on, in Eq. (4). In this situation, the macroscopic response of the dielectric is the polarization

$$P = -\sum ex(\omega, 2\omega, 3\omega, \dots) \quad (5)$$

where $x(\omega, 2\omega, 3\omega, \dots)$ is the solution of Eq. (4) in the frequency domain, and the summation extends over all the dipole moments per unit volume. In terms of the electric field E the polarization may be written as

$$P = \epsilon_0 (\chi^{(1)} E + \chi^{(2)} E^2 + \chi^{(3)} E^3 + \dots) \quad (6)$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, and so on, represent the susceptibility coefficients. Fig. 3 (top left) illustrates schematically the polarization as a function of the electric field.

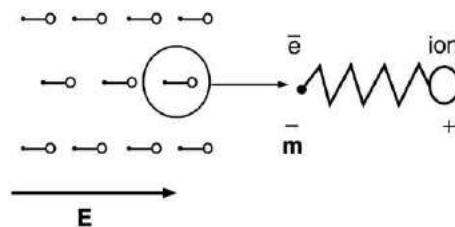


Fig. 2 Electric dipoles in a dielectric medium under an external electric field.

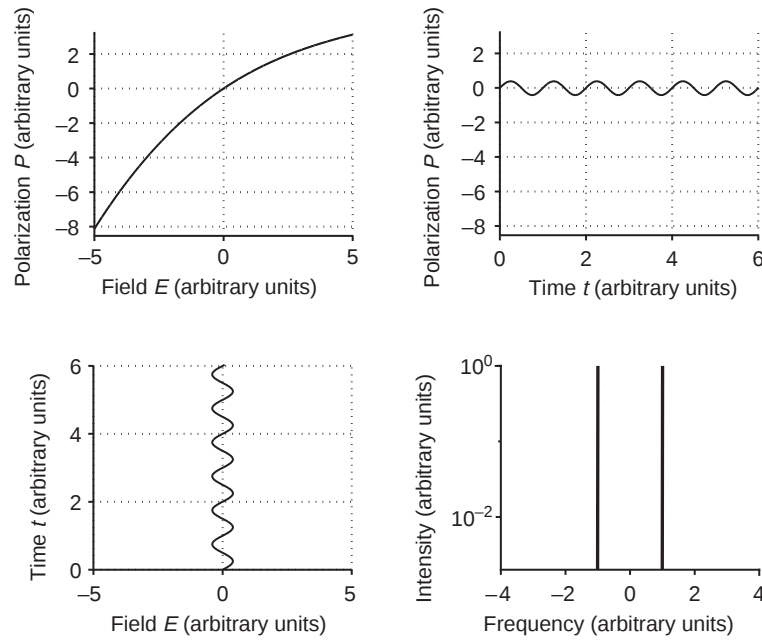


Fig. 3 Polarization induced by an electric field of small amplitude. Nonlinear dependence of the polarization as a function of field amplitude (top left) and time dependence of the input electric field (bottom left). Time dependence of the induced polarization (top right) and corresponding intensity spectrum (bottom right).

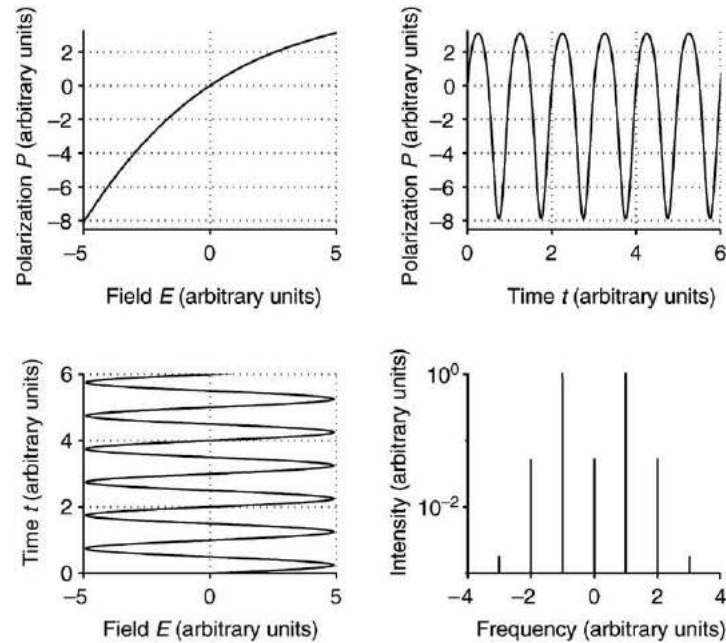


Fig. 4 Polarization induced by an incoming electric field of large amplitude.

In particular, one can clearly observe that when the amplitude of the incoming electric field is sufficiently small (bottom left in Fig. 3), the polarization (top right) is proportional to the electric field (bottom left), thus implying that the electric dipole radiates a wave having the same frequency as that of the incoming light (bottom right). On the other hand, Fig. 4 shows that for an electric field of large amplitude, the polarization is no longer proportional to the electric field, leading to radiation at harmonic frequencies (see Fig. 4, bottom right).

This nonlinear behavior leads to a host of important phenomena in optical fibers, which are useful for many optical systems but detrimental for other systems.

Parametric Phenomena in Optical Fibers

In anisotropic materials, the leading nonlinear term in the polarization, i.e., the $\chi^{(2)}$ term, leads to phenomena such as the harmonic generation or optical rectification. This $\chi^{(2)}$ term vanishes in homogeneous isotropic materials such as cylindrical optical fibers, and there the leading nonlinear term becomes the $\chi^{(3)}$ term. Thus, most of outstanding nonlinear phenomena in optical fibers originate from the third-order nonlinear susceptibility $\chi^{(3)}$. Some of those phenomena are described below.

Optical Kerr Effect

The optical Kerr effect is probably the most important nonlinear effect in optical fibers. This effect induces an intensity dependence of the refractive index, which leads to a vast wealth of fascinating phenomena such as self-phase modulation (SPM), cross-phase modulation (CPM), four-wave mixing (FWM), modulational instability (MI) or optical solitons. The Kerr effect can be conveniently described in the frequency domain through a direct analysis of the polarization, which takes the following form:

$$P_{\text{NL}}(\omega) = \frac{3}{4} \epsilon_0 \chi^{(3)}(\omega) |E(\omega)|^2 E(\omega) \quad (7)$$

The constant 3/4 comes from the symmetry properties of the tensor $\chi^{(3)}$. Setting $P_{\text{NL}}(\omega) = \epsilon_0 \epsilon_{\text{NL}} E(\omega)$, where $\epsilon_{\text{NL}} = \frac{3}{4} \chi^{(3)} |E|^2$ is the nonlinear contribution to the dielectric constant, the total polarization takes the form

$$P(\omega) = P_{\text{L}} + P_{\text{NL}} = \epsilon_0 [\chi^{(1)}(\omega) + \epsilon_{\text{NL}}] E(\omega) \quad (8)$$

As Eq. (8) shows, the refractive index n , at a given frequency ω , is given by

$$n^2 = 1 + \chi^{(1)} + \epsilon_{\text{NL}} = (n_0 + \Delta n_{\text{NL}})^2 \quad (9)$$

with $n_0^2 = 1 + \chi^{(1)}$. In practice $\Delta n_{\text{NL}} \ll n_0$, and then, the refractive index is given by

$$n(\omega, |E|^2) = n_0(\omega) + n_2^e |E|^2 \quad (10)$$

where n_2^e is the nonlinear refractive index defined by $n_2^e = 3\chi^{(3)}/(8n_0)$. The linear polarization P_{L} is responsible for the frequency dependence of the refractive index, whereas the nonlinear polarization P_{NL} causes an intensity dependence of the refractive index, which is referred to as the optical Kerr effect. Knowing that the wave intensity I is given by $I = \alpha |E|^2$, with $\alpha = \frac{1}{2} \epsilon_0 c n_0$, the refractive index can be then rewritten as

$$n(\omega, I) = n_0(\omega) + n_2 I \quad (11)$$

with $n_2 = n_2^e / \alpha = 2n_2^e / (\epsilon_0 c n_0)$. For fused silica fibers one has typically: $n_2 = 2.66 \times 10^{-20} \text{ m}^2 \text{ W}^{-1}$. For example, an intensity of $I = 1 \text{ GW cm}^{-2}$ leads to $\Delta n_{\text{NL}} = 2.66 \times 10^{-7}$, which is much smaller than $n_0 \approx 1.45$.

Four-Wave Mixing

The four-wave mixing (FWM) process is a third-order nonlinear effect in which four waves interact through an energy exchange process. Let us consider two intense waves, $E_1(\omega_1)$ and $E_2(\omega_2)$, with $\omega_2 > \omega_1$, called pump waves, propagating in an optical fiber. Hereafter we consider the simplest case when waves propagate with the same polarization. In this situation the total electric field is given by

$$E_{\text{tot}}(\mathbf{r}, t) = E_1 + E_2 = A_1(\omega_1) \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t)] + A_2(\omega_2) \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t)] \quad (12)$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wavevectors of the fields E_1 and E_2 , respectively. Eq. (7), which gives the nonlinear polarization induced by a single monochromatic wave, remains valid provided that the frequency spacing between the two waves is relatively small, i.e., $|\Delta\omega| = |\omega_2 - \omega_1| \ll \omega_0 = (\omega_1 + \omega_2)/2$. In this context, Eq. (7) leads to

$$P_{\text{NL}} \approx \frac{3}{4} \epsilon_0 \chi^{(3)}(\omega_0) |E_{\text{tot}}|^2 E_{\text{tot}} \quad (13)$$

Substituting Eq. (12) in Eq. (13) yields

$$P_{\text{NL}} = 2n_0 n_2 \epsilon_0 [(|E_1|^2 + 2|E_2|^2)E_1 + (|E_2|^2 + 2|E_1|^2)E_2 + E_1^2 E_2^* + E_1^* E_2^2] \quad (14)$$

The term $|E_1|^2 E_1$ in the right-hand side of Eq. (14) represents the self-induced Kerr effect on the wave ω_1 . The term $|E_2|^2 E_1$ which corresponds to a modification of the refractive index seen by the wave ω_1 , due to the presence of the wave ω_2 , represents the cross-Kerr effect. Thus, the refractive index at frequency ω_1 depends simultaneously on the intensities of the two pumps, i.e., $n = n(\omega_1, |E_1|^2, |E_2|^2)$. Similarly the third and fourth terms in the right-hand side of Eq. (14) illustrate the self-induced and cross-Kerr effects on the wave ω_2 , respectively. Note that the self-induced Kerr effect is responsible for a self-phase modulation, which induces a spectral broadening of a pulse propagating through the optical fiber, whereas the cross-Kerr effect leads to a cross-phase modulation that induces a spectral broadening of one wave in the presence of a second wave. The two last terms in the right-hand side of Eq. (14) correspond to the generation of new waves at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, respectively. The wave with the smallest frequency $2\omega_1 - \omega_2 = \omega_s$ is called the Stokes wave, whereas the wave with the highest frequency $2\omega_2 - \omega_1 = \omega_{\text{as}}$ is called the anti-Stokes wave. These two newly generated waves interact with the two pumps through an energy exchange process. This interaction is commonly referred to as a four-wave mixing process.

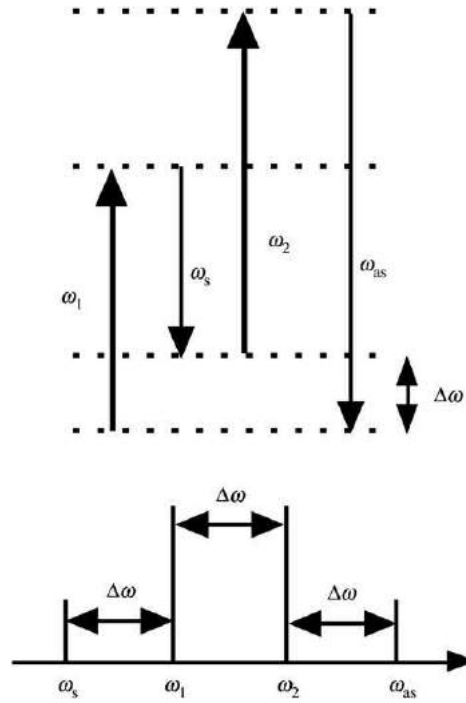


Fig. 5 Schematic diagram representing a non-degenerate four-wave mixing process (top) and the corresponding frequency spectrum (bottom).

From a quantum mechanical point of view, four-wave mixing corresponds to a process in which two photons with frequencies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_s and ω_{as} respectively. The new waves are generated at frequencies ω_s and ω_{as} such that

$$\omega_1 + \omega_2 = \omega_s + \omega_{as} \quad (15)$$

which states that the total energy is conserved during the interaction. But the condition for this FWM process to occur is that the total momentum is conserved, that is,

$$\Delta \mathbf{k} = \mathbf{k}_s + \mathbf{k}_{as} - \mathbf{k}_1 - \mathbf{k}_2 = 0 \quad (16)$$

or equivalently,

$$n(\omega_s)\omega_s + n(\omega_{as})\omega_{as} - n(\omega_1)\omega_1 - n(\omega_2)\omega_2 = 0 \quad (17)$$

Eq. (16) is known as the phase-matching condition. Fig. 5 displays a schematic diagram of a four-wave mixing process with the corresponding frequency spectrum.

FWM with different pump frequencies $\omega_1 \neq \omega_2$ is called 'nondegenerate FWM', whereas the case $\omega_1 = \omega_2$ is referred to as 'degenerate FWM', or more simply, as three-wave mixing.

Conclusion

Optical fibers constitute a key device for many areas of optical sciences, and in particular for ultrafast optical communications. The nonlinear phenomena that arise through the nonlinear refractive index change (induced mainly by the Kerr effect) have been largely investigated these last two decades for applications such as all-optical wavelength conversion, parametric amplification, generation of new optical frequencies or ultrahigh repetition rate pulse trains. However, in many other optical systems such as optical communication systems, these nonlinear phenomena become harmful effects, and in this case control processes are being developed in order to suppress or at least reduce their impact in the system.

See also: Nonlinear Optics. Nonlinear Optics in Photonic Crystal Fibers

Further Reading

Agrawal, G.P., 2001. *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Bloembergen, N., 1977. *Nonlinear Optics*. Reading, MA: Benjamin.

- Boyd, R.W., 1992. *Nonlinear Optics*. San Diego, CA: Academic Press.
- Butcher, P.N., Cotter, D.N., 1990. *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.
- Dianov, E.M., Mamyshev, P.V., Prokhorov, A.M., Serkin, V.N., 1989. *Nonlinear Effects in Optical Fibres*. Switzerland: Harwood Academic Publishers.
- Hellwarth, R., 1977. Third-order optical susceptibilities of liquids and solids. *Progress in Quantum Electronics* 5, 1–68.
- Pocholle, J.P., Papuchon, M., Raffy, J., Desurvire, E., 1990. Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22, 187–268.
- Shen, Y.R., 1984. *Principles of Nonlinear Optics*. New York: Wiley.

Basic Concepts of Optical Amplifiers

MFS Ferreira, University of Aveiro, Aveiro, Portugal

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

g_0	Peak value of the gain coefficient at peak	m^{-1}
$g(\nu)$	Gain coefficient	m^{-1}
G_0	Unsaturated amplifier gain	
$G(\nu)$	Amplifier gain	
L	Amplifier length	m
n_{sp}	Spontaneous emission factor	
P	Signal power	W
P_{in}	Input signal power	W
P_{in}^{s}	Input saturation power	W
P_{out}	Output signal power	W
$P_{\text{out}}^{\text{s}}$	Output saturation power	W
P_{sat}	Saturation power	W
$S_{\text{sp}}(\nu)$	ASE noise spectral density	J
Δf	Detector bandwidth	Hz
$\Delta \nu_0$	Bandwidth of the gain coefficient	Hz
ν	Optical frequency	Hz
ν_0	Atomic transition frequency	Hz

Introduction

The transmission distance of a fiber-optic communication system is limited by fiber loss and dispersion. For long-haul lightwave systems, the loss limitations have traditionally been overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electric signal. Because of the complexity and high cost of such regenerators, the need for optical amplifiers became obvious in the mid-1980s. The optical amplifier is ideally a transparent box that provides gain and is also insensitive to the bit rate, modulation format, power and wavelength of the signal passing through it.

Several means of obtaining optical amplification has been suggested since the 1970s, including direct use of the transmission fiber as gain medium through nonlinear effects, semiconductor amplifiers, or doping optical waveguides with an active material (rare-earth ions), that could provide gain. Due to the spectacular results on erbium-doped fiber amplifiers, which are particularly suitable in the third transmission window (around $1.5 \mu\text{m}$), an intense worldwide research activity on optical amplifiers has developed. As a consequence, the development of erbium-doped fiber amplifiers has reached an industrial level, and commercial devices are now available.

Semiconductor amplifiers, on the other hand, have the same technical basis as semiconductor lasers. Although the strong nonlinearity of semiconductor amplifiers degrades the performances of transmission systems, the state-of-art semiconductor devices seem to be the most interesting amplifiers for transmission in the second transmission window (around $1.3 \mu\text{m}$).

Amplifier Gain and Bandwidth

In a perfect amplifier, the amplification process would be insensitive to the bit rate, modulation format, power, state of polarization, wavelength, and optical bandwidth of the signal passing through it. On the other hand, no interaction would take place if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength (or frequency) of the incident signal, but also on the electromagnetic field intensity at any point inside the amplifier. Details of wavelength and intensity dependence of the optical signal depend on the amplifying medium.

We consider a case in which the gain medium is modeled as a homogeneously broadened two-level system. In such medium, the gain coefficient (i.e., the gain per unit length) can be written as:

$$g(\nu, P) = \frac{g_0}{1 + \frac{(\nu - \nu_0)^2}{\Delta \nu_0^2} + \frac{P}{P_{\text{sat}}}} \quad (1)$$

where g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, ν the optical frequency, ν_0 the atomic transition frequency, $\Delta \nu_0$ the 3 dB local gain bandwidth, P the optical power of the signal, and P_{sat} the saturation power,

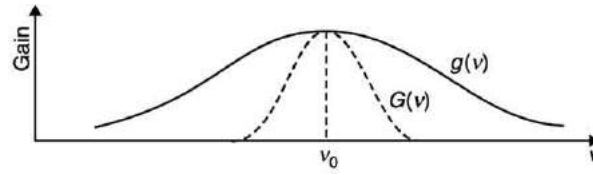


Fig. 1 Gain coefficient profile $g(v)$ and the corresponding amplifier gain spectrum $G(v)$.

which depends on the gain medium parameters. It must be emphasized that Δv_0 and P_{sat} refer to the local gain. However, from the communication system point of view, it is more desirable to use the related concepts of amplifier bandwidth and amplifier saturation power that will be evaluated below.

The amplifier gain G is defined as

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} \quad (2)$$

where P_{in} is the input power and P_{out} the output power of a continuous wave (CW) signal being amplified. The amplifier gain G may be found by using the relation:

$$\frac{dP}{dz} = g(v, P)P \quad (3)$$

where $P(z)$ is the optical power at a distance z from the amplifier input end.

If the signal power obeys the condition $P \ll P_{\text{sat}}$ throughout the amplifier, the gain coefficient given by Eq. (1) can be considered independent of the signal power. In such a case, the amplifier is said to be operated in the unsaturated regime and works as a linear device. The gain coefficient presents in this situation a Lorentzian profile that is characteristic of homogeneously broadened two-level systems. However, the gain spectrum of actual amplifiers can deviate significantly from the Lorentzian profile.

The solution of Eq. (3) in the unsaturated regime is an exponentially growing signal power, given by

$$P(z) = P_{\text{in}} \exp(gz) \quad (4)$$

For an amplifier length L , we then find that the linear amplifier gain is

$$G(v) = \exp(gL) = \exp \left[\frac{g_0 L}{1 + (v - v_0)^2 / \Delta v_0^2} \right] \quad (5)$$

Both the amplifier gain $G(v)$ and the gain coefficient $g(v)$ are maximum when $v = v_0$. However, $G(v)$ decreases much faster than $g(v)$ with the signal detuning $v - v_0$, because of the exponential dependence of G on g . As a consequence, the amplifier bandwidth Δv_A , which is defined as the FWHM of $G(v)$, is much smaller than the gain bandwidth Δv_0 (Fig. 1). This can result in signal distortion in the case where a broadband optical signal is transmitted through the amplifier. From Eq. (5) we can obtain the following relation between the amplifier bandwidth and the gain bandwidth:

$$\Delta v_A = \Delta v_0 \sqrt{\frac{\ln 2}{g_0 L - \ln 2}} \quad (6)$$

Gain Saturation

An important limitation of the nonideal amplifier is related with the power dependence of the gain coefficient given by Eq. (1). This property is known as gain saturation and it appears when the signal power ratio P/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P becomes comparable to the saturation power P_{sat} , the amplifier gain G will also decrease.

Assuming that $v = v_0$ and substituting g from Eq. (1) in Eq. (3) gives

$$\frac{dP}{dz} = \frac{g_0 P}{1 + P/P_{\text{sat}}} \quad (7)$$

Considering the initial condition $P(0) = P_{\text{in}}$, we obtain from Eqs. (2) and (7) the following implicit relation for the amplifier gain:

$$(1 - G) \frac{P_{\text{in}}}{P_{\text{sat}}} = \ln \left(\frac{G}{G_0} \right) \quad (8)$$

where $G_0 = \exp(g_0 L)$. The input saturation power P_{in}^s is defined as the input power for which the amplifier gain G is reduced by a factor of 2 from its unsaturated value G_0 (Fig. 2). Indeed, it is obtained by using $G = G_0/2$ in Eq. (8):

$$P_{\text{in}}^s = \frac{2 \ln(2) P_{\text{sat}}}{(G_0 - 2)} \quad (9)$$

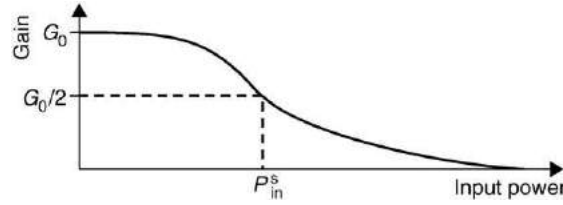


Fig. 2 Saturated amplifier gain as a function of the input power.

As observed from [Eq. \(9\)](#), the input saturation power P_{in}^s does not coincide with P_{sat} . The output saturation power is given by $P_{out}^s = G_0 P_{in}^s / 2$. In practice, $G_0 \gg 2$ and P_{out}^s is found to be smaller than P_{sat} by about 30%.

Gain saturation can be seen as a serious limitation, particularly for multichannel communication systems. However, the self-regulating effect of gain saturated amplifiers can be useful in long-haul lightwave communication systems, including many concatenated amplifiers. In fact, if the signal level in a chain of amplifiers is unexpectedly increased along the chain, the saturation effect causes a lower gain provided by the following amplifiers and vice versa for a sudden signal power decrease.

Amplifier Noise

Besides the bandwidth and gain saturation limitations, another property must be considered concerning practical optical amplifiers. In fact, optical amplifiers always add spontaneously emitted photons to the signal during the amplification process. Those photons are amplified besides the signal photons so that, at the amplifier output, an amplified spontaneous emission (ASE) noise is presented. Since spontaneous emission always takes place, ASE noise is unavoidable and does not depend on the amplifier temperature. This is one of the most important differences between optical and electrical amplifiers, where amplifier noise is of thermal origin and can be reduced by lowering the amplifier temperature.

The ASE determines a degradation of the signal-to-noise ratio (SNR). The SNR degradation is usually characterized by the amplifier noise figure, which is defined as the SNR ratio between input and output:

$$NF = \frac{SNR_{in}}{SNR_{out}} \quad (10)$$

The SNR is usually referred to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. Therefore, the noise figure as defined in [Eq. \(10\)](#) would usually depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. We will consider the case of an ideal detector, whose performance is limited by shot noise only.

The SNR of the input signal is simply determined by the detection shot noise and can be written as:

$$SNR_{in} = \frac{P_{in}}{2h\nu\Delta f} \quad (11)$$

where Δf is the detector bandwidth.

To evaluate the term SNR_{out} , one should add the contribution of spontaneous emission to the receiver noise. The ASE noise spectral density is assumed to be constant and can be written as

$$S_{sp}(\nu) = (G - 1)n_{sp}h\nu \quad (12)$$

where G the amplifier gain and

$$n_{sp} = \frac{N_1}{N_1 - N_0} \quad (13)$$

is known as the spontaneous emission factor or the population inversion factor. In [Eq. \(13\)](#) N_0 and N_1 are the atomic populations for the ground and excited states, respectively.

Considering a low noise amplifier, the signal power impinging the photodetector is larger than the optical noise power and the shot noise power. As a consequence, the electrical noise, due to the signal-ASE beating, is the dominant contribution and the SNR of the amplified signal is given by:

$$SNR_{out} \approx \frac{GP_{in}}{4S_{sp}\Delta f} \quad (14)$$

Using [Eqs. \(11\)–\(14\)](#), the amplifier noise figure defined by [Eq. \(10\)](#) becomes:

$$NF = 2n_{sp} \frac{G - 1}{G} \approx 2n_{sp} \quad (15)$$

where the approximation holds when the gain is much higher than one. In the case of an ideal amplifier, $n_{sp} = 1$ and [Eq. \(15\)](#) show that the SNR is degraded by 3 dB. For most practical amplifiers, NF can be as large as 6–8 dB.

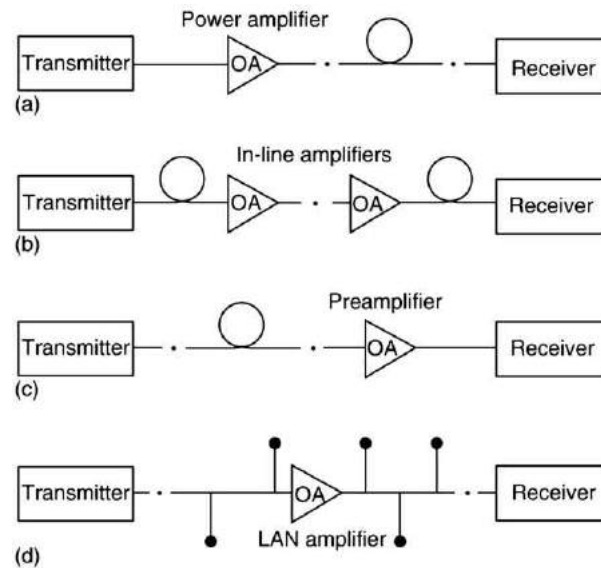


Fig. 3 Four generic configurations for incorporating optical amplifiers into transmission systems: (a) as a power amplifier; (b) as in-line amplifiers; (c) as a preamplifier; (d) for compensation of distribution losses in local-area networks.

Basic Amplifier Configurations

The relative importance of the different limiting factors discussed above depends on the actual amplifier application. **Fig. 3** shows the four basic system configurations envisioned for the incorporation of optical amplifiers. The first configuration is to place the amplifier immediately following the laser transmitter to act as a power amplifier or booster (**Fig. 3(a)**). The main purpose of such amplifiers is to boost the signal power, which can provide an increase of the transmission distance by 100 km or more. Since the signal input power is typically large (0.1–1.0 mW), the key parameter for the power amplifier will be to maximize the saturation output power and not necessarily the absolute gain.

The second configuration is to place the amplifier in-line and perhaps incorporated at one or more places along the transmission path (**Fig. 3(b)**), replacing the electronic regenerators. The in-line amplifier corrects for periodic signal attenuation and may exist in a cascade form. The use of in-line optical amplifiers is particularly attractive for multichannel communication systems, since they can amplify all channels simultaneously.

The third configuration consists of using the amplifier immediately before the receiver, so it functions as a preamplifier (**Fig. 3(c)**). The purpose of such an amplifier is to improve the receiver sensitivity. The main figures of merit are high gain and low amplifier noise, because the entire amplifier output is immediately detected.

In local-area networks (LANs), distribution losses often limit the number of possible nodes. The fourth application of optical amplifiers consists of using them for compensating such distribution losses (**Fig. 3(d)**).

See also: Optical Amplifiers: SOAs

Further Reading

- Agrawal, G.P., 1992. *Fiber-Optic Communication Systems*. New York: Wiley.
- Bjarklev, A., 1993. *Optical Fiber Amplifiers: Design and System Applications*. Boston, MA: Artech House.
- Desurvire, E., 1994. *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.
- Green Jr, P.E., 1993. *Fiber Optic Networks*. Englewood Cliffs, NJ: Prentice Hall.
- Iannone, E., Matera, F., Mecozzi, A., Settembre, M., 1998. *Nonlinear Optical Communication Networks*. New York: Wiley.
- Kazovsky, L., Benedetto, S., Willner, A., 1996. *Optical Fiber Communication Systems*. Boston: Artech House.
- Shimada, S., Ishio, H. (Eds.), 1994. *Optical Amplifiers and their Applications*. New York: Wiley.

Erbium Doped Fiber Amplifiers for Lightwave Systems

P Bollond, JDS Uniphase Corporation, Ewing, NJ, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The telecommunications industry has undergone a revolution since the 1980s, by using glass optical fibers for the transmission of information encoded as pulses of light. A single telecommunications-grade optical fiber has been shown to support the propagation of more than 1 Tbit per second (1×10^{12} pulses per second) over distances comparable to typical city separations (> 100 km). From this technology, optical fiber links have evolved to become a network on a planet-wide scale, and form the physical backbone of the information age.

Erbium-doped fiber amplifiers (EDFAs) are an enabling technology for optical fiber communication networks. They have several important properties that make them the amplification component of choice in long-distance commercial data transport networks. Erbium ions deposited in silica-based glass allow amplification in the lowest loss region of commercial-grade optical fiber (< 0.25 dB/km from approximately 1530–1620 nm). Erbium-doped fiber (EDF) is manufactured in a form that allows for low-loss fusion splicing to standard communications fiber. Compact semiconductor laser diodes are available to excite the erbium ions into an amplification state. The long lifetime of the excited state of erbium provides the ability to simultaneously amplify multiple wavelength channels without significant cross-channel interference. Multiple channel systems have been deployed with more than 100 optical channels (or wavelengths) through each EDFA, and this has allowed network capacities to be dramatically increased. The low noise properties of the EDFA also allow networks to be constructed with many amplified spans before the optical signal has to be electronically regenerated. Practically unlimited transmission distance has been demonstrated using a small number of optical soliton channels through periodically amplified EDFA lightwave systems.

The development history of the EDFA can be traced back to the first optical amplifier. In 1962, a neodymium-based fiber amplifier was invented that operated at 1064 nm. During the 1980s, the need for an optical amplifier at telecommunications wavelengths initiated research at many locations throughout the world. In 1987, the University of Southampton (UK) was first to demonstrate an EDFA that had optical gain at 1550 nm, and during the following years the design of erbium-doped fiber was optimized for this application. In 1989, a practical semiconductor laser diode became available to pump EDF, and the first compact optical fiber amplifier modules soon appeared for commercial deployment. The traditional method of signal regeneration, prior to 1990, was to use electronics to detect the optical signal after each transmission span, recover the digital signal, and then retransmit using another laser diode. The EDFA allows practical wavelength division multiplexing (WDM) of multiple optical signals with all optical signal amplification, and provides significant performance and cost advantages over electronic regeneration.

EDFAs have emerged from the laboratory to be widely deployed in communication networks. EDFAs are used to boost transmitted power of the signal lasers (booster amplifier), amplify signals in transit to compensate losses sustained in the fiber (line amplifier), or amplify signals before a receiver (pre-amplifier). Typically, the amplifier module is specifically manufactured for particular systems that are mounted on electronic circuit boards. These circuit packs are then housed in central offices (local telephone exchanges), remote 'repeater huts', or even in undersea 'bottles' as part of a transoceanic cable. The high cost of network failure requires that the manufactured EDFA modules comply with stringent reliability criteria, to provide a useful operating lifetime greater than 25 years when subject to extreme environmental conditions.

Amplifiers are constructed for incorporation into either existing fiber links as part of an upgrade, or for newly planned systems. Because of the high cost of installing new fiber into the ground and securing property rights, it has become economically desirable to upgrade many existing fiber links rather than to build new systems. Transmission cables usually have many pairs of individual optical fibers, some of which will not initially be transporting data, and these 'un-lit' or 'dark fibers' can be activated as consumer demand increases over time. Typically, each fiber of a pair is used to carry either 'east'- or 'west'-bound traffic. Around city areas metropolitan area networks can be expanded in this way, but for long-distance links (long haul networks with distances > 1000 km) operation is designed for a larger number of channels (40 to 120 wavelengths) at higher data rates (10 or 40 GHz per channel) over specialized transmission cables containing low numbers of fiber pairs. Typically communication systems are designed to meet certain cost targets, expressed as dollars per Gbit/s per km, for total transmission distances. The total transmission distance is limited by the optical signal to noise ratio (OSNR) degradation after each fiber-amplifier link, with a smaller permissible degradation at higher bit rates. The required gain and OSNR performance for the system is then translated to an EDFA module specification.

This article discusses EDFA design and applications, and shows the elements involved in producing reliable modules for commercial lightwave systems.

Components for EDFAs

EDFAs are comprised of passive optical components, erbium-doped fiber, and pump lasers. Passive components are chosen to meet optical and environmental specifications, while the erbium-doped fiber is selected based on the optical power, gain, and noise figure requirements. Pump lasers are a key influence on the price and performance of optical amplifiers.

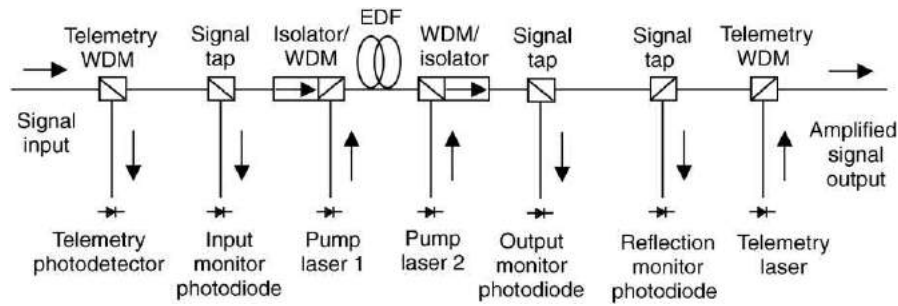


Fig. 1 Single-stage EDFA with features.

The amplifier module is typically connected to the transmission fiber using fiber connectors. These polished fiber connectors have a higher insertion loss and reflectivity than fusion splicing, but allow for easy deployment in the network. Internal optical components are fusion spliced together to provide low loss, low back reflection, high strength, and high reliability joins. Fusion splicing is tailored to particular fiber types, so that optical components with dissimilar fiber types are joined with the lowest loss.

The signal and pump radiation is combined with low loss using optical components called wavelength division multiplexers (see Fig. 1). These components are based on fused fiber or interference filter based technology. Fused fiber WDMs offer the lowest insertion loss (<0.1 dB is commercially available) but is restricted to widely spaced wavelengths (e.g., 980 nm pump and 1550 nm signal). Interference filter based WDMs are available for closely spaced wavelengths (e.g., 1480 nm pump and 1530 nm signal) and have a very low wavelength dependent insertion loss (flatness). Interference filters can be designed to produce sharp low-pass, high-pass, or bandpass type filters suitable for combining closely spaced wavelengths, as well as broader peaks suitable for lowering the gain in particular signal wavelength regions to produce gain flattened amplifiers.

Signal reflections can cause an amplifier to act as a laser, and this detrimental effect is eliminated in EDFAs by using optical isolators. In an isolator the signal light is coupled out of the single-mode fiber through a graded index (GRIN) lens and passes through a nonlinear crystal before being coupled back into the optical fiber. The isolator consists of a birefringent rutile (TiO_2) or Yttrium OrthoVanadate (YVO_4) wedge, followed by a Yttrium Iron Garnet (YIG) Faraday rotator, followed by another birefringent wedge. The YIG crystal is surrounded by a permanent magnet that rotates the light's polarization by 45 degrees. The 45-degree polarization rotation, coupled with the two birefringent wedges, ensures that light is efficiently coupled to the output fiber but not in the reverse direction. Commercial isolators are available with low insertion losses across the signal band, with some samples below 0.35 dB. Note that the YIG crystal works well for the 1480 nm pump and 1530–1620 nm signal bands, but currently there is no suitable isolator material that covers 980 to 1550 nm and this puts some limitations on certain EDFA designs.

The isolator design has been extended to make multiport circulators. A three-port circulator has the input into port 1 and output of port 2, light entering port 2 is directed to port 3, and light entering port 3 is blocked with an isolator. The circulator allows for adding and dropping of individual channels when a narrow bandwidth reflective filter is placed on port 2. Circulators also can be used to separate co-propagating and counter-propagating traffic on a single transmission fiber before amplification is done.

Erbium-doped fiber has two strong absorption bands around 980 nm and 1480 nm that are suitable for achieving a population inversion in the erbium ion. The 980 nm wavelength allows low noise amplification, while 1480 nm lasers provide higher EDFA output. 980 nm pump lasers usually incorporate fiber Bragg gratings to stabilize the laser wavelength to match the narrow EDF absorption peak, and also have lower drive current requirements. A 1480 nm pump can provide more amplification, since there are more photons in each mW of laser power at 1480 nm than 980 nm. High-output power pump lasers incorporate thermoelectric coolers (TECs) within the pump laser package and can provide fiber coupled power greater than 500 mW. Both high-power lasers have been qualified to meet the most stringent reliability standards, with typical mean time before failure beyond 25 years. Low-cost 980 nm pump lasers, that do not have TECs, are also available in smaller packages and can supply up to ~ 200 mW. By using wavelength division or polarization combining techniques, it is possible to further increase the available pump power in the erbium fiber.

The EDFA module is assembled into a package that may contain passive optical components, several meters of coiled EDF, pump lasers, photo-detectors, and electronic circuit boards. In some cases, it is advantageous to thermally stabilize the components so that the amplifier performance can be maintained when exposed to extreme environmental conditions. This may occur when the central office's environmental control is compromised (e.g., air conditioning failure). In particular, EDF can exhibit undesirable spectral gain changes if the ambient temperature changes by more than 5°C .

Pump lasers with internal TECs can dissipate more than 5 Watts of heat per laser and since cooling fans usually do not have the required reliability, passive cooling is commonly used in the module in the form of a metal heat sink. The amplifier module's size can be compact, limited by the height of optical components or pump laser diodes, or by the size of a built-in heat sink. The module is designed to survive conditions of electrostatic discharge, humidity, temperature, thermal shock, extreme vibration, and other stresses that may be inadvertently present during operation in the field. In addition, all material in the EDFA module is also qualified against problems with out-gassing (e.g., hydrogen release), combustion and chemical or biological exposure.

The Single-Stage Amplifier

A simple single-stage EDFA consists of an erbium-doped fiber spool with signal and pump combining multiplexer. The fiber is optically pumped by 980 nm and/or 1480 nm laser(s), whose light is coupled into the signal fiber by a passive component called an optical multiplexer. The pump wavelengths are readily absorbed by erbium ions embedded in silica raising them to an excited state. Amplification occurs when, stimulated by a nearby signal photon, an excited erbium ion relaxes back to the ground state producing a second, identical signal photon. The erbium ion can be approximated as a three-level atomic system that can be completely inverted by a 980 nm photon, to provide the lowest noise amplification. In contrast, the 1480 nm pump will excite the erbium ion directly into the upper laser level as a two-level system, and because of rapid spontaneous emission from this level, the maximum inversion in this case cannot exceed approximately 75%. Note that as the pump photons are absorbed, the inversion will be nonuniform along the EDF length.

There are two signal wavelength regions commonly amplified by EDFAs, the C-band (conventional band) from approximately 1528 to 1565 nm, and the L-band (long band) from approximately 1570 to 1620 nm. Amplification in the C-band readily occurs when moderate pump power is available, and relies on the erbium ion's spectral absorption and emission wavelength window that is suited to high levels of atomic inversion. L-band amplification is also achieved with moderate pump powers, but because of the lower absorption and emission cross-sections, similar gain is reached using approximately five times more EDF with a lower average inversion. The C-band amplifier is typically less costly because less EDF is used, while high-concentration erbium fibers are available specifically for L-band EDFAs.

An EDFA's most critical performance parameters are its amplified signal output power (typically stated in dBm) and its noise figure (stated in dB). Output power is mainly determined by total pump power and the amplifier internal loss. The noise figure (NF) is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output.

A single-stage amplifier typically has 1 or 2 pump lasers but can have more if polarization- or wavelength-pump-combining is implemented for higher power. When the pump radiation propagates in the same direction as the signal, the amplifier is co-pumped, while counter-pumping denotes the case when the pump laser propagates against the signal. For a single pump, a co-pumping 980 nm laser minimizes the NF (suitable for a pre-amplifier) while counter-pumped 1480 nm architecture optimizes output power at some expense to the NF (suitable for a booster amplifier). Recent semiconductor pump laser improvements have enabled 980 nm pump lasers with output powers >500 mW to be commercially available, allowing most single- or dual-stage EDFAs to be energized by one laser.

Single-stage designs can be enhanced, as shown in Fig. 1. To control the network an optical signal may be used as a telemetry or network supervisory channel, and this is removed with a filter. Telemetry wavelengths are usually outside of the useful EDF amplification window, and commonly range from 1500–1520 nm and from 1620–1640 nm. An isolator may be used at the input and/or output to prevent pump laser or amplified spontaneous emission from the erbium-doped fiber 'leaking' into the transmission path. Optical taps may be included to provide information about signal spectra at the input and output sides. Their feedback can be used to control pump laser biases for tuning output power, monitoring amplifier performance, or simply to trigger alarms. In addition, a reflection monitor is sometimes placed at the output to observe backwards propagating optical signals arising from reflections. Electronics in the module will continuously monitor the pump lasers and photodetectors and, for example, can place the module in an 'eye-safe' low-output power (<10 mW) mode within milliseconds if a transmission fiber break is detected through increased back-reflected optical power.

The amplifier has a signal input power of -30 to -10 dBm (1–100 μ W) for each optical channel, and has a gain of 25 dB to compensate for a typical span loss of 100 km optical fiber link. This signal level allows high powers at the receiver photodetector for high-quality signal detection, yet is low enough to avoid nonlinear propagation effects in the transmission fiber. The typical total output signal power of an 80 channel (wavelength) C-band EDFA is less than 200 mW (+23 dBm) to avoid stimulated Raman scattering (SRS) in standard transmission fiber. The use of improved low nonlinearity transmission fibers and longer transmission distances can lead to specified total EDFA output powers to be greater than 400 mW (+26 dBm). To provide the best performance, the amplifier usually will allow only a single direction of propagation, with bi-directional communications systems using one transmission fiber and circulators to separate 'east' and 'west' traffic into individual EDFAs.

Fig. 2 shows the results of a numerical simulation for a single-stage amplifier as a function of the input power and EDF length. The amplifier was assumed to have EDF with peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 100 mW at 980 nm. The amplifiers signal loss before and after the EDF was taken to be 0.8 and 1.2 dB, respectively. As amplifier input power increases there is a decrease in gain provided by the medium since there is a fixed amount of pump power available. The maximum gain is usually achieved near 1530 nm where the difference between the EDF's emission and absorption cross-sections is greatest. Selecting the length of EDF is critical to achieving the desired performance, and this can be examined using numerical simulation for a wide range of design options.

Undersea systems, with their long distances and large costs of network failure, place stringent design requirements on EDFAs. These systems mostly use single-stage designs with emphasis on low noise operation. This can be achieved by eliminating most of the optical components prior to the EDF, and also by co-propagating a strong 980 nm pump using a low loss fused fiber WDM. Undersea repeaters are spaced by 30 to 80 km, shorter than terrestrial networks, with each channel operated at higher power to achieve multi-thousand kilometer distances. For example, the trans-Pacific TPC-5J cable spans 8600 km from Coos Bay, Oregon (USA) to Ninomiya (Japan), using EDFAs spaced every 33 km. The tight electrical power budgets available to each repeater (powered from land) necessitate using pump lasers without TECs. During installation, the EDFA will experience large mechanical

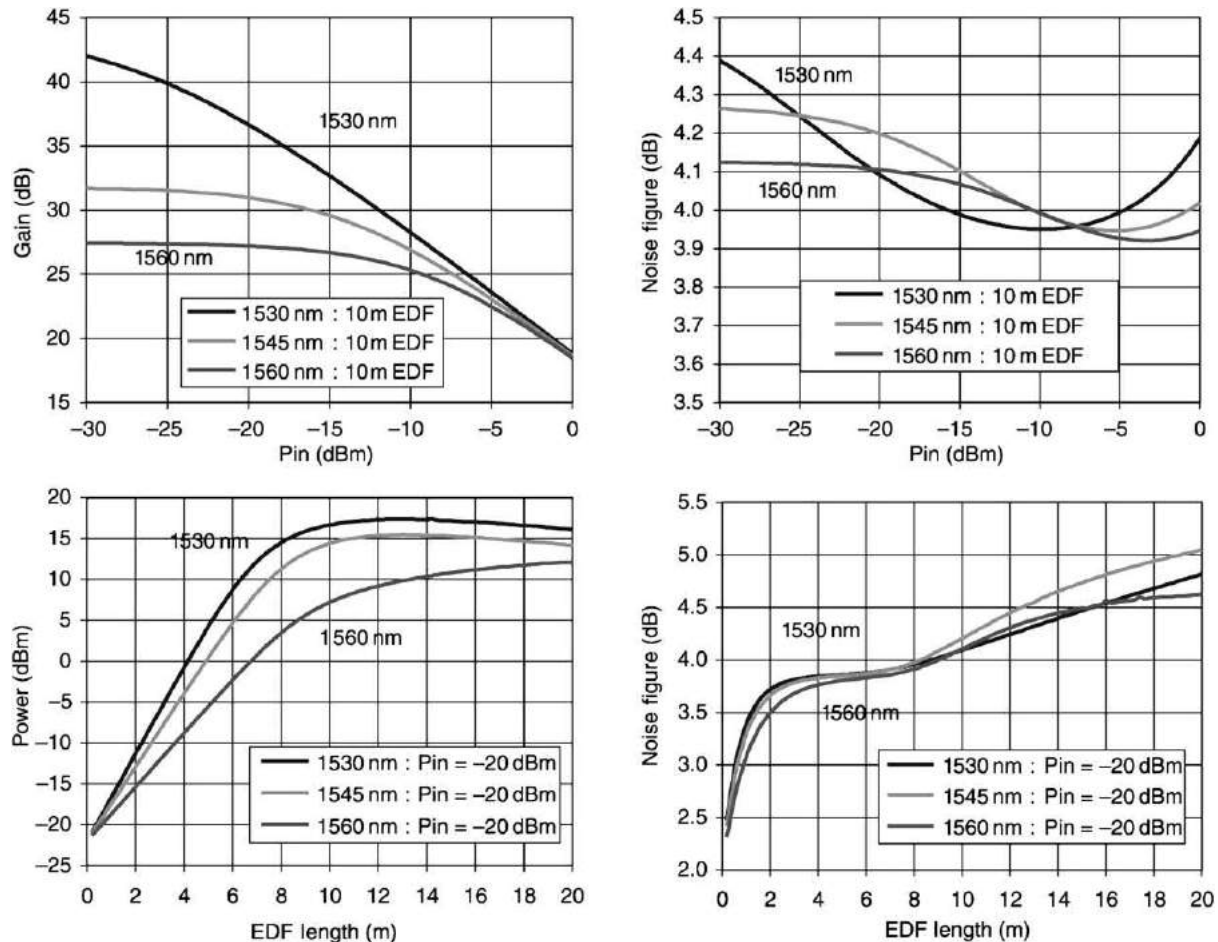


Fig. 2 Typical single-channel performance of a single-stage EDFA.

shock, as the cable and metal repeater bottles are unwound aboard ship and dropped into the ocean. Other design considerations are done for operation at the constant ambient temperature of the ocean bottom ($\sim +2$ to 4°C) or for seasonal temperature changes on the continental shelves, where the optics and EDF will operate at 15 to 30 degrees above ambient due to heating from the electronics.

Single-stage optical amplifiers are suited for a wide range of applications such as single-channel amplifiers, simple WDM amplifiers, and low-cost amplifiers. Using high-power pump lasers or combined pump laser schemes allows such amplifiers to deliver output powers >20 dBm. However, single-stage amplifiers cannot meet the requirements of all telecommunications architectures, leading to increasing demand for multiple-stage EDFAs that are discussed below.

Multiple-Stage EDFAs

The most common implementation of a multiple-stage EDFA is to improve the noise and output power characteristics by imbedding an optical isolator between two sections of EDF. The isolator blocks backward-traveling amplified spontaneous emission to improve the efficiency of the amplifier. This is shown in [Fig. 3](#) for two common single-pump designs. The first is the 'pump by-pass' design where the residual pump radiation from the first stage is redirected around the mid-stage isolator. This is often used when pumping with 980 nm since signal band optical isolators will not transmit 980 nm radiation. This design can introduce problems when using low-quality components as small signal levels can propagate around the pump by-pass fiber to create multiple path interference (MPI) effects. Although MPI effects can be eliminated by using a pump splitting coupler to directly pump each EDF section, the pump by-pass design makes the most efficient use of available pump power. The second design is called the 'pump through' design, and is used when pumping at 1480 nm since both the pump and signal band (e.g., 1550 nm) photons will transmit through commercial isolators with only small loss. Note that for cost and space constraints within the module, it is sometimes advantageous to use hybrid optical components, e.g., an isolator and WDM can be combined into one compact package.

An EDFA with multiple input signals will have a very nonuniform output gain profile. This is a consequence of the erbium ion's wavelength-dependent emission and absorption cross-sections in the host glass material. [Fig. 4](#) shows gain spectra for both C-band

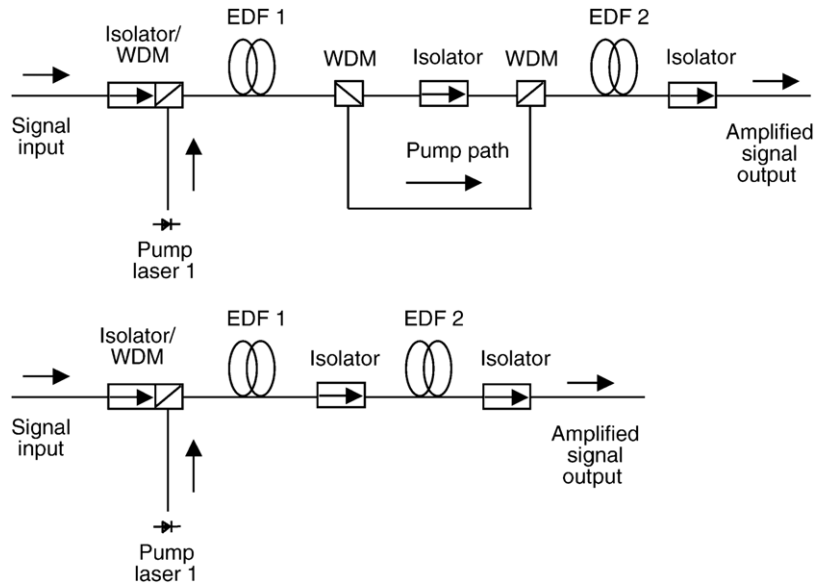


Fig. 3 Imbedded Isolator EDFA designs.

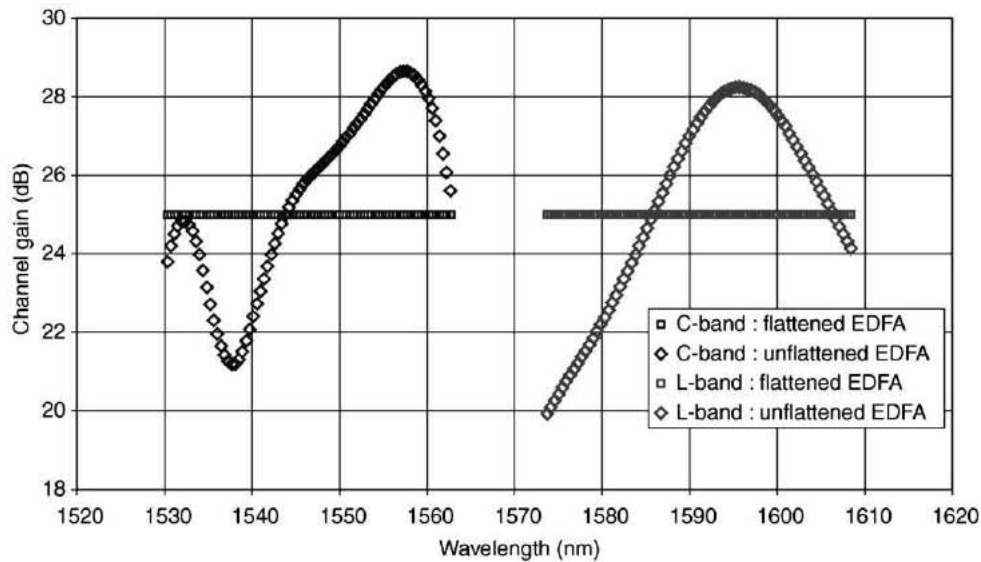


Fig. 4 Gain spectrum of a multi-staged optical amplifier with and without gain flattening filters.

and L-band EDFAs with and without gain-flattening filters (GFFs). The gain spectrum is dependent on the EDF's chemical composition and EDFA design features. For an amplifier with 25 dB gain, a GFF with peak loss less than 10 dB is usually needed to correct for these gain deviations. Although many technologies are available for GFFs (e.g., thin-film interference filters, Bragg gratings, tapered fiber filters, etc.) the basis function of these technologies is usually not identical to the EDF's gain profile, and this mismatch results in gain flatness error. However, when using these technologies, manufactured EDFAs can have a gain error (flatness) less than 1.0 dB over bandwidths greater than 35 nm.

Early EDFAs for WDM systems used 16 optical channels spaced by 200 GHz in the 1540 to 1560 nm part of the C-band. In this case, the EDF gain profile is relatively smooth and no GFFs are needed to achieve a 1.0 dB gain flatness specification. It is important to note that for a particular level of signal and pump power, a longer length of EDF in the EDFA will produce more gain at the long wavelength end of the spectrum, hence by shortening the EDF length in Fig. 4 the result will be relatively flat gain from approximately 1540 to 1560 nm. As EDFA technology has matured, the EDF can be the bandwidth-limiting component in a system, and it has become necessary to use GFFs to realize useable bandwidths of up to 55 nm when using regular silica-based EDF. The most economic EDFAs operate in the C-band where relative to the L-band, pump laser efficiency is highest, EDF lengths are shortest, and the chromatic dispersion of most installed transmission fiber limits nonlinear optical effects.

In single-stage amplifiers, the gain-flattening filter is at the amplifier output, and this results in lost output power. Wideband gain-flattened amplifiers usually have a multiple EDF stage design, with the GFF component between two EDF stages. The EDF sections that follow the attenuating GFF provide additional amplification, to give high power out of the EDFA module that has large internal losses.

In addition to GFF components inside the EDFA, an attenuator can be used to compensate for input signal power changes, which will produce a spectral gain rotation or tilt. The attenuator reduces the power on all optical channels equally and allows the amplifier to operate in a 'fixed gain' mode. In WDM systems, optical amplifiers may also need mid-stage access where the signal is fed into an external device between the amplifier stages. Reasons for doing this include monitoring, adding or dropping of individual channels, and dispersion compensation. Since additional losses up to 10 dB are introduced in the amplification path, the amplifier design has to be optimized for those losses.

Typical two-stage gain-flattened optical amplifier architecture is shown in Fig. 5. Input and output couplers and telemetry WDMs can be included if required. A single pump can be split and shared between stages to save cost. When multiple pumps are needed, the most common configuration is a 980 nm pump laser co-pumping the first stage (EDF1) for low noise and a 1480 nm laser counter-pumping the second spool (EDF2) for high gain. A significant loss element, e.g., a gain-flattening filter, add/drop module or dispersion compensation module, is situated between the stages. Note that mid-stage access can be located before or after the gain flattening or between additional EDF stages. The gain spectrum of multistage gain-flattened EDFAs is shown in Fig. 6, showing the gain equalization possible using a multistage amplifier design and gain-flattening filters. Using deeper GFFs can increase the usable bandwidth of the amplifiers, but this requires higher-power pump lasers to maintain the same gain and optical signal-to-noise ratio (OSNR) performance.

The two-stage EDFA shown in Fig. 5 highlights a common problem for amplifier design. Given high-grade optical components and pump lasers, what length of erbium-doped fiber should be used to give the best amplifier performance? This question is best answered by using the results of extensive numerical simulations of the optical amplifier.

Fig. 7 shows the results of numerical simulation for a two-stage amplifier as a function of the first and second EDF stage lengths. The amplifier was assumed to have EDF with a peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 130 mW at 980 nm in the first stage and 160 mW at 1480 nm in the second stage. The total input power was assumed to be -2.5 dBm for 80 channels distributed from 1530 nm to 1563 nm. Note that each of the 80 channels is separated by 100 GHz, and will generally support a long-distance communications system with each wavelength modulated at 10 GHz.

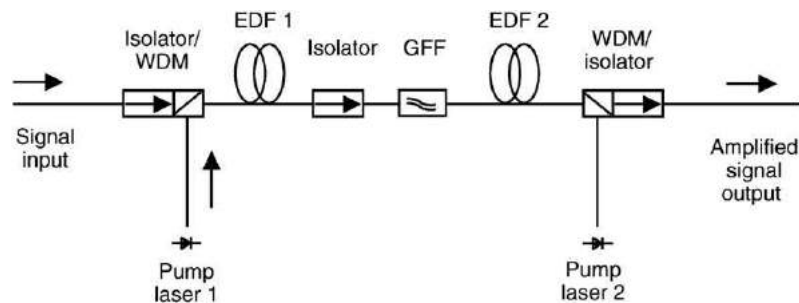


Fig. 5 Typical multistage gain-flattened optical amplifier architecture.

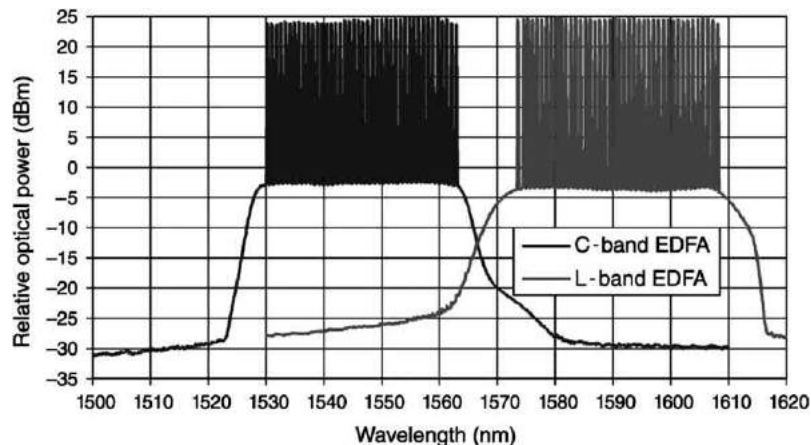


Fig. 6 Measured gain and noise spectra of multistaged gain-flattened C-band and L-band EDFAs.

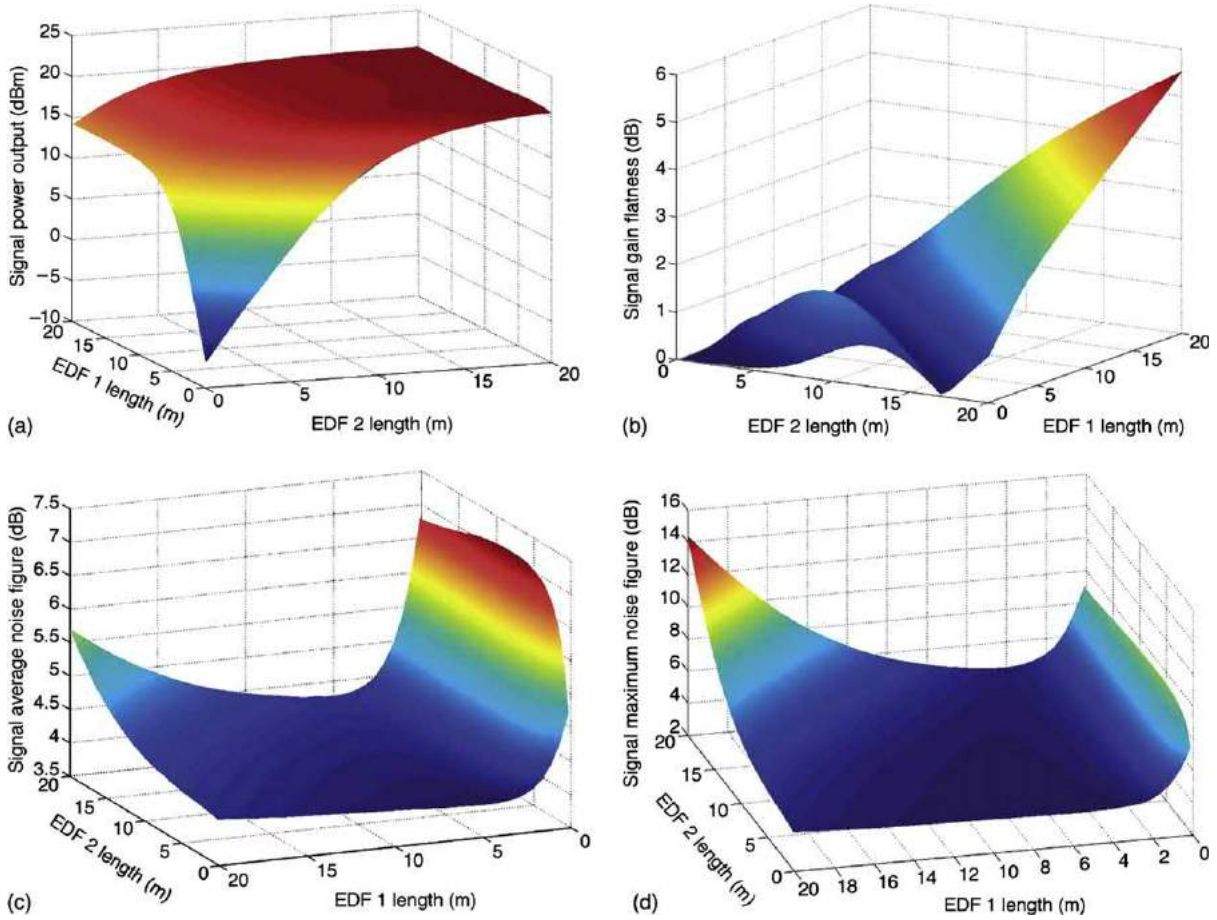


Fig. 7 (a) Total output power, (b) spectral flatness, (c) average noise figure, and (d) maximum channel noise figure of a two-stage optical amplifier with a gain-flattening filter.

Fig. 7(a) shows that the optical output power is largest when both EDF stages are approximately 20 meters. For this design, EDF lengths significantly longer than 20 meters will produce less output power, since the pump radiation will have been completely absorbed in the first few meters of EDF. The spectral gain flatness is shown in **Fig. 7(b)**. For this design, approximately 17 meters of EDF was needed to produce a flat gain spectrum, i.e., each optical channel had the same gain. Other combinations of EDF lengths were not matched to the particular gain-flattening filter and produced large gain variations across the band. In general as the amplifier's total EDF length increases the gain spectrum exhibits a positive tilt (e.g., longer wavelength channels will experience more gain). **Figs. 7(c, d)** show the average channel and maximum channel NF, respectively. From a system design perspective, a link can be limited by the optical channel with the lowest optical signal-to-noise ratio (OSNR) or highest noise figure, and this is a critical parameter of interest when designing optical amplifiers. In practice the best amplifier design is a compromise between high-gain, low-gain flatness, and low NF. From extensive numerical simulations, for the two-stage gain flattened amplifier example it was for approximately 4 meters in EDF stage 1, and 13 meters in stage 2.

Successful commercial EDFA products are more than a single amplifier design that can accommodate all applications. In some cases an economical solution is a modular design approach, where smaller capacity, lower-cost amplifiers are initially installed in a network. Additional gain can be added to the basic amplifier for longer-distance spans or as network traffic increases over time. This 'pay as you grow' approach can be done using additional gain stages, or pump lasers, and can be upgraded without interruption of revenue-generating network traffic.

Low-Noise Design of EDFAs

A key design feature of commercial EDFAs is low-noise operation since the degradation of the OSNR limits the reach of the system. The NF of an EDFA can be defined as

$$NF = SNR_{in}/SNR_{out} \quad (1)$$

The NF can also be calculated from

$$NF = \frac{1}{G} \times \left[\frac{P_{ASE}(v_s, L)}{h\nu_s} + 1 \right] \quad (2)$$

where P_{ASE} is the amplified spontaneous emission (ASE) noise power at the signal frequency v_s for an EDFA of gain G (h =Planck's constant). An alternative equation for the NF for a single section of EDF is

$$NF = \int_0^L \frac{\gamma_e(v_s, z) - \gamma_a(v_s, z)}{P_s(z)} P_s(0) dz + 1 \quad (3)$$

where γ_e and γ_a are the EDF emission and absorption factors at the signal frequency v_s , at distance z along the amplifying fiber. Eq. (3) shows that low NF can result from rapid signal gain in the initial fiber along the EDF. This result indicates that co-propagating pump and signal photons, both into the same end of the EDF, combined with the maximum inversion obtained by using a 980 nm pump, will result in low-noise amplification. The NF is typically expressed in dB units, and using a full quantum mechanical theory, the lowest possible NF for a fully inverted EDF with a large signal gain can be shown to be approximately 3 dB. Actual EDFAs will only approach the '3 dB quantum noise limit' when the signal is significantly lower than the available pump power (e.g., low channel count applications) so that high inversion can be maintained along the EDF length. Optical components in the EDFA prior to the first EDF will attenuate the signal but not the noise, since ASE is only generated in the EDF, and this input stage loss will directly add to the EDFA's NF.

The use of 1480 nm pump lasers can also result in a further NF penalty of 0.3 to 1.5 dB since complete inversion cannot be achieved when the erbium ions are operating as a two-level atomic system and more ASE is generated. The ASE, like the signal gain, has a wavelength-dependent spectral profile and this results in each channel having a different NF. Typically, the optical channel with the largest NF (lowest OSNR) is used to define the EDFA's noise performance. EDFAs with low-input signal power (e.g., -15 dBm) and high gain (e.g., 25 dB) may be described as low noise when the NF is <4 dB. For large total input signal power (e.g., +2.5 dBm for high channel count operation) low noise may be NF <5 dB.

For an amplifier with multiple EDF sections, the noise figure can be calculated from

$$NF_{\text{system}} = \frac{NF_1}{L_1} + \frac{NF_2}{L_1 G_1 L_2} + \dots + \frac{NF_N}{L_1 G_1 L_2 G_2 \dots L_{N-1} G_{N-1}} \quad (4)$$

where L_i is the signal loss prior to the gain G_i of the EDF stage i . From Eq. (4), multistage amplifiers with large inter-stage losses will have numerically larger NFs, but the total module NF is dominated by the first EDF stage. Undersea EDFAs, with a single short EDF section, very small L_i and 980 nm pumping, can have NF <3.5 dB.

Advanced EDFA Functions

An important trend in optical amplifiers, and systems in general, is the move towards dynamic control. Several technologies are being considered and the primary idea is to dynamically control the intensity level of each optical channel or wavelength to correct for any gain/loss inequalities in the network. This is particularly important as channel counts and bit rates rise concurrently with the functionality demanded of optical communication networks. A common requirement in this category is fast amplifier response so that in a network where individual optical channels can be added or dropped there are no power distortions at other surviving wavelengths. In a typical drop event, in less than 1 microsecond, 31 of 32 wavelengths can be completely removed from the input to the EDFA by an add/drop module located elsewhere in the network. This event will cause a total input power decrease of 15 dB and without rapid pump power adjustment the remaining channel will be excessively amplified.

Erbium ions have an excited state lifetime of approximately 10 ms, and this sets the time-scale for the transient signal response. A network transient usually manifests as a sharp optical amplifier (OA) gain fluctuation lasting up to several tens of microseconds that will pulse large signal levels through the entire optical network. Dynamic network loading necessitates controlling temporal gain transients, and this may be achieved by using a microprocessor in the amplifier module. These intelligent EDFA modules can rapidly adjust the pump laser power based on feedback from internal photo detectors to control gain transients.

The need for dynamic control also exists in the frequency domain. Wideband amplifiers have a gain variation among the optical channels of typically up to 1 dB over all operating conditions, and since these EDFAs are concatenated over many spans the electronic receiver circuit at the end of the network will have to detect optical channels with widely varying power levels. This can limit the number of spans that the network can bridge before electronic regeneration becomes necessary. This problem can be corrected using a dynamic gain equalizing filter (DGEF) to actively attenuate individual optical channels levels. A DGEF can be embedded inside an amplifier to create a loss-less component, and can be used to equalize channel powers for an entire network.

The problems of accumulation of gain ripple become more apparent in systems that employ amplifiers based on stimulated Raman scattering. A conventional EDFA amplified system may have 10 spans transmitting 100 optical channels at 10 Gbit/s each over a distance of approximately 800 km before electronic regeneration becomes necessary. To span longer distances with higher capacity, Raman amplifiers are used to amplify the transmission fiber along with EDFAs. This system architecture has the advantage of a higher effective OSNR for the individual spans, and has been demonstrated for distances over 4000 km with 40 Gbit/s channels. However, the variability of the Raman amplifier gain, due to network loading and differing transmission fiber properties, necessitates the use of loss-less DGEFs.

High-capacity transmission uses individual channels that are each modulated up to typical 2.5, 10, or 40 GHz, that requires that the optical amplifiers have low polarization mode dispersion (PMD) and low chromatic dispersion (CD). Polarization mode dispersion distorts the temporal spread of an ultra-short pulse as different polarization states travel at different speeds through the transmission fiber and amplifier components. PMD is particularly a problem in the first generation of installed transmission fiber and through polarization components (e.g., isolators) that are not PMD compensated. Without PMD compensation, 40 Gbit/s transmission, for example, can be limited to very short distances (several km). Fortunately, PMD compensators have been developed for upgrading aged installed networks, and also amplifiers are available with very low PMD.

To achieve long-distance transmission, a chromatic dispersion map is produced for all the optical spans, and by using a dispersion compensating module in each amplifier, the net dispersion across the channel wavelength window is controlled. Although optical amplifiers have approximately 1000 times less optical fiber in them compared to the transmission span that it is amplifying, chromatic dispersion in the amplifier can also be of concern for very long-haul networks. The CD of the amplifier is primarily in the EDF, and is specific to the EDF's geometric and chemical composition. Furthermore, since the erbium ion is a resonant atomic system, there may be a pump and signal power contribution to the dispersion (resonant dispersion) that could degrade network performance. Characterization of these optical effects is typically done on manufactured 'field grade' amplifiers as part of a quality control process.

Conclusion

As optical networks evolve EDFAs will continue to be an enabling technology for higher capacity and more dynamic communications networks. EDFA technology has already advanced to accommodate multiple channels, to span several wavelength windows and to provide features such as dispersion compensation, gain-transient suppression, and dynamic gain flatness. As future network architectures are introduced, to incorporate Raman amplification and additional wavelength windows, the demands placed upon EDFA design will continue to expand.

See also: Optical Amplifiers: SOAs

Further Reading

- Agrawal, G., 1989. *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
- Agrawal, G., 1997. *Fiber Optic Communication Systems*. New York: John Wiley and Sons.
- Becker, P.C., Olsson, N.A., Simpson, J.R., 1999. *Erbium-doped Fiber Amplifiers – Fundamentals and Technology*. San Diego, CA: Academic Press.
- Desurvire, E., 1995. *Erbium Doped Fiber Amplifiers*. New York: John Wiley and Sons.
- Pedersen, B., *et al.*, 1991. The design of erbium-doped fiber amplifiers. *Journal of Lightwave Technology* 9 (9), 1105.
- Sudo, S., 1997. *Optical Fiber Amplifiers: Materials, Devices and Applications*. Boston, MA: Artech House.

All-Optical Signal Regeneration

O Leclerc, Alcatel Research & Innovation, Marcoussis, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The breakthrough of optical amplification, combined with the techniques of wavelength division multiplexing (WDM) and dispersion management, have made it possible to exploit a sizeable fraction of the optical fiber bandwidth (several terahertz). Systems based on 10 Gbit/s per channel bit-rate and showing capacities of several terabit/s, with transmission capabilities of hundreds or even thousands of kilometers, have reached the commercial area.

While greater capacities and spectral efficiencies are likely to be reached with current technologies, there is potential economic interest in reducing the number of wavelength channels by increasing the channel rate (e.g., 40 Gbit/s). However, such fourfold increase in the channel bit-rate clearly results in a significant increase in propagation impairments, stemming from the combined effects of noise accumulation, fiber dispersion, fiber nonlinearities, and inter-channel interactions and contributing to two main forms of signal degradation. The first one is related to the amplitude domain; power levels of marks and spaces can suffer from random deviations arising from interaction between signal and amplified spontaneous emission (ASE) noise or with signals from other channels through cross-phase modulation (XPM) from distortions induced by chromatic dispersion. The second type of signal degradations occurs in the time domain; time position of pulses can also suffer from random deviations arising from interactions between signal and ASE noise through fiber dispersion. Preservation of high power contrast between '1' and '0', and of both amplitude fluctuations and timing jitter below some acceptable levels, are mandatory for high transmission quality, evaluated through bit-error-rate (BER) measurements or estimated by Q-factors. Moreover, in future optical networks, it appears mandatory to ensure similar but high optical signal quality at the output of whatever nodes in the networks, as to enable successful transmission of the data over arbitrary distance.

Among possible solutions to overcome such systems limitations is the implementation of Optical Signal Regeneration, either in-line for long-haul transmission applications or at the output of network nodes. Such Signal Regeneration performs, or should be able to perform, three basic signal-processing functions that are Re-amplifying, Re-shaping, and Re-timing, hence the generic acronym '3R' (Fig. 1). When Re-timing is absent, one usually refers to the regenerator as '2R' device, which has only re-amplifying and re-shaping capabilities. Thus, full 3R regeneration with retiming capability requires clock extraction.

Given system impairments after some transmission distance, two solutions remain for extending the actual reach of an optical transmission system or the scalability of an optical network. The first consists in segmenting the system into independent trunks, with full electronic repeater/transceivers at interfaces (we shall refer to this as 'opto-electronic regeneration' or O/E Regeneration forthwith). The second solution, i.e., all-optical Regeneration, is not the optical version of the first which would have higher

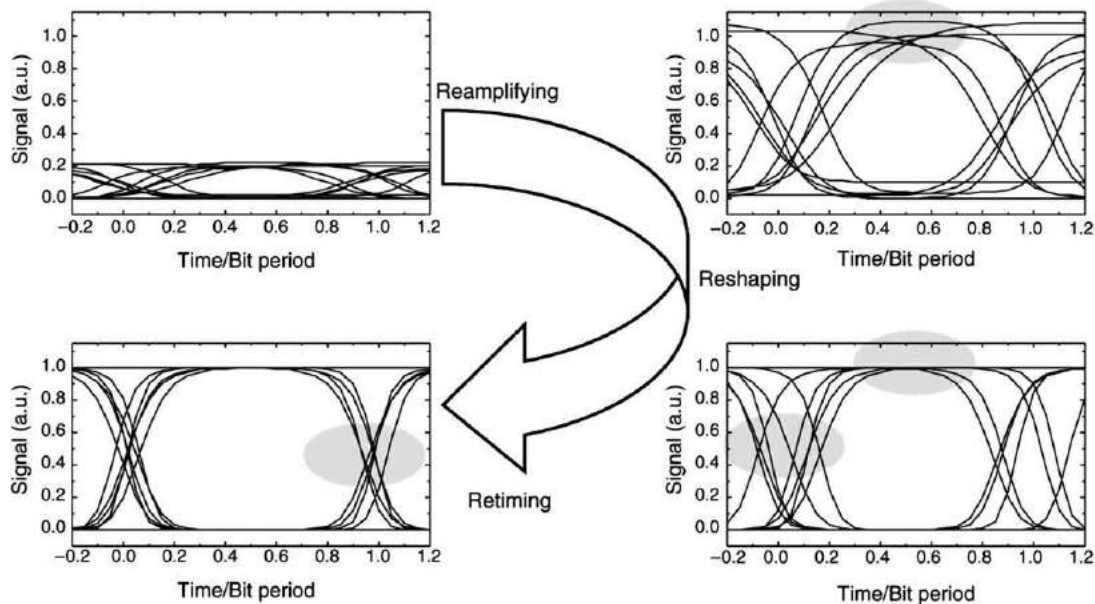


Fig. 1 Principle of 3R regeneration, as applied to NRZ signals. (a) Re-Amplifying; (b) Re-Shaping; (c) Re-Timing. NB: Such eye diagrams can be either optical or electrical eye diagrams.

bandwidth capability but still performs the same signal-restoring functions with far reduced complexity. At this point, it should be noted that Optical 3R techniques are not necessarily void of any electronic functions (e.g., when using electronic clock recovery and O/E modulation), but the main feature is that these electronic functions are narrowband (as opposed to broadband in the case of electronic regeneration).

Some key issues have to be considered when comparing such Signal Regeneration approaches. The first is that today's and future optical transmission systems or/and networks are WDM networks. Under this condition, the WDM compatibility – or the fact that any Regeneration solution can simultaneously process several WDM channels – represents a key advantage. The maturity of the technology – either purely optical or opto-electronic – also plays an important role in the potential (pre-)development of such solutions. But the main parameter that will decide the actual technology (and also technique) relies on the tradeoff between actual performance of the regeneration solutions and their costs (device and implementation), depending on the targeted applications (long-haul system, medium haul transport, wide area optical network, etc.).

In this article, we review the current alternatives for all-optical Signal Regeneration, considering both theoretical and experimental performance and practical implementation issues. Key advantages and possible drawbacks of each solutions are discussed, to sketch the picture in this field. However, first we must focus on some generalities about Signal Regeneration and the way to define (and qualify) such regenerator performance. In a second part, we will detail the currently-investigated optical solutions for Signal Regeneration with a specific highlight for semiconductor-based solutions using either semiconductor optical amplifiers (SOA) technology or newly-developed saturable absorbers. Optical Regeneration techniques based on synchronous modulation will also be discussed in a third section. The conclusion will summarize the key features of each solution, so as to underline the demanding challenge optical components are facing in this application.

Generalities on Signal Regeneration

Principles

In the general case, Signal Regeneration is performed using a decision element exhibiting a nonlinear transfer function. Provided with a threshold level and when associated with an amplifier, such an element then performs the actual Re-shaping of the incoming data (either in the electrical or optical domain) and completes a 2R Signal Regenerator. Fig. 2 shows the generic structure of such a Signal Regenerator in the general case as applied to non return to zero (NRZ) data. A clock recovery block can be added (dotted lines) to provide the decision element with time reference and hence perform the third R (Re-timing) of full Signal 3R Regeneration. At this point, it should be mentioned that the decision element can operate either on electrical signals (standard electrical DFF) provided that optical $\rightarrow$ electrical and electrical $\rightarrow$ optical signal conversion stages are added or directed onto optical signals using the different techniques described below. The clock signal can be of an electrical nature, as for electrical decision element in O/E regenerator – or either an electrical or a purely optical signal in all-optical regenerators.

Prior to reviewing and describing the various current technology alternatives for such Optical Signal Regeneration, the issue of the actual characterization of regenerator performance needs to be explained and clarified. As previously mentioned, the core element of any Signal Regenerator is the decision element showing a nonlinear transfer function that can be of varying steepness. As will be seen in Fig. 3, the actual regenerative performance of the regenerator will indeed depend upon the degree of nonlinearity of the decision element transfer function.

Fig. 3 shows the principle of operation of a regenerator incorporating a decision element with two steepnesses of the nonlinear transfer function. In any case, the '1' and '0' symbols amplitude probability densities (PD) are squeezed after passing through the decision element. However, depending upon the addition of a clock reference for triggering the decision element, the symbol arrival time PD will be also squeezed (clocked DECISION=3R regeneration) or enlarged (no clock REFERENCE=2R regeneration) resulting in conversion of amplitude fluctuations to time position fluctuations.

As for system performance – expressed through BER – regenerative capabilities of any regenerator simultaneously depend upon both the output amplitude and arrival time PD of the '1' and '0' symbols. In the unusual case of 2R regeneration (no clocked decision), a tradeoff has then to be derived, considering both the reduction of amplitude PD and the enlarged arrival time PD induced by the regenerator, to ensure sufficient signal improvement. In Fig. 3(a), we consider a step function for the transfer function of the decision element. In this case, amplitude PD are squeezed to Dirac PD after the decision element, and depending upon addition or not of a clock reference, arrival time PD is reduced (3R) or dramatically enlarged (2R). In Fig. 3(b), the decision element exhibits a moderately nonlinear transfer function. This results in an asymmetric and less-pronounced squeezing of the

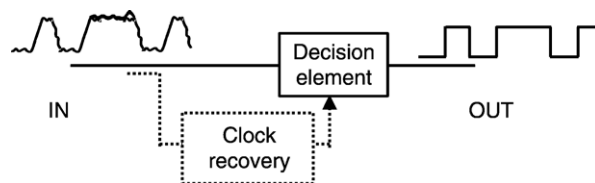


Fig. 2 Generic structure of Signal 2R/3R Regenerator based on Decision Element (2R) and Decision Element and Clock Recovery (3R).

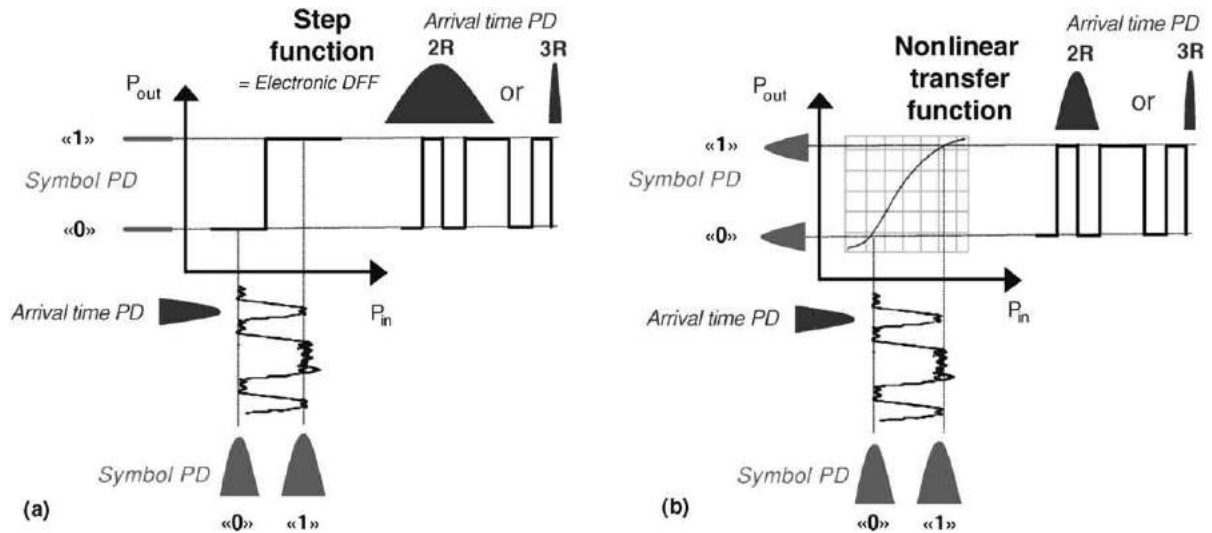


Fig. 3 Signal Regeneration process using Nonlinear Gates. (a) Step transfer function (=Electronic DFF); (b) 'moderately' nonlinear transfer function. As an illustration of the Regenerator operation '1' and '0' symbols amplitude probability density (PD) and arrival time probability density (PD) are shown in light gray and dark gray, respectively.

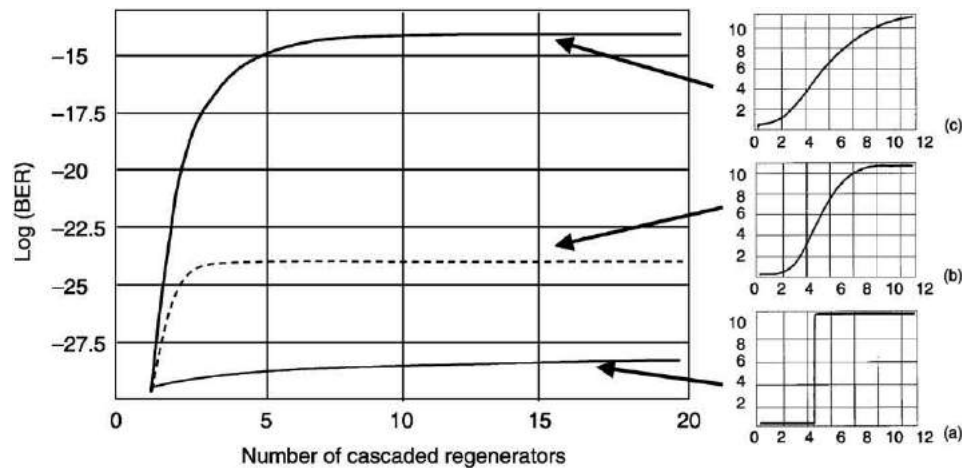


Fig. 4 Evolution of the BER with concatenated regenerators for nonlinear gates with nonlinear transfer function of decreasing depths from case (a)–(c) (step function).

amplitude PD compared to the previous case, but in turn results in a significantly less enlarged arrival time PD when no clock reference is added (2R regeneration). Comparison of these two decision element of different nonlinear transfer function indicates that for 3R regeneration applications, the more nonlinear the transfer function of the decision element the better performance, the ideal case being the step function. In the case of 2R regeneration applications, a tradeoff between the actual reduction of the amplitude PD and enlargement of timing PD is to be derived and clearly depends upon the actual shape of the nonlinear transfer function of the decision element.

Qualification of Signal Regenerator Performance

To further illustrate the impact of the actual shape of the nonlinear transfer function of the decision element in 3R application, the theoretical evolution of BER with number of concatenated regenerators have been plotted for regenerators having different nonlinear responses. **Fig. 4** shows the numerically calculated evolution of the BER of a 10 Gbit/s NRZ signal with fixed optical signal-to-noise ratio (OSNR) at the input of the 3R regenerator, as a function of the cascaded regenerator incorporating nonlinear gates with nonlinear transfer function of different depths. From **Fig. 4**, can be seen different behaviors depending on the nonlinear function shape. As previously stated, the best regeneration performance is obtained with an ideal step function (case a), which is actually the case for O/E regenerator using electronic decision flip-flop (DFF). In that case, BER linearly increase (i.e., more errors)

in the cascade. Conversely, when nonlinearity is reduced (cases (b) and (c)), both BER and noise accumulate, until the concatenation of nonlinear functions reach some steady-state pattern, from which BER linearly increases. Concatenation of nonlinear devices thus magnifies shape differences in their nonlinear response, and hence their regenerative capabilities.

Moreover, as can be seen in Fig. 4, all curves standing for different regeneration efficiencies pass through a common point defined after the first device. This clearly indicates that it is not possible to qualify the regenerative capability of any regenerator when considering the output signal after only one regenerator. Indeed, the BER is the same for either a 3R regenerator or a mere amplifier if only measured after a single element. This originates from the initial overlap between noise distributions associated with marks and spaces, that cannot be suppressed but only minimized by a single decision element through threshold adjustment.

As a general result, the actual characterization of the regenerative performance of any regenerator should *in fine* be conducted considering a cascade of regenerators. In practice this can easily be done with the experimental implementation of the regenerator under study in a recirculating loop. Moreover, such an investigation tool will also enable access to the regenerator performance with respect to the transmission capabilities of the regenerated signal, which should not be overlooked.

Let us now consider the physical implementation of such all-optical Signal Regenerators, along with the key features offered by the different alternatives.

All-Optical 2R/3R Regeneration Using Optical Nonlinear Gates

Prior to describing the different solutions for all-optical in-line Optical Signal Regeneration, it should be mentioned that since the polarization states of the optical data signals cannot be preserved during propagation, it is required that the regenerator exhibits an extremely low polarization sensitivity. This clearly translates to a careful optimization of all the different optical components making up the 2R/3R regenerator. It should be noted that this also applies to the O/E solutions but is of limited impact, since only the photodiode has to be polarization insensitive.

Fig. 5 illustrates the generic principle of operation of an all-optical 3R Regenerator using optical nonlinear gates. Contrary to what occurs in O/E, regenerator where the extracted clock signal drives the electrical decision element, the incoming and distorted optical data signal triggers the nonlinear gate, hence generating a switching window which is applied to a newly generated optical clock signal so as to reproduce the initial data stream on the new optical carrier.

In the case of 2R Optical Signal Regeneration, a continuous wave (CW) signal is substituted for the synchronized optical clock pulses. As previously mentioned, the actual regeneration performance of the 2R/3R devices will mainly depend upon the nonlinearity of the transfer function of the decision element but in 3R applications the quality of the optical clock pulses has also to be considered. In the following, we describe current solutions to explain the two main building blocks of all-optical Signal Regenerators: the decision element (i.e., nonlinear gate) and the clock recovery (CR) elements.

Optical Decision Element

In the physical domain, optical decision elements with ideal step response – as for electrical DFF – do not exist. Different nonlinear optical transfer functions, approaching more or less the ideal case, can be realized in various media such as fiber, SOA, electro-absorption modulators (EAM), and lasers. Generally, as described below, the actual response (hence the regenerative properties of the device) of such optical gates directly depends upon the incoming signal instantaneous power. Under these conditions, it appears essential to add an adaptation stage so as to reduce intensity fluctuations (as caused by propagation or crossing routing/switching node) and provide the decision element with fixed power conditions. In practice, this results in the addition of a control circuitry (either optical or electrical) in the Re-amplification block, whose complexity directly depends on actual system environment (ultra-fast power equalization for packet-switching applications and compensation of slow power fluctuations in transmission applications).

As previously described the decision gate performs Re-shaping (and Re-timing when clock pulses are added) of the incoming distorted optical signal, and represent the regenerator's core element. Ideally, it should also act as a transmitter with characteristics ensuring the actual propagation of the regenerated data stream. In that respect, the chirp possibly induced by the optical decision gate onto the regenerated signal – and the initial quality of the optical clock pulses in 3R applications – should be carefully considered (ideally by means of loop transmission) as to adequately match line transmission requirements.

Different solutions for the actual realization of the optical decision element have been proposed and extensively investigated using, for example, cross gain modulation in semiconductor optical amplifier (SOA) devices but the most promising and flexible

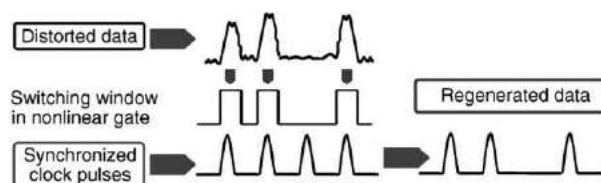


Fig. 5 Principle of operation of an all-Optical Signal 3R Regenerator using nonlinear gates.

devices probably are interferometers, for which, descriptions of the generic principle of operation follows. Consider a CW signal (probe) at λ_2 wavelength injected into an optical interferometer, in which one arm incorporates a nonlinear medium in which an input signal carried by λ_1 wavelength (command) is, in turn, injected. Such a signal, at λ_1 wavelength, induces a phase shift through cross-phase modulation (XPM) in this arm of the interferometer, the amount depending upon power level P_{in,λ_1} . In turn, such phase modulation (PM) induces amplitude modulation (AM) on the signal at λ_2 wavelength when recombined at the output of the interferometer and translates the information carried by wavelength λ_1 onto λ_2 . Under these conditions, such optical gates clearly act as wavelength converters (it should be mentioned that Wavelength Conversion is not necessarily equivalent to Regeneration; i.e., a linear transfer function performs suitable Wavelength Conversion but by no means Signal Regeneration).

Optical interferometers can be classified according to the nature of the nonlinearity exploited to achieve a π phase shift. In the case of fiber-based devices, such as the nonlinear optical loop mirror (NOLM), the phase shift is induced through the Kerr effect in an optical fiber. The key advantage of fiber-based devices such as NOLM lies in the near-instantaneous (fs) response of the Kerr nonlinearity, making them very attractive for ultra-high bit-rate operation (≥ 160 Gbit/s). Polarization-insensitive NOLMs have been realized, although with the same drawbacks concerning integrability. With recent developments in highly nonlinear (HNL) fibers, however, the required NOLM fiber length could be significantly reduced, hence dramatically reducing environmental instability.

A second type of device is the integrated SOA-based Mach-Zehnder interferometers (MZI). In MZIs, the phase shift is due to the effect of photo-induced carrier depletion in the gain saturation regime of one of the SOAs. The control and probe can be launched in counter- or co-directional ways. In the first case, no optical filter is required at the output of the device for rejecting the signal at λ_1 wavelength but operation of the MZI is limited by its speed. At this point, one should mention that the photo-induced modulation effects in SOAs are intrinsically limited in speed by the gain recovery time, which is a function of the carrier lifetime and the injection current. An approach referred to as differential operation mode (DOM) and illustrated on Fig. 6, which takes advantage of the MZI's interferometric properties, makes it possible to artificially increase the operation speed of such 'slow' devices up to 40 Gbit/s.

As discussed earlier, the nonlinear response is a key parameter for regeneration efficiency. Combining two interferometers is a straightforward means to improve the nonlinearity of the decision element transfer function, and hence regeneration efficiency. This approach was validated at 40 Gbit/s using a cascade of two SOA-MZI, (see Fig. 7 (left)). Such a scheme offers the advantage of restoring data polarity and wavelength, hence making the regenerator inherently transparent. Finally, the second conversion stage can be used as an adaptation interface to the transmission link achieved through chirp tuning in this second device.

Such an Optical 3R Regenerator was upgraded to 40 Gbit/s, using DOM in both SOA-MZIs with validation in a 40 Gbit/s loop RZ transmission. The 40 Gbit/s eye diagram monitored at the regenerator output after 1, 10, and 100 circulations are shown in Fig. 7 (right) and remain unaltered by distance. With this all-optical regenerator structure, the minimum OSNR tolerated by the regenerator (1 dB sensitivity penalty at 10^{-10} BER) was found to be as low as 25 dB/0.1 nm. Such results clearly illustrate the high performance of this SOA-based regenerator structure for 40 Gbit/s optical data signals.

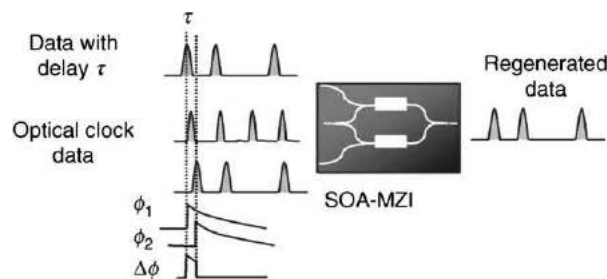


Fig. 6 Schematic and principle of operation of SOA-MZI in differential mode.

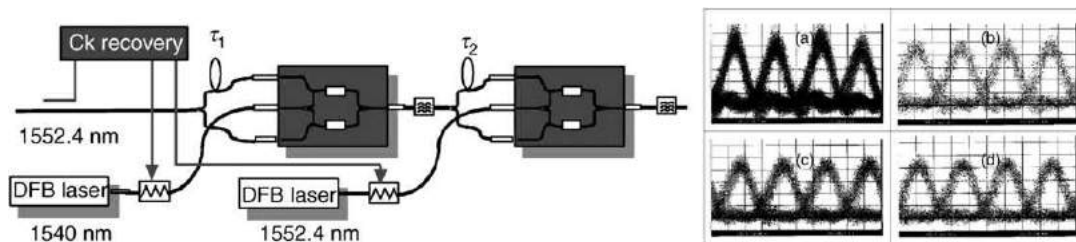


Fig. 7 Optimized structure of a 40 Gbit/s SOA-based 3R regenerator. 40 Gbit/s eye diagram evolution: (a) B-to-B; (b) 1 lap; (c) 10 laps; (d) 100 laps.

Such a complex mode of operation for addressing 40 Gbit/s bit-rates will probably be discarded when we consider the recent demonstration of standard-mode wavelength conversion at 40 Gbit/s, which uses a newly-designed active-passive SOA-MZI incorporating evanescent-coupling SOAs. The device architecture is flexible in the number of SOAs, thus enabling easier operation optimization and reduced power consumption, leading to simplified architectures and operation for 40 Gbit/s optical 3R regeneration.

Based on the same concept of wavelength conversion for Optical 3R Regeneration, it should be noted many devices have been proposed and experimentally validated as wavelength converters at rates up to 84 Gbit/s, but with cascability issues still to be demonstrated to assess their actual regenerative properties.

Optical Clock Recovery (CR)

Next to the decision, the CR is a second key function in 3R regenerators. One possible approach for CR uses electronics while another only uses optics. The former goes with OE conversion by means of a photodiode and subsequent EO conversion through a modulator. This conversion becomes more complex and power-hungry as the data-rate increases. It is clear that the maturity of electronics gives a current advantage to this approach. But considering the pros and cons of electronic CR for cost-effective implementation, the all-optical approach seems more promising, since full regenerator integration is potentially possible with reduced power consumption. In this view, we shall focus here on the optical approach and more specifically on the self-pulsating effect in three-sections distributed feedback (DFB) lasers or more recently in distributed Bragg reflector (DBR) lasers. Recent experimental results illustrate the potentials of such devices for high bit rates (up to 160 Gbit/s), broad dynamic range, broad frequency tuning, polarization insensitivity, and relatively short locking times (1 ns). This last feature makes these devices good candidates for operation in asynchronous-packet regimes.

Optical Regeneration by Saturable Absorbers

We next consider saturable absorbers (SA) as nonlinear elements for optical regeneration. Fig. 8 (left) shows a typical SA transfer function and illustrates the principle of operation. When illuminated with an optical signal with peak power below some threshold (P_{sat}), the photonic absorption of the SA is high and the device is opaque to the signal (low transmittance). Above P_{sat} , the SA transmittance rapidly increases and asymptotically saturates to transparency (passive loss being overlooked). Such a nonlinear transfer function only applies to 2R optical regeneration.

Different technologies for implementing SAs are available, but the most promising approach uses semiconductors. In this case, SA relies upon the control of carrier dynamics through the material's recombination centers. Parameters such as on-off contrast (ratio of transmittance at high and low incident powers), recovery time ($1/e$) and saturation energy, are key to device optimization. In the following, we consider a newly-developed ion-irradiated MQW-based device incorporated in a micro-cavity and shown on Fig. 8 (right). The device operates as a reflection-mode vertical cavity, providing both a high on/off extinction ratio by canceling reflection at low intensity and a low saturation energy of 2 pJ. It is also intrinsically polarization-insensitive. Heavy ion-irradiation of the SA ensures recovery times (at $1/e$) shorter than 5 ps (hence compatible with bit-rate above 40 Gbit/s), while maintaining a dynamic contrast in excess of 2.5 dB at 40 GHz repetition rate.

The regenerative properties of SA make it possible to reduce cumulated amplified spontaneous emission (ASE) in the '0' bits, resulting in a higher contrast between mark and space, hence increasing system performance. Yet SAs do not suppress intensity noise in the marks, which makes the regenerator incomplete. A solution for this noise suppression is optical filtering with nonlinear (soliton) pulses. The principle is as follows. In absence of chirp, the soliton temporal width scales in the same way as the reciprocal of its spectral width (Fourier-transform limit) times its intensity (fundamental soliton relation). Thus, an increase in pulse intensity corresponds to both time narrowing and spectral broadening. Conversely, a decrease in pulse intensity corresponds to time broadening and spectral narrowing. Thus, the filter causes higher loss when intensity increases, and lower loss when intensity decreases. The filter thus acts as an automatic power control (APC) in feed-forward mode, which causes power

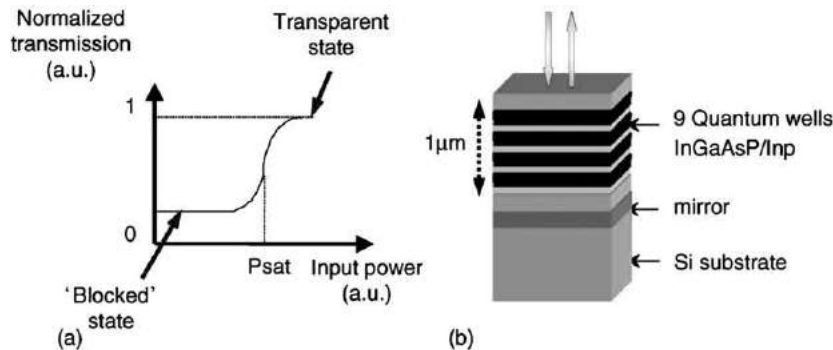


Fig. 8 (a) Saturable Absorber (SA) ideal transfer function. (b) Structure of Multi-Quantum Well SA.

stabilization. The resulting 2R regenerator (composed by the SA and the optical filter) is fully passive, which is of high interest for submarine systems where the power consumption must be minimal, but it does not include any control in the time domain (no Re-timing).

System demonstrations of such passive SA-based Optical Regeneration have been reported with a 20 Gbit/s single-channel loop experiment. Implementation of the SA-based 2R Regenerator with 160 km-loop periodicity made it possible to double the error-free distance ($Q = 15.6$ dB or 10^{-9} BER) of a 20 Gbit/s RZ signal. So as to extend the capability of passive 2R regeneration to 40 Gbit/s systems, an improved configuration was derived from numerical optimization and experimentally demonstrated in a 40 Gbit/s WDM-like, dispersion-managed loop transmission, showing more than a fourfold increase in the WDM transmission distance at 10^{-4} BER (1650 km without the SA-based regenerator and 7600 km when implementing the 2R regenerator with 240 km periodicity).

Such a result illustrates the potential high interest of such passive optical 2R regeneration in long-haul transmission (typically in noise-limited systems) since the implementation of SA increases the system's robustness to OSNR degradation without any extra power consumption. Reducing both saturation energy and insertion loss along with increasing dynamic contrast represent key future device improvements. Regeneration of WDM signals from the same device, such as one SA chip with multiple fibers implemented between Mux/DMux stages, should also be thoroughly investigated. In this respect, SA wavelength selectivity in quantum dots could possibly be advantageously exploited.

Synchronous Modulation Technique

All-optical 3R regeneration can be also achieved through in-line synchronous modulation (SM) associated with narrowband filtering (NF). Fig. 9 shows the basic layout of such an Optical 3R Regenerator. It is composed of an optical filter followed by an Intensity and Phase Modulator (IM/PM) driven by a recovered clock. Periodic insertion of SM-based modulators along the transmission link provides efficient jitter reduction and asymptotically controls ASE noise level, resulting in virtually unlimited transmission distances. Re-shaping and Re-timing provided by IM/PM intrinsically requires nonlinear (soliton) propagation in the trunk fiber following the SM block. Therefore, one can refer to the approach as distributed optical regeneration. This contrasts with lumped regeneration, where 3R is completed within the regenerator (see above with Optical Regeneration using nonlinear gates), and is independent of line transmission characteristics.

However, when using a new approach referred to 'black box' optical regeneration (BBOR), it is possible to make the SM regeneration function and transmission work independently in such a way that any type of RZ signals (soliton or non-soliton) can be transmitted through the system. The BBOR technique includes an adaptation stage for incoming RZ pulses in the SM-based regenerator, which ensures high regeneration efficiency regardless of RZ signal format (linear RZ, DM-soliton, C-RZ, etc.). This is achieved using a local and periodic soliton conversion of RZ pulses by means of launching an adequate power into some length of fiber with anomalous dispersion. The actual experimental demonstration of the BBOR approach and its superiority over the 'classical' SM-based scheme for DM transmission was experimentally investigated in 40 Gbit/s DM loop transmission. Under these conditions, one can then independently exploit dispersion management (DM) techniques for increasing spectral efficiency in long-haul transmission, while ensuring high transmission quality through BBOR.

One of the key properties of the SM-based all-optical Regeneration technique relies on its WDM compatibility. The first (Fig. 10, left) and straightforward solution to apply Signal Regeneration to WDM channels amounts to allocating a regenerator to each WDM channel. The second consists in sharing a single modulator, thus processing the WDM channels at once in serial fashion. This approach requires WDM synchronicity, meaning that all bits be synchronous with the modulation, that can be achieved either by use of appropriate time-delay lines located within a DMux/Mux apparatus (Fig. 10, upper right), or by making

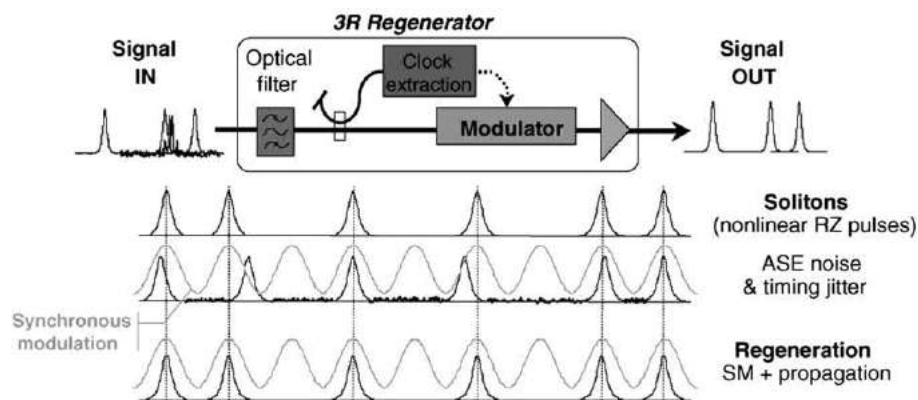


Fig. 9 Basic layout of the all-optical Regenerator by Synchronous Modulation and Narrowband Filtering and illustration of the principle of operation.

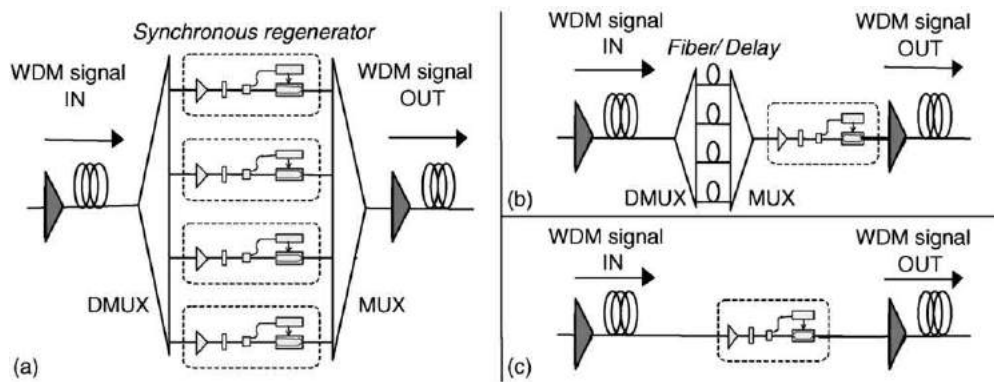


Fig. 10 Basic implementation schemes for WDM all-Optical Regeneration. (a) parallel asynchronous; (b) serial re-synchronized; (c) serial self-synchronized.

the WDM channels inherently time-coincident at specific regenerator locations (**Fig. 10, bottom right**). Clearly, the serial regeneration scheme is far simpler and cost-effective than the parallel version; however, optimized re-synchronization schemes still remain to be developed for realistic applications. Experimental demonstration of this concept was assessed by means of a 4×40 Gbit/s dispersion-managed transmission over 10 000 km ($\text{BER} < 5.10^{-8}$) in which a single modulator was used for the simultaneous regeneration of the 4 WDM channels.

Considering next all-optical regeneration schemes with ultra-high speed potential, a compact and loss-free 40 Gbit/s Synchronous Modulator, based on optically-controlled SOA-MZI, was proposed and loop-demonstrated at 40 Gbit/s with an error-free transmission distance in excess of 10 000 km. Moreover, potential ultra-high operation of this improved BBOR scheme was recently experimentally demonstrated by means of a 80 GHz clock conversion with appropriate characteristics through the SOA-MZI. One should finally mention all fiber-based devices such as NOLM and NALM for addressing ultra-high speed SM-based Optical Regeneration, although no successful experimental demonstrations have been reported so far in this field.

Conclusion

In summary, optical solutions for Signal Regeneration present many key advantages. These are the only advantages to date to possibly ensure WDM compatibility of the regeneration function (mostly 2R related). Such optical devices clearly exhibits the best 2R regeneration performance (wrt to O/E solutions) as a result of the moderately nonlinear transfer function (which in turn can be considered as a drawback in 3R applications), but the optimum configuration is still to be clearly derived and identified depending upon the system application. Optics also allow to foresee and possibly target ultrafast applications above 40G, for signal regeneration if needed. Among the current drawbacks, one should mention the relative lack of wavelength/formats flexibility of these solutions (compared to O/E solutions). It is complex or difficult to restore the input wavelength or address any C-band wavelength at the output of the device or to successfully regenerate modulation formats other than RZ. In that respect, investigations should be conducted to derive new optical solutions capable of processing more advanced modulation formats at 40G. Finally, the fact that the nonlinear transfer function of the optical is in general triggered by the input signal instantaneous power also turns out to be a drawback since it requires control circuitry. The issue of cost (footprint, power consumption, etc.) of these solutions, compared to O/E ones, is still open. In this respect, purely optical solutions incorporating all-optical clock recovery, the performance of which is still to be technically assessed, are of high interest for reducing costs. Complete integration of an all-optical 2R/3R regenerator or such parallel regenerators onto a single semiconductor chip should also contribute to make all-optical solutions cost-attractive even though acceptable performance of such fully integrated devices is still to be demonstrated.

From today's status concerning the two alternative approaches for in-line regeneration (O/E or all-optical), it is safe to say that the choice between either solution will be primarily dictated by both engineering and economical considerations. It will result from a tradeoff between overall system performance, system complexity and reliability, availability, time-to-market, and rapid returns from the technology investment.

Further Reading

- Bigo, S., Leclerc, O., Desurvire, E., 1997. All-optical fiber signal processing for solitons communications. *IEEE Journal of Selected Topics on Quantum Electronics* 3 (5), 1208–1223.
- Dagens B, Labrousse A, Fabre S, *et al.* (2002) New modular SOA-based active-passive integrated Mach-Zehnder interferometer and first standard mode 40 Gb/s all-optical wavelength conversion on the C-band. paper PD 3.1, Proceedings of ECOCOL 02, Copenhagen.
- Durhuus, T., Mikkelsen, B., Joergensen, C., Danielsen, S.L., Stubkjaer, K.E., 1996. All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14 (6), 942–954.

- Lavigne B, Guerber P, Brindel P, Balmeffre E and Dagens B (2001) Cascade of 100 optical 3R regenerators at 40 Gbit/s based on all-active Mach–Zehnder interferometers. paper We.F.2.6, Proceedings of ECOC 2001, Amsterdam.
- Leclerc, O., Lavigne, B., Chiaroni, D., Desurvire, E., 2002. All-optical regeneration: Principles and WDM implementation. Kaminow, I.P., Li, T. (Eds.), 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. 732–784. chap. 15.
- Leuthold, J., Raybon, G., Su, Y., 2002. 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. Electronic Letters 38 (16), 890–892.
- Öhlén, P., Berglind, E., 1997. Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters. IEEE Photon Technol. Letters 9 (7), 1011–1013.
- Otani T, Suzuki M and Yamamoto S (2001) 40 Gbit/s optical 3R regenerator for all-optical networks. We.F.2.1, Proceedings of ECOC 2001, Amsterdam.
- Oudar, J.-L., Aubin, G., Mangeney, J., 2004. Ultra-fast quantum-well saturable absorber devices and their application to all-optical regeneration of telecommunication optical signals. Annales des Télécommunications 58 (11–12),
- Sartorius, B., Mohrle, M., Reichenbacher, S., 1997. Dispersive self-Q-switching in self-pulsating DFB lasers. IEEE Journal of Quantum Electronics 33 (2), 211–218.
- Ueno Y, Nakamura S, Sasaki J, *et al.* (2001) Ultrahigh-speed all-optical data regeneration and wavelength conversion for OTDM systems. Th.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Heterodyning

T-C Poon, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

© 2005 Elsevier Ltd. All rights reserved.

Heterodyning, also known as frequency mixing, is a frequency translation process. Heterodyning has its root in radio engineering. The principle of heterodyning was discovered in the late 1910s by radio engineers experimenting with radio vacuum tubes. Russian cellist and electronic engineer Leon Theremin invented the so-called Thereminvox, which was one of the earliest electronic musical instruments that generates an audio signal by combining two different radio signals. American electrical engineer Edwin Armstrong invented the so-called super-heterodyne receiver. The receiver shifts the spectrum of the modulated signal, that is, the frequency contents of the modulated signal, so as to demodulate the audio information from it. Commercial amplitude modulation (AM) broadcast receivers nowadays are super-heterodyne type. After illustrating the basic principle of heterodyning by some simple mathematics, we will discuss some examples on how the principle of heterodyning is used in radio and in optics.

For a simple case, when signals of two different frequencies are heterodyned or mixed, the resulting signal produces two new frequencies, the sum and difference of the two original frequencies. Fig. 1 illustrates how signals of two frequencies are mixed or heterodyned to produce two new frequencies, $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$, by simply multiplying the two signals $\cos(\omega_1 t + \theta_1)$ and $\cos(\omega_2 t)$, where ω_1 and ω_2 are radian frequencies of the two signals and θ_1 is the phase angle between the two signals. Note that when the frequencies of the two signals to be heterodyned are the same, i.e., $\omega_1 = \omega_2$, the phase information of $\cos(\omega_1 t + \theta_1)$ can be extracted to get $\cos \theta_1$ if we use an electronic lowpass filter (LPF) to filter out the term $\cos(2\omega_1 t + \theta_1)$. This is shown in Fig. 2. Heterodyning is often referred to as homodyning for the mixing of two signals of the same frequency.

For a general case of heterodyning, we can have a signal represented by $s(t)$ with its spectrum $s(\omega)$, which is given by the Fourier transform of $s(t)$, and when it is multiplied by $\cos(\omega_2 t)$, the resulting spectrum is frequency-shifted to new locations in the frequency domain as:

$$\mathcal{F}\{s(t)\cos(\omega_2 t)\} = \frac{1}{2}S(\omega - \omega_2) + \frac{1}{2}S(\omega + \omega_2) \quad (1)$$

where $\mathcal{F}\{s(t)\} = S(\omega)$ and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform of the quantity being bracketed. The spectrum of $s(t)\cos(\omega_2 t)$, along with the spectrum of $s(t)$, is illustrated in Fig. 3. It is clear that multiplying $s(t)$ with $\cos(\omega_2 t)$ is a process of heterodyning, as we have translated the spectrum of the signal $s(t)$. This process is known as modulation in communication systems.

In order to appreciate the process of heterodyning let us now consider, for example, heterodyning in radio. In particular, we consider AM. While the frequency content of an audio signal (from 0 to around 3.5 kHz) is suitable to be transmitted over a pair of wires or coaxial cables, it is, however, difficult to be transmitted in air. By impressing the audio information onto a higher

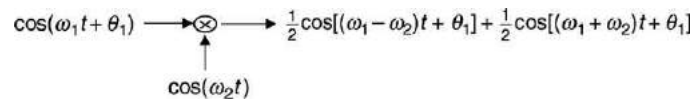


Fig. 1 Heterodyning of two sinusoidal signals.

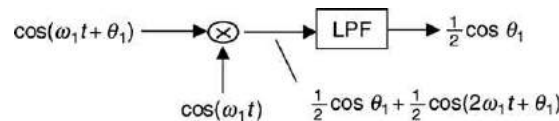


Fig. 2 Heterodyning becomes homodyning when the two frequencies to be mixed are the same: homodyning allows the extraction of the phase information of the signal.

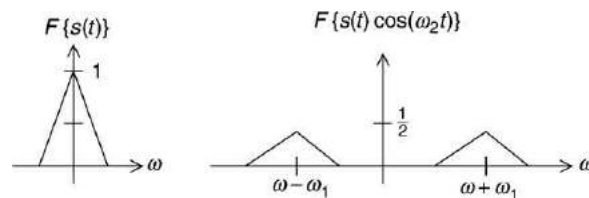


Fig. 3 Spectrum of $s(t)$ and $s(t)\cos(\omega_2 t)$. It is clear that spectrum of $s(t)$ has been translated to new locations upon multiplying a sinusoidal signal.

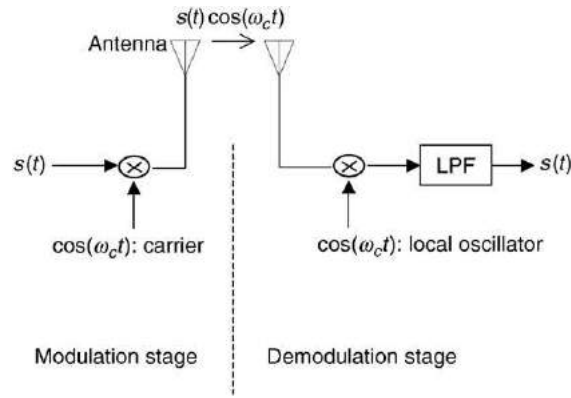


Fig. 4 Radio link: heterodyning in the demodulation stage is often known as heterodyne detection.

frequency, say 550 kHz, as one of the broadcast bands in AM radio, i.e., by modulating the audio signal, the resulting modulated signal can now be transmitted using antennas of reasonable dimensions. In order to recover (demodulate) the audio signal from the modulated signal, the modulated signal is first received by an antenna in a radio receiver, multiplied by the so-called local oscillator with the same frequency as the signal used to modulate the audio signal, then followed by a lowpass filter to eventually obtain the audio information back. The situation is illustrated in Fig. 4. We shall now describe the process mathematically with reference to Fig. 4.

We denote $s(t)$ as the audio signal. After multiplying by $\cos(\omega_c t)$, which is usually called a carrier in radio in the modulation stage, where ω_c is the radian frequency of the carrier. The modulated signal $s(t)\cos(\omega_c t)$ is transmitted via an antenna. When the modulated signal is received in the demodulation stage, the modulated signal is then multiplied by the local oscillator of signal waveform $\cos(\omega_c t)$. The output of the multiplier is given by

$$s(t)\cos(\omega_c t)\cos(\omega_c t) = s(t)\frac{1}{2}[1 + \cos(2\omega_c t)] \quad (2)$$

Note that the two signals in the demodulation stage, $s(t)\cos(\omega_c t)$ and $\cos(\omega_c t)$, are heterodyned to produce two new frequencies, the sum $\omega_c + \omega_c = 2\omega_c$ to give the term $\cos(2\omega_c t)$ and the difference $\omega_c - \omega_c = 0$ to give the term $\cos(0) = 1$. Since the two frequencies to be multiplied in the demodulation stage are the same, this is homodyning as explained above. Now by performing lowpass filtering (LPF), we can recover our original audio information $s(t)$. Note that if the frequency of the local oscillator in the demodulation stage within the receiver is higher than the frequency of the carrier used in the modulation stage, heterodyning in the receiver is referred to as super-heterodyning, which most receivers nowadays use for amplitude modulation. In general, we have two heterodyning processes in the radio system just described, one in the modulation stage and the other in the demodulation stage. However, it is unusual to speak of heterodyning in the modulation stage, and so we just refer to the process in the demodulation stage as 'heterodyne detection.'

In summary, heterodyne detection can extract the information from a modulated signal and can also extract the phase information of a signal if homodyning is used. We shall see, in the next section, how optical heterodyne detection is employed.

Optical information is usually carried by coherent light such as a laser. Let $\psi_p(x, y)$ be a complex amplitude of the light field, which may physically represent a component of the electric field. We further assume that the light field is oscillating at temporal frequency ω_0 . Therefore, we can write the light field as $\psi_p \exp(i\omega_0 t)$. By taking the real part of $\psi_p \exp(i\omega_0 t)$, i.e., $\text{Re}[\psi_p \exp(i\omega_0 t)]$, we recover the usual physical real quantity. For a simple example, if we let $\psi_p = A \exp(-ik_0 z)$, where A is some constant and k_0 is the wavenumber of the light, $\text{Re}[\psi_p \exp(i\omega_0 t)] = \text{Re}[A \exp(-ik_0 z) \exp(i\omega_0 t)] = A \cos(\omega_0 t - k_0 z)$, which is a plane wave propagating along the positive z direction in free space. To detect light energy, we use a photodetector (PD), as shown in Fig. 5. Assuming a plane wave for simplicity, as $\psi_p = A$, we have also taken $z=0$ at the surface of the photodetector. As the photodetector responds to intensity, i.e., $|\psi_p|^2$, which gives the current, i , as output by spatially integrating the intensity:

$$i \propto \int_S |\psi_p \exp(i\omega_0 t)|^2 dx dy = A^2 S \quad (3)$$

where S is the surface area of the photodetector. We can see that the photodetector current is proportional to the intensity, A^2 , of the incident light. Hence the output current varies according to the intensity of the optical signal intensity. This mode of photodetection is called direct detection or incoherent detection in optics.

Let us now consider the heterodyning of two plane waves on the surface of the photodetector. We assume an information-carrying plane wave, also called the signal plane wave, $A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)] \times \exp(-ik_0 x \sin \phi)\}$, and a reference plane wave, $A_r \exp(i\omega_0 t)$, or called a local oscillator in radio. The situation is shown in Fig. 6.

Note that the frequency of the signal plane wave is ω_s higher than that of the reference signal, and it is inclined at an angle ϕ with respect to the reference plane wave, which is normal incident to the photodetector. Also, the information content $s(t)$ is in the phase of the signal wave. In the situation shown in Fig. 6, we see that the two plane waves are interfered on the surface of the

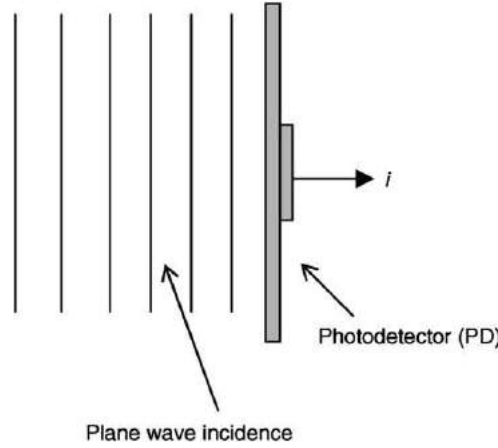


Fig. 5 Optical direction detection or optical incoherent detection.

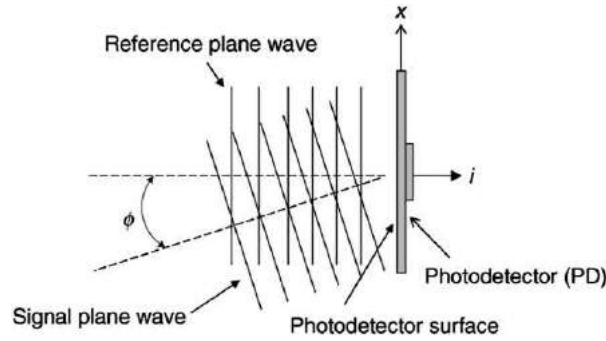


Fig. 6 Optical heterodyne detection.

photodetector, giving the total light field ψ_t , given by

$$\psi_t = A_r \exp(i\omega_0 t) + A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)]\} \times \exp(-ik_0 x \sin \phi) \quad (4)$$

Again, the photodetector responds to intensity, giving the current

$$i \propto \int_S |\psi_t|^2 dx dy = \int_{-a}^a \int_{-a}^a [A_r^2 + A_s^2 + 2A_r A_s \cos(\omega_s t + s(t) - k_0 x \sin \phi)] dx dy \quad (5)$$

where we have assumed that the photodetector has a $2a \times 2a$ square area. The current can be evaluated to be:

$$i(t) \propto 2a(A_r^2 + A_s^2) + 4A_r A_s \frac{\sin(k_0 a \sin \phi)}{k_0 \sin \phi} \cos(\omega_s t + s(t)) \quad (6)$$

The current output has two parts: the DC current and the AC current. The AC current at frequency ω_s is commonly known as the heterodyne current. Note that the information content $s(t)$ originally embedded in the phase of the signal plane wave has now been preserved and transferred to the phase of the heterodyne current. The above process is called optical heterodyning. In optical communications, it is often referred to as optical coherent detection. In contrast, if the reference plane wave has not been used for the detection, we have the incoherent detection. The information content carried by the signal plane wave would be lost, as it is evident that for $A_r = 0$, the above equation gives only a DC current at a value proportional to the intensity of the plane wave, A_s^2 .

Let us now consider some of the practical issues encountered in coherent detection. Again, the AC part of the current given by the above equation is the heterodyne current $i_{\text{het}}(t)$, given by

$$i_{\text{het}}(t) \propto A_r A_s \frac{\sin(k_0 a \sin \phi)}{k_0 \sin \phi} \cos(\omega_s t + s(t)) \quad (7)$$

We see that since the two plane waves propagate in slightly different directions, the heterodyne current output is degraded by a factor of

$$\frac{\sin(k_0 a \sin \phi)}{k_0 \sin \phi} = a \operatorname{sinc}(k_0 a \sin \phi)$$

where $\operatorname{sinc}(x) = \sin(x)/x$. For small angles, i.e., $\sin \phi \approx \phi$, the current amplitude falls off as $\operatorname{sinc}(k_0 a \phi)$. Hence, the heterodyne current is at a maximum when the angular separation between the signal plane wave and the reference plane wave is zero, i.e., the

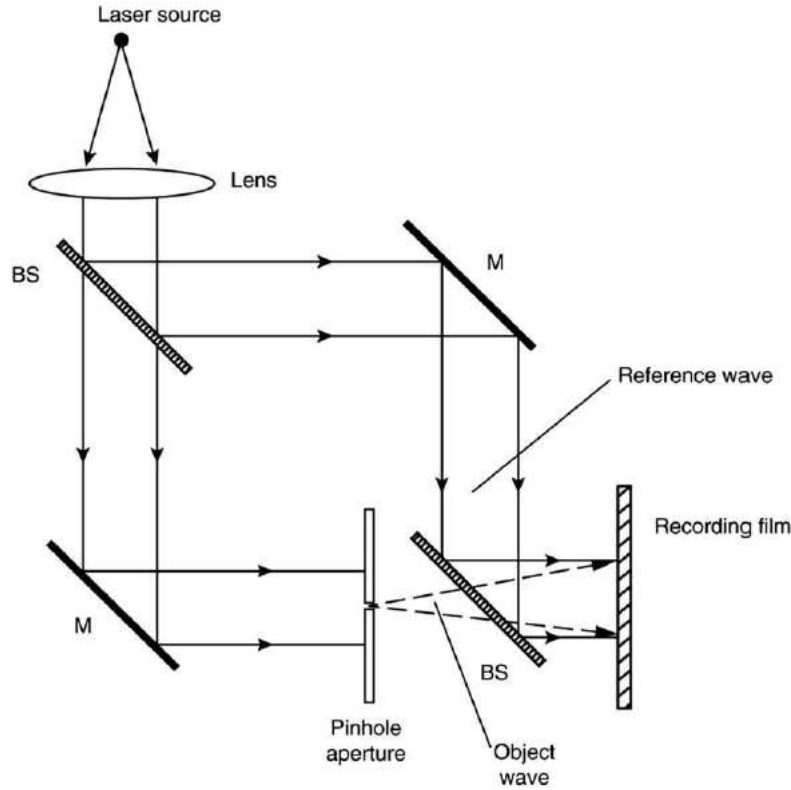


Fig. 7 Holographic recording of a point source object.

two plane waves are propagating exactly in the same direction. The current will go to zero when $k_0 a \phi = \pi$, or $\phi = \lambda_0 / 2a$, where λ_0 is the wavelength of the laser light. To see how critical it is for the angle ϕ to be aligned in order to have any heterodyne output, we assume the size of the photodetector $2a = 1\text{cm}$ and the laser used is red, i.e., $\lambda_0 \approx 0.6\mu\text{m}$; ϕ is calculated to be about 2.3×10^{-3} degrees. Hence to be able to work with coherent detection, we need to have precise optomechanical mounts for angular rotation.

Finally we describe a famous application of heterodyning in optics – holography. As we have seen, the mixing of two light fields with different temporal frequencies will produce heterodyne current at the output of the photodetector. But we can record the two spatial light fields of the same temporal frequency with photographic films instead of using electrical devices. Photographic films respond to light intensity as well. We discuss holographic recording of a point source object as a simple example. **Fig. 7** shows a collimated laser split into two plane waves and recombined by the use of two mirrors (M) and two beamsplitters (BS). One plane wave is used to illuminate the pinhole aperture (our point source object), and the other illuminates the recording film directly. The plane wave that is scattered by the point source object, which is located z_0 away from the film, generates a diverging spherical wave. This diverging wave is known as an object wave in holography. The object wave arising from the point source object on the film is given by, according to Fresnel diffraction:

$$\psi_0 = A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \times \exp\left[-ik_0(x^2 + y^2)/2z_0\right] \exp(i\omega_0 t) \quad (8)$$

This object wave is a spherical wave, where A_0 is the amplitude of the spherical wave.

The plane wave that directly illuminates the photographic plate is known as a reference wave, ψ_r . For the reference plane wave, we assume that the plane wave has the same phase with the point source object at a distance z_0 away from the film. Its field distribution on the film is, therefore, $\psi_r = A_r \exp(-ik_0 z_0) \exp(i\omega_0 t)$, where A_r is the amplitude of the plane wave. The film now records the interference of the reference wave and the object wave, i.e., what is recorded on the film as a 2D pattern is given by $t(x, y) \propto |\psi_r + \psi_0|^2$. The resulting recorded 2D pattern, $t(x, y)$, is called the hologram of the object. This kind of recording is known as holographic recording, distinct from a photographic recording in which the reference wave does not exist and hence only the object wave is recorded. Now:

$$\begin{aligned} t(x, y) &\propto |\psi_r + \psi_0|^2 = |A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \exp[-ik_0(x^2 + y^2)/2z_0] \exp(i\omega_0 t)|^2 \\ &= A + B \sin\left\{\frac{k_0}{2z_0} [(x^2 + y^2)/2z_0]\right\} \end{aligned} \quad (9)$$

where A and B are some inessential constants. Note that the result of recording two spatial light fields of the same temporal

frequency preserves the phase information (noticeably the depth parameter z_0) of the object wave. This is considered optical homodyning as it is clear from the result shown in Fig. 2.

The intensity distribution being recorded on the film, upon being developed, will have transmittance given by the above equation. This expression is called the Fresnel zone plate, which is the hologram of a point source object and we shall call it the point-object hologram. Fig. 8 shows the hologram for a particular value of z_0 and k_0 . The importance of this phase-preserving recording is that when we illuminate the hologram with a plane wave, called the reconstruction wave, a point object is reconstructed at the same location of the original point source object if we look towards the hologram as shown in Fig. 9, that is the point source is reconstructed at a distance z_0 from the hologram as if it were at the same distance away from the recording film. For an arbitrary 3D object, we can imagine the object as a collection of points, and therefore, we can see that we have a collection of Fresnel zone plates on the hologram. Upon reconstruction, such as the illumination of the hologram by a plane wave, the observer would see a 3D virtual object behind a hologram.

As a final example, we discuss a state-of-the-art holographic recording technique called optical scanning holography, which employs the use of optical heterodyning and electronic homodyning to achieve real-time holographic recording without the use of films. We will take the holographic recording of a point object as an example. Suppose we superimpose a plane wave and a spherical wave of different temporal frequencies and use the resulting intensity pattern as an optical scanning beam to raster scan a point object located z_0 away from the point source that generates the spherical wave. The situation is shown in Fig. 10. Point C is the point source that generates the spherical wave (shown with dashed lines) on the pinhole object, our point object. This point source, for example, can be generated by a focusing lens. The solid parallel rays represent the plane wave. The interference of the two waves generates a Fresnel zone plate type pattern on the pinhole object:

$$I_s(x, y; z_0, t) = \left| A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \exp\left[-ik_0(x^2 + y^2)/2z_0\right] \exp[i(\omega_0 + \Omega)t] \right|^2$$

$$= 1 + \left(\frac{1}{\lambda_0 z_0}\right)^2 + \frac{1}{\lambda_0 z_0} \sin\left[\frac{\pi}{\lambda_0 z_0}(x^2 + y^2) - \Omega t\right] \quad (10)$$

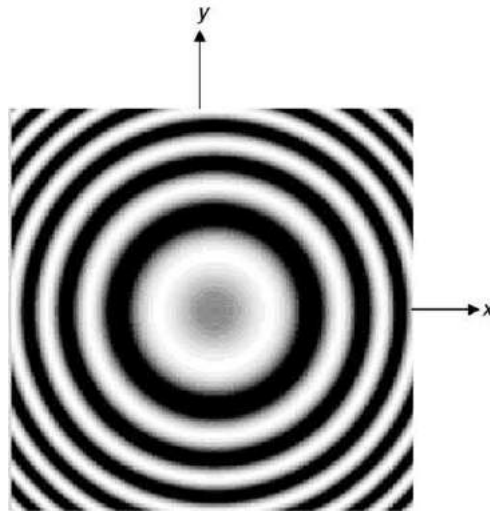


Fig. 8 Point-object hologram.

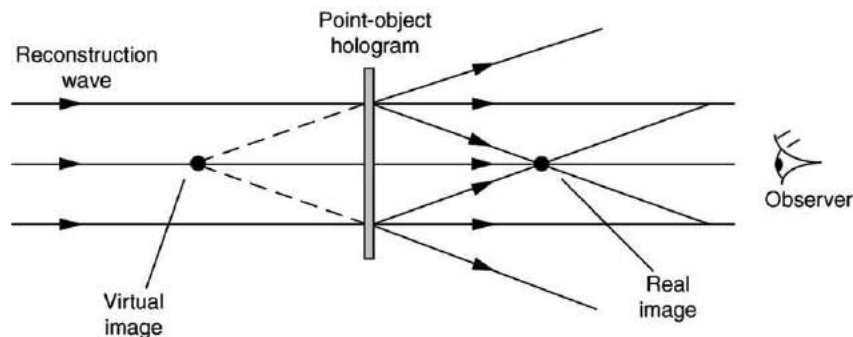


Fig. 9 Reconstruction of a point-object hologram.

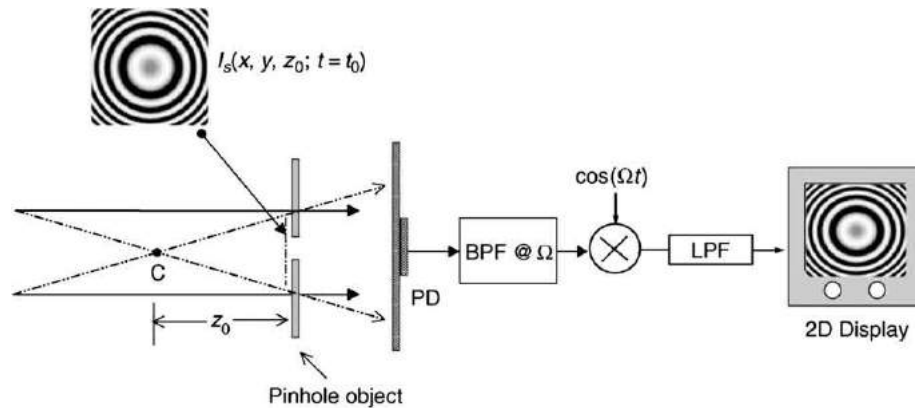


Fig. 10 Optical scanning holography (use of optical heterodyning and electronic homodyning to record holographic information upon scanning the object in two dimensions).

assuming $A_r = A_0 = 1$ for simplicity. Note that the plane wave is at temporal frequency ω_0 , and the spherical wave is at temporal frequency $\omega_0 + \Omega$. The expression $I_s(x, y; z_0, t)$ is a temporally modulated Fresnel zone plate and is known as the time-dependent Fresnel zone plate (TDFZP). If we freeze time, say at $t = t_0$, we have the Fresnel zone plate pattern on the pinhole object as shown in Fig. 10. However, if we let the time run in the above expression, physically we will have running zones that would be moving away from the center of the pattern. These running zones are the result of optical heterodyning of the plane wave and the spherical wave of different temporal frequencies. It is this TDFZP that is used to raster scan a 3D object to obtain holographic information of the scanned object and such technique is called optical scanning holography.

Upon scanning by the TDFZP, the photodetector, captures all the transmitted light and delivers a current, which consists of a heterodyne current at frequency Ω . After electronic bandpass filtering (BPF) at Ω , the heterodyne current is homodyned electronically by $\cos(\Omega t)$ and lowpass filtered (LPF) to extract the phase information of the current. When this final scanned and processed current is display in a 2D display, as shown in Fig. 10, we have the Fresnel zone plate, which is the hologram of the pinhole object, on the display. We can photograph this 2D display to obtain a transparency and have it illuminated by a plane wave to have the reconstruction as shown in Fig. 9, or, since the hologram is now in electronic form, we can store it in a PC and reconstruct it digitally. When holographic reconstruction is performed digitally, we have so-called digital holography.

See also: All-Optical Multiplexing/Demultiplexing

Further Reading

- Banerjee, P.P., Poon, T.C., 1991. *Principles of Applied Optics*. MA: Irwin.
 Palais, J.C., 1988. *Fiber Optic Communications*. New Jersey: Prentice-Hall.
 Poon, T.C., Banerjee, P.P., 2001. *Contemporary Optical Image Processing with MATLAB*. Oxford: Elsevier.
 Poon, T.C., Qi, Y., 2003. Novel real-time joint-transform correlation using acousto-optic heterodyning. *Applied Optics* 42, 4663–4669.
 Poon, T.C., Wu, M.H., Shinoda, K., Suzuki, Y., 1996. Optical scanning holography. *Proc. IEEE* 84, 753–764.
 Pratt, W.K., 1969. *Laser Communication Systems*. New York: Wiley.
 Stremler, F.G., 1990. *Introduction to Communication Systems*. MA: Addison-Wesley.
 VanderLugt, A., 1992. *Optical Signal Processing*. New York: Wiley.

Terahertz Lasers

Benjamin S Williams, University of California, Los Angeles, CA, United States

Qing Hu, MIT, Cambridge, MA, United States

© 2018 Elsevier Ltd. All rights reserved.

Optically Pumped Molecular Gas Lasers

The first commercially available terahertz lasers (more traditionally known as far-infrared or sub-millimeter wave lasers) were gas lasers, where stimulated emission takes place between rotational levels of excited vibrational states of low-pressure molecular gasses. The first such demonstration was made in 1963, where lasing in water vapor excited with a pulsed electric discharge was observed by Crocker *et al.* (1964). Gas lasers continue to be an important THz laser source, although now the most common arrangement is to use optical pumping by a CO₂ gas laser (Chang and Bridges, 1970), which significantly increases the pumping selectivity and efficiency (Fig. 1). A large variety of discrete lines can be obtained between 0.1 and 8 THz (Inguscio *et al.*, 1986), although most strongly pumped lines are below 3 THz. Continuous-wave power levels of milliwatts are not uncommon, with powers of hundreds of milliwatts possible for the strongest lines (e.g. the methanol lines at 118.8 and 184.3 μm). There have been many review articles published on far-IR gas lasers, including (Jacobsson (1989), Inguscio *et al.* (1986), Dodel (1999)), and they are commercially available from several vendors.

The selection of laser frequencies is limited by available gasses, some of which are “inconvenient” to handle from a safety and environmental point of view. Only discrete tuning of the lasers is available because of the limited transition lines, however harmonic generators can be used to generate tunable sidebands (Farhoomand *et al.*, 1985). Although these laser systems present limitations in terms of size and weight, this has not prevented them from being flown into space. For example, a far-IR gas laser was used as a 2.5 THz local oscillator source in the Earth Observing System (EOS) Aura satellite, in the Microwave Limb Sounder heterodyne instrument. This laser delivered ~ 30 mW of cw power while consuming 120 W of electrical power (Mueller *et al.*, 2007).

It is notable that THz molecular gas lasers can operate at room temperature, even though the THz photon energy ($h\nu \sim 1\text{--}30$ meV) is smaller than the thermal energy $k_B T$. This is in contrast to the THz semiconductor lasers to be discussed below, all of which require varying degrees of cryogenic cooling. A naïve assumption is sometimes made that a population

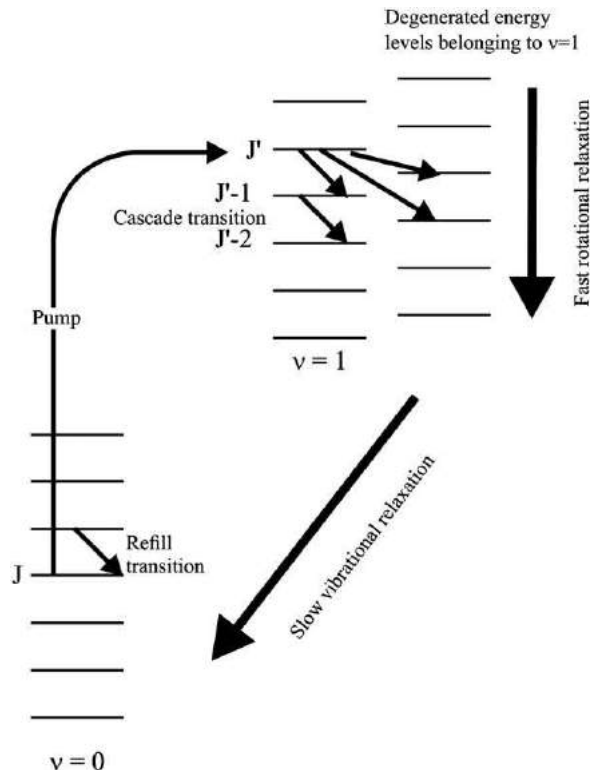


Fig. 1 Schematic of energy levels and transitions in a generic far-IR molecular gas laser. The pump photon excites the selected molecule from its vibration ground state to an excited vibrational state. FIR laser action takes place between rotational levels in the excited state. Figure from Jacobsson, S., 1989. Optically pumped far infrared lasers. *Infrared Phys.* 29, 853–874.

inversion is not possible if $(h\nu) < k_B T$). However, one must remember that a pumped laser system is by its nature not in thermal equilibrium. A population inversion is possible as long as the lower state lifetime is shorter than the upper to lower state relaxation time ($\tau_1 < \tau_{21}$). Rotational transitions in a low-pressure gas have very little broadening (~ 1 – 20 MHz), due to the weak scattering processes (Jacobsson, 1989). As a result, the energy states are very distinct, and can maintain different relaxation rates. Second, the narrow linewidths lead to large cross sections, and high peak levels of gain even for modest population inversion levels.

The inherent limitations of the molecular gas lasers (i.e. size, weight, limited tunability, and low power efficiencies etc.) have created interest in developing terahertz semiconductor lasers. One need look no further than the success of the interband diode lasers in the visible and near-infrared wavelengths, to see the inspiration for developing (potentially) low cost, compact, coherent sources. However, the terahertz photon energies are 100–1000 times smaller than visible photons - appropriate materials that have both a small bandgap and can support a population inversion simply do not exist in nature. As a result, semiconductor terahertz laser research has focused on unconventional sources of stimulated emission. We will highlight three of these here: the germanium intra-valence band lasers, silicon impurity-state lasers, and quantum-cascade lasers.

Germanium Intra-Valence Band Lasers

The first type of THz semiconductor laser demonstrated was the hot-hole p-Ge laser. In this device, lasing action results from a hole population inversion in bulk p-Ge that is established between the light and heavy hole bands due to a “streaming motion” that takes place in crossed electric and magnetic fields (Golka *et al.*, 1991). For low lattice temperatures (< 20 K), holes have kinetic energies far below the optical phonon energy (37 meV), and their energy relaxation is mediated by (relatively) slow acoustic phonons and impurity scatterings. However, for critical ratios of the electric to magnetic field strength, heavy holes are less bent by the magnetic field due to their heavier mass and therefore can acquire more energy from the electric field and be accelerated to have sufficient kinetic energy to emit an optical phonon, and quickly relax to lower energies. Light holes acquire insufficient energy and do not scatter by the optical phonon. This acts as a pumping mechanism which establishes a population inversion between the light hole and heavy hole bands (Fig. 2).

Large fields are required: typical electric field strengths are from 1 to 2 kV/cm and magnetic fields greater than 1 T. Hence, a liquid-helium cooled superconducting magnet is typically used. The gain is relatively low, as the pumping mechanism is highly nonselective and inefficient, and the upper state lifetime is still relatively short. Free-carrier absorption is also an issue, and the doping must be kept low ($< 10^{15} \text{ cm}^{-3}$). Nonetheless, due to the large total semiconductor volume, large peak powers of several Watts have been obtained in pulsed mode. Absent a cavity mode selection mechanism, broadband lasing is obtained (spanning over 10 – 20 cm^{-1}) that can be tuned from 1–4 THz by the applied electric/magnetic fields. Its broadband gain has proven useful for mode-locked operation, where pulses with a 100-ps width have been obtained (Hovenier *et al.*, 2000). However the need for a strong magnetic field, high voltage, and cryogenic operation ($T < 20$ K) has limited the utility of this source. Traditionally, due to high power consumption and low efficiency the maximum duty cycle is usually limited to less than 10^{-4} , although a report of up to 5% duty cycle has been demonstrated (Bründermann *et al.*, 2000).

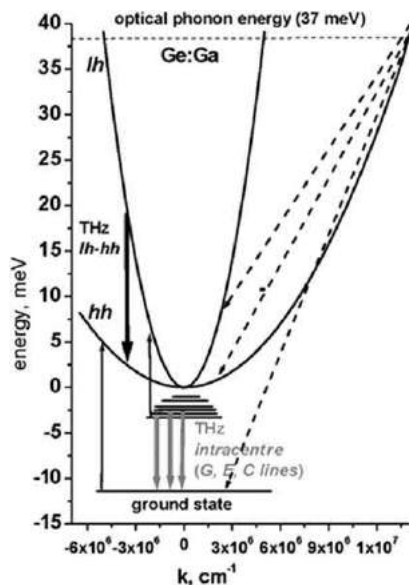


Fig. 2 Schematic of intra-valence band transitions in p-Ge lasers with the hole energies plotted in positive direction. Dashed arrows correspond to relaxation of holes due to interaction with optical phonons; the bold black downward arrow indicates stimulated emission from light- to heavy-hole subband. Figure from Hübers, H.-W., Pavlov, S.G., Shastin, V.N., 2005. Terahertz lasers based on germanium and silicon. *Semicond. Sci. Technol.* 20, S211–S221.

A related device is the strained p-Ge resonant state laser. *cw* lasing that was tunable with pressure from 2.5 to 10 THz was demonstrated (Goussev *et al.*, 1999). Power levels of tens of microwatts were observed and operation takes place at liquid helium temperatures. In this device, application of strain lifts the degeneracy of the light-hole and heavy-hole band such that the 1 s impurity state of the heavy-hole band is brought into resonance with the light-hole band. This leads to a population inversion between the heavy hole 1 s state and the light-hole impurity states, which are depopulated by electric field ionization. No magnetic field is necessary. A similar laser was demonstrated in 2000 in SiGe/Si quantum wells, where the mechanism of lasing is the same but the strain is provided instead by the epitaxial mismatch (Kagan *et al.*, 2000; Blom *et al.*, 2001). This is a promising method which eliminates the need for externally applied strain, but the power level is still quite low and high voltage (300–1500 V) is required.

Silicon Impurity State Lasers

Silicon is not ordinarily considered a promising material for semiconductor lasers, due to its indirect bandgap. However, *n*-type silicon crystals can be used to make optically pumped 4-level terahertz lasers, based upon donor intracenter radiative transitions. Reviews are given in Hübers *et al.*, (2005), Pavlov *et al.* (2013). Silicon is an attractive material for THz applications in general; unlike III-V semiconductors it is non-polar and has low lattice absorption in the THz range. Various donor impurities have been demonstrated for lasing, including As, P, Sb, and Bi. Excepting a few outliers, lasing has mostly been observed between 1–2 THz and 5–7 THz as shown in Fig. 3. Population inversion is obtained between states based upon varying rates of electron relaxation mediated by acoustic and optical phonons. Cavities are formed by polishing bulk crystals with sizes of several millimeters on a side. If desired, stress can be applied to modify the lasing frequency or improve the performance.

A schematic of the energy states and relaxation pathways is shown in Fig. 3 for several different donor species. Typical donor densities are in the range of 10^{15} – 10^{16} cm⁻³. Pumping occurs via photoionization of the impurity ground state into the conduction band, typically performed using a mid-infrared CO₂ laser. After the electron is excited high into the conduction band, it relaxes via emission of optical and acoustic phonons until it is captured in the high-excited donor states, whereupon it relaxes via emission of acoustic phonons along specific paths that depend upon the energy structure of the particular donor. For example, in Si:Sb and Si:P this relaxation ends in a long lived state ($2p_0$) with an estimated lifetime of hundreds of picoseconds. Lasing occurs between the $2p_0 \rightarrow 1s$ (T_2), where the lower radiative state is estimated to have a lifetime of 10–100ps.

Typical pump threshold intensities are in the range of 10–300 kW/cm², depending upon the donor type, concentration, and stress. Peak powers on the order of milliwatts have been measured. In order to avoid thermal ionization of the impurities, and thermal backfilling of the lower radiative state, operation requires cryogenic temperatures typically < 10 K, although operation up to 30 K has been observed for Si:Bi lasers. Furthermore, decay dynamics of the nonequilibrium phonon distribution limits the laser operation to pulsed mode – for example in Si:P and Si:Sb lasers emission is quenched after ~150 ns. The operating temperature may be potentially improved by the use of impurities with deeper energy states, so as to increase the energy barrier for thermal backfilling. Research is ongoing to better understand the complexities of the relaxation physics, as well as to improve the performance.

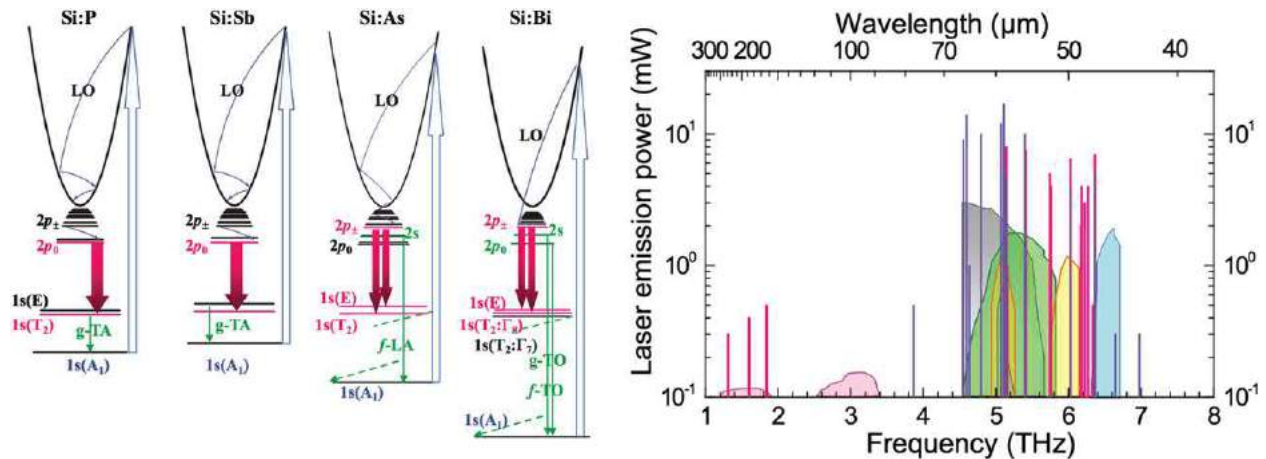


Fig. 3 (left) Laser schemes realized in Si: P, Si: Sb, Si:As and Si:Bi when pumped with a CO₂ laser into the conduction band. The straight blue arrows up indicate the optical pumping and the bold red arrows down are for stimulated emission. Resonant coupling with intervalley phonons is shown by green vertical arrows down. (right) Various reported laser lines from various optically pumped Si impurity state lasers. Figures from Pavlov, S.G., Zhukavin, R.K., Shastin, V.N., H.-W., 2013. The physical principles of terahertz silicon lasers based on intracenter transitions. *Phys. Status Solidi B* 250, 9–36.

Quantum-Cascade Lasers

At present, the most active area in THz laser research is in the field of terahertz quantum-cascade (QC) lasers. The QC-laser is a unipolar intraband laser in which photons are generated as electrons within the conduction band make radiative transitions between quantized subbands engineered into heterostructure quantum wells. By designing sophisticated sequences of coupled quantum wells, “one dimensional artificial molecules” are created with energy levels of the subbands designed to emit photons in the mid-infrared or terahertz spectral range. Furthermore, it is possible to achieve a fine degree of control over electron transport, such that electrons can be injected (usually) via resonant tunneling into an upper radiative state, and can be selectively depopulated from the lower state with some combination of tunneling and scattering processes. The radiative cross-section of the radiative transition can likewise be engineered by controlling the overlap and parity of the various wavefunctions.

The first THz quantum-cascade lasers were demonstrated in 2001 and shortly thereafter (Köhler *et al.*, 2002; Rochat *et al.*, 2002; Williams *et al.*, 2003); their technological progress has been summarized in various review articles (Williams, 2007; Kumar, 2011; Vitiello *et al.*, 2015). The most common material system is GaAs/Al_xGa_{1-x}As heterostructures grown by molecular beam epitaxy, although other material systems have been demonstrated as well (e.g. InGaAs/AlInAs, InGaAs/GaAsSb, etc.). A common feature of these polar III-V semiconductors, is that they all exhibit strong electron interaction with longitudinal optical (LO) phonons (which typically have energies from 30–50 meV). At this time various THz QC-lasers currently have been demonstrated between 1.2 and 5.6 THz (and at frequencies as low as 0.6 THz in a very strong magnetic field).

So far cryogenic operation is still required for THz QC-lasers; the current temperature record is $T_{\text{max}}=200$ K in pulsed mode (<1% duty cycles), and $T_{\text{max}}=129$ K in cw mode without the assistance of a magnetic field. By applying a strong magnetic field perpendicular to the planes of quantum wells, additional quantization can be achieved in the transverse dimensions, which reduces the available scattering volume in the momentum space and consequently increases the upper-state lifetimes. As a result, a significant increase of the maximum operating temperature is achieved, from 165 K to 225 K, and the latter is still the record of all the solid-state THz lasers at this writing (Wade *et al.*, 2009). While efforts to improve temperature performance of the active material continue, with ongoing improvements in cryogen-free coolers, and considering the modest power consumption requirements, even operation in the 40–90 K range is very feasible for many applications. For devices operating above 77 K in cw, milliwatt level powers are typical, although in some cases very high powers have been observed. Current record results stand at over 2 W peak power in pulsed mode, and 230 mW in continuous-wave (cw) mode for devices cooled by liquid helium. Furthermore, wall-plug power efficiencies (WPE) greater than 1% have been achieved by several groups.

Active Regions

The heart of the quantum-cascade laser is the multiple-quantum-well active region. Most active region designs can be grouped into three categories: bound-to-continuum (BTC), resonant-phonon (RP) designs, as well as hybrid designs which combine features of each. Schematics for the bandstructure and energy levels are shown in Fig. 4.

The key to the BTC design is the coupling of several quantum wells together to create a superlattice which supports a “miniband” of subband states when the appropriate electric field is applied. The radiative transition takes place between an isolated upper “bound” state that resides above the miniband, and a lower state which is the top state of the miniband. Generally speaking, intra-miniband scattering is favored over scattering out of the bound state, which creates a population inversion at lower temperatures. Owing to the relatively small widths of the minibands (about 15–20 meV) for THz QCLs, LO-phonons are not directly involved in the depopulation process. Since the energy drop per module is relatively small, and the oscillator strength is large, this design is often characterized by low threshold voltages and currents, although performance tends to drop off rapidly with rising temperature above 100 K.

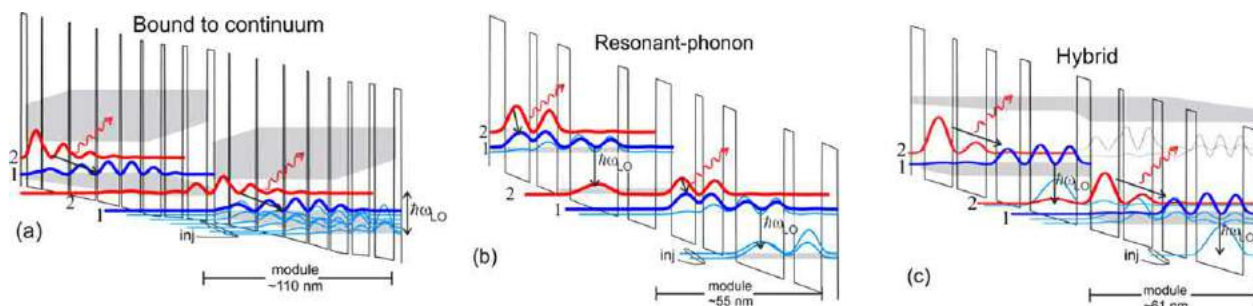


Fig. 4 Schematic figure of the conduction band edge (GaAs/Al_xGa_{1-x}As heterostructures) and energy levels of three types of active regions for quantum-cascade lasers. Two repeated modules are shown in each case. Energy levels are plotted with the probability density of the bound-state. Radiative transition occurs between levels 2 and 1. Also shown for (b) and (c) are electron relaxation via emission of longitudinal optical phonons. Figure adapted from Williams, B.S., 2007. Terahertz quantum cascade lasers. *Nat. Photon.* 1, 517–524.

The other major active region type is the resonant phonon (RP) scheme. This design, a reservoir/injector state is designed to reside 36 meV (one LO phonon energy) below the lower radiative state. This encourages fast depopulation via electron-LO-phonon scattering into the reservoir state. In order to prevent simultaneous depopulation of the upper state, the energy levels are designed so that electrons are selectively removed via resonant tunneling from the lower state only. In general, RP and hybrid designs have better temperature performance than the BTC designs. Good RP/hybrid designs regularly operate above 130 K and even as high as 200 K in pulsed mode. The explicit inclusion of an LO-phonon scattering event for depopulation results in a more robust depopulation mechanism, as well as a larger energetic barrier to thermal backfilling of electrons from the reservoir.

Several issues are limiting further improvements of the temperature performance of THz QC-lasers so far. Unlike gas lasers, electrons within QC-lasers exist within a high density solid-state matrix: they interact strongly with the lattice, other carriers, and have a large density of states available for scattering. As a result, lifetimes are picoseconds or shorter, and the energy level broadenings are on the order of hundreds of GHz or larger. This not only reduces the peak gain cross-section, but also tends to blur energy levels, making it difficult to maintain selective injection and depopulation essential to achieve a population inversion. This problem is exacerbated for THz QC-lasers at lower frequencies (< 2 THz), as the energy spacing becomes comparable with the level broadening.

The second major issue is thermally activated scattering and leakage that degrade the upper state lifetime for high electron temperatures. For example, as electrons in the upper subband acquire sufficient in-plane kinetic energy, they may emit an LO-phonon and relax to the lower subband, or tunnel into the continuum. There is active research underway to suppress these mechanisms, which primarily involves the development of novel active region designs. In addition, new heterostructure material systems are also under investigation. For example, GaN-based materials are attractive since GaN has a much larger LO-phonon energy (90 meV) compared to most III-Vs, which should in principle suppress thermally activated LO-phonon scattering. An even more radical approach is to develop quantum-dot based cascade lasers: the discrete density of states is predicted to suppress nonradiative and dephasing scattering to improve high temperature operation. Nonetheless, no lasers have been reported to date using these new material/structure approaches.

Waveguides and Cavities

The biggest difference between THz QC-lasers and semiconductor lasers at shorter wavelengths lies in the techniques for waveguiding. Conventional dielectric waveguides (as used for semiconductor diode lasers and mid-IR QCLs) are impractical for THz QC-lasers. Due to the $\lambda/2$ scaling of free-carrier loss, the use of doped cladding layers introduces excessive loss at THz wavelengths. Hence, two types of unique waveguides for THz QC-lasers have been developed that minimize the overlap of the mode with doped semiconductor, and instead use metal layers partially or fully for optical confinement; unlike at shorter wavelengths losses are quite modest for noble metals in the terahertz range. These two schemes are shown in Fig. 5.

The first type is the surface-plasmon (SP) waveguide, which involves the growth of a thin (0.2–0.8 μm thick) heavily doped layer underneath the 10 μm -thick GaAs/AlGaAs quantum-well active region, but on top of a semi-insulating GaAs substrate. The resulting mode is a compound surface plasmon tightly confined by the top metal contact, and loosely bound to the heavily doped lower plasma layer. The mode extends far into the substrate (by tens to hundreds of microns); since the substrate is semi-insulating, the free carrier loss is minimal. The downside is that ridges narrower than $\sim 100 \mu\text{m}$ tend to squeeze the mode out of the active region and into the substrate, which effectively puts a floor on the minimum device area (and power dissipation), which in turn limits the maximum achievable cw operating temperature.

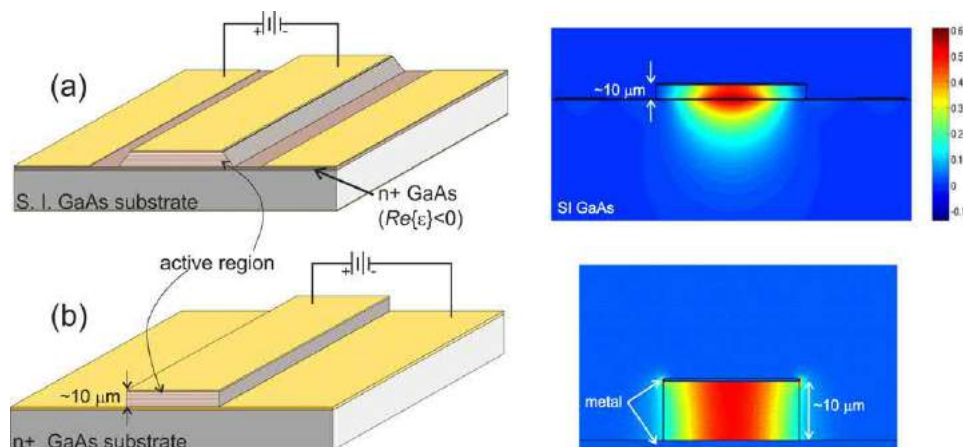


Fig. 5 Schematic of terahertz QC-laser waveguides. Schematic diagram (left) and typical two-dimensional mode intensity pattern (right) for (a) surface plasmon waveguides and (b) metal-metal waveguides. $\text{Re}(\epsilon) < 0$ indicates that the real part of the permittivity is less than zero in the heavily doped GaAs laser. Figure from Williams, B.S. 2007. Terahertz quantum cascade lasers. *Nat. Photon.* 1, 517–524.

An alternative to the SP waveguide is the so-called “metal-metal” (MM) waveguide, in which the waveguide mode is tightly confined between metal cladding placed immediately above and below the $\sim 10\ \mu\text{m}$ thick epitaxial active region. The overall result resembles a microstrip transmission line. MM waveguides tend to have the best high-temperature performance, mostly as a result of lower overall losses (both absorption and radiative) compared to the SP waveguide. Furthermore, the strong modal confinement of MM waveguides allows both the vertical and lateral dimensions to be made much smaller than the wavelength. This feature is unique for a new genre of lasers termed “photonic wire lasers”, in which a large fraction of mode propagates outside of the solid core. This in turn reduces the total thermal dissipation and required cooling power, which was key to observation of the highest temperature cw operation (129 K) in a metal-metal waveguide (Wienold *et al.*, 2014). Furthermore, it is possible to mechanically perturb the portion of the mode outside the core using MEMS actuators, which has allowed broad single-mode tuning of up to 330 GHz (Qin *et al.*, 2011).

While metal-metal waveguides are preferred in terms of temperature performance, if a MM ridge waveguide is simply cleaved to form a Fabry-Pérot cavity, it performs poorly as an edge emitting laser. The cleaved facet radiates as a sub-wavelength sized aperture, which exhibits an extremely divergent beam with a poor radiation efficiency. Hence, an active area of research has emerged in the design of novel cavity types to allow high quality beams simultaneously with high output power.

A straightforward approach is to mount a silicon lens flush to the metal-metal waveguide facet, which helps to improve the beam quality and improve the impedance matching. Another class of approaches uses Bragg scattering from periodic structures to reshape the beam. Second-order distributed feedback (DFB) and 2D photonic crystal cavities have demonstrated surface emission; the beam is much improved since the waveguide surface, not the facet, is now the radiating aperture. One challenge that emerges is that the DFB laser prefers to lase in the mode with the lowest total losses, which usually is a weakly radiating (high-Q) band-edge mode with low output power. A variety of schemes are under investigation, including graded photonic heterostructures that force the laser to lase in the strongly radiating symmetric band-edge mode, and other structures (e.g. quasi-periodic crystals, dual slit gratings, and others) that break the lattice symmetry to increase radiative efficiency.

End-fire QC-lasers based upon 3rd-order DFBs or antenna-coupled DFB cavities are another attractive option; they can achieve very narrow far field beams even when the transverse waveguide cross section is sub-wavelength (Amanti *et al.*, 2009). The beam is formed by a linear phased array of scatters along the length of the cavity; the beam divergence scales inversely with square root of the cavity length for a perfectly phase-matched 3rd-order DFB laser. Hence, this allows narrow photonic wire cavities to be used, which further keeps the power dissipation low for good cw operation. Slope efficiencies can be made large by including antenna-like structures into the cavity. These approaches are particularly effective for milliwatt-scale power output.

Still another class of approaches involves the phase locking of multiple smaller emitting elements to achieve high quality beams and to scale up the power by power combining. While arrays of 2nd order DFBs have been demonstrated, it is challenging to phase lock more than a few elements. Two new approaches address this issue. One is to use radiative coupling to couple large arrays of sub-wavelength sized QC-laser cavities – for appropriate lattice spacings the individual cavities will prefer to lase in an in-phase collective supermode that provides a directive high quality beam with high slope efficiency (Kao *et al.*, 2016). If this array is biased slightly below threshold, it can be used as a reflective THz amplifier. The other approach is similar, but deliberately uses arrays of sub-wavelength cavities with low-quality factor (so they do not self-oscillate) to form a reflective amplifying “metasurface”. When this metasurface is placed into an external cavity, these emitters lock to the cavity mode (Xu *et al.*, 2015). Such a device, known as a metasurface vertical-emitting-cavity surface-emitting-laser (VECSEL), not only allows scalable power with a high-quality beam, but detailed control over the phase, spectral, and polarization response of the metasurface.

THz QC-Laser Frequency Combs

A rapidly evolving area of current research is the development of broadband, multi-mode THz QC-lasers that operate as frequency combs (Burghoff *et al.*, 2014). QC-lasers can be designed to exhibit gain over very large fractional gain bandwidths. This feature emerges naturally because certain designs of active region exhibit emission from multiple broadened transitions; furthermore, the active region can be made deliberately inhomogeneous by sandwiching multiple stacks, each intended to emit at different wavelengths (Rösch *et al.*, 2015). If the intracavity dispersion is properly managed (so that each longitudinal mode sees the same round trip time), then the multiple modes will interact via four-wave-mixing to establish a phase coherent comb. This is a different mechanism than the traditional method of comb generation based upon ultrafast mode-locked lasers - instead of a circulating intracavity pulse, the intensity inside the laser cavity is much closer to a constant value.

Perhaps the largest application is rapid, high-precision dual-comb spectroscopy, in which two combs (signal and local oscillator) with slightly detuned comb tooth spacings are mixed. This technique effectively mimics the functionality of Fourier-Transform Spectroscopy, but without the need for a mechanically moving mirror. The best reported comb bandwidths are approximately 1 THz, which still falls short of the octave spanning bandwidth needed to allow f-2f self-stabilization of the comb. For this reason, a major focus of current research is the effort to increase the comb bandwidth, by engineering both the active region and the cavity dispersion, in order to increase the gain bandwidth while simultaneously managing dispersion.

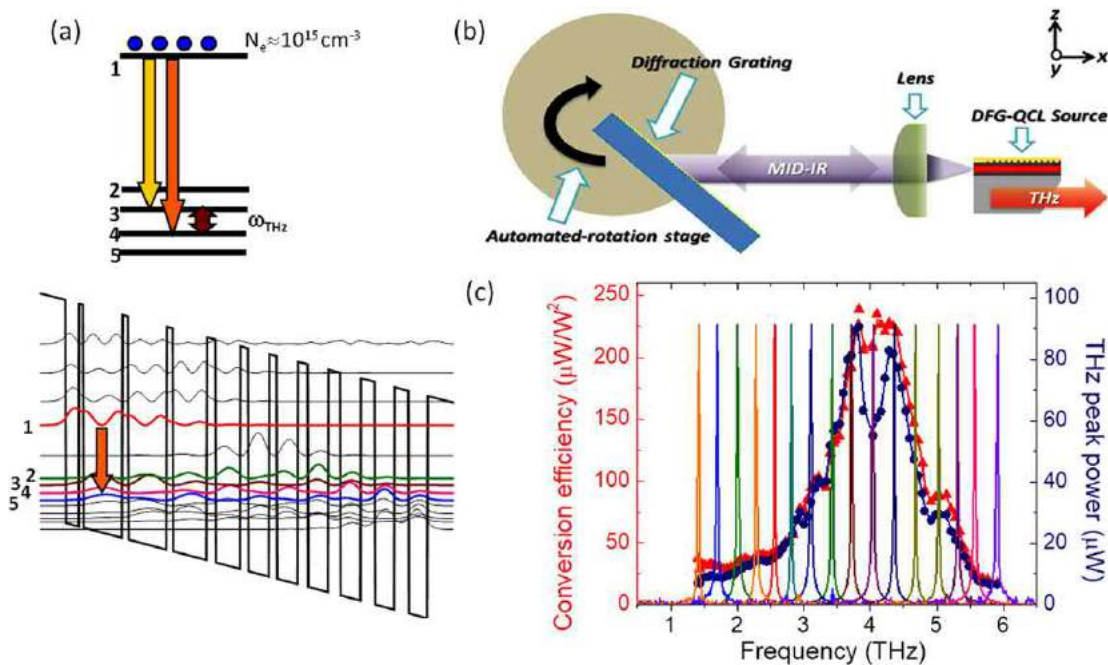


Fig. 6 (a) Schematic of mid-IR QC-laser active region band diagram, indicating the levels primarily involved with the DFG process. (b) Schematic of external cavity setup for tunable THz DFG. (c) Spectra of THz DFG over tuning range. Figure adapted from Vijayraghavan, K., Jiang, Y., Jang, M. *et al.*, 2013. Broadly tunable terahertz generation in mid-infrared quantum cascade lasers. *Nat. Commun.* 4, 2021.

Nonlinear Generation of THz Radiation in Mid-IR QC-Lasers

There exists a separate approach for generating THz radiation in QC-lasers that uses nonlinear downconversion from shorter wavelengths rather than direct lasing. Specifically, room-temperature THz radiation has been generated using intracavity difference frequency generation (DFG) within mid-infrared QC-lasers lasing at two wavelengths (Belkin *et al.*, 2007). Generation of THz radiation in nonlinear crystals via downconversion has a long history - however usually the limited strength of the available $\chi^{(2)}$ nonlinearity has required large laser intensities; as a result these techniques have required high power pulsed lasers (both Q-switched and mode-locked) in the visible and near-IR.

Mid-IR QC-lasers with wavelengths near 10 μm are more attractive, since the smaller pump photon energy results in an improved Manley-Rowe ratio, and the lasers themselves are physically smaller. Furthermore, the quantum-cascade laser material itself can be engineered to possess an extremely large resonant $\chi^{(2)}$ nonlinearity based upon the mid-IR intersubband transitions (Fig. 6). Ordinarily, resonant nonlinearities are accompanied by large linear absorption losses, but because this nonlinearity is associated with the same intersubband transitions that produce the mid-IR laser gain, the resonant absorption is avoided. This allows efficient mixing to take place within the laser cavity, where the circulating intensities are high. This approach leverages the greater technological maturity of mid-IR lasers in the 4–10 μm range, where room-temperature continuous-wave operation is possible with hundreds of mW or more.

While the initial demonstrations of THz DFG had very low conversion efficiency and sub-microwatt pulsed powers, significant advances have taken place recently. Currently, the highest reported THz power levels at room-temperature are 1.9 mW of peak power in pulsed mode, and 14 μW of power in cw mode (Lu and Razeghi, 2016). These sources are particularly promising for applications which require broad tunability, since only modest tuning of one of the mid-IR modes translates to large tuning of the THz beat note. For example, tuning of a single mode from 1.7–5.3 THz was shown in (Vijayraghavan *et al.*, 2013). Work is underway to further increase the nonlinear conversion efficiency, improve the efficiency of out-coupling the THz radiation from the cavity, as well as the increase the range and ease of tunability.

See also: Broadband Terahertz Sources. THz Molecular Spectroscopy

References

- Amanti, M.I., Fischer, M., Scalari, G., Beck, M., Faist, J., 2009. Low-divergence single-mode terahertz quantum cascade laser. *Nature Photon.* 3, 586–590.
 Belkin, M.A., Capasso, F., Belyanin, A., *et al.*, 2007. Terahertz quantum-cascade-laser source based on intracavity difference-frequency generation. *Nature Photon.* 1, 288–292.

- Blom, A., Odnoblyudov, M.A., Cheng, H.H., Yassievich, I.N., Chao, K.A., 2001. Mechanism of terahertz lasing in SiGe/Si quantum wells. *Appl. Phys. Lett.* 79, 713.
- Bründermann, E., Chamberlin, D.R., Haller, E.E., 2000. High duty cycle and continuous terahertz emission from germanium. *Appl. Phys. Lett.* 76, 2991.
- Burghoff, D., Kao, T.-Y., Han, N., *et al.*, 2014. Terahertz laser frequency combs. *Nat. Photon.* 8, 462–467.
- Chang, T.Y., Bridges, T.J., 1970. Laser action at 452, 496, and 541 μm in optically pumped CH_3F . *Opt. Commun.* 9, 423–426.
- Crocker, A., Gebbie, H.A., Kimmitt, M.F., Mathias, L.E.S., 1964. Stimulated emission in the far infra-red. *Nature* 201, 250–251.
- Dodel, G., 1999. On the history of far-infrared (FIR) gas lasers: thirty-five years of research and application. *Infrared Phys. Technol.* 40, 127–139.
- Farhoomand, J., Blake, G.A., Frerking, M.A., Pickett, H.M., 1985. Generation of tunable sidebands in the far-infrared region. *Proc. SPIE* 598, 84–87.
- Golka, S., Pflügl, C., Schrenk, W., Strasser, G., 1991. Special issue – Far-infrared semiconductor lasers. *Opt. Quantum Electron.* 23, S111.
- Gousev, Y.P., Altukhov, I.V., Korolev, K.A., *et al.*, 1999. Widely tunable continuous-wave THz laser. *Appl. Phys. Lett.* 75, 757.
- Hovenier, J.N., Diez, M.C., Klassen, T.O., *et al.*, 2000. The p-Ge terahertz laser properties under pulsed- and mode-locked operation. *IEEE Trans. Microwave Theory Tech.* 48, 670.
- Hübers, H.-W., Pavlov, S.G., Shastin, V.N., 2005. Terahertz lasers based on germanium and silicon. *Semicond. Sci. Technol.* 20, S211–S221.
- Inguscio, M., Moruzzi, G., Evenson, K.M., Jennings, D.A., 1986. A review of frequency measurements of optically pumped lasers from 0.1 to 8 THz. *J. Appl. Phys.* 60, R161.
- Jacobsson, S., 1989. Optically pumped far infrared lasers. *Infrared Phys* 29, 853–874.
- Kagan, M.S., Altukhov, I.V., Sinis, V.P., *et al.*, 2000. Terahertz emission of SiGe/Si quantum wells. *Thin Solid Films* 380, 237.
- Kao, T.-Y., Reno, J.L., Hu, Q., 2016. Phase-locked laser arrays through global antenna mutual coupling. *Nat. Photon.*
- Köhler, R., Tredicucci, A., Beltram, F., *et al.*, 2002. Terahertz semiconductor-heterostructure laser. *Nature* 417, 156.
- Kumar, S., 2011. Recent progress in terahertz quantum cascade lasers. *IEEE J. Sel. Top. Quantum Electron.* 17, 38–47.
- Lu, Q., Razeghi, M., 2016. Recent advances in room temperature, high-power terahertz quantum cascade laser sources based on difference-frequency generation. *Photonics* 3, 42.
- Mueller, E.R., Henschke, R., Williams, E., *et al.*, 2007. Terahertz local oscillator for the Microwave Limb Sounder on the Aura satellite. *Appl. Opt.* 46, 4907–4915.
- Pavlov, S.G., Zhukavin, R.K., Shastin, V.N., Hübers, H.-W., 2013. The physical principles of terahertz silicon lasers based on intracenter transitions. *Phys. Status Solidi B* 250, 9–36.
- Qin, Q., Reno, J.L., Hu, Q., 2011. MEMS-based tunable terahertz wire-laser over 330GHz. *Opt. Lett.* 36, 692–694.
- Rochat, M., Ajili, L., Willenberg, H., *et al.*, 2002. Low-threshold terahertz quantum-cascade lasers. *Appl. Phys. Lett.* 81, 1381.
- Rösch, M., Scalari, G., Beck, M., Faist, J., 2015. Octave-spanning semiconductor laser. *Nat. Photon.* 9, 42–47.
- Vijayraghavan, K., Jiang, Y., Jang, M., *et al.*, 2013. Broadly tunable terahertz generation in mid-infrared quantum cascade lasers. *Nat. Commun.* 4, 2021.
- Vitiello, M.S., Scalari, G., Williams, B.S., Denatale, P., 2015. Quantum cascade lasers: 20 years of challenges. *Opt. Express* 23, 5167–5182.
- Wade, A., Federov, G., Smirnov, D., *et al.*, 2009. Magnetic-field-assisted terahertz quantum cascade laser operating up to 225 K. *Nat. Photon.* 3, 41–45.
- Wienold, M., Röben, B., Schrottke, L., *et al.*, 2014. High-temperature, continuous-wave operation of terahertz quantum-cascade lasers with metal-metal waveguides and third-order distributed feedback. *Opt. Express* 22, 3334–3348.
- Williams, B.S., 2007. Terahertz quantum cascade lasers. *Nat. Photon.* 1, 517–524.
- Williams, B.S., Callebaut, H., Kumar, S., Hu, Q., Reno, J.L., 2003. 3.4-THz quantum cascade laser based on longitudinal-optical-phonon scattering for depopulation. *Appl. Phys. Lett.* 82, 1015.
- Xu, L., Curwen, C.A., Hon, P.W.C., *et al.*, 2015. Metasurface external cavity laser. *Appl. Phys. Lett.* 107, 221105.

THz Molecular Spectroscopy

Zbigniew Kisiel, Institute of Physics of the Polish Academy of Sciences, Warsaw, Poland

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The THz region is nominally defined by the International Telecommunication Union (ITU) as covering 0.3–3 THz. It has been exploited by several branches of molecular spectroscopy, each of which describes it in terms of its customary units of either frequency, wavelength or wavenumber. **Table 1** provides a convenient summary.

The most useful and, at the same time, the most challenging molecular spectroscopy application of the THz region is high resolution spectroscopy in the gas phase. This takes advantage of the narrow linewidths of molecular absorption lines and of the ability of contemporary instrumental techniques to deliver molecular spectra, which are only limited by molecular properties and not by instrumental factors. The apparent widths of room temperature lines free from broadening effects are at the MHz level for THz frequency transitions, so that frequency control of significantly better than one part in 10^7 is mandatory. Absorption coefficients of such transitions are usually smaller or much smaller than 10^{-3} per cm of absorption path so that considerable source stability and multiple techniques for enhancing the recovered signal are also required.

Rotational Spectroscopy

The principal physical phenomenon observed in THz molecular spectra is quantized molecular rotation. In the microscopic world rotational energy can only take on very well defined values, or energy levels. Those are labeled by using quantum numbers corresponding to the total molecular angular momentum and to its projection onto a selected axis in the molecule. The rotational energies also depend on the molecular moment of inertia, which is expressed in terms of its components I_a , I_b , I_c along a unique set of center of mass Cartesian axes, called principal axes a , b , c . Transitions between rotational energy levels are allowed if the molecule possesses a permanent electric dipole moment, μ , and then its components μ_a , μ_b , μ_c , along the three principal axes are of relevance. Non-zero value of each such component leads to specific selection rules, or allowed changes in quantum number values for absorption of electromagnetic radiation inducing a change in rotational energy. Of relevance to all types of rotational spectra is the fact that frequencies of transitions depend directly on rotational constants, A , B , C , which are inverse quantities to the three principal moments of inertia. Since the moments of inertia only depend on atomic masses and on their positions in the molecule, their spectroscopic determination provides a route to precise experimental molecular structures. Furthermore, rotational transitions of a given molecule form a very characteristic and precisely measurable fingerprint pattern. This enables various analytical applications, as the high specificity allows unambiguous determination of molecular compositions in complex gas phase mixtures. This is actually the only method for identifying and quantifying the presence of molecules in the interstellar medium, currently at 200 or so species. Precise measurement of line intensities also allows quantitative determination of molecular abundances.

There are two limiting spectroscopic molecular rotation cases, depending on the general molecular shape: whether it is close to a cylinder (prolate rotor) or to a disk (oblate rotor). Rotational energies (in frequency units) are then approximately given by:

$$\text{prolate rotor } (A > B = C): \quad E_{JK} = BJ(J+1) + (A-B)K^2 + \text{smaller terms} \quad (1)$$

$$\text{oblate rotor } (A = B > C): \quad E_{JK} = BJ(J+1) + (B-C)K^2 + \text{smaller terms} \quad (2)$$

where J and K are the quantum numbers corresponding to the total molecular angular momentum and to its projection onto the molecular symmetry axis, respectively. The selection rules in both limiting cases are identical, being $\Delta J = 1$ and $\Delta K = 0$, and leading to a deceptively simple expression for transition frequencies:

$$\nu = 2B(J+1) + \text{smaller terms} \quad (3)$$

Most molecules are actually asymmetric rotors, with all three rotational constants different. Their quantum mechanical description uses the values of K for the two limiting symmetric top cases, renamed K_a (prolate) and K_c (oblate) to label asymmetric

Table 1 Limits of the THz region in units used in various branches of molecular spectroscopy

	Lower bound	Upper bound	Unit
Frequency	0.3	3	THz
	300	3000	GHz
Wavelength	1	0.1	mm
	1000	100	μm
Wavenumber	10	100	cm^{-1}

rotor energy levels, by means of the notation J_{KaKc} . The frequencies of the most relevant transition types can also be usefully predicted by replacing $2B$ in Eq. (3) with $(B + C)$ for a prolate asymmetric rotor or by $(A + B)$ for an oblate one. In practice, precise reproduction of measured transition frequencies requires also inclusion of terms from several other effects, such as distortion of molecular structure on rotation (centrifugal distortion), or splitting due to the presence in the molecule of atoms with non-zero nuclear quadrupole (nuclear quadrupole hyperfine splitting).

From the point of view of THz spectroscopy disk type asymmetric rotors are less relevant due to a more compressed energy level structure as implied by Eq. (2). On the other hand, prolate rotors can have sizable values of the A rotational constant, and Eq. (1) allows a much faster increase in rotational energies with increasing values of quantum number K_a , leading to significant numbers of THz region transitions. This is also partly due to basic spectroscopic properties associated with the dipole moment components responsible for the observed transitions. The selection rules leading to the K independent transitions described by Eq. (3) arise from the non-zero dipole moment component directed close to the symmetry axis of the molecule (μ_a for a prolate, μ_c for an oblate rotor). In an asymmetric rotor the third, μ_b , dipole moment component is possible and is associated with rather more complex selection rules for transitions, which introduce frequency dependence on the K_a quantum number, and thus directly on the value of the potentially sizable A rotational constant. In this way the $\mu_b \neq 0$ case can give rise to plentiful b -type rotational transitions in the THz.

As a boundary example: in the very light water molecule μ_b is the only non-zero dipole moment component and the rotational constants are $A=835.8$, $B=435.3$, $C=278.1$ GHz. In consequence, only two significant rotational transitions of water, $6_{16} \leftarrow 5_{23}$ at 22.235 GHz and $3_{13} \leftarrow 2_{20}$ at 183.410 GHz, fall below the THz region, whereas the actual water rotational spectrum extends through the THz region and beyond. While this is an important case, it is not typical, as rotational constants for molecules with significant THz region spectra are most often in the 1–10 GHz region. A more representative example is the relatively rigid acrylonitrile molecule, $\text{CH}_2=\text{CHC}\equiv\text{N}$, which gives rise to the THz spectrum displayed in Fig. 1. The rotational constants in this case are $A=49.95$, $B=4.97$, $C=4.51$ GHz and lead to plentiful transitions allowed by non-zero μ_a and μ_b dipole moment components, due to the orientation of the molecular dipole moment shown in Fig. 2. At 1 THz the b -type transitions dominate and comprise of bR -branch ($\Delta J=1$) and bQ -branch ($\Delta J=0$) transitions. While it is not possible to derive simple expressions for frequencies of bR -branch transitions those involve an $(A+B)(J+1)$ leading term. A more successful approximate relation for the dense bands made up of bQ -branch transitions is:

$$\nu = [A - (B + C)](2K_a + 1) + \text{smaller terms} \quad (4)$$

Of course, an appropriate treatment of the problem with a suitable quantum mechanical Hamiltonian allows reproduction of experimental frequencies to measurement accuracy, but this is the domain of dedicated computer programs. Nevertheless, it is easy to see from the above how a sizable value of the rotational constant A can result in a rich THz region spectrum. The relative intensities of a -type and b -type transitions and their temperature dependence are illustrated in Fig. 2. The a -type transitions lose relevance in relation to b -type transitions simply because at a given THz frequency the former involve a much greater value of J .

THz region rotational transitions in excited vibrational states of molecules, can also be of appreciable intensity, as has already been seen in Fig. 1. The intensities of such transitions will be reduced relative to those in the ground state by the standard Boltzmann population factor of $e^{-\Delta E/kT}$. At room temperature, rotational transitions in an excited vibrational state positioned 100 cm^{-1} above the ground state will have over 60% of the intensity of the ground state transitions, while the $\nu_{11}=1$ state lines in

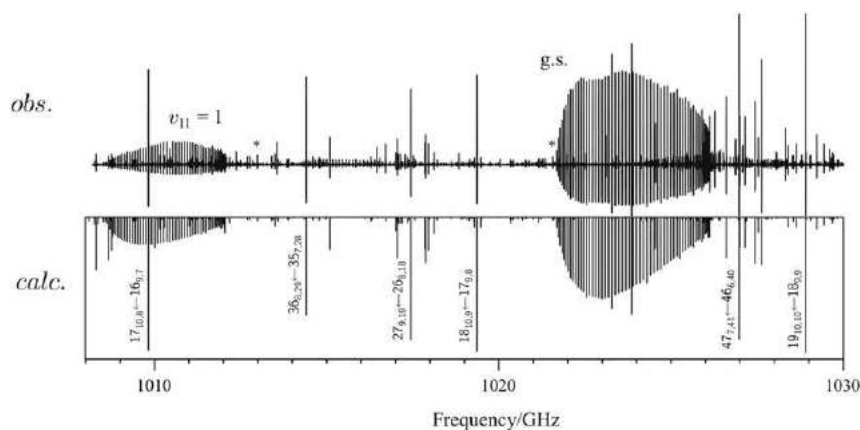


Fig. 1 The room temperature spectrum of acrylonitrile recorded at 1 THz. It consists almost entirely of b -type rotational transitions in the ground state (g.s.) and in the lowest excited vibrational states (such as the marked $\nu_{11}=1$). Quantum numbers for the strongest bR -branch ($\Delta J=1$) transitions are indicated on the figure, while the two bands of dense lines are bQ -branch transitions ($\Delta J=0$) for $K_a=11 \leftarrow 10$. The strongest a -type transitions in this region are marked with asterisks. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. *Journal of Molecular Spectroscopy* 258, 26–34.

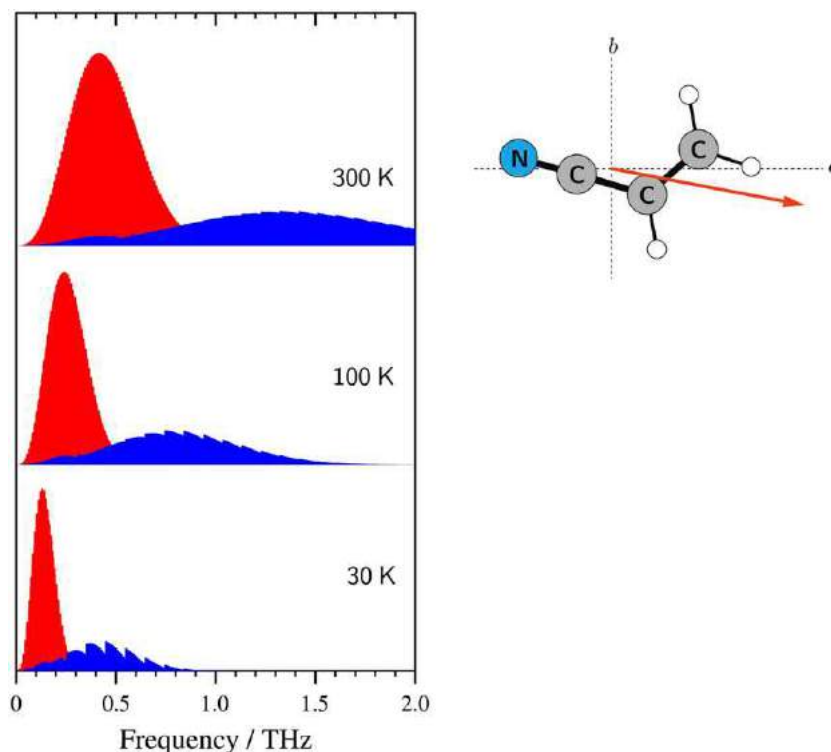


Fig. 2 The nature of the THz rotational spectrum of acrylonitrile and its temperature dependence. The dipole moment of the molecule (with orientation depicted by the arrow) is made up of a large μ_a component giving rise to stronger but lower frequency transitions (red bands), and a smaller μ_b component resulting in weaker but higher frequency transitions (blue bands). At the lower, astrophysical temperatures, the nominally weaker b -type transitions dominate above 0.5 THz. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. *Journal of Molecular Spectroscopy* 258, 26–34.

Fig. 1 originate from a state with vibrational energy of 228 cm^{-1} and are consequently at 33% of ground state intensity. The situation becomes more complex when the molecule has a clearly defined internal rotor, which can have several equivalent orientations relative to the remaining molecular frame. The most common case is that of a C_{3v} -symmetry internal rotor (typically a CH_3 group) in a C_s -symmetry molecule (containing a symmetry plane). The existence of three equivalent minima possible on internal rotation (torsional motion) of the methyl group gives rise to a splitting of the ground state, and of each excited vibrational state, into two substates, called A and E after the C_3 group classification of such motion. Separate rotational level stacks are possible for each substate, giving rise to two sets of rotational transitions of comparable intensity. For a sizable internal rotation barrier the frequencies of A and E-substate transitions are not too different resulting in characteristic A–E doubling of lines in the spectrum. When the internal rotation barrier is lower the vibrational E–A splitting increases and for the important methanol molecule is close to 9 cm^{-1} . Since methanol is a rather light molecule with comparable masses of the internal rotor and of the OH frame the frequencies of A and E ground state transitions become much more difficult to analyse.

Rotation Vibration Spectroscopy

Pure rotation is not the only phenomenon of relevance to high resolution THz molecular spectroscopy. A related one is vibrational spectroscopy with resolved rotational structure, in the form of transitions between rotational level stacks belonging to different vibrational states. Typical vibrational energy level differences are in excess of 100 cm^{-1} so that transitions between vibrational states with different vibrational quantum numbers typically fall in the infrared region and are relatively rare in the THz. But there are special cases, one of which is an important class of molecules endowed with a relatively low-barrier inversion motion. When two symmetry equivalent orientations of the molecule are possible then quantum tunneling through the separating barrier leads to splitting of each vibrational state of the inversion motion into two components. In this way vibrational state $v=0$ is split into 0^+ and 0^- substates, $v=1$ into 1^+ and 1^- substates, etc. Rotation-vibration transitions between a pair of such substates are allowed by a vibrational dipole moment along the inertial axis connecting the two minima. The most celebrated case of this type is the inversion splitting spectrum between the 0^+ and 0^- substates of the ammonia molecule, which consists of a multitude of transitions centered at 23.8 GHz. This was the first microwave molecular spectrum observed as far back as the 1930s, and caused some confusion since it was not a pure rotation spectrum. The tunneling splitting between the next higher, 1^+ and 1^- , inversion doublet of ammonia is 1056 GHz so that the respective rotation-vibration transitions fall in that region and have been studied by THz spectroscopy.

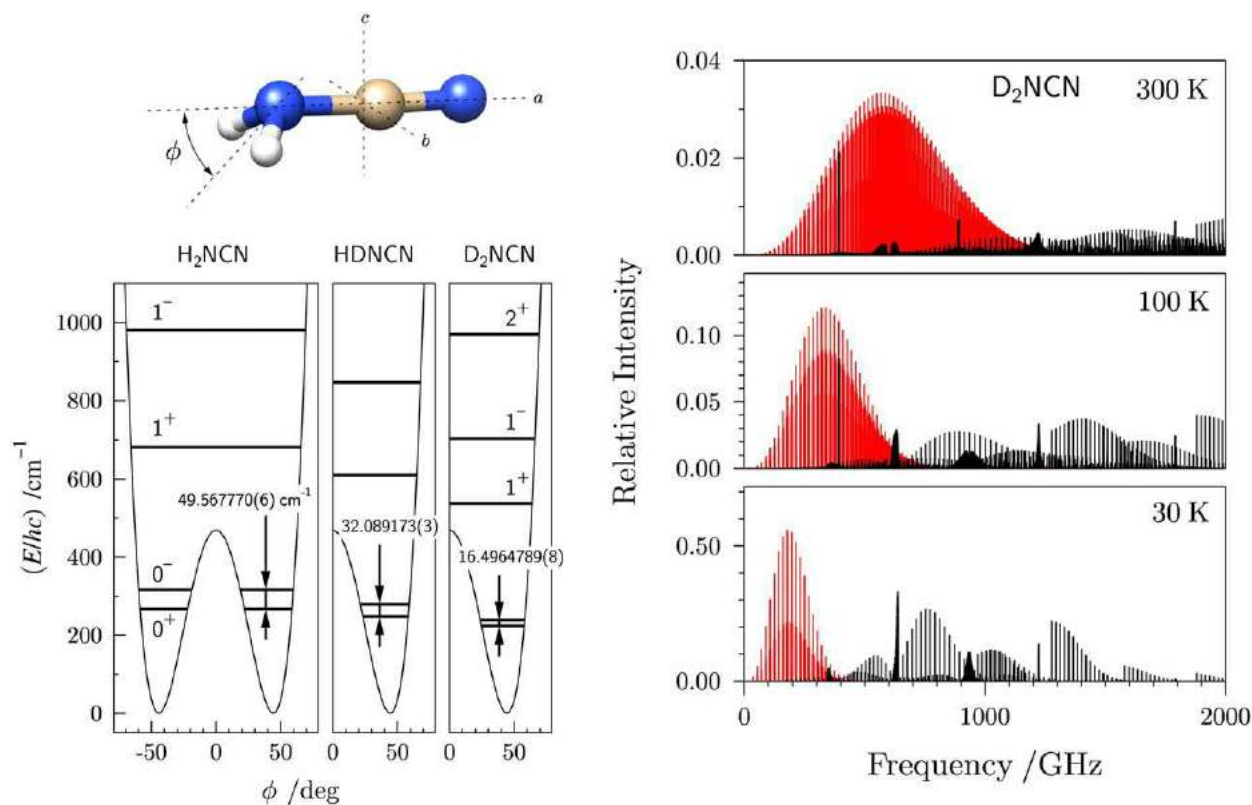


Fig. 3 The origin and temperature dependence of THz transitions for the cyanamide molecule. The pure rotation spectrum (red in the right panel for D_2NCN) arises from the large permanent μ_a dipole moment component, while higher frequency lines (black in the right panel) are due to c -type rotation-vibration transitions between the two inversion substates, 0^+ and 0^- , of the ground state. Global analysis of the available transitions allowed precise determination of inversion splitting (marked in the left panel) illustrating its considerable isotope dependence. Reproduced from Kisiel, Z., Kraśnicki, A., Jabs, W., *et al.*, 2013. Rotation and rotation-vibration spectroscopy of the $0^+ - 0^-$ inversion doublet in deuterated cyanamide. *Journal of Physical Chemistry A* 117, 9889–9898 with permission.

A related molecule to ammonia is cyanamide, $H_2NC\equiv N$, in which the inversion motion at the pyramidal nitrogen nucleus is characterized by a significantly lower barrier of below 500 cm^{-1} . In consequence, the $0^+ - 0^-$ inversion splitting ranges from 0.495 to 1.486 THz (Fig. 3) and creates ample opportunities for THz molecular spectroscopy. Furthermore, cyanamide is a light molecule so that its pure rotation spectrum is intertwined with the vibration-rotation spectrum, even for the relatively heavy $D_2NC\equiv N$ isotopologue, as shown in Fig. 3.

Laboratory Molecular Spectroscopy

The ability to record THz molecular spectra in the laboratory is critically determined by the techniques used to generate highly controlled electromagnetic radiation in this frequency region. Generation needs to be complemented by suitable methods of detection and, in an associated step, by efficient extraction of molecular signals. Various modulation techniques are employed for the latter, generally making use of the narrow natural width of molecular absorption lines. Linewidths have a complex dependence on quantum numbers of the involved transitions but can, in general, be regarded to range from 1 MHz full-width at half height (FWHH) at the lower end of the THz region, to 10 MHz FWHH at the upper end. Several influential techniques of laboratory spectroscopy are described below, chosen for their current relevance, or promise for further molecular spectroscopy applications.

Sources

There are many sources that give access to the THz frequency region, but molecular spectroscopy requires high quality performance on routine operation. There is considerable premium on resolution and on broadband access, namely recording of spectra over considerable frequency regions. The spectra contain many features that are not amenable to direct interpretation, and can often be satisfactorily understood only upon a protracted off-line analysis. The highest resolution is possible with tunable monochromatic radiation, although promising time domain microwave techniques are appearing. Somewhat lower resolution, but very broad

Table 2 THz region coverage available with the cascaded harmonic multiplication sources used at the Jet Propulsion Laboratory^a

Synthesizer frequency (GHz)	N ^b	Output frequency (GHz)	Synthesizer frequency (GHz)	N	Output frequency (GHz)
16.1–17.8	18	290–320	17.6–20.0	54	950–1080
13.0–17.7	30	390–530	14.7–17.2	72	1055–1240
14.8–17.0	36	534–612	12.1–13.2	108	1305–1425
12.6–14.8	54	680–800	13.1–14.8	108	1420–1600
12.8–14.2	60	770–850	14.8–17.1	108	1600–1850
11.8–12.9	72	850–930	16.5–18.5	108	1780–2000
			15.2–17.2	162	2470–2780

^aYu, S. Personal communication.^bThe total multiplication factor of the frequency of the driving microwave synthesizer.

frequency coverage is also possible with Fourier transform far-infrared interferometers, especially those using synchrotron radiation sources.

High-resolution spectroscopic access to the THz region was pioneered in the 1950s by the research group of Gordy at Duke University, Durham, United States (see Chapter 2.1 of Rao (1976) for a review), who carried out many seminal studies including those of the many isotopologues of the water molecule. They used point contact diode multiplication of radiation from microwave klystrons and point contact diode detectors, a technology that was adopted in many other laboratories. The use of klystrons and of oscilloscope based detection limited the spectroscopic access to relatively narrow windows. Nevertheless, there are similarities between this approach and the technique that has over the recent years turned out to be the optimum source for routine THz molecular spectroscopy. This is cascaded harmonic multiplication of microwave radiation, in which a 20 GHz region microwave synthesizer drives a stack of active and passive microwave multipliers. The principal advantage is that it is left to the synthesizer to provide frequency stabilization and tunability, which is then simply transferred to THz frequencies. In practice, years of development have gone into optimizing the multiplier combinations, and into ensuring harmonic purity. The broadest coverage has been developed at the Jet Propulsion Laboratory (JPL), Pasadena, United States (Drouin *et al.*, 2005) as summarized in Table 2. Usable output powers range from the mW level at the low frequency end to sub- μ W level at the highest frequency bands. The total multiplication factor of the driving synthesizer frequency is the product of the multiplication factors of the individual multipliers in the stack and, for example, $18 = 6 \times 3$ for the lowest frequency source in Table 2, and $108 = 6 \times 2 \times 3 \times 3$ for the 1.3–2.0 THz sources. This is a reasonably mature technology and cascaded multiplication sources in many configurations are available commercially.

A very attractive class of sources are high frequency backward wave oscillators (BWO). These are vacuum tubes, developed in the Soviet Union, and revealed to the spectroscopic community in the 1970s in publications of the research group at the Institute of Applied Physics, Nizhny Novgorod, Russia (then Gorky, Soviet Union) (reviewed in Chapter 2.2 of Rao (1976)). These sources generate THz radiation directly, have relatively high output power, excellent electronic tunability, and linewidth estimated at less than 100 Hz. The drawbacks are that high voltage and a sizable magnetic field are required for operation, and it is necessary to provide external frequency stabilization by means of phase locking techniques. Devices with operation of up to 1.4 THz have been produced and have been used in many spectroscopic laboratories throughout the world. The most prolific adopters have been the research group at the University of Cologne, Germany (as reviewed in Schlemmer *et al.* (2009)) who used such sources in both fundamental generation mode and with harmonic multiplication. The unique properties of these BWOs have also allowed development of several novel spectrometer designs: the fast scan spectrometer at The Ohio State University, Columbus, United States, and the resonator spectrometer for broadband measurements of atmospheric absorption lines at the Nizhny Novgorod laboratory. Both designs are discussed further below.

More specialized, high-resolution access to the THz region is also possible by using difference frequency laser radiation, or laser sideband generation. One such approach has been to use a frequency stabilized far-infrared molecular laser and mix it with a 300 GHz BWO source, by means of a rooftop optical mixer (Schlemmer *et al.*, 2009). This approach allowed μ W level electronically swept frequency operation near the end of the THz region.

Broadband access to most of the THz region has also been possible for a considerable time with far-infrared configurations of Fourier-transform infrared (FTIR) interferometers. The available resolution has been steadily increasing with the increase in optical path difference available from the movable mirror of the interferometer. Currently, resolutions in the region of 0.001 cm^{-1} (30 MHz) are possible. The considerable length of the optical path associated with the very long travel of the movable mirror requires a strong radiation source for sample illumination, so that the majority of such instruments are installed on beamlines of synchrotron radiation sources. A comparison of the 1.8 THz molecular spectrum of the same molecule recorded with microwave techniques and with optical interferometric techniques is shown in Fig. 4. While the advantages of microwave based techniques are clearly apparent it should be borne in mind that FTIR gives access to practically the whole THz region in one continuous spectrum. The difference in resolution and frequency measurement precision between the two traces in Fig. 4 is only an order of magnitude and may in some cases be acceptable.

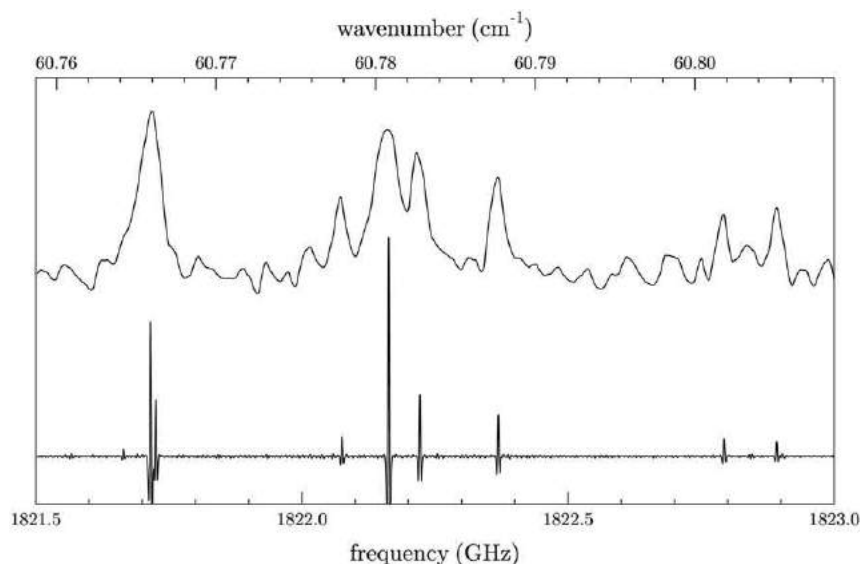


Fig. 4 Comparison of the room temperature THz spectrum of acrylonitrile recorded with microwave multiplication techniques (lower) (Kisiel *et al.*, 2012) and by using a Fourier transform interferometer with an 8.8 m optical path difference installed on the AILES beamline at the SOLEIL synchrotron source (upper) (Kisiel *et al.*, 2015).

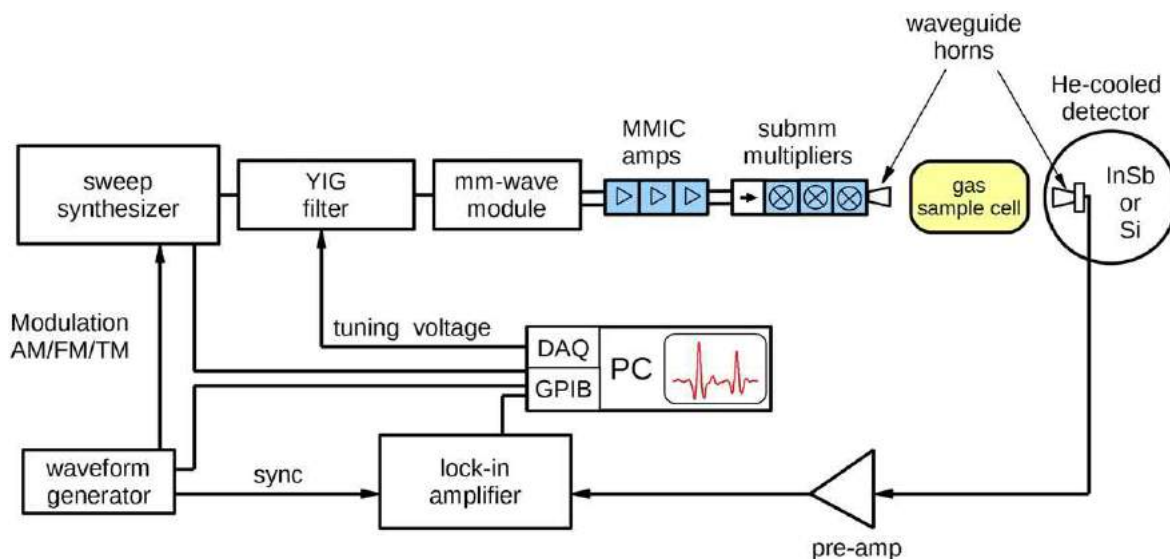


Fig. 5 Schematic diagram of the JPL cascaded harmonic multiplication THz spectrometer, utilizing various types of source modulation and several different detectors. Reproduced from Drouin, B.J., Maiwald, F.W., Pearson, J.C., 2005. Application of cascaded frequency multiplication to molecular spectroscopy. *Review of Scientific Instruments* 76, 093113 with permission.

Detection and the Cascaded Multiplication Spectrometer

The preferred detection method is to use liquid helium cooled bolometers, typically of the Indium Antimonide (InSb) type. A single detector of this type provides coverage of the complete THz region at the sensitivity and bandwidth ample for most studies. Room temperature, zero-bias GaAs Schottky diode detectors are also an attractive alternative for applications below 1 THz, where freedom from the need to use liquid helium offsets the somewhat lower sensitivity.

Detection of molecular signals is intimately associated with various stratagems aiming to increase signal to noise ratio and attempting to eliminate unwanted baseline variation. The most typical experimental setup is an absorption experiment, as illustrated in Fig. 5. THz radiation is passed through an absorption cell containing the gaseous molecular sample, and the transmitted signal is detected. Since molecular absorption coefficients are relatively small the first step is to increase absorption length, by using reasonably long sample cells, typically 1–3 m in length. The most common further technique is electronic source modulation.

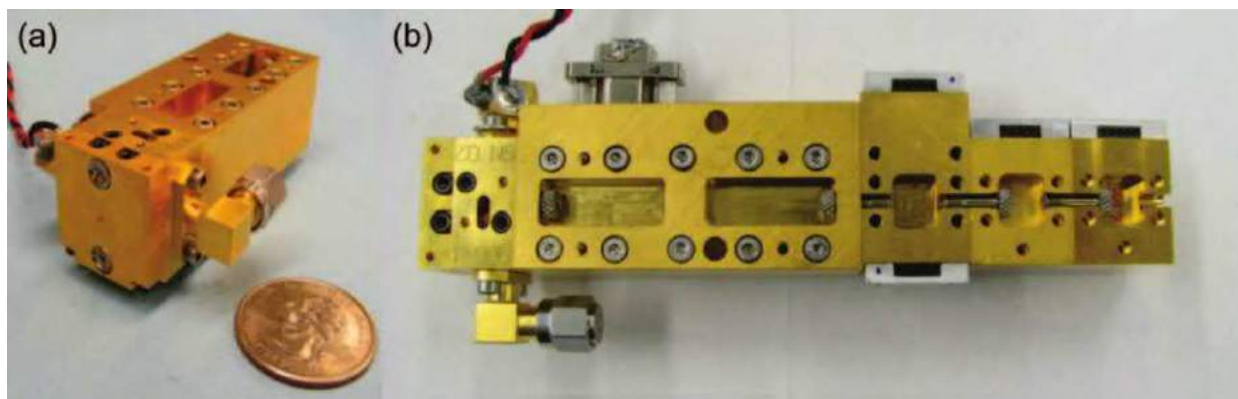


Fig. 6 Illustration of the small size of the 2.7 THz cascaded multiplication source: (a) the three final stage triplers with a 0.69×0.69 mm output port in the center of the output plate (b) the final 27x output multipliers with preceding GaAs and GaN simplifiers. Reproduced from Pearson, J.C., Drouin, B.J., Maestrini, A., *et al.*, 2011. Demonstration of a room temperature 2.48–27.5 GHz coherent spectroscopy source. *Review of Scientific Instruments* 82, 093105 with permission.

The frequency of an otherwise monochromatic source is modulated by an amount comparable to widths of absorption lines. In this way there will be a small alternating signal at the detector in the presence of a molecular absorption line and little or no signal in the absence of a line. A suitable type of phase sensitive detection will then provide separation from the large direct current detector signal proportional to the unabsorbed radiation, at the same time considerably improving the spectroscopic baseline. Two types of source modulation and of associated phase sensitive detection are possible. The first is standard frequency modulation (FM) and detection at twice the frequency f of the rate at which this modulation is applied. In this $2f$ detection the observed line profile is a second derivative of the natural line profile, as in the bottom trace in Fig. 4. This has the added bonus that the width of the central peak of the second derivative of a typical near-Lorentzian absorption profile is close to a third of width of the undifferentiated profile, introducing some useful resolution enhancement. The second source modulation method is to use toneburst modulation. In this case modulation is applied in an on-off mode determined by a square wave of frequency f and phase sensitive detection is also at this frequency. The resulting line profile is also second derivative in appearance, although it is made up of positive and negative lobes comparable in width to the natural waveform. Traditionally, toneburst modulation (TM) was somewhat easier to implement than the FM method for detection of broader THz region, although the distinction has largely disappeared. Finally, in high accuracy lineshape work the signal distortion introduced by source modulation is not acceptable so that amplitude modulation (AM), with either a mechanical chopper or by electronic means, is employed. A complete spectrometer design using a cascaded harmonic multiplication source is shown in Fig. 5. The miniature size of a THz cascaded multiplication source can be assessed from Fig. 6.

The performance of a source modulation cascaded multiplication spectrometer is critically dependent on the available THz power. The power profile of the source is particularly important at higher frequencies, where it results from the operation of many multiplication/amplification stages. The critical dependence of the actual spectral profile on the available power is shown in Fig. 7. The calculated methanol spectrum for the frequency region 2.45–2.75 THz has relatively uniform intensity of the strongest lines, confirming that the intensity profile of the spectrum in Fig. 7 is an instrumental effect due to the source power characteristics shown in the top part of this figure. Nonetheless, Fig. 7 covers a very considerable fragment of the molecular spectrum of methanol recorded at high resolution, while the much narrower 4 GHz width inset displays relative intensities of lines at close to prediction. At this smaller scale other instrumental effects on intensities can also be important. These are standing waves from reflections between the source and the detector, and from unavoidable reflections at all optical surfaces crossed by the radiation, especially the cell windows. If necessary, suitable techniques can be used to rescale the intensity of broadband spectra in order to remove the effect of the source power variation.

The Chirped Pulse THz Spectrometer

Broadband spectroscopy based on chirped pulse excitation has recently revolutionized microwave spectroscopy of supersonic expansion. In this method the sample is first excited by means of a pulse containing an internal frequency sweep and the resulting free induction decay (FID) signal emitted by the sample is then collected and Fourier transformed to give the frequency domain spectrum. High resolution time domain methods based on chirped pulse excitation have now been advanced to the millimeter wave and the THz regions. Generation of THz frequencies is by means of a multiplier chain and detection is by recording the time domain response signal with a fast digital oscilloscope, after downconversion with a similar multiplier chain. Chirped pulse molecular spectroscopy in the THz region has been demonstrated by Neill *et al.* (2013) as shown for methanol in Fig. 8. This spectrum was obtained in a record acquisition time of 58 μ s, and results from 232 sequential measurements. Each measurement took 250 ns and consisted of a 25 ns excitation pulse, followed by 225 ns of acquisition. Each excitation pulse was a chirp covering

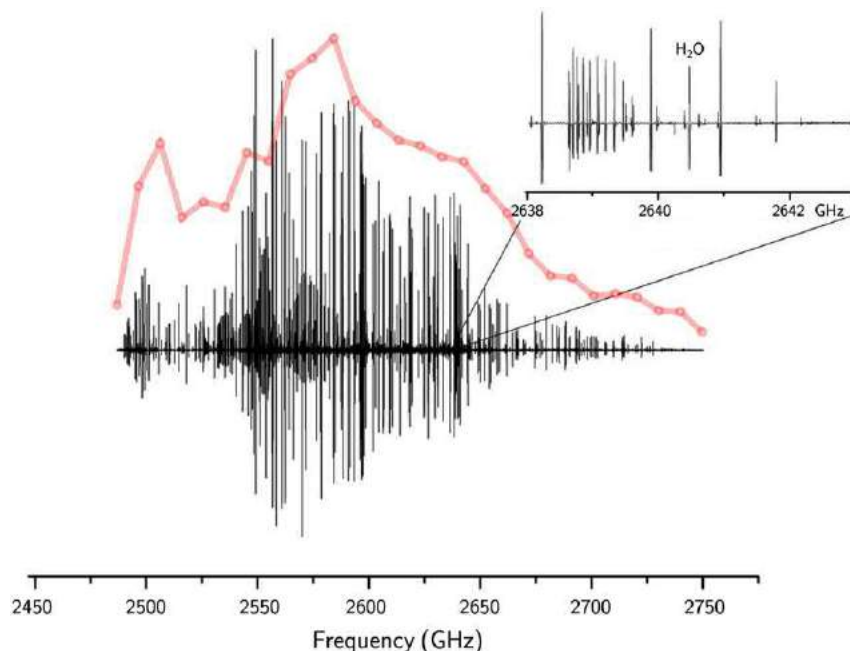


Fig. 7 Illustration of the broad, high resolution coverage available with a single cascaded frequency multiplication source. The spectrum in the lower panel is that of methanol at room temperature and spans 270 GHz at Doppler limited resolution, as shown in more detail in the 4 GHz width inset (reproduced from Pearson, J.C., Drouin, B.J., Yu, S., Gupta, H., 2011. Microwave spectroscopy of methanol between 2.48 and 2.77 THz. *Journal of the Optical Society of America B* 28, 2549–2577). The overall intensity profile is largely determined by the power performance of the source, with typical measured output in the 1–14 μW range plotted in red. The overall intensity profile is largely determined by the power performance of the source, with typical measured output in the 1–14 μW range plotted in red. Reproduced from Pearson, J.C., Drouin, B.J., Maestrini, A., *et al.*, 2011. Demonstration of a room temperature 2.48–27.5 GHz coherent spectroscopy source. *Review of Scientific Instruments* 82, 093105 with permission.

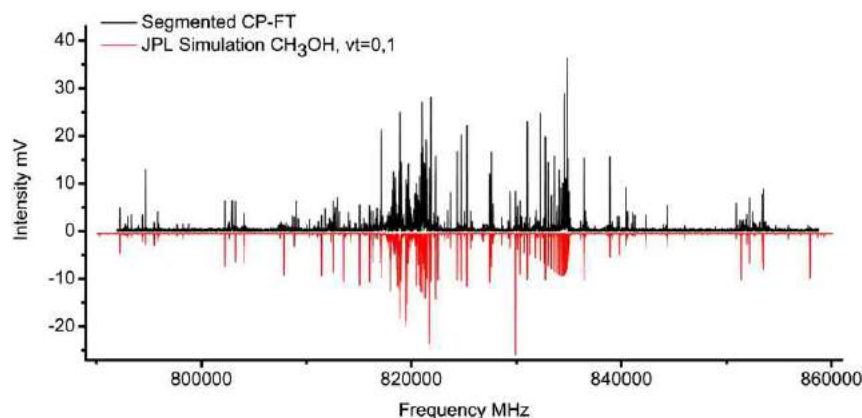


Fig. 8 Example of chirped pulse, Fourier transform access to the THz region. The spectrum is that of methanol and covers 67 GHz at a signal to noise ratio of 120:1. It results from Fourier transform of a single measurement cycle consisting of 232 free-induction decays recorded sequentially over a total 58 μs acquisition time. Reproduced from Neill, J.L., Harris, B.J., Steber, A.L., *et al.*, 2013. Segmented chirped-pulse Fourier transform submillimeter spectroscopy for broadband gas analysis. *Optics Express* 21, 19743–19749 with permission.

288 MHz in the frequency domain. This was obtained by generating a chirped pulse of 4 MHz bandwidth with programmed arbitrary waveform generator, and then subjecting it to 72 times frequency multiplication. The center frequency was ramped for successive pulses in order to provide continuous frequency coverage by means of the 232 FID segments. The technique shows considerable promise, even though some teething problems of spurious signals and additional instrumental line broadening are still to be ironed out. Comparison in **Fig. 8** between the experimental spectrum and the simulation shows that there is still considerable local variation in sensitivity, as apparent in the profile of the Q-branch near 835 GHz, or in the near disappearance of some predicted strong lines. Nevertheless, sufficient system stability for time-domain signal averaging has already been demonstrated and very rapid THz analytical applications appear feasible. Further refinement of such designs and broader adoption may

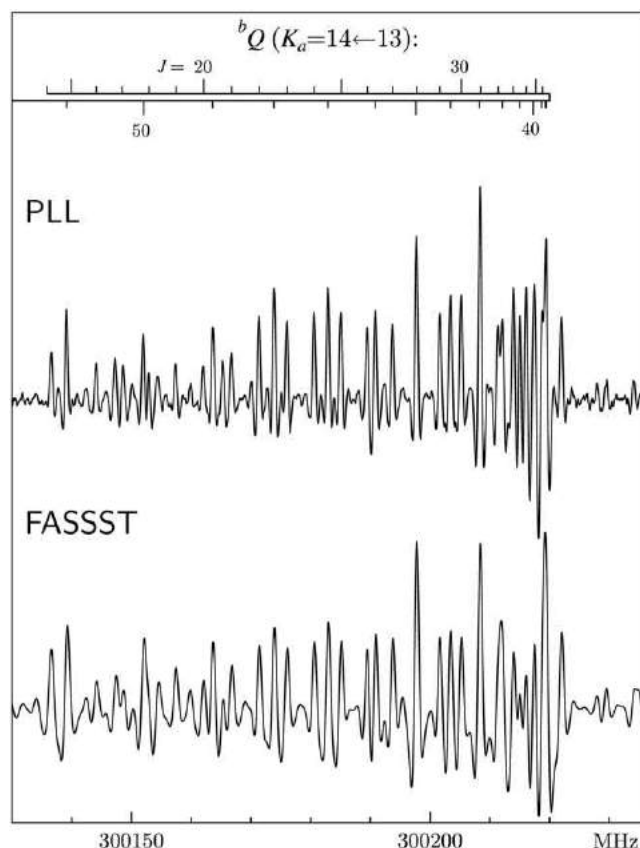


Fig. 9 Comparison of trans-gauche diethyl ether spectrum obtained in two different modes of using the BWO sources: under standard phase lock loop frequency control (PLL) or in the fast scanning mode calibrated with Fabry-Perot resonator fringes and SO₂ reference lines (FASSST). The FASSST method offered continuous access to many tens of GHz of the rotational spectrum at scan speeds exceeding 10 GHz s⁻¹, with only a small decrease in resolution. Reproduced from Medvedev, I., Winnewisser, M., De Lucia, F.C., *et al.*, 2004. The millimeter- and submillimeter-wave spectrum of the trans-gauche conformer of diethyl ether. *Journal of Molecular Spectroscopy* 228, 314–328.

be expected, while the ultimate limitation appears to be imposed by the decay rate of the free induction signal, found in the discussed work to be described by a time constant of 100 ns.

The FASSST Spectrometer

This is one of the spectrometer designs making use of the specific properties of high frequency BWO sources. It was recognized that these sources have very low noise on fast sweep operation so that sweep speeds exceeding 10 GHz s⁻¹ were achievable without significantly compromising resolution. At such sweep speeds the source has to be operated in free running mode so that external, post-measurement frequency calibration was employed. This was achieved by means of fringes of a 15 m long folded Fabry-Perot resonator (9.5 MHz fringe spacing) and absolute calibration was transferred from a reference cell containing SO₂ gas. The fringe channel and the SO₂ spectrum were recorded simultaneously with the main spectrum. Considerable signal filtering is also required, as well as averaging of up and down scans in order to eliminate low level frequency nonlinearities. This spectrometer was developed at The Ohio State University, Columbus, United States (see [Medvedev *et al.* \(2004\)](#) and the references cited therein) and the authors coined the name “Fast Scan Submillimeter Spectroscopic Technique” (FASSST) which is the acronym by which it is known. The use of three different BWO sources allowed recording of practically continuous 110–370 GHz spectra for many molecules. Calibration against a more standard spectrometer employing a phase locked source revealed that the fast scanning mode introduced only a small penalty in resolution and frequency accuracy, see [Fig. 9](#). While frequencies in the original FASSST spectrometer only overlap with the lower limit of the THz region, the technique was powerfully extended by multiplying the output of the 300 GHz source with a frequency tripler. Frequency calibration was carried out at the fundamental frequency, while frequency tripling allowed over 200 GHz of spectroscopic coverage around 1 THz.

The Broadband Resonator Spectrometer

Resonator spectrometers are a class of spectrometer in which the quality factor (*Q*) of an optical resonator is affected by the medium filling the resonator. The most robust measurement is that of the shape of a given resonator fringe, which yields precise

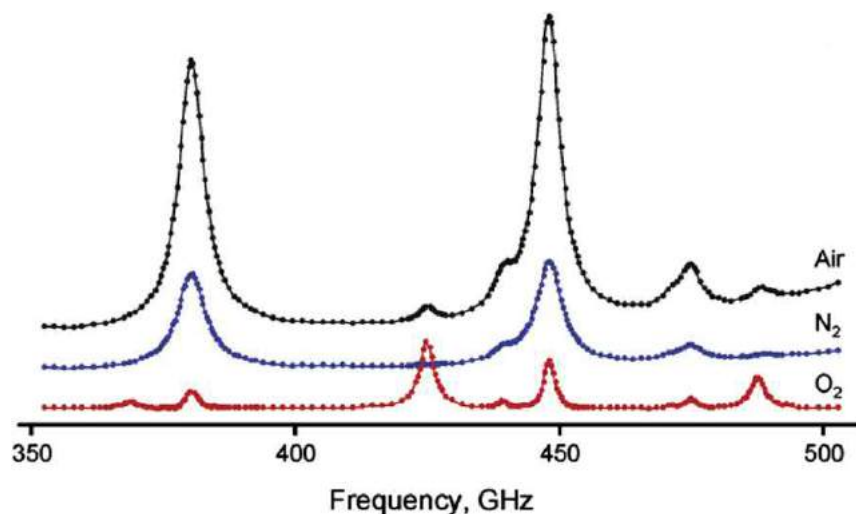


Fig. 10 Absorption spectra of the atmospheric gas composition and of individual atmospheric constituents at atmospheric pressure and with varying water vapor content (measured at 6.6%, 1.9% and near 0% relative humidity for air, N₂, and O₂, respectively). The spectra are composed of precise point by point absolute absorption measurements made with a single source of a BWO based resonator spectrometer. Reproduced from Krupnov, A.F., Tretyakov, M.Y., Belov, S.P., *et al.*, 2012. Accurate broadband rotational BWO-based spectroscopy. *Journal of Molecular Spectroscopy* 280, 110–118.

values of molecular absorption coefficients and broadening parameters. A powerful recent version of this type of spectrometer design is also based on the broad tunability and clean output power of high frequency BWOs (Krupnov *et al.*, 2012). In this case a phase locked BWO is used to make measurements on successive resonator fringes for a resonator of fixed physical size, without any need for moving any of the resonator mirrors. This technique allows measurement of gas phase rotational spectra in the important atmospheric pressure regime, Fig. 10, where the strongest lines in the three spectra are those of water. Nitrogen does not have any lines, but contributes to pressure broadening. The lines at 368.5, 424.8, and 487.2 GHz, which are visible only in the O₂ and in the air spectrum, are a spin triplet for the $N=3 \leftarrow 1$ rotational transition of the oxygen molecule. Oxygen is a rare molecule in which its electronic ground state is a triplet state ($^3\Sigma$) due to the presence of two unpaired electrons. In such case, even though O₂ does not have a permanent electric dipole moment, it will have pure rotational transitions allowed by the magnetic dipole moment. The N quantum number accounts also for the electronic angular momentum, which is combined with the usual rotational angular momentum. The widths of molecular lines in Fig. 10 are in the region of 5 GHz, and are four orders of magnitude greater than in the low pressure regime. This makes it clear why the standard signal modulation methods discussed above are not applicable and a specialized technique is necessary. The incentive for its development was that the atmospheric THz molecular spectrum is of particular applied importance as will be seen below in the discussion of astrophysical applications.

THz Lamb-Dip Spectroscopy

This is a method at the other extreme to that described immediately above, as it is a variant of molecular spectroscopy in which precision and resolution are enhanced over those available with the more routine techniques. Lamb-dip spectroscopy, or saturation spectroscopy, is usually associated with laser spectroscopy since it relies on the use of high radiation power density. In a longitudinal THz molecular absorption cell it is most convenient to achieve sufficient power density by a double pass arrangement, where a rooftop reflector is placed at the end of the cell, allowing detection to be at the cell entrance. In such case molecules with a zero Doppler velocity component relative to the probing radiation (moving perpendicular to the cell axis) will show preferentially decreased absorption at the transition frequency. Their lower energy levels participating in the transition will be rapidly depopulated and further photons arriving at this frequency will not be absorbed. At frequencies on the slopes of the Doppler absorption profile the absorption on the forward and backward pass will correspond to molecules moving with different velocity vectors so that depopulation by radiation will be much less efficient. In this way a frequency sweep will produce a narrow peak at the absorption maximum directed toward zero absorption (the Lamb-dip). In many cases high effective radiation power can also be set up through unavoidable Fabry–Perot type resonances in the absorption cell. For this reason observation of Lamb dips on absorption lines of water is a relatively common occurrence when the cell is pumped down, but trace water is still desorbing from the cell walls. Nevertheless, a well designed Lamb-dip experiment as described by Cazzoli and Pazzarini (2013) allows frequency measurement precision at the kHz level, as is implicit from the examples in Fig. 11. The conditions for successful saturation measurements require sample pressures of less than 1 mTorr and reduced frequency modulation widths comparable with the achievable sub 100 kHz linewidths of the Lamb-dip components. Such results have clear relevance to derivation of very precise spectroscopic constants, including those implicated in various types of small scale spectroscopic splitting, that is not resolvable by other means.

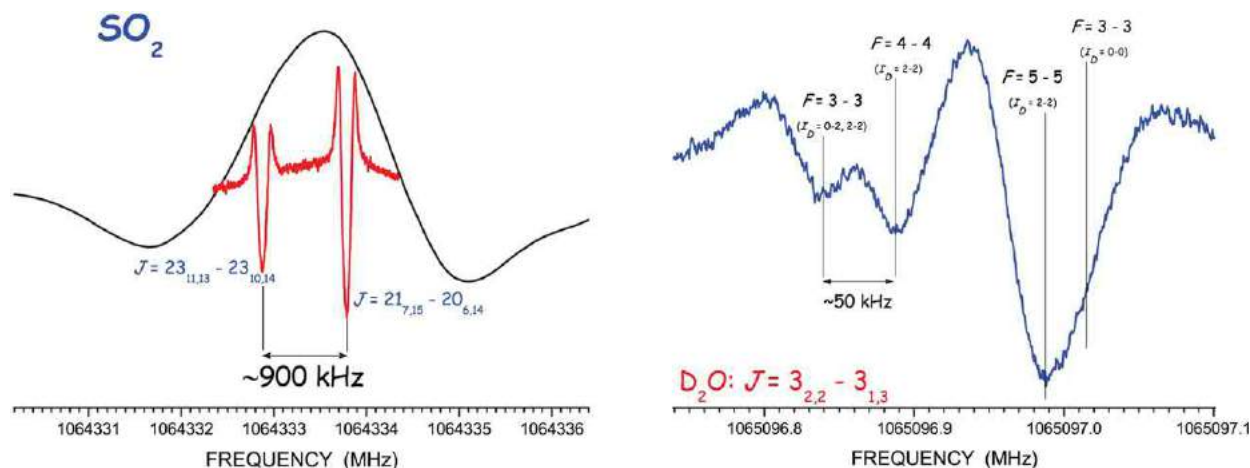


Fig. 11 Illustration of the enhanced resolution possible on application of the Lamb-dip technique in the THz region. The left panel shows how two transitions of SO_2 blended in the standard Doppler profile limited spectrum (black trace) can be completely resolved (red trace). The right panel demonstrates the resolution of components in a transition of D_2O that result from nuclear quadrupole hyperfine splitting due to the two deuterium nuclei. Reproduced from Cazzoli, G., Puzzarini, C., 2013. Sub-doppler resolution in the THz frequency domain: 1 kHz accuracy at 1 THz by exploiting the Lamb-dip technique. *Journal of Physical Chemistry A* 117, 13759–13766 with permission.

Table 3 THz region bands available at the ALMA observatory

Band number	Frequency (GHz)	Wavelength (mm)
7	275–373	0.804–1.091
8	385–500	0.600–0.779
9	602–720	0.417–0.498
10	787–950	0.316–0.381

Astrophysical Molecular Spectroscopy

In this case the emphasis is on detection. Extra-terrestrial THz spectra are mostly emission spectra resulting from excitation by distant sources, so that recording such spectra involves efficient collection of incoming radiation by means of suitable antennas and then extracting the spectra by heterodyne techniques. Precision in antenna construction is paramount since adhering to even the $\lambda/10$ rule of surface precision for a radio telescope dish of several meter diameter is challenging for sub-millimeter wavelengths. The premier radio telescope installation is currently the Atacama Large Millimeter Array (ALMA), which is operated on the Chajnantor plain in the Atacama desert in Chile, at an altitude of 5000 m. This consists of fifty movable 12-meter diameter antennas made to 25 μm surface precision, which can be configured into various baseline configurations for increased spatial resolution. Heterodyne detection is implemented with the use of liquid helium cooled downconversion “back ends” allowing simultaneous coverage of a signal bandwidth of up to 7.5 GHz. Detection is in the form of position/intensity/frequency “data cube”, so that mapping of the spatial distribution of signals from several molecules is possible on the basis of a single observation.

ALMA makes available four frequency bands for THz spectroscopic studies, as listed in Table 3. Since the observatory is subject to atmospheric absorption its location was chosen in a particularly dry, high-altitude area in order to minimize the absorption from the atmospheric spectrum already shown in the laboratory version in Fig. 10. The actual atmospheric absorption at the ALMA site is reproduced in Fig. 12 and it is immediately apparent how THz atmospheric molecular absorption lines affect the coverage chosen for the various bands. The main culprits in this case are water lines, as listed in Fig. 12, although the 60 GHz oxygen absorption also enforces the gap between bands B1 and B2. Many other molecular lines are also visible in Fig. 12, although those do not pose such a hindrance to observations. Those lines are mainly due to ozone, which is an important atmospheric constituent, and such lines are, for example, the ‘grass’ between 230 GHz and the 325 GHz water line, the multiplets around 655 GHz, and the multiplets around 835 GHz.

Although ALMA is currently the most capable ground based astrophysical THz observatory, it has had several ground based predecessors, and THz molecular spectroscopy has also been possible with satellite observatories. The advantage of satellite observatories is their immunity to atmospheric absorption, but the disadvantage has been relatively short operational lifetime due to limits on liquid helium coolant supply used to achieve the highest detection sensitivity. The best known THz capable satellite has been the Herschel Space Observatory, carrying a 3.5 m diameter collecting mirror and a heterodyne spectroscopy instrument

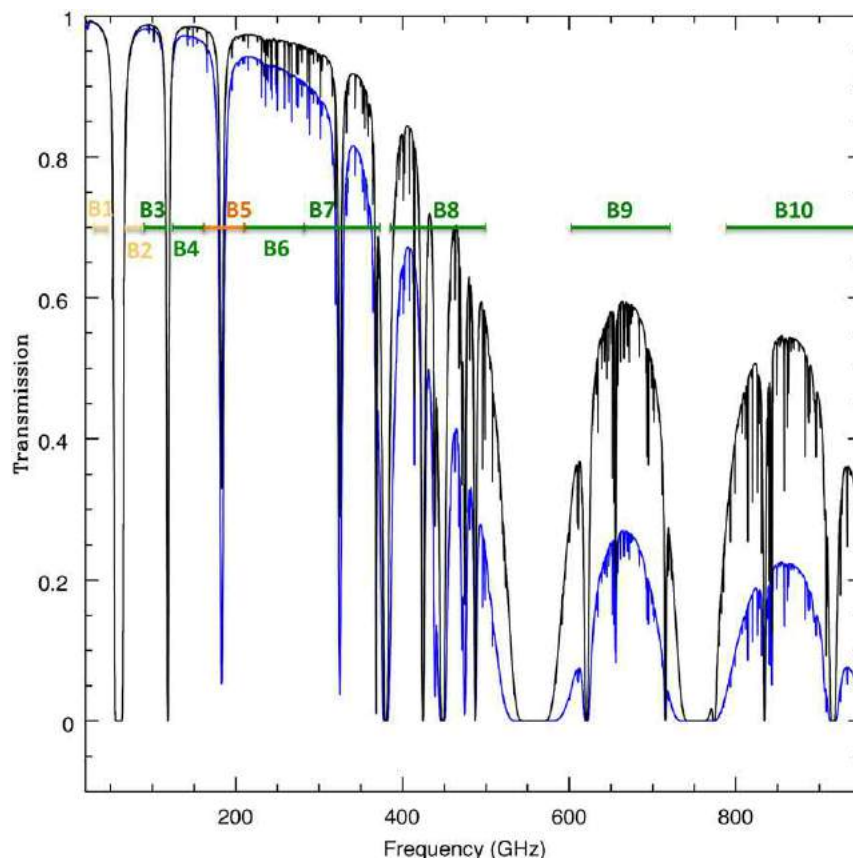


Fig. 12 Comparison of the band choice of the ALMA observatory with the atmospheric water spectrum recorded in its location and corresponding to 50% (blue) and 12.5% (black) of the available observing conditions. The water THz spectrum commences with the relatively benign $5_{14} \leftarrow 4_{22}$ line at 325.2 GHz, but it then contains the fully attenuating lines, $4_{13} \leftarrow 3_{21}$ at 380.2 GHz, $1_{10} \leftarrow 1_{01}$ at 556.9 GHz, and $2_{11} \leftarrow 2_{02}$ at 752 GHz, which enforce the three gaps in coverage between bands B7 to B10. Note that all of the atmospheric lines from the laboratory resonator spectrum shown in Fig. 10 are visible. Reproduced from Schieven, G. (Ed.), 2016. Observing With ALMA – A Primer, ALMA Doc. 4.1, Ver. 3. Available at: <https://almascience.eso.org/documents-and-tools/cycle4/alma-early-science-primer>.

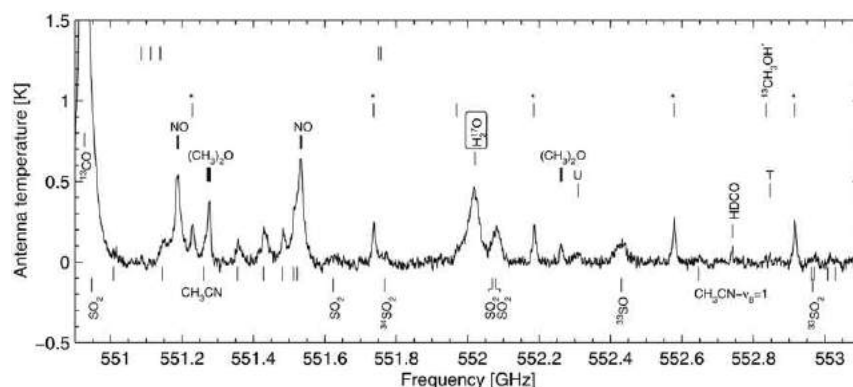


Fig. 13 A small excerpt from the Odin satellite spectral survey of the Orion KL interstellar molecular cloud core. Unlabeled markers placed at the intensity level of 0.9 K identify a Q -branch line sequence of CH_3OH in the first excited torsional state. These lines and the strong ^{13}CO and H_2^{17}O lines indicate how increasing sensitivity of astrophysical spectra requires complete understanding also of transitions in rare isotopic species, and in excited vibrational states of many molecules. Reproduced from Olofsson, A.O.H., Persson, C.M., Koning, N., *et al.*, 2007. A spectral line survey of Orion KL in the bands 486–492 and 541–577 GHz with the Odin satellite I. The observational data. *Astronomy and Astrophysics* 476, 791–806 with permission.

(HIFI) covering 480–1280 GHz and 1419–1910 GHz. Herschel was operated from 2009 to 2013, when its 2300 liters of onboard helium coolant ran out. There have been many scientific results from Herschel, even after its lifetime, such as identification of water vapor on the dwarf planet Ceres.

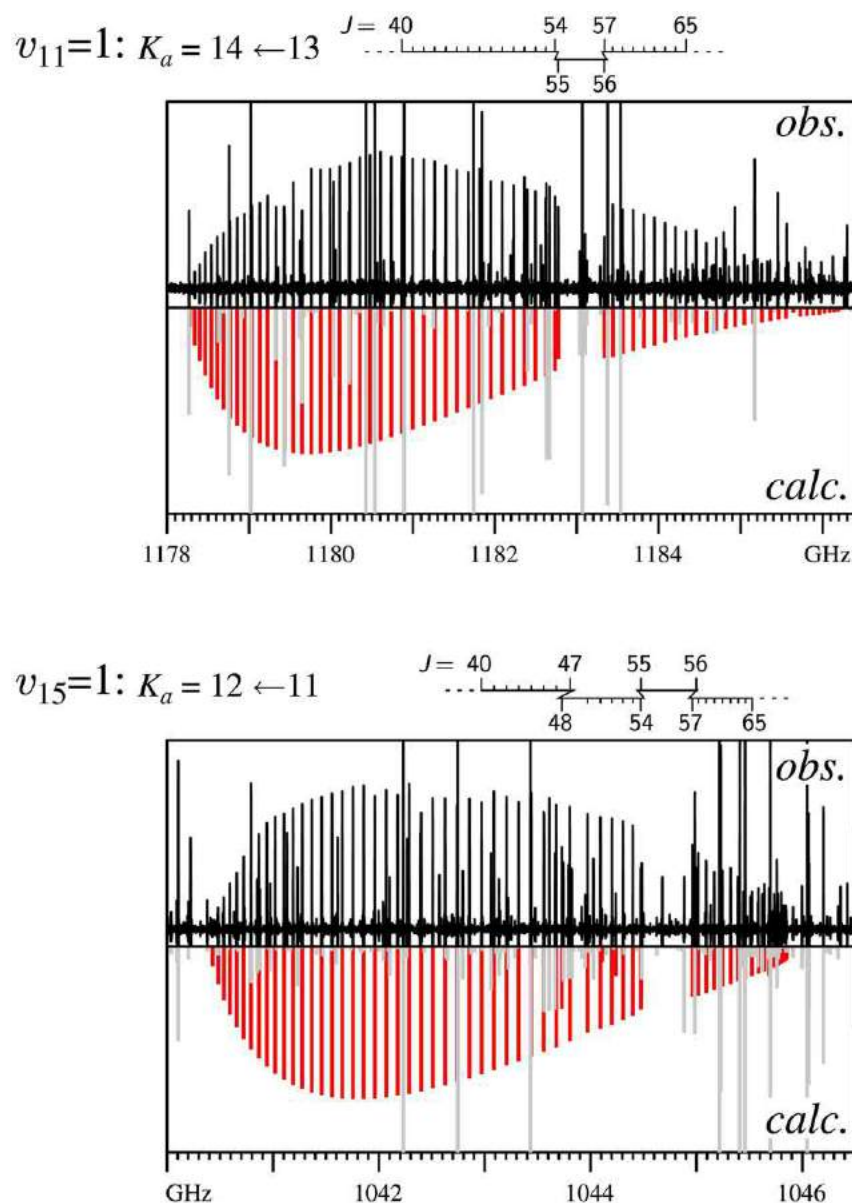


Fig. 14 Example of clearly discernible perturbations in the THz spectrum of acrylonitrile. The striking gaps in bQ -branch bands in two different vibrational states, $v_{11}=1$ (upper) and $v_{15}=1$ (lower) are the result of matching resonance interactions, which take place for specific values of J between 54 and 57. Such interactions are successfully accounted for in the complete quantum mechanical treatment of the problem, as shown in the calculated spectrum. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2012. Broadband rotational spectroscopy of acrylonitrile: Vibrational energies from perturbations. *Journal of Molecular Spectroscopy* 280, 134–144.

A preview of the rich astrophysical molecular spectroscopy in the THz is visible in [Fig. 13](#), which displays only a small fragment of a spectroscopic survey of the diffuse molecular cloud in the Orion nebula. This result comes from the Odin satellite, equipped with a moderate 1.1 m diameter telescope operating at 486–580 GHz. It is noteworthy that understanding this spectrum requires the knowledge not only of the main molecular species, but also of their, sometimes rare, isotopes (^{17}O has a terrestrial abundance of 0.04%) and of transitions in excited vibrational states. In the spectrum in [Fig. 13](#) this concerns relatively few molecules, but more sensitive observational instruments such as ALMA require such information for a large number of molecules.

Influence of the THz Region on Molecular Spectroscopy

The advances in experimental techniques, which currently allow broadband access to high resolution molecular spectra in the THz region, have created new challenges for spectroscopists. The first is simply the need for efficient handling of spectra as illustrated by

the fact that high-resolution coverage of the whole THz region at 0.1 MHz point spacing requires 27 million data points and such spectrum will contain well in excess of one hundred thousand lines. There is, therefore, the need to efficiently navigate very long spectra, to compare spectra with predictions, and to set up data sets for programs fitting spectroscopic constants in a Hamiltonian suitable for the problem at hand. Various computer programs have been developed to meet this challenge and those are generally available for use by the spectroscopic community on specialized websites.

Measurement of THz molecular spectra usually took place as an extension of previous lower frequency studies of a given molecule. The primary advance brought in by the THz data was increased numerical precision in accounting for transition frequencies of considerably greater numbers of measured lines. But there have also been cases where THz measurements allowed access to qualitatively new molecular information, which is was not revealed at lower frequencies. One such example is for the already discussed acrylonitrile molecule and is illustrated in Fig. 14. The rotational spectrum of acrylonitrile has been subjected to many preceding spectroscopic studies, but it was only when transitions around 1 THz were measured that several intriguing features became apparent. It does not take special insight to notice that there are unusual gaps in the otherwise uniform progressions of Q-branch transitions in the two lowest excited vibrational states of the molecule. The pertinent transitions are not missing but are shifted due to interactions resulting from near coincidence in energy of rotational levels in the two vibrational states (Kisiel *et al.*, 2012). Similar, but less spectacular, interactions were observed between the ground state and the $\nu_{11}=1$ vibrational state, also at frequencies close to 1 THz, but for values of the J quantum number exceeding 100 (Kisiel *et al.*, 2009). A suitable quantum mechanical analysis of such interactions allowed precise determination of the vibrational frequencies of the two lowest normal modes of acrylonitrile: $\nu_{11}=228.29986(2)$ and $\nu_{15}=332.67811(2)$ cm^{-1} , even though the actual vibrational transitions from the ground state fall outside the THz region. This analysis was later confirmed by synchrotron based high resolution rotation vibration spectroscopy (Kisiel *et al.*, 2015).

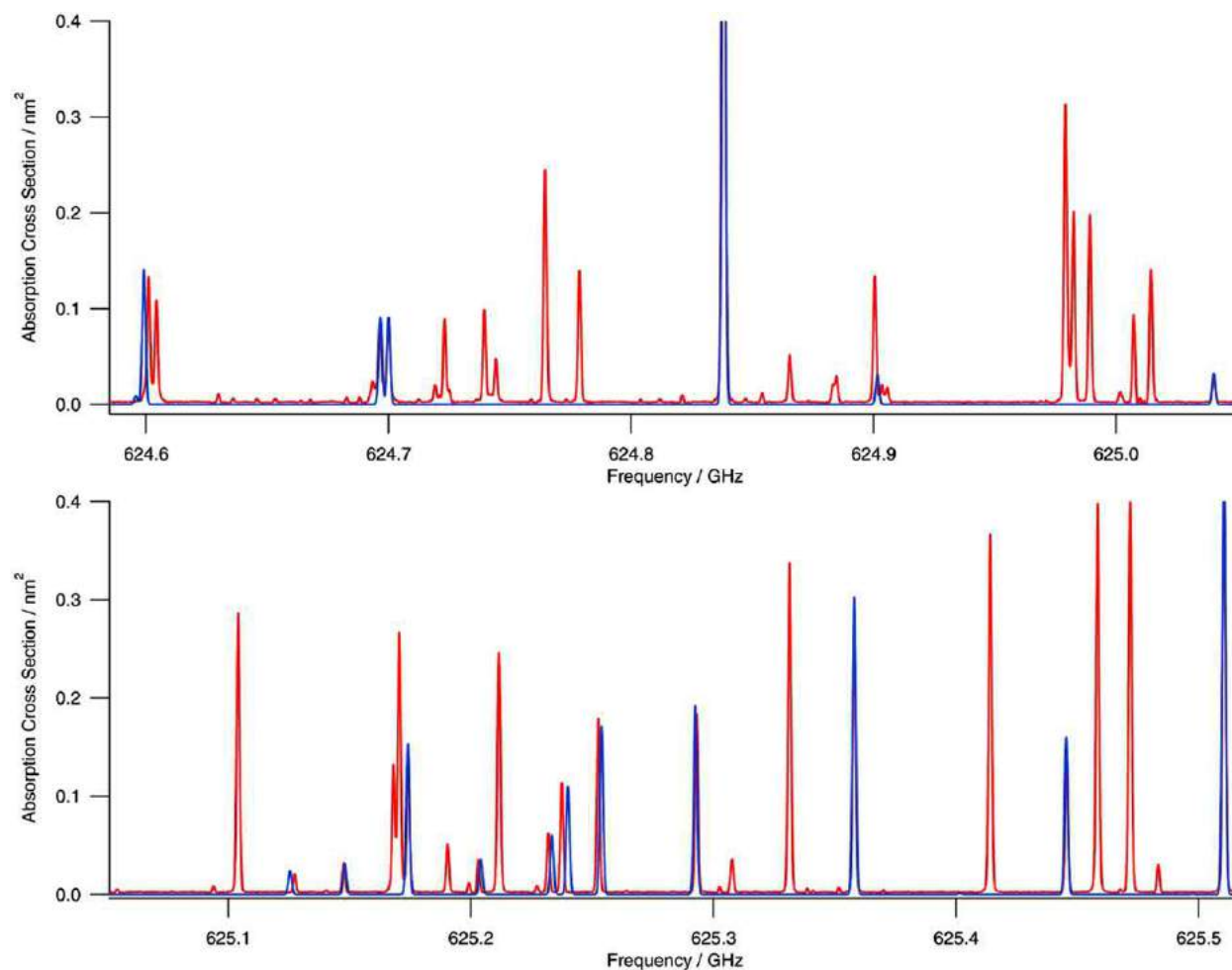


Fig. 15 Comparison of the experimentally derived, absolute intensity calibrated, 300 K spectrum of methanol (red) with prediction from the available catalog data (blue). The experimental trace is a point by point simulation based on recording 166 experimental spectra over the 248–398 K temperature range. The number of inconsistencies between the two traces indicates that considerable progress still has to be made in understanding this complex spectrum. Reproduced from Fortman, S.M., Neese, C.F., De Lucia, F.C., 2014. The complete, temperature resolved experimental spectrum of methanol (CH_3OH) between 560 and 654 GHz. *Astrophysical Journal* 782, 1–8 with permission.

Molecular Spectroscopy Databases and Their Verification

The molecules with extensive THz spectra, which are of greatest applied interest, such as water and methanol, are some of the most challenging molecules for quantum mechanics. The extensive coverage of their spectra that is now possible and the need to reproduce the experimental frequencies to within their experimental uncertainty require very specialized techniques of constructing and fitting quantum mechanical Hamiltonians. The number of parameters of fit becomes considerable and a ratio of the number of fitted lines to the number of fitted parameters exceeding 100 is considered satisfactory. This means that for data sets reaching tens of thousands of lines the number of fitted parameters runs into hundreds.

At the same time most users of such data, such as the astrophysical community, are not familiar with the spectroscopic details but simply require line lists of frequencies and temperature scalable intensities. In order to cater for this demand several very useful databases have been established. These are the JPL Catalog, the Cologne Database for Molecular Spectroscopy (CDMS), and the High-resolution transmission molecular absorption database (HITRAN). The first two databases are oriented more towards the astrophysical user, while the development of HITRAN is geared to the needs of atmospheric spectroscopy.

The databases are continuously updated as new experimental data is acquired, but their coverage of a given molecule is still far from complete. The complexity of some spectroscopic problems, especially of the spectrum of methanol, is such that transition frequencies that are not among the fitted data, but are only predicted, may not yet be sufficiently reliable. For this reason dedicated measurements have been undertaken with the specific aim of database verification, both in transition frequency and in absolute intensity. This was carried out for several molecules of astrophysical importance by recording their laboratory spectra at multiple temperatures. Suitable processing allowed derivation of an absolute intensity spectrum for any desired temperature, and the result for the methanol spectrum at 0.6 THz is shown in Fig. 15. Comparison with the data available in the JPL and CDMS catalogs (see Relevant Websites section) reveals that there are still significant gaps in the coverage of transitions and that there is still room for improvement in the (frequency, intensity) parameters of some of the catalog lines.

Conclusions

Relatively routine experimental access to the THz region and the resulting opportunities for high-resolution molecular spectroscopy allowed considerable advances in spectroscopy of several important molecules. The astrophysical community is an important consumer of such data and it has stimulated much laboratory work, largely in an attempt to characterize the spectroscopic ‘weeds’ that hinder identification of new molecules. At the same time significant improvement in characterization of rotational and vibrational properties of several key molecules has been reached, although the abundance of spectroscopic data shows that there are still significant gaps in identification of observed lines and there is still much further work ahead.

References

- Cazzoli, G., Puzzarini, C., 2013. Sub-doppler resolution in the THz frequency domain: 1 kHz accuracy at 1 THz by exploiting the Lamb-dip technique. *Journal of Physical Chemistry A* 117, 13759–13766.
- Drouin, B.J., Maiwald, F.W., Pearson, J.C., 2005. Application of cascaded frequency multiplication to molecular spectroscopy. *Review of Scientific Instruments* 76, 093113.1–10.
- Kisiel, Z., Martin-Drumel, M.-A., Pirali, O., 2015. Lowest vibrational states of acrylonitrile from microwave and synchrotron radiation spectra. *Journal of Molecular Spectroscopy* 315, 83–91.
- Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. *Journal of Molecular Spectroscopy* 258, 26–34.
- Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2012. Broadband rotational spectroscopy of acrylonitrile: Vibrational energies from perturbations. *Journal of Molecular Spectroscopy* 280, 134–144.
- Krupnov, A.F., Tretyakov, M.Y., Belov, S.P., *et al.*, 2012. Accurate broadband rotational BWO-based spectroscopy. *Journal of Molecular Spectroscopy* 280, 110–118.
- Medvedev, I., Winniewisser, M., De Lucia, F.C., *et al.*, 2004. The millimeter- and submillimeter-wave spectrum of the trans-gauche conformer of diethyl ether. *Journal of Molecular Spectroscopy* 228, 314–328.
- Neill, J.L., Harris, B.J., Steber, A.L., *et al.*, 2013. Segmented chirped-pulse Fourier transform submillimeter spectroscopy for broadband gas analysis. *Optics Express* 21, 19743–19749.
- Rao, K.N. (Ed.), 1976. *Molecular Spectroscopy: Modern Research*, vol. II. New York: Academic Press. (Chapter 2).
- Schlemmer, S., Giesen, T., Lewen, F., Winniewisser, G., 2009. High-resolution laboratory terahertz spectroscopy and applications to astrophysics. In: Laane, J. (Ed.), *Frontiers of Molecular Spectroscopy*. Amsterdam: Elsevier, pp. 241–265.

Relevant Websites

- <http://www.almaobservatory.org/>
ALMA.
- <http://www.almascience.org/>
ALMA.
- <http://www.astro.uni-koeln.de/cdms/>
CDMS – Cologne Database for Molecular Spectroscopy.

<http://www.istokmw.ru/vakuumnie-generatori-maloy-moshnosti/>

Examples of THz region BWO tubes produced by Istok Company, Russia.

<http://vadiodes.com/en/products/custom-transmitters>

Examples of THz transmitter modules produce by Virginia Diode Inc., USA.

<http://www.esa.int/SPECIALS/Herschel>

Herschel.

<https://www.cfa.harvard.edu/hitran/>

High-resolution transmission molecular absorption database, HITRAN.

<http://hitran.iao.ru/>

HITRAN on the web.

<https://spec.jpl.nasa.gov/>

JPL Molecular Spectroscopy, home of the JPL Catalog and of the SPFIT/SPCAT program suite.

<http://www.mwl.sci-nnov.ru/2016.html>

Microwave Spectroscopy Laboratory at Nizhny Novgorod.

<http://www.nasa.gov/herschel>

NASA.

<http://www.snsb.se/en/Home/Space-Activities-in-Sweden/Satellites/Odin/>

Odin satellite.

<http://info.ifpan.edu.pl/~kisiel/prospe.htm>-PROSPE

Programs for Rotational SPECTroscopy.

Broadband Terahertz Sources

Kang Liu, University of Rochester, Rochester, NY, United States

Xi-Cheng Zhang, University of Rochester, Rochester, NY, United States; ITMO University, Saint-Petersburg, Russia; and Capital Normal University, Beijing, China

© 2018 Elsevier Ltd. All rights reserved.

Photo-Induced Carriers in Solid Materials

Photoconductive Antenna

Broadband Terahertz (THz) pulses can be generated by a biased photoconductive (PC) antenna excited with femtosecond optical pulses. The PC antenna is one of the most widely used components for THz generation and detection. Fig. 1(a) gives a schematic demonstration of a typical THz generation PC antenna structure. It is composed of two metal electrodes deposited on a semiconductor substrate. When generating THz radiation, a DC voltage is applied across the electrodes. The femtosecond optical pulses which have photon energy higher than the semiconductor energy band gap will radiate the area between the electrodes and generate photo-induced free carriers in the substrate. Sometimes, free carriers can also be formed using excitation pulses with photon energy lower than the semiconductor band gap, due to multi-photon absorption in the substrate. Once the free carriers are formed, they will be accelerated by the electric field between the electrodes, and a photocurrent J is produced. As the photocurrent varies with time, it radiates electromagnetic wave pulses with an electric field E_{THz} at the far field that can be express as (Zhang and Xu, 2010)

$$E_{\text{THz}} \propto \frac{\partial J(t)}{\partial t} \quad (1)$$

Fig. 1(b) plots the calculated results of a photocurrent induced by an optical pulse and its corresponding THz far-field waveform with typical material and physical parameters, which should give a straightforward intuition regarding the relation between the optical pulse, the photocurrent, and the THz radiation electric field. The emitted THz electric field has a polarization parallel to the direction of the bias field, which can be flipped by switching the direction of the bias field.

The power and bandwidth of the THz emission from PC antenna largely depend on the structure of the electrodes (Tani *et al.*, 1997). Fig. 2 shows the temporal and spectral amplitude of THz radiation from three different electrode structures: dipole, bow tie and strip line, deposited on low-temperature grown GaAs. The THz pulse from the strip line structure has a shorter pulse duration, corresponding to a spectrum extending up to 4 THz with a FWHM of about 1.3 THz.

The output power of a PC emitter also depends on the bias voltage and the optical pump power. When the optical pump power is low and the bias field is weak, the THz radiation field amplitude increases linearly with these two parameters. However, increasing the bias voltage has a limitation because a high electric field may result in the breakdown of the substrate material. For example, the breakdown field of LT-GaAs is reported as ~ 300 kV/cm (Tani *et al.*, 1997). In addition, the THz radiation output power will saturate as the optical pump power reaches a certain level, for the screening effect of the bias field by the photo-induced carriers.

Ever since the first introduction of THz generation by PC antenna in the 1980s (Auston *et al.*, 1984a), this technique has undergone dramatic developments, due to the soaring amount of research on the THz-time-domain-spectroscopy techniques. Up to now, the strongest broadband THz pulse energy emitted by PC antenna was reported to be $8.3 \mu\text{J}$ with a peak electric

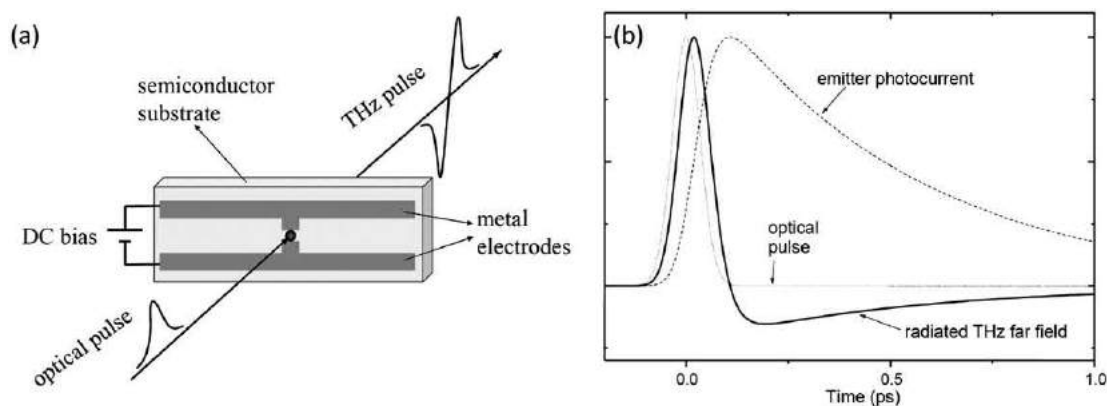


Fig. 1 (a) Schematic diagram of THz pulse emission from a PC antenna excited by a femtosecond laser pulse and (b) calculated photocurrent (dashed line) vs. time. Reprinted from Lee, Y.S. 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 61.

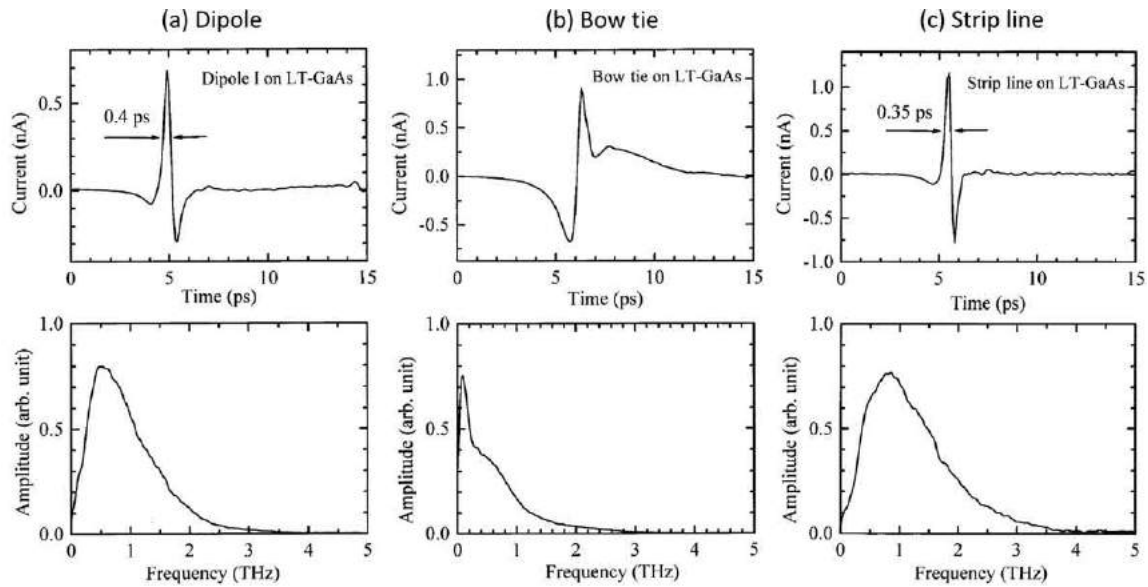


Fig. 2 THz radiation pulse shapes and amplitude spectra from PC emitters with (a) dipole, (b) bow tie and (c) strip line. Reprinted from Tani, M., Matsuura, S., Sakai, K., Nakashima, S.-i., 1997. Emission characteristics of photoconductive antennas based on low-temperature-grown GaAs and semi-insulating GaAs. *Applied Optics* 36, 7853–7859.

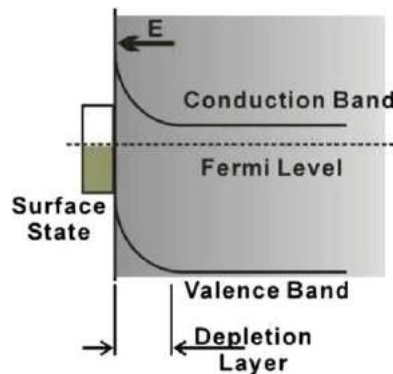


Fig. 3 Schematic demonstration of band bending and surface field of a n-type GaAs wafer. Reprinted from Zhang, X.-C., Xu, J., 2010. *Introduction to THz Wave Photonics*. New York: Springer, p. 32.

field of 331 kV/cm, achieved with an encapsulated interdigitated ZnSe Large Aperture Photo-Conductive Antenna (LAPCA) (Ropagnol *et al.*, 2016). A 20 THz broadband generation using semi-insulating GaAs interdigitated photoconductive antennae was also reported in 2014 by Hale *et al.*, (2014).

Built-in Field in Semiconductor

THz generation from the surface of semiconductor illuminated by femtosecond laser was reported by Zhang *et al.* (1990). This generation takes advantage of the surface built-in field of semiconductors, such as GaAs (Zhang and Xu, 2010). The Fermi level of the surface state at the interface between the material and the air is different from the Fermi level of the bulk material, which leads to the semiconductor band bending close to the interface. In this area, which is also called the depletion layer, forms a built-in surface field that causes the carriers to drift toward the inside of the bulk material and eventually a balance is reached between the drift and the diffusion of the free carriers. Fig. 3 shows the band bending and surface field in an n-type GaAs wafer. When a femtosecond laser pulse excites the semiconductor, photo-induced carriers will start drifting again under the built-in electric field, and therefore emit THz radiation just like what happens in the PC antenna. Since the surface field of a n-type GaAs is in opposite direction with a p-type GaAs, the THz pulse generated from p-type GaAs will have an opposite polarity compared to the one from n-type GaAs. The THz radiation amplitude depends on the optical pump pulse incident angle. The amplitude reaches the maximum as the angle approaches the Brewster angle.

Photo-Dember Effect

With the semiconductor materials that have small or no built-in field effect, THz radiation can still be generated due to photo-Dember effect (Zhang and Xu, 2010). When a semiconductor material surface is illuminated by a femtosecond laser pulse with photon energy higher than the material band gap, a large amount of free electron hole pairs are generated with an inhomogeneous distribution. The asymmetric distribution of the carriers on the surface causes them to diffuse toward the inside of the bulk material. Due to the fact that electrons and holes have different mobility, an ultrafast spatial separation between the electrons and holes is formed, and results in the radiation of THz pulses. InAs has been considered as an important semiconductor THz emitter due to its high electron mobility. When the THz generation process from InAs was studied, it was found that both the THz radiation from n-type and p-type InAs have the same polarity, identical to the THz pulse polarity from n-type GaAs, which cannot be explained by the mechanism of built-in surface field THz generation, but can be addressed by photo-Dember effect (Gu *et al.*, 2002). Actually, both effects exist in semiconductor materials when excited by ultrafast laser pulses. But the dominating effect depends on the semiconductor band gap structure and the property of the laser pulses.

Accelerated Free Electrons

Free-Electron Lasers

The free-electron lasers are outstanding broadband THz sources featuring ultra intense THz radiation brightness, normally several orders of magnitude higher than the laser-driven table-top THz sources. It has broad applications from chemical and biological imaging to driving nonlinear effects in the matter with THz radiation.

A relativistic electron bunch under acceleration emits cone-shaped radiation pattern in the direction of the electrons' velocity (Lee, 2010). There are many different ways to generate radiation by accelerating/decelerating the electron beam: bending the beam in a strong magnetic field (coherent synchrotron radiation and coherent edge radiation, CSR and CER), passing the beam through a small aperture (coherent diffraction radiation, CDR), or passing the beam from vacuum into a medium (coherent transition radiation, CTR, or *bremstrahlung*) (Wu *et al.*, 2013). Fig. 4 presents a schematic demonstration of the coherent THz radiation from a linear particle accelerator (LINAC). An ultrafast femtosecond pulse releases an ultrashort electron bunch from an electron source, which normally is either a photocathode electron gun or a semiconductor surface. In the accelerator, the electron bunch gains relativistic energy, 10–100 MeV, and the THz radiation can be generated through smashing the electron bunch onto a metal target or bending the electron trajectory by a magnetic field.

Many accelerator-based THz source facilities have been constructed worldwide, including the energy-recovered linac (ERL) in the Jefferson Laboratory (Carr *et al.*, 2002), the Source Development Lab (SDL) of the Brookhaven National Laboratory (Shen *et al.*, 2007), the Linac Coherent Light Source (LCLS) and the Facility for Advanced Accelerator Experimental Tests (FACET) of the SLAC National Accelerator Laboratory (Wu *et al.*, 2013), Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung (BESSY) and Deutsches Elektronen Synchrotron (DESY) in Germany (Abo-Bakr *et al.*, 2002). The strongest coherent broadband THz radiation electric field from accelerator-based facilities has been generated by the SLAC National Accelerator Laboratory. The peak electric field at a THz focus has reached 4.4 GV m^{-1} (0.44 V Å^{-1}) (Wu *et al.*, 2013). Notice that an electric field in the V Å^{-1} regime is comparable to the dipole field of an ionic bond, and can lead to many research topics involving the interaction between material and extreme THz radiation (Dhillon *et al.*, 2017).

Nonlinear Crystals

Optical Rectification

Optical Rectification (OR) is a second order nonlinear effect that is widely exploited for broadband THz pulse generation. Compared with PC antennae, OR crystals do not require application of high voltage on the material, making it one of the most

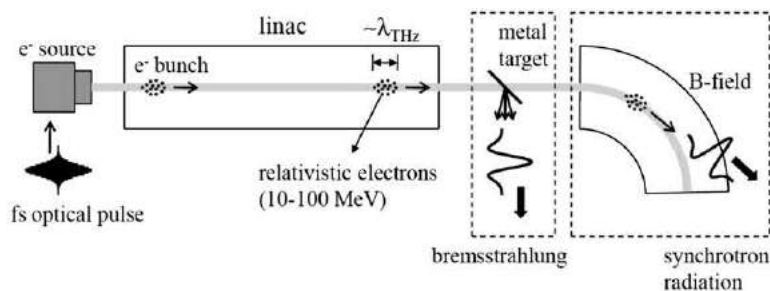


Fig. 4 Coherent THz radiation from relativistic electrons in a linear accelerator (linac). Reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 103.

efficient and convenient intense coherent broadband table-top THz sources (Hafez *et al.*, 2016; Auston *et al.*, 1984b; Hu *et al.*, 1990).

When an intense optical beam passes through a nonlinear material, electric polarization P induced in the material can be expressed as a power series in the field strength E of the optical beam (Boyd, 2008):

$$P = \epsilon_0(\chi^{(1)}E + \chi^{(2)}EE + \chi^{(3)}EEE + \chi^{(4)}EEEE + \dots) \quad (2)$$

where ϵ_0 is the permittivity of free space, $\chi^{(1)}$ is the linear susceptibility, the quantities $\chi^{(2)}$ and $\chi^{(3)}$ are known as the second- and third- order nonlinear optical susceptibilities respectively. Among the different orders of nonlinear polarizations, the second order polarization components include:

$$\begin{aligned} P(2\omega_1) &= \epsilon_0\chi^{(2)}E_1^2(\text{SHG}) \\ P(2\omega_2) &= \epsilon_0\chi^{(2)}E_2^2(\text{SHG}) \\ P(\omega_1 + \omega_2) &= 2\epsilon_0\chi^{(2)}E_1E_2(\text{SFG}) \\ P(\omega_1 - \omega_2) &= 2\epsilon_0\chi^{(2)}E_1E_2^*(\text{DFG}) \\ P(0) &= 2\epsilon_0\chi^{(2)}(E_1E_1^* + E_2E_2^*)(\text{OR}) \end{aligned} \quad (3)$$

ω_1, ω_2 , E_1 and E_2 are respectively the frequencies and the electric fields of the two distinct frequency components of the incident beam. Here different polarization components are labeled by their corresponding physical process, such as second-harmonic generation (SHG), sum-frequency generation (SFG), difference-frequency generation (DFG), and optical rectification (OR) that we shall focus on in this section.

In essence, OR is the generation of a quasi-DC polarization in a nonlinear material when it has been passed through by an intense optical beam (see Fig. 5(e)), which results in a DC electric field proportional to the beam intensity. It can be understood as a second order phenomenon that is the reverse process of the electro-optic effect (Dexheimer, 2007). If the light beam that induces OR is a pulse instead of a continuous wave, the electric field generated by OR will become a time varying function related with the pulse envelope, thus radiate electromagnetic waves (Zhang and Xu, 2010). The far field radiation electric field should be proportional to the second derivative of OR induced time varying electric field:

$$E_{\text{THz}}(t) \propto \frac{\partial^2 P(0, t)}{\partial t^2} = \chi^{(2)} \frac{\partial^2 I(t)}{\partial t^2} \quad (4)$$

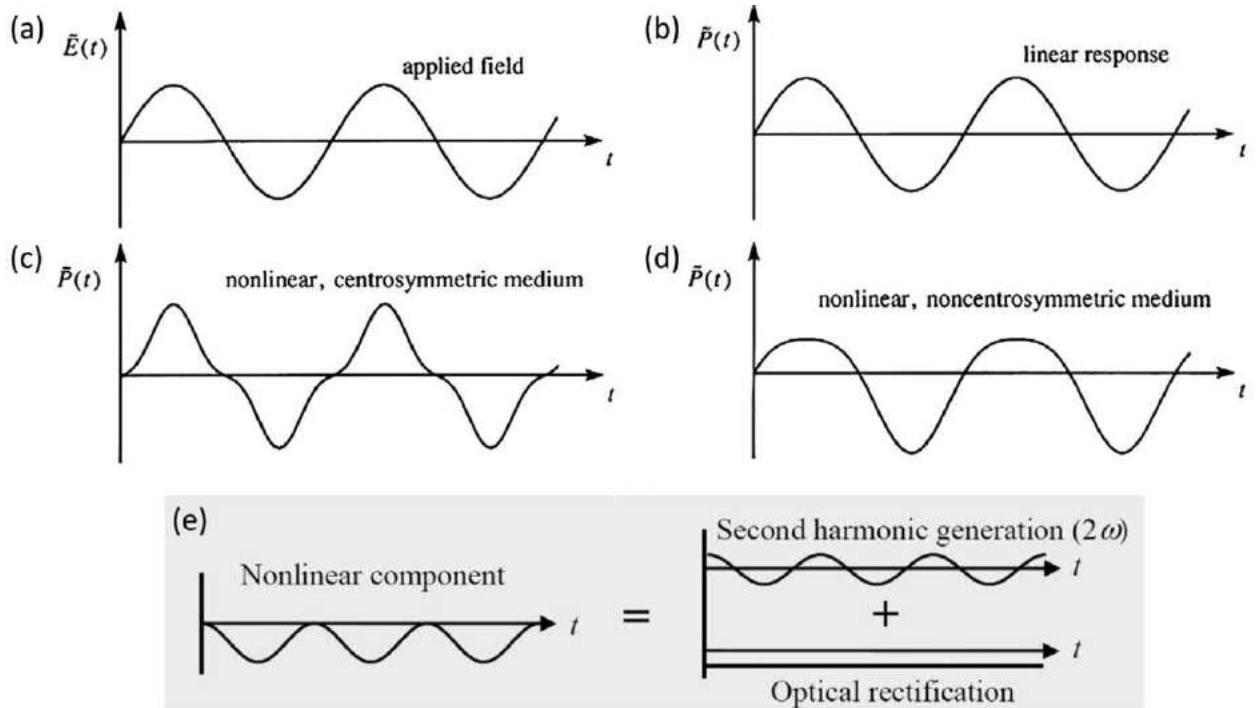


Fig. 5 (a–d) Waveforms associated with the atomic response (Reprinted from Boyd, R.W., 2010. *Nonlinear Optics*. Elsevier Science, p. 44) and (e) the decomposition of nonlinear response in a noncentrosymmetric medium. Reprinted from Lee, Y.S., 2010. *Principles of Terahertz Science and Technology*. New York: Springer, p. 76.

Here, $I(t)$ is the time dependent optical beam intensity. When the optical pulse duration is in picosecond or sub-picosecond level, the radiation frequency is within THz range.

Due to the fact that the second order nonlinear susceptibility $\chi^{(2)}$ vanishes in the centrosymmetric medium (Boyd, 2008), OR requires the material to have a noncentrosymmetric structure, such as the zinc-blende structure (see Fig. 6). The typical zinc-blende crystals that have been used in THz generations and detections include ZnTe, GaAs, GaP, etc.

Other than the susceptibility that is based on the crystal structure, the THz radiation efficiency, waveform and bandwidth depend on many factors, including laser beam properties, phase matching conditions, the crystal thickness, orientation, absorption and dispersion, etc.

In a nonlinear process such as optical rectification, one of the most crucial factors is the phase matching condition, which requires the energy and momentum conservation of the participating electromagnetic waves. For OR THz generation process, it can be described as follows:

$$\begin{cases} \omega_{O1} - \omega_{O2} = \Omega_{\text{THz}} \\ k_{O1} - k_{O2} = k_{\text{THz}} \end{cases} \quad (5)$$

where ω_{O1} and ω_{O2} are the two frequency components that participate in the OR, k_{O1} and k_{O2} are their corresponding wave vectors. If we divide the first equation with the second one, we get:

$$\frac{\partial \omega_O}{\partial k_O} = \frac{\Omega_{\text{THz}}}{k_{\text{THz}}} \quad (6)$$

$$v_{G,O} = v_{ph,\text{THz}} \quad (7)$$

where $v_{G,O}$ is the group velocity of the optical pulse, and $v_{ph,\text{THz}}$ is the phase velocity of THz. Simply put, when the optical pulse group velocity equals the phase velocity of THz wave, the phase matching condition of OR THz generation is satisfied. That is why this condition is also called velocity-matching condition.

ZnTe is the most popular crystal for THz generation, because in this material the group velocity of femtosecond laser pulse around 800 nm (Ti: Sapphire laser center wavelength) matches very well with phase velocity of THz wave. For the optical wavelength $\lambda = 812$ nm and the THz frequency 1.69 THz, $n_{gr}(812 \text{ nm}) = n_T(1.69 \text{ THz}) = 3.22$ in ZnTe (Lee, 2010). Table 1 lists the optical wavelengths for OR THz generation velocity-matching at 2 THz in several typical zinc-blende crystals.

If a crystal system is highly symmetric, many of its second order nonlinear susceptibility tensor elements vanish. Among the nonvanishing ones, only a few of them are independent. For example, ZnTe has the crystal class of $\bar{4}3m$, which has three nonvanishing contracted matrix (Boyd, 2008) elements and only one is independent:

$$d_{ij} = \begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix} \quad (8)$$

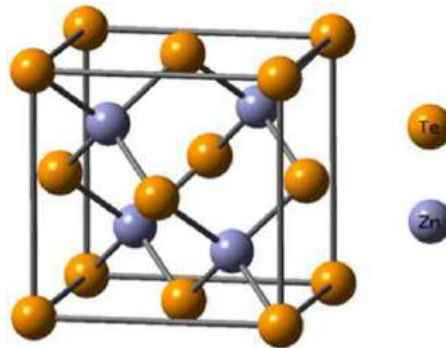


Fig. 6 ZnTe crystal structure. Reprinted from Kurban, M., Erkoç, Ş., 2016. Mechanical properties of CdZnTe nanowires under uniaxial stretching and compression: A molecular dynamics simulation study. Computational Materials Science 122, 295–300.

Table 1 Optical wavelength for velocity-matching in zinc-blende crystals at 2 THz

	<i>ZnTe</i>	<i>CdTe</i>	<i>GaP</i>	<i>InP</i>	<i>GaAs</i>
Wavelength (μm)	0.8	0.97	1.0	1.22	1.35

Source: From Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 90.

When a linear polarization optical beam is passing through the crystal, the OR induced polarization P can be expressed as follows:

$$\begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix} \begin{pmatrix} E_x^2 \\ E_y^2 \\ E_z^2 \\ 2E_yE_x \\ 2E_zE_x \\ 2E_zE_y \end{pmatrix} \quad (9)$$

where x, y, z scales represent the $[100]$, $[010]$ and $[001]$ crystal axis, respectively.

For a normal incidence case of a (110) cut ZnTe crystal, as shown in Fig. 7(a), a linearly polarized incident beam is propagating along the $[110]$ axis and the polarization direction has an angle θ with $[001]$ axis, then the THz radiation intensity can be expressed as (Lee, 2010):

$$I_{\text{THz}}(\theta) = \frac{3}{4} I_{\text{THz}}^{\text{max}} \sin^2 \theta (4 - 3 \sin^2 \theta) \quad (10)$$

where the $I_{\text{THz}}^{\text{max}}$ is achieved when $\theta = \sin^{-1} \sqrt{\frac{2}{3}}$. Fig. 7(b) plots the THz radiation intensity as a function of θ .

The spectral bandwidth of the THz generation from electro-optic (EO) crystals is mainly limited by the absorption at THz frequency of the materials. From 5–10 THz, the absorption is normally dominated by the transverse-optical (TO) phonon resonances, whereas at lower frequencies, the absorption is attributed to more complicated combination of second-order phonon processes. ZnTe, for example, has a strong fundamental TO phonon resonance at 5.32 THz. At 1.6 and 3.7 THz, there are two major absorption bands as a mixture of LO-LA (longitudinal-optical and -acoustical) differences at the X and L points of the Brillouin zone (Schall *et al.*, 2001). Fig. 8 shows the measured (solid line) absorption coefficient of ZnTe at THz frequency as well

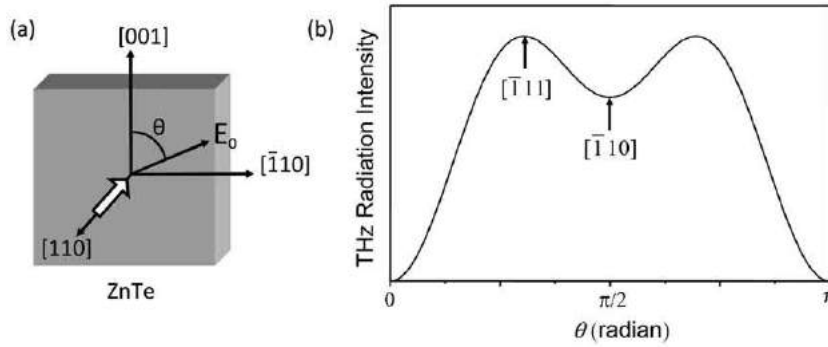


Fig. 7 (a) A linearly polarized optical wave is incident on a (110) ZnTe crystal with normal angle, θ is the angle between the optical field and the $[001]$ axis; (b) THz radiation intensity from ZnTe as a function of θ . Modified and reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 76.

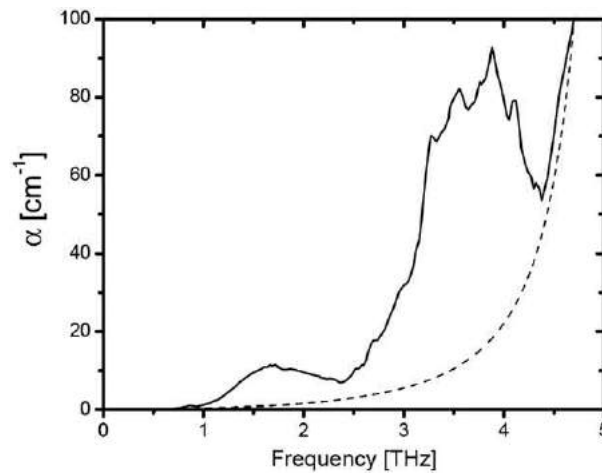


Fig. 8 Measured (solid line) power absorption coefficient (cm^{-1}) for ZnTe crystal compared with the calculated (dashed line) absorption for the TO-phonon line at room temperature. Reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 90.

Table 2 Lowest TO-phonon lines of EO crystals

	<i>ZnTe</i>	<i>CdTe</i>	<i>GaP</i>	<i>InP</i>	<i>GaAs</i>	<i>GaSe</i>	<i>LiNbO₃</i>	<i>LiTaO₃</i>
ν_{TO} (THz)	5.3 ^a	4.3 ^a	11 ^b	9.2 ^c	8.1 ^d	6.4 ^d	7.7 ^e	4.2 ^f

^aSchall *et al.* (2001).^bLeitenstorfer *et al.* (1999).^cDebernardi (1998).^dKuroda *et al.* (1987).^eSeiji *et al.* (2002).^fSeiji *et al.* (2003).

Source: From Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 91.

Table 3 Properties of common materials for OR (Wu and Zhang, 1996)

Crystal	EO coefficient (pmV ⁻¹)	Index of refraction	THz index of refraction	THz absorption coefficient (cm ⁻¹)
ZnTe	$r_{41}=4.0$ (0.633 μm)	2.85 (0.8 μm)	~ 3.17	1.3
LiNbO ₃	$r_{33}=30.9$ $r_{51}=32.6$ (0.633 μm)	$n_o=2.29$, $n_e=2.18$ (0.633 μm)	$n_o \sim 6.8$, $n_e \sim 4.98$	16
LiTaO ₃	$r_{33}=r_{51}=30.5$ (0.82 μm)	$n_o=2.176$, $n_e=2.18$ (0.633 μm)	$n_o \sim 6.5$, $n_e \sim 6.4$	46
CdTe	$r_{41}=4.5$ (1.00 μm)	2.84 (0.8 μm)	~ 3.23	4.8
DAST	$r_{11}=160$ (0.82 μm)	$n_o=2.46$, $n_e=1.70$ (0.820 μm)	~ 2.4	150
GaSe	1.7 (0.8 μm)	2.85 (0.8 μm)	~ 3.72	0.07
GaAs	$r_{41}=1.43$ (1.15 μm)	3.61 (0.886 μm)	~ 3.4	0.5

Source: From Hafez, H.A., Chai, X., Ibrahim, A. *et al.*, 2016. Intense terahertz radiation and their applications. Journal of Optics 18, 093004.

as the calculated absorption (dashed line) due to the TO phonon (Gallot *et al.*, 1999). Table 2 lists the lowest TO phonon lines of some commonly used EO crystals (Schall *et al.*, 2001; Leitenstorfer *et al.*, 1999; Debernardi, 1998; Kuroda *et al.*, 1987; Seiji *et al.*, 2002, 2003).

Tilted Pulse Front

Although ZnTe is one of the most popular EO crystals for THz generation, as shown in Table 3, its EO coefficient is relatively small compared with some other crystals, such as inorganic EO crystal LiNbO₃ and organic crystal 4-N-methylstilbazolium tosylate (DAST) (Hafez *et al.*, 2016). This section focuses on the tilted pulse front technique associated with THz generation from LiNbO₃, and THz generation from DAST and other organic crystals will be introduced in the next section.

Yang *et al.* (1971) first demonstrated the THz generation by OR with picosecond laser pulses, using LiNbO₃ crystal as the excited material. However, the conventional OR method of sending the pump beam at normal incidence to the crystal and generating THz radiation in the forward direction does not work efficiently with LiNbO₃, due to the large mismatch between the optical group velocity and the THz phase velocity in the material. The optical group refractive index is $n_o=2.3$, and the THz refractive index is $n_T=5.2$ (Lee, 2010). In order to conquer this mismatch, Hebling *et al.* (2002) proposed the pulse front tilting technique, which later became a standard procedure for THz generation from LiNbO₃ crystal.

In analogy to the Cherenkov radiation (Auston *et al.*, 1984b), in LiNbO₃ crystal, a femtosecond laser beam with considerably smaller beam size than the THz wavelength can be seen as a point source moving faster than the THz radiation, as shown in Fig. 9(a). If the optical beam size is not negligible compared to THz wavelength, when the pulse front is aligned with the Cherenkov cone, like in Fig. 9(b), the optical pulse front will propagate with the THz radiation at the same speed. Therefore, the velocity matching is fulfilled. This angle θ_c can be calculated using the optical group and THz phase refractive indices in LiNbO₃ as following:

$$\theta_c = \cos^{-1} \left(\frac{v_T}{v_O} \right) = \cos^{-1} \left(\frac{n_O}{n_T} \right) \cong 64^\circ \quad (11)$$

The tilt of the optical pulse front can be achieved with a diffractive grating (Hebling *et al.*, 2008). After the grating, an imaging system can be used to image the beam area on the grating onto the emitting surface of LiNbO₃ crystal. Fig. 10 shows a typical experimental setup for pulse front tilting THz generation.

LiNbO₃ based pulse front tilting setup is one of the most intense table-top ultrafast THz sources. Hirori *et al.* (2011) reported single cycle THz pulse with amplitude exceeding 1 MVcm⁻¹ at room temperature. Huang *et al.* (2013) demonstrated that with cryogenically cooled LiNbO₃, the THz generation energy conversion efficiency can be increased about 3 times compared to room temperature crystal. Lange *et al.* (2014) have used 1.5 MVcm⁻¹ intense THz field generated by cryogenically cooled LiNbO₃ to excite extremely nonperturbative nonlinearities in GaAs.

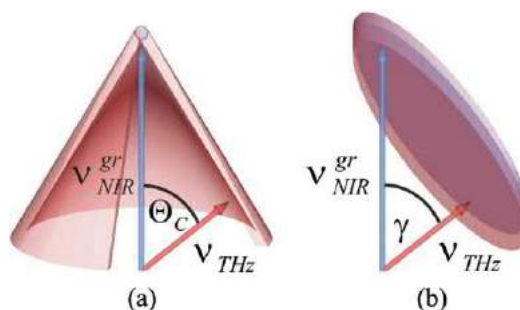


Fig. 9 (a) Cherenkov radiation: THz radiation is emitted as a cone with an angle θ_c . The velocity matching condition is satisfied if the excitation beam size is very small. (b) Pulse front tilting: velocity matching is automatically satisfied without upper limit of the excitation beam size, γ is the angle between the infrared beam velocity and the THz radiation velocity. Reprinted from Hebling, J., Almási, G., Kozma, I.Z., Kuhl, J., 2002. Velocity matching by pulse front tilting for large-area THz-pulse generation. *Optics Express* 10, 1161–1166.

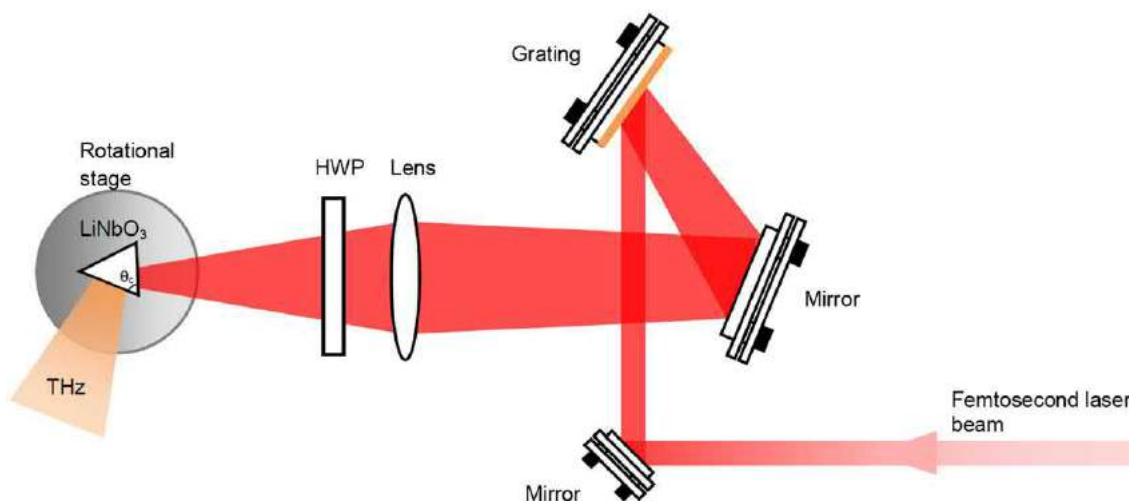


Fig. 10 Experimental setup for pulse tilting technique. HWP: Half Wave Plate.

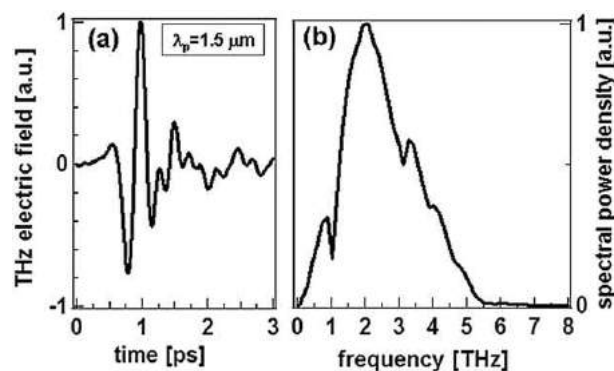


Fig. 11 (a) THz pulse from DAST. The corresponding THz spectrum is plotted in (b). Reprinted from Hauri, C.P., Ruchert, C., Vicario, C., Ardana, F., 2011. Strong-field single-cycle THz pulses generated in an organic crystal. *Applied Physical Letters* 99, 161116.

Organic Crystals

Organic crystals have recently become a great interest for ultra intense THz pulse generation. Zhang *et al.* (1992) first demonstrated the THz pulse generation from 4-N-methylstilbazolium tosylate (DAST). In 2011, a high quality THz beam generated from DAST with maximum electric field of 1.35 MVcm^{-1} was demonstrated (Hauri *et al.*, 2011), as shown in Fig. 11. However, although the THz spectrum extended up to 5 THz, a TO phonon absorption of DAST can be clearly seen at 1.1 THz. To avoid this phonon absorption, other than cooling the crystal to suppress the vibration modes, alternative organic crystals such as

2-(3-(4-hydroxystyryl)-5,5-dimethylcyclohex-2-enylidene)malononitrile (OH1) and 4-N,N-dimethylamino-40-N0-methylstilbazolium 2,4,6-trimethylbenzenesulfonate (DSTMS) can be used. In 2014, Shalaby *et al.* generated an extremely bright THz bullet with peak field up to 83 MVcm^{-1} (Shalaby and Hauri, 2015), which holds the current record of THz generation from an EO crystal. Meanwhile, the demanding conditions for growing good organic crystals in large dimension as well as the low damage threshold of organic crystals bring challenges to their wide use as THz sources.

Air-Plasma

Single-Color Method

Filamentation is the process by which a high-intensity beam self-focuses through nonlinear processes and collapses. This plasma channel stabilizes the beam at very small diameters (30 to 100 μm) and maintains high intensities over ranges much longer than the Rayleigh length of a traditional, geometrically focused beam. Fig. 12 shows a laser induced filamentation in the air.

Intense broadband THz generation from ionized gas drew a great amount of attention ever since it was first demonstrated by Hamster *et al.* (1993). In this work, strong femtosecond laser pulses were focused into a Helium gas target, generating strong emission of THz pulses with a conversion efficiency of less than 10^{-6} . The mechanism behind the THz radiation is the Ponderomotive force at the focus of the laser beam. The Ponderomotive force is experienced by a charged particle in an inhomogeneous oscillating electro-magnetic field, pushing the particle toward the area of low optical intensity regardless of the sign of the charge (Morales and Lee, 1974). However, since the ions are many orders of magnitude heavier compared to electrons, these forces generate a large density difference between ionic and electronic charges, and this charge separation results in a powerful electromagnetic transient, as schematically shown in Fig. 13.

The following relation between the emitted THz power from a one-color plasma and the excitation laser parameters was proposed by Hamster *et al.* (1994):

$$P_{\text{THz}} \propto \left(\frac{W}{R_0}\right)^2 \left(\frac{\lambda}{\tau}\right)^4 \quad (12)$$

where W is the laser pulse energy, R_0 is the $1/e^2$ radius of the laser beam at the focus, λ is laser wavelength and τ is the pulse duration. Therefore, this equation predicts a strong dependence of the THz power on the wavelength and the pulse duration of the excitation laser. In the presence of strong bias field, the generation efficiency can be improved to comparable to that of THz radiation from semiconductor surface (Löffler *et al.*, 2000).

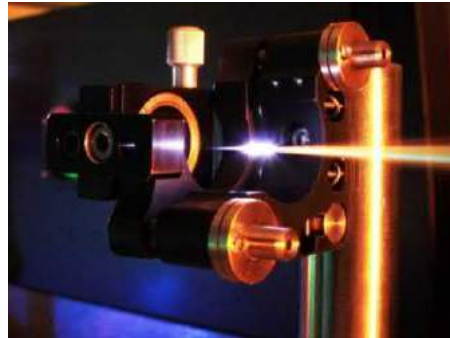


Fig. 12 Laser-induced air plasma (center bright line) emits an intense THz field.

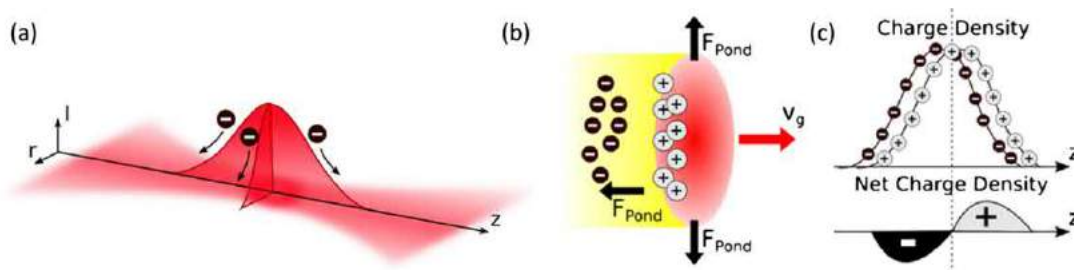


Fig. 13 (a) Laser propagates along z direction, the Ponderomotive force created by the gradient of the electric field drives free electrons. (b) Ponderomotive force drives electrons away in the plasma, leave the heavy ions relatively stationary. (c) Net charge density creates electrical dipole along z direction (Liu *et al.*, 2015).

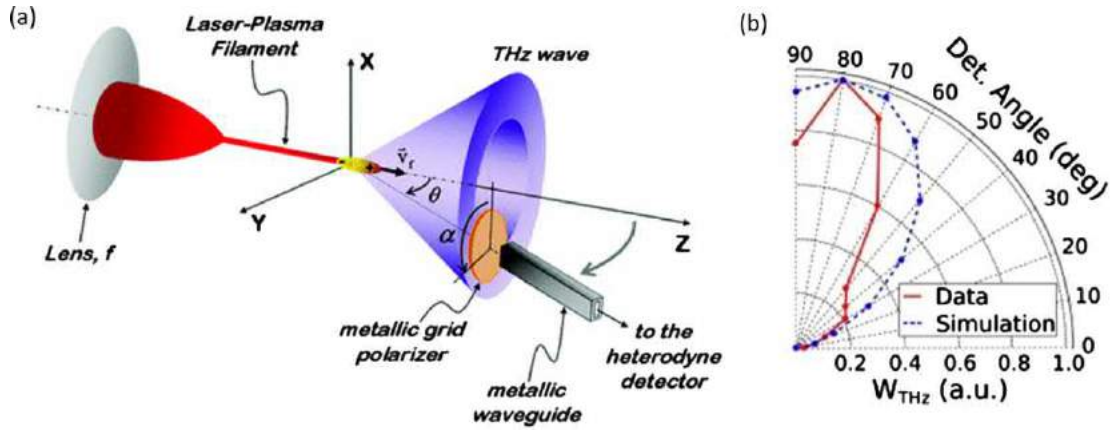


Fig. 14 (a) The experimental setup for measuring the THz radiation pattern from a laser-induced filament (Reprinted from D'Amico, C., Houard, A., Franco, M., *et al.*, 2007. Conical forward THz emission from femtosecond-laser-beam filamentation in air. *Physical Review Letters* 98, 235002). (b) THz radiation pattern from a microplasma measured (red solid line) and calculated from Eq. (13) (blue dotted line). Reprinted from Buccheri, F., Zhang, X.-C., Terahertz emission from laser-induced microplasma in ambient air. *Optica* 2, 366–369.

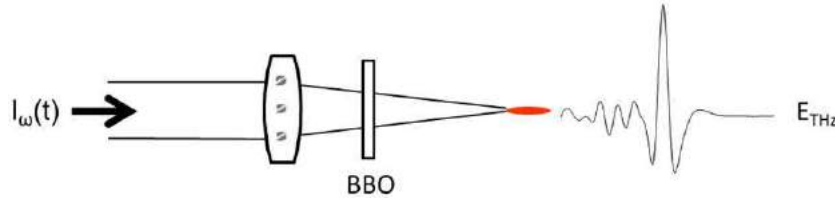


Fig. 15 Schematic demonstration of experimental setup for THz generation from two-color laser induced air plasma.

At the focus of the laser beam, the ionization front is moving in the wake of the laser pulse at light velocity, therefore generate THz radiation in a Cherenkov radiation cone shape in the forward direction, as shown in Fig. 14(a). The angular energy density distribution of the THz radiation can actually be expressed as (Amico *et al.*, 2008):

$$f(\omega, \theta, L) = \frac{\sin^2 \theta}{(1 - \cos \theta)^2} \sin^2 \left(\frac{L\omega}{2c} (1 - \cos \theta) \right) \quad (13)$$

where ω is the radiated THz frequency, θ is the emission angle, L is the longitudinal length of the plasma column, and c is the speed of light. As we can see, the THz radiation angular distribution has something to do with the plasma length. To explore the extreme case of THz radiation from "microplasma," where the plasma length is at μm scale, Buccheri *et al.* demonstrated that with such a small length of plasma, the THz radiation cone angle opens up to about 80 degrees (Buccheri and Zhang, 2015), as shown in Fig. 14(b).

Two-Color Method

All the THz generation schemes from gas plasma mentioned in last section use only one color laser pulses as the excitations. However, the mixing of laser pulses (fundamental ω) and their second harmonics (SH) 2ω to create the plasma has shown to provide enhancement of THz generation efficiency by several orders of magnitude (Cook and Hochstrasser, 2000). Fig. 15 illustrates a basic experimental setup for the two-color air plasma generation method. The SH of the laser pulse is usually generated by using a type-I β -barium borate (BBO) crystal. The THz radiation intensity is maximized when the fundamental and SH polarizations are parallel, and is almost negligible when they are perpendicular (Kress *et al.*, 2004; Xie *et al.*, 2006).

THz radiation from two-color laser induced air plasma can be phenomenologically described as Four Wave Mixing (FWM) process, which is a third order nonlinear optical process. The model can be expressed as:

$$E_{THz}(t) \propto \chi^{(3)} E_{2\omega}(t) E_{\omega}^*(t) E_{\omega}^*(t) \quad (14)$$

where E_{THz} , E_{ω} and $E_{2\omega}$ are respectively the THz, the fundamental and the second harmonic (SH) electric field and $\chi^{(3)}$ is an effective third-order nonlinear susceptibility of the plasma filament. This relation has been experimentally confirmed by Xie *et al.* (2006) (see Fig. 16). It indicates that above the ionization threshold, the generated THz field is proportional to intensity of the fundamental laser and is also proportional to the square root of the second harmonic laser.

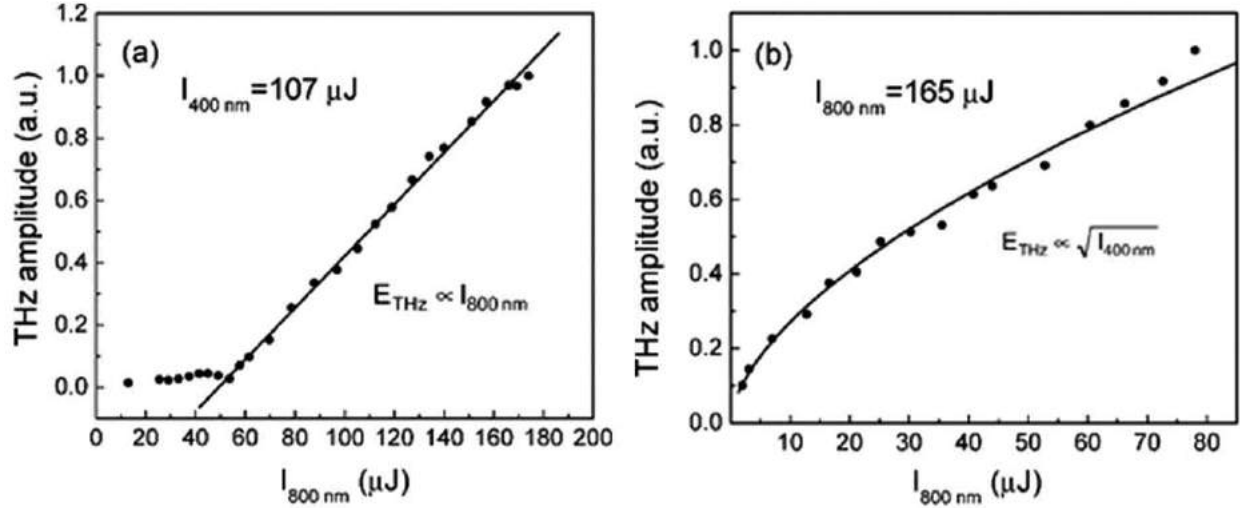


Fig. 16 (a) THz field as a function of laser intensities of 800 nm laser beam (a) and 400 nm laser beam (b) in THz wave generation by combination of 800 and 400 nm fs laser beams. Reprinted from Xie, X. Dai, J., Zhang, X.-C., 2006. Coherent control of THz wave generation in ambient air. *Physical Review Letters* 96, 075005.

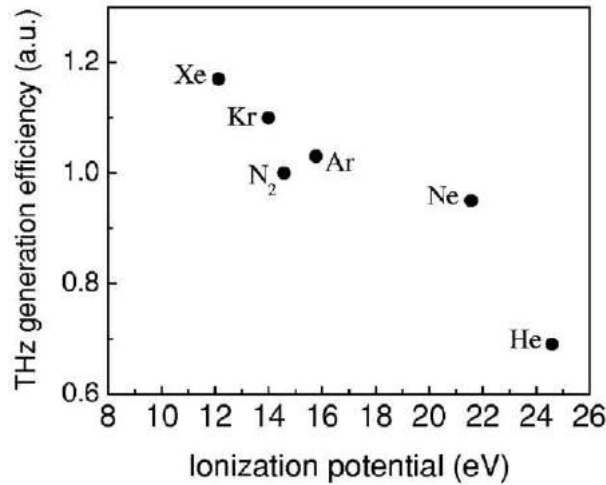


Fig. 17 Terahertz wave generation efficiency of six inert gases vs ionization potential at 20 Torr. Reprinted from Chen, Y., Yamaguchi, M., Wang, M., Zhang, X.-C., 2007. Terahertz pulse generation from noble gases. *Applied Physics Letters* 91, 251116.

Other than the ambient air, noble gases can also be used in the two-color plasma THz generation scheme. As shown in [Fig. 17](#), the generation efficiencies from various noble gases highly depend on their ionization potentials ([Chen et al., 2007](#)). The strongest THz emitter among them appeared to be Xenon.

The generated THz field from two-color air plasma can be modified by changing the phase between the fundamental pulses and the SH pulses. In the FWM model, this can be expressed as:

$$E_{\text{THz}}(t) \propto \chi^{(3)} E_{2\omega}(t) E_{\omega}^*(t) E_{\omega}^*(t) \cos((\varphi)) \quad (15)$$

where (φ) denotes the phase difference between ω and 2ω . Due to the dispersion of the two colors in air, the phase shift between these two color beams varies as they propagate. Therefore, the phase (φ) can be easily controlled by shifting the β -BBO longitudinally ([Kress et al., 2004](#)). However, due to the laser intensity change toward the focus and the spatial limit between the focusing optic and the focus, simply shifting the β -BBO cannot control the phase (φ) precisely or significantly enough for certain purposes, such as remote generation of THz ([Dai et al., 2011](#)). In those cases, phase compensators including a pair of silica wedges were exploited in order to fine tune the phases and polarizations of the ω and 2ω beams ([Dai et al., 2009](#)). [Fig. 18](#) lists two types of phase compensators, which respectively control the two color beams in one beam path or in two separate paths. With the help of phase compensator, the generated THz field strength as well as the pulse polarity can be modified explicitly, as shown in [Fig. 19](#)

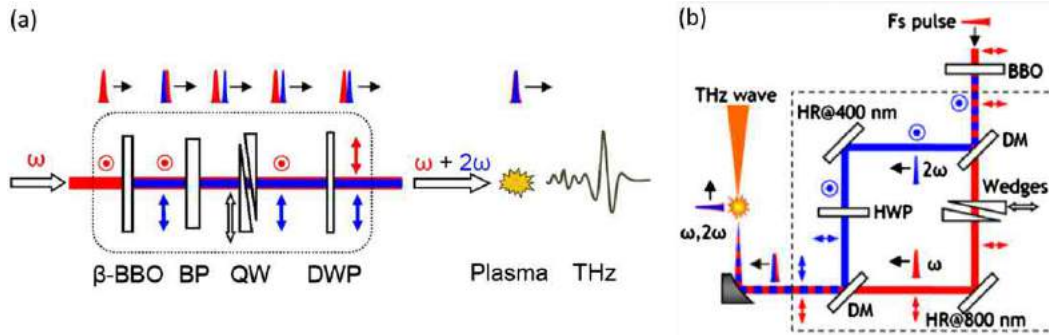


Fig. 18 (a) Inside the dashed line is the in-line PC (phase compensator). (b) Schematic illustration of the PC incorporated with a wedge pair: DM used to separate or recombine ω and 2ω beams; HWP used to control the polarization of the 2ω beam. DWP, dual-wavelength waveplate; BP, birefringent plate (α -BBO); QW, quartz wedges; DM, dichroic mirror. Reprinted from Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. *Physical Review Letters* 103, 023001 and Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. *IEEE Journal of selected topics in Quantum Electronics* 17, 183.

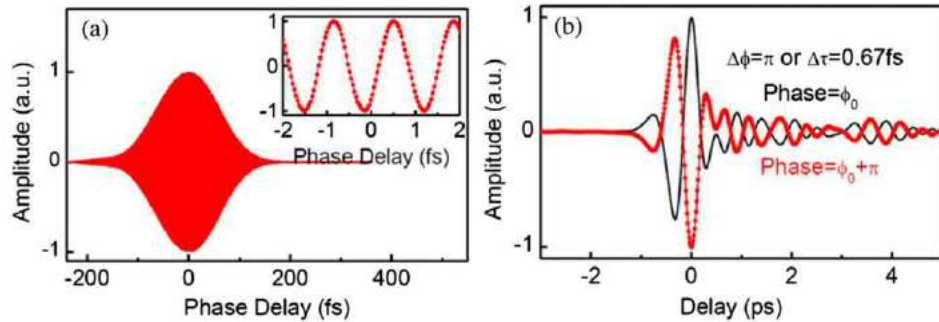


Fig. 19 (a) Phase scan obtained by continuously changing the insertion of the fused silica wedge pair while keeping the delay line for the THz waveform at the peak THz electric field. (b) Two THz waveforms with opposite polarities due to π phase shift induced by changing the relative insertion of the wedge pair in the PC. $\Delta(\phi)$ is the phase change on relative phase difference between the two colors to generate these two THz waveforms, and $\Delta\tau$ is the corresponding time delay. Reprinted from Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. *IEEE Journal of selected topics in Quantum Electronics* 17, 183.

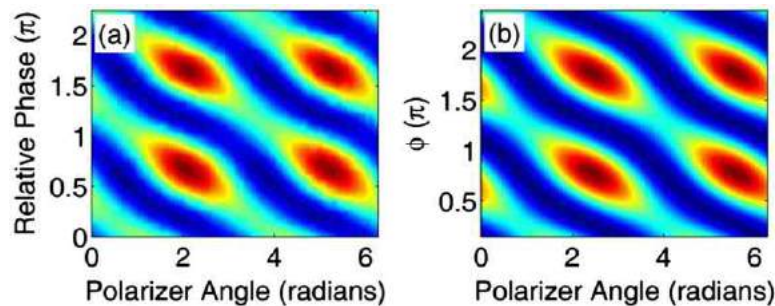


Fig. 20 THz intensity vs. THz polarizer angle and the relative phase between the ω and 2ω pulses with left-handed circularly polarized ω pulse and elliptically polarized 2ω pulse. (a) the experimental results, and (b) the corresponding simulation result. Reprinted from Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. *Physical Review Letters* 103, 023001.

(Dai *et al.*, 2011). Further examination on how the various pump pulse polarizations affect the THz generation field was also enabled due to the phase compensator technique, as shown in Fig. 20 (Dai *et al.*, 2009).

With ultra-short femtosecond laser pulses, an ultra-broadband THz wave with spectrum extending up to 30 THz can be achieved (Clough *et al.*, 2012; Kim *et al.*, 2008). Such a broad bandwidth requires broadband THz detection methods, such as THz-air-biased-coherent-detection (ABCD) (Dai *et al.*, 2011) or interferometric detection (Kim *et al.*, 2008), to reflect the complete

spectral properties of the THz source. Fig. 21 compares the THz radiation from air-plasma detected by ZnTe crystal and THz-ABCD method. In 2014, maximum THz fields from two-color laser induced plasma reaching 8 MV/cm has been demonstrated (Oh *et al.*, 2014), covering the spectrum from 0.1–10 THz.

Although many experiments support the FWM model as the principle mechanism for THz generation from two-color laser induced air-plasma, the third order nonlinearity susceptibility $\chi^{(3)}$ of bond or free electrons is too small to explain the high intensity of THz radiation. Therefore, a transient photocurrent model was developed to explain coherent terahertz emission from air excited by a symmetry-broken laser field composed of ω and 2ω laser pulses (Kim *et al.*, 2007, 2008). In this model, a nonvanishing transverse plasma current is produced when the bound electrons are stripped off the ions by an asymmetric laser field, e.g. a mixture of two-color laser field with a proper relative phase. This photocurrent transient, occurring on the timescale of the photoionization, can thus produce electromagnetic radiation at THz frequencies. The electron dynamics right after the ionization were treated classically by Kim *et al.* (2007, 2008), whereas a full quantum mechanical simulation was reported by Karpowicz and Zhang (Dai *et al.*, 2009).

For elongated two-color filamentation, the THz radiation forms a cone shape due to off-axis phase matching (You *et al.*, 2012). The THz yield and angular distribution depend highly on the filament dimensions, and the dephasing length l_d , over which the THz radiation polarity remains the same. l_d can be expressed as

$$l_d = \left(\frac{\lambda}{2}\right)(n_\omega - n_{2\omega})^{-1} \quad (16)$$

where λ is the wavelength at ω , n_ω and $n_{2\omega}$ are respectively the refractive index at frequency ω and 2ω . For filament with electron density of $\sim 10^{16} \text{ cm}^{-3}$ in ambient air, the dephasing length is about 22 mm.

As shown in Fig. 22, for a filament comparable to or shorter than the dephasing length, THz waves generated along the filament have both positive polarity and negative polarity. Those components interfere with each other constructively or destructively in the

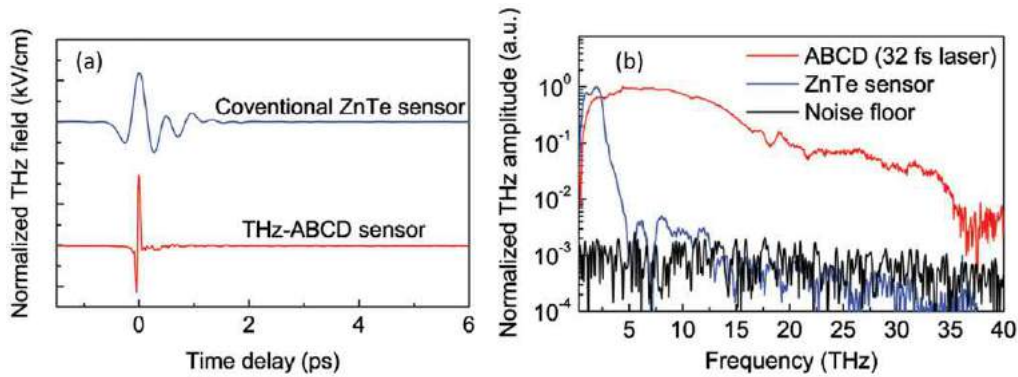


Fig. 21 (a) Time-resolved THz signals generated and detected using dry nitrogen gas as compared to conventional EO crystal detection in ZnTe. The probe beam for air detection has energy of 85 μJ and pulse duration of 32 fs. (b) Corresponding spectra after Fourier transform. Reprinted from Clough, B., Dai, J., Zhang, X.-C., 2012. Laser air photonics: Beyond the terahertz gap. *Materials Today* 15, 50–58.

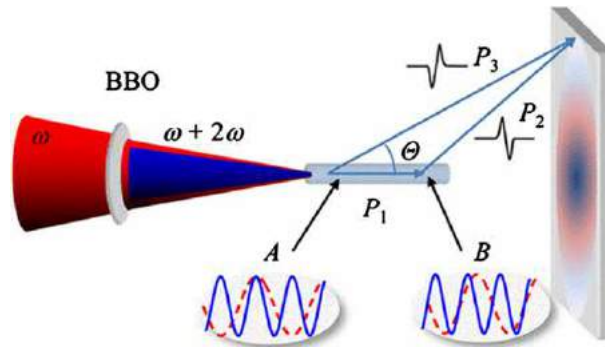


Fig. 22 Schematic of THz emission from a long, two-color laser-induced filament. The phase difference between 800 nm (dashed red curves) and 400 nm (solid blue curves) pulses along the filament results in a periodic oscillation of microscopic current amplitude and polarity. The resulting far-field THz radiation is determined by interference between the waves emitted from the local sources along the filament. P_1 , P_2 and P_3 are respectively the optical path along the different directions as shown in the figure, θ is the angle between P_3 and P_1 . Reprinted from You, Y.S., Oh, T.I., Kim, K.Y., 2012. Off-axis phase-matched terahertz emission from two-color laser-induced plasma filaments. *Physical Review Letters* 109, 183902.

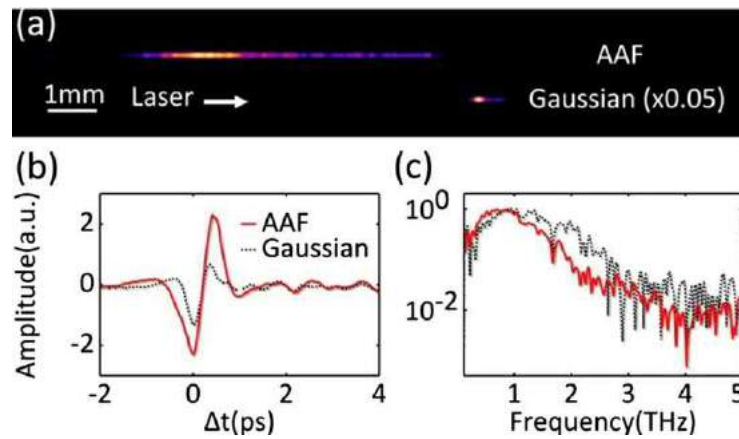


Fig. 23 (a) Fluorescence image comparison between an AAF beam plasma and a Gaussian plasma (intensity reduced by 20 times), both generated with pump pulse energy of 0.5 mJ. (b) Measured THz waveforms and (c) their corresponding normalized spectra generated by the two plasmas in (a). Reprinted from Liu, K., Koulouklidis, A.D., Papazoglou, D.G., Tzortzakis, S., Zhang, X.-C., 2016. Enhanced terahertz wave emission from air-plasma tailored by abruptly autofocusing laser beams. *Optica* 3, 605.

Table 4 List of peak THz electric field strength ranges from different sources. LA-PC, large-aperture photoconductive antenna; LA-EO, large-aperture electro-optic crystal

THz sources	1–10 kVcm^{-1}	10–100 kVcm^{-1}	0.1–1 MVcm^{-1}	1–10 MVcm^{-1}	10–100 MVcm^{-1}
LA-PC (Ropagnol <i>et al.</i> , 2016)		331 kVcm^{-1}			
Free electron lasers (Wu <i>et al.</i> , 2013)					44 MVcm^{-1}
LA-EO (Blanchard <i>et al.</i> , 2007)		230 kVcm^{-1}			
LiNbO ₃ (Lange <i>et al.</i> , 2014)			1.5 MVcm^{-1}		
Organic crystal (Shalaby and Hauri, 2015)					83 MVcm^{-1}
Air Photonics (Oh <i>et al.</i> , 2014)				8 MVcm^{-1}	

far field, depending on their path difference which is related to the radiation angle, forming a donut shape far field radiation profile. You *et al.* (2012) also found out that the generation yield increases almost linearly with the plasma filament length, so one can increase the THz generation by simply extending the plasma length. Recently, different methods to tailor the filamentation toward more intense THz generation have been reported, including concatenating two plasmas into one (Manceau *et al.*, 2010) and using an abruptly autofocusing (AAF) optical beam (Liu *et al.*, 2016) (Fig. 23).

Summary

After introducing the main broadband THz pulse emitters, in this section we would like to present a chart comparing the maximum THz radiation amplitude of each source as of 2017 (Zhang *et al.*, 2017). Notice the large-aperture (LA) EO crystals listed do not include LiNbO₃-based tilted pulse front setup or organic crystals (Table 4).

See also: Strong-Field Terahertz Excitations in Semiconductors. Terahertz Lasers

References

- Abo-Bakr, M., Feikes, J., Holldack, K., Wüstefeld, G., Hübers, H.W., 2002. Steady-state far-infrared coherent synchrotron radiation detected at BESSY II. *Physical Review Letters* 88, 254801.
- Amico, C.D., Houard, A., Akturk, S., *et al.*, 2008. Forward THz radiation emission by femtosecond filamentation in gases: Theory and experiment. *New Journal of Physics* 10, 013015.
- Auston, D.H., Cheung, K.P., Smith, P.R., 1984a. Picosecond photoconducting Hertzian dipoles. *Applied Physics Letters* 45, 284–286.
- Auston, D.H., Cheung, K.P., Valdmann, J.A., Kleinman, D.A., 1984b. Cherenkov radiation from femtosecond optical pulses in electro-optic media. *Physical Review Letters* 53, 1555–1558.
- Blanchard, F., Razzari, L., Bandulet, H.C., *et al.*, 2007. Generation of 1.5 μJ single-cycle terahertz pulses by optical rectification from a large aperture ZnTe crystal. *Optics Express* 15, 13212–13220.

- Boyd, R.W., 2008. *Nonlinear Optics*. Elsevier Science.
- Buccheri, F., Zhang, X.-C., 2015. Terahertz emission from laser-induced microplasma in ambient air. *Optica* 2, 366–369.
- Carr, G.L., Martin, M.C., McKinney, W.R., *et al.*, 2002. High-power terahertz radiation from relativistic electrons. *Nature* 420, 153–156.
- Chen, Y., Yamaguchi, M., Wang, M., Zhang, X.-C., 2007. Terahertz pulse generation from noble gases. *Applied Physics Letters* 91, 251116.
- Clough, B., Dai, J., Zhang, X.-C., 2012. Laser air photonics: Beyond the terahertz gap. *Materials Today* 15, 50–58.
- Cook, D.J., Hochstrasser, R.M., 2000. Intense terahertz pulses by four-wave rectification in air. *Optics Letters* 25, 1210.
- Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. *Physical Review Letters* 103, 023001.
- Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. *IEEE Journal of selected topics in Quantum Electronics* 17, 183.
- Debernardi, A., 1998. Phonon linewidth in III-V semiconductors from density-functional perturbation theory. *Physical Review B* 57, 12847–12858.
- Dexheimer, S.L., 2007. *Terahertz Spectroscopy: Principles and Applications*. Taylor & Francis.
- Dhillon, S.S., Vitiello, M.S., Linfield, E.H., *et al.*, 2017. The 2017 terahertz science and technology roadmap. *Journal of Physics D: Applied Physics* 50, 043001.
- Gallot, G., Zhang, J., McGowan, R.W., Jeon, T.-I., Grischkowsky, D., 1999. Measurements of the THz absorption and dispersion of ZnTe and their relevance to the electro-optic detection of THz radiation. *Applied Physics Letters* 74, 3450–3452.
- Gu, P., Tani, M., Kono, S., Sakai, K., Zhang, X.-C., 2002. Study of terahertz radiation from InAs and InSb. *Journal of Applied Physics* 91, 5533–5537.
- Hafez, H.A., Chai, X., Ibrahim, A., *et al.*, 2016. Intense terahertz radiation and their applications. *Journal of Optics* 18, 093004.
- Hale, P.J., Madeo, J., Chin, C., *et al.*, 2014. 20 THz broadband generation using semi-insulating GaAs interdigitated photoconductive antennas. *Optics Express* 22, 26358–26364.
- Hamster, H., Sullivan, A., Gordon, S., Falcone, R.W., 1994. Short-pulse terahertz radiation from high-intensity-laser-produced plasmas. *Physical Review E* 49, 671–677.
- Hamster, H., Sullivan, A., Gordon, S., White, W., Falcone, R., 1993. Subpicosecond, electromagnetic pulses from intense laser-plasma interaction. *Physical Review Letters* 71, 2725.
- Hauri, C.P., Ruchert, C., Vicario, C., Ardana, F., 2011. Strong-field single-cycle THz pulses generated in an organic crystal. *Applied Physics Letters* 99, 161116.
- Hebling, J., Almási, G., Kozma, I.Z., Kuhl, J., 2002. Velocity matching by pulse front tilting for large-area THz-pulse generation. *Optics Express* 10, 1161–1166.
- Hebling, J., Yeh, K.-L., Hoffmann, M.C., Bartal, B., Nelson, K.A., 2008. Generation of high-power terahertz pulses by tilted-pulse-front excitation and their application possibilities. *Journal of the Optical Society of America B* 25, B6–B19.
- Hirori, H., Doi, A., Blanchard, F., Tanaka, K., 2011. Single-cycle terahertz pulses with amplitudes exceeding 1 MV/cm generated by optical rectification in LiNbO₃. *Applied Physics Letters* 98, 091106.
- Huang, S.-W., Granados, E., Huang, W.R., *et al.*, 2013. High conversion efficiency, high energy terahertz pulses by optical rectification in cryogenically cooled lithium niobate. *Optics Letters* 38, 796–798.
- Hu, B.B., Zhang, X.C., Auston, D.H., Smith, P.R., 1990. Free-space radiation from electro-optic crystals. *Applied Physics Letters* 56, 506–508.
- Kim, K.Y., Glowina, J.H., Taylor, A.J., Rodriguez, G., 2007. Terahertz emission from ultrafast ionizing air in symmetry-broken laser fields. *Optics Express* 15, 4577–4584.
- Kim, K.Y., Taylor, A.J., Glowina, J.H., Rodriguez, G., 2008. Coherent control of terahertz supercontinuum generation in ultrafast laser–gas interactions. *Nature Photonics* 2, 605.
- Kress, M., Ler, T.L., Eden, S., Thomson, M., Roskos, H.G., 2004. Terahertz-pulse generation by photoionization of air with laser pulses composed of both fundamental and second-harmonic waves. In: *Optics Letters*, 29. . p. 1120.
- Kuroda, N., Ueno, O., Nishina, Y., 1987. Lattice-dynamical and photoelastic properties of GaSe under high pressures studied by Raman scattering and electronic susceptibility. *Physical Review B* 35, 3860–3870.
- Lange, C., Maag, T., Hohenleutner, M., *et al.*, 2014. Extremely nonperturbative nonlinearities in GaAs driven by atomically strong terahertz fields in gold metamaterials. *Physical Review Letters* 113, 227401.
- Lee, Y.S., 2010. *Principles of Terahertz Science and Technology*. New York: Springer.
- Leitenstorfer, A., Hunsche, S., Shah, J., Nuss, M.C., Knox, W.H., 1999. Detectors and sources for ultrabroadband electro-optic sampling: Experiment and theory. *Applied Physics Letters* 74, 1516–1518.
- Liu, K., Buccheri, F., Zhang, X.-C., 2015. THz science and technology of micro-plasma. *Physics (Chinese Wuli)* 44, 497–502.
- Liu, K., Koulouklidis, A.D., Papazoglou, D.G., Tzortzakos, S., Zhang, X.-C., 2016. Enhanced terahertz wave emission from air-plasma tailored by abruptly autofocusing laser beams. *Optica* 3, 605.
- Löffler, T., Jacob, F., Roskos, H.G., 2000. Generation of terahertz pulses by photoionization of electrically biased air. *Applied Physics Letters* 77, 453–455.
- Manceau, J.M., Massauti, M., Tzortzakos, S., 2010. Strong terahertz emission enhancement via femtosecond laser filament concatenation in air. *Optics Letters* 35, 2424–2426.
- Morales, G.J., Lee, Y.C., 1974. Ponderomotive-force effects in a nonuniform plasma. *Physical Review Letters* 33, 1016–1019.
- Oh, T.I., Yoo, Y.J., You, Y.S., Kim, K.Y., 2014. Generation of strong terahertz fields exceeding 8 MV/cm at 1 kHz and real-time beam profiling. *Applied Physics Letters* 105, 041103.
- Ropagnol, X., Khorasaninejad, M., Raeiszadeh, M., *et al.*, 2016. Intense THz pulses with large ponderomotive potential generated from large aperture photoconductive antennas. *Optics Express* 24, 11299–11311.
- Schall, M., Walther, M., Uhd Jepsen, P., 2001. Fundamental and second-order phonon processes in CdTe and ZnTe. *Physical Review B* 64, 094301.
- Seiji, K., Hideaki, K., Seizi, N., Mitsuo Wada, T., 2003. Dielectric properties of ferroelectric lithium tantalate crystals studied by terahertz time-domain spectroscopy. *Japanese Journal of Applied Physics* 42, 6238.
- Seiji, K., Naoki, T., Hideaki, K., Mitsuo Wada, T., Seizi, N., 2002. Terahertz time domain spectroscopy of phonon-polaritons in ferroelectric lithium niobate crystals. *Japanese Journal of Applied Physics* 41, 7033.
- Shalaby, M., Hauri, C.P., 2015. Demonstration of a low-frequency three-dimensional terahertz bullet with extreme brightness. *Nature Communications* 6, 5976.
- Shen, Y., Watanabe, T., Arena, D.A., *et al.*, 2007. Nonlinear cross-phase modulation with intense single-cycle terahertz pulses. *Physical Review Letters* 99, 043901.
- Tani, M., Matsuura, S., Sakai, K., Nakashima, S.-I., 1997. Emission characteristics of photoconductive antennas based on low-temperature-grown GaAs and semi-insulating GaAs. *Applied Optics* 36, 7853–7859.
- Wu, Q., Zhang, X.C., 1996. Ultrafast electro-optic field sensors. *Applied Physics Letters* 68, 1604–1606.
- Wu, Z., Fisher, A.S., Goodfellow, J., *et al.*, 2013. Intense terahertz pulses from SLAC electron beams using coherent transition radiation. *Review of Scientific Instruments* 84, 022701.
- Xie, X., Dai, J., Zhang, X.-C., 2006. Coherent control of THz wave generation in ambient air. *Physical Review Letters* 96, 075005.
- Yang, K.H., Richards, P.L., Shen, Y.R., 1971. Generation of far-infrared radiation by picosecond light pulses in LiNbO₃. *Applied Physics Letters* 19, 320–323.
- You, Y.S., Oh, T.I., Kim, K.Y., 2012. Off-axis phase-matched terahertz emission from two-color laser-induced plasma filaments. *Physical Review Letters* 109, 183902.
- Zhang, X.C., Hu, B.B., Darrow, J.T., Auston, D.H., 1990. Generation of femtosecond electromagnetic pulses from semiconductor surfaces. *Applied Physics Letters* 56, 1011–1013.
- Zhang, X.C., Ma, X.F., Jin, Y., *et al.*, 1992. Terahertz optical rectification from a nonlinear organic crystal. *Applied Physics Letters* 61, 3080–3082.
- Zhang, X.C., Shkurinov, A., Zhang, Y., 2017. Extreme terahertz science. *Nature Photonics* 11, 16–18.
- Zhang, X.-C., Xu, J., 2010. *Introduction to THz Wave Photonics*. New York: Springer.

Terahertz Detectors

Antoni Rogalski, Military University of Technology, Warsaw, Poland

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Terahertz (THz) range of electromagnetic spectrum still presents a challenge for both electronic and photonic technologies and is often described as the final unexplored area of spectrum. This radiation is frequently treated as the spectral region within frequency range (ν) ≈ 0.1 – 10 THz.

The past 20 years have seen a revolution in THz systems, as advanced materials research provided new and higher-power sources, and the potential of THz for advanced physics research and commercial applications was demonstrated. Numerous recent breakthroughs in the field have pushed THz research into the centre stage. As examples of milestone achievements can be included the development of THz time-domain spectroscopy (TDS), THz imaging, and high-power THz generation by means of nonlinear effects (Siegel, 2002; Bründermann *et al.*, 2012; Saeedkia, 2013; Sizov and Rogalski, 2010). Research involved with THz technologies is now receiving increasing attention, and devices exploiting this wavelength band are set to become increasingly important in diverse range of human activity applications (e.g., security, biological, drugs and explosion detection, gases fingerprints, imaging, etc.). Nowadays, the THz technology is also of much use in fundamentals science, such as nanomaterials science and biochemistry. This is based on the fact that THz frequencies correspond to single and collective excitations in nanoelectronic devices and collective dynamics in biomolecules.

General Classification of Terahertz Detectors

The majority of detectors can be classified in two broad categories: photon detectors and thermal detectors.

Photon Detectors

In photon detectors the radiation is absorbed within the material by interaction with electrons either bound to lattice atoms or to impurity atoms or with free electrons. The radiation can be also absorbed by electrons localized in the quantum wells or in the minibands of superlattices. The observed electrical output signal results from the changed electronic energy distribution. The photon detectors show a selective wavelength dependence of response per unit incident radiation power.

Depending on the nature of the interaction, the class of photon detectors is further sub-divided into different types. The most important are: intrinsic detectors, extrinsic detectors, and photoemissive (Schottky barriers). Different types of detectors are described in detail in the monograph *Infrared Detectors* (Rogalski, 2011). Fig. 1 shows spectral detectivity characteristics of different types of detectors.

Photodetectors that utilize excitation of an electron from the valence to conduction band are called intrinsic detectors. Those which operate by exciting electrons into the conduction band or holes into the valence band from impurity states within the band (impurity-bound states in energy gap, quantum wells or quantum dots), are called extrinsic detectors. In comparison with intrinsic photoconductivity, the extrinsic photoconductivity is far less efficient because of limits in the amount of impurity that can be introduced into semiconductor without altering the nature of the impurity states. Intrinsic detectors are most common at the short wavelengths, below $20\text{ }\mu\text{m}$.

A key difference between intrinsic and extrinsic detectors is that extrinsic detectors require much cooling to achieve high sensitivity at a given spectral response cutoff in comparison with intrinsic detectors. Low-temperature operation is associated with longer-wavelength sensitivity in order to suppress noise due to thermally induced transitions between close-lying energy levels. The long wavelength cutoff can be approximated as $T_{\text{max}} = 300\text{ K}/\lambda_c [\mu\text{m}]$, where λ_c is the cut-off wavelength.

Thermal Detectors

The second class of detectors is composed of thermal detectors – their operation principles are briefly described in Table 1. In a thermal detector the incident radiation is absorbed to change the material temperature, and the resultant change in some physical property is used to generate an electrical output. The detector is suspended on lags, which are connected to the heat sink (see figures inside of Table 1). The signal does not depend upon the photonic nature of the incident radiation. Thus, thermal effects are generally wavelength independent; the signal depends upon the radiant power (or its rate of change) but not upon its spectral content. Attention is directed toward three approaches which have found the greatest utility in infrared technology, namely, bolometers, pyroelectric, and thermoelectric effects. In pyroelectric detectors a change in the internal electrical polarization is measured, whereas in the case of thermistor bolometers a change in the electrical resistance is measured.

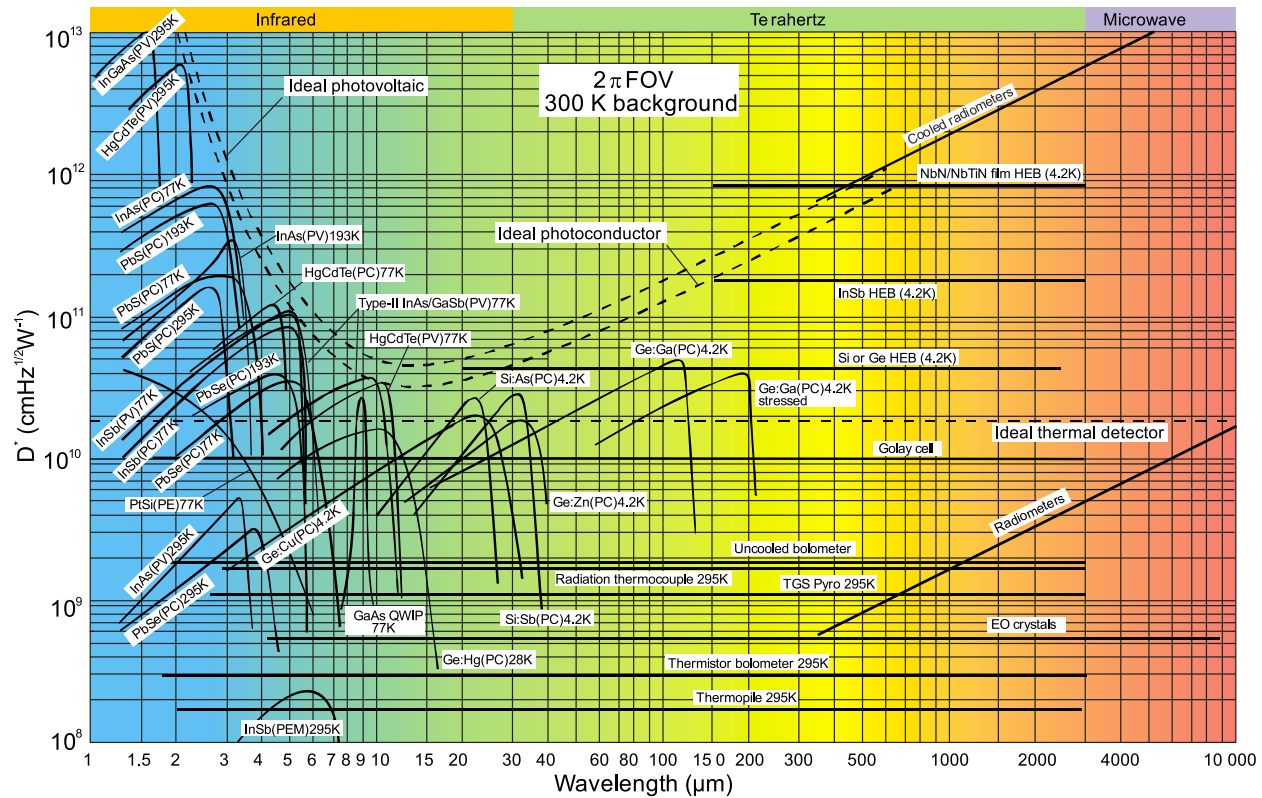


Fig. 1 Comparison of the D^* of various available detectors when operated at the indicated temperature. Chopping frequency is 1000 Hz for all detectors except the thermopile (10 Hz), thermocouple (10 Hz), thermistor bolometer (10 Hz), Golay cell (10 Hz) and pyroelectric detector (10 Hz). Each detector is assumed to view a surrounding hemisphere (2π field of view) at a temperature of 300 K. Theoretical curves for the background-limited D^* (dashed lines) for ideal photovoltaic and photoconductive detectors and thermal detectors are also shown. PC, photoconductive detector; PEM, photoelectromagnetic detector; PV, photovoltaic detector; HEB, hot electron bolometer.

Bolometers may be divided into several types. The most commonly used are the metal, the thermistor, and the semiconductor bolometers. A fourth type is the superconducting bolometer. This bolometer operates on a conductivity transition in which the resistance changes dramatically over the transition temperature range.

The key trade-off with respect to conventional thermal detectors is between sensitivity and response time. The detector sensitivity is often expressed by noise equivalent temperature ($NEDT$) represented by the temperature change, for incident radiation, that gives an output signal equal to the rms noise level. The thermal conductance is an extremely important parameter, since the noise equivalent temperature difference ($NEDT$) is proportional to $(G_{th})^{1/2}$, but the thermal response time of the detector, τ_{th} is inversely proportional to G_{th} . Therefore, a change in thermal conductance due to improvements in material processing techniques improves sensitivity at the expense of time response.

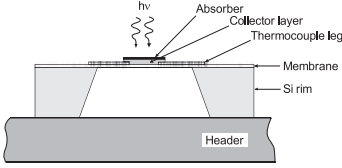
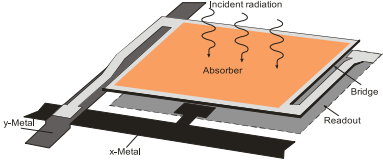
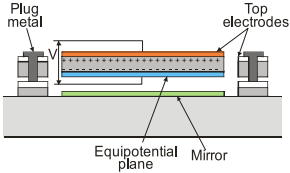
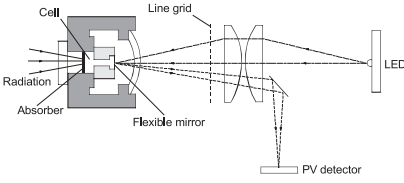
Detectors for Room Temperature THz Imaging

Particular attention in development of THz imaging systems is devoted to the realization of sensors with a large potential for real-time imaging while maintaining a high dynamic range and room-temperature operation. CMOS process technology is especially attractive due to their low price tag for industrial, surveillance, scientific, and medical applications. However, CMOS THz imagers developed thus far have mainly operated single detectors based on lock-in technique to acquire raster-scanned imagers with frame rates on the order of minutes. With this mind, much of recent developments are directed towards three types of focal plane arrays (FPAs):

- Schottky barrier diodes (SBDs) compatible with the CMOS process,
- field effect transistors (FETs) relying on plasmonic rectification phenomena, and
- adaptation of infrared bolometers to the THz frequency range.

An important issue for a FPA is pixel uniformity. It appears however, that the production of monolithically-integrated detector arrays encounters so many technological problems that the device-to-device performance variations and even the percentage of non-functional detectors per chip tend to be unacceptably high.

Table 1 Thermal detectors

Mode of operation	Schematic of detector	Operation and properties
Thermopile		The thermocouple is usually a thin, blackened flake connected thermally to the junction of two dissimilar metals or semiconductors. Heat absorbed by the flake causes a temperature rise of the junction, and hence a thermoelectric electromotive force is developed which can be measured. Although thermopiles are not as sensitive as bolometers and pyroelectric detectors, they will replace these in many applications due to their reliable characteristics and good cost/performance ratio. Thermocouples are widely used in spectroscopy.
Bolometer Metal Semiconductor		The bolometer is a resistive element constructed from a material with a very small thermal capacity and large temperature coefficient so that the absorbed radiation produces a large change in resistance. The change in resistance is like the photoconductor; however, the basic detection mechanisms are different. In the case of a bolometer, radiant power produces heat within the material, which in turn produces the resistance change. There is no direct photon-electron interaction.
Superconductor Hot electron		Initially, most bolometers were the thermistor type made from oxides of manganese, cobalt, or nickel. At present microbolometers are fabricated in large format arrays for thermal imaging applications. Some extremely sensitive low-temperature semiconductor and superconductor bolometers are used in THz region.
Pyroelectric detector		The pyroelectric detector can be considered as a small capacitor with two conducting electrodes mounted perpendicularly to the direction of spontaneous polarization. During incident of radiation, the change in polarization appears as a charge on the capacitor and a current is generated, the magnitude of which depends on the temperature rise and the pyroelectrical coefficient of the material. The signal, however, must be chopped or modulated. The detector sensitivity is limited either by amplifier noise or by loss-tangent noise. Response speed can be engineered making pyroelectric detectors useful for fast laser pulse detection, however with proportional decrease in sensitivity.
Golay cell		The Golay cell consists of an hermetically sealed container filled with gas (usually xenon for its low thermal conductivity) and arranged so that expansion of the gas under heating by a photon signal distorts a flexible membrane on which a mirror is mounted. The movement of the mirror is used to deflect a beam of light shining on a photocell and so producing a change in the photocell current as the output. In modern Golay cells the photocell is replaced by a solid state photodiode and light emitting diode is used for illumination. The performance of the Golay cell is only limited by the temperature noise associated with the thermal exchange between the absorbing film and the detector gas; consequently, the detector can be extremely sensitive with $D^* \approx 3 \times 10^9 \text{ cm Hz}^{1/2} \text{ W}^{-1}$, and responsivities of 10^5 to 10^6 V/W. The response time is quite long, typically 15 ms.

The performance of monolithically integrated detector arrays with room temperature THz detectors is summarized in [Table 2](#). SPDs respond to the THz electric field and usually generate an output current or voltage through a quadratic term in their current-voltage characteristics. In general, the noise equivalent power (NEP) of SBD and FET detectors is better than that of Golay cells and pyroelectric detectors around 300 GHz. Both the pyroelectric and the bolometer FPAs with detector response times in the millisecond time range are not suited for heterodyne operation. FET detectors are clearly capable in heterodyne detection with improving sensitivity. Diffraction aspects predicts FPAs for higher frequencies (0.5 THz and above) and in conjunction with large $f/\#$ optics.

Table 2 Parameter of some uncooled THz detectors

Device type	Electrical responsivity (V/W)	Conditions	NEP (W/Hz ^{1/2})
Schottky diodes			
ErAs/InGaAlAs spiral planar antenna	–	Zero bias, 639 GHz	4.0×10^{-12} NEDT = 120 mK
InGaAs log-spiral antenna	~200 for system estimate 10^3 intrinsic for the diode	0.8 THz	5.0×10^{-12}
VDI Model: WR2.8 ZBD	1500	260–400 GHz	2.7×10^{-12}
VDI Model: WR1.5ZBD	750	500–750 GHz	5.1×10^{-12}
VDI Model: WR1.0 ZBD	200	750–1100 GHz	20×10^{-12}
VDI Model: WR0.65 ZBD	100	1100–1700 GHz	40×10^{-12}
Bolometers			
Hg _{0.8} Cd _{0.2} Te HEB	0.30 at 17 mV bias, 36 GHz	Room temperature	2.2×10^{-9} for 17 mV bias, 35 GHz
	96 for 0.89 THz, 13 mV bias		7.4×10^{-9} for 0.89 THz, 12 mV bias
SixGey:H	170	0.934 THz, uncooled	0.2×10^{-9}
Vanadium oxide	–	Uncooled	320×10^{-12} at 4.3 THz, 9×10^{-13} @ 7.5–14 μ m
Niobum film	21	3.6 mA bias, 1 kHz mod, 300K	1.10×10^{-10}
Ti, antenna-coupled microbolometer	–	10 kHz chop, 1.04 mA bias, 300K	1.5×10^{-11}
Nb ₅ N ₆	400	0.4 mA bias, >10 kHz	9.8×10^{-12}
Vanadium oxide array	1.5×10^4	1 V bias, 130 μ m, uncooled	2.00×10^{-10}
Nb, polyimide, antenna coupled	450	<1 THz	1.5×10^{-11}
Al/Nb; antenna coupled	85	1 kHz mod, 1.6 mA bias	2.5×10^{-11}
Free-standing Nb bridge antenna coupled	210 (average over 10 devices)	650 GHz	12.5×10^{-12}
Pyroelectrics			
Philips P5219 deuterated L-alanine TGS	321	10 Hz mod; amplifier with gain of 4.8, 91 GHz	3.1×10^{-8}
QMC instr	18,300	10 Hz mod;	4.4×10^{-10}
	1200	1.89 THz, <20 Hz mod	
LiTaO ₃	–	530 GHz, Melectron Model SPH-45	2.0×10^{-9}
Golay cells			
Tydex Golay Cell GC-1X	100 000	21 Hz chper	1.4×10^{-10}
Microtech Instruments	10000	12.5 Hz chopper	10×10^{-8}
Micro-array, layer by layer, polimer membranes over Si	–	30 Hz mod 105 GHz	300×10^{-9}
Tydex Golay cell, 6-mm-diameter diamond window		10 Hz mod	7.0×10^{-10}
CMOS-based and plasma detectors			
BiCMOS SiGe, 0.25 μ m HBT	Current R_i 1 A/W at 0.7 THz	3 $\times$ 5 array, chopper 125 kHz	50×10^{-12} at 0.7 THz
BiCMOS SiGe, 0.25 μ m NMOS	Voltage R_v 80 kV/W at 0.6 THz	3 $\times$ 5 array, chopper 16 kHz	300×10^{-12} at 0.6 THz
CMOS SiGe, 65 nm NMOS	Voltage R_v 140 kV/W at 0.87 THz	32 $\times$ 32 array, chopper 5 kHz	100×10^{-12} at 0.87 THz
CMOS SiGe, 65 nm NMOS	Voltage R_v 0.8 kV/W at 1 THz	3 $\times$ 5 array, chopper 1 kHz	66×10^{-12} at 1 THz
CMOS-SBD, 130 nm	Voltage R_v 0.323 kV/W at 0.28 THz	4 $\times$ 4 array, chopper 1 kHz	29×10^{-12} at 0.28 THz
CMOS-SBD, 65 nm	–	1 element; 1 MHz mod	42×10^{-12} at 0.86 THz
CMOS, 150 nm, NMOS	Voltage R_v at 4.1 THz	1 element	133×10^{-12} at 4.1 THz
InGaAs HEMT	Voltage R_v 23 kV/W at 200 GHz	1 element	0.5×10^{-12} at 200 GHz
Asymmetric dual-grating gate InGaAs HEMT	Voltage R_v 6.4 kV/W at 1.5 THz	1 element	50×10^{-12} at 1.5 THz

Below, a short description of different kinds of uncooled THz detectors is presented (Brown and Segovia-Vargas, 2015).

Schottky Barrier Diodes

In spite of achievements of other kind of detectors for THz waveband, the Schottky barrier diodes (SBDs) are among the basic elements in THz technologies. They are used either in direct detection and as nonlinear elements in heterodyne receiver mixers

operating in temperature range of 4–300 K. The cryogenically cooled SBDs were used in mixers preferably in 1980s and early 1990s and then they have been replaced widely by superconductor-insulator-superconductor (SIS) or hot electron bolometer (HEB) mixers, in which mixing processes are similar to that observed in SBDs, but, e.g., in SIS structures the rectification process is based on quantum-mechanical photon-assisted tunnelling of quasiparticles (electrons). The nonlinearity of SBD I - V characteristic (the current increases exponentially with the applied voltage) is the prerequisite for mixing to occur.

Historically first Schottky-barrier structures were pointed contacts of tapered metal wires (e.g., a tungsten needle) with a semiconductor surface (the so-called crystal detectors). Due to limitation of whisker technology, such as constraints on design and repeatability, starting in the 1980s, the efforts were made to produce planar Schottky diodes with air-bridge fingers (see Fig. 2(a)). This design has been the most important steps toward a practical Schottky diode mixer for THz frequency applications, with several thousand diodes on a single chip and where parasitic losses such as the series resistance and the shunt capacitance are minimized. To achieve good performance at high frequencies the diode area should be small. Reducing junction area one reduces junction capacitances to increase operating frequency. But at the same time one increases the series resistance.

Using advanced technology, the diodes are integrated with many passive circuit elements (impedance matching, filters, and waveguide probes) onto the same substrate. By improving the mechanical arrangement and reducing loss, the planar technology is pushed well beyond 300 GHz up to several THz. For example, Fig. 2(b) shows photographs of a bridged four-Schottky diodes' chip arrayed in a balanced configuration to increase power handling.

Recently, an alternative method of Schottky barrier formation has been elaborated by molecular beam epitaxy (MBE) in-situ deposition of a semimetal (ErAs) on semiconductor (InGaAs/InAlAs on InP substrates) to reduce the imperfections that give rise to excess low-frequency noise, particularly $1/f$ noise (Brown *et al.*, 2006). Excellent NEP performance for this III-V semiconductor SBD has been reported ($1.4 \text{ pW/Hz}^{1/2}$ at 100 GHz). By using interband tunnelling, a heterojunction backward diode demonstrated 49.7 kV/W responsivity and 0.18 pW/Hz NEP at 94 GHz (Zhang *et al.*, 2011). More recently Han *et al.* (2012) have demonstrated fully functional CMOS imager operating near or in the sub-millimeter-wave frequency range (see Fig. 3). The 4×4 array increases the imaging speed by 4–8 times, due to fewer mechanical scan steps.

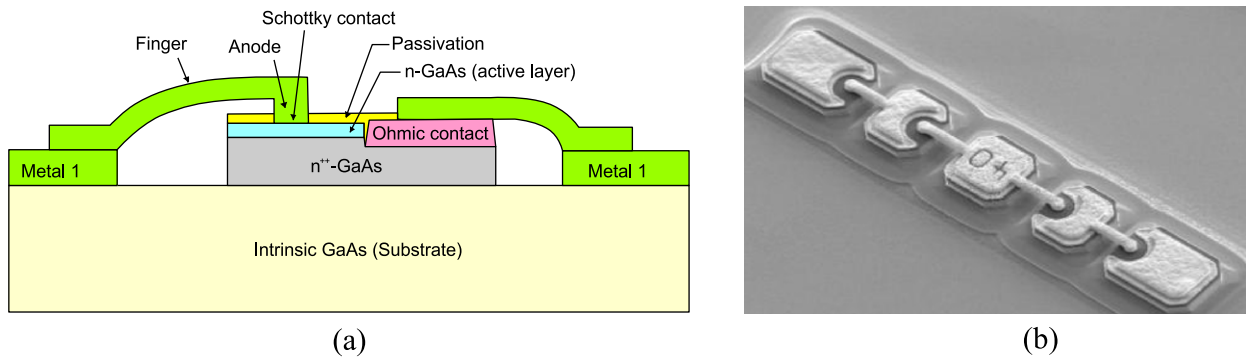


Fig. 2 GaAs Schottky barrier diode: (a) schematic of a planar diode and (b) a four-diode chip array.

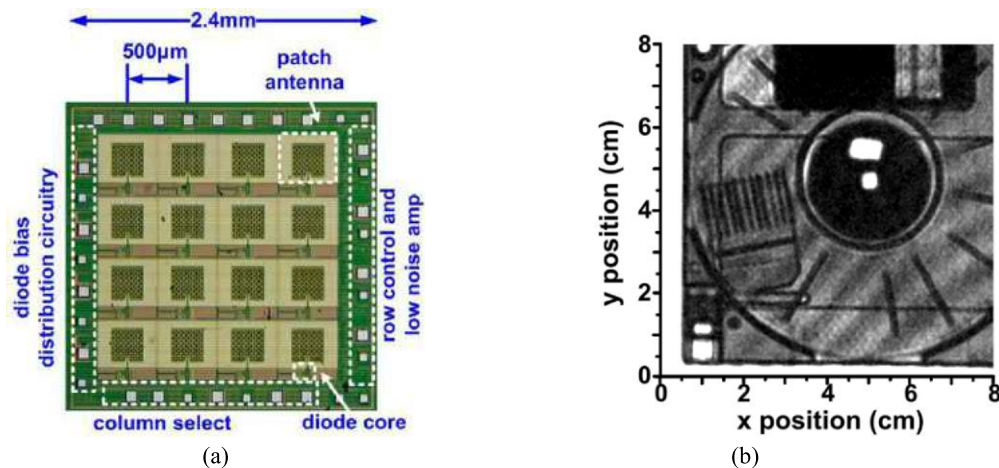


Fig. 3 CMOS SBD 280-GHz imager: die photos of the array (a) and an image of music greeting card obtained. After Han, R., Zhang, Y., Kim, Y., *et al.*, 2012. 280 GHz and 860 GHz image sensors using Schottky-barrier diodes in $0.13 \mu\text{m}$ digital CMOS. IEEE International Solid-State Circuits Conference, 254–256.

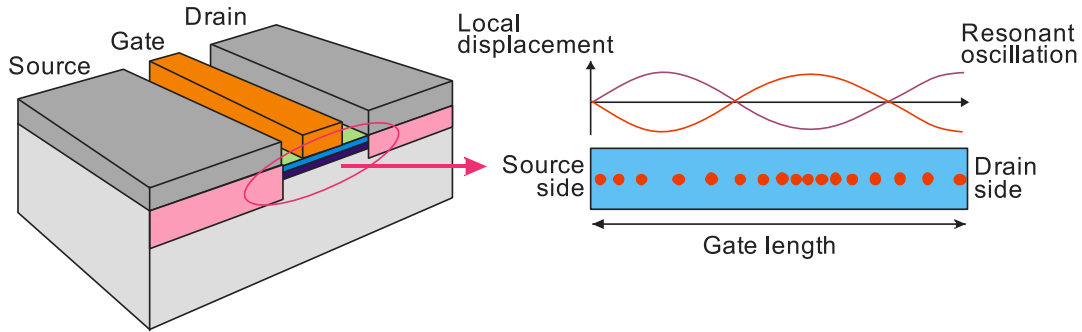


Fig. 4 Plasma oscillations in a field effect transistor.

Field Effect Transistor and CMOS-Based Detectors

The use field effect transistors (FETs) as detectors of THz radiation was first proposed by [Dyakonov and Shur \(1993\)](#) on the basis of formal analogy between the equations of the electron transport in a gated two-dimensional transistor channel and those of shallow water, or acoustic waves in music instruments. [Fig. 4](#) schematically shows the resonant oscillation of plasma waves in gated region of FET.

The detection by FETs is due to nonlinear properties of the transistor, which lead to the rectification of the ac current induced by the coming radiation. As a result, a photoresponse appears in the form of a dc voltage between source and drain. This voltage is proportional to the radiation intensity (photovoltaic effect). A big improvement in sensitivity can be obtained by adding a proper antenna or a cavity coupling. The FETs can be used both for resonant (resonant case of high electron mobility, when plasma oscillation modes are excited in the channel) and non-resonant (broadband) THz detection in the case of low mobility, where plasma oscillations are overdamped ([Knap and Dyakonov, 2013](#)).

The large-scale interest in using FETs as THz detectors started around 2004 after the first experimental demonstration of sub-THz and THz detection in silicon-CMOS FETs. Two years later it was shown that Si-CMOS FETs can reach a *NEP* value competitive with the best conventional room temperature THz detectors. At present the advantages of Si-CMOS FET technology (room temperature operation, very fast response time, easy on-chip integration with read-out electronics and high reproducibility) lead to the straight-forward array fabrication ([Knap et al., 2014](#)).

Recently, the first CMOS FPA used to capture transmission-mode THz video streams in real-time without the need for raster scanning and source modulation has been fabricated. A camera with 32×32 pixel array fully integrated in a 65-nm CMOS process technology has been demonstrated ([Al Hadi et al., 2012](#)) (see right side of [Fig. 5](#)). Each 80- μm array pixel consists of a differential on-chip ring antenna coupled to NMOS direct detector operated well-beyond its cut-off frequency. The camera chip has been packed together with a 41.7-dBi silicon lens in a $5 \times 5 \times 3 \text{ cm}^3$ camera module. In continuous-wave illumination the camera achieves an responsivity of 100–200 kV/W and a total *NEP* of 10–20 $\text{nW/Hz}^{1/2}$ up to 500 fps at 856 GHz.

Microbolometers

An impressive promising technology is also coming from commercially available microbolometer arrays. Adaptation of infrared microbolometers to the THz frequency range after the successful demonstration of active THz imaging in 2006 ([Lee et al., 2006](#)) entailed that in the period 2010–2011 three different companies/organizations announced cameras optimized for the $>1\text{-THz}$ frequency range: NEC (Japan) ([Oda, 2010](#)), INO (Canada) ([Bolduc et al., 2011](#)) and Leti (France) ([Nguyen et al., 2012](#)). The number of vendors is expected to increase soon.

Different designs of THz bolometer pixels have been proposed. NEC's pixel is divided into two parts (see left side of [Fig. 5](#)): a silicon readout integrated circuit (ROIC) in the lower part, and a suspended microbridge structure in the upper part. The microbridge has a two-storied structure. The bottom is composed of a diaphragm and two legs, while the top (eaves) structure is formed on the diaphragm to increase the sensitive area and fill factor. The diaphragm and the eaves absorb THz radiation. The diaphragm is composed of VO_x bolometer thin film, SiN_x passivation layers and TiAlV electrodes, while the eaves structure is composed of SiN_x layer and TiAlV thin film THz absorption layer.

A schematic of one Leti pixel of an amorphous silicon microbolometer array is shown in top centre of [Fig. 5](#). The 50- μm pitch is associated with quasi-double-bowtie antennas to a thermometer microbridge structure derived from the standard IR bolometer. The membrane is suspended over the substrate by arms and pillars. In order to enhance the antenna gain, an equivalent quarter-wavelength resonant cavity is realized under antennas with an 11- μm thick SiO_2 layer deposited over the metallic reflector. To ensure electric contact between the bolometer pillars and CMOS metal upper contacts, the vias are etched through an 11- μm cavity and then metalized.

[Fig. 6](#) summarizes the *NEP* values for bolometer FPAs fabricated by three vendors. The FPAs optimized for 2–5 THz exhibit impressive *NEP* values below 100 $\text{pW/Hz}^{1/2}$. It can be seen that wavelength dependence of *NEP* is quite flat below 200 μm . Further improvement of performance is possible by increasing number of pixels, modification of antenna design while preserving pixel pitch, ROIC and technological stack.

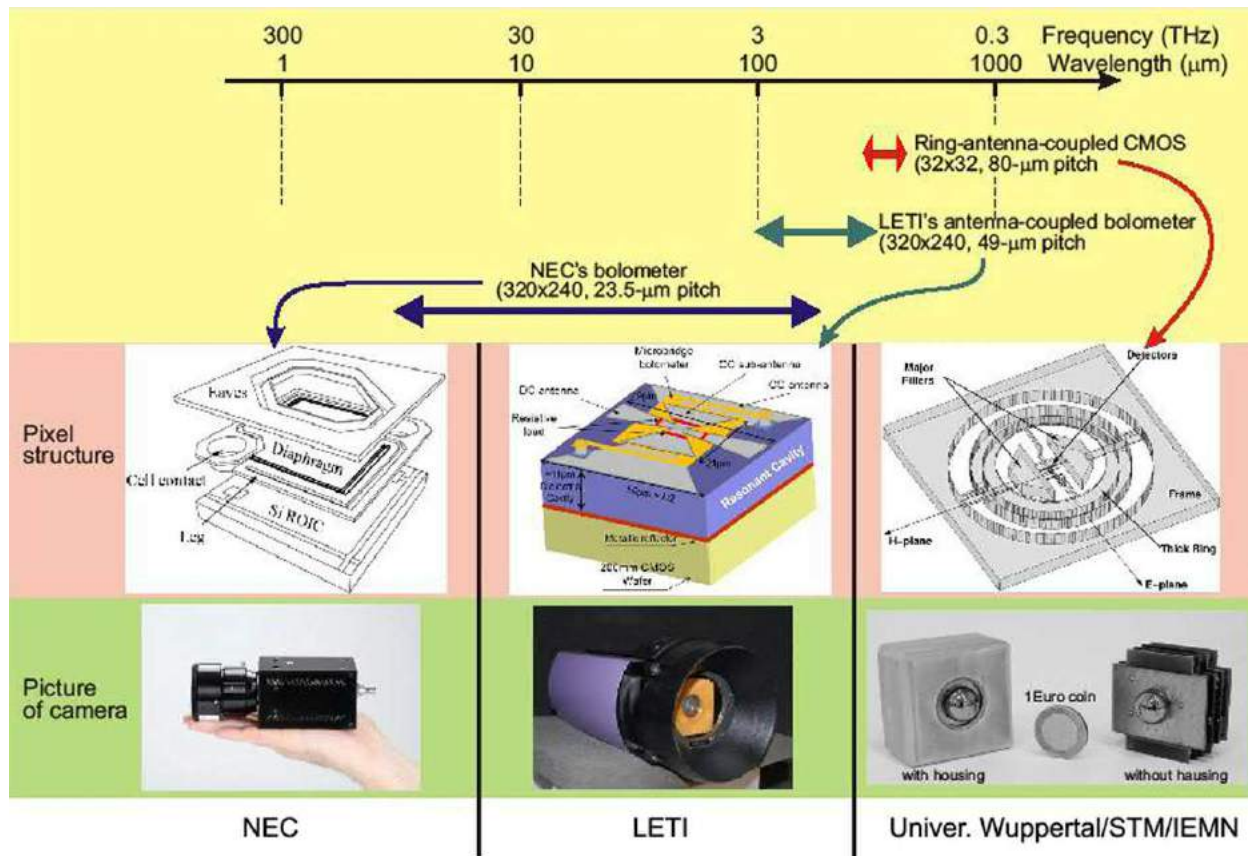


Fig. 5 Development status of uncooled THz focal plane arrays.

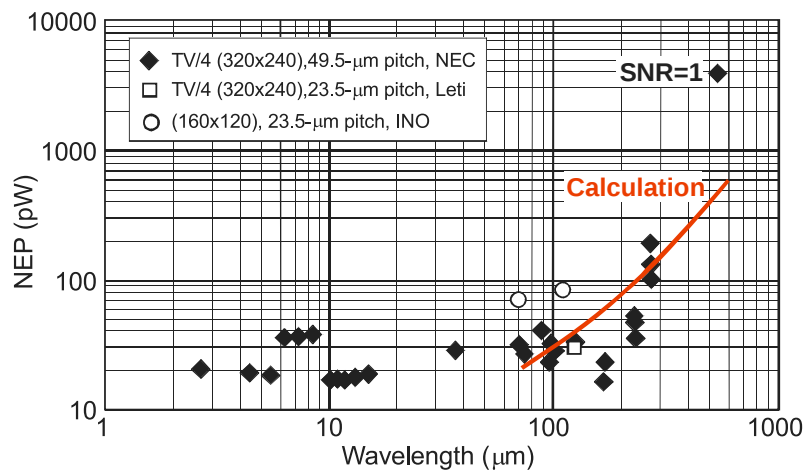


Fig. 6 Spectral dependence of *NEP* for bolometer THz FPAs.

Extrinsic Semiconductor Detectors

Research and development of extrinsic IR photodetectors have been ongoing for more than 50 years. In the 1950s and 1960s, germanium could be made purer than silicon. Today, the problems with producing pure Si have been largely solved. Si has several advantages over Ge; for example, three orders of magnitude higher impurity solubilities are attainable, hence thinner detectors with better spatial resolution can be fabricated from silicon. Si has a lower dielectric constant than Ge, and the related device technology of Si has now been more thoroughly developed, including contacting methods, surface passivation, and mature MOS and CCD technologies. Moreover, Si detectors are characterized by superior hardness in nuclear radiation environments.

For wavelengths longer than 40 μm there are no appropriate shallow dopants for silicon; therefore, germanium devices are still of interest for very long wavelengths. Germanium photoconductors have been used in a variety of infrared astronomical experiments, both airborne and space-based at wavelength ranging from 3 to more than 200 μm . Very shallow donors, such as Sb, and acceptors, such as B, In, or Ga, provide cut-off wavelengths in the region of 100 μm . Fig. 7 shows the spectral response of the extrinsic germanium and silicon photoconductors.

Ge:Ga photoconductors are the best low background photon detectors for the wavelength range from 40 to 120 μm . Application of uniaxial stress along the [100] axis of Ge:Ga crystals reduces the Ga acceptor binding energy, extending the cutoff wavelength to $\approx 240 \mu\text{m}$. At the same time, the operating temperature must be reduced to less than 2 K.

The standard planar hybrid architecture, commonly used to construct near and mid-infrared focal-plane arrays (Rogalski, 2011), is not suitable for far IR detectors where readout glow, lack of efficient heat dissipation, and thermal mismatch between the detector and the readout could potentially limit their performance. Usually, the far-infrared arrays have a modular design with many modules stacked together to form a 2-dimensional array.

The Infrared Astronomical Satellite (IRAS), the Infrared Space Observatory (ISO), and for the far-infrared channels the Spitzer Space Telescope (Spitzer) have all used bulk germanium photoconductors. In the Spitzer mission a 32×32 -pixel Ge:Ga unstressed array was used for the 70- μm band, while the 160 μm band had a 2×20 array of stressed detectors. The detectors are configured in the so-called Z-plane to indicate that the array has substantial size in the third dimension.

An innovative integral field spectrometer, called the Field Imaging Far-Infrared Line Spectrometer (FIFI-LS) was developed for the Herschel Space Observatory and SOFIA – see Fig. 8. To accomplish this, the instrument has two 16×25 Ge:Ga arrays, unstressed for the 45–110 μm range and stressed for the 110–210 μm range.

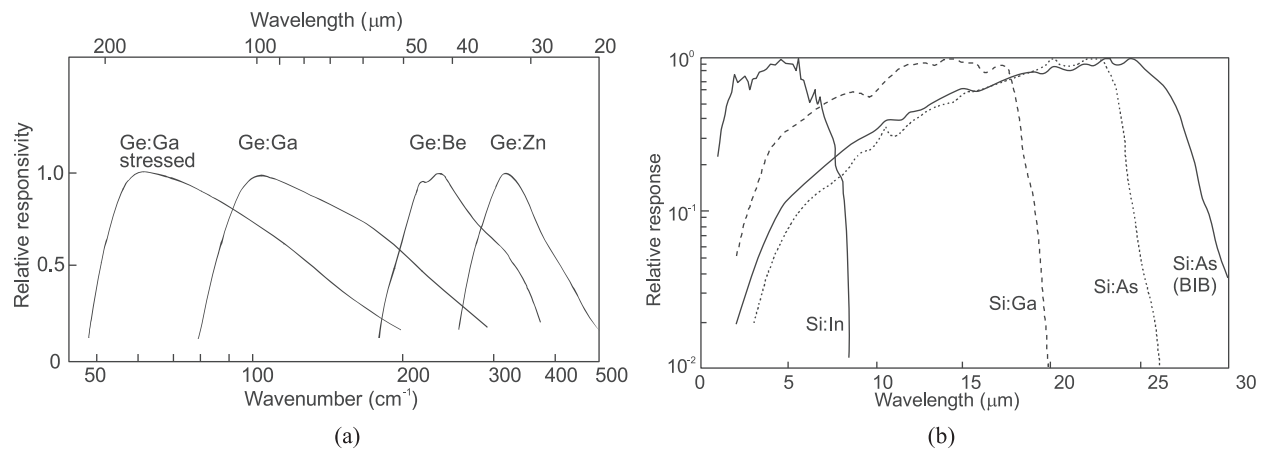


Fig. 7 Relative spectral response of some germanium (a) and silicon (b) extrinsic photoconductors. For comparison also response of Si:As BIB is shown.

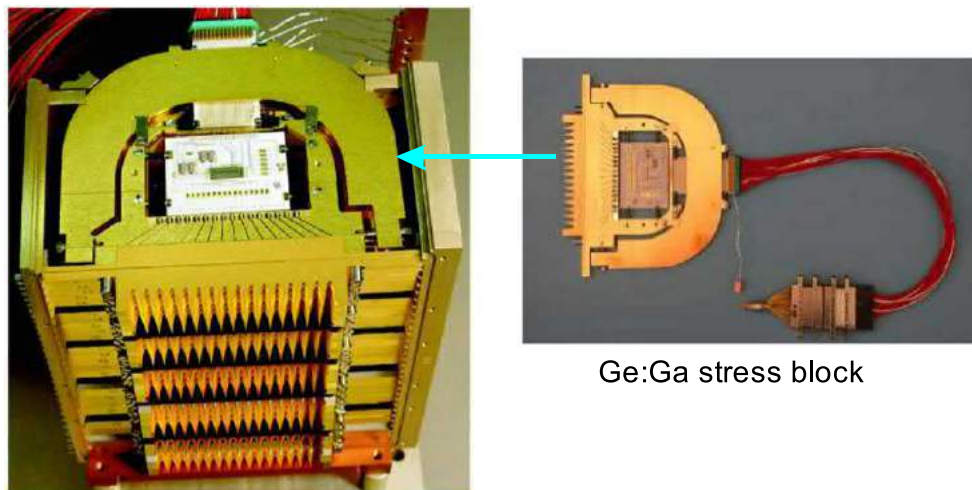


Fig. 8 PACS photoconductor focal plane array. The 25 stressed and low-stress modules of PACS instrument (corresponding to 25 spatial pixels) in the red and blue arrays are integrated into their housing. Available at: <http://fifi-ls.mpg-garching.mpg.de/detector.html>.

The Photodetector Array Camera and Spectrometer (PACS) is one of the three science instruments on ESA's far infrared and sub-millimeter observatory – Herschel Space Laboratory. Apart from two Ge:Ga photoconductor arrays, it employs two filled silicon bolometer arrays with 16×32 and 32×64 pixels, respectively, to perform integral-field spectroscopy and imaging photometry in the 60–210 μm wavelength regime. Median NEP values are $8.9 \times 10^{-18} \text{ W/Hz}^{1/2}$ for the stressed and $2.1 \times 10^{-17} \text{ W/Hz}^{1/2}$ for the unstressed detectors, respectively. The detectors are operated at $\sim 1.65 \text{ K}$. The readout electronics is integrated into the detector modules – each linear module of 16 detectors is read out by a cryogenic amplifier/multiplexer circuit in CMOS technology but operates at temperature 3–5 K.

Blocked Impurity Band Detectors

One of the major problems in the design of extrinsic photoconductors is that the doping concentration is driven by conflicting requirements: the doping concentration needs to be as high as possible to get high photon-absorption coefficients (the doping concentration is limited by hopping conduction induced by direct transfer of charge carriers from one impurity site to the next in heavily-doped semiconductors). In contrast a low doping concentration is also desirable to achieve a low electrical conductivity which in turn reduces Johnson noise.

In 1979 M. Petroff and D. Stapelbroeck, working at the Rockwell International Science Centre, invented what they referred to as Blocked-Impurity Band (BIB) detectors. These detectors were developed to provide significantly reduced nuclear radiation sensitivity and improved performance compared to extrinsically doped silicon photoconductors. BIBs have some excellent properties which make them extremely useful for astronomical applications.

BIB detectors overcome the limitation of the doping density present in a standard extrinsic photoconductor by placing a thin intrinsic (undoped) silicon blocking layer between a heavily doped active layer and a planar contact. The active region of detector structure, usually based on epitaxially grown n-type material, is sandwiched between a higher doped degenerate substrate electrode and an undoped blocking layer. Doping of the active layer is high enough for the onset of an impurity band in order to display a high quantum efficiency for impurity ionization (in the case of Si:As BIB, the active layer is doped to $\approx 5 \times 10^{17} \text{ cm}^{-3}$). The device exhibits a diode-like characteristic, except that photoexcitation of electrons takes place between the donor impurity and the conduction band.

The design of BIB detectors offers a number of advantages over conventional extrinsic photoconductors: the high absorption coefficient of the absorbing layer means that detectors with comparatively small active volumes can be made, providing low susceptibility to cosmic rays without compromising quantum efficiency. Also, due to the heavy doping of the active layer, the impurity band increases in width, therefore effectively decreasing the energy gap between the impurity band and the conduction band.

The main application of BIB arrays today is for ground- and space-based far-infrared astronomy – Si-As BIB performance are gathered in Table 3. The arrays should be operated under the most uniform possible conditions, in the most benign and constant environment possible. Array performance is strongly affected by background levels. Extrinsic silicon arrays for high background applications are less developed than that for low background applications.

The largest extrinsic infrared detector arrays are manufactured for astronomy owing to investments from NASA and the National Science Foundation. At present Raytheon Vision Systems (RVS), DRS Technologies, and Teledyne Imaging Sensors (formerly Rockwell Scientific Company) supply the majority of IR arrays used in astronomy, and between them the most important are BIB detector arrays. Impressive progress has been achieved especially in Si:As BIB array technology with formats as large as 2048×2048 and pixels as small as $18 \mu\text{m}$.

Table 3 Performance of Si:As BIB FPAs fabricated in several formats for both ground and space based applications

Parameter	Si:As BIB	Phenix	MIRI	Aquarius-1k
Application/Users	Ground-based telescopes, ESO, Univer. of Tokyo	Space telescopes, JAXA	Space telescopes, JAXA, NASA	Ground-based telescopes, ESO, Univer. of Arizona
Format	320×240	1024×1024 , 2048×2048	1024×1024	1024×1024
Pixel size	$50 \mu\text{m}$	$25 \mu\text{m}$	$25 \mu\text{m}$	$30 \mu\text{m}$
ROIC type	DI	SFD	SFD	SFD
Fill factor	$\geq 95\%$	$\geq 95\%$	$\geq 98\%$	$\geq 98\%$
ROIC input referred noise	$< 1000e^-_{\text{RMS}}$	$6\text{--}20e^-_{\text{RMS}}$	$< 10\text{--}30e^-$	Low gain $< 1000e^-_{\text{RMS}}$ High gain $< 100e^-_{\text{RMS}}$
Integration capacity	$7 \text{ or } 20 \times 10^6 e^-$	$3 \times 10^6 e^-$	$2 \times 10^5 e^-$	$1 \text{ or } 11 \times 10^6 e^-$
Max. frame rates	100–500 Hz	0.1 Hz	0.1 Hz	120 Hz
Number of outputs	16 or 32	4	4	16 or 64
Packaging	LCC	LCC	Module	Module – 2 side butttable

DI, direct injection; e^- , electron; LCC, leadless chip carrier; RMS, root mean square; SFD, source follower per detector.

Source: Mills, R., Beuville, E., Corrales, E., *et al.*, 2011. Evolution of large format impurity band conductor focal plane arrays for astronomy applications. Proceedings of SPIE 8154, 81540R.

Semiconductor Bolometers

Cooled silicon bolometers demonstrate broadband and nearly flat spectral response in the 1–3000 μm wavelength range. They are easier to fabricate with high operability, good uniformity, and lower cost, but has low operating temperature (4.2–0.3 K). During fabrication of bolometers, their area, operating temperature, thermal time constant, and thermal conductance are adjusted to meet the specific design requirements. The present day technology exists to produce arrays of hundreds of pixels that are operated in many experiments including NASA Pathfinder ground based instruments, and balloon experiments.

The thermistors are typically fabricated by lithography on membranes of Si or SiN. The impedance is selected to a few $\text{M}\Omega$ to minimize the noise in JFET amplifiers operated at about 100 K. Limitation of this technology is assertion of thermal mechanical, and electrical interface between the bolometers at 100–300 mK and the amplifiers at ≈ 100 K. Usually, JFET amplifiers are sited on membranes which isolate them so effectively that the environment remains at much lower temperatures (about 10 K) – see Fig. 9. In addition, the equipment at 10 K is itself thermally isolated from nearby components at 0.1–0.3 K.

In bolometer metal films that can be continuous or patterned in a mesh absorb the photons. The patterning is designed to select the spectral band, to provide polarization sensitivity, or to control the throughput. Different bolometer architectures are used. In close-packed arrays and spider web, the pop-up structures or two-layer bump bonded structures are fabricated.

At present high-performance bolometer arrays for the far IR and sub-mm spectral ranges are available. For example, the Herschel/PACS instrument uses a 2048-pixel array of bolometers and is an alternative to JFET amplifiers (Mills *et al.*, 2011). The architecture of this array is vaguely similar to the direct hybrid mid-infrared arrays, where one silicon wafer is patterned with bolometers, each in the form of a silicon mesh.

Pair Braking Photon Detectors

One of the methods of photon detection consists in using superconducting materials. If the temperature is far below the transition temperature, T_c , most of electrons in them are banded into Cooper pairs. Photons with energies exceeding the binding Cooper pair energies in the superconductor, 2Δ (each electron must be supplied an energy Δ), can break these pairs producing quasiparticles (electrons) (see inset of Fig. 10(a)). When the bias voltage is increased to the gap voltage, the Cooper pairs on one side of the junction can break up into two quasiparticles, which then tunnel to the other side of the junction before recombining, resulting in a sharp increase in current. This process resembles the interband absorption in semiconductors, with the energy gap equal 2Δ , when the photons are absorbed and electron-hole pairs are created.

Several structures of pair braking detectors which use different ways to separate quasiparticles from Cooper pairs have been proposed. Among them are: superconductor-insulator-superconductor (SIS) and superconductor-insulator-normal metal (SIN) detectors and mixers, radio frequency (RF) kinetic inductance detectors, and superconducting quantum interference device

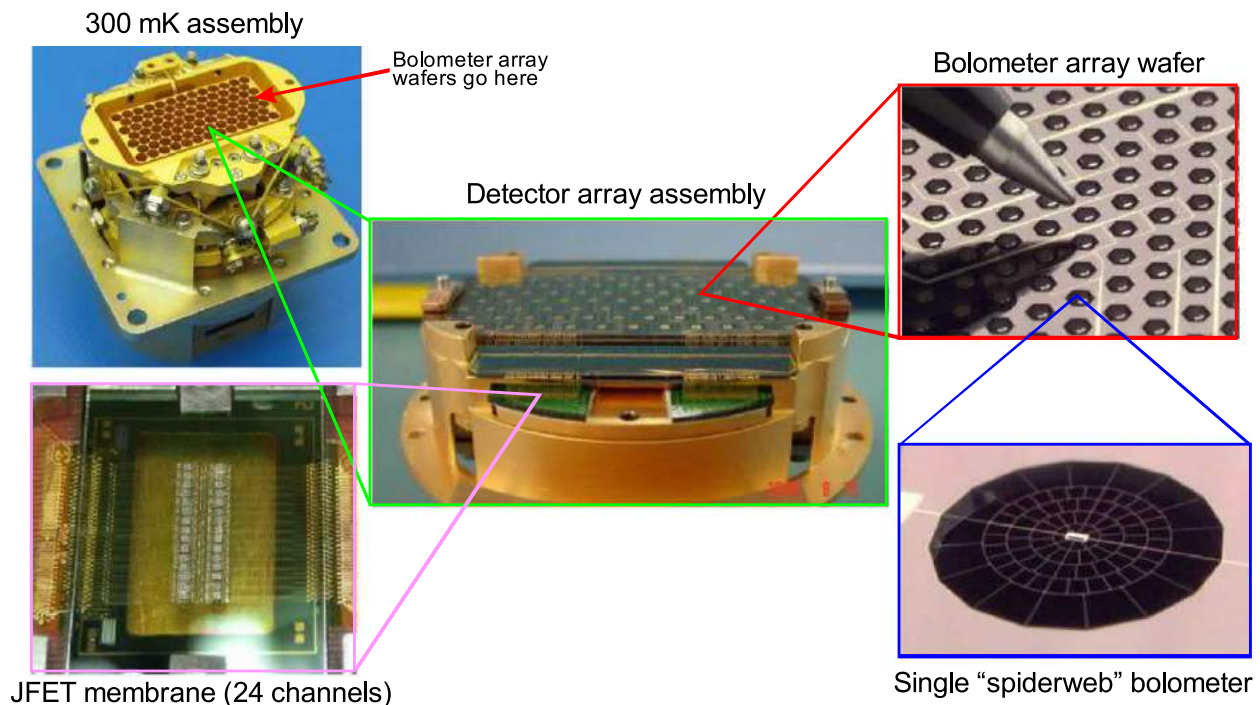


Fig. 9 Bolometer array of the Spectral and Photometric Imaging Receiver (SPIRE). Available at: <http://herchel.jpl.nasa.gov/spireInstrument.shtml>.

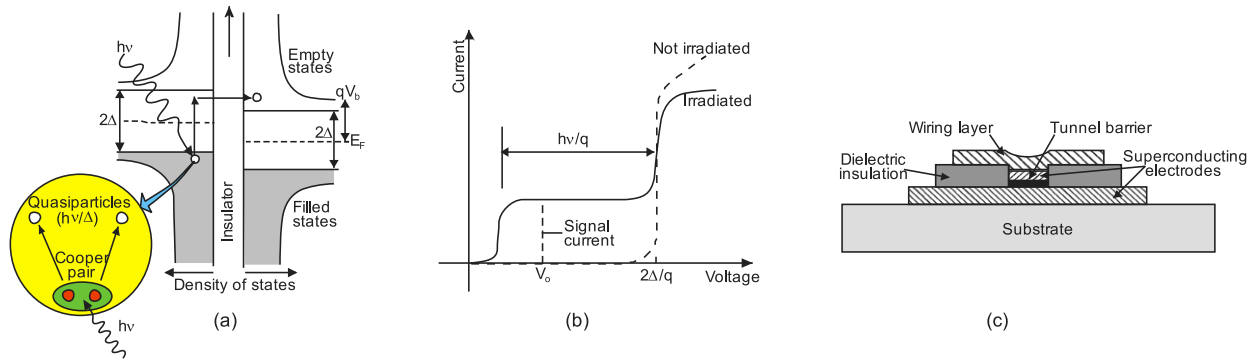


Fig. 10 SIS junction: (a) energy diagram with applied bias voltage and illustration of photon assisted tunnelling, (b) current-voltage characteristic of a non-irradiated and irradiated barrier (the intensity of the incident radiation is measured as an excess of the current at a certain bias voltage V_0 – schematic creation of quasiparticle is shown inside (a)), and (c) a cross section of a typical SIS junction.

(SQUID) kinetic inductance detectors. Superconducting detectors offer many benefits: outstanding sensitivity, lithographic fabrication, and large array sizes, especially through the development of multiplexing techniques. The basics physics of these devices and progress in their developments are described in [Zmuidzinas and Richards \(2004\)](#).

The SIS detector is a sandwich of two superconductors separated by a thin (≈ 20 Å) insulator, what is schematically shown in [Fig. 10\(c\)](#). Nb and NbTiN are almost exclusively used as superconductors for the electrodes. For a standard junction process, the base electrode is 200-nm sputtered Nb, the tunnel barrier is made using a thin 5-nm sputtered Al layer which is either thermally oxidized (Al_2O_3) or plasma nitridized (AlN). The counterelectrode is 100-nm sputtered Nb or reactively sputtered NbTiN. Typical junction areas are about $1 \mu\text{m}^2$.

SIS tunnel junctions are mainly used as mixers in heterodyne type mm and sub-mm receivers, because of their strong non-linear I-V characteristic. They can be also used as direct detection detectors. Operating temperature of SIS junctions is below 1 K; typically $T \leq 300$ mK. However, up to now SIS detectors are difficult to integrate into large arrays.

Progress in development of large-format, high-sensitivity focal plane arrays is especially promising with two detector technologies: transition-edge superconducting (TES) bolometers (see next section) and microwave kinetic inductance detectors (MKIDs) based on different principles of superconductivity. Multiple instruments are currently in development based on arrays up to 10,000 detectors using both time-domain multiplexing (TDM) and frequency-domain multiplexing (FDM) with superconducting quantum interference devices (SQUIDs) ([Bock, 2009](#)). Both sensors show potential to realize the very low $\sim 10^{-20}$ W/Hz^{1/2} sensitivity needed for space-borne spectroscopy.

A MKID is essentially a high-Q resonant circuit made out of either superconducting microwave transmission lines or a lumped element LC resonator (fabricated from thin aluminium and niobium films). In the first case a meandered quarter-wavelength strip of superconducting material is coupled by means of a coupling capacitance to a coplanar waveguide through-line used for excitation and readout. Lumped element are instead created from an LC series resonant circuit inductively coupled to a microstrip feed line placed in a high frequency resonant circuit (see [Fig. 11](#)). Photons hitting an MKID break Cooper pairs, which changes the surface impedance of the transmission line or inductive element producing a number of quasiparticles. This causes the resonant frequency and quality factor to shift an amount proportional to the energy deposited by the photon. The amplitude (c) and phase (d) of a microwave excitation signal sent through the resonator. The change in the surface impedance of the film following a photon absorption event pushes the resonance to lower frequency and changes its amplitude. The energy of the absorber photon can be determined from the degree of phase and amplitude shift. The readout is almost entirely at room temperature and can be highly multiplexed; in principle hundreds or even thousands of resonators could be read out on a single feedline ([Day et al., 2003](#); [Mazin, 2009](#)).

Superconducting HEB and TES Detectors

Among superconducting detectors used for terahertz downconversion, hot-electron-bolometer (HEB) mixers have attracted the attention. Their low local-oscillator (LO) power consumption (less than $1 \mu\text{W}$), near-quantum-limited noise performance, and ease of fabrication have placed them above competing technologies in the quest for implementation of large-format heterodyne arrays.

In principle, HEB is quite similar to the transition-edge sensor (TES) bolometer, where small temperature changes caused by the absorption of incident radiation strongly influence resistance of biased sensor near its superconducting transition. The main difference between HEBs and ordinary bolometers is the speed of their response. High speed is achieved by allowing the radiation power to be directly absorbed by the electrons in the superconductor, rather than using a separate radiation absorber and allowing the energy to flow to the superconducting TES via phonons, as ordinary bolometers do. After photon absorption, a single electron initially receives the energy $h\nu$, which is rapidly shared with other electrons, producing a slight increase in the electron temperature. In the next step, the electron temperature subsequently relaxes to the bath temperature through emission of phonons.

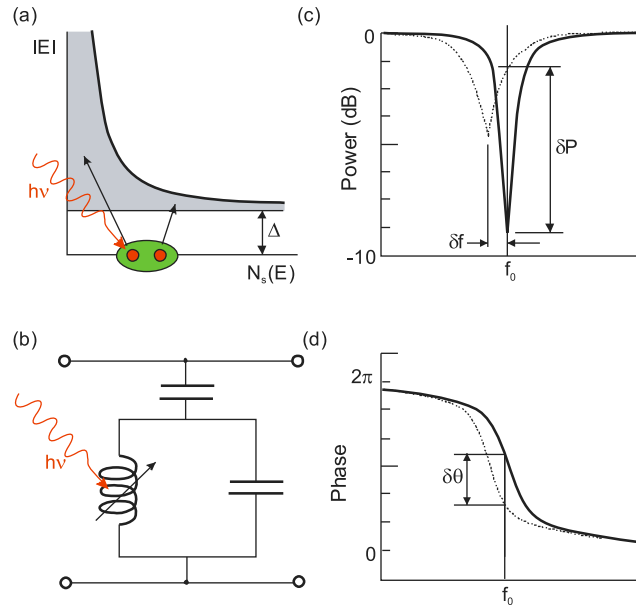


Fig. 11 An illustration of the operational principle behind a MKID.

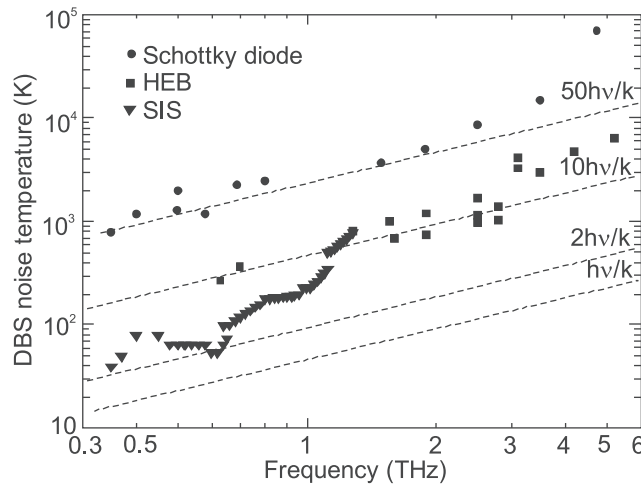


Fig. 12 DSB noise temperature of Schottky diode mixers, SIS mixers, and HEB mixers operated in terahertz spectral band. Reproduced from Hübers, H.-W., 2008. Terahertz heterodyne receivers. *IEEE Journal of Selected Topics in Quantum Electronics* 14, 378–391.

In comparison with TES, the thermal relaxation time of the HEB's electrons can be made fast by choosing a material with a large electron-phonon interaction. The development of superconducting HEB mixers has led to the most sensitive systems at frequencies in the terahertz region, where the overall time constant has to be a few tens of picoseconds. These requirements can be realized with a superconducting microbridge made from NbN, NbTiN, or Nb on a dielectric substrate (Gol'tsman *et al.*, 2005).

Generally, in terahertz receivers, the noise of a mixer is quoted in terms of a single-sideband (SSB), T^{SSB} , or double-sideband (DSB), T^{DSB} , mixer noise temperature. The DSB noise temperatures achieved with Schottky diode mixers, SIS mixers, and HEB mixers operated in terahertz spectral band are presented in Fig. 12 (Hübers, 2008). The noise temperature of SBD receivers has essentially reached a limit of about $50 \, hv/k$ in frequency range below 3 THz. Above 3 THz, there occurs a steep increase, mainly due to increasing losses of the antenna and reduced performance of the diode itself.

Fig. 13(a) presents an example cross-section view of NbN mixer chip. About 150-nm thick Au spiral structure is connected to the contacts pads. The superconducting NbN film extends underneath the contact layer/antenna. The central area of a mixer chip shown in Fig. 13(b) is manufactured from a 3.5-nm thick superconducting NbN film on a high resistive Si substrate (Gol'tsman *et al.*, 2005). The active NbN film area is determined by the dimensions of the 0.2- μm gap between the gold contact pads. The NbN microstrip is integrated with a planar antenna patterned as log-periodic spiral.

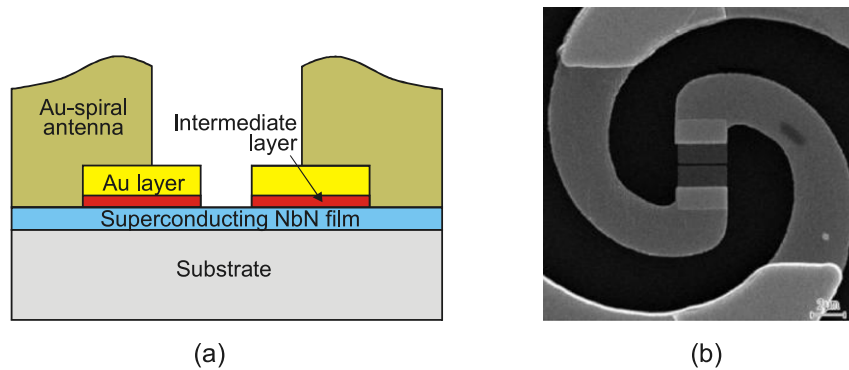


Fig. 13 NbN HEB mixer chip: (a) cross section view and (b) SEM micrograph of the central area of mixer.

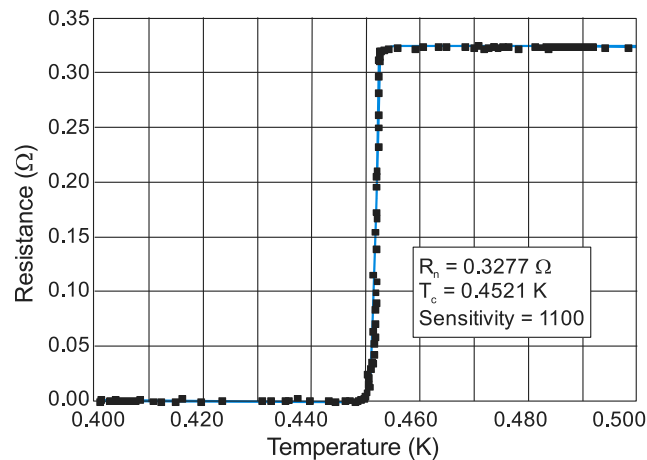


Fig. 14 Resistance vs. temperature for a high-sensitivity TES Mo/Au bilayer with superconducting transition at 444 mK.

The name of the transmission-edge sensor (TES) bolometer is derived from its thermometer, which is based on thin superconducting films held within transition region, where it change from the superconducting to the normal state over a temperature range of a few milliKelvin (see Fig. 14) (Benford and Moseley, 2000). Changes in temperature transition can be set by using a bilayer film consisting of a normal material and a layer of superconductor (e.g., thin Mo/Au, Mo/Cu, Ti/Au, etc.). Such design enables diffusion of the Cooper pairs from the superconductor into the normal metal and makes it weakly superconducting – this process is called the proximity effect. As a result, the transition temperature is lowered relative to that for the pure superconducting film ($T < 200$ mK). Thus in principle, the TES bolometers are quite similar to the HEBs.

TES bolometers are superior to current-biased particle detectors in terms of linearity, resolution, and maximum count rate. At present, these detectors can be applied for THz photons counting because of their high sensitivity and low thermal time constant. Membrane isolated TES bolometers are capable of reaching a phonon $NEP \approx 4 \times 10^{-20}$ W/Hz^{1/2}. The current generation of sub-orbital experiments largely rely on TES bolometers. Important feature of this sensor is that it can operate in wide spectral band, between the radio and gamma rays.

The most ambitious example of TES bolometer array is that used in the submillimetre camera SCUBA (Submillimetre Common-User Bolometer Array) – 2 with 10,240 pixels (SCUBA-2, 2006). The camera operated at wavelengths of 450 and 850 μ m has been mounted on the James Clerk Maxwell Telescope in Hawaii. Each SCUBA-2 array is made of four side-butttable sub-arrays, each with 1280 (32 $\times$ 40) transition-edge sensors.

Conclusions

The THz detectors will receive increasing importance in a very diverse range of applications including detection of biological and chemical hazardous agents, explosive detection, building and airport security, radio astronomy and space research, biology and medicine. The future sensitivity improvement of THz instruments will come with the use of large format arrays with readouts in the focal plane to provide the vision demands of high resolution spectroscopy. One of promising solution is room temperature THz imaging achieved with low-cost CMOS detectors and microbolometer.

References

- Al Hadi, R., Sherry, H., Grzyb, J., *et al.*, 2012. A 1 k-pixel video camera for 0.7–1.1 terahertz imaging applications in 65-nm CMOS. *IEEE Journal of Solid-State Circuits* 47, 2999–3012.
- Benford, D.J., Moseley, S.H. 2000. Superconducting transition edge sensor bolometer arrays for submillimeter astronomy. In: *Proceedings of the International Symposium on Space and THz Technology*. Available at: www.eecs.umich.edu/~jeast/benford_2000_4_1.pdf.
- Bock, J.J., 2009. Superconducting detector arrays for far-infrared to mm-wave astrophysics. Available at: <http://cmbpol.uchicago.edu/depot/pdf/white-paper-j-bock.pdf>.
- Bolduc, M., Terroux, M., Tremblay, B., *et al.*, 2011. Noise-equivalent power characterization of an uncooled microbolometer-based THz imaging camera. *Proceedings of SPIE* 8023.80230C-1–10.
- Brown, E.R., Segovia-Vargas, D., 2015. Principles of THz direct detection. In: Carpintero, G., Garcia Muñoz, L.E., Hartnagel, L., Preu, S., Räisänen, A.V. (Eds.), *Semiconductor Terahertz Technology: Devices and Systems at Room Temperature Operation*. Wiley, pp. 212–253.
- Brown, E.R., Young, A.C., Zimmerman, J., Kazemi, H., Gossard, A.C., 2006. High-sensitivity, quasi-optically-coupled semimetal-semiconductor detectors at 104 GHz. *Proceedings of SPIE* 6212, 621205.
- Bründermann, E., Hübers, H.-W., Kimmitt, M.F., 2012. *Terahertz Techniques*. Heidelberg: Springer-Verlag.
- Day, P., LeDuc, H.G., Mazin, B.A., Vayonakis, A., Zmuidzinas, J., 2003. A broadband superconducting detector suitable for use in large arrays. *Nature* 425, 817–821.
- Dyakonov, M., Shur, M.S., 1993. Shallow water analogy for a ballistic field effect transistor: new mechanism of plasma wave generation by the dc current. *Physical Review Letters* 71, 2465–2468.
- Gol'tsman, G.N., Vachtomin, Yu.B., Antipov, S.V., *et al.*, 2005. NbN phonon-cooled hot-electron bolometer mixer for terahertz heterodyne receivers. *Proceedings of SPIE* 5727, 95–106.
- Han, R., Zhang, Y., Kim, Y., *et al.*, 2012. 280 GHz and 860 GHz image sensors using Schottky-barrier diodes in 0.13 μm digital CMOS. *IEEE International Solid-State Circuits Conference*, 254–256.
- Hübers, H.-W., 2008. Terahertz heterodyne receivers. *IEEE Journal of Selected Topics in Quantum Electronics* 14, 378–391.
- Knap, W., Coquillat, D., Dyakonova, N., *et al.*, 2014. Terahertz plasma field effect transistors. In: Perenzoni, M., Paul, D.J. (Eds.), *Physics and Applications of Terahertz Radiation*. Dordrecht: Springer, pp. 77–100.
- Knap, W., Dyakonov, M.I., 2013. Field effect transistors for terahertz applications. In: Saeedkia, D. (Ed.), *Handbook of Terahertz Technology*. Cambridge: Woodhead Publishing, pp. 121–155.
- Lee, A.W.M., Williams, B.S., Kumar, S., Hu, Q., Reno, J.L., 2006. Real-time imaging using a 4.3-THz quantum cascade laser and a 320×240 microbolometer focal-plane array. *IEEE Photonics Technology Letters* 18, 1415–1417.
- Mazin B.A., 2009. Microwave kinetic inductance detectors: The first decade. *The Thirteenth International Workshop on Low Temperature Detectors-LTD13*. AIP Conference Proceedings 1185 (1), 135–142.
- Mills, R., Beuville, E., Corrales, E., *et al.*, 2011. Evolution of large format impurity band conductor focal plane arrays for astronomy applications. *Proceedings of SPIE* 8154.81540R-1–81540R-10.
- Nguyen, D.-T., Simoons, F., Ouvrier-Bufferet, J.-L., Meilhan, J., Coutaz, J.-L., 2012. Broadband THz uncooled antenna-coupled microbolometer array – electromagnetic design, simulations and measurements. *IEEE Transactions on Terahertz Science and Technology* 2, 299–305.
- Oda, N., 2010. Uncooled bolometer-type terahertz focal-plane array and camera for real-time imaging. *Comptes Rendus Physique* 11, 496–509.
- Rogalski, A., 2011. *Infrared Detectors*, second ed. Boca Raton: CRC Press.
- Saeedkia, D. (Ed.), 2013. *Handbook of Terahertz Technology for Imaging, Sensing and Communications*. Oxford: Woodhead Publishing Limited.
- SCUBA-2, 2006. *The Royal Observatory*. Available at: <http://www.roe.ac.uk/ukac/projects/scubatwo/>.
- Siegel, P.H., 2002. Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques* 50, 910–928.
- Sizov, F., Rogalski, A., 2010. THz detectors. *Progress in Quantum Electronics* 34, 278–347.
- Zhang, Z., Rajavel, R., Deelman, P., Fay, P., 2011. Sub-micro area heterojunction backward diode millimeter-wave detectors with $0.18 \text{ pW/Hz}^{1/2}$ noise equivalent power. *IEEE Microwave and Wireless Components Letters* 21, 267–269.
- Zmuidzinas, J., Richards, P.L., 2004. Superconducting detectors and mixers for millimeter and submillimeter astrophysics. *Proceedings of IEEE* 92, 1597–1616.

Gravitational Wave Detection

Marie-Anne Bizouard, Paris-Saclay University, Orsay, France

Nelson Christensen, Carleton College, Northfield, MN, United States and University of Côte d'Azur, Nice, France

© 2018 Elsevier Ltd. All rights reserved.

Introduction

In the later part of the 19th century, Albert Michelson performed extraordinary experiments that shook the foundations of the physics world. Michelson's precise determination of the speed of light was an accomplishment that takes great skill to reproduce today. Edward Morley teamed up with Michelson to measure the velocity of Earth with respect to the aether. The interferometer that they constructed was exquisite, and through amazing experimental techniques the existence of the aether was disproved. The results of Michelson and Morley led to a revolution in physics, and provided evidence that helped Albert Einstein to develop the general theory of relativity. Now the Michelson interferometer has provided dramatic confirmation of Einstein's theory of general relativity through the direct detection by the Laser Interferometer Gravitational-Wave Observatory (LIGO) of gravitational waves, and the observation of black holes (Abbott *et al.*, 2016a,b).

An accelerating electric charge produces electromagnetic radiation – light. It should come as no surprise that an accelerating mass produces gravitational light, namely, gravitational radiation (or gravitational waves). In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational waves because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational waves in 1916, and Taylor and Weisberg (1989) indirectly confirmed their existence through observations of the orbital decay of the binary pulsar 1913 + 16 system. The direct detection of gravitational waves has been difficult, and has literally taken decades of tedious experimental work to accomplish. The only possibility for producing detectable gravitational waves comes from extremely massive objects accelerating up to relativistic velocities. The gravitational waves that have been detected so far have come from the coalescence of binary black hole systems. For example, GW150914 was produced by the merger of a $29 M_{\odot}$ black hole and a $36 M_{\odot}$ black hole some 1.3×10^9 light-years away. The total energy radiated in gravitational waves was equivalent to $3 M_{\odot} c^2$, with a peak luminosity of 3.6×10^{56} ergs/s.

Other possibly detectable gravitational wave sources are also astrophysical: supernovae, pulsars, neutron star binary systems, newly formed black holes, or even the Big Bang. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravitational waves from the Big Bang would provide unique information of the universe at its earliest moments. Observations of core-collapse supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies up to hundreds of Hertz, and the signals from these objects will help to decipher their characteristics. Gravitational waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter; they would also help to explain the mechanism that produces short gamma ray bursts. The observation of black hole formation from these binary systems, and the ringdown of the newly formed black hole as it approaches a perfectly spherical shape, would be the *coup de grâce* for the debate on black hole existence, and the ultimate triumph for general relativity.

Advanced LIGO (Aasi *et al.*, 2015) and Advanced Virgo (Acernese *et al.*, 2015) are second generation interferometric gravitational wave detectors. Initial LIGO and Virgo conducted observations from 2002 through 2010. Advanced LIGO and Advanced Virgo will ultimately have better sensitivities, by a factor of 10, over their initial designs. They will search for gravitational waves from 10 Hz up to a few kilohertz. Their target sensitivities will allow them to observe signals from the coalescence of binary neutron star systems ($1.4 M_{\odot} - 1.4 M_{\odot}$) out to distances of 200 Mpc for Advanced LIGO and 150 Mpc for Advanced Virgo. The mergers of more massive binary black holes systems will extend much farther.

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational waves will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational waves will deform the fabric of spacetime. Just like electromagnetic radiation there are two polarizations for gravitational waves. Let us imagine a linearly polarized gravitational wave propagating in the z -direction, $h(z, t) = h_0 e^{i(kz - \omega t)}$. The fabric of space is stretched due to the strain created by the gravitational wave. Consider a length L_0 of space along the x -axis. In the presence of the gravitational wave the length oscillates like

$$L(t) = L_0 + \frac{h_0 L_0}{2} \cos(\omega t)$$

Hence, there is a change in its length of

$$\Delta L_x = \frac{h_0 L_0}{2} \cos(\omega t)$$

A similar length L_0 of the y -axis oscillates, like

$$\Delta L_y = -\frac{h_0 L_0}{2} \cos(\omega t)$$

One axis stretches, while the perpendicular one contracts, and then vice versa, as the wave propagates through. Consider the relative change of the lengths of the two axes (at $t=0$),

$$\Delta L = \Delta L_x - \Delta L_y = h_{0+} L_0$$

or

$$h_{0+} = \frac{\Delta L}{L_0}$$

So the amplitude of a gravitational wave is the amount of strain that it produces on spacetime. The other gravitational wave polarization ($h_{0\times}$) produces a strain on axes 45 degree from (x, y) . Imagine some astrophysical event produces a gravitational wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L_0 . The first gravitational wave observed by LIGO had an amplitude of $h \sim 10^{-21}$ with a frequency at peak gravitational wave strain of 150 Hz (Abbott *et al.*, 2016a). The magnitude of a gravitational wave falls off as $1/r$, so it will be impossible to observe events that are too far away. However, when the detectors' sensitivity is improved by a factor of n , the rate of signals should grow as n^3 (the increase of the observable volume of the universe). This is because the gravitational wave detectors to be discussed below measure signals from all directions; they cannot be pointed, but reside in a fixed position on the surface of the Earth.

A Michelson interferometer, with arms aligned along the x and y axes, can measure small phase differences between the light in the two arms. Therefore this type of interferometer can turn the length variations of the arms produced by a gravitational wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational wave detectors have evolved. Imagine a gravitational wave of amplitude h is incident on an interferometer. The change in the arm length will be $\Delta L \sim h L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. The Advanced LIGO and Advanced Virgo detectors will measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, much smaller than an atomic nucleus. The recent observation of gravitational waves has been one of the most spectacular accomplishments in experimental physics, and has been greeted with much excitement across the globe.

The history of the attempt to measure gravitational waves has been long. The realization that gravitational waves might be detectable crystallized as a result of the Conference on the Role of Gravitation in Physics, Chapel Hill, NC in 1957 (De Witt, 1957). Pirani (2009) had recently published a paper, and then gave presentation at the conference. He showed that the relative acceleration of particle pairs can be associated with the Riemann tensor. The interpretation of the Chapel Hill attendees was that nonzero components of the Riemann tensor were due to gravitational waves. Pirani, Richard Feynmann, and Hermann Bondi came up with the *sticky bead* argument (Bondi, 1957), essentially showing that gravitational waves exist and can be detected. Joe Weber, of the University of Maryland, was also at the Chapel Hill Conference, and from this inspiration he started to think about gravitational wave detection.

A few years after, in the early 1960s, Weber initiated the first experimental attempts to detect gravitational waves. Weber used a 1400 kg aluminum cylinder; a gravitational wave would excite the fundamental mechanical oscillation mode of the bar (Cho, 2016). The idea of using a Michelson interferometer to detect gravitational waves is almost as old as Weber's bar detector. In 1962 two Soviet physicists, Pustovoit and Gertsenshtein, noted that the use of a Michelson interferometer would be a possible means to detect gravitational waves over a frequency range that was broader than the Weber bars. In addition, the authors noted that the interferometers would have a sensitivity that would potentially be better than the Weber bars (Pustovoit and Gertsenshtein, 1963).

Also in the early 1960s, Weber and his student, Robert Forward, also considered using a Michelson interferometer to detect gravitational waves. After completing his PhD with Weber, Forward worked with Hughes Research Laboratories. It was at Hughes that Forward first constructed a Michelson interferometer to be used as a gravitational wave detector. Forward used earphones to listen to the motion of the interference signal (Forward, 1978). The engineering of signal extraction for modern interferometers is obviously far more complex.

At the time Forward was implementing an interferometric gravitational wave detector, Rainer Weiss at MIT produced a thorough investigation into not only how a Michelson interferometer could be used to detect gravitational waves, but also a systematic and comprehensive investigation into the noise sources that would constrain such a measurement (Weiss, 1972). However, as opposed to Forward's design where the laser beam traveled down the arms of the interferometer once, Weiss proposed a system where the laser beam would bounce back and forth multiple times in the interferometer, thereby increasing the effective arm length (and increasing the strain sensitivity of the detector). This is known as a Herriott optical delay line system. In what could be considered as the most important part of Weiss's (1972) presentation, he systematically listed and quantified the most important noise sources in an interferometric gravitational wave detector. These noise sources included amplitude noise on the laser source (including shot noise), frequency noise of the laser source, thermal noise in the masses and their suspension systems, radiation pressure noise from the laser light, seismic noise, noise due to residual gas in the vacuum system housing the interferometric detector, cosmic ray noise, gravitational-gradient noise, and residual electric and magnetic field noise. This comprehensive description of a realistic broadband interferometric gravitational wave detector initiated the experimental effort that has led to the present day LIGO and Virgo detectors.

Subsequently there was rapid activity on the construction of prototype laser interferometric gravitational wave detectors. The goal was to make prototypes that demonstrated the technology needed to construct kilometer-length interferometers. In the late 1970s, the Max Planck group in Garching, near Munich, Germany, created a 3 m arm length detector (Billing *et al.*, 1979), and

a 30 m detector in the early 1980s (Shoemaker *et al.*, 1988); this instrument was the first which showed a correspondence between noise models and performance in the spirit of Weiss's (1972) paper. Also, in the early 1980s Weiss and his team at MIT built laser interferometric gravitational wave detector with a Herriott optical delay line system and 1.5 m length arms. The Munich interferometers were also Herriott delay lines. The early 1980s also saw the construction of a 10 m prototype in Glasgow, Scotland (Ward *et al.*, 1985). The uniqueness of this system was that instead of the Herriott delay lines in the arms it used resonant optical cavities (Drever *et al.*, 1981, 1983), an idea by Ron Drever (then at Glasgow, before moving to Caltech). These Fabry–Perot cavities, to be discussed below, also have the light bounce back and forth a number of times, but in a resonant fashion within an optical cavity. A similar interferometer, albeit with 40 m arms, was then constructed at Caltech in the late 1980s (Spero, 1986). Fabry–Perot cavities were eventually incorporated into the design of LIGO and Virgo. The work on all of these prototypes were absolutely critical in establishing the technology needed for LIGO (Abramovici *et al.*, 1992). Similarly in the 1980s, the work on lasers, laser stabilization and interferometer optics in Orsay, France, plus the research on vibration isolation systems in Pisa, Italy, helped to create Virgo (Bradaschia *et al.*, 1990).

Numerous collaborations are building and operating second generation interferometers in order to detect gravitational waves. Advanced LIGO in the United States consists of two 4 km interferometers located in Livingston, Louisiana, and Hanford, Washington (Aasi *et al.*, 2015). Advanced LIGO started observations in 2015, and will be working over the coming years to achieve its design sensitivity, with the goal to reach it by 2019. The European Advanced Virgo is a 3 km interferometer near Pisa, Italy (Acernese *et al.*, 2015), and will start acquiring data in 2017, and will also be aiming for its target sensitivity in the coming years. GEO-600, a German-British collaboration, is a 600 m detector near Hanover, Germany (Affeldt *et al.*, 2014), and is currently operational. KAGRA is the Japanese 3 km interferometer that is presently under construction, and should commence observations in 2019 (Somiya, 2012). There will be a third 4 km LIGO interferometer, LIGO-India (Unnikrishnan, 2013), located in India, with the goal to be operational by 2024. All of the kilometer-length detectors will be attempting to detect gravitational waves with frequencies from 10 Hz up to a few kilohertz.

As will be described below, there are a number of terrestrial noise sources that will inhibit the performance of the interferometric detectors. The sensitivity of detection increases linearly with interferometer arm length, which implies that there could be advantages to constructing a gravitational wave detector in space. This is the goal of the laser interferometer space antenna (LISA) consortium. The plan is to deploy three satellites in a heliocentric orbit with a separation of about 2.5×10^6 km. LISA is a European Space Agency project, with a target launch date of 2034. LISA will observe gravitational waves in a frequency band from below 10^{-4} Hz to above 10^{-1} Hz. Due to the extremely long baseline, LISA is not strictly an interferometer, as most light will be lost as the laser beams expand, while traveling such a great distance. Instead, the phase of the received light will be detected and used to lock the phase of the light that is reemitted by another laser. Much of the technology needed for LISA to succeed was recently demonstrated with the LISA Pathfinder mission; in this mission the relative acceleration between two test masses was measured to be 5.2 ± 0.1 fm/s²/√Hz for frequencies between 0.7 and 20 mHz (Armano *et al.*, 2016).

Interferometer Configurations

The Michelson interferometer is the tool to be used to detect a gravitational wave. Fig. 1 shows a basic optical setup. The beamsplitter and the end mirrors would be suspended by wires, and effectively free to move in the plane of the interferometer. The arms have lengths L_1 and L_2 that are roughly equal on a kilometer scale. With a laser power P and wavelength λ incident on the beamsplitter, the light exiting the dark port of the interferometer is

$$P_{\text{out}} = P \sin^2 \left[\frac{2\pi}{\lambda} (L_1 - L_2) \right]$$

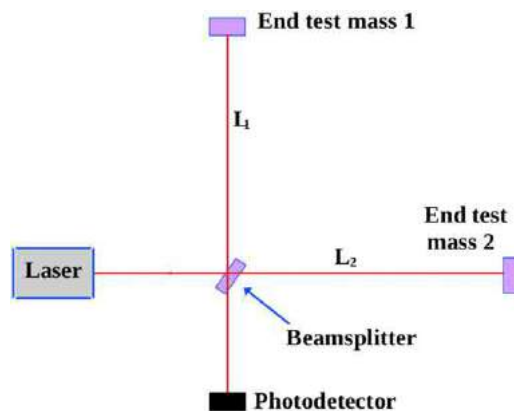


Fig. 1 A basic Michelson interferometer. The photodetector receives light exiting the dark port of the interferometer and hence the signal.

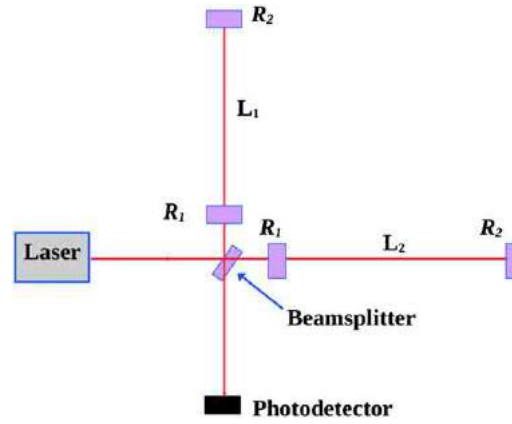


Fig. 2 A Michelson interferometer with Fabry-Perot cavities in each arm. The front cavity mirrors have reflectivity R_1 , while the end mirrors have $R_2 \sim 1$. By using Fabry-Perot cavities Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) will increase the effective arm length by a factor of 140.

The interferometer operates with the condition that in the absence of excitation the light exiting the dark port is zero. This would be the case for a simple and basic interferometer. If E_0 is the amplitude of the electric field from the laser, and assuming the use of a 50–50 beamsplitter, the electric field (neglecting unimportant common phase shifts) for the light incident on the photodetector would be

$$E_{\text{out}} = E(e^{i\delta\phi_1} - e^{i\delta\phi_2}) \approx i \frac{E_0}{2} (\delta\phi_1 - \delta\phi_2) = iE_0 \frac{2\pi}{\lambda} (L_1 - L_2)$$

A gravitational wave of optimal polarization normally incident upon the interferometer plane will cause one arm to decrease in length, while the other increases. The Michelson interferometer acts as a gravitational wave transducer; the stretching and squeezing of the spacetime between the mirrors results in more light exiting the interferometer dark port. The mirrors in the interferometer are suspended via fibers so that they are free to move under the influence of the gravitational wave, acting like relativistic freely falling masses.

An interferometer's sensitivity to gravitational waves increases with arm length, but geographical, physical, and financial constraints will limit the size of the arms. If there could be some way to bounce the light back and forth to increase the effective arm length it would increase the detector performance. Fabry-Perot cavities do just that. When they are on resonance they have a storage time for the light of

$$\tau_s = \frac{2L(R_1 R_2)^{1/4}}{[c(1 - \sqrt{R_1 R_2})]}$$

where (R_1 and R_2 are power reflection coefficients). **Fig. 2** shows the system of a Michelson interferometer with Fabry-Perot cavities. This gravitational wave interferometer design was proposed in the late 1970s by Ron Drever, and subsequently tested by his research group in the early 1980s (Drever *et al.*, 1981, 1983). The far mirror R_2 has a very high reflectivity ($R_2 \sim 1$) in order to ultimately direct the light back toward the beamsplitter. The front mirror reflectivity R_1 is such that LIGO's effective arm length increases from $L=4$ km to $L \sim 560$ km. The optical properties of the mirrors of the Fabry-Perot cavities must be exquisite in order to achieve success. The mirror substrates are made via a combination of super-polishing for small scale smoothness, and then ion-beam milling for large scale uniformity. The coatings (doped tantala) for the mirrors are ion-beam sputtered, multilayer dielectrics. The mirrors for Advanced LIGO (and Advanced Virgo) were coated by Laboratoire des Matériaux Avancés (LMA, Lyon, France). Advanced LIGO's mirrors were tested, and the surface errors are between 0.08 and 0.23 nm, with absorption between 0.2 (parts per million) and 0.4 ppm. In terms of both absorption and scattering, the Advanced LIGO arm round-trip loss goal is less than 75 ppm. The radius of curvature for the input mirrors (R_1) is 1934 m, while for the end mirrors (R_2) it is 2245 m, with a radius of curvature spread between -1.5 and 1.0 m (Aasi *et al.*, 2015). A LIGO test mass (and therefore a Fabry-Perot mirror) can be seen in **Fig. 3**. The mirrors for Advanced Virgo have similar exquisite properties (Acernese *et al.*, 2015).

In 1888 Michelson and Morley, with their interferometer, had a sensitivity that allowed the measurement of 0.02 of a fringe, or about 0.126 rad. The Advanced LIGO interferometers during the first observing run O1 have already demonstrated a phase noise spectral density of

$$\phi(f) = 5 \times 10^{-11} \text{ radian}/\sqrt{\text{Hz}}$$

for frequencies around 150 Hz. Assuming a 150 Hz signal with a 150 Hz bandwidth this implies a phase sensitivity of $\Delta\phi = 6.1 \times 10^{-10}$ rad. There has been quite an evolution in interferometry since Michelson's time.

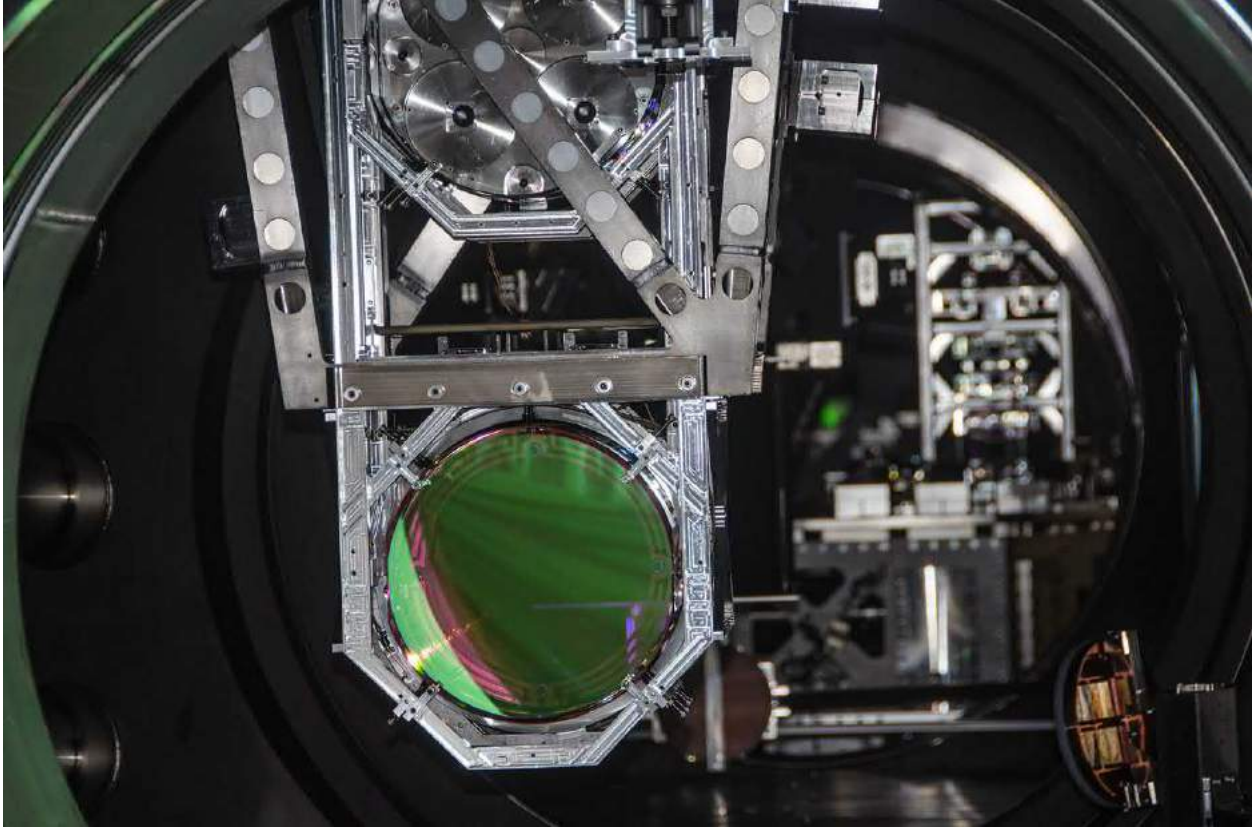


Fig. 3 A picture of the input test mass (mirror R_1) for Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) within its vibration isolation suspension system. The fused silica component is 40 kg, 34 cm in diameter, and 20 cm thick. Photograph courtesy of LIGO/Caltech/MIT.

The noise sources that inhibit the interferometer performance are discussed below. However, let us consider one's ability to measure the relative phase between the light in the two arms. The Heisenberg uncertainty relation for light with phase ϕ and photon number N is $\Delta\phi\Delta N \sim 1$. For a measurement lasting time τ using laser power P and frequency f , the photon number is $N = P\lambda\tau/hc$ (here h is Planck's constant), and with Poisson statistics describing the light $\Delta N = \sqrt{N} = \sqrt{P\lambda\tau/hc}$. Therefore

$$\Delta\phi\Delta N = \frac{2\pi}{\lambda}\Delta L\sqrt{P\lambda\tau/hc} = 1$$

implies that

$$\Delta L = \frac{1}{2\pi}\sqrt{hc\lambda/P\tau}$$

With more light power the interferometer can measure smaller distance displacements and achieve better sensitivity. Advanced LIGO and Advanced Virgo will use about 200 W of laser light. However, there is a nice trick one can use to produce more light circulating in the interferometer, namely power recycling (Meers, 1988). Fig. 4 displays the power recycling interferometer design. The interferometer operates such that virtually none of the light exits the interferometer dark port, and the bulk of the light returns toward the laser. An additional mirror, R_r , in Fig. 4, recycles the light. The Advanced LIGO goals are to have 125 W actually impinging on the recycling mirror R_r , creating 5.2 kW upon the beamsplitter, and 750 kW within the Fabry-Perot cavities in each of the interferometer's arms. For Advanced LIGO, recycling will increase the effective light power by another factor of 42. Advanced Virgo has a similar design. The higher circulating light power therefore improves the sensitivity of these interferometric detectors. It is also interesting to note that the GEO-600 detector has been operating for years using squeezed light, namely quantum states of light, as a way to reduce the noise below the shot noise limit (Affeldt et al., 2014).

There is one additional modification to the interferometer system that can further improve sensitivity, but only at a particular frequency. A further Fabry-Perot system can be made by installing what is called a signal recycling mirror (SRM); this would be mirror R_s in Fig. 5 (Meers, 1988). Imagine the light in arm 1 of the interferometer, and that it acquires phase as the arm expands due to a gravitational wave. The traveling gravitational wave's oscillation will subsequently cause arm 1 to contract, while arm 2 expands. If the light that was in arm 1 could be sent to arm 2, while it is expanding, then the beam would acquire additional phase. This process could be repeated over and over. Mirror R_s serves this purpose, with its reflectivity defining the storage time for light in each interferometer arm. The storage time defined by the cavity formed by the SRM, R_s , and the mirror at the front of the interferometer arm cavity, R_1 , determines the resonance frequency. Signal recycling will give a substantial boost to interferometer

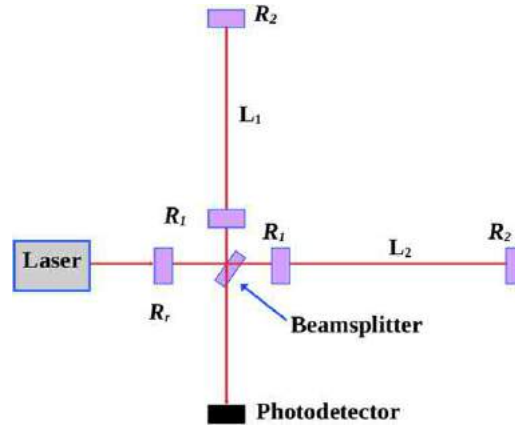


Fig. 4 A power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light would exit the interferometer through the light port and head back to the laser. Installation of the recycling mirror with reflectivity R_r sends the light back into the system. A Fabry–Perot cavity is formed between the recycling mirror and the first mirror (R_1) of the arms. For Advanced LIGO this strategy will increase the power circulating in the interferometer by a factor of 42.

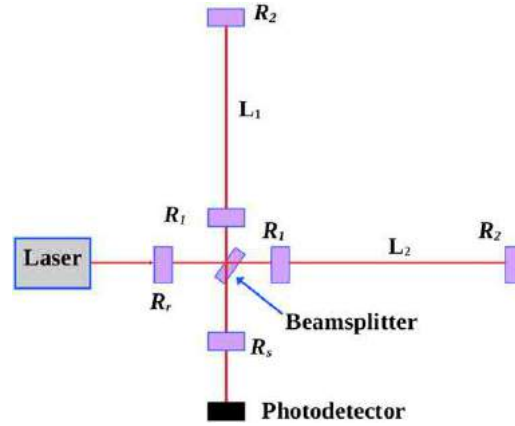


Fig. 5 A signal recycled and power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light containing the gravitational wave signal would exit the interferometer via the dark port and head to the photodetector. Installation of the SRM with reflectivity R_s sends the light back into the system. The phase of the light acquired from the gravitational wave will build up at a particular frequency determined by the reflectivity R_s .

sensitivity at a particular frequency, and will eventually be implemented in all the main ground-based interferometric detectors. The Advanced LIGO and Advanced Virgo interferometers are infinitely more complex than the relatively simple systems displayed in the figures of this paper.

Fig. 6(a) presents an aerial view of the LIGO site at Hanford, Washington. The magnitude of the 4 km system is apparent. **Fig. 6(b)** displays the Virgo detector with its 3 km, located near Pisa, Italy.

Noise Sources and Interferometer Sensitivity

If the interferometers are to detect distance displacements less than 10^{-18} m then they must be isolated from a host of deleterious noise sources. Seismic disturbances should not shake the interferometers. Thermal excitation of components will affect the sensitivity of the detector and should be minimized. The entire interferometer must be in an adequate vacuum in order to avoid fluctuations in gas density that would cause changes in the index of refraction and hence a modification of the optical path length. The laser intensity and frequency noise must be minimized. The counting statistics of photons influences accuracy. If ever there was a detector that must avoid Murphy's law this is it; little things going wrong cannot be permitted if such small distance displacements are to be detected. The target noise sensitivity for the Advanced LIGO interferometers is displayed in **Fig. 7**.

In the best of all worlds the interferometer sensitivity will be limited by the counting statistics of the photons. A proper functioning laser will have its photon number described by Poisson statistics, or shot noise; if the mean number of photons arriving per unit time is N then the uncertainty is $\Delta N = \sqrt{N}$, which as noted above implies an interferometer displacement



(a)



(b)

Fig. 6 (a) Aerial view of the Laser Interferometer Gravitational-Wave Observatory (LIGO) Hanford, Washington site. The vacuum enclosure at Hanford contains the 4 km interferometer. Photograph courtesy of LIGO/Caltech/MIT. (b) Aerial view of the Virgo detector, with 3 km arms, located near Pisa, Italy. Photograph courtesy of the European Gravitational Observatory.

sensitivity of

$$\Delta L = \frac{1}{2\pi} \sqrt{\frac{hc\lambda}{P\tau}}$$

(where P is the light power impinging on the beamsplitter) or a spectral density of

$$\Delta L(f) = \frac{1}{2\pi} \sqrt{\frac{hc\lambda}{P}}$$

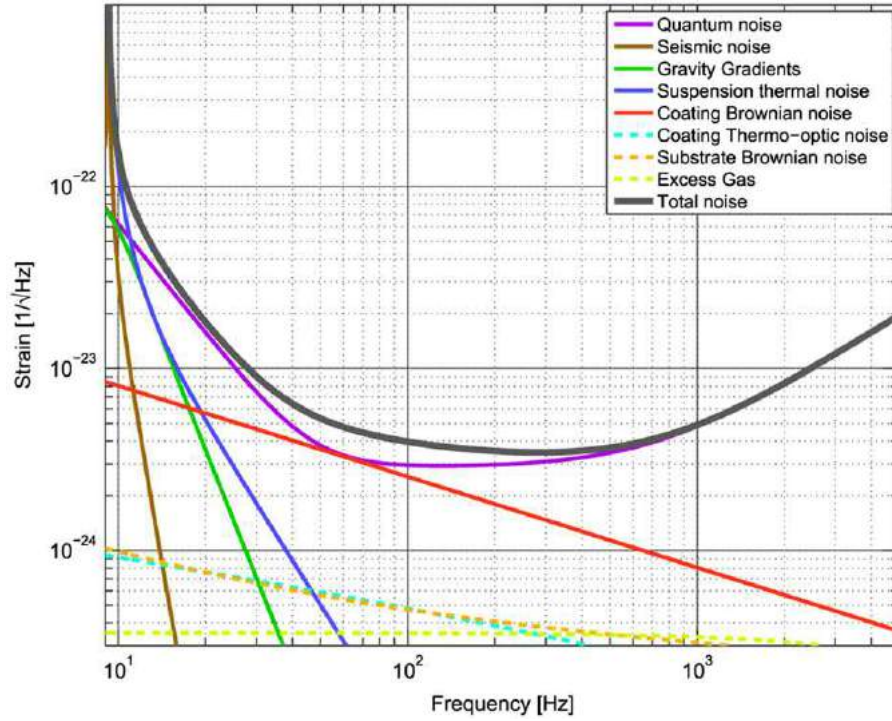


Fig. 7 The target spectral density of the noise for the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) system. Advanced LIGO will be limited by seismic noise at low frequencies (~ 10 Hz), thermal noise (from the suspension system and the coatings of the mirrors) in the intermediate regime (~ 100 Hz). Radiation pressure will also be a dominating noise source from ~ 10 to ~ 100 Hz, while photon shot noise will be the limiting noise thereafter. Together the radiation pressure noise and the shot noise are referred to as quantum noise. Other sources of noise are also noted in the figure. This figure is from Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. *Classical and Quantum Gravity* 32. Available at: <http://stacks.iop.org/0264-9381/32/i=7/a=074001.s>; which should be consulted for a more expansive description of the limiting noise sources for Advanced LIGO.

in units of $\text{m}/\sqrt{\text{Hz}}$. Note also that the sensitivity increases as the light power increases. The reason for this derives from the statistics of repeated measurements. The relative lengths of the interferometer arms could be measured, once, by a photon. However, the relative positions are measured repeatedly with every photon from the laser, and the variance of the mean decreases as $\sqrt{N}$ where N is the number of measurements (or photons) involved. The uncertainty in the difference of the interferometer arm lengths is therefore inversely proportional to photon number, and hence the laser's power. In terms of strain sensitivity this would imply

$$h(f) = \frac{1}{2\pi L} \sqrt{\frac{hc\lambda}{P}}$$

This assumes the light just travels down the arm and back once. With Fabry-Perot cavities the light is stored, and the typical photon takes many trips back and forth before exiting the system. In order to maximize light power the end mirrors ($R_2 \sim 1$) and the strain sensitivity is improved to

$$h(f) = \frac{1}{4\pi\tau_s} \sqrt{\frac{\pi\hbar\lambda}{Pc}}$$

where the Fabry-Perot cavity storage time τ_s was defined above.

As the frequency of gravitational waves increases the detection sensitivity will decrease. If the gravitational wave causes the interferometer arm length to increase, then decrease, while the photons are still in the arm cavity, then the phase acquired from the gravitational wave will be washed away. This is the reason why interferometer sensitivity decreases as frequency increases, and explains the high frequency behavior seen in Fig. 7. Taking this into account, the strain sensitivity is

$$h(f) = \frac{1}{4\pi\tau_s} \sqrt{\frac{\pi\hbar\lambda}{Pc}} (1 + (4\pi f\tau_s)^2)^{1/2}$$

and f is the frequency of the gravitational wave.

If the gravitational wave is to change the interferometer arm length then the mirrors that define the arm must be free to move like freely falling masses. In systems like Advanced LIGO and Advanced Virgo, wires suspend the mirrors; each mirror is like a pendulum. The mirrors and the wires that suspend them are a monolithic-fused silica assembly, with the wires annealed and

welded to the sides of the mirrors. The pendulum itself is the first component of an elaborate vibration isolation system. Seismic noise will be troublesome for the detector at low frequencies. The spectral density of the seismic noise is about $x(f) = (10^{-9} \text{ m}/\sqrt{\text{Hz}})(10 \text{ Hz}/f)^2$ for $f > 10 \text{ Hz}$ (the low frequency observational limit for Advanced LIGO and Advanced Virgo) (Saulson, 1994). A simple pendulum, by itself, acts as a motion filtering device. Above its resonance frequency a pendulum filters motion with a transfer function like $T(f) \propto (f_0/f)^2$, where f_0 is the resonant frequency for the pendulum. The mirrors for Advanced LIGO will actually be suspended by a four-stage pendulum system. The various gravitational wave detector collaborations have different vibration isolation designs. The mirrors in these interferometers are suspended in elaborate vibration isolation systems, which may include multiple pendulums, isolation stacks, and isolated optical tables. For example, for many years super-attenuators have made Virgo the most sensitive gravitational wave detector in the low frequency regime (below $\sim 40 \text{ Hz}$) (Acernese *et al.*, 2015). Active feedback is used on some parts of the isolation system to control seismic noise below $\sim 10 \text{ Hz}$. Seismic noise will be the limiting factor for interferometers seeking to detect gravitational waves in the vicinity of $\sim 10 \text{ Hz}$, as can be seen in the sensitivity curve presented in Fig. 7.

Due to the extremely small distance displacements that these systems are trying to detect it should come as no surprise that thermal noise is a problem. This noise enters through a number of components in the system. The two most serious thermal noise sources are the wires suspending the mirrors in the pendulum, and the mirrors themselves, especially the optical coatings on the mirror surfaces. Consider the wires; there are a number of notes at which they can oscillate (i.e., violin modes). At temperature T each mode will have energy of $k_B T$, but distributed over a band of frequencies determined by the quality factor (or Q) of the material. Low-loss (or high- Q) materials work best; for the violin modes of the wires there will be much noise at particular frequencies (in the hundreds of hertz). For the Advanced LIGO mirrors the first violin mode is at 510 Hz , while the vertical stretching mode of the wires is at $\sim 9 \text{ Hz}$.

The best sensitivity for Advanced LIGO and Advanced Virgo occurs around $\sim 100 \text{ Hz}$. The limiting source of noise in this region (along with quantum noise) is due to Brownian noise in the optical coatings on the mirror surfaces. There is a tremendous amount of on-going research to try and reduce the mechanical dissipation in the optical coatings. The Japanese detector KAGRA, which is currently under construction, will have its mirrors and the bottom parts of its suspension system cooled to 20 K (Somiya, 2012) in order to reduce thermal noise.

The frequency noise of the laser can couple into the system to produce length displacement noise in the interferometer. With arm lengths of $\sim 4 \text{ km}$, it will be impossible to hold the length of the two arms absolutely equal. The slightly differing arm spans will mean that the light sent back from each of the two Fabry-Perot cavities will have slightly differing phases. As a consequence, great effort is made to stabilize the frequency of the light entering the interferometer. The Advanced LIGO laser can be seen in Fig. 8. The primary laser is a nonplanar ring-oscillator (NPRO). This beam is then amplified to 35 W with a medium power oscillator, and then up to 220 W with a high power oscillator; see Aasi *et al.* (2015) for more details. For Advanced LIGO, the laser is locked and held to a specific frequency by use of signals from a reference cavity, a mode cleaner cavity, and the interferometer.

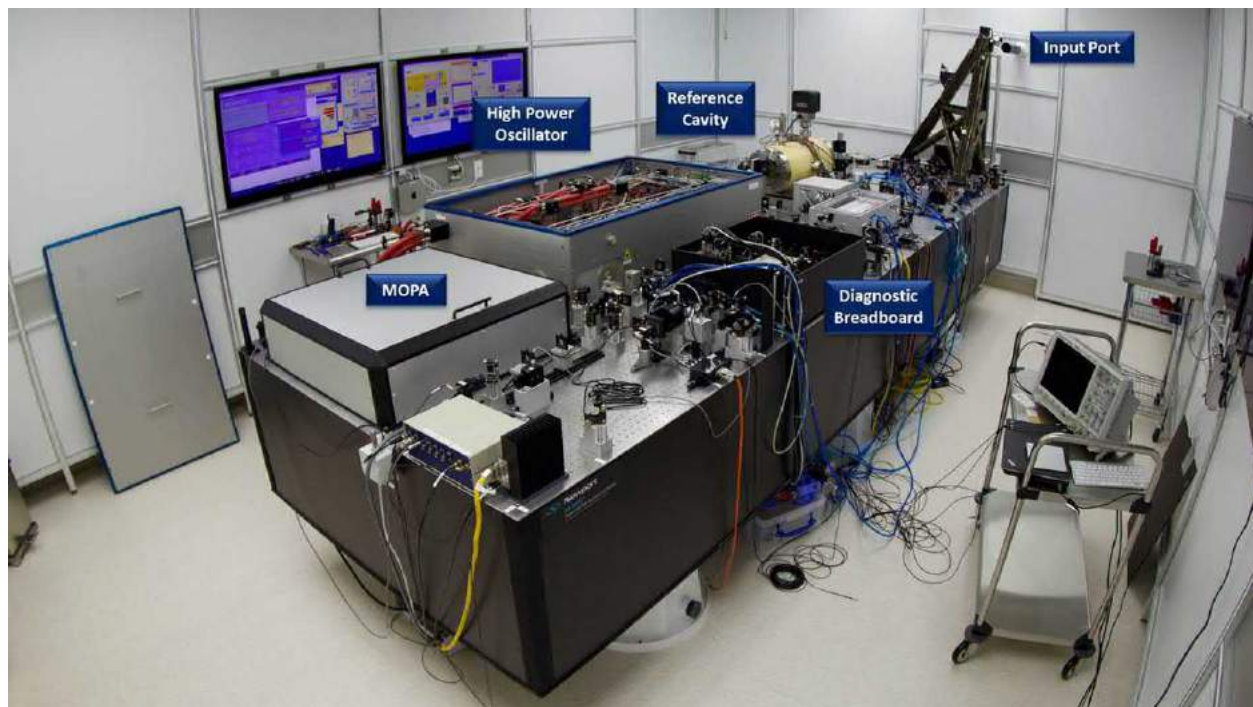


Fig. 8 The Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) laser system. This is a multistage Nd:YAG system that can deliver 200 W . Photograph courtesy of LIGO/Caltech/MIT.

For low frequency stabilization the temperature of the NPRO is adjusted. At intermediate frequencies adjustment is made by signals to a piezoelectric transducer within the NPRO cavity. At high frequencies the noise is reduced with the use of an electro-optic crystal. The Advanced LIGO lasers currently have a frequency noise of $1 \times 10^{-6} \text{ Hz}/\sqrt{\text{Hz}}$ at 100 Hz; this requirement is needed by Advanced Virgo too.

It is also important to worry about the stability of the laser power for the interferometric detectors. The goal is to be quantum noise limited at frequencies within the observational frequency band. The Nd:YAG power amplifiers used are pumped with an array of laser diodes, so the light power is controlled through feedback to the laser diodes. The Advanced LIGO requirements for the fluctuations on the power P are $\Delta P/P < 2 \times 10^{-9}/\sqrt{\text{Hz}}$ at 10 Hz; Advanced Virgo's requirements are similar. In these laser interferometric gravitational wave detectors, the spatial quality of the light is ensured through the use of an input mode cleaning cavity. Advanced LIGO uses an isosceles triangular array of mirrors with the two base mirrors separated by 0.465 m and the third mirror displaced by 16.24 m. The length of the input mode cleaner for Advanced Virgo is 143.424 m. The optical system for Advanced LIGO is displayed in Fig. 9. Aside from the laser and the phase modulator, the entire optical system is an ultra-high vacuum. Note that at the output of the interferometer there is a SRM. Given the reflectivity of the mirror, and the phase of the light when arriving at there, it is possible to enhance the gravitational wave signal at a particular frequency by feeding it back into the interferometer for enhancement. The output mode cleaner is used to improve the spatial quality of the output beam, and remove modulation frequency sidebands, before the photodetection.

The target light powers for Advanced LIGO are displayed in Fig. 9. When Advanced LIGO attains its target sensitivity there will be 750 kW within the Fabry–Perot cavities. Advanced Virgo's light powers will be similar. With such a large amount of power the

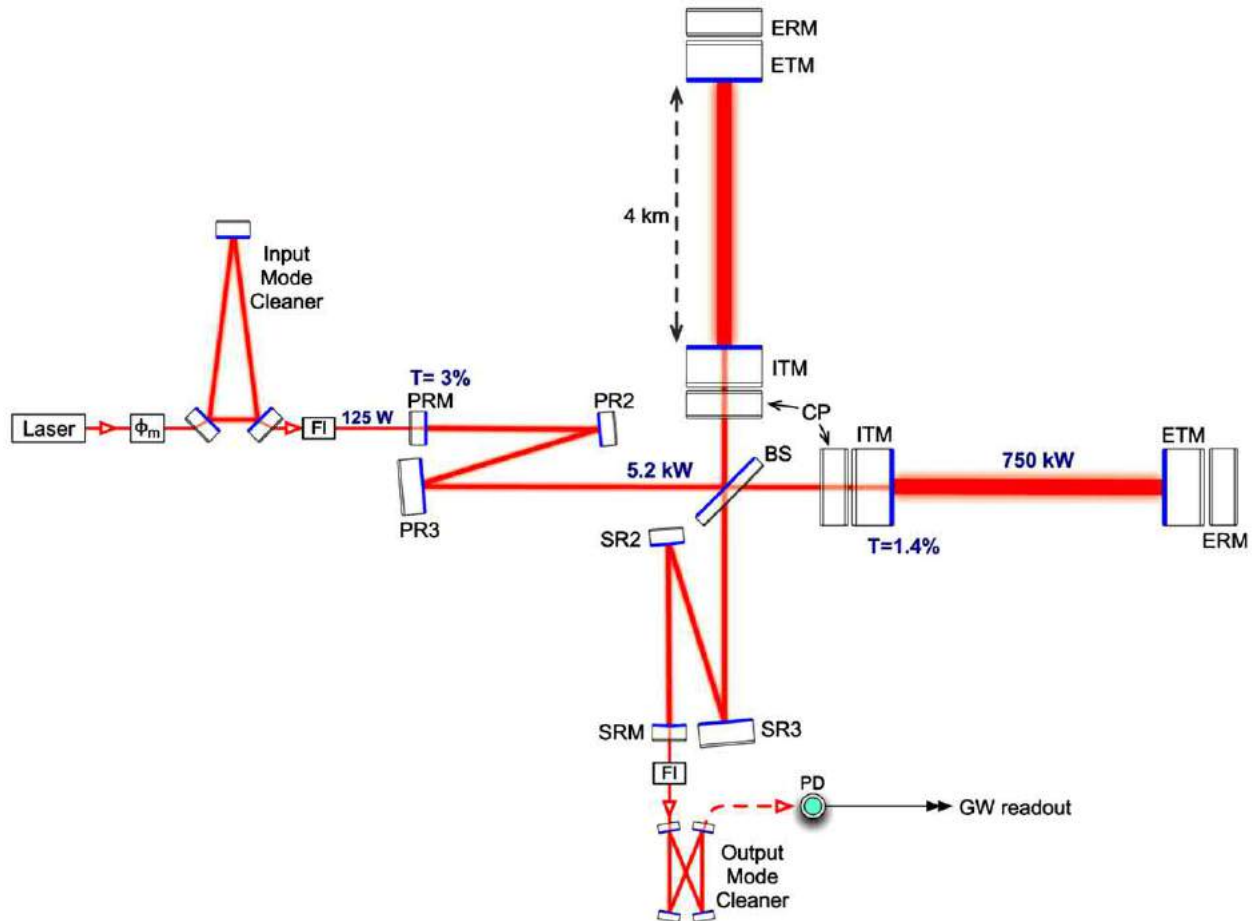


Fig. 9 The Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) optical system. The laser (~ 200 W) light propagates from the stabilized laser through a phase modulator (ϕ_m) to the input mode cleaner, then through a Faraday isolator (FI) to the power recycling mirror (PRM). The folding mirrors, PR2 and PR3, direct the light to the beamsplitter (BS) input of the interferometer. Note that approximately 125 W of light impinges upon the power recycling mirror, resulting in 5.2 kW at the input port to the beamsplitter. The Fabry–Perot cavities are formed with the input test mass (ITM), which is coupled to a compensation plate (CP), and the end test mass (ETM) which is coupled to an end reaction mass (ERM). The Fabry–Perot cavities will contain 750 kW of light power. Note too that the output signal from the interferometer can itself be recycled and amplified at specific frequencies, dependent on the reflectivity of the SRM; SR2 and SR3 are folding mirrors. The output beam also has its spatial features cleaned with the output mode cleaner before the light falls upon a photodetector (PD). The figure is from Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. *Classical and Quantum Gravity* 32. Available at: <http://stacks.iop.org/0264-9381/32/i=7/a=074001.s>.

mirrors will actually get heated and lightly change their shape. Ring heaters encircle the input test masses and the end test masses, and are used to correct the shape of the masses. A CO₂ laser beam is also sent onto the surface of the compensation plate to provide further corrections to the thermal lens of the input test mass. The compensation plate serves as a reaction mass for the input test mass in its isolation suspension system; the same is true for the end reaction mass with respect to the end test mass.

Parametric instabilities are another consequence of high power operation for Advanced LIGO and Advanced Virgo. This is an interaction between a mechanical oscillation mode of the mirror and higher order optical modes via light scattering. This can be a nonlinear process and can prevent the interferometers from operating at higher powers. Advanced LIGO has observed parametric instabilities (Evans *et al.*, 2015). To eliminate the excited modes one can heat the mirror to slightly change its shape, thereby changing the mechanical oscillation mode with respect to an excited optical mode. Advanced LIGO uses electrostatic actuators to move the masses; the actuators can also be used to damp excited mechanical modes.

The immense number of photons, coupled with the fact that the photon arrival times are random (Poisson statistics) means that radiation reaction noise will be important. This can be seen in the low frequency component of the quantum noise in Fig. 7. The low frequency radiation reaction noise, plus the high frequency shot noise, combine to create the total quantum noise in the interferometer. At high frequencies the shot noise decreases with laser power, while at low frequencies radiation reaction noise and the basic interferometer's quantum noise is a tradeoff between these two effects. While this quantum noise seems to be an unavoidable noise source, quantum states of light (namely squeezed states of light) can reduce this noise. This reduction of noise was demonstrated by initial LIGO (Aasi *et al.*, 2013) where squeezed light reduced the noise below the shot noise level for frequencies above 150 Hz; for higher frequencies a 2.15 dB (28%) reduction in the shot noise was observed. The GEO-600 gravitational wave detector now uses squeezed light continuously, and has achieved an impressive 3.7 dB reduction in the shot noise level (Affeldt *et al.*, 2014). The use of quantum states of light is one of the ways that Virgo and LIGO hope to reduce their noise in the years to come.

Gravitational Wave Detection GW150914

On September 14, 2015, at 09:50:45 UTC a gravitational wave was detected directly for the first time. The gravitational wave was first observed at the LIGO Livingston Observatory (Louisiana), and then 7 ms later at the LIGO Hanford Observations (Washington). An on-line signal search algorithm identified the signal in 3 min. An off-line examination of the data using a template-based search for compact binary coalescence signals identified the gravitational wave with a signal-to-noise ratio of 24

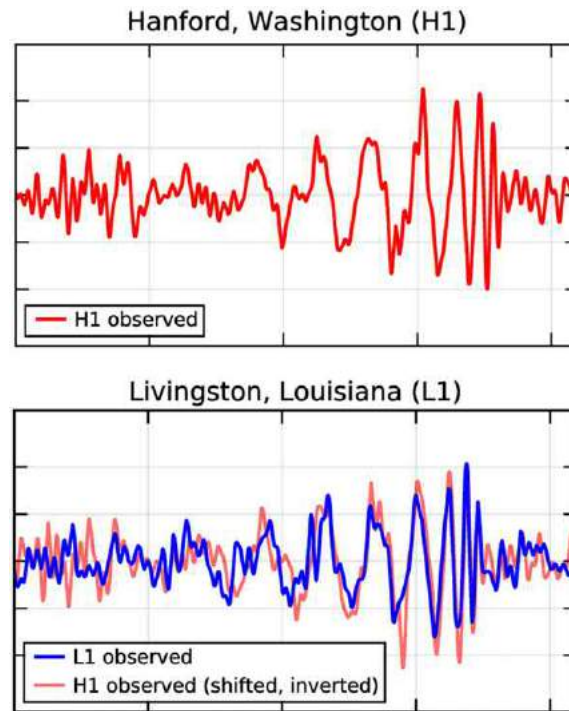


Fig. 10 The measured gravitational wave signal GW150914 as observed at the two Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) interferometric detectors. The data has been bandpass filtered (35–350 Hz), and the gravitational wave signal is clearly observable by eye. Top: the signal as observed from the LIGO Hanford detector. Bottom: the signal as observed from the LIGO Livingston detector (blue). In addition, the Hanford signal (red) is superimposed after it has been displaced by 7 ms and inverted (due to the relative orientation of the two detectors). The similarity of the two measured signals is clearly visible. Figures courtesy of the LIGO Open Science Center (losc.ligo.org).

(Abbott *et al.*, 2016a). Parameter estimation routines were used to determine that the gravitational wave signal was emitted from the merger of two black holes with masses of $36 M_{\odot}$ and $29 M_{\odot}$. The newly created black hole had a mass of $62 M_{\odot}$, meaning that the total energy of gravitational wave emitted was equivalent to $3 M_{\odot} c^2$. The system was 1.3 billions light-years away from us when it merged.

The measured gravitational wave signal, GW150914, from the two LIGO detectors is displayed in Fig. 10 (Abbott *et al.*, 2016a). The peak amplitude of GW150914 is $h \sim 10^{-21}$ which corresponds to a displacement of the interferometers' arms of $\Delta L \sim 2 \times 10^{-18}$ m. The exquisite sensitivity of these interferometers can be seen from these numbers. In addition to GW150914, during Advanced LIGO's first observing run two other gravitational wave events were observed (also stellar mass binary black hole mergers) (Abbott *et al.*, 2016b,c).

Conclusions

The observation of gravitational waves using Michelson interferometers testifies to the utility of these devices, and to the scientists that have made them work so well. More than a hundred years ago Michelson succeeded in carrying off experiments of amazing difficulty as he measured the speed of light and disproved the existence of the aether. Gravitational wave detection is an experiment worthy of Michelson. In addition, the detection of gravitational waves a century after their prediction by Albert Einstein, and his observation that they will never be observed, testifies to the tremendous progress that has been made in technology, notably here in optics. A new window into the universe has been created, gravitational wave astronomy.

LIGO, Virgo, GEO, and KAGRA are creating a new type of telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. This has started with the unexpected observation of gravitational waves produced by binary black hole systems with tens of solar masses. Physicists do know that there will be other signals that they can predict: binary systems containing neutron stars, for example. It is suspected that short gamma ray bursts come from the coalescence of binary neutron stars, or neutron star – black hole binary systems. A core-collapse supernova will produce a burst of gravitational waves that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravitational waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort. The exciting initial observations of gravitational waves have been made, but it is just the beginning of a new astronomy.

Acknowledgements

NC is supported by National Science Foundation grant PHY-1505373. This article has been assigned LIGO Document number P1700043.

References

- Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. Classical and Quantum Gravity 32, 074001. Available at: <http://stacks.iop.org/0264-9381/32/i=7/a=074001.s>.
- Aasi, J., Abadie, J., Abbott, B., *et al.*, 2013. Enhanced sensitivity of the LIGO gravitational wave detector by using squeezed states of light. Nature Photonics 7, 613–619.
- Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016a. Physical Review Letters 116 (6), 061102. Available at: <http://link.aps.org/doi/10.1103/PhysRevLett.116.061102>.
- Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016b. Physical Review Letters 116 (24), 241103. Available at: <http://link.aps.org/doi/10.1103/PhysRevLett.116.241103>.
- Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016c. Physical Review X 6 (4), 041015. Available at: <https://link.aps.org/doi/10.1103/PhysRevX.6.041015>.
- Abramovici, A., Althouse, W.E., Drever, R.W.P., *et al.*, 1992. Science, 256. pp. 325–333. Available at: <http://science.sciencemag.org/content/256/5055/325>; ISSN: 0036-8075.
- Acernese, F., Agathos, M., Agatsuma, K., *et al.*, 2015. Classical and Quantum Gravity 32, 024001. Available at: <http://stacks.iop.org/0264-9381/32/i=2/a=024001>.
- Affeldt, C., Danzmann, K., Dooley, K.L., *et al.*, 2014. Classical and Quantum Gravity 31, 224002. Available at: <http://stacks.iop.org/0264-9381/31/i=22/a=224002>.
- Armano, M., Audley, H., Auger, G., *et al.*, 2016. Physical Review Letters 116 (23), 231101. Available at: <http://link.aps.org/doi/10.1103/PhysRevLett.116.231101>.
- Billings, H., Maischberger, K., Rudiger, A., *et al.*, 1979. Journal of Physics E: Scientific Instruments 12, 1043. Available at: <http://stacks.iop.org/0022-3735/12/i=11/a=010>.
- Bondi, H., 1957. Nature 179, 1072. Available at: <http://dx.doi.org/10.1038/1791072a0>.
- Bradaschia, C., Fabbro, R.D., Virgilio, A.D., *et al.*, 1990. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 289, 518–525. Available at: <http://www.sciencedirect.com/science/article/pii/016890029091525G>; ISSN: 0168-9002.
- Cho, A., 2016. Science. doi:10.1126/science.aaf4057.
- De Witt, C.M. (Ed.), 1957. Proceedings: Conference on the Role of Gravitation in Physics, Chapel Hill, NC.
- Drever, R.W.P., Ford, G.M., Hough, J., *et al.*, 1983. A gravity-wave detector using optical cavity sensing. In: Schmutzer, E. (Ed.), Proceedings of the Ninth International Conference on General Relativity and Gravitation, Jena, July 1980, pp. 265–267.
- Drever, R.W.P., Hough, J., Munley, A.J., *et al.*, 1981. Optical Cavity Laser Interferometers for Gravitational Wave Detection. Berlin; Heidelberg: Springer, pp. 33–40. Available at: http://dx.doi.org/10.1007/978-3-540-38804-3_4; ISBN: 978-3-540-38804-3.
- Evans, M., Gras, S., Fritschel, P., *et al.*, 2015. Physical Review Letters 114 (16), 161102. Available at: <http://link.aps.org/doi/10.1103/PhysRevLett.114.161102>.
- Forward, R.L., 1978. Physical Review D 17 (2), 379–390. Available at: <http://link.aps.org/doi/10.1103/PhysRevD.17.379>.

- Meers, B.J., 1988. Physical Review D 38 (8), 2317–2326. Available at: <https://link.aps.org/doi/10.1103/PhysRevD.38.2317>.
- Pirani, F.A.E., 2009. Republication of: On the physical significance of the Riemann tensor. General Relativity and Gravitation 41, 1215–1232.
- Pustovoit, V., Gertsenshtein, M., 1963. Soviet Physics JETP-USSR 16, 433–435.
- Saulson, P., 1994. Fundamentals of Interferometric Gravitational Wave Detectors. Available at: <https://books.google.fr/books?id=F87sCgAAQBAJ>; ISBN: 9789814501903.
- Shoemaker, D., Schilling, R., Schnupp, L., *et al.*, 1988. Physical Review D 38 (2), 423–432. Available at: <https://link.aps.org/doi/10.1103/PhysRevD.38.423>.
- Somiya, K., 2012. Classical and Quantum Gravity 29, 124007. Available at: <http://stacks.iop.org/0264-9381/29/i=12/a=124007>.
- Spero, R., 1986. The Caltech laser-interferometric gravitational wave detector. In: Ruffini, R., (Ed.), Fourth Marcel Grossmann Meeting on General Relativity, pp. 615–620.
- Taylor, J., Weisberg, J., 1989. Further experimental tests of relativistic gravity using the binary pulsar PSR 1913 + 16. Astrophysical Journal 345, 434–450.
- Unnikrishnan, C.S., 2013. International Journal of Modern Physics D 22, 1341010. Available at: <http://www.worldscientific.com/doi/abs/10.1142/S0218271813410101>.
- Ward, H., Hough, J., Newton, G.P., *et al.*, 1985. IEEE Transactions on Instrumentation and Measurement IM-34, pp. 261–265. ISSN: 0018-9456.
- Weiss, R., 1972. Electromagnetically coupled broadband gravitational antenna. Technical Reports MIT Quarterly Report of the Research Laboratory for Electronics. Available at: <https://dcc.ligo.org/LIGO-P720002/public/main>.

Cavity QED Effects in Molecular Systems

Stéphane Kéna-Cohen, Polytechnique Montréal, Department of Engineering Physics, Montreal, QC, Canada

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic thin films have found widespread use in display technology. The ease with which these materials can be processed and the versatility afforded by molecular design have been two of the main catalysts for the spread of organic light-emitting diode (OLED) technology. These characteristics have also made organic molecules attractive platforms for studying light-matter interaction in various regimes. The nature of optical excitations in molecular materials depends greatly on the type and structure of the molecular system. The underlying building block can range from small molecule, to polymer, to macromolecules such as proteins. In an isolated molecule, the optical absorption energy differs from the energy difference between the lowest unoccupied molecular orbital and highest occupied molecular orbital by the electron-hole pair (or exciton) binding energy. Furthermore, when strong intermolecular interactions are present, such as in certain polymers or crystalline aggregates, energy levels are broadened into bands and wavefunctions can become delocalized over many molecules. In this both of these limits, cavity QED effects have been exploited to modify the absorption and emission properties of organic molecules. For example, they can be used to increase the fluorescence efficiency of organic molecules by several orders of magnitude, to tailor the angular emission of molecules in organic thin films and to create hybrid light-matter modes called exciton-polaritons where the cavity photon and molecular excited state become intrinsically linked.

Modification of Molecular Spontaneous Emission

The possibility to modify molecular fluorescence by changing the optical environment was beautifully demonstrated in a series of pioneering experiments by Drexhage *et al.* where the radiative decay rate of fluorophore was measured as a function of its distance from a metal surface. In the quantum theory, the change in rate can be understood to occur from a modification by the metal of the local density of optical states (LDOS) at the emitter position. In the electric dipole approximation, Fermi's Golden rule predicts a radiative rate:

$$\gamma = \frac{\pi\omega_0 p^2}{3\hbar\epsilon_0} \rho(\mathbf{r}_0, \omega_0)$$

where ω_0 is the emitter transition frequency, $\mathbf{p} = p\hat{n}$ is its transition dipole moment and $\rho(\mathbf{r}_0, \omega_0)$ the density of optical states evaluated at the emitter position $\mathbf{r}_0$. The LDOS consists of a sum of the normalized field intensities, performed over all of the normal modes supported by the optical environment. The form of the LDOS will favor emission into normal modes with large field intensities at the emitter position and be inhibited in those with vanishing intensities.

Classically, one can understand the change in radiative rate as being due to the action of the radiation field scattered by the environment back onto the molecular dipole. To quantitatively explain Drexhage's experiments, Chance, Prock and Silbey introduced the powerful Dyadic Green's function approach. The Dyadic Green's function, $\mathbf{G}(\mathbf{r}, \mathbf{r}')$, a tensor, can be used to express the electric field, $\mathbf{E}(\mathbf{r})$, resulting from a given current density, $\mathbf{J}(\mathbf{r})$, using:

$$\mathbf{E}(\mathbf{r}) = i\omega\mu_0 \int_V \mathbf{G}(\mathbf{r}, \mathbf{r}') \mathbf{J}(\mathbf{r}') d^3\mathbf{r}'$$

It is obtained by solving the inhomogeneous wave equation within the studied structure with the appropriate boundary conditions. Then, for a molecule that is modeled as a localized oscillating current density (a point dipole), the electric field generated by this dipole is simply given by:

$$\mathbf{E}(\mathbf{r}) = \mu_0\omega^2 \mathbf{G}(\mathbf{r}, \mathbf{r}_0) \cdot \mathbf{p}$$

Given that the rate of energy dissipation due to the oscillating charge density is:

$$W = -\frac{1}{2} \int_V \text{Re}\{\mathbf{J}^* \cdot \mathbf{E}\} d^3\mathbf{r}$$

we find that the energy dissipation rate of the dipole is:

$$\dot{W} = \frac{\omega}{2} \text{Im}\{\mathbf{p}^* \cdot \mathbf{E}(\mathbf{r}_0)\} \propto \hat{n} \cdot \text{Im}\{\mathbf{G}(\mathbf{r}_0, \mathbf{r}_0)\} \cdot \hat{n}$$

where $\hat{n}$ is a unit vector along the dipole. The radiative decay rate is thus be related to the imaginary part of the Dyadic Green's function evaluated at the dipole position. Physically, this term contains two contributions: one due to the typical dipole radiation and a second due to the scattered radiation field. In particular, this term can be directly related to the LDOS:

$$\rho(\mathbf{r}_0) = \frac{6\omega_0}{\pi c^2} \hat{n} \cdot \text{Im}\{G(\mathbf{r}_0, \mathbf{r}_0)\} \cdot \hat{n}$$

Note that changes in the optical environment also lead to frequency shifts of the emitted radiation, similar to the Lamb shift, but these tend to be very small.

Molecules in Plasmonic Resonators

Plasmonic resonators are particularly interesting for enhancing the fluorescence or phosphorescence rates of molecules given the large field intensities that can be obtained within a small mode volume. Although mode volume is implicitly included in the definition of the LDOS, its role can be explicitly highlighted using Purcell's formula. In plasmonic resonators, several unique features must be considered that are usually absent in dielectric structures. Firstly, the excitation of modes with field components in the metal leads to losses due to Joule heating. These non-radiative losses, which occur with a rate γ_{nr} , can be significant and should not be confused with the intrinsic non-radiative (e.g., vibrational) loss rate of the molecule, $\gamma_{\text{nr}}^{\text{mol}}$. Rather, part of the radiation emitted by the dipole will be quenched by the metal. As a result, emitters coupled to plasmonic cavities will always suffer from (often significantly) less than unity quantum efficiencies. Secondly, the open nature of these systems can lead to significant enhancement of the excitation fields, which also lead to enhanced fluorescence. Finally, as for Fabry-Pérot optical cavities, plasmonic resonators intrinsically change the emitter radiation pattern and this change can significantly modify the measured fluorescence enhancement, which will depend on the experimental configuration. The interplay between fluorescence enhancement and quenching was highlighted in an experiment by Anger *et al.* where single molecule fluorescence was studied as a function of distance from a spherical gold nanoparticle in a geometry where this distance could be actively tuned. The results are shown in Fig. 1, where it can be seen that the fluorescence rate, which is a function of both the increased excitation field and the modified quantum efficiency reaches a maximum enhancement of ~ 7 at a distance 5 nm from the metal surface, beyond which it drops rapidly due to increased quenching.

The enhancements in fluorescence rate can be dramatic for molecules with a low intrinsic quantum yield. In the presence of a resonator, the modified quantum efficiency is $\eta = \gamma_r / (\gamma_r + \gamma_{\text{nr}} + \gamma_{\text{nr}}^{\text{mol}}) \simeq \gamma_r / \gamma$, where $\gamma_r \equiv \gamma - \gamma_{\text{nr}}$. The approximation holds whenever the Purcell factor is large enough to dominate any internal loss. The Purcell factor is defined as $F_p = \gamma / \gamma_0$, where γ_0 is the radiative rate without the presence of the antenna. By using a dye with $\eta = 2.5\%$ within the gap of a bowtie antenna, the Moerner group were able to demonstrate a thousand-fold fluorescence rate enhancement. In particular the structure reduced the fluorescence lifetime to below 10 ps, as compared to the molecule's intrinsic lifetime of 275 ps. In this work, η was increased to a maximum of 25% and limited by the fraction of emission leading to Joule heating.

One of the most interesting proposals for achieving large Purcell factors, while maintaining high quantum yield is the use of optical patch antennas proposed by Esteban *et al.* These can maintain quantum efficiencies of nearly 50%, combined with Purcell

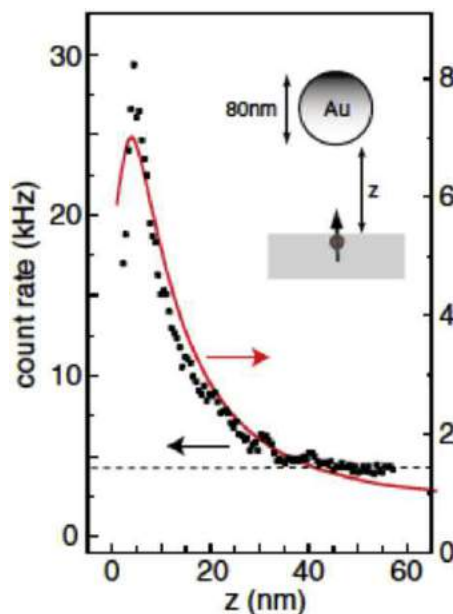


Fig. 1 Fluorescence count rate as a function of distance between an 80 nm-diameter gold nanoparticle and a single molecule of Nile Blue in a 2-nm-thick PMMA layer. The increase in local density of optical states (LDOS) near the nanoparticle leads to an increased count rate as it approaches the molecule. Very close to the nanoparticle, however, radiation occurs mostly in non-radiative modes, which leads to a quenching of the molecular emission. Reproduced from Anger, P., Bharadwaj, P., Novotny, L., *et al.*, 2006. Enhancement and quenching of single-molecule fluorescence. *Physical Review Letters* 96, 113002.

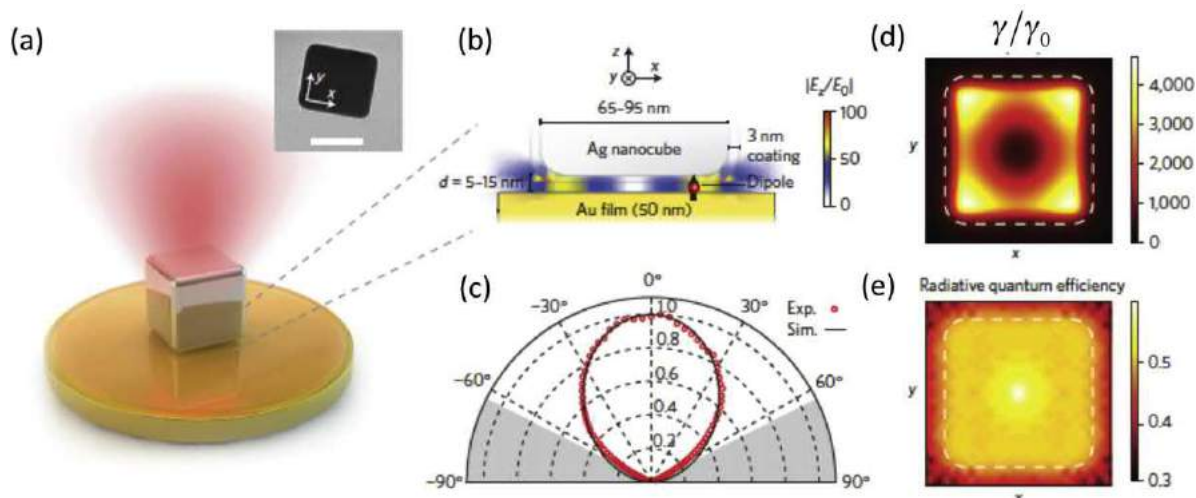


Fig. 2 (a) Schematic of the nanocube patch antenna from Akselrod *et al.* (the inset shows a transmission electron microscope image of a nanocube). (b) The electric field profile in the z -direction is shown for the fundamental plasmonic gap mode. (c) The experimental and simulated radiation profiles are shown. These show a maximum at normal incidence from the substrate, which is useful for out-coupling. (d) The calculated Purcell factor as a function of position for a vertically oriented dipole, which shows rate enhancements as large as 4000. The Purcell factor is closely related to field profile, which is peaked in the corners of the cube. (e) The radiative quantum efficiency for a vertically oriented dipole as a function of position. The quantum efficiency is remarkably high, staying above 50% for nearly all dipole positions. Reproduced from Akselrod, G.M., Argyropoulos, C., Hoang, T.B., *et al.*, 2014. Probing the mechanisms of large Purcell enhancement in plasmonic nanoantennas. *Nature Photonics* 8, 835.

factors of approaching a thousand. The structures consist of a small metaling disk (the patch) placed tens of nanometers above a much larger disk (the ground plane). The behavior of these structures can be understood by considering the modal structure of the infinite metal-insulator-metal waveguide. Such a waveguide supports a strongly confined mode in the gap, which does not exhibit a cutoff as a function of gap thickness. As a result, relatively large Purcell effects are possible, but this mode does not radiate. The wavelength scale patch serves to create discrete resonances as a function of patch width. These serve to further increases the Purcell factor, while allowing for efficient radiation out of the structure. A variant of this structure, consisting of a metallic nanocube (the patch) above an infinite metal plane was used by Akselrod *et al.* to demonstrate $F_p > 2000$ for vertically oriented dipoles, while maintaining a high quantum yield and directionality. The calculated field profile, Purcell factors and radiative quantum efficiencies of these antennas are shown in Fig. 2.

Dramatic changes in radiative decay rates can have important consequences for the photophysical processes within a molecule. Enderlein first considered the possible role of the Purcell effect on the photobleaching of organic molecules. Namely if photobleaching occurs from the excited state with a given rate, efficient radiation will reduce the time spent in the excited state and thus reduce the probability of a photobleaching. This can then dramatically increase the number of photons emitted before photobleaching. For many molecules, the photophysics are slightly different from those initially proposed by Enderlein, but the outcome is the same. In practice, photobleaching often occurs from the triplet state following intersystem crossing from the excited singlet state. For large Purcell factors, the triplet yield after optical excitation can be dramatically decreased due to the much faster radiative pathway, which can now compete effectively with intersystem crossing.

Cavity Effects in OLEDs

The typical OLED architecture consists of several organic layers deposited on top of an indium tin oxide (ITO) anode, which sits atop a glass substrate. Organic layer thicknesses tend to range between 100 and 150 nm and the structures are typically capped with an aluminum cathode. It is not obvious that cavity effects need to be carefully accounted for in such an innocuous structure. There is no obvious cavity and the emissive layers are typically located 50 nm away from the metal cathode. Saito *et al.* used the approach of Chance *et al.* to calculate the decay rate of an emitter in proximity to the cathode and found that significant reductions in lifetime could be observed for distances up to 150 nm away. This reduction is principally due to the excitation of surface plasmon-polaritons at the metal-organic interface by the dipole's near field components. Bulovic *et al.* highlighted the importance of dipole orientation on the angular emission profile and polarization behavior of OLEDs. In particular, they found that the presence of the metal cathode is sufficient to lead to weak microcavity effects and as a result, small changes in organic layer thicknesses are sufficient to shift the observed emission wavelength by ~ 15 nm. They also confirmed experimentally the polarization splitting in the emission, which occurs at high angle and the predominance of TM-polarized radiation.

Hobson *et al.* presented a thorough analysis of light extraction in an OLED geometry, by using the method of Chance, Prock and Silbey. They calculated the power dissipated via the different possible pathways: radiation into air, into the glass substrate,

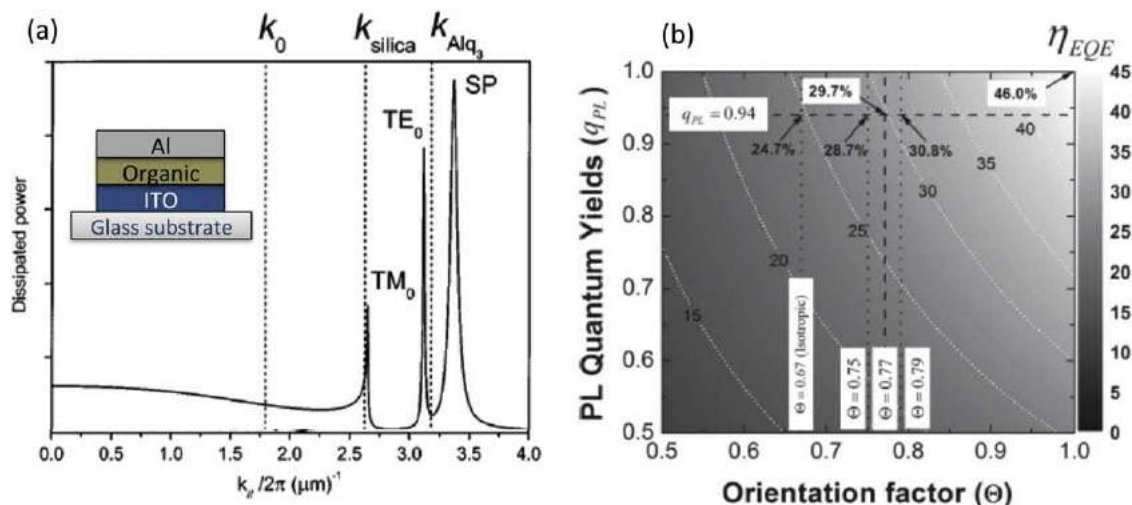


Fig. 3 (a) Power dissipation spectrum as a function of in-plane momentum $k_{\parallel}$ for a 100-nm-thick model small-molecule organic light-emitting diode (OLED) with randomly oriented dipoles. The emitters are located 50 nm away from the cathode. The critical wavevectors for air, silica and the organic Alq_3 are indicated above the figure. A schematic OLED structure is shown in the inset. The fraction of power dissipated into a given mode is given integrating under each peak and the power waveguided in the substrate corresponds to that between k_0 and k_{silica} . Reproduced from Hobson, P.A., Wasey, J.A.E., Sage, I., *et al.*, 2002. The role of surface plasmons in organic light-emitting diodes. *IEEE Journal of Selected Topics in Quantum Electronics* 8 (2), 378. (b) Contour plot of the calculated external quantum efficiency (EQE) for a realistic OLED as a function of the dipole orientation factor and intrinsic quantum yield of the emitter. The dashed lines correspond measured values for $\text{Ir}(\text{ppy})_2(\text{acac})$, which has an intrinsic quantum yield of 0.94 and a horizontal to vertical dipole ratio of 0.77:0.23 ($\Theta = 0.77$). ITO, indium tin oxide. Reproduced from Kim, S.-Y., Jeong, W.-I., Mayr, C., *et al.*, 2013. Organic light-emitting diodes with 30% external quantum efficiency based on a horizontally oriented emitter. *Advanced Functional Materials* 23, 3896.

into waveguided modes or into surface plasmon modes as shown in Fig. 3(a) for the case of a model 100-nm-thick small-molecule OLED. Importantly, they correctly found that near-field coupling to surface plasmons are a considerable source of loss even for emitters located more than 50 nm away from the cathode. For polymer-based systems, coupling to the surface plasmon mode is typically weaker than for isotropically oriented small molecules, because the transition dipole moments are predominantly in-plane and cannot radiate efficiently in the TM-polarized surface plasmon mode.

A large number of studies have since focused on strategies to out-couple waveguided light from the substrate and active layers, such as using microlens arrays, corrugated structures, scattering films or low-index grids. Recently, however, the importance of molecular orientation has been reemphasized. In particular, the Kim and Brütting groups have studied the importance of molecular orientation on out-coupling efficiency. Fig. 3(b) shows the maximum theoretical external quantum efficiency (EQE) of a state-of-the-art green phosphorescent OLED as a function of dipole orientation factor and PL quantum efficiency of the emitter. One can see that the maximum out-coupling efficiency, and concomitant EQE for an ideal isotropic emitter tends to be around $\sim 25\%$. However, this value can be increased to 46% for horizontally oriented dipoles. As a result, several small molecule guest–host combinations have since been identified that can lead to external quantum efficiencies surpassing 30% due to a natural tendency for their transition dipole moments to align in-plane.

Finally, the ability to spectrally control the emission in microcavity OLEDs where the ITO has been replaced or complemented by a thin metal anode has further advantages. Such microcavities have been used to increase the emitted light in the forward direction and more importantly, to tune the emission wavelength or increase spectral purity. For example, Sony’s “Super Top Emission” architecture introduced this strategy in combination with a single white OLED design to produce full color displays.

Strong Light–Matter Coupling

The results of the previous sections, which were based on Fermi’s golden rule, were obtained by treating light-matter interaction – in those cases, within the electric dipole approximation – using perturbation theory. This approach accurately predicts the absorption and emission dynamics of molecules at short times compared to the Rabi frequency of the combined light-matter system. In practice, this approach tends to give the correct results because of strong dissipation or short coherence times that damp any long-time oscillatory behavior (Rabi oscillations). In optical cavities where only a few discrete modes interact with the molecules, this approach can fail. Interaction of the molecule with the vacuum electromagnetic field occurs at a rate given by the so-called vacuum Rabi frequency, Ω . In a uniformly filled cavity with ideal mirrors:

$$\hbar\Omega = 2\sqrt{\frac{Ne^2f}{4m\epsilon_0\epsilon_b}}$$

where N is the molecular number density, f is the oscillator strength, and ε_b is the background dielectric constant. There is no explicit dependence on mode volume because of the use of molecular density. The corresponding Rabi energy $\hbar\Omega$ can be as high as ~ 1 eV, which easily exceeds the typical photon and exciton linewidths (or equivalently dissipation rates) in organic microcavities. As a result, it is relatively simple to reach the strong light–matter coupling regime. The explicit factor of 2 emphasizes the fact that this definition corresponds to the energy splitting of new light–matter eigenmodes of the system: the lower and upper exciton-polaritons (LP and UP). For each in-plane wavevector, LP and UPs are formed, which are coherent superpositions of the molecular excitons (that with the correct wavevector) and the corresponding cavity photon mode. Their dispersion relation, which is shown in Fig. 4(a), is inherited mostly from the typical Fabry–Perot dispersion and shows typical anticrossing behavior around the bare exciton resonance. For planar structures, this dispersion can be readily measured as a function of angle given the correspondence between angle of incidence and in-plane wavevector. This contrasts strongly with bulk exciton-polaritons, which can also exist in molecular crystals at low-temperature. Lidzey *et al.* were the first to observe the formation of polaritons in an organic microcavity at room-temperature using a porphyrin dye in a host matrix.

Because of the large number of molecules as compared to the number of photon modes supported in such structures, there exists several eigenstates of the molecule-photon Hamiltonian that are not coupled to the photon mode: the so-called dark states. These play an important role in dictating the dynamics under nonresonant excitation. Polaritonic modes can also be modified by the presence of structural and energetic disorder, effects that were considered analytically by Agranovich *et al.* and numerically by Michetti and La Rocca. The overarching conclusion is that in the photonic parts of the polariton branches, disorder does not cause a strong deviation from the ideal polariton mode, but very close to $k=0$ and in the flat (exciton-like) regions of the polariton branches, disorder causes a strong spatial localization of the modes.

As for inorganic polaritons, a strong motivation for studying organic microcavities in the strong-coupling regime has been to realize low-threshold sources of coherent light: polariton lasers. Indeed, if a large population of polaritons can be accumulated near the energy minimum at $k=0$, Bose stimulation will favor further population of this state. When scattering into this state exceeds the loss rate, a polariton condensate will form. At this point, the state becomes macroscopically occupied and spontaneously acquires a phase, which corresponds to a breaking of U(1) symmetry. Because photons emitted through the cavity mirrors preserve both the phase and momentum of the underlying polaritons, this results in both spatially and temporally coherent emission. Fig. 4(b) shows an angle-resolved contour plot of the luminescence emitted from a strongly-coupled organic microcavity pumped below and above the condensation threshold. Above threshold, the luminescence is emitted from the bottom of the LP and a small blueshift can be observed due to polariton nonlinearities.

The detailed mechanisms for relaxation from a high-energy molecular excited state created by a laser pump to the minimum of the LP are still under investigation. Measurements show that this relaxation occurs principally through the dark states. Although the decay of polaritons occurs on a rapid timescale given by the polariton lifetime (typically 0.01–1 ps), time-resolved photoluminescence measurements show a decay on the timescale of the reservoir lifetime. This strongly suggests that these are the initially excited states from which polaritons must then be populated. Various mechanisms have been proposed to explain relaxation from

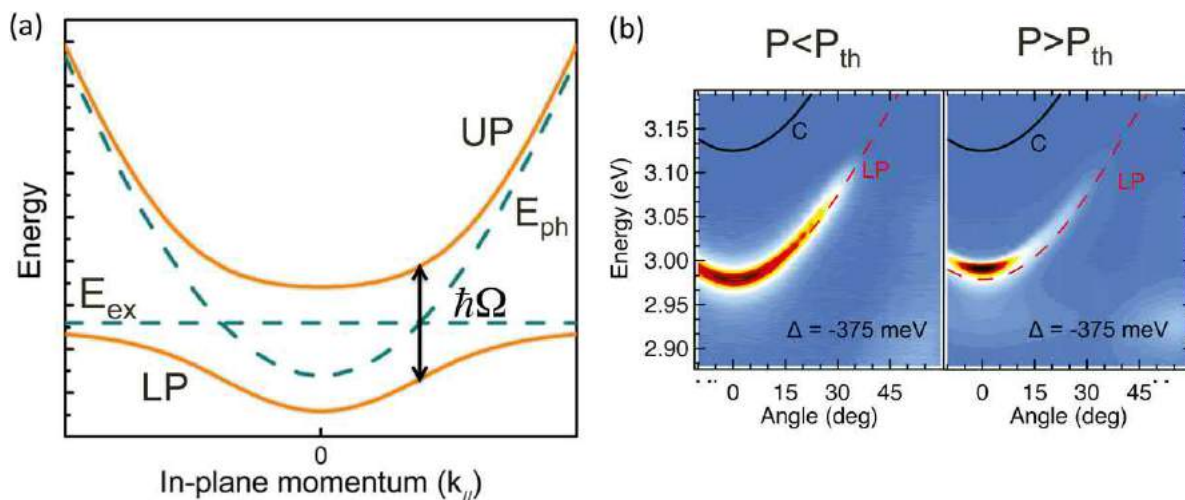


Fig. 4 (a) The dispersion relation of upper and lower polaritons (UP and LP) is shown as solid lines. The dispersion relations of the bare exciton (E_{ex}) and cavity photon (E_{ph}) are shown as dashed lines. The polariton branches show typical anticrossing behavior around the exciton energy. The minimum splitting occurs at the degeneracy point where both branches are split by the Rabi energy. (b) Contour plot of the angle-resolved photoluminescence from a strongly-coupled microcavity pumped below and above the condensation threshold. Below threshold the emission intensity decreases monotonically when going up from the minimum of the LP branch. Above threshold, all of the emission comes from the bottom of the LP and a small blueshift can be observed. The color scales are normalized. The measured irradiance above threshold is approximately 10^4 times larger than that below threshold. Modified from Daskalakis, K.S., Maier, S.A., Murray, R., *et al.*, 2014. Nonlinear interactions in an organic polariton condensate. *Nature Materials* 13, 271.

the dark states to polaritonic states including the emission of localized molecular vibrations, pure dephasing mechanisms due to low-energy vibrations and radiative emission through uncoupled vibrational modes of the ground electronic state.

The realization of cavities with low-enough dissipation rates have allowed for polariton condensates to be demonstrated in microcavities containing organic single-crystals (Kéna-Cohen *et al.*), disordered organic and polymer films (Daskalakis *et al.* and Plümhof *et al.*) and macromolecules (Dietrich *et al.*). One particularly interesting feature of microcavities in the strong coupling regime is the nonlinearity due which arises due to a depletion of the ground state. This nonlinearity gives us rise to effective repulsive interactions between polaritons and between polaritons and dark excitons. The first has been used to demonstrate room-temperature superfluidity of polaritons, while the second gives rise to modulation instabilities and can be used to create artificial potential barriers using a pump laser. Until now, the nonlinearities observed in organic microcavities remain much weaker than those in inorganics and thus require much larger densities of polaritons to show dramatic effects.

Another fascinating possibility is that of using organic microcavities in the strong coupling regime to modify molecular properties or chemical reaction rates. For example, the Ebbesen group has shown a significantly reduced rate of photoisomerization for a photochromic dye in a strongly coupled microcavity compared to the out-of-cavity case. Indeed, theoretical work by Galego *et al.* has shown how the modification of potential energy surfaces in the strongly coupled regime can indeed suppress photochemical reactions due to a stabilization of the molecular structure. More recently, strong coupling of optically active vibrational modes within planar optical cavities has also been demonstrated. This approach is particularly promising for the future control of chemical reaction pathways as it could allow for bond-specific targeting.

Conclusions

This article reviewed selected applications of cavity QED to solid-state molecular systems. We saw how in the weak coupling regime, optical resonators allow for the control of the spontaneous emission rate and radiation pattern. These effects have important applications for the realization of fluorescent markers, OLEDs and single photon sources. In the strong coupling regime, hybrid light-matter eigenstates-the polaritons-form with their own unique properties. These possess strong nonlinearities and have applications for the realization of low-threshold organic lasers and the control molecular properties and chemical reactions. We should also note that optical cavities are also routinely used in other applications which have not been covered here such as molecular spectroscopy, conventional organic lasers and photon Bose-Einstein condensation with liquid dye solutions. However, in these applications, QED effects play a lesser role.

See also: Organic Lasers

Further Reading

- Agranovich, V.M., Litniskaia, M., Lidzey, D.G., 2003. Cavity polaritons in microcavities containing disordered organic semiconductors. *Physics Reviews B* 67, 085311.
- Bharadwaj, P., Deutsch, B., Novotny, L., 2009. Optical antennas. *Advances in Optics and Photonics* 1, 438–483.
- Brutting, W., Frischeisen, J., Schmidt, T.D., Scholz, B.J., Mayr, C., 2012. Device efficiency of organic light-emitting diodes: Progress by improved light outcoupling. *Physica Status Solidi: A* 210, 44–65.
- Bulović, V., Khalfin, V.B., Gu, G., *et al.*, 1998. Weak microcavity effects in organic light-emitting devices. *Physics Review B* 58, 3730–3740.
- Carusotto, I., Ciuti, C., 2013. Quantum fluids of light. *Reviews of Modern Physics* 85, 299–366.
- Chance, R.R., Prock, A., Silbey, R., 1978. Molecular fluorescence and energy transfer near interfaces. In: Prigogine, I., Rice, S.A. (Eds.), *Advances in Chemical Physics*, vol. 37. Hoboken, NJ: John Wiley & Sons, Inc., pp. 1–65.
- Daskalakis, K.S., Maier, S.A., Kéna-Cohen, S., 2016. Polariton condensation in organic semiconductors. In: Bozhevolnyi, S.I., Martín-Moreno, L., García-Vidal, F. (Eds.), *Quantum Plasmonics*. Berlin Heidelberg: Springer, pp. 151–163. Volume 185 of the Springer Series in Solid-State Sciences.
- Enderlein, J., 2002. Theoretical study of single molecule fluorescence in a metallic nanocavity. *Applied Physics Letters* 80, 315–317.
- Esteban, R., Teperik, T.V., Greffet, J.J., 2010. Optical patch antennas for single photon emission using surface plasmon resonances. *Physical Review Letters* 104, 026802.
- García-Vidal, G.J.F., Feist, J., 2015. Cavity-induced modifications of molecular structure in the strong coupling regime. *Physics Reviews X* 5, 041022.
- Hobson, P.A., Wedge, S., Wasey, J.A.E., Sage, I., Barnes, W.L., 2002. Surface plasmon mediated emission from organic light-emitting diodes. *Advance Materials* 14, 1393–1396.
- Hutchison, J.A., Schwartz, T., Genet, C., Devaux, E., Ebbesen, T.W., 2012. Modifying chemical landscapes by coupling to vacuum fields. *Angewandte Chemie International Edition* 51, 1592–1596.
- Kéna-Cohen, S., Forrest, S.R., 2012. Exciton-polaritons in organic semiconductor microcavities. In: Sanvitto, D., Timofeev, V. (Eds.), *Exciton Polaritons in Microcavities*. Berlin Heidelberg: Springer, pp. 349–375. Volume 172 of the Springer Series in Solid-State Sciences.
- Novotny, L., Hecht, B., 2012. *Principles of Nano Optics*. Cambridge: Cambridge University Press.
- Saito, S., Tsutsui, T., Era, M., *et al.*, 1993. Progress in organic multilayer electroluminescent devices. *Proceedings of SPIE* 1910, 212–221.

Modification of the Spontaneous Transition Rate in Confined Space

In order to understand the modification of the spontaneous emission rate in an external cavity-like structure, it must be remembered that in quantum electrodynamics this rate is proportional to the density of modes of the electromagnetic field, i.e., the vacuum field fluctuations above the atomic transition frequency ω_0 . As a consequence the spontaneous emission rate is increased if the atom is surrounded by a cavity tuned to the transition frequency. Conversely, the decay rate decreases when the cavity is mistuned. This fact was already recognized in the pioneering days of nuclear magnetic resonance by Purcell.

The spontaneous decay rate of the atom in the cavity, γ_c , will then be enhanced in relation to that in free space, γ_f , by a factor given by the ratio of the corresponding mode densities

$$\frac{\gamma_c}{\gamma_f} = \frac{\rho_c(\omega_0)}{\rho_f(\omega_0)} = \frac{2\pi Q}{V_c \omega_0^3} = \frac{Q \lambda_0^3}{4\pi^2 V_c}$$

where V_c is the volume of the cavity and Q is the quality factor of the cavity which expresses the 'sharpness' of the mode. For low-order cavities $V_c \approx \lambda_0^3$ this means that the spontaneous emission rate is increased by roughly a factor of Q . As mentioned already, when the cavity is detuned, the decay rate will decrease. In this case the atom cannot emit a photon, since the cavity is not able to accept it, and the energy will stay with the atom, i.e., its decay is inhibited.

To change the decay rate of an atom, in principle no resonator has to be present; any conducting surface near the radiator affects the mode density and, therefore, the spontaneous radiation rate. Parallel conducting planes can alter the emission rate somewhat but can only reduce the rate by a factor of 2 owing to the existence of modes, independent of the plate separation with an electric field perpendicular to the planes.

In order to demonstrate experimentally the modification of the spontaneous decay rate, it is not necessary to go to single-atom densities. The experiments where the spontaneous emission is inhibited can also be performed with large atom numbers. However, in the opposite case, when the increase of the spontaneous rate is observed, a large number of excited atoms may disturb the experiment by induced transitions. The first experimental work on inhibited spontaneous emission was done by Drexhage, Kuhn and Schäfer in 1974. The fluorescence of a thin dye film near a mirror was investigated. A reduction of the fluorescence decay by up to 25% was observed. Later experiments with microcavities filled with dye solutions were performed by de Martini *et al.* These experiments demonstrated that thresholdless lasing is possible in such systems.

Inhibited spontaneous emission was observed by Gabrielse and Dehmelt. In these neat experiments with a single electron stored in a Penning trap they observed that cyclotron orbits show lifetimes which are up to 10 times longer than that calculated for free space. The electrodes of the trap form a cavity which decouples the cyclotron motion from the vacuum field leading to the longer lifetime.

A new stage in experiments on cavity QED was reached when it was recognized that highly excited alkaline atoms are very suitable for these experiments. The main quantum numbers are in the range $n=20-60$. Those so-called Rydberg atoms couple very strongly to the radiation field as will be discussed later. The transitions to neighboring states are in the microwave region. The cavities can therefore be built as low-order cavities with dimensions on the order of the wavelength of the transitions being in the mm or cm regions. The spontaneous rate of these transitions in free space is small owing to the low value of their transition frequency, therefore an enhancement is possible. Experiments with Rydberg atoms on the inhibition of spontaneous emission have been performed by Kleppner and coworkers and by Haroche and coworkers. In the latter experiment a 3.4 μm transition of the Cs atom was suppressed.

The first observation of enhanced atomic spontaneous emission in a resonant cavity was published by Haroche *et al.* This experiment was performed with Rydberg atoms of Na excited in the 23s state in a niobium superconducting cavity resonant at 340 GHz. Cavity tuning-dependent shortening of the lifetime was observed. The cooling of the cavity had the advantage of totally suppressing the black-body field. The latter effect is completely absent if optical transitions are observed, however, in this case it is more difficult to obtain low-order cavities. The first experiments on optical transitions were performed by Feld and collaborators. They succeeded in demonstrating the enhancement of spontaneous transitions even in higher-order optical cavities.

In modern semiconductor devices both electronic and optical properties can be tailored with a high degree of precision. Therefore, electron-hole systems producing recombination radiation analogous to radiating atoms can be localized in cavity-like structures, e.g., in quantum wells. Thus optical microcavities of half or full wavelength size are obtained. Both suppression and enhancement of spontaneous emission in semiconductor microcavities were demonstrated in experiments by Yamamoto and collaborators. Yablonoivitch *et al.* proposed the use of photonic band gap structures in semiconductors. In those systems spontaneous emission is forbidden in certain frequency regions owing to the special material properties.

Energy Shift in Confined Space

In the previous section we focused on changes of radiation rates of atoms near conducting walls or in cavity-like structures. Next we will discuss the more subtle phenomenon of energy shifts. While radiation rates are modified by the influence of the vacuum field on the real emission of photons the energy shifts are caused by the modification of a virtual photon interaction.

Resonant and nonresonant phenomena have to be distinguished. The resonant self-energy shift of a decaying atomic dipole in the vicinity of a conducting wall can be determined from the average polarization energy produced by the image dipole field. For distances comparable to the wavelength the near-field condition is satisfied, resulting in the z^{-3} dependence of the static dipole–dipole interaction characteristic for the van der Waals energy; under far-field conditions the distance dependence is given by z^{-1} . The polarization of the atom by the nonresonant parts of the broadband electromagnetic field causes energy shifts, of which the Lamb shift is the most prominent. In the sense of the nonrelativistic treatment by Bethe the major contribution of that shift can be described as being a result of the emission and reabsorption of virtual photons. It is plausible that just as the real emission of a photon is modified in confined space, so also is the virtual process. The latter ‘real’ radiation energy shift is thus a consequence of vacuum fluctuations only. It is identical with the energy shift predicted by Casimir and Polder and is analogous to the better-known result of Casimir on the force between two plane neutral conducting plates.

The question of modification of atomic energies in confined space has recently found considerable interest and many calculations of the phenomenon have been performed. Direct application to the energy shift of Rydberg atoms, which are of special interest for experimental studies, was performed by Barton in 1987. He calculated the direct electrostatic interaction with a conducting wall and the radiation induced (retarded) effects. The result is that in the case of two parallel plates the electrostatic effect is dominant when the distance L between the conducting plates is small, $L < n^3 a_0 / \alpha$ (n is the principal quantum number of the Rydberg atom, a_0 the Bohr radius, and α the fine-structure constant), and the radiative effect plays the major role when large distances are used, $L > n^3 a_0 \alpha$.

Experiments to measure the van der Waals force of conducting planes have been performed by Hinds and coworkers. They could clearly demonstrate the z^{-3} dependence of the van der Waals force. The retarded effects were not detectable at large distances but at small distances from the wall a deviation from the z^{-3} dependence was found which was attributed to the retarded QED potential.

The energy shift of rubidium Rydberg atoms in confined space has been measured by Marrocco *et al.* in an experiment with extreme high-resolution using the Ramsey experiment two-field method. The atoms are excited in s-Rydberg states ($n \approx 30$) by two-photon transitions using the light of an ultrastable dye laser, the linewidth of which was less than 10 Hz. The laser intensity was enhanced in a folded cavity locked to the laser frequency. Both interaction zones necessary for the Ramsey method were enclosed in this cavity. Between the two interaction regions the atoms pass through a pair of conducting plates, the distance between which can be changed. The shift of the Ramsey interference was measured as a function of the plate distance. Using the Ramsey method has the advantage that the shift can be determined without a direct probing of the atoms in the space between the plates. A level shift on the order of about 150 Hz could be measured, in very good agreement with theory.

Maser Operation

If the rate of atoms crossing a cavity exceeds the cavity damping rate ω/Q the photon released by each atom is stored long enough to interact with the next atom. Here ω stands for the cavity frequency and Q for the quality factor of the cavity. The atom–field coupling becomes stronger and stronger as the field builds up and evolves to a steady state. Using Rydberg atoms with a large field–atom coupling constant leads to a new kind of maser which operates with exceedingly small numbers of atoms and photons. The photons corresponding to transitions between neighboring Rydberg atoms are in the microwave region at about 20–100 GHz. Atomic fluxes as small as a few atoms per second have generated maser action, as could be demonstrated by Walther *et al.* of the University of Munich in 1985 using a superconducting cavity. For such a low flux there is never more than a single atom in the resonator – in fact, most of the time the cavity is empty of atoms. It should also be mentioned that in the case of such a setup a single resonant photon is sufficient to saturate the maser transition.

The scheme for this one-atom maser or micromaser is shown in Fig. 1. The setup represents the simplest system in radiation–atom interaction: a single atom is interacting with a single mode of the radiation field. This device was used to verify the complex dynamics of a single atom in a quantized field predicted by the Jaynes–Cummings model. All of the features are explicitly a consequence of the quantum nature of the electromagnetic field: the statistical and discrete nature of the photon field leads to new characteristic dynamics such as collapse and revivals in the exchange of a photon between the atom and the cavity mode. The frequency of this exchange is usually called the Rabi flopping frequency. The field in the cavity is measured through the number of atoms in the lower state of the maser transition. Due to the strong coupling between the atom and maser field both are entangled. The strong coupling can also be used to entangle subsequent atoms.

The steady-state field of the micromaser shows sub-Poisson statistics. This is in contrast to regular masers and lasers where coherent fields (Poisson statistics) are observed. The reason for nonclassical radiation being produced is that a fixed interaction time of the atoms is chosen, leading to careful control of the atom–field interaction dynamics.

Under steady-state conditions, the photon statistics $P(n)$ of the field of the micromaser are essentially determined by the pump parameter $\theta = N_{\text{ex}}^{1/2} \Omega t_{\text{int}}$, denoting the angular rotation of the pseudospin vector of the interacting atom. Here, N_{ex} is the average number of atoms that enter the cavity during the decay time of the cavity τ_{cav} , Ω the vacuum Rabi flopping frequency, and t_{int} is the

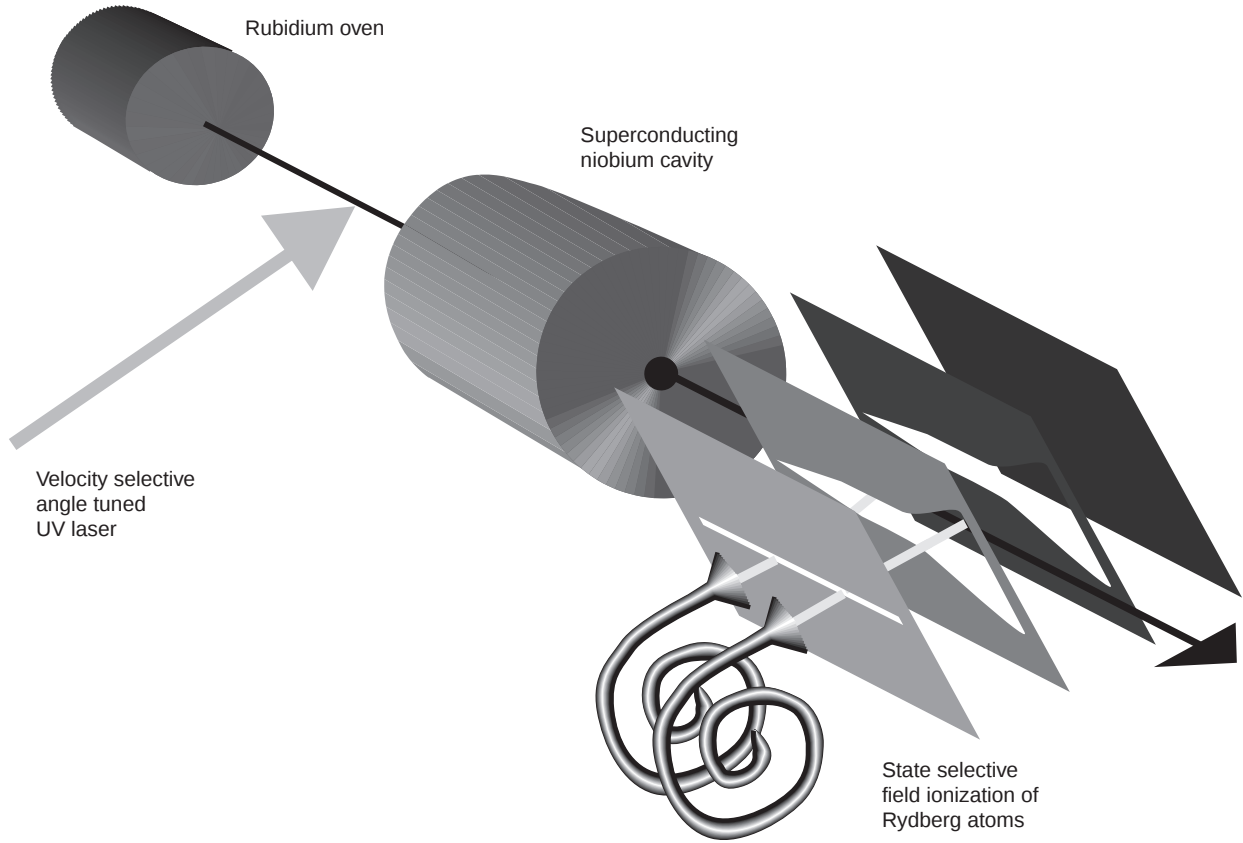


Fig. 1 Micromaser setup of rubidium Rydberg atoms in the 63p state. The velocity of the atoms is controlled by exciting a velocity subgroup of atoms in the atomic beam. The atoms in the upper and lower maser levels are selectively detected by field ionization. The number of photons deposited in the cavity is determined through the number of atoms in the lower state (61d).

atom–cavity interaction time. The normalized photon number of the maser $\langle v \rangle = \langle n \rangle / N_{\text{ex}}$ shows the following generic behavior (see Fig. 2). It suddenly increases at the maser threshold value $\theta = 1$ and reaches a maximum for $\theta = 2$ (denoted by 1 in Fig. 2). The behavior at threshold corresponds to the characteristics of a continuous phase transition. As θ increases further, the normalized averaged photon number $\langle v \rangle / N_{\text{ex}}$ decreases to a minimum slightly below $\theta = 2\pi$ and then abruptly increases to a second maximum (3a in Fig. 2). This general type of behavior recurs roughly at integer multiples of 2π but becomes less pronounced with increasing θ . The reason for the periodic maxima of the average photon number is that for integer multiples of $\theta = 2\pi$ the pump atoms perform an integer number of full Rabi flopping cycles, and start to flip over at a slightly larger value of θ , thus leading to enhanced photon emission. The periodic maxima for $\theta = 2\pi, 4\pi$, and so on may be interpreted as first-order phase transitions.

The photon statistics of the maser radiation is usually characterized by the Q -parameter introduced by Mandel:

$$Q_{\text{field}} = \{ (\langle n^2 \rangle - \langle n \rangle^2) / \langle n \rangle \} - 1$$

It can be seen that $Q_{\text{field}} = 0$ corresponds to a Poissonian photon distribution. Q_{field} for the micromaser is plotted as a dotted line in Fig. 2. The value drops below zero in the region 2a, 2b, etc. This shows the highly sub-Poissonian character of the one-atom-maser field being present over a large range of parameters.

The reason for the sub-Poissonian atomic statistics is the following. A changing flux of atoms changes the Rabi frequency via the stored photon number in the cavity. Adjusting the interaction time allows the phase of the Rabi nutation cycle to be chosen such that the probability of the atoms leaving the cavity in the upper maser level increases when the flux, and hence the photon number in the cavity, is raised. This leads to sub-Poissonian atomic statistics since the number of atoms in the lower state decreases with increasing flux and photon number in the cavity. This feedback mechanism can be neatly demonstrated when the anticorrelation of atoms leaving the cavity in the lower state is investigated. Measurements of this anticorrelation phenomenon could be made.

The fact that anticorrelation is observed shows that the atoms in the lower state are more equally spaced than expected for a Poissonian distribution of the atoms in the beam. This means, for example, that when two atoms enter the cavity close to each other, the second one performs a transition to the lower state with reduced probability.

The interaction with the cavity field thus leads to an atomic beam with atoms in the lower maser level showing number fluctuations which are up to 40% below those of a Poissonian distribution found in usual atomic beams. This is interesting

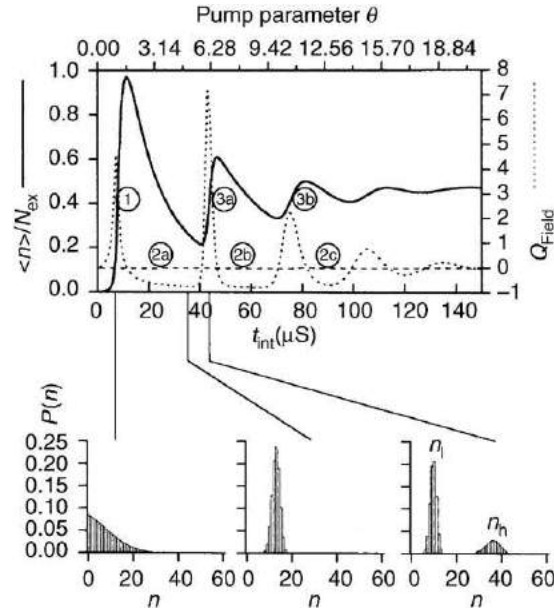


Fig. 2 One-atom maser or micromaser characteristics. The upper part of the figure shows the average photon number versus interaction time (solid curve) and the photon number fluctuations represented by the Q -factor (dotted line). Both curves are determined by the photon exchange dynamics between atom and field. The pump parameter gives the angular rotation of the pseudospin vector of the interacting atom (e.g., 2π corresponds to a full rotation, i.e., the atom is again in the upper level). In the lower part of the figure the steady-state photon number distribution $P(n)$ is shown for three values of t_{int} . The distribution on the left side corresponds to the maser threshold, that on the right gives an example of the double-peaked distribution associated with the quantum jump behavior. In this situation, the atom is back in the excited state and can again emit, leading to a higher steady-state photon number n_h . With increasing t_{int} this part will grow and n_l will decrease and disappear. A new jump occurs in the region 3b.

because atoms in the lower level have emitted a photon to compensate for cavity losses inevitably present. Although this process is induced by dissipation giving rise to fluctuations, the atoms still obey sub-Poissonian statistics.

Generation of Number States (Fock States) of the Radiation Field

When the micromaser is operated at low temperatures, a very interesting feature is observed in the inversion of the population of the two maser levels called trapping states. They are a steady-state feature of the maser field; they occur in the micromaser as a direct consequence of field quantization in a cavity. At low cavity temperatures the number of black-body photons, n_{th} , in the cavity mode is reduced and trapping states begin to appear; at higher temperatures they are washed out by the thermal photons. The trapping states show up when the atom–field coupling Ω and the interaction time t_{int} are chosen such that in a cavity field with n_q photons the atoms undergo an integer number k of Rabi cycles. This is summarized by the condition

$$\Omega t_{\text{int}} \sqrt{n_q + 1} = k\pi \quad (1)$$

for the trapping state, denoted by $(n_q k)$. When Eq. (1) is satisfied, the cavity photon number is left unchanged after interaction of an atom and hence the photon number is ‘trapped’. This will occur regardless of the atomic pump rate N_{ex} . The trapping state is therefore characterized by the upper-bound photon number n_q and the number of integer multiples of full Rabi cycles k . In this situation the field is stabilized. Whenever a photon disappears, e.g., due to dissipation, the next incoming atom experiences a changed Rabi nutation frequency and emits a photon with high probability. At the trapping condition a quantum nondemolition situation is present. Through the dynamics of the Rabi nutation the field is measured, without any net change of the field.

The build-up of the cavity field can be seen in Fig. 3, where the emerging atom inversion is plotted against the interaction time and pump rate. At low atomic pump rates (low N_{ex}) the maser field cannot build up and the maser exhibits Rabi oscillations due to the interaction with the vacuum field. At the positions of the trapping states, the field increases until it reaches the trapping-state condition. This is manifested as a reduced emission probability and hence as a dip in the atomic inversion. Once in a trapping state, the maser will remain there regardless of the pump rate.

Under ideal conditions the micromaser field in a trapping state is a Fock state, but when the micromaser is operated in a continuous wave mode, the field state is very fragile and highly sensitive to external influences and experimental parameters. However, in contrast to continuous-wave operation, in pulsed operation trapping states are more stable and more practical and can be used over a much broader parameter range than in continuous-wave operation.

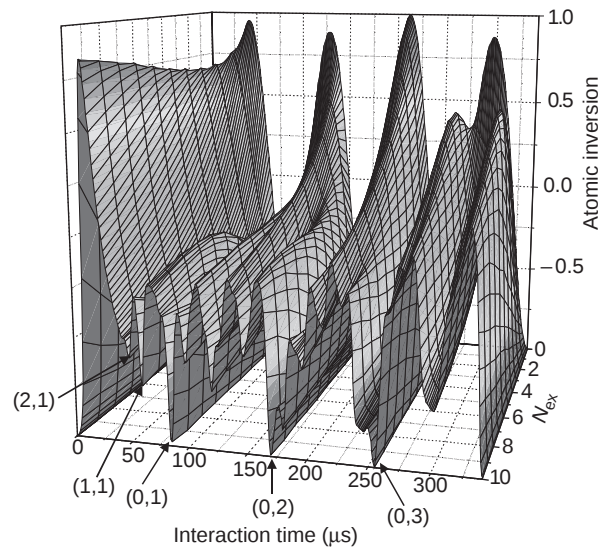


Fig. 3 Behavior of the micromaser at low temperature showing the trapping states, manifested as valleys along the N_{ex} axis. For the designation of the trapping states see text.

To demonstrate the principle, Fig. 4 shows a simulation of a sequence of 20 pulses of the pumping atoms in which an average of seven excited atoms per pulse are present. Under the trapping condition only a single emission event occurs, producing a single lower state atom which leaves a single photon in the cavity. Since the atom–cavity system is in the trapping condition, the emission probability is reduced to zero and the photon number is stabilized. Consequently, excited state atoms following the emitting atom stay in the upper maser level. The variation of the time when an emission event occurs during the atom pulses in Fig. 4 is due to the Poissonian spacing of upper-state atoms entering the cavity and the stochasticity of the quantum process.

The lower part of Fig. 4 shows the photon number distribution resulting from this process. It was demonstrated experimentally by Brattke *et al.* that a single photon number state could be generated with a success rate of 85%. By improving the experimental parameters one can expect to prepare single-photon Fock states in 98% of the pulses by this method.

Other Cavity Experiments

Besides the micromaser, there are also other experiments using Rydberg atoms in cavities and the strong interaction of these atoms with cavities which lead to interesting applications. One example is an experiment by S. Haroche, J. M. Raimond *et al.*, who succeeded in realizing a Schrödinger cat state in a cavity. Studies on the decoherence of this state could be conducted. The experiments are quite important in exploring the boundaries between the quantum and classical worlds. Another example is the quantum nondemolition detection of a single photon in a cavity by S. Haroche employing the dispersion level shift of probing atoms.

There is an interesting equivalence between an atom interacting with a single-mode field and a quantum particle in a harmonic potential, as was first pointed out by D. F. Walls and H. Risken; this connection results from the fact that the radiation field is quantized on the basis of the harmonic oscillator. The Jaynes–Cummings dynamics can therefore also be observed with trapped ions, as recently demonstrated in a series of beautiful experiments by D. Wineland *et al.* They also produced a Schrödinger cat state by preparing a single trapped ion in a superposition of two spatially separated wavepackets which are formed by coupling different vibrational quantum states in the excitation process.

Besides the experiments in the microwave region, a single-atom laser emitting in the visible range has also been realized by M. Feld *et al.* Furthermore, cavity quantum electrodynamic effects have been studied in the optical spectral region by J. L. Kimble *et al.*

Today's technology in microfabrication of semiconductor diode structures allows the realization of low-order cavity structures for diode lasers. In these systems the spontaneous emission is controlled in the same way as in the micromaser. Since spontaneous decay is a source of strong losses, control of this phenomenon leads to highly efficient laser systems. Quantum control of spontaneous decay thus has important consequences for technical applications. This topic will be briefly described in the next section.

Microlasers

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Perot resonator to λ/n (where n stands for the refractive index) while reducing the lateral dimensions to a range of the same order of magnitude. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis.

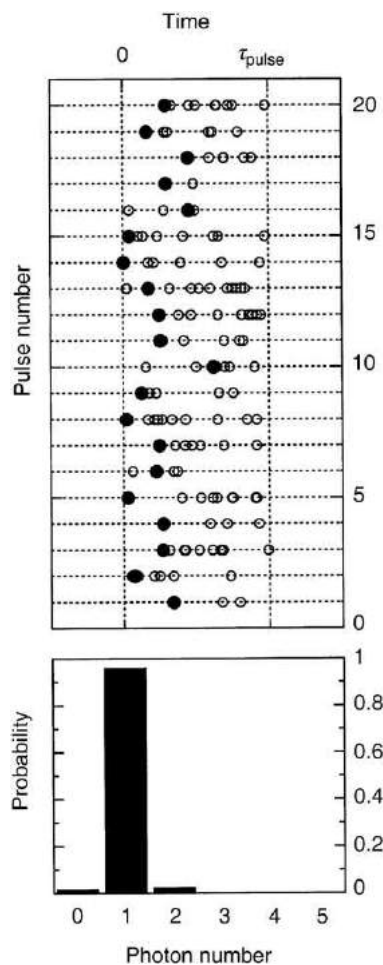


Fig. 4 A simulation of a subset of 20 subsequent atom bunches generated by pulsed laser excitation with the associated probability distribution for photons in the cavity or lower-state atom production (filled circles represent lower-state atoms and open circles represent excited-state atoms). The start and end of each pulse is indicated by the vertical dotted lines marked 0 and τ_{pulse} , respectively. For details, see Battke S, Varcoe BTH and Walther H (2001) Generation of photon number states on demand via cavity, quantum electrodynamics. *Physical Review Letters* 86: 3534–3537.

The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Perot configuration. Because spontaneous emission is a major source of energy loss, speed limitations, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. If the fraction of spontaneous emission coupled into the lasing mode is made close to one, the ‘thresholdless’ laser is obtained, in which the light output increases almost linearly with the pump power instead of exhibiting a sharp turn-on at the pump threshold.

Semiconductor microcavities provide high-Q Fabry–Perot cavities for both basic studies and potential applications. Molecular beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15–20 pairs of quarter-wave layers of $\text{Al}_{0.2}\text{Ga}_{0.8}\text{As}$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is maximum.

Since the microcavities have an extremely low threshold, their efficiency will also be high. Low-cost, high-density light source arrays and photonic circuits are made possible by the small size and low power consumption of such resonators. It will be possible to produce entire wafers containing millions of microlasers with a multipole arrangement instead of having to cleave each individual semiconductor laser as at present. This improvement will lead to higher yields and lower cost per element. Another advantage of surface-emitting microcavity sources is the efficient optical coupling of their stable symmetric mode patterns into optical fibers or waveguides. These lasers are thus certain to spark a revolution in optical communication in the future.

Conclusion

This article reviewed experiments in cavity quantum electrodynamics. The experiments shed new light on our understanding of the radiation interaction of atoms. The achievable strong coupling between atoms and radiation leads to the possibility to generate

entanglement between the generated radiation field and the atoms, presenting a basis for interesting applications in connection with, e.g., quantum computing and quantum information processing. Furthermore Fock states of the radiation field can be generated. The control of spontaneous decay finally leads to interesting new laser systems offering high efficiency.

Further Reading

- Berman, P.R., 1994. Cavity quantum electrodynamics. In: Bederson, B., Bates, D. (Eds.), *Advances in Atomic, Molecular, and Optical Physics*. Boston, MA: Academic Press. Supplement 2.
- Englert, B.-G., Löffler, M., Benson, O., *et al.*, 1998. Entangled atoms in micromaser physics. *Fortschritte der Physik* 46, 897–926.
- Haroche, S., 1992. Cavity quantum electrodynamics. In: Dalibard, J., Raimond, J.M., Zinn-Justin, J. (Eds.), *Fundamental Systems in Quantum Optics*. Amsterdam: North-Holland, pp. 767–935.
- Haroche, S., Kleppner, D., 1989. Cavity quantum electrodynamics. *Physics Today* 42, 24–30.
- Kimble, H.J., Carnal, O., Georgiades, N., *et al.*, 1995. Quantum optics with strong coupling. In: Wineland, D.J., Wieman, C.E., Smith, S.J. (Eds.), *Atomic Physics*, vol. 14. New York: American Institute of Physics, pp. 314–335.
- Marrocco, M., Weidinger, M., Sang, R.T., Walther, H., 1998. Quantum electrodynamic shifts of Rydberg energy levels between parallel metal plates. *Physical Review Letters* 81, 5784–5787.
- Meschede, D., 1992. Radiating atoms in confined space: from spontaneous emission to micromasers. *Physics Reports* 211, 201–250.
- Meystre, P., 1992. Cavity quantum optics and the quantum measurement process. In: Wolf, E. (Ed.), *Progress in Optics*, vol. 30. Amsterdam: North-Holland.
- Walther, H., 1992. Experiments on cavity quantum electrodynamics. *Physics Reports* 219, 263–281.

Cavity QED in Semiconductors

M Kira, W Hoyer, and SW Koch, Philipps-University, Marburg, Germany
G Khitrova and HM Gibbs, University of Arizona, Tucson, AZ, USA

© 2018 Elsevier Inc. All rights reserved.

Introduction

According to classical electromagnetism, propagating light fields obey Maxwell's equations under which electric and magnetic fields exchange their strength in an oscillatory manner. If a propagating field is followed at a single reference point, it is convenient to represent the electric field as a complex number $|E(t)|\exp(-i\omega t)$ where $|E(t)|$ is the magnitude of the field and the phase of the field $\exp(-i\omega t)$ oscillates in time t with the optical frequency ω . Fig. 1 illustrates how the classical field is determined by one single point in the complex plane; consequently, the phase and amplitude of the field are known exactly. According to quantum mechanics, this simple picture has to be altered because the Heisenberg uncertainty relation states that complementary quantities like phase and amplitude cannot be determined precisely at the same time. As a result, the real and imaginary parts of the field can be determined only with accuracy $\Delta\text{Re}[E]$ and $\Delta\text{Im}[E]$, respectively, and the Heisenberg uncertainty principle determines the best possible accuracy $\Delta\text{Re}[E] \Delta\text{Im}[E] \geq 1$. To incorporate the fundamental inaccuracy, the electric field has to be defined by using complex-valued distributions or equivalently wavefunctions as shown in Fig. 1. This quantization procedure leads to the Schrödinger equation for light analogous to that for particles. The field called quantum optics investigates the quantum electrodynamics (QED) features of light.

In general, light does not propagate freely in space, but it interacts strongly with the surrounding matter. For example, light can be absorbed, and its energy can be converted into excitation of the matter. As a result, the light may be slowed down, reflected, scattered, or diffracted. In order to explain the implications of the light-matter interaction in detail, one obviously has to apply a quantum mechanical description also for the matter. This approach determines the so-called eigenstates of the matter, so that the matter may occupy only certain discrete states with discrete energies. The Rydberg series of a hydrogen atom is a typical example. Similar discrete energy levels are also found for the quantized light; these levels are often referred to as photons which heuristically describe the particle properties of the light. The lowest-order light-matter interaction consists of processes where one photon is absorbed (emitted) while the matter is simultaneously excited (de-excited) from one eigenstate to another. This generic type of interaction leads to a microscopic description of the optical effects mentioned above. The magnitude of these effects depends on the strength of the interaction. By implementing different cavity configurations, the reflected light can be forced to travel across the same matter many times. As a result, the light-matter interaction is enhanced by a factor proportional to the multiple passes of the light. Thus, a cavity can be efficiently used to amplify optical phenomena. Fig. 2 shows cavity configurations commonly used for semiconductors. Semiconductor cavity QED investigates quantum optical effects in semiconductor systems by enhancing them with cavity mirrors. In general, quantum optical features produce classically unexpected effects which typically stem from: (a) the discrete nature of eigenstates of light and matter; (b) the superposition principle of quantum mechanics; and (c) the Heisenberg uncertainty principle. In this article, we review both theoretical and experimental advances made so far to predict, observe, and control quantum optical effects in semiconductors with respect to effects (a)–(c).

During the last few decades, atomic quantum optics has already developed toward QED investigations, and semiconductor optics is developing rapidly into the same regime. Advanced atomic QED theories have successfully explained and guided intriguing experimental developments like laser cooling, atom condensation, and photon teleportation. However, the atomic

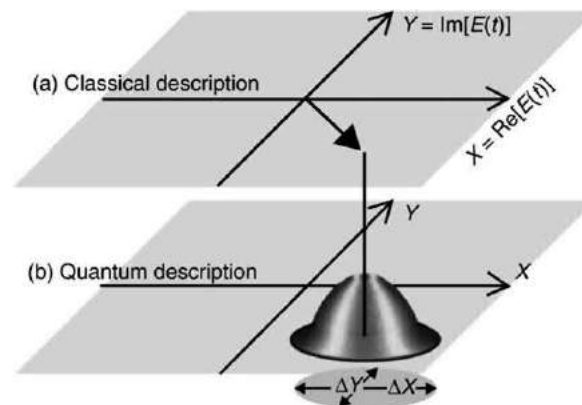


Fig. 1 (a) The classically described light field $E(t)$ is a single point in the complex XY -plane; the corresponding (X, Y) vector has a definite phase and angle. (b) Quantum mechanically described light is a distribution with fluctuations ΔX and ΔY described by the shaded circle.

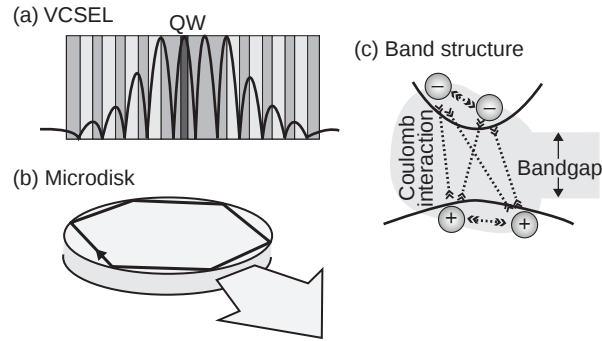


Fig. 2 (a) A typical vertical-cavity surface-emitting laser (VCSEL) structure; a stack of quarter-wavelength layers with different refractive index provides a mirror such that the light intensity (solid line) is strongly confined inside the cavity. A quantum well (QW) is positioned at the field maximum. (b) In a microdisk laser, light encounters ideal total reflection when the edge of the disk is reached. The so-called whispering gallery mode emits (leaks) light from the edge. (c) A schematic illustration of semiconductor band structure, bandgap, and Coulomb interaction for an electro-hole system.

approach can mostly be applied to describe dilute and only weakly interacting atomic gases since relatively simple models of few-level systems are used to describe the material. The elementary electronic excitations in semiconductors consist of electrons and holes (missing electrons) lifted into the conduction and valence bands, respectively. The corresponding eigenstates form continuous energy bands with the band-gap energy separation indicated in Fig. 2. Since electrons and holes have opposite charges, they experience Coulomb attraction whereas bare electrons or holes repel each other. Under favorable conditions, the attractive Coulomb interaction prevails and atom-like bound electron-hole pairs (excitons) may be formed. However, since the Coulomb force has an infinite range and an electron-hole system is typically dense, excitons cannot be treated as weakly interacting quasiparticles. As a result, the atomic QED approaches cannot be used to describe quantum optics of semiconductors in general. Thus, this article concentrates mostly on semiconductor QED theory which fully includes the interaction of charge carriers, i.e., electrons and holes, as well as quantum features of light.

The development work on quantum devices and processes is vital for advancements in several key technologies such as computers and telecommunications. The continuous decrease in component dimensions leads to a microscopic structure size (less than $1\ \mu\text{m}$); at the same time, increase in the device performance is accompanied by ultrafast operation time (faster than $1\ \text{ps}$). Due to these development trends, the properties of components and processes are becoming more and more quantum mechanical. Although the description of the quantum processes is complicated, the microscopic behavior offers entirely new operational functionality such as massively parallel quantum computers. These possibilities are based on the controlled manipulation of the quantum mechanical state of the light and matter. Due to rapid recent developments, optically coupled semiconductor devices have a great potential to become technologically successful and commercially viable quantum components.

Quantized States: Observation via Strong Light-Matter Interaction

In order to enhance the light-matter interaction, we choose a system where the semiconductor is placed inside a cavity in the position where the field intensity is maximum. For an empty cavity, the field intensity has a strong dependence upon frequency with a strong resonance at ω_R when the cavity length is an integer number of light wavelengths in the material. The width of the resonance is directly related to the quality of the cavity as determined by the number of back-and-forth reflections of light inside the cavity. If one chooses a vertical-cavity surface-emitting laser (VCSEL) structure, the most efficient coupling is obtained by placing a thin planar semiconductor structure at the field maximum as shown in Fig. 2. If the semiconductor is planar and thin, the structure is called a quantum well because electrons and holes are confined in one direction in analogy to the standard particle-in-a-box problem of fundamental quantum mechanics. As a result, the energy levels of a quantum well are discrete in the confinement direction. If the structure is narrow enough, the system dynamics is confined to the lowest quantum-well level such that the carriers are quasi-two-dimensional due to the free in-plane motion. The optical response of such a system to a weak classical probe beam has been successfully analyzed with the so-called semiconductor Bloch equations. For the quantum-well system alone, the absorption spectrum may have a sharp resonance; this is often referred to as an excitonic resonance since it is located below the fundamental bandgap energy at a position corresponding to the exciton binding energy. Obviously, the light-matter coupling becomes large when the exciton and cavity resonances coincide. However, since these resonances are coupled, the optical response is altered; we observe a splitting into two absorption peaks instead of the original degenerate resonances. This phenomenon is commonly referred to as normal-mode coupling; in general, it is a quite common feature in quantum mechanics that degenerate states split due to an additional interaction. Fig. 3 shows the first experimental observation of normal-mode coupling in semiconductors.

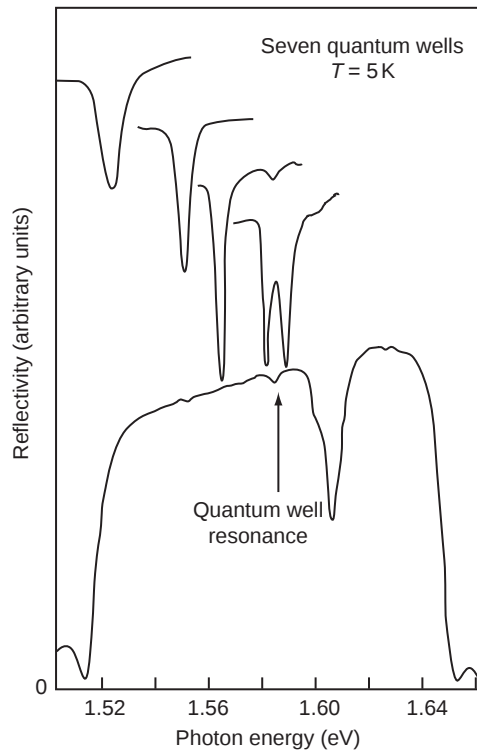


Fig. 3 Probe reflectivity of a seven-quantum-well microcavity structure. The various curves correspond to the energy difference between exciton and cavity resonances. When these resonances become degenerate, the reflectivity shows two resonances corresponding to the normal-mode splitting. From Weisbuch C, Nishioka M, Ishikawa A and Arakawa Y (1992) Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Physical Review Letters* 69: 3314. Copyright (2004) by the American Physical Society.

Since a semiconductor is a strongly interacting many-body system, several effects alter the specific character of an excitonic resonance when electrons and holes are excited. The attractive interaction between electrons and holes becomes weaker for increased carrier density since the surrounding charges screen the bare Coulomb interaction. A particular exciton may also experience scattering from nearby electrons, holes, or other excitons via the Coulomb interaction. Furthermore, electrons and holes are Fermions, which requires that any given carrier state can only be occupied once. This poses fundamental limits on how many carriers can coexist in a specific volume; above a certain limit, additional occupation is prevented, and so-called Pauli blocking is observed. Since an exciton consists fundamentally of Fermionic constituents, eventually Pauli blocking effects become important even for excitons. Due to a combination of these many-body effects, the excitonic resonance is weakened by an increasing carrier density. **Fig. 4** shows a comparison between theory and experiment of the excitonic absorption spectrum for different carrier densities. For elevated densities, the exciton resonance is broadened and its height reduced. When the same investigations are repeated in a microcavity, we observe that the normal-mode peaks decrease in height but their separation is roughly unchanged for moderate densities. In general, the normal-mode splitting is proportional to the oscillator strength of the excitonic absorption, i.e., the area under the resonance. Thus, the observed constant splitting suggests an unchanged oscillator strength for moderate densities. The invariant oscillator strength was unexpected and can only be explained by using a theory which includes the Coulomb interaction of carriers microscopically. When the carrier density is increased even further, the exciton resonance is completely bleached, and the microcavity transmission has only a single peak, at the cavity mode energy; this is commonly referred to as the weak-coupling regime, in contrast to the nonperturbative strong coupling with two transmission peaks. The strong coupling regime provides several intriguing phenomena; for example, parametric amplification of emission has been demonstrated by applying simultaneously multiple light beams to the sample.

The ultimate cavity QED limit of normal-mode coupling follows when the light-matter interaction is so strong that even a single photon leads to splitting (genuine strong coupling). We briefly outline some aspects of the quantum statistical limit known from atomic physics. This situation can be analyzed with the so-called Jaynes-Cummings model where only two states of the atom are included together with a single light mode. In this case, the interaction between a single photon and an atom leads to a normal-mode splitting g_0 . If two photons interact with one atom, the splitting increases to $2^{1/2}g_0$; more generally coupling between n photons and an atom leads to a splitting of $n^{1/2}g_0$. In the QED limit, the matter response is very nonlinear because one can detect the quantized nature of light directly from the energy splitting. This limit has already been reached in atomic cavity QED, whereas this regime is hard to approach with semiconductors because they behave more like multi-atom systems. If one has N atoms in a cavity, the first photon excites one atom, but there are N different ways of doing this, so the splitting is $N^{1/2}g_0$. If a second photon

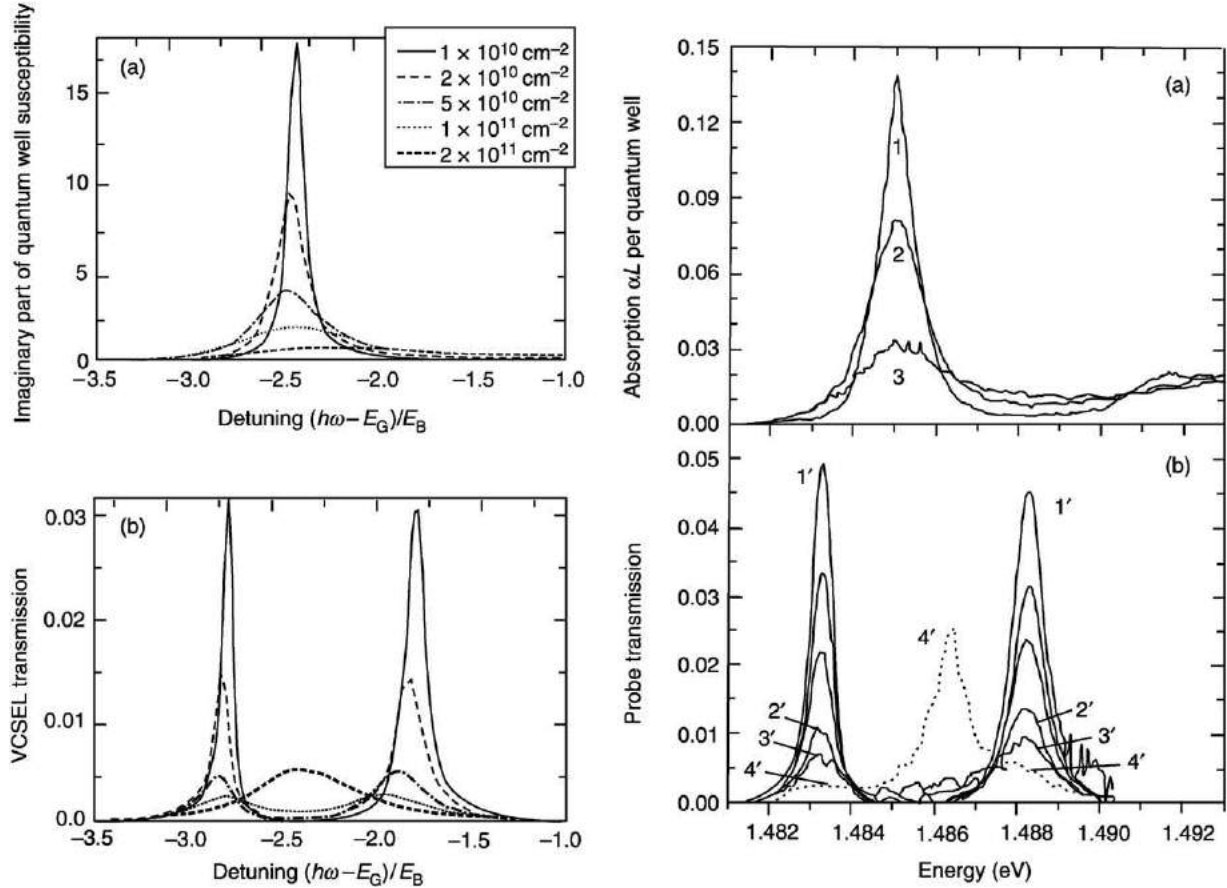


Fig. 4 Left column: (a) microscopically computed bare quantum-well absorption spectrum, i.e., the imaginary part of the susceptibility, as a function of carrier density. (b) Calculated transmission of the quantum-well microcavity for corresponding densities. The right column shows experimentally determined (a) absorption and (b) microcavity transmission for the same conditions as in the calculations. From Jahnke F, Kira M, Koch SW, *et al.* (1996) Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Physical Review Letters* 77: 5257. Copyright (2004) by the American Physical Society.

arrives, it interacts with the remaining unexcited atoms rather than the excited one, so that the splitting remains $N^{1/2}g_0$ as long as the number of photons is much smaller than N . Thus, in an N -atom system, or equivalently in semiconductors, the quantized light effects are much harder to observe than for a single atom. Currently, the normal-mode splitting in a semiconductor microcavity is explainable by the classical features of light, so that only the quantized nature of the matter is observed. In order to approach the ultimate cavity QED limit, one obviously has to reduce N and increase the light-matter coupling. In the future, this objective might be possible in quantum-dot/nanocavity systems where the semiconductor is confined in all spatial directions.

When an excited semiconductor is not under any influence of external classical light fields, it can still emit light via spontaneous emission resulting from the recombination of electron-hole pairs. The resulting light emission is called photoluminescence. Since the emission process takes place in a strongly interacting many-body system, one has to systematically include the Coulomb interaction and Fermionic features. The emission properties can be consistently described by the so-called semiconductor luminescence equations. When the microcavity photoluminescence spectrum is investigated, one observes a similar normal-mode splitting as for the transmission studies. However, this normal-mode coupling is not in the ultimate cavity QED limit even though the quantum fluctuations of the light field trigger the spontaneous emission. Nevertheless, the luminescence shows an interesting new transition as a function of the excitation level. Fig. 5 contains a comparison between theory and experiment of normal-mode peak heights and positions as a function of carrier density. For low densities, the high-energy peak is lower in height but it overtakes the low-energy peak for moderate carrier densities that are still below lasing threshold. This threshold-like overtaking was attributed to Bose action, involving exciton formation, final-state stimulation, and Bose condensation. The inset shows an estimate of just how Bosonic the excitons are, i.e., how much they actually behave as independent atoms; the value unity corresponds to the fully Bosonic situation. The nonlinear luminescence transition takes place at a density regime where the underlying Fermionic electron and hole contributions become important (commutator is roughly 0.5). A more detailed analysis with the semiconductor luminescence equations shows that Fermionic carrier nonlinearities in a detuned cavity are responsible for the experimental observations. This example of a misinterpretation based on a Bosonic analysis stresses how important it is to include the Coulomb interaction and Fermion character of carriers when analyzing the properties of semiconductors.

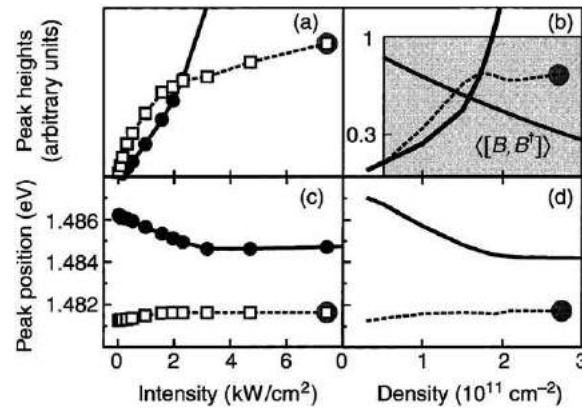


Fig. 5 Measured microcavity luminescence intensities (a) and peak energies (c) versus excitation intensity for the high-energy peak (solid line) and the low-energy peak (dashed line). The results of the microscopic theory are shown in (b) and (d). From Kira M, Jahnke F, Koch SW, *et al.* (1997) Quantum theory of nonlinear semiconductor microcavity luminescence explaining “boson” experiments. *Physical Review Letters* 79: 5170. Copyright (2004) by the American Physical Society.

Superposition Principle: Observation of Quantum Optical Correlations

The above normal-mode-coupling investigations have shown interesting nonlinear effects, but they also revealed that the ultimate QED limit – demonstrating discrete states of light directly – has not yet been reached. Alternatively, a QED investigation may concentrate on other fundamental features of quantum mechanics. One possibility is to study the implications of the superposition principle which states that the joint properties of a light–matter system can always be expressed as a superposition that combines light and matter wavefunctions. The most dramatic consequence can be observed in so-called entangled wavefunctions which cannot be factorized into light and matter parts. In a factorized wavefunction, the light and matter parts are independent whereas entangled wavefunctions conditionally connect light and matter wavefunctions. To elaborate the subtle details of entanglement, we first analyze a simple example. Assume that light can be in two different polarization states, $|\sigma^+\rangle$ or $|\sigma^-\rangle$, and the matter part is either excited $|\text{up}\rangle$ or de-excited $|\text{down}\rangle$. A wavefunction, $|\text{up}\rangle + |\text{down}\rangle$ $|\sigma^+\rangle + |\sigma^-\rangle$, is clearly a superposition of the fundamental states, and at the same time the light and matter parts are completely factorized. However, the wavefunction $|\text{up}\rangle |\sigma^+\rangle + |\text{down}\rangle |\sigma^-\rangle$ is entangled, since the light and matter parts cannot be separated. In the entangled state, measurement on the light state will conditionally determine a definite state of the matter, whereas the factorized state has no such conditionality. The entanglement has far-reaching consequences which cannot be understood with a classical analysis. For example, the principles of teleportation and quantum computing follow directly from the controlled manipulation of different parts of the entangled wavefunction. For atomic systems, wavefunction entanglement has been demonstrated and utilized in several experiments. For semiconductors, the direct manipulation and detection of the wavefunction seems difficult due to the overwhelming complexity of the many-body wavefunctions. Once again, the entanglement and the wavefunction are simplest for low-dimensional structures; direct entanglement effects have been demonstrated recently in a single quantum dot and between a pair of dots. Also for more complicated semiconductor systems, such as a quantum well in a microcavity, the entanglement can be observed as correlations between light and matter. In this case, the existence of QED correlations basically means that light and matter properties depend conditionally on each other. In general, the direct entanglement and QED correlation investigations have a large development potential for semiconductors.

When a semiconductor is excited with an external light pulse, entanglement-related correlations couple the classical and luminescence emission dynamics. The resulting dynamic interplay between the semiconductor Bloch and luminescence equations is mainly mediated by the correlation between photons and electron–hole densities. In the following, we analyze a microcavity configuration where a strong pump pulse generates large QED correlations. The effect of the correlations is then measured by the response of a weak probe beam. Fig. 6 shows a comparison of theory and experiment of probe reflection in a configuration where the pump and probe do not have any spectral overlap. The measured probe reflection displays long-living oscillations as a function of time delay, i.e., phase difference, between the pump and the probe. Only by including the QED correlation in the theory can the oscillatory probe reflection be explained. When the QED contributions are omitted from the theory, the phase difference does not have any effect on the probe reflection. Thus, this experiment–theory example represents a direct observation of the cavity QED effects in semiconductors. Fig. 6 also shows more pronounced oscillations when two phase-locked pump pulses provide the excitation; again the full QED theory explains the enhanced oscillation features.

The QED features can also alter the normal-mode coupling characteristic of a weak probe beam. Fig. 7 shows a comparison of theory and experiment for a situation where the pump spectrum is located between the two normal-mode coupling resonances. Both the theory and experiment show a new third resonance which follows the energetic position of the pump. The third peak is a direct consequence of the QED correlations. Since the third peak is generated from the QED field–carrier correlations generated by

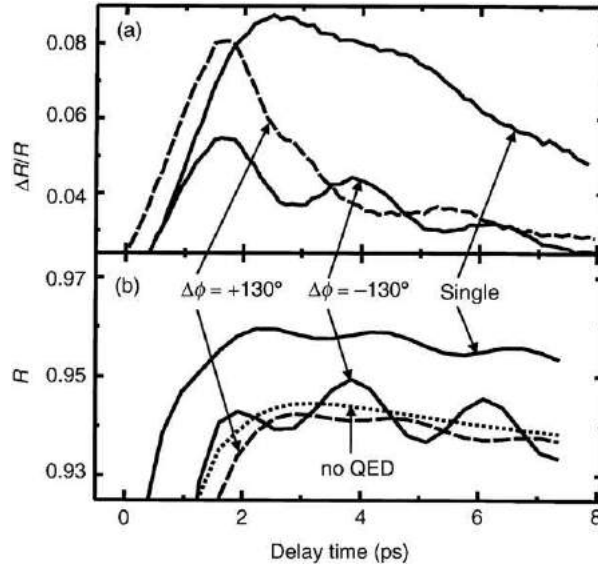


Fig. 6 (a) Measured differential reflectivity and (b) computed reflection probability of the probe pulse as function of probe delay with respect to the excitation pulse. For two-pulse excitations, $\pm 130^\circ$ relative phase shifts are used. The dotted line is computed without the QED correlations. From Lee Y-S, Norris TB, Kira M, *et al.* (1999) Quantum correlations and intraband coherences in semiconductor cavity QED. *Physical Review Letters* 83: 5338. Copyright (2004) by the American Physical Society.

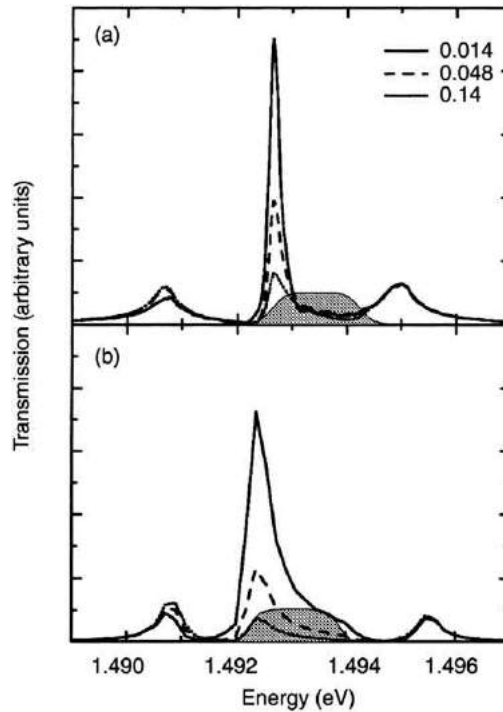


Fig. 7 Dependence of probe transmission on probe intensity (nJ cm^{-2}) at a constant pump intensity of 330 nJ cm^{-2} for the microcavity. (a) Experiment, and (b) theory. The shaded regions indicate the pump spectrum. From Ell C, Brick P, Hübner M, *et al.* (2000) Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Physical Review Letters* 85: 5392. Copyright (2004) by the American Physical Society.

the pump pulse, its magnitude follows the intensity of the pump. As the probe intensity is decreased the relative effect of the QED correlations is increased so that the third peak grows in the probe transmission.

In these QED investigations, the cavity plays an important role since the mirror feedback leads to a significant amplification of the QED effects through the enhanced light-matter coupling. From a practical point of view, normal-mode coupling also provides well-separated spectral features which can be used as classical-emission reference points.

Heisenberg Uncertainty Principle: Squeezing of Light Emission

Quantized light effects can also be investigated by measuring the light fluctuations of the field directly. To maximize the quantum effects, we analyze the field fluctuations in the emission directions where the classical field vanishes. This situation can be realized in planar quantum-well structures which are nearly free of disorder. In such systems, a light pulse propagating perpendicular to the structure leads to transmission and reflection of classical light only along the excitation axis. If the detection is performed at an angle, the so-called secondary emission is purely quantum mechanical. However, the classical and quantum emissions are coupled in the same way as for the QED correlation study. The resulting fluctuations of secondary emission obey the Heisenberg uncertainty principle; in the following, the special quantum features of these fluctuations are investigated.

In the full analysis, we determine the variance ΔX and ΔY of the emission as shown in Fig. 8 (see also Fig. 1). The Heisenberg uncertainty principle requires that the quadrature fluctuations obey $\Delta X \Delta Y \geq 1$. For a specific quadrature, the minimum uncertainty limit is usually defined to be $\Delta X = 1$ or $\Delta Y = 1$. If ΔX and ΔY are different, the observed light field is squeezed; and if the variance in one quadrature is less than one, the field is squeezed below the minimum uncertainty limit. In both cases, the field has a strong quantum nature; in particular, squeezing below unity suggests that measurements can be more accurate than the standard quantum limit for that quadrature.

To illustrate the behavior of the quantum fluctuations, we excite the quantum well resonantly with a relatively strong pulse. The emission is detected at an angle of 45° away from the excitation axis. Since the excitation pulse is relatively strong, the carrier density starts to oscillate during the pulse because Fermionic carriers can be excited only once; thus, further excitation actually leads to de-excitation. In other words, when these states become almost fully excited, the pulse can no longer excite the system, and we observe periodic de-excitation and excitation analogous to the Rabi-flopping of a strongly excited two-level system. Fig. 8 shows the exciting pulse and corresponding quadrature fluctuations during the excitation process. As long as the pulse is present, the squeezing is at the 4% level and oscillates with the Rabi-flopping frequency. The quadrature fluctuations clearly show squeezing below the minimum uncertainty limit. Hence, the field has obvious quantum properties.

Similar squeezing and quantum characteristics statistics have been predicted and observed in the photon statistics of resonance fluorescence from two-level atoms subjected to an intense coherent light beam. The quantum properties of the scattered light in both quantum-well and atomic emission are enhanced when the driving field forces the system to oscillate between the excited and de-excited state. Just as described above, when the system becomes de-excited, it can no longer emit an additional photon; this inhibition manifests itself as sub-Poissonian photon statistics and squeezing in the mode quadratures. When the excitation pulse re-excites the system, such restrictions are no longer imposed. Thus, the field properties show quantum features oscillating with the Rabi-flopping frequency. Squeezing effects can also be observed with excitation by an electric current; for example, amplitude squeezing of diode laser emission has been demonstrated by controlling the electron statistics of the current flow.

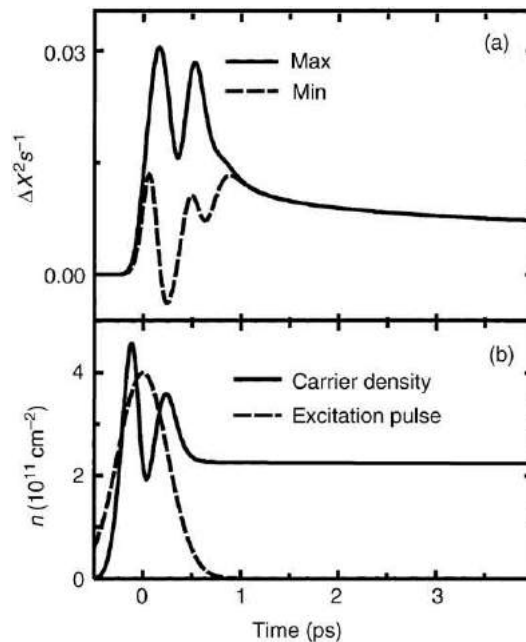


Fig. 8 (a) Time-resolved maximum (solid line) and minimum (dashed line) deviations of the mode quadratures from the vacuum value of unity. (b) Corresponding time-resolved carrier density (solid line) and excitation pulse (dashed line). From Kira M, Jahnke F and Koch SW (1999) Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Physical Review Letters* 82: 3544. Copyright (2004) by the American Physical Society.

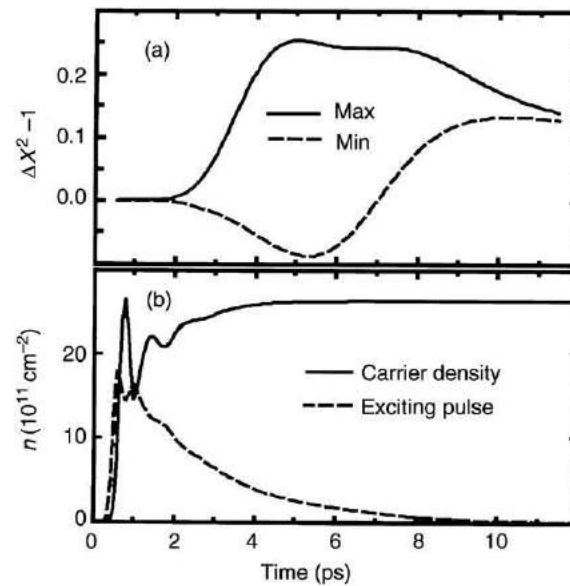


Fig. 9 Squeezed light from a quantum well inside a microcavity: (a) the time evolution of maximum and minimum of the quadrature fluctuations; (b) the corresponding density and light field intensity at the quantum well.

The squeezing investigations demonstrate that some QED effects are expected to be seen without a cavity. **Fig. 9** shows the quadrature fluctuations when the cavity is added to the squeezing analysis. We observe qualitatively similar squeezing, but now the level of squeezing is enhanced up to 30% compared to the minimum uncertainty. This is once again a demonstration how QED effects can be amplified by using an increased light–matter coupling provided by the cavity.

Summary

This article overviews some of the most intriguing features of cavity quantum electrodynamics in semiconductors. Even though the related research has started only recently, several important quantum phenomena have already been predicted and measured: (a) the discrete quantum mechanical states of light and/or matter can be measured with strong coupling; (b) the consequences of the superposition principle have been detected as light–matter entanglement; and (c) the Heisenberg uncertainty principle has been tested via the squeezing of light. In all cases, the cavity enhances light–matter coupling leading to more pronounced quantum effects. Compared to atomic systems, the semiconductor has strong Coulomb correlation effects due to the relatively high density, which makes the analysis challenging. However, the semiconductor can also provide new mechanisms like the photon/carrier-density correlations which trigger new unexpected quantum phenomena.

The field of semiconductor quantum optics is very active and is developing rapidly. It benefits both from advances in computer capabilities and in semiconductor crystal growth. The increasing computer capacity allows more profound, accurate, and realistic modeling of semiconductor structures. At the same time, advances in crystal growth technology provide us with improved samples which are almost disorderless. In particular, the growth of quantum wells with narrower exciton linewidths and quantum dots with larger dipole moments and reduced dephasing rates may be achieved, which is certain to make quantum features increasingly apparent and unavoidable.

All these research efforts eventually focus on producing devices utilizing quantum mechanical principles. One of the main objectives endeavors to develop quantum logic components for building blocks of quantum computers. Similar expanding possibilities can be expected, e.g., for accuracy of detection, device efficiency, and component design in general. Considering all of this, we are almost guaranteed to see new astonishing advancements; in this bright future, semiconductor cavity QED research will most likely be a pre-eminent element.

See also: Cavity QED

Further Reading

Allen, L., Eberly, J.H., 1987. *Optical Resonance and Two-Level Atoms*, 1st edn. New York: Dover.
 Banyai, L., Koch, S.W., 1993. *Semiconductor Quantum Dots*, 1st edn. Singapore: World Scientific.
 Berman, P.R. (Ed.), 1994. Several articles in *Cavity Quantum Electrodynamics*, 1st edn., Boston, MA: Academic Press.

- Cohen-Tannoudji, C., Dupont-Roc, J., Grynberg, G., 1989. *Photons and Atoms*, 3rd edn. New York: Wiley.
- Haug, H., Koch, S.W., 2004. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific.
- Khitrova, G., Gibbs, H.M., Jahnke, F., Kira, M., Koch, S.W., 1999. Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Reviews of Modern Physics* 71, 1591–1639.
- Kira, M., Jahnke, F., Hoyer, W., Koch, S.W., 1999. Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Progress in Quantum Electronics* 23, 189–279.
- Scully, M.O., Zubairy, M.S., 1999. *Quantum Optics*, 1st edn. Cambridge, UK: Cambridge University Press.
- Walls, D.F., Milburn, G.J., 1994. *Quantum Optics*, 1st edn. New York: Springer-Verlag.

Silicon Qubits

Thaddeus D Ladd, HRL Laboratories, LLC, Malibu, CA, United States

Malcolm S Carroll, Sandia National Laboratory, Albuquerque, NM, United States

© 2018 Elsevier Inc. All rights reserved.

Motivation for Silicon Qubits

The first motivation for basing qubits on silicon is the obvious foundation of classical microelectronics. Although silicon quantum computers would operate in a fundamentally different way from classical computers – for example, at cryogenic temperatures – still the level of development in material quality, crystal growth, and fabrication methodologies for silicon is unrivaled by any other material in the world. Leveraging even a small fraction of the worldwide investment in silicon for qubit development could potentially put silicon-based qubits far ahead of other solid-state alternatives.

The second, less obvious reason for choosing silicon is the remarkably clean magnetic environment witnessed by spins in highly purified and isotopically enriched silicon material. Fortuitously, 95.3% of the naturally occurring isotopes of Si nuclei (^{28}Si and ^{30}Si) are spin-0. These nuclei therefore have a “closed shell” of nuclear moments, providing no external magnetic field whatsoever. Add to this the possibility of intrinsic silicon with part-per-billion chemical quality and the system is remarkably close to “vacuum” with respect to magnetic noise properties.

The most dramatic experimental demonstrations of this “semiconductor vacuum” are from inductively and optically probed magnetic resonance experiments in bulk silicon crystals which have been isotopically enriched to contain 99.9995% ^{28}Si . These experiments probe dilute phosphorus impurities (at a density of 10^{15} cm^{-3} , i.e. one out of every 50 million atoms), which serve as donors in the material. For conduction electrons, these are closely analogous to a suspended single-charged ion, although it is worth recalling that these “ions” are suspended in a highly quiet material featuring negligible motion due to lattice vibrations without any need for laser cooling, laser trapping, or electrodes, in contrast to trapped ion or neutral atom qubit experimental setups.

In cryogenic, electron spin resonance (ESR) experiments, electron spins precess in an applied magnetic field, kicked off by a microwave pulse. The spinning electrons dephase first and foremost due to quasi-static inhomogeneities in the local magnetic field, an effect readily reversed by spin-echo techniques, which periodically invert the relative phases accrued by static rotation speed differences. At high levels of enrichment (e.g., 99.9995% ^{28}Si) and once inhomogeneity is removed as a source of dephasing, the most important term causing dephasing is the dipole-dipole couplings between the dilute phosphorus atoms themselves. By applying a magnetic field gradient, these dipole-dipole effects can also be reduced. In [Tyryshkin et al. \(2012\)](#), it is shown that electron spins may be estimated to precess in phase for nearly a minute in this material.

Phosphorus impurities are optically addressable as well. They have a hydrogen-like energy structure in the THz energy range, which is inconvenient for both electronics and optics, but still a noteworthy structure for qubit studies ([Litvinenko et al., 2015](#)). They also feature a set of sharp optical transitions corresponding to the transition from an atomic-hydrogen-like neutral donor to a molecular-hydrogen-like bound excitonic state. These transitions are atomically sharp in isotopically enhanced silicon, enabling the ability to polarize and measure the hyperfine couplings to the spin of the ^{31}P nucleus. Again using radio-frequency pulses to refocus dephasing effects from inhomogeneous magnetic fields, these nuclei can be observed to precess without loss of phase coherence for hours, including at room temperature ([Saeedi et al., 2013](#)).

These bulk experiments are a testament to the capability to preserve spin coherence in silicon, and are not replicated in any other material. But while a “vacuum” may be an apt description for the magnetic environment seen by conduction electrons, unfortunately the electronic structure of silicon is not completely “vacuum”-like. The electron structure of crystalline silicon produces an indirect bandgap. The energy minima for electrons in the conduction band are not at crystal momentum $k=0$, but rather along the six crystalline axes of the cubic structure, providing an anisotropy quite unlike a vacuum. These sixfold degenerate minima of the conduction band are referred to as valleys, and their degeneracy poses a problem for qubits since they can represent an uncontrolled degree of freedom for electrons that prevents clean control of qubit states. For the phosphorus impurity, the neutral donor ground state is an equal superposition of valley states which is energetically separated from other states by an energy of 11.7 meV in unstrained bulk ([Zwanenburg et al., 2013](#)). While this effectively breaks the valley degeneracy for this system, a further consequence is that the atomic-like structure of the zero-phonon donor-bound exciton transitions comes with an extreme optical inefficiency: decay of the donor-bound exciton state is dominated first by Auger recombination, in which excitonic recombination ionizes the impurity, and second by broad phonon-assisted transitions.

Valley physics is more severe for electrons bound by large, shallow potentials in planar silicon devices. The valley degeneracy is partially broken by tensile strain in heterostructure stacks or by vertical electrostatic confinement in close proximity to an oxide interface, but while these effects raise four of the six valley states by 10s to 100s of meV, the remaining two valley states are much closer to degenerate. The splitting of these last two valleys is determined by atomic-scale details of the structure and the magnitude of the vertical electric field; assuring a sufficiently large value of this splitting is a key design constraint in silicon qubits.

Within a given valley or valley superposition, conduction electrons in silicon are reasonably well described by an effective mass model, with an in-plane effective mass of $0.19m_0$. This number is substantially higher than in GaAs, an earlier material studied for

spin qubits, requiring much higher demands on the lithographic resolution of structures for trapping or controlling electrons relative to GaAs. These limits are, however, within the limits of e-beam lithography and even modern implementations of photolithography. A substantial number of approaches to lithographically defined qubits have been introduced in the last 20 years. A technical review of many of these proposals and the progress on implementing them was compiled by Zwanenburg *et al.* (2013).

In this article, we highlight what the authors consider three of the most important categories of silicon qubits for silicon-based quantum computing, most of which have made substantial progress in demonstrating coherent operation since the Zwanenburg *et al.* review was published. These three principle categories of qubits are single phosphorus impurities, metal-oxide-semiconductor (MOS)-based dots, and dots based on heterostructures of strained silicon quantum wells with SiGe barriers. In what follows we will elaborate on these three essential categories, followed by categorization of the ways of controlling these qubits.

Three Principal Material Types of Silicon Qubits

The three qubit categories are summarized in Fig. 1. All three types ultimately rely on the spins of trapped conduction electrons, although most variations of phosphorus-based qubits heavily employ the ^{31}P nuclear spin as well. The first principle difference is the choice for how to confine electrons in the z -direction, defined as the substrate growth direction, normal to the silicon wafer. In the case of phosphorus impurities, it is the Coulomb attraction of the phosphorus impurity, which appears positively charged when replacing a silicon atom. This potential strongly attracts the electron, resulting in a binding energy of 45.6 meV (Zwanenburg *et al.*, 2013). In contrast, metal-oxide-semiconductor (MOS) quantum dots trap electrons in an electric potential defined by the hard “wall” of an oxide and the electric field created by the bending of the silicon bandgap due to the oxide interface. This potential is the same as that of the channel of a silicon MOS Field Effect Transistor (MOSFET), an extremely successful device forming the basis of modern microelectronics. The ground state of this approximately triangular potential sits at energies 10s of meV above the conduction band. SiGe quantum dots are similar to the MOS case, except in this system the barrier is pushed away from the surface of the semiconductor material, being defined instead by the band-offset between the strained silicon quantum well and a SiGe barrier, typically engineered to be larger than 150 meV (Zwanenburg *et al.*, 2013).

The trade-offs of the three approaches to vertical confinement have largely to do with disorder. A key advantage of the phosphorus system is that every phosphorus impurity provides effectively the same potential in a silicon lattice, assuming the impurity is sufficiently removed from surfaces or dielectric interfaces. This is in contrast to the quantum dot systems, where the exact energy of the vertical confinement may be impacted by random fluctuations in the amorphous oxide in the MOS case or by the heterostructure interface in the SiGe case.

A related distinguishing feature is the method for horizontal electron confinement. Here, the phosphorus system confines electrons horizontally in principally the same way as it does vertically: by its $1/r$ Coulomb potential. In contrast, the MOS and SiGe quantum dot systems rely on electrostatic gates to provide a horizontal potential profile for electrons. There are a number of strategies for designing these gates. Example scanning electron microscope (SEM), cross-sectional transmission electron microscope (TEM), and scanning tunneling microscope (STM) images of fabricated gates for all three vertical confinement types are shown in Fig. 2.

Most recent silicon quantum dot qubits are enhancement-mode devices, which are designed for an empty MOS or quantum well channel. A global field gate or individual dot gates pull the confined electrons from ohmic contacts. This contrasts with depletion mode structures, more typical in the AlGaAs/GaAs system, in which the system is doped such that a gas of electrons would reside in the channel region in equilibrium, but then negatively biased gate electrodes push electrons away from the dot regions, leaving behind a controlled number of charges. Depletion mode behaves poorly in silicon due to the disorder created by the dopants, which is more poorly screened in silicon than in GaAs. Besides avoiding disorder, however, enhancement-mode

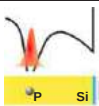
Type	Sketch	Main advantage	Main challenge
Donor		Lowest disorder and largest valley splitting	Multi-qubit fabrication and operation
MOS		Most similar to commercial transistors	Disordered potential
SiGe		Clean epitaxial barrier (low disorder)	Valley degeneracy

Fig. 1 Table indicating the three principle types of silicon qubits; the sketch column gives a rough image of the band structure as a black line, the electron wave function in red, and the material stack below. The confinement illustrated is referred to as “vertical” in the text (i.e. it the confinement in the growth direction), however the growth stack is here drawn horizontally, so that the energy of the band-structure above may be drawn vertically.

devices have the advantage that they could in principle confine electrons using only a single gate per dot. Most architectures mix enhancement and depletion operation using multiple gates per qubit. These multiple gates are useful for crafting each electron's confining potential, in part to overcome the effects of disorder in the vertically confining structure and in part to improve flexibility of electron control.

For all three silicon qubit categories, at least one form of control – under strong consideration since the seminal proposals of Loss and DiVincenzo (1998) for quantum dots and Kane (1998) for phosphorus donors – is the kinetic exchange interaction. This interaction relies on the ability to tune the energy level of a dot or donor based on a gate directly above it (called the A-gate in the Kane proposal and in this article, but often referred to as a plunger gate), and to control the interaction between nearby spins either via A-gates or with an additional gate controlling a tunnel barrier between the dots (called the J-gate in the Kane proposal and in this article, but often referred to as an exchange gate).

The kinetic exchange interaction is a consequence of quantum tunneling in the Coulomb blockade regime and Pauli exclusion (Zwanenburg *et al.*, 2013). The essential physics is that the spin-singlet state of a pair of electrons (with total angular momentum $S=0$ and an antisymmetric spin state) is capable of tunneling into the ground state for two electrons on a single dot or donor site, since the resulting state is fully antisymmetric. Spin-triplet states (with total angular momentum $S=1$ and a symmetric spin state) would, in contrast, result in a disallowed symmetric doubly-occupied ground state, and are therefore forbidden from tunneling unless the dots are detuned in energy by a sufficient amount to allow tunneling into excited orbital or valley states. The kinetic exchange interaction lowers the energy of the singlet state relative to the triplet state due to the singlet's allowed tunneling process. Electrical control of tunneling via gate voltages modulate this interaction, either by increasing the tunnel coupling directly (via J-gates) or by bringing the chemical potentials of the quantum dots or donors closer to a resonance condition (via A-gates).

Kinetic exchange provides one of the key proposed mechanisms for coupling silicon qubits, but there are others which will be addressed later in this article. Nonetheless, we will frame our discussion in terms of the basic A-gate/J-gate control of quantum dot chemical potentials and exchange energies for the sake of a meaningful comparison of the three basic categories. Fig. 2(a) provides the simplest visual realization of the A/J or plunger/exchange gate concept, with the visible “paddle” gates intending to serve the role of A gates and the straight gates between intending to serve the role of J gates. While harder to see in those images of dot-based devices employing overlapping gates, similar A/J functionality, in addition to horizontal confinement, is intended here as well.

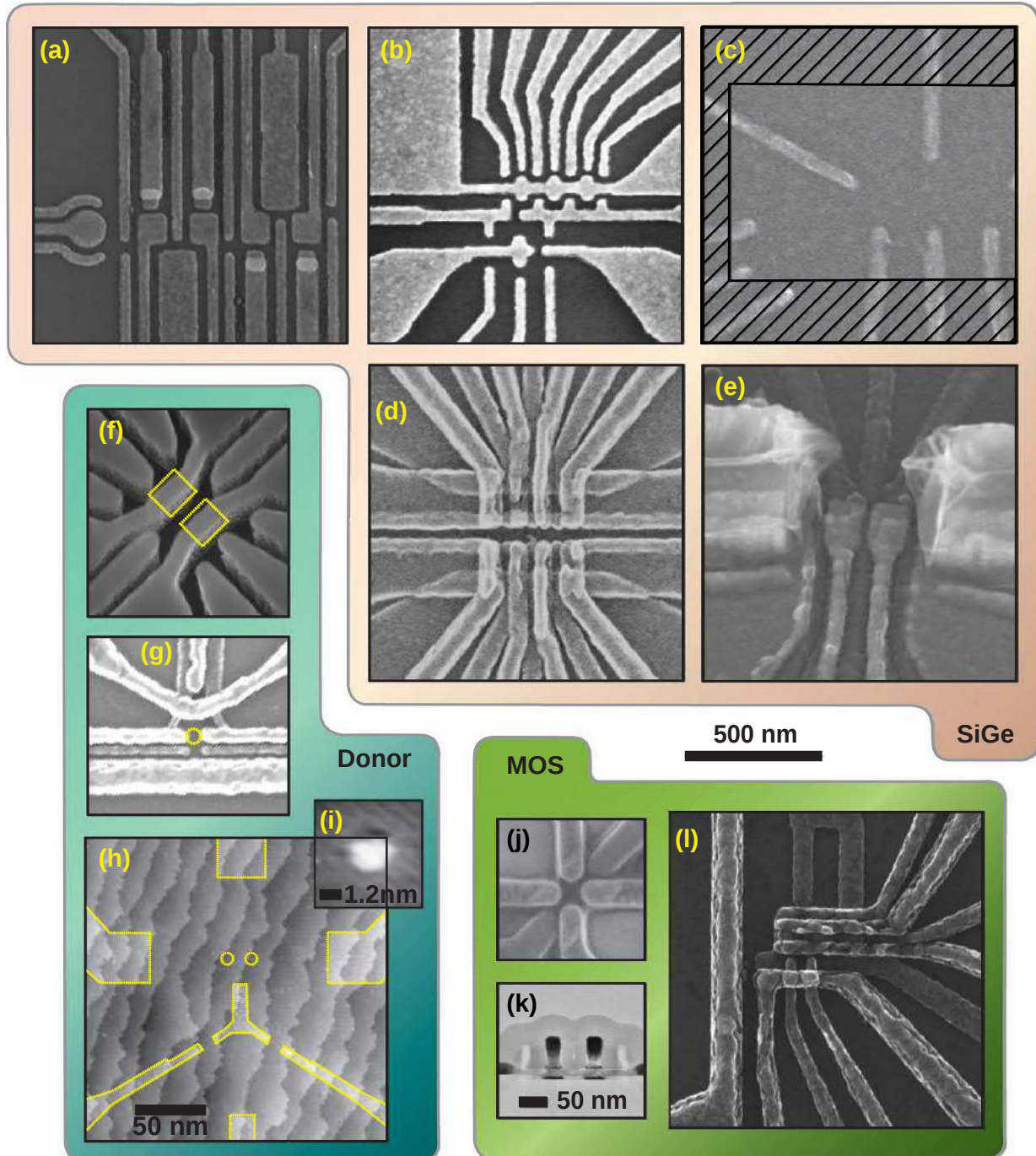
Due to their naturally occurring confining potentials, phosphorus impurities would appear to have an advantage relative to dots in that the requirement of multiple gates for horizontal confinement is relaxed. However, the phosphorus system can introduce substantially more strict fabrication requirements compared to quantum dots, especially in multi-qubit systems. The phosphorus potential results in extremely small electron wavefunctions. The Bohr radius of an electron on a donor is approximately 2.5 nm. This means that in order to implement a reasonably strong exchange interaction entirely in the bulk, theory estimates that A-gates would need to be separated by approximately 10–20 nanometer pitch (Zwanenburg *et al.*, 2013). An additional challenge is that the nondegenerate phosphorus valley state creates a lattice-scale wavefunction oscillation, which results in a kinetic exchange interaction that varies from crystal-lattice site to crystal-lattice site (Zwanenburg *et al.*, 2013). Some early donor qubit designs, as a consequence, proposed extremely challenging donor and electrode placement, both vertically and horizontally (Hollenberg *et al.*, 2006; Testolin *et al.*, 2007).

Alternative long-range coupling schemes between donors, as we describe below when we elaborate on coupling mechanisms, might be employed to circumvent many of the strict requirements of exchange-based donor qubit architectures. Such approaches could use standard silicon foundry processing, using ion implantation to place an average number of donors within the placement requirements of these proposed schemes. For exchange-based donor approaches, however, a new fabrication technique with atom-scale precision is under parallel development. Using a scanning-tunneling-microscope (STM) to do lithography on a hydrogen-terminated Si surface, combined with exposure of the patterned surfaces to phosphine, a quasi-three-dimensional molecular-beam-epitaxy capability with spatial resolution at a single lattice site has been demonstrated (Fuechle *et al.*, 2012). See Fig. 2(h) and (i) for qubit devices including both donor qubit islands and electrode gates fabricated this way (Watson *et al.*, 2015, 2017). The benefit of STM-defined phosphorus gates is not limited to their small size and pitch: the atom-by-atom fabrication leads to conducting channels that have shown substantially lower $1/f$ voltage noise than other lithographically defined gate stacks (Shamim *et al.*, 2016), which is a critical feature given that such $1/f$ noise may ultimately limit the fidelity of silicon qubits employing exchange interactions. The extremely promising and rapidly evolving STM-lithography approach, however, is still immature relative to the much more firmly established silicon foundry processes with respect to yield, integration, and manufacturing.

The appeal of quantum dot approaches relative to the phosphorus system is the engineering advantage of defining dots entirely by traditional lithography. The MOS system is especially promising in this regard due to its close affinity to the ubiquitous complementary-MOS (CMOS) transistor. A further advantage of MOS relative to the SiGe system is that the large vertical electric field present in this system breaks the valley degeneracy from a two-fold degeneracy to valley splittings that can easily exceed 300 μeV , corresponding to temperatures of 3.5 K (Gamble *et al.*, 2016). This large energy means that thermally initialized electron states have negligible excited valley population at typical dilution refrigerator temperatures. This vertical field is highly reduced in the SiGe system, leading to a less certain energy difference between the two lowest valley states. While comfortable valley splittings in excess of 50 μeV have been observed and used for high quality initialization, control, and read-out in some devices (Eng *et al.*, 2015), in other devices the valley splitting is too small for measurable qubit behavior (Borselli *et al.*, 2011; Zajac *et al.*, 2015).

Despite the challenge of valley splitting, the primary reason to consider SiGe-based dots relative to MOS-based dots has to do with noise and disorder. MOS-based dots place the qubit wavefunction against amorphous thermal SiO_2 . While this is certainly the most developed and studied semiconductor-oxide interface in history, even the highest quality oxides have intrinsic interfacial

disorder. There is inherent atomic mismatch between the two materials. MOS transistors also routinely observe $1/f$ character trap noise. Oxide disorder, therefore, might have adverse impact on the yield of MOS dots and the $1/f$ noise may limit the fidelity of MOS-dot operation. Si/SiGe dots hope to reduce the effect of interfacial disorder by pushing the qubit wavefunction away from any oxide or metal interfaces and into a region of cleaner epitaxial crystalline semiconductor. Indirect measures of disorder such as mobility and the density at which transport is first observed (i.e., the metal insulator transition) are both appreciably lower in the strained-Si channels of the SiGe-based structures. A flexible gate design in MOS may enable tuning configurations to compensate for this difference in electrostatic disorder and indeed high quality single MOS dots have been demonstrated (Yang *et al.*, 2013). It is unclear whether this can be extended to bigger multi-dot networks. With respect to separation of the electron channel from imperfect amorphous interfaces with SiGe, to date, this has not completely eliminated the presence of $1/f$ noise, presumably coming from the remote gate and dielectric stack of the device. Recent measures of charge noise in MOS quantum dots (Freeman *et al.*, 2016) and recent qubits that include a MOS interface (Harvey-Collard *et al.*, 2017) show noise amplitudes comparable to



SiGe devices (Jock *et al.*, 2017). As of the date of this article, the relative performance of MOS vs. SiGe quantum dots relative to both interfacial disorder and charge noise remains unclear.

Following the choice of how to trap electrons, qubits are further defined by how those electrons are controlled. Again, there are a few principle strategies for defining and controlling qubits which we now define, noting that they are not exclusive, as well-engineered qubits are likely to use a combination of control mechanisms.

Memory, Control, and Readout Mechanisms

Different types of silicon qubits make different choices for how to store quantum information, how to perform quantum logic, and how to read out quantum states. See Table 1 for a summary of the qubit types which we detail below.

For memory, the first basic decision is whether to store information as charge or spin, the latter being by far the favored choice due to its much longer coherence time. However, when performing quantum logic, spin qubits are often partially converted to charge qubits. We therefore discuss both charge and spin qubits, and then proceed to indicate their control mechanisms.

Charge Qubits

The simplest form of silicon qubit is the charge qubit. For this, consider two dots (or a dot and a donor) in charge states (n, m) and $(n + 1, m - 1)$, where the two numbers indicate the number of conduction electrons or valence holes in each dot. The simplest such qubit would encode information on whether an electron is on the left dot (or donor) or the right, e.g. the charge states $(1, 0)$ and $(0, 1)$. A key motivation for this type of qubit is the simplicity of control, measurement and initialization. For control, direct modulation of gate voltages will reliably move charges between dots with speeds typically limited only by the control electronics employed. For measurement, single-electron-transistor or similar devices are capable of sensing the motion of single charges at low-temperature (Zwanenburg *et al.*, 2013; Gonzalez-Zalba *et al.*, 2015), enabling a direct read-out of this qubit's state. By sweeping voltage bias of confinement gates, broad ranges of charge numbers n and m can be detected by observing the tunneling of

Table 1 Summary of memory and control mechanisms for silicon qubits discussed in this article.

Memory		Single-qubit control	Multiqubit control	
Charge		Direct charge motion	Capacitive couplings	
Spin	Multi-spin	Exchange interactions		Exchange and contact hyperfine interactions
		Single-triplet oscillations in gradients		
	Single-spin	ESR with AC magnetic fields	Magnetic dipole-dipole couplings	
		ESR with AC electric fields and gradients		

Fig. 2 Sample micrographs of silicon qubits. Qubits in the pink box employ SiGe heterostructures beneath the gate structures shown for gate confinement. Qubits in the blue box employ single donors. Qubits in the green box employ MOS confinement. All images are from SEMs except (h) and (i) which are STM images and (k) which is a TEM. All images are roughly scaled to the 500 nm scalebar shown in the center of the figure, except where indicated otherwise. Devices in Fig. (a)-(c) use Ti/Au gates, while those in (d), (e), (g), and (l) employ aluminum. Devices in (f) and (g) indicate windows for phosphorus donor implantation as dashed yellow regions. (a) Quadruple dot, image courtesy Mark Eriksson, University of Wisconsin, United States; simplified and adapted from Ward *et al.* (2016). (b) Triple dot structure fabricated by Christian Volk, Center for Quantum Devices, University of Copenhagen, 2016. (c) Double dot, image courtesy Kenta Takeda, University of Tokyo; see Takeda *et al.* (2016). doi: 10.1126/sciadv.1600694. The cross hatched region indicates the region in which a micromagnet made from deposited cobalt provides a magnetic gradient across the double-dot. A similar structure, without micromagnet, was demonstrated by HRL in Maune *et al.* (2012). (d) Six dot structure using overlapping gates, image courtesy Jason Petta, Princeton University; see Zajac *et al.* (2015). A similar structure was employed at HRL in Reed *et al.* (2016). (e) Double dot structure using overlapping gate with visible cobalt micromagnet. Image courtesy the Vandersypen group, copyright TU Delft, the Netherlands. (f) Polysilicon quantum dot and readout gate structure from Sandia National Laboratory; the quantum dot couples to an underlying phosphorus donor; see Harvey-Collard *et al.* (2017). (g) Gates for SET readout-channel for a single phosphorus device, courtesy Andrea Morello, UNSW; see Muhonen *et al.* (2014). (h) A phosphorus device in which both the qubit and the gates, outlined by dashed yellow lines, are fabricated by STM lithography. Courtesy of T.F. Watson from the Simmons Group, Center of Excellence for Quantum Computation and Communication Technology; see Watson *et al.* (2015). (i) A zoom into the yellow circle of the device in (h). (j) Quadruple quantum dot device in a fully depleted silicon-on-insulator (FDSOI) nanowire field-effect transistor. The gates shown are polysilicon separated by silicon nitride; electrons accumulate in corner states between the underlying silicon nanowire and an oxide surrounding them. Image courtesy Fernando Gonzalez-Zalba, Hitachi Cambridge Laboratory; see Betz *et al.* (2016). (k) TEM of a CMOS nanowire in the source/drain direction; the central features are two gates made out of TiN/polysilicon on SiO₂ and Hf-based dielectric. They are surrounded by silicon nitride spacers. See Maurand *et al.* (2016). (l) Aluminum MOS triple-dot structure. See Veldhorst *et al.* (2014). Image courtesy Andrew Dzurak, UNSW.

single charges from nearby baths, enabling initialization into a single (n,m) state. The tunneling amplitude between a pair of dots may be voltage-controlled via the energy detuning or the size of the electrostatic tunnel barrier between the dots; indeed driving voltages at different rates may enable any superposition of qubit states via voltage control and appropriately controlled Landau-Zener-Stueckelberg oscillations (Ward *et al.*, 2016; Gonzalez-Zalba *et al.*, 2016).

Charge qubits are often used to study basic properties of quantum dots (Wang *et al.*, 2013), as well as provide proving grounds for the electrostatic interactions which may provide controlled-logic gates between qubits. For example, Ward *et al.* (2016) indicates a controlled-phase gate between a pair of single-electron charge qubits defined in a Si/SiGe heterostructure, using the device layout shown in Fig. 2(a).

The disadvantage of a charge qubit is its sensitivity to charge noise. Charge noise may result from any noisy electric field, which may originate from fluctuators in the materials used in the device (e.g. $1/f$ noise from dielectrics) or from the control gates (e.g. Johnson noise from control circuitry). While the magnitude of charge noise will vary quite a bit between devices and experimental set-ups, typical dephasing times of charge qubits due to charge noise is in the range of 0.1–10 nanoseconds (Zwanenburg *et al.*, 2013). Given typical constraints of the electronics used to control qubits, these timescales are simply too fast for high fidelity operation.

Spin Qubits

A comparably simple qubit is the single-spin qubit. A single electron spin bound to a donor or to a quantum dot is naturally a two-level quantum system, corresponding to spin-up and spin-down. As a fundamental magnetic dipole, a bound spin-1/2 has no sensitivity to electric fields, and interacts only magnetically. In silicon and silicon heterostructures, conduction-band electrons have a g -factor very close to 2, indicative of the small bulk spin-orbit interaction in silicon. Spin relaxation times typically exceed seconds at reasonable magnetic fields (Zwanenburg *et al.*, 2013).

The key advantage of the single-spin implementation relative to charge-qubit implementations is access to the long coherence times as observed in bulk systems. Recently, a single phosphorus qubit with single-spin encoding, implemented in silicon isotopically enhanced to contain only 800 ppm ^{29}Si , showed a static dephasing time of 270 μs , extended to a multiple-spin-echo decoherence time exceeding 30 s (Muhonen *et al.*, 2014), despite its proximity to the aluminum gates forming the readout mechanism (see Fig. 2(g)). Similarly, a single spin qubit in an MOS dot, similar to that in Fig. 2(l), also in 800 ppm ^{29}Si , showed comparable times with static dephasing of 120 μs extended to a multiple-spin-echo decoherence time of 28 ms (Veldhorst *et al.*, 2014). Single-spin control for this system inherits the high-fidelity control of spin-resonance, enabling single-qubit gates with 99.6% fidelity as measured by randomized benchmarking (Veldhorst *et al.*, 2014).

A key consideration for single-spin encodings is that they typically require a large magnetic field, and their control requires careful tracking of each spin's Larmor precession in order to achieve phase-synchronous control. Ultimately, some limit to qubit coherence will exist due to the stability of the applied magnetic field or of the local oscillator. It is possible to circumvent these issues while still maintaining the coherence of spin qubits by using multi-spin qubits, in which qubit states are encoded by the total angular momentum of a set of spins. For example, the two-spin singlet-triplet qubit encodes information in the total angular momentum $S=0$ (singlet) and $S=1$ (triplet) subspaces of the two spins. A three-spin encoding combines a singlet-triplet pair with a third spin, assuring a total angular momentum $S=1/2$; a four-spin encoding likewise assures a total angular momentum $S=0$. These encodings are used in a way which imposes a constant total projection of angular momentum for all qubit states. In this way the qubit spin states can be brought to degeneracy for use as memory, assuring that the states do not evolve any phases relative to one another due to any constant or globally fluctuating magnetic fields. For this reason these qubits are referred to as decoherence-free subspaces or subsystems (DFS) (DiVincenzo *et al.*, 2000).

Spin Control by ESR and EDSR: To control single-spins, an obvious option is to use magnetic resonance, in which a large static applied magnetic field is present and transverse microwave magnetic fields are applied resonant with the electron Larmor precession of about 28 GHz per tesla of applied field. A clear technical difficulty with this mode of operation is that the controlling microwaves will typically not be localized, affecting multiple spins at once. To circumvent this difficulty, individual spins must have their resonant frequency shifted relative to their neighbors. In phosphorus devices, such a shift is available from the longitudinal component of the gate-controlled hyperfine interaction with the ^{31}P nucleus (Kane, 1998; Laucht *et al.*, 2015), enabling resonance shifts of order MHz for reasonable gate voltage variations. In MOS and SiGe devices, g -factor inhomogeneity has been shown to result from spin-orbit effects in the quantum dot barrier, enabling dot-to-dot and voltage-tunable g -factor variations again on the order of MHz (Kawakami *et al.*, 2014; Veldhorst *et al.*, 2015).

Even more localized microwave control is possible using electric dipole spin resonance (EDSR), in which magnetic field gradients via micromagnets or spin-orbit effects convert AC electric fields to magnetic fields. If a magnetic field gradient is present, then voltage-induced motion of the electron in that gradient may mimic the effect of transverse microwave fields, except with the clear advantage of maintaining localization to the electron undergoing spatial modulation. In SiGe devices, fabricated cobalt micromagnets have provided this gradient (Takeda *et al.*, 2016; Kawakami *et al.*, 2014), as shown in the devices of Fig. 2(c) and (e). Another possibility here is the use of hole spins, which naturally have a stronger spin-orbit interaction in confined nanostructures such as that in Fig. 2(k), enabling EDSR without additional gradients (Maurand *et al.*, 2016).

Spin Control by Kinetic Exchange: The kinetic exchange mechanism enables a conversion between charge qubits (including, unfortunately, sensitivity to charge noise) and spin-qubits. It does this by using voltages to attempt to generate a superposition of $(2,0)$ and $(1,1)$ charge states (reminiscent of the charge qubit), the amplitudes and energies of which are different for

Pauli-allowed spin singlets and Pauli-excluded spin triplets in the (2,0) state. This spin-to-charge conversion is useful since it enables initialization, control, and measurement of spin-qubits without requiring high magnetic fields or magnetic field gradients.

The action of kinetic exchange is to reduce the energy of a singlet, which may be considered a single-qubit control axis for multiple spins encoded in a DFS. For a pair of spins in a singlet-triplet qubit, this voltage-induced interaction may therefore be considered a rotation about the z -axis of the multi-spin qubit's Bloch sphere. To control other axes in this case, another mechanism is needed. Examples of successful introduction of a second interaction for the additional control axes are magnetic gradients, arising from either deliberately introduced magnetic field gradients (Takeda *et al.*, 2016), single nuclear spin (Harvey-Collard *et al.*, 2017), or an ensemble of nuclear spins (Maune *et al.*, 2012). The latter case has been a fruitful qubit implementation in the nuclear-spin-rich system of GaAs (Nichol *et al.*, 2017).

Remarkably, if DFS encodings are used employing at least three spins per qubit, the exchange interaction can provide universal quantum logic, including multi-qubit gates, with no other control mechanisms. This was shown in DiVincenzo *et al.* (2000), which shows a pulse sequence locally equivalent to controlled-phase for two three-spin encoded DFS qubits. More recently, pulse sequences have been found which provide exchange-only multiqubit control without any specification made on the m -quantum number for either qubit, potentially allowing operation in much lower magnetic fields (Fong and Wandzura, 2011). This exchange-only encoding works best in highly uniform magnetic fields, and is particularly important for silicon where nuclear-induced gradients may be made very small via the use of isotopic enhancement. The basic operation of exchange-only triple-dots qubits in 800 ppm ^{29}Si , using a gate layout similar to that in Fig. 2(d), has been demonstrated in Eng *et al.* (2015) and Reed *et al.* (2016). This encoded triple-dot spin qubit showed a ^{29}Si -limited static dephasing time of 3 μs , extended via exchange echo to over 600 μs at high magnetic field (Eng *et al.*, 2015), with operational fidelity limited by charge noise (Reed *et al.*, 2016).

Initialization and Readout of Spin Qubits: While the inductive techniques of magnetic resonance are excellent and highly mature for control, the readout-mechanisms of traditional magnetic resonance are many orders of magnitude too insensitive. One effective way to initialize and measure a single spin is via spin-selective tunneling to or from a bath. In particular, if a bath of electrons has its Fermi energy set via voltage to be between the two Zeeman sublevels of the dot spin, one spin will have available states for tunneling and the other will not. This basic principle depends on operating at a field and temperature such that the electron Zeeman energy $g\mu_B B$ vastly exceeds thermal energy $k_B T$, where the temperature here is that of the electrons in the bath (including any noise introduced from electronics, etc.). This is reasonable at fields of about a tesla, which also corresponds to convenient microwave ESR control frequencies of 3–30 GHz. This method of measurement has been employed for single-shot single-spin qubits in MOS (Veldhorst *et al.*, 2015), phosphorus (Muhonen *et al.*, 2014; Watson *et al.*, 2017), and SiGe (Kawakami *et al.*, 2014) systems.

Multi-spin encodings enable initialization and measurement using the Pauli-blockade mechanism, which enables the preparation and detection of singlets of spin pairs. This mechanism depends on distinguishing spin-dependent electron tunneling regimes, which requires electron temperatures much less than any excited state in the system. For SiGe qubits, low-lying valley states can easily prohibit this type of measurement, but nonetheless Pauli-blockade singlet-triplet measurement has been demonstrated in both SiGe and MOS dots (Zwanenburg *et al.*, 2013). Attempts to partially circumvent the effects of small valley splitting by operating in a filled shell (i.e., (3,1)–(4,0) occupation) have also been shown in coupled donor-dot and multi-dot systems (Harvey-Collard *et al.*, 2017).

Qubit Couplings and Hybridizations

Coupling of qubits is necessary for computation. Four predominant coupling mechanisms in silicon qubit device architectures are: (1) contact hyperfine between electron and nuclear spins, (2) capacitive or electric dipole coupling; (3) spin dipole coupling; and (4) kinetic exchange. Shuttling of electrons has been considered as a complementary function with coupling mechanisms for donors and quantum dots. This is especially of interest for short range mechanisms that can restrict layout flexibility such as exchange and contact hyperfine (Skinner *et al.*, 2003; Witzel *et al.*, 2015; Pica *et al.*, 2016). Shuttling of electrons through quantum dot networks has been experimentally demonstrated in GaAs (Baart *et al.*, 2016; Fujita *et al.*, 2017).

The selection of coupling mechanism has a strong influence on both the physical manifestation (e.g., layout and signal control) as well as performance (e.g., speed and dominant error sources). Coupling mechanisms, less obviously, also allow redefinition of the single qubit encoding through qubit hybridizations. Hybridizations often seek “the best of both worlds,” circumventing a challenge of one of the principal encodings through merging properties of a second qubit's properties. The invention of the transmon for Josephson-junction-based qubits is an example of the extraordinary success that can come from hybridization (Houck *et al.*, 2009).

Contact Hyperfine Interaction

The Fermi contact hyperfine interaction is ubiquitously important for both quantum dot and donor qubits because it is the dominant mechanism through which nuclear spins interact with the electron spin (i.e., either directly through a strong coupling to a ^{31}P nucleus or through overlap with trace ^{29}Si or ^{73}Ge). The contact hyperfine term, therefore, plays a role in the Hamiltonian of any qubit leading to contributions to line width or resonant frequency. In donors for which the electron spin is the primary

encoded qubit, as discussed earlier, the contact hyperfine is used as a tuning mechanism through the Stark shift (Kane, 1998; Laucht *et al.*, 2015).

Alternatively, the nuclear spin of the phosphorus atom has been proposed as either the data or memory qubit in several computing schemes (for example see Kane (1998), Skinner *et al.* (2003), Hill *et al.* (2015), Witzel *et al.* (2015), Pica *et al.* (2016)), due to several important advantages: it features extremely long coherence times, it can be isolated from the electric environment when the donor is ionized, and very high fidelity NMR rotations are possible. Furthermore, it naturally occurs with 100% abundance as the spin-1/2 ^{31}P isotope and it is a qubit that, therefore, cannot be lost (i.e., there is always a spin 1/2 system present). A strong coupling between the electron and a nucleus is “built-in” to the donor system, with a substantial amplitude of 117.53 MHz (Zwanenburg *et al.*, 2013). This provides a convenient spin-spin coupling to this long-lived quantum memory.

The Kane proposal (Kane, 1998) and many variations of it published since, have indicated device architectures using electrons to mediate interactions between donor nuclear spins through the contact hyperfine interaction, although of course all proposals further need a mechanism for donor-donor coupling either through shuttling or other mechanisms that we discuss below. It is also worthwhile to note that the hyperfine contact term has already enabled repeated quantum-non-demolition measurement of the nuclear spin, offering extremely high initialization and read-out fidelity, while magnetic resonance techniques offer excellent control fidelity (Muhonen *et al.*, 2014). As an example process exploiting these high fidelity operations, Bell inequality violations in this electron-nuclear system were recently demonstrated (Dehollain *et al.*, 2016).

Spin Qubit Coupling with Exchange

In the highly influential Loss/DiVincenzo (Loss and DiVincenzo, 1998) and Kane proposals (Kane, 1998), the kinetic exchange interaction is used as the entangling mechanism between single-spin qubits. As the name suggests, this interaction “exchanges” spins, enabling a “full swap” if tuned to a π -pulse. Spin-information can be moved across a device by a series of such swaps. Of course, if pulsed for half the duration, spins may be entangled. The sqrt-SWAP interaction between a pair of spins provides maximal entanglement; a controlled-NOT (CX) or controlled-phase (CZ) gate could then be accomplished using two sqrt-SWAP gates, with the additional use of single-spin rotations implemented via the methods described in the previous sections. Alternatively, exchange in the presence of Zeeman inhomogeneity due either to micromagnetic gradient fields or g -factor inhomogeneity can provide a CZ gate. Two-qubit gates for single spin qubits were recently demonstrated this way in MOS dots (Veldhorst *et al.*, 2015) and SiGe dots (Watson *et al.*, 2017; Zajac *et al.*, 2017). Sensitivity to charge noise was observed, providing one of the key present challenges in improving control fidelity of this type of coupling in QDs.

Exchange is a short-range interaction of order of the size of the electron wavefunctions. A possible longer term challenge for quantum dots and an immediate challenge for donors is the lithographic layout of the dense packing of electrodes needed to control the tunnel couplings, establish electron occupation, and provide readout and initialization capabilities. More complex layouts, such as extensions from present 1D layouts to 2D layouts, introduces challenges for exchange based architectures (Veldhorst *et al.*, 2016; Vandersypen *et al.*, 2017). For direct coupling of donor electron spins, the lithographic challenge is exaggerated both because of the more tightly confined electron wavefunction in the bulk and because of crystal-orientation-dependent tunnel coupling strengths. An alternate approach to couple donors through ionizing or hybridizing the donor electron to a larger surface quantum dot state (Kane, 1998; Calderon *et al.*, 2009) has been proposed to ease the lithographic requirements; recent experimental validations in this direction demonstrate coherent donor-quantum-dot hybrids (Harvey-Collard, *et al.*, 2017).

Variations of multispin and charge qubit hybridization can enable a variety of possibilities. One example is an encoding that resembles the triple-dot exchange-only qubit, but employs three electrons in two dots. An excited valley-orbit state in one of the two-dots effectively provides the third state of exchange-only control, with its energy splitting acting as an effectively constant exchange interaction. One axis of control is provided by voltage-controlled exchange-like interactions, while the other is provided by this excited-state splitting. Despite relying on what is ultimately a charge-state splitting for one axis of control, the coherence in these qubits can be extended to time scales of microseconds at the cost of substantially slowing down the qubit rotation times (Thorgrimsson *et al.*, 2017). Another example combining exchange with single spin qubits is that single spin qubit encodings can be combined with the singlet-triplet-readout methods of exchange qubits to provide direct means to make the parity measurements associated with quantum error correction (Jones *et al.*, 2016; Veldhorst *et al.*, 2016).

Capacitive and Electric Dipole Coupling

Charge qubits, as well as spin-qubit encodings with a charge qubit component, offer a natural long-range coupling scheme. The electric field created by charge displacement in one qubit can be used to control the state by displacing the charge of another qubit. At short range, this is effectively a quantum cross-capacitance effect; at larger distances, it has the character of an electric dipole-dipole coupling. Capacitive coupling has enabled multiqubit logic in silicon-based charge qubits, such as the one shown in Fig. 2(a) (Ward *et al.*, 2016). Multispin qubits may be coupled this way, for example between singlet-triplet qubits that exploit the dipole difference between the (2,0) and (1,1) charge states to influence a neighboring qubit’s charge states. The approach has been successfully demonstrated with nearest neighbor qubits reaching approximately 90% entanglement fidelity in a GaAs system (Nichol *et al.*, 2017). The distance between the qubits was approximately 500 nm in this case providing a modest relaxation of space constraints compared to exchange coupling.

Such coupling mechanisms may be especially pertinent for donor-based systems, since exchange couplings are challenging to engineer in this system. In one proposal, the electric dipole of a phosphorus impurity is “stretched” by the action of an A-gate, enabling electric control of a long-distance dipole-dipole coupling. Since this long-range coupling has a weak spatial dependence in comparison to exchange, these coupling mechanisms may allow phosphorus to be fabricated through a controlled ion implantation process, with the inevitable placement straggle compensated for by gate calibration (Tosi *et al.*, 2017). An electric dipole may also be created through the combination of a phosphorus impurity and an exchange-coupled quantum dot above it. Coherent quantum dot coupling to a single donor potential has been recently demonstrated, including characterization of the charge noise in that system (Urdampilleta *et al.*, 2015; Harvey-Collard *et al.*, 2017). Key to the success of these electric dipole coupling approaches is assuring that the mechanisms which provide the charge sensitivity for long-distance coupling may be rapidly modulated to prevent undue decoherence from charge noise.

Capacitive couplings may be enhanced or lengthened using extra hardware. Increasing the coupling range using floating gates has been proposed (Trifunovic *et al.*, 2012), although the small predicted effects of dissipation in these gates awaits experimental validation. Superconducting coupling elements seem especially promising, since the coupling of charge qubit components of qubits and the management of charge noise have been well addressed in the superconducting qubit community. These ideas are beginning to be adopted by silicon qubits, and superconductors may offer a powerful spin-charge hybridization offering benefits such as longer-range coupling. Conversion of exchange qubits to charge qubits allow charge-induced coupling to microwave fields in superconducting resonators, providing rapid spin-spin interactions across a chip; the recently demonstrated strong coupling of a single electron charge qubit in a SiGe device to a superconducting resonator is a critical step in this direction (Mi *et al.*, 2017).

Electric dipoles with optical oscillation frequencies could enable optical connections between silicon qubits, which would be especially convenient for those applications in optical quantum communication requiring memory and quantum logic. Unfortunately, silicon’s indirect bandgap drastically reduces this system’s efficiency for most existing concepts for spin-photon entanglement employing near-bandgap excitons, as demonstrated in III-V semiconductors. However, optically efficient emitters do exist in silicon and may provide future qubits with practical optical interfaces; a recent proposal employing chalcogen donors is provided in Morse *et al.* (2017).

Magnetic Dipole Coupling

One of the potential advantages of single spin encodings is the long decoherence times due to the relatively good decoupling from environmental factors like charge noise. One approach to fully realize the long coherence times of spins is to altogether avoid all coupling schemes that have a charge qubit component. The spin is a magnetic dipole and the magnetic dipole-dipole interaction therefore provides a mechanism of coupling qubits not subject to charge noise. A critical challenge in exploiting dipole-dipole couplings is that they are long-range and “always on,” making scheduled control challenging. Also they are slow, typically requiring millisecond interaction times. Nonetheless, this mechanism is appealing to donors because the dipole-dipole coupling avoids the atomic precision fabrication requirement for exchange and a system may exploit the long memory times of ^{31}P nuclear spins. An early proposal for dipole coupled donor quantum computing argues that using a long range “always on” component is manageable combined with g-factor tuning (de Sousa *et al.*, 2004). The Stark shift of the donor contact hyperfine and g-factor has been modelled (Rahman *et al.*, 2009) and Stark shifted resonant frequency have been experimentally demonstrated both in ensembles and single donor cases (Wolfowicz *et al.*, 2014; Laucht *et al.*, 2015). A more recent architecture proposes a more direct modulation of the dipolar interaction through use of precisely timed ionization and deionization of donors to control the magnetic dipole-dipole coupling, holding information on the associated ^{31}P nuclear spins (Hill *et al.*, 2015). This can substantially reduce the complication of long range coupling. In another recent proposal, mechanical motion of a scanning stage of ^{31}P modulates the dipole-dipole coupling strengths (O’Gorman *et al.*, 2016). Although these architectures no longer require atomic-scale placement and gating, they still depend on highly uniform energy levels for a regular array of gated donors and the two qubit rotations are slower because of the weaker interaction strengths.

The Future of Silicon Qubit Systems

In this article, we have outlined three principal ways of fabricating single-electron silicon systems, and three principal ways of encoding qubits on those systems. These fabrication and encoding techniques can furthermore be hybridized and combined producing improved features. Obviously, a large amount of design space exists for constructing systems of silicon qubits, and it remains an active, worldwide area of research to optimize the many trade-offs between the many possibilities. At the time of writing, the largest published coherently coupled silicon qubit system contains only two entangled qubits, leaving a long road ahead to the size and scale required for fault-tolerant quantum computing. Many concepts have been proposed for ways to scale, although many fundamental demonstrations remain to be proven prior to assurance that any such scaling path will succeed. However, the incredible scaling success of classical silicon microprocessors based on CMOS provides continual and dramatic encouragement of the notion that once the basic problems of charge-noise limits in control fidelities and valley- or disorder-limits in device yield are solved, a viable path for scaling to a useful technology will be found.

Acknowledgement

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

See also: Cavity QED

References

- Baart, T.A., *et al.*, 2016. Single-spin CCD. *Nat. Nanotechnol.* 11, 330.
- Betz, A.C., *et al.*, 2016. Reconfigurable quadruple quantum dots in a silicon nanowire transistor. *Appl. Phys. Lett.* 108, 203108.
- Borselli, M.G., *et al.*, 2011. Pauli spin blockade in undoped Si/SiGe two-electron double quantum dots. *Appl. Phys. Lett.* 99, 063109.
- Calderon, M.J., Saraiva, A., Koiller, B., Das Sarma, S., 2009. Quantum control and manipulation of donor electrons in Si-based quantum computing. *J. Appl. Phys.* 105, 122410.
- Dehollain, J.P., *et al.*, 2016. Bell's inequality violation with spins in silicon. *Nat. Nanotechnol.* 11, 242.
- de Sousa, R., *et al.*, 2004. Silicon quantum computation on magnetic dipolar coupling. *Phys. Rev. A* 70, 052304.
- DiVincenzo, D.P., *et al.*, 2000. Universal quantum computation with the exchange interaction. *Nature* 408, 339.
- Eng, K., *et al.*, 2015. Isotopically enhanced triple-quantum-dot qubit. *Sci. Adv.* 1, e1500214.
- Fong, B.H., Wandzura, S.M., 2011. Universal quantum computation and leakage reduction in the 3-qubit decoherence free subsystem. *Quantum. Inf. Comput.* 11, 1003–1018.
- Freeman, B.M., Schoenfeld, J.S., Jiang, H.W., 2016. Comparison of low frequency charge noise in identically patterned Si/SiO₂ and Si/SiGe quantum dots. *Appl. Phys. Lett.* 108, 253108.
- Fuechsle, M., *et al.*, 2012. A single atom transistor. *Nat. Nanotechnol.* 7, 242.
- Fujita, T., Baart, T.A., Reichl, C., Wegscheider, W., Vandersypen, L.M.K., 2017. Coherent shuttle of electron-spin states. *npj Quantum Inf.* 3, 22.
- Gamble, J.K., *et al.*, 2016. Valley splitting of single-electron Si MOS quantum dots. *Appl. Phys. Lett.* 109, 253101.
- Gonzalez-Zalba, M.F., Barraud, S., Ferguson, A.J., Betz, A.C., 2015. Probing the limits of gate-based charge sensing. *Nat. Commun.* 6, 6084.
- Gonzalez-Zalba, M.F., *et al.*, 2016. Gate-sensing coherent charge oscillations in a silicon field-effect transistor. *Nano Lett.* 16, 1614.
- Harvey-Collard, P., *et al.*, 2017. Coherent coupling between a quantum dot and a donor in silicon. *Nat. Commun.* 8, 1029.
- Hill, C.D., *et al.*, 2015. A surface code quantum computer in silicon. *Sci. Adv.* 1, e1500707.
- Hollenberg, L., *et al.*, 2006. Two-dimensional architectures for donor-based quantum computing. *Phys. Rev. B* 74, 045311.
- Houck, A.A., Koch, J., Devoret, M.H., Girvin, S.M., Schoelkopf, R.J., 2009. Life after charge noise: Recent results with transmon qubits. *Quantum Inf. Process* 8, 105.
- Jock, R.M., *et al.*, 2017. Probing low noise at the MOS interface with a spin-orbit qubit. *arXiv:1707.04357*.
- Jones, C., *et al.*, 2016. A logical qubit in a linear array of semiconductor quantum dots. *arXiv:1608.06335*.
- Kane, B.E., 1998. A silicon-based nuclear spin quantum computer. *Nature* 393, 133.
- Kawakami, E., *et al.*, 2014. Electrical control of a long-lived spin qubit in a Si/SiGe quantum dot. *Nat. Nanotechnol.* 9, 666.
- Laucht, A., *et al.*, 2015. Electrically controlling single-spin qubits in a continuous microwave field. *Sci. Adv.* 1, e1500022.
- Litvinenko, K.L., *et al.*, 2015. Coherent creation and destruction of orbital wavepackets in Si:P with electrical and optical read-out. *Nat. Commun.* 6, 6549.
- Loss, D., DiVincenzo, D.P., 1998. Quantum computation with quantum dots. *Phys. Rev. A* 57, 120.
- Maune, B.M., *et al.*, 2012. Coherent singlet-triplet oscillations in a silicon-based double quantum dot. *Nature* 481, 344.
- Maurand, R., *et al.*, 2016. A CMOS silicon spin qubit. *Nat. Commun.* 7, 13575.
- Mi, X., *et al.*, 2017. Strong coupling of a single electron in silicon to a microwave photon. *Science* 355, 156.
- Morse, K.J., 2017. A photonic platform for donor spin qubits in silicon. *Sci. Adv.* 3, e1700930.
- Muhonen, J.T., *et al.*, 2014. Storing quantum information for 30 seconds in a nanoelectronic device. *Nat. Nanotechnol.* 9, 986.
- Nichol, J.M., *et al.*, 2017. High-fidelity entangling gate for double-quantum-dot spin qubits. *npj Quantum Inf.* 3, 3.
- O'Gorman, J., *et al.*, 2016. A silicon-based surface code quantum computer. *npj Quantum Inf.* 2, 15019.
- Pica, G., *et al.*, 2016. Surface code architecture for donors and dots in silicon with imprecise and nonuniform qubit couplings. *Phys. Rev. B* 93, 034306.
- Rahman, R., *et al.*, 2009. Gate-induced g-factor control and dimensional transition for donors in multivalley semiconductors. *Phys. Rev. B* 80, 155301.
- Reed, M.D., *et al.*, 2016. Reduced sensitivity to charge noise in semiconductor spin qubits via symmetric operation. *Phys. Rev. Lett.* 116, 110402.
- Saeedi, K., *et al.*, 2013. Room-temperature quantum bit storage exceeding 39 minutes using ionized donors in silicon-28. *Science* 342, 830.
- Shamim, S., *et al.*, 2016. Ultralow-noise atomic-scale structures for quantum circuitry in silicon. *Nano Lett.* 16, 5779.
- Skinner, A.J., Davenport, M.E., Kane, B.E., 2003. Hydrogenic spin quantum computing in silicon: A digital approach. *Phys. Rev. Lett.* 90, 087901.
- Takeda, K., *et al.*, 2016. A fault-tolerant addressable spin qubit in a natural silicon quantum dot. *Sci. Adv.* 2, e1600694.
- Testolin, M.J., *et al.*, 2007. Robust controlled-NOT gate in the presence of large fabrication-induced variations of the exchange interaction strength. *Phys. Rev. A* 76, 012302.
- Thorgrimsson, B., *et al.*, 2017. Extending the coherence of a quantum dot hybrid qubit. *npj Quantum Inf.* 3, 32.
- Tosi, G., *et al.*, 2015. Silicon quantum processor with robust long-distance qubit couplings. *Nat. Commun.* 8, 450.
- Trifunovic, L., *et al.*, 2012. Long-distance spin-spin coupling via floating gates. *Phys. Rev. X* 2, 011006.
- Tyryshkin, A.M., *et al.*, 2012. Electron spin coherence exceeding seconds in high-purity silicon. *Nat. Mater.* 11, 143.
- Urdampilleta, R.M., *et al.*, 2015. Charge dynamics and spin blockade in a dybrid double quantum dot in silicon. *Phys. Rev. X* 5, 031024.
- Vandersypen, L.M.K., *et al.*, 2016. Interfacing spin qubits in quantum dots and donors – hot, dense and coherent. *npj Quantum Inf.* 3, 34.
- Veldhorst, M., *et al.*, 2014. An addressable quantum dot qubit with fault-tolerant control-fidelity. *Nat. Nanotechnol.* 9, 981.
- Veldhorst, M., *et al.*, 2015. A two-qubit logic gate in silicon. *Nature* 526, 410.
- Veldhorst, M., Eerink, H.G.J., Yang, C.H., Dzurak, A.S., 2016. Silicon CMOS architecture for a spin-based quantum computer. *arXiv:1609.09700*.
- Wang, K., Payette, C., Dovzhenko, Y., Deelman, P.W., Petta, J.R., 2013. Charge relaxation in a single electron Si/SiGe double quantum dot. *Phys. Rev. Lett.* 111, 046801.
- Ward, D.R., *et al.*, 2016. State-conditional coherent charge qubit oscillations in a Si/SiGe quadruple quantum dot. *npj Quantum Inf.* 2, 16032.
- Watson, T.F., *et al.*, 2015. High-fidelity rapid initialization and read-out of an electron spin via the single donor d-charge state. *Phys. Rev. Lett.* 115, 166806.
- Watson, T.F., *et al.*, 2017. A programmable two-qubit quantum processor in silicon. *arXiv:1708.04214*.
- Witzel, W.M., Montaño, I., Muller, R.P., Carroll, M.S., 2015. Multi-qubit gates protected by adiabaticity and dynamical decoupling applicable to donor qubits in silicon. *Phys. Rev. B* 92, 081407(R).
- Wolfowicz, G., *et al.*, 2014. Conditional control of donor nuclear spins in silicon using Stark shifts. *Phys. Rev. Lett.* 113, 157601.

- Yang, C.H., Rossi, A., Ruskov, R., *et al.*, 2013. Spin-valley lifetimes in a silicon quantum dot with tunable valley splitting. *Nat. Commun.* 4, 2069.
- Zajac, D.M., Hazard, T.M., Mi, X., Wang, K., Petta, J.R., 2015. A reconfigurable gate architecture for Si/SiGe quantum dots. *Appl. Phys. Lett.* 106, 223507.
- Zwanenburg, F.A., *et al.*, 2013. Silicon quantum electronics. *Rev. Mod. Phys.* 85, 961.
- Zajac D.M., *et al.*, 2017. Quantum CNOT gate for spins in silicon. [arXiv:1708.03530](https://arxiv.org/abs/1708.03530).

Entanglement and Quantum Information

PG Kwiat, University of Illinois at Urbana-Champaign, Urbana, IL, USA

DFV James, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2018 Elsevier Inc. All rights reserved.

Glossary

Entanglement a property of quantum systems consisting of two or more distinct subsystems, often separate particles. When transformations on one system do not affect the other, the state is said to be separable; when this is not the case, and the quantum state of the overall system cannot be resolved into separate states of the individual pieces, the system is entangled.

Pure and mixed states when the probability amplitudes specifying a quantum state are deterministic, the state is said to be pure; when the amplitudes are random quantities, the state is mixed. The distinction is similar to that of coherent and partially coherent optical fields.

Quantum computer an information processing device that exploits quantum mechanical phenomena to greatly enhance computational power for certain problems. Some mathematical problems thought to be intractable on conventional computers can, in theory, be performed efficiently on a quantum computer.

Quantum cryptography (also known as quantum key distribution) a quantum information protocol by which two parties acquire a shared series of random numbers (a cryptographic key) by exploiting the quantum nature of

light. Absolute security can be proven by virtue of the indivisibility and uncopy-ability of individual quanta.

Quantum state tomography a means by which quantum states can be determined experimentally by performing a series of appropriate measurements on multiple identically prepared systems. From such measurements, the elements of the density matrix, which fully specifies the state, may be inferred.

Quantum superdense coding a quantum information protocol by which two bits of classical information may be communicated by a single quantum bit, initially part of an appropriate entangled quantum state.

Quantum teleportation a quantum information protocol by which the unknown quantum state of one particle can be transferred to another distant particle, using a pair of entangled particles, a projective measurement, and exchange of two bits of classical information.

Qubit (quantum bit) a two-level quantum mechanical system, which constitutes the building blocks of quantum information processing devices. Examples include a spin-1/2 particle, an atom with two well-isolated levels, or the polarization of a photon.

Quantum information is an emerging field of technology that encompasses the application of fundamental quantum mechanical phenomena, such as entanglement, to tasks in information processing and communications. The importance of optical technology in quantum information is not surprising, given the success of quantum optics in the investigation of these fundamental phenomena and the pre-eminent role of optics in modern communications. This article describes some of the important advances in photonic quantum information and discusses the prospects for the future.

Basic Notions: Qubits and Entanglement

Qubits

Quantum information is usually confined to two-level quantum systems, which are referred to as quantum bits, or qubits. The qubits form the register of a quantum computer or carry the information in a quantum communications channel. Physical examples of qubits under active study include the spin degrees of freedom of an electron or spin-1/2 nucleus, two well-isolated energy levels of an atom or trapped ion, and the polarization degrees of freedom of a photon; here we concentrate on the latter (Recently some attention has been devoted to quantum information based on continuous variables, such as those arising in connection with 'squeezed light'. However, here we will focus on discrete systems, i.e., qubits.). The two states of each qubit are generically denoted by $|0\rangle$ and $|1\rangle$ (although, because of our concentration on optical polarization, we shall use the horizontal and vertical states $|H\rangle$ and $|V\rangle$ interchangeably with $|0\rangle$ and $|1\rangle$ when describing experiments). An arbitrary state of a single qubit, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, is specified by the complex probability amplitudes α and β associated with the two possible states, constrained by $|\alpha|^2 + |\beta|^2 = 1$. States for which α and β are deterministic quantities are referred to as pure states; when they are stochastic, the state is mixed, and must be described by a density matrix, e.g., $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, where p_i is the probability of the pure state $|\psi_i\rangle$. Pure states are a useful but unattainable ideal, just like a perfectly coherent laser beam; in reality even the best-prepared quantum states have some level of mixedness, which can be quantified by the entropy: $S \equiv -\text{Tr}(\rho \ln \rho) = -\sum_i \lambda_i \ln \lambda_i$, where λ_i are the eigenvalues of the matrix ρ .

Multiqubit States and Entanglement

For two qubits, the possible pure states of the system are specified by four probability amplitudes: (Qubits are assumed to be distinguishable particles, and so the wavefunctions describing their states need not be symmetrized under exchange of particles.) $|\psi_2\rangle = \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$, with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$, and, e.g., $|00\rangle \equiv |0\rangle \otimes |0\rangle$ denoting the state in which the first and the second qubits are both in state $|0\rangle$. In general, 2^n amplitudes are needed to characterize a pure state of n qubits. Immediately we can see one advantage of quantum systems for information storage: if each distinct probability amplitude is regarded as a data register, the size of the memory grows exponentially with the number of qubits. Further, if one were to flip just one qubit, for example the second, the state would be transformed into $|\psi'_2\rangle = \beta|00\rangle + \alpha|01\rangle + \delta|10\rangle + \gamma|11\rangle$, where all of the amplitudes have been affected. This simple example demonstrates the notion of quantum parallelism, one of the most powerful properties of quantum information processors.

If the two-particle state can be written as a product of two single-particle states, i.e., if $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle = (A|0\rangle + B|1\rangle) \otimes (C|0\rangle + D|1\rangle)$, then these two particles can be considered as separate, unconnected entities: the state is said to be separable (When mixture is also present, the state is separable if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes \rho_i^B$, i.e., as a sum of product states for systems A and B.). When the state cannot be written in this form, it is called an entangled state. Entanglement is one of the most fascinating fundamental properties of quantum systems: Erwin Schrödinger described it as ‘the characteristic trait of quantum mechanics ... that enforces its entire departure from classical lines of thought.’ Consider performing a measurement on the second of the two qubits, to establish which of its two states the particle was in, leaving the first qubit untouched. The result would collapse the state of the first qubit into one of two possibilities: either $|\phi_0\rangle = (\alpha|0\rangle + \gamma|1\rangle)/\sqrt{P_0}$ if the outcome of the measurement were 0 (which occurs with probability $P_0 = |\alpha|^2 + |\gamma|^2$) or $|\phi_1\rangle = (\beta|0\rangle + \delta|1\rangle)/\sqrt{P_1}$ if it were 1 (with probability $P_1 = |\beta|^2 + |\delta|^2$). Thus the state of the first qubit is instantaneously projected into a specific state by performing a measurement on the second qubit, which can be in a completely different location. This example demonstrates the nonlocality of quantum mechanics, which has been confirmed experimentally in some depth by a number of elegant quantum optics experiments over the last 30 years.

The fidelity $F = |\langle\Phi|\Psi\rangle|^2$ characterizes the overlap of two states (The generalized fidelity between mixed states ρ_A and ρ_B is $F = |\text{Tr}\sqrt{\sqrt{\rho_A}\rho_B\sqrt{\rho_A}}|^2$). The fidelity of $|\phi_0\rangle$ and $|\phi_1\rangle$ is a way to quantify the amount of entanglement in the initial two-qubit state. If the final two states are identical, the fidelity will be unity, and the initial state is separable; conversely if the two final states are orthogonal and occur with equal probability, then in a sense the initial state is maximally entangled, since the largest possible nonlocal influence would occur due to the measurement. Mathematically, the fidelity of the two final states is $|\langle\phi_0|\phi_1\rangle|^2 = 1 - C^2/4P_0P_1$, where $C = 2|\alpha\delta - \beta\gamma|$ is called the concurrence of the state. If $C=0$, the state is separable; if $C=1$, the state is maximally entangled (The concurrence can also be generalized to mixed states, although it has a much more complicated expression. The quantification of entanglement for mixed states with more than two subsystems is presently an active area of research.).

As discussed below, entanglement forms the heart of a number of quantum information protocols, such as dense coding, teleportation, and one type of quantum key distribution (also known as quantum cryptography), and large-scale entangled states of many qubits seem to be a requirement for the more ambitious goal of practical quantum computing.

Creating Entangled States Experimentally

Entangled states can currently be created in a controlled manner using technologies such as ion traps, cavity quantum electrodynamics, and optical spontaneous parametric down-conversion (SPDC). Down-conversion is a nonlinear optical process by which an incident ‘pump’ photon can be split (or down-converted) into a pair of longer-wavelength daughter photons (historically called ‘signal’ and ‘idler’) in a crystal possessing a $\chi^{(2)}$ nonlinearity, such as beta-barium borate (BBO). Mathematically, the process is described using the creation and annihilation operators of the field modes ($\hat{a}_s^\dagger, \hat{a}_s$), thus:

$$|\Phi_{\text{out}}\rangle = \exp[-i \sum_{p,s,i} c_{p,s,i} (\hat{a}_p^\dagger \hat{a}_s \hat{a}_i + \hat{a}_i^\dagger \hat{a}_s^\dagger \hat{a}_p)] |\Phi_{\text{in}}\rangle$$

where the triple sum is over all field modes (subscripts p , s , and i refer to pump, signal, and idler modes, respectively) and $c_{p,s,i}$ is a coefficient linearly dependent on the second-order nonlinear susceptibility $\chi_{ijk}^{(2)}$ and also on the birefringent properties of the crystal. Further, $c_{p,s,i}$ will be negligibly small unless both total photon energy and momentum are conserved (i.e., $\omega_p = \omega_s + \omega_i$ and $\vec{k}_p = \vec{k}_s + \vec{k}_i$, where $\vec{k}$ is the intracrystal momentum).

One particularly efficient method for using this phenomenon to create entangled states is as follows. For a specific geometry (type-I phase matching), the daughter photons emerge from the crystal with identical polarizations (perpendicular to the parent polarization and the crystal optic axis) on opposite sides of a cone that is centered about the pump beam (Fig. 1(a)). Because each photon is in a definite state of polarization, the two photons are not in an entangled state. Two crystals, aligned with their axes of symmetry oriented at 90° to each other, as shown in Fig. 1(b), can then be used to create an entangled state (Another widely used method employs type-II phase matching in a single crystal. The down-conversion photons are emitted with perpendicular polarizations along a pair of cones. Pairs emitted along particular directions are in the maximally polarization-entangled state $|\psi_{-}\rangle = \frac{1}{\sqrt{2(|HV\rangle - |HV\rangle)}}$). Such a source was used in the first demonstrations of superdense coding and quantum teleportation.). With crossed crystals, two processes are possible: the parent photon can down-convert in the first crystal to yield two vertically polarized photons, or it can down-convert in the second crystal to yield two horizontally polarized photons. Because it is impossible to

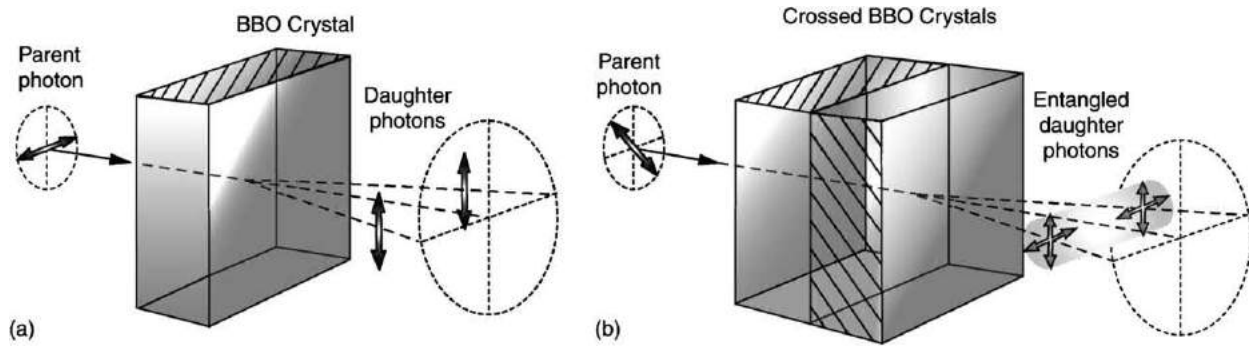


Fig. 1 An entangled-photon source. (a) For a given orientation of the nonlinear crystal, a horizontally polarized parent photon produces a pair of vertically polarized daughters. The daughters emerge on opposite sides of an imaginary cone. The cone's axis is parallel to the original direction taken by the parent photon. The two daughter photons are not in an entangled state. Reorienting the BBO crystal by 90° will produce a pair of horizontally polarized daughters if a vertically polarized pump beam is used. (b) Passing a photon polarized at 45° through two crossed BBO crystals can produce two photons in an entangled state. Because of the Heisenberg uncertainty principle, there is no way to tell in which crystal the parent photon 'gave birth,' and so a coherent superposition of two possible outcomes results: a pair of vertically polarized photons or a pair of horizontally polarized photons. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

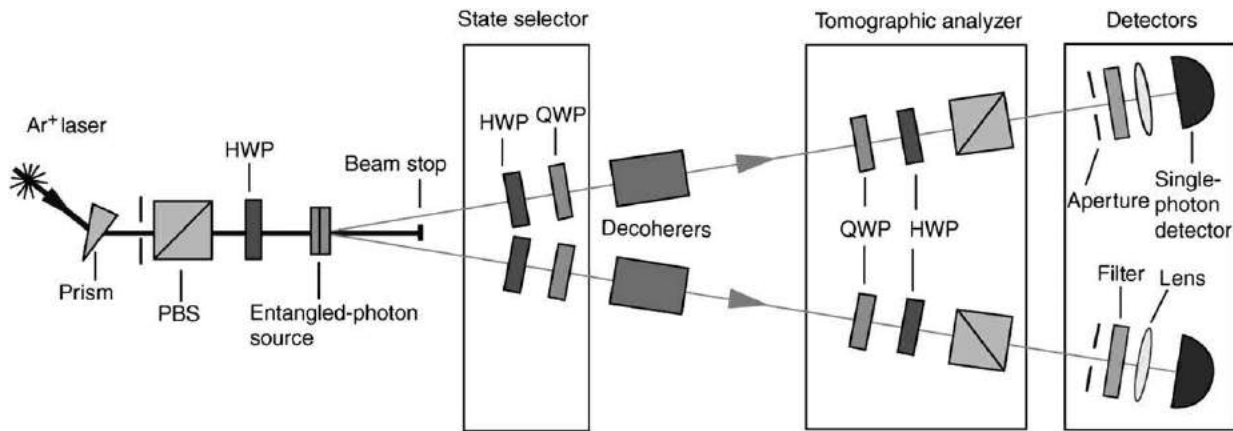


Fig. 2 Creating and measuring two-photon entangled states. The 'pump' photons are created, for example, in an argon ion laser and are linearly polarized with a polarizing beamsplitter (PBS). The half-wave plate (HWP) rotates the polarization state before the photon enters the pair of nonlinear crystals that constitute the entangled-photon source; the initial angle of pump polarization controls the entanglement of the pair produced. Each photon's polarization state can be altered at will by the subsequent HWP and quarter-wave plate (QWP). The decoherers following the state selection allow the production of mixed photon states. The optical elements (QWP, HWP, and PBS) in the tomographic analyzer allow the measurement of each photon in an arbitrary basis. Coincidence measurements of photons allow the quantum state to be determined. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

distinguish which of these processes has occurred, the state of the daughter photons is a coherent quantum-mechanical superposition of the states that would arise from each crystal alone; the output of the crossed crystals is photons in the maximally entangled state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle)$. (Note that in an arbitrary basis $|\theta\rangle \equiv \cos\theta|H\rangle + \sin\theta|V\rangle$ and $|\theta^\perp\rangle \equiv -\sin\theta|H\rangle + \cos\theta|V\rangle$, this maximally entangled state has the form $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|\theta\theta\rangle + |\theta^\perp\theta^\perp\rangle)$, demonstrating that the nonlocal correlations are present regardless of the bases used to represent the state).

Fig. 2 shows how this basic source can be adapted to produce any pure quantum state of two photons by placing rotatable half- and quarter-wave plates (which can be used to transform the polarization state of a single photon) before the crystal pair and in the paths of the two daughter photons. To create mixed states, a long birefringent crystal can be used to delay one polarization component with respect to the other. If the delay is longer than the coherence time of the photons, the horizontal and vertical components are effectively decohered; that is, the phase relationship between the different states is destroyed.

The figure also shows schematically the apparatus required for measuring the quantum state. In classical optics, four Stokes parameters are required to specify the polarization of a single beam (i.e., an ensemble of uncorrelated photons); for a pair of photons, 16 projective measurements, each with different wave plate settings, is required. From these 16 measurements, all of the elements of the 4×4 density matrix describing the (in general, mixed) state of the photon pairs can be deduced. This is an example of quantum state tomography, a technique that has found application to a number of quantum optical systems.

Quantum Key Distribution

Two parties, historically known as Alice and Bob, want to have a secret conversation. A generic, classical encryption protocol would begin when Alice and Bob convert their messages to separate binary streams of 0's and 1's, which are then encrypted and decrypted with a set of secret 'keys' known only to the two. Each key is a random string of 0's and 1's that is as long as the binary string comprising each message. To encrypt, Alice (the sender) sequentially adds each bit of the key to each bit of her message, using addition modulo 2. She then sends the encrypted message over a public channel to Bob, who decrypts it by simply repeating the addition modulo 2 of the key to the message. This type of encryption, known as a one-time pad, is currently the only provably secure encryption protocol. But the one-time pad is effective only if Alice and Bob never reuse the key, and more obviously, if the key remains secret. A potential eavesdropper, Eve, cannot be allowed to glean any part of the bit stream that makes up the key. Therein lies a central problem of cryptography – how can secret keys be created and then securely distributed?

Quantum key distribution (QKD) exploits the fundamentally indivisible nature of photons to perform this task. There are a variety of QKD protocols; here we describe one that employs entangled photon pairs (see Fig. 3). Alice and Bob use a source such as the one described above to produce maximally entangled photons in the state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle) = \frac{1}{\sqrt{2}}(|45^\circ, 45^\circ\rangle + |-45^\circ, -45^\circ\rangle)$. One photon goes to Alice and the other to Bob. For each pair, Alice and Bob randomly and independently analyze their respective photons (using a polarizing beamsplitter) in the H/V or the $45^\circ/-45^\circ$ basis. They record a bit value of 0 for all H or 45° results, and a 1 for all V or -45° results. After a sufficient number of measurements (dictated by the length of the key), Alice and Bob have a public discussion, e.g., over the internet. For each detected photon, they announce which basis they used for the measurement, but not the actual measurement result. Whenever they made the same basis choice (50% of the time), the correlations of the entangled state ensure their measured bit values agree. By contrast, they discard the results when they used different bases, because their measurements are completely uncorrelated (see Table 1).

An eavesdropper (Eve) cannot tap the line, as she might with conventional communications, because of the indivisibility of individual photons and the fact that arbitrary quantum systems cannot be accurately cloned (The no-cloning proof is very simple. If we have a copying operation such that $|0\rangle|c\rangle \rightarrow |0\rangle|0\rangle$ and $|1\rangle|c\rangle \rightarrow |1\rangle|1\rangle$ ($|c\rangle$ is the initial state of the copier), then by linearity, a superposition becomes an entangled state, instead of two copies: $\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}|c\rangle \rightarrow \frac{(|0\rangle|0\rangle+|1\rangle|1\rangle)}{\sqrt{2}} \neq \frac{(|0\rangle+|1\rangle)}{\sqrt{2}}\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}$). If Eve steals Bob's photon (a 'denial-of-service' attack), the photon's information never becomes part of the key. Thus, although a wiretap would reduce the rate of the transmission, it would not jeopardize the security of the key. Eve can try to intercept the photon, measure it, and send another one to Bob. But any measurement Eve would make to determine the photon's polarization state would perturb the photon and collapse the entangled state. The photon she sends to Bob would therefore only be 'classically' correlated with Alice's photon. Consequently, Eve's intervention necessarily induces additional errors into Bob's key, which Alice and Bob can detect by

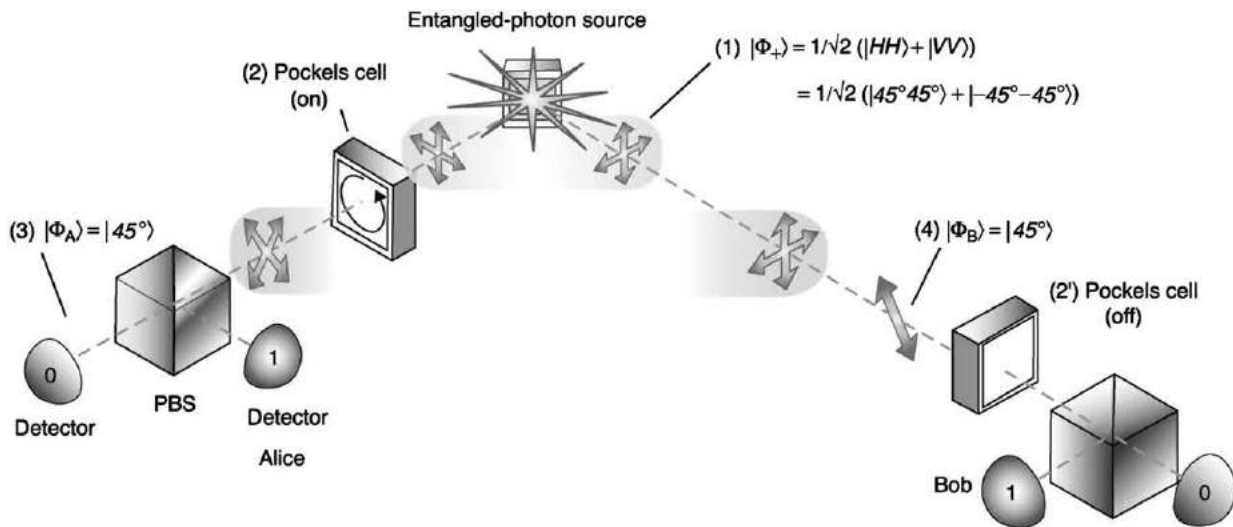


Fig. 3 Quantum cryptography using entangled photons. Entangled photons can be used to create a pair of identical cryptographic keys. One member of an entangled pair (1) is sent to Alice, and the other to Bob. Alice and Bob each randomly and independently analyze their respective photon in one of two linear polarization bases with a polarizing beamsplitter (PBS); the basis can be actively chosen using a Pockels cell before the PBS, as shown, to rotate the polarization (alternatively, the 'choice' could be made by directing the photon onto a nonpolarizing 50/50 beamsplitter; in the transmitted path, one analysis basis is used, in the reflected path, the other is used). In the example shown, Alice used the $45^\circ/-45^\circ$ basis (2), and measured 45° polarization (3), thus projecting Bob's photon into the identical state (4). Since he chose the H/V analysis basis (2'), he is equally likely to detect a 0 or a 1. By subsequent public discussions Alice and Bob determine the events for which they used the same analysis basis (and discard the other events). For these events Alice and Bob will have obtained identical measurement results, which they may interpret as raw key material. After classical error correction and privacy amplification techniques are applied, the remaining string is the sought-after shared secret key. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

Table 1 Polarization entanglement-based quantum cryptography protocol

The analysis basis	H/V	H/V	45°/–45°	H/V	45°/–45°	45°/–45°	H/V	45°/–45°
Alice's measurement result ^a	<i>H</i>	–	–45°	<i>V</i>	45°	45°	<i>V</i>	–45°
Alice's bit value	0	–	1	1	0	0	1	1
Bob's analysis basis	H/V	45°/–45°	45°/–45°	45°/–45°	H/V	45°/–45°	H/V	HV
Bob's measurement result	<i>H</i>	–45°	–45°	45°	<i>H</i>	–	<i>V</i>	<i>H</i>
Bob's bit value	0	1	1	0	0	–	1	0
Public discussion: Both photons detected and same basis used?	yes	no	yes	no	no	no	yes	no
Remaining secret key	0		1				1	

^aIn some cases Alice or Bob may not detect a photon due to loss or detector inefficiency. These events simply do not contribute to the key data.

Table 2 Method to convert the initial state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ into any of the four Bell states

<i>Alice's operation</i>	<i>Polarization transformation</i>	<i>Resultant two-qubit state</i>
$\hat{I}$, the identity	$H \rightarrow H; V \rightarrow V$	$ \Phi_+\rangle = \frac{1}{\sqrt{2}}(00\rangle + 11\rangle)$
$\hat{\sigma}_x$	$H \rightarrow V; V \rightarrow H$	$ \Psi_+\rangle = \frac{1}{\sqrt{2}}(01\rangle + 10\rangle)$
$i\hat{\sigma}_y$	$H \rightarrow V; V \rightarrow -H$	$ \Psi_-\rangle = \frac{1}{\sqrt{2}}(01\rangle - 10\rangle)$
$\hat{\sigma}_z$	$H \rightarrow H; V \rightarrow -V$	$ \Phi_-\rangle = \frac{1}{\sqrt{2}}(00\rangle - 11\rangle)$

publicly revealing a small subset of their actual key. Unfortunately, even with no eavesdropper, the encryption keys created by any real-world quantum cryptography system typically possess a few-percent errors. To make sure their key is secure, Alice and Bob ascribe all errors to Eve and then estimate the maximum amount of information available to the eavesdropper. They then use a classical privacy amplification protocol to reduce Eve's knowledge of the secret key to less than one bit by reducing the length of the key. It has been proven that if the initial error probability per bit is greater than $\sim 15\%$, no secret bits will remain after error detection and privacy amplification.

Researchers are working to make entanglement-based quantum cryptography more practical. Also, a number of longer distance quantum cryptography demonstrations using weak pulses have been carried out, over tens of kilometers in fibers and in free space. In fact, the first commercially available systems have recently been announced (in Europe and the United States).

Quantum Superdense Coding

It is possible for Alice to send Bob two bits of classical information using a single qubit in the quantum superdense coding protocol. Suppose that Alice and Bob share one qubit each of an entangled pair in the maximally entangled state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, where we have returned to the generic labeling scheme for the qubit states. Alice has two classical bits of information, which is equivalent to one of four choices. She can encode this information by applying one of four possible transformations on her qubit; a suitable set of transformations are the three Pauli matrices and the identity (i.e., do nothing). This set of operations can be performed experimentally on photon qubits quite easily, e.g., using waveplates, as indicated in [Table 2](#). The four resultant two-qubit states are all maximally entangled and are orthonormal. They form a special basis for the two-qubit states, the Bell basis, which is of particular use in analyzing entanglement.

Alice now sends her qubit to Bob, e.g., through an optical fiber. Bob can then perform a Bell state analysis (Complete discrimination of all four Bell states is presently an unsolved technical problem: for photons one needs either a nonlinear interaction (which is typically very weak) or to exploit so-called 'hyper-entanglement' involving other entangled degrees of freedom of the photons.). on his qubit pair, i.e., a projective measurement of the two-qubit state in the Bell basis. The result immediately reveals the choice of operation Alice made, and in effect, two bits of classical information have been encoded on a single qubit.

Quantum Teleportation

Another application of entanglement is quantum state teleportation ([Fig. 4](#)), in which the infinite amount of information contained in an arbitrary qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is transferred by communication of two bits of classical information between Alice and Bob, if they also share two qubits in the maximally entangled state $|\Psi_-\rangle$. The three-photon initial state (i.e., the input photon plus the two entangled photons) is

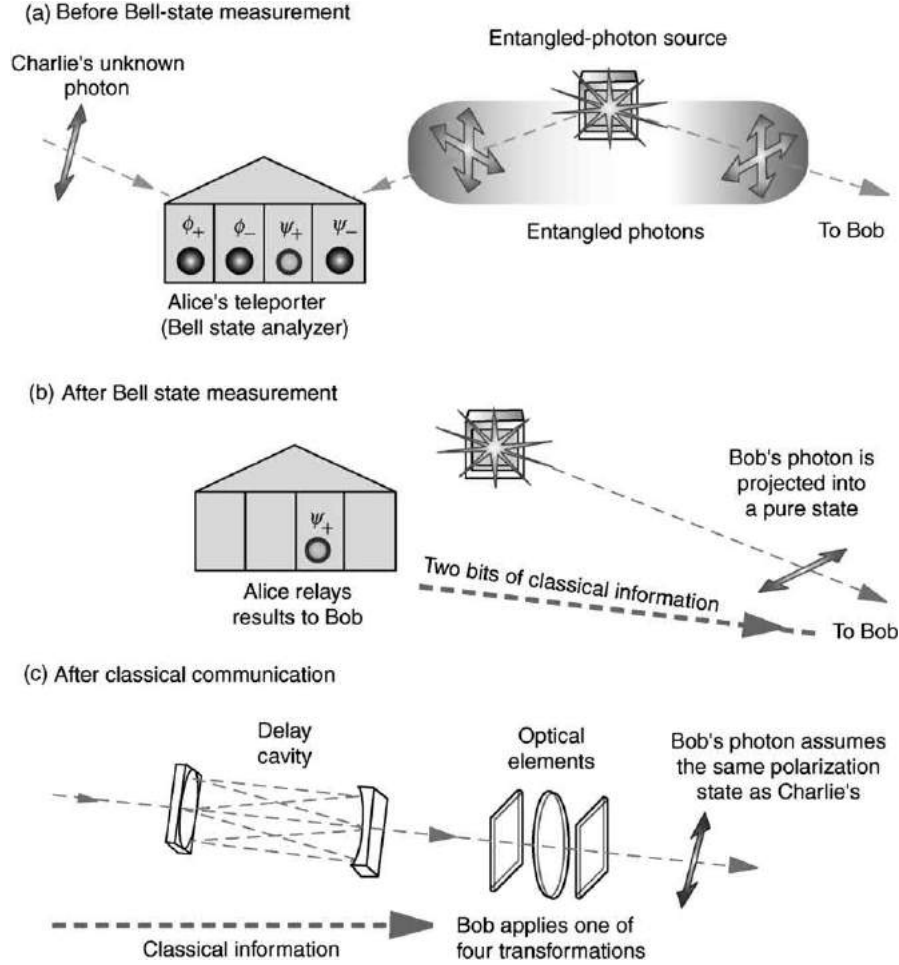


Fig. 4 Quantum state teleportation. (a) Teleportation requires an entangled photon source, a Bell state analyzer and a classical communications channel. One entangled photon goes to Bob and the other to Alice, who also receives a photon of unknown polarization state. (b) Alice performs a joint polarization measurement of the two photons and relays the result to Bob using two classical bits of information. The photon going to Bob is projected into a pure state as a result of Alice's measurement. (c) Upon receiving Alice's classical information, Bob performs a simple transformation on his photon (which he has been storing), such as a rotation of the polarization vector. The resulting state of his photon is then identical to the unknown state Alice wished to teleport. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

$$|\psi_0\rangle = (\alpha|H\rangle + \beta|V\rangle) \otimes \frac{1}{\sqrt{2}(|HV\rangle - |VH\rangle)},$$

which can be rewritten with the first two photons (the input plus the first half of the entangled pair) represented by the Bell state basis:

$$|\psi_0\rangle = \frac{1}{2} \{ |\Phi_+\rangle \otimes (-\beta|H\rangle + \alpha|V\rangle) + |\Phi_-\rangle \otimes (\beta|H\rangle + \alpha|V\rangle) + |\Psi_+\rangle \otimes (-\alpha|H\rangle + \beta|V\rangle) + |\Psi_-\rangle \otimes (-\alpha|H\rangle - \beta|V\rangle) \}.$$

Now suppose that a Bell state measurement is performed on the first two photons. The third photon is immediately projected into one of four possible states, which can be transformed back into the state of the original input photon by a simple operation, e.g., with a waveplate. For example, if the Bell state measurement produced the result $|\Psi_+\rangle$, the third photon is immediately 'collapsed' into the pure state $|\psi_3\rangle = -\alpha|H\rangle + \beta|V\rangle$. By applying a π -phase shift to the horizontal polarization (relative to the vertical), $|\psi_3\rangle$ can be transformed into the original input state: $|\psi_3\rangle = \alpha|H\rangle + \beta|V\rangle$. Alice (who does the Bell state measurement) need only communicate to Bob (who wants to receive the teleported photon) which of the four Bell states she measured. Note that during the entire teleportation procedure, neither Alice nor Bob can obtain any idea of the values of the parameters α and β , which specify the state. Also, because Alice's measurement collapses the unknown state, there is only a single copy at the end of the protocol.

Quantum state teleportation was first demonstrated experimentally by Zeilinger and coworkers (University of Innsbruck, Austria). The group was able to determine two of the four Bell states unambiguously (the other two states gave the same experimental signature) and proved for those cases that the state of the input photon could indeed be transferred to Bob. Other

experiments have realized modified forms of quantum teleportation in different systems. For example, Kimble's group (California Institute of Technology) teleported the coherent state of an optical mode using squeezed (rather than polarization-entangled) light. Researchers have recently suggested how teleportation might form the basis of a distributed network of quantum communication channels, and how it might enable quantum computing in all-optical systems.

Other Application of Optical Entanglement

Quantum Computing

The most challenging and powerful application of quantum information is large-scale quantum computing. Two features that make quantum information processing potentially powerful are the exponentially large Hilbert space, which gives quantum registers very large capacities, and quantum parallelism, which means that data processing tasks can be performed very efficiently. One fundamental drawback, which severely constrains useful applications of quantum computers, is that the final measurement can only produce a number of classical bits equal to the number of qubits. Thus, quantum computers are limited to performing tasks in which a small amount of information is meant to be gleaned from processing a large amount of data; examples include searching an unstructured database for a specific entry (Grover's algorithm) or finding the periodicity of a function (the quantum Fourier transform). This second task is central to Shor's factor-finding algorithm, the most famous quantum computing algorithm to date.

A practicable quantum computer technology must have at least the following features, first identified by DiVincenzo:

- A set of well-characterized, distinguishable qubits to form a quantum data storage register.
- The ability to initialize the qubits of the register in a simple fiducial state.
- Decoherence times that are much longer than the time needed to perform logical operations.
- The ability to perform any single qubit operation on any qubit in the register, and the ability to perform two-qubit conditional logic gates (such as the CNOT gate: $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \rightarrow \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle$). Together, these operations constitute a universal set of quantum gates, from which all other gates can be synthesized.
- The ability to measure each qubit.

All-optical schemes have been used to implement small quantum algorithms. However, most of the approaches are not scalable, due to the limitations of non-linear optics to perform a CNOT gate at the single-photon level. Recent proposals have suggested that very high efficiency single-photon detectors, along with sources of single photons 'on-demand,' allow scalable quantum computing with only linear optics; preliminary two-qubit gates have been experimentally demonstrated. A number of other promising candidate technologies that meet DiVincenzo's five requirements are being pursued vigorously. For example, the ability to create multiqubit entanglement and perform reliable measurements on trapped ions cooled and manipulated by lasers has recently been demonstrated. Solid state systems offer the possibility of scalability, and a number of schemes, such as quantum dots, isolated impurities with nuclear spins, and superconducting quantum interference devices (SQUIDs), are being investigated. It is possible that the final quantum computing technology may take the form of a hybrid between various present approaches.

Lithography

Lithography, in which a pattern is optically imaged onto some photoresistive material, is the primary method of manufacturing microscale or nanoscale electronic devices. An inherent limitation of this process is that details smaller than a wavelength of light cannot be written reliably. However, quantum state entanglement might circumvent this limitation. Under the right circumstances, the interference pattern formed by beams of entangled photons can have half the classical fringe spacing. Quantum lithography requires two beams of photons, a coherent superposition consisting of the state in which two photons are in beam A while none are in B, and the state in which no photon is in beam A while two photons are in B. Such number-entangled states can be made in the laboratory, and the predictions about fringe spacings have been verified. However, other obstacles must be overcome to surpass current classical-lithography techniques.

Two-Photon Imaging and Microscopy

At present, two-photon microscopy is widely used to produce high-resolution images, often of biological systems. However, the classical light sources (lasers) used for the imaging have random spreads in the temporal and spatial distributions of the photons, and the light intensity must be very high if two photons are to intersect within a small enough volume to cause a detectable excitation. Such high intensity can damage the system under investigation. Because the temporal and spatial correlations may be much stronger between members of an entangled photon pair, much weaker light sources could be used, which would be much less damaging to the systems being observed. The development of such systems is currently an active area of research.

Further Reading

- Berman, G.P., Doolen, G.D., Mainieri, R., Tsifrinovitch, V., 1998. *Introduction to Quantum Computers*. Singapore: World Scientific.
- Bouwmeester, D., Ekert, A.K., Zeilinger, A. (Eds.), 2000. *The Physics of Quantum Information*. Berlin: Springer-Verlag.
- Braunstein, S.L., Lo, H.K. (Eds.), 2001. *Scalable Quantum Computers: Paving the Way to Realization*. Berlin: Wiley-VCH.
- DiVincenzo, D.P., 2000. The physical implementation of quantum computation. *Fortschritte der Physik* 48, 771–783.
- Gisin, N., Ribordy, G., Tittel, W., Zbinden, H., 2002. Quantum cryptography. *Review of Modern Physics* 74, 145–195.
- Kok, P., Braunstein, S.L., Dowling, J.P., 2002. *Optics and Photonics News* 13 (9), 24–27.
- Lo, H.K., Popescu, S., Spiller, T. (Eds.), 1998. *Introduction to Quantum Computation and Information*. Singapore: World Scientific.
- Macchiavello, C., Palma, G.M., Zeilinger, A. (Eds.), 2000. *Quantum Computation and Quantum Information Theory*. Singapore: World Scientific.
- Mandel, L., Wolf, E., 1995. *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.
- Nielsen, M.A., Chuang, I.L., 2000. *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press.
- Williams, C.P., Clearwater, S.H., 1998. *Explorations in Quantum Computing*. Santa Clara: Telos.
- Yariv, A., 1989. *Quantum Electronics*. New York: Wiley.

Applications in Semiconductors

HM van Driel and JE Sipe, University of Toronto, Toronto, ON, Canada

© 2018 Elsevier Inc. All rights reserved.

Introduction

Interference phenomena are well-known in classical optics. In the early 19th century, Thomas Young showed conclusively that light has wave properties, with the first interference experiment ever performed. He passed quasi-monochromatic light from a single source through a pair of double slits using the configuration of Fig. 1(a) and observed an interference pattern consisting of a series of bright and dark fringes on a distant screen. The intensity distribution can be explained only if it is assumed that light has wave or phase properties. In modern terminology, if E_1 and E_2 are the complex electric fields of the two light beams arriving at the screen, by the superposition principle of field addition the intensity at a particular point on the screen can be written as

$$I \propto |E_1 + E_2|^2 = |E_1|^2 + |E_2|^2 + |E_1||E_2| \cos(\phi_1 - \phi_2) \quad (1)$$

where $\phi_1 - \phi_2$ is the phase difference of the beams arriving at the screen. While this simple experiment was used originally to demonstrate the wave properties of light, it has subsequently been used for many other purposes, such as measuring the wavelength of light. However, for our purposes, here the Young's double slit apparatus can also be viewed as a device that redistributes light, or controls its intensity at a particular location. For example, let's consider one of the slits to be a source, with the other slit taken to be a gate, with both capable of letting through the same amount of light. If the gate is closed, the distant screen is nearly uniformly illuminated. But if the gate is open, at a particular point on the distant screen there might be zero intensity or as much as four times the intensity emerging from one slit because of interference effects, with all the light merely being spatially redistributed. In this sense the double slit system can be viewed as a device to redistribute the incident light intensity. The key to all this is the superposition principle and the properties of optical phase.

The superposition principle and interference effects also lie at the heart of quantum mechanics. For example, let's now consider a system under the influence of a perturbation with an associated Hamiltonian that has phase properties. In general the system can

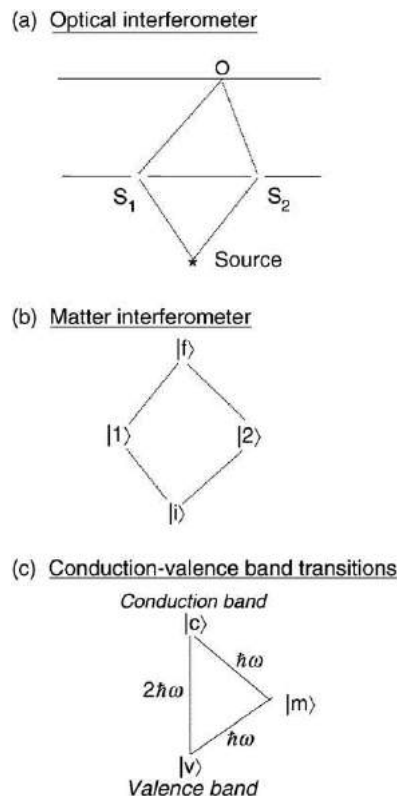


Fig. 1 (a) Interference effects in the Young's double slit experiment; (b) illustration of the general concept of coherence control via multiple quantum mechanical pathways; (c) interference of single- and two-photon transitions connecting the same valence and conduction band states in a semiconductor.

evolve from one quantum state $|i\rangle$ to another $|f\rangle$ via multiple pathways involving intermediate states $|m\rangle$. Because of the phased perturbation, interference between those pathways can influence the system's final state. If a_m is the (complex) amplitude associated with a transition from the initial to the final state, via intermediate virtual state $|m\rangle$, then for all possible intermediate states the overall transition probability, W , can be written as

$$W = \left| \sum_m a_m \right|^2 \quad (2)$$

In the case of only two pathways, as illustrated in Fig. 1(b), W becomes

$$\begin{aligned} W &= |a_1 + a_2|^2 \\ &= |a_1|^2 + |a_2|^2 + |a_1||a_2| \cos(\phi_1 - \phi_2) \end{aligned} \quad (3)$$

where ϕ_1 and ϕ_2 are now the phases of the two transition amplitudes. While this expression strongly resembles that of Eq. (1), here ϕ_1 and ϕ_2 are influenced by the phase properties of the perturbation Hamiltonian that governs the transition process; these phase factors arise, for example, if light fields constitute the perturbation. But overall it is clear that the change in the state of the system is affected by both the amplitude and phase properties of the perturbation. Analogous with the classical interferometer, equality of the transition amplitudes leads to maximum contrast in the transition rate. Although it has been known since the early days of quantum mechanics that the phase of a perturbation can influence the evolution of a system, generally phase has only been discussed in a passive role. Only in recent years has phase been used as a control parameter for a system, on the same level as the strength of the perturbation itself.

Coherence Control

Coherence control, or quantum control as it is sometimes called, refers to the active process through which one can use phase-dependent perturbations originating with, for example, coherent light waves, to control one or more properties of a quantum system, such as state population, momentum, or spin. This more general perspective of interference leads to a picture of interference of matter waves, rather than simply light waves, and one can now speak of an effective 'matter interferometer'. Laser beams, in particular, are in a unique position to play a role in such processes, since they offer a macroscopic 'phase handle' with which to create such effects. This was recognized by the community of atomic and molecular scientists who emphasized the active manifestations of quantum interference effects, and proposed that branching ratios in photochemical reactions might be controlled through laser-induced interference processes. The use of phase as a control parameter also represents a novel horizon of applications for the laser since most previous applications had involved only amplitude (intensity).

Of course, just as the 'visibility' of the screen pattern for a classical Young's double slit interferometer is governed by the coherence properties of the two slit source, the evolution of a quantum system reflects the coherence properties of the perturbations, and the degree to which one can exercise phase control via external perturbations is also influenced by the interaction of the system of interest with a reservoir – essentially any other degrees of freedom in the system – which can cause decoherence and reduce the effectiveness of the 'matter interferometer'. Since the influence of a reservoir will generally increase with the complexity of a system, one might think that coherence control can only be effectively achieved in simple atomic and molecular systems. Indeed, some of the earliest suggestions for coherence control of a system involved using interference between single and three photon absorption processes, connecting the same initial and final states with photons of frequency 3ω and ω , respectively, and controlling the population of excited states in atomic or diatomic systems. However, it has been difficult to extend this 'two color' paradigm to more complex molecular systems. Since the reservoir leads to decoherence of the system overall, shorter and shorter pulses must be used to overcome decoherence effects. However, short pulses possess a large bandwidth with the result that, for complex systems, selectivity of the final state can be lost. One must therefore consider using interference between multiple pathways in order to control the system, as dictated by the details of the unperturbed Hamiltonian of the system.

For a polyatomic molecule or a solid, the complete Hamiltonian is virtually impossible to determine exactly and it is therefore equally difficult to prescribe the optimal spectral (amplitude and phase) content of the pulses that should be used for control purposes. An alternative approach has therefore emerged, in which one foregoes knowledge of the eigenstates of the Hamiltonian and details of the different possible interfering transition paths in order to achieve control of the end state of a system. This branch of coherence control has come to be known as optimal control. In optimal control one employs an optical source for which one can (ideally) have complete control over the spectral and phase properties. One then uses a feedback process in which experiments are carried out, the effectiveness of achieving a certain result is determined, and the pulse characteristics are then altered to obtain a new result. A key component is the use of an algorithm to select the new pulse properties as part of the feedback system. In this approach the molecule teaches the external control system what it 'requires' for a certain result to be optimally achieved. While the 'best' pulse properties may not directly reveal details of the multiple interference process required to achieve the optimal result, this technique can nonetheless be used to gain some insight into the properties of the unperturbed Hamiltonian and, regardless of such understanding, achieve a desirable result. It has been used to control chemical reaction rates involving several polyatomic molecules, with considerable enhancement in achieving a certain product relative to what can be done using simple thermodynamics.

Coherence Control in Semiconductors

Since coherence control of chemical reactions involving large molecules has represented a significant challenge, it was generally felt that ultrafast decoherence processes would also make control in solids, in general, and semiconductors, in particular, difficult. Early efforts therefore focused on atomic-like situations in which the electrons are bound and associated with discrete states and long coherence times. In semiconductors, the obvious choice is the excitonic system, defect states, or discrete states offered by quantum wells. Population control of excitons and directional ionization of electrons from quantum wells has been clearly demonstrated, similar to related population control of and directional ionization from atoms. Other manifestations of coherence control of semiconductors include control of electron–phonon interactions and intersub-band transitions in quantum wells.

Among the different types of coherence control in semiconductors is the remarkable result that it is possible to coherently control the properties of free electrons associated with continuum states. Although optimal control might be used for some of these processes, we have achieved clear illustrations of control phenomena based on the use of harmonically related beams and interference of single- and two-photon absorption processes connecting the continuum valence and conduction band states as generically illustrated in Fig. 1(c). Details of the various processes observed can be found elsewhere. Of course, the momentum relaxation time of electrons or holes in continuum states is typically of the order of 100 fs at room temperature, but this time is sufficiently long that it can permit phase-controlled processes. Indeed, with respect to conventional carrier transport, this ‘long time’ lapse is responsible for the typically high electrical mobilities in crystalline semiconductors such as Si and GaAs. In essence, a crystalline semiconductor with translational symmetry has crystal momentum as a good quantum number, and selection rules for scattering prevent the momentum relaxation time from being prohibitively small. Since electrons or holes in any continuum state can participate in such control processes, one need not be concerned about pulsewidth or bandwidth, unless the pulses were to be so short that carriers of both positive and negative effective mass were generated within one band.

Coherent Control of Electrical Current Using Two Color Beams

We now illustrate the basic principles that describe how the interference process involving valence and conduction band states can be used to control properties of a bulk semiconductor. We do so in the case of using one or two beams to generate and control carrier population, electrical current, and spin current. In pointing out the underlying ideas behind coherence control of semiconductors, we will skip or suppress many of the mathematical details which are required for a full understanding but which might obscure the essential physics. In particular we consider how phase-related optical beams with frequencies ω and 2ω interact with a direct gap semiconductor such that $\hbar\omega < E_g < 2\hbar\omega$ where E_g is the electronic bandgap. For simplicity we consider exciting electrons from a single valence band via single-photon absorption at 2ω and two-photon absorption at ω . As is well known, within an independent particle approximation, the states of electrons and holes in semiconductors can be labeled by their vector crystal momentum $\mathbf{k}$; the energy of states near the conduction or valence band edges varies quadratically with $|\mathbf{k}|$. For one-photon absorption the transition amplitude can be derived using a perturbation Hamiltonian of the form $H = e/mc\mathbf{A}^{2\omega} \cdot \mathbf{p}$ where $\mathbf{A}^{2\omega}$ is the vector potential associated with the light field and $\mathbf{p}$ is the momentum operator. The transition amplitude is therefore of the form

$$a^{2\omega} \propto E^{2\omega} e^{i\phi_{2\omega}} p_{cv} \quad (4)$$

where p_{cv} is the interband matrix element of $\mathbf{p}$ along the field ($E^{2\omega}$) direction; for illustration purposes the field is taken to be linearly polarized. The overall transition rate between two particular states of the same $\mathbf{k}$ can be expressed as $W_1 \propto a^{2\omega}(a^{2\omega})^* \propto I^{2\omega}|p_{cv}|^2$ where $I^{2\omega}$ is the intensity of the beam. This rate is independent of the phase of the light beam as well as the sign of $\mathbf{k}$. Hence the absorption of light via single-photon transitions generally populates states of equal and opposite momentum with equal probability or, equivalently, establishes a standing electron wave with zero crystal momentum. This is not surprising since photons possess very little momentum and, in the approximation of a uniform electric field, do not give any momentum to an excited electron. The particular states that are excited depends on the light polarization and crystal orientation. However, the main point is that while single-photon absorption can lead to anisotropic filling of electron states, the distribution in momentum space is not polar (dependent on sign of $\mathbf{k}$). Similar considerations apply to the excited holes, but to avoid repetition we will focus on the electrons only.

For two-photon absorption involving the ω photons and connecting states similar to those connected with single-photon absorption, one must employ the 2nd order perturbation theory using the Hamiltonian $H = e/mc\mathbf{A}^\omega \cdot \mathbf{p}$. For two-photon absorption there is a transition from the valence band to an (energy nonallowed) intermediate state followed by a transition from the intermediate state to the final state. To determine the total transition rate one must sum over all possible intermediate states. For semiconductors, by far the dominant intermediate state is the final state itself, so that the associated transition amplitude has the form

$$a^\omega \propto (E^\omega e^{i\phi_\omega} p_{cv})(E^{2\omega} e^{i\phi_{2\omega}} p_{cc}) \propto (E^\omega)^2 e^{2i\phi_\omega} p_{cv} \hbar k \quad (5)$$

where the matrix element p_{cc} is simply the momentum of the conduction band state ($\hbar k$) along the direction of the field. Note that unlike an atomic system, p_{cc} is nonzero, since Bloch states in a crystalline solid do not possess inversion symmetry. If two-photon absorption acts alone, the overall transition rate between two particular $\mathbf{k}$ states would be $W_2 \propto (I^\omega)^2 |p_{cv}|^2 k^2$. As with single-photon absorption, because this transition rate is independent of the sign of k , two-photon absorption leads to production of electrons with no net momentum.

When both single- and two-photon transitions are present simultaneously, the transition amplitude is the sum of the transition amplitudes expressed in Eq. (3). The overall transition rate is then found using Eq. (1) and yields $W = W_1 + W_2 + \text{Int}$ where the interference term Int is given by

$$\text{Int} \propto E^{2\omega} E^\omega E^\omega \sin(\phi_{2\omega} - 2\phi_\omega)k \quad (6)$$

Note, however, that the interference effect depends on the sign of k and hence can be constructive for one part of the conduction band but destructive for other parts of that band, depending on the value of the relative phase, $\Delta\phi = \phi_{2\omega} - 2\phi_\omega$. In principle one can largely eliminate transitions with $+\mathbf{k}$ and enhance those with $-\mathbf{k}$. This effectively generates a net momentum for electrons or holes and hence, at least temporarily, leads to an electrical current in the absence of any external bias. The net momentum of the carriers, which is absent in the individual processes, must come from the lattice. Because the carriers are created with net momentum during the pulses one has a form of current injection that can be written as

$$dJ/dt \propto E^{2\omega} E^\omega E^\omega \sin(\Delta\phi) \quad (7)$$

where J is the current density. This type of current injection is allowed in both centrosymmetric and noncentrosymmetric materials. The physics and concepts behind the quantum interference leading to this form of current injection are analogous with the Young's double slit experiment. Note as well that if this 'interferometer' is balanced (single- and two-photon transition rates are similar), then it is possible to control the overall transition rate with great contrast. Roughly speaking, one balances the two arms of the 'effective interferometer' involved in the interference process, one 'arm' corresponding to the one-photon process and the other to the two-photon process. As an example, let's consider the excitation of GaAs, which has a room-temperature bandgap of 1.42 eV (equivalent to 870 nm). For excitation of GaAs using 1550 and 775 nm light under balanced conditions, the electron cloud is injected with a speed close to 500 km s^{-1} . Under 'balanced' conditions, nearly all the electrons are moving in the same direction. Assuming that the irradiance of the 1550 nm beam is 100 MW cm^{-2} , while that of the second harmonic beam is only 15 kW cm^{-2} (satisfying the 'balance' condition), with Gaussian pulse widths of 100 fs, one obtains a surprisingly large peak current of about 1 kA cm^{-2} for a carrier density of only 10^{14} cm^{-3} if scattering effects are ignored. When scattering is taken into account, the value of the peak current is reduced and the transient current decays on a time-scale of the momentum relaxation time. Fig. 2(a) shows an experimental setup which can be used to demonstrate coherence control of electrical current using the above parameters. Fig. 2(a) shows the region between a pair of electrodes on GaAs being illuminated by a train of harmonically related pulses. Fig. 2(b)

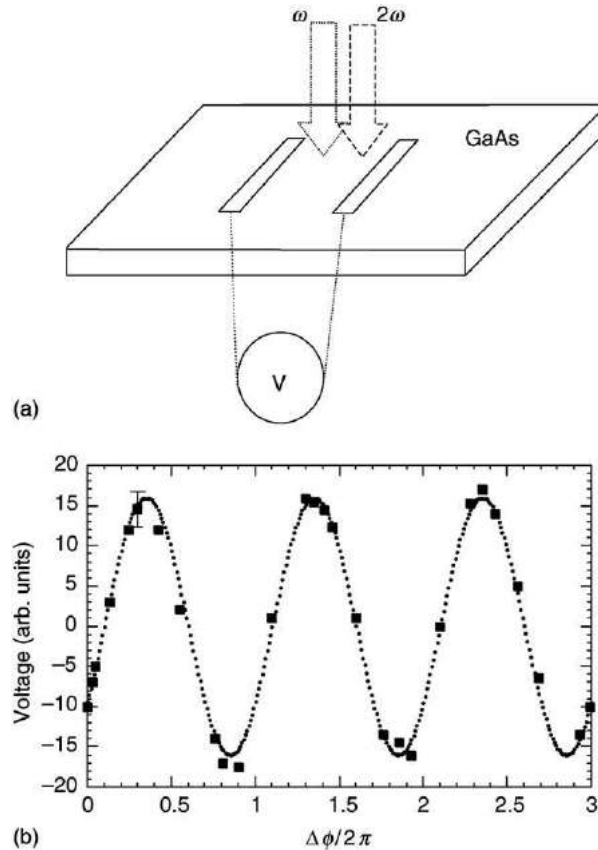


Fig. 2 (a) Experimental setup to measure steady-state voltage (V) across a pair of electrodes on GaAs with the intervening region illuminated by phased radiation at ω and 2ω ; (b) induced voltage across a pair of electrodes on a GaAs semiconductor as a function of the phase parameter associated with two harmonically related incident beams.

illustrates how the steady-state experimental voltage across the capacitor changes as the phase parameter $\Delta\phi$ is varied. Transient electrical currents, generated though incident femtosecond optical pulses, have also been detected through the emission of the associated Terahertz radiation.

The current injection via coherence control and conventional cases (i.e., under a DC bias) differs with respect to their evolution. The current injected via coherent control has an onset determined by the rise time of the optical pulses. In the case of normal current production, existing carriers are accelerated by an electric field, and the momentum distribution is never far from isotropic. For a carrier density of 10^{14} cm^{-3} a DC field $\sim 80 \text{ kVcm}^{-1}$ is required to produce a current density of 1 kA cm^{-2} . In GaAs, with an electron mobility of $8000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ this current would occur about $1/2 \text{ ps}$ after the field is ‘instantaneously’ turned on. This illustrates that the coherently controlled phenomenon efficiently and quickly produces a larger current than can be achieved with the redirecting of statistically distributed electrons.

A more detailed analysis must take into account the actual light polarization, crystal symmetry, crystal face, and orientation relative to the optical polarization. For given optical intensities of the two beams, the maximum electrical current injection in the case of GaAs occurs for linearly polarized beams both oriented along the (111) or equivalent direction. However, large currents can also be observed with the light beams polarized along other high symmetry directions.

Coherent Control of Carrier Density, Spin Population, and Spin Current Using Two Color Beams

The processes described above do not exhaust the coherent control effects that can be observed in bulk semiconductors using harmonic beams. Indeed, for noncentrosymmetric materials, certain light polarizations and crystal orientations allow one to coherently control the total carrier generation rate with or without generating an electrical current. In the case of the cubic material GaAs, provided the ω beam has electric field components along two of the three principal crystal axes, with the 2ω beam having a component along the third direction, one can coherently control the total carrier density. However, the overall degree of control here is determined by how ‘noncentrosymmetric’ the material is.

The spin degrees of freedom of a semiconductor can also be controlled using two color beams. Due to the spin–orbit interaction, the upper valence bands of a typical semiconductor have certain spin characteristics. To date, optical manipulation of electron spin has been largely based on the fact that partially spin-polarized carriers can be injected in a semiconductor via one-photon absorption of circularly polarized light from these upper valence bands. In such carrier injection – where in fact two-photon absorption could be used as well – spins with no net velocity are injected, and then are typically dragged by a bias voltage to produce a spin-polarized current. However, given the protocols discussed above it should not come as a surprise that the two color coherence scheme, when used with certain light polarizations, can coherently control the spin polarization, making it dependent on $\Delta\phi$. Furthermore, for certain polarization combinations it is also possible to generate a spin current with or without an electrical current. Given the excitement surrounding the field of spintronics, where the goal is the use of the spin degree of freedom for data storage and processing, the control of quantities involving the intrinsic angular momentum of the electron is of particular interest.

Various polarization and crystal geometries can be examined for generating spin currents with or without electrical current. For example, with reference to Fig. 3(a), for both beams propagating in the z (normal) direction of a crystal and possessing the same circular polarization, an electrical current can be injected in the (xy) plane, at an angle from the crystallographic x direction dependent on the relative phase parameter, $\Delta\phi$; this current is spin-polarized in the z direction. As well, the injected carriers have a $\pm z$ component of their velocity, as many with one component as with the other; but those going in one direction are preferentially spin-polarized in one direction in the (xy) plane, while those in the other direction are preferentially spin-polarized in the opposite direction. This is an example of a spin current in the absence of an electrical current.

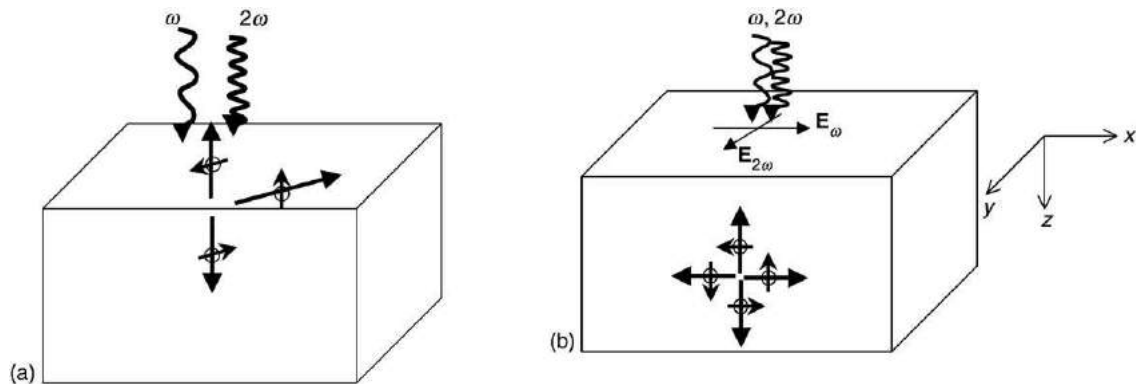


Fig. 3 (a) Excitation of a semiconductor by co-circularly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the x - y plane in a direction dependent on $\Delta\phi$ while a pure spin current is generated in the beam propagation (z) direction. (b) Excitation of a semiconductor by orthogonally, linearly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the direction of the fundamental beam polarization as well as along the beam propagation (z) direction.

Such a pure spin current that is perhaps more striking is observable with the two beams cross linearly polarized, for example with the fundamental beam in the x direction and the second harmonic beam in the y direction as shown in Fig. 3(b). Then there is no net spin injection; the average spin in any direction is zero. And for a vanishing phase parameter $\Delta\phi$ there is no net electrical current injection. Yet, for example, the electrons injected with a $+x$ velocity component will have one spin polarization with respect to the z direction, while those injected with a $-x$ component to their velocity will have the opposite spin polarization with respect to the z direction.

The examples given above are the spin current analog of the two-color electrical current injection discussed above. There should also be the possibility of injecting a spin current with a single beam into crystals lacking center-of-inversion symmetry.

The simple analysis presented above relies on simple quantum mechanical ideas and calculations using essentially nothing more than Fermi's Golden Rule. Yet the process involves a mixing of two frequencies, and can therefore be thought of as a nonlinear optical effect. Indeed, one of the key ideas to emerge from the theoretical study of such phenomena is that an interpretation of coherence control effects alternate to that provided by the simple quantum interference picture that is provided by the usual susceptibilities of nonlinear optics. These susceptibilities are, of course, based on quantum mechanics, but the macroscopic viewpoint allows for the identification and classification of the effects in terms of 2nd order nonlinear optical effects, 3rd order optical effects, etc. Indeed, one can generalize many of the processes we have discussed above to a hierarchy of frequency mixing effects or high-order nonlinear processes involving multiple beams with frequency $n\omega, m\omega, p\omega, \dots$ with n, m, p being integers. Many of these schemes require high intensity of one or more of the beams, but can occur in a simple semiconductor.

Coherent Control of Electrical Current Using Single Color Beams

It is also possible to generate coherence control effects using beams at a single frequency (ω), if one focuses on the two orthogonal components (e.g., x and y) polarization states and uses a noncentrosymmetric semiconductor of reduced symmetry such as a strained cubic semiconductor or a wurtzite material such as CdS or CdSe. In this case, interference between absorption pathways associated with the orthogonal components can lead to electrical current injection given by

$$dJ/dt \propto E^\omega E^\omega \sin(\phi_\omega^x - \phi_\omega^y) \quad (8)$$

Since in this process current injection is linear in the beam's intensity, the high intensities necessary for two-photon absorption are not necessary. Nonetheless, the efficacy of this process is limited by the fact that it relies on the breaking of center-of-inversion symmetry of the underlying crystal. It is also clear that the maximum current occurs for circularly polarized light and that right and left circularly polarized light lead to a difference in sign of the current injection. Finally, for this particular single beam scheme, when circularly polarized light is used the electrical current is partially spin polarized.

Conclusions

Through the phase of optical pulses it is possible to control electrical and spin currents, as well as carrier density, in bulk semiconductors on a time-scale that is limited only by the rise time of the optical pulse and the intrinsic response of the semiconductor. A few optically based processes that allow one to achieve these types of control have been illustrated here. Although at one level one can understand these processes in terms of quantum mechanical interference effects, at a macroscopic level one can understand these control phenomena as manifestations of nonlinear optical phenomena. For fundamental as well as applied reasons, our discussion has focused on coherence control effects using the continuum states in bulk semiconductors, although related effects can also occur in quantum dot, quantum well, and superlattice semiconductors. Applications of these control effects will undoubtedly exploit the all-optical nature of the process, including the speed at which the effects can be turned on or off. The turn-off effects, although not discussed extensively here, are related to transport phenomena as well as momentum scattering and related dephasing processes.

See also: Foundations of Coherent Transients in Semiconductors

Further Reading

- Brumer, P., 1995. Laser control of chemical reactions. *Scientific American* 272, 56.
 Rabitz, H., de Vivie-Riedle, R., Motzkus, M., Kompa, K., 2000. Whither the future of controlling quantum phenomena? *Science* 288, 824.
 Shah, J., 1999. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. New York: Springer.
 van Driel, H.M., Sipe, J.E., 2001. Coherent control of photocurrents in semiconductors. In: Tsen, T. (Ed.), *Ultrafast Phenomena in Semiconductors*. New York: Springer Verlag, p. 261.



Encyclopedia of Modern Optics

Second Edition

Editors in Chief

Bob Guenther
Duncan Steel

**ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION**

ENCYCLOPEDIA OF MODERN OPTICS SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 2

Spectroscopy ■ Terahertz ■ Optics of Semiconductors and 2D Materials
■ Nonlinear Optical Spectroscopy ■ Metamaterials and Plasmonics
■ Lasers, UV Lasers, Random lasers



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 2

Editorial Board	v
List of Contributors for Volume 2	ix
Editors in Chief	xi
Introduction	xiii

VOLUME 2

Transient Holographic Grating Techniques in Chemical Dynamics	<i>E Vauthey</i>	1
Atomic Physics	<i>G Kurizki, AG Kofman, and D Petrosyan</i>	12
Terahertz Physics of Semiconductor Heterostructures	<i>Juraj Darmo and Karl Unterrainer</i>	19
Strong-Field Terahertz Excitations in Semiconductors	<i>Ulrich Huttner, Rupert Huber, Mackillo Kira, and Stephan W Koch</i>	33
Rydberg States in Semiconductors	<i>Manfred Bayer and Marc Assmann</i>	40
Two-Dimensional Coherent Spectroscopy of Transition Metal Dichalcogenides	<i>Galan Moody</i>	52
Excitons in Magnetic Fields	<i>Kankan Cong, G Timothy Noe II, and Junichiro Kono</i>	63
Ultrafast Studies of Semiconductors	<i>J Shah</i>	82
Band Structure and Optical Properties	<i>W Zawadzki</i>	87
Excitons	<i>I Galbraith</i>	93
Quantum Wells and GaAs-Based Structures	<i>P Blood</i>	98
Recombination Processes	<i>PT Landsberg</i>	110
Coherent Terahertz Sources	<i>L Wang</i>	118
Using Ultrafast Optical Spectroscopy to Unravel the Properties of Correlated Electron Materials	<i>Rohit P Prasankumar, Dmitry A Yarotski, and Antoinette J Taylor</i>	123
Tutorial on Multidimensional Coherent Spectroscopy	<i>Mark Siemens</i>	150
Two-Dimensional Infrared (2D IR) Spectroscopy	<i>Lauren E Buchanan and Wei Xiong</i>	164
Two-Dimensional Electronic Spectroscopy	<i>Yin Song, Xiaoqin Li, and Jennifer P Ogilvie</i>	184
Multidimensional Terahertz Spectroscopy	<i>Michael Woerner, Klaus Reimann, and Thomas Elsaesser</i>	197
Second Harmonic Generation Spectroscopy of Hidden Phases	<i>Liuyan Zhao, Darius Torchinsky, John Harter, Alberto de la Torre, and David Hsieh</i>	207
Saturated Absorption Spectroscopy for Diode Laser Locking	<i>Bachana Lomsadze</i>	227
Attosecond Spectroscopy	<i>Agnieszka Jaroń-Becker and Andreas Becker</i>	233
Nonlinear Spectroscopies	<i>SR Meech</i>	244
Alternative Plasmonic Materials	<i>Gururaj V Naik</i>	252
Raman Lasers	<i>Marco Santagiustina</i>	265
GaN Lasers	<i>Harumasa Yoshida</i>	271
Optically Pumped Semiconductor Lasers	<i>Jerome V Moloney and Alexandre Laurain</i>	280

Optical Parametric Amplifiers	<i>Giulio Cerullo, Sandro De Silvestri, and Cristian Manzoni</i>	290
Mode-Locked Lasers	<i>Ladan Arissian and Jean-Claude Diels</i>	302
Few-Cycle and Attosecond Lasers	<i>Francesca Calegari and Caterina Vozzi</i>	311
Chirped Pulse Amplification	<i>GA Mourou</i>	324
Carbon Dioxide Laser	<i>CR Chatwin</i>	325
Dye Lasers	<i>FJ Duarte and A Costela</i>	336
Edge Emitters	<i>JJ Coleman</i>	350
Excimer Lasers	<i>JJ Ewing</i>	358
Metal Vapor Lasers	<i>DW Coutts</i>	368
Noble Gas Ion Lasers	<i>WB Bridges</i>	376
Planar Waveguide Lasers	<i>S Bhandarkar</i>	384
Up-Conversion Lasers	<i>A Brenier</i>	394
Thin Disk Lasers	<i>Mikhail Larionov</i>	407
Microchip Lasers	<i>John J Zayhowski</i>	415
Supercontinuum Generation	<i>James R Taylor</i>	424
Infrared Transition Metal Solid-State Lasers	<i>Kenneth L Schepler</i>	435
UV Lasers	<i>Yushi Kaneda</i>	446
Single-Frequency Lasers	<i>Xiushan Zhu</i>	451
Semiconductor Lasers	<i>Stephan W Koch and Martin R Hofmann</i>	462

LIST OF CONTRIBUTORS FOR VOLUME 2

Ladan Arissian
University of New Mexico, Albuquerque, NM, United States

Marc Assmann
Technical University of Dortmund, Dortmund, Germany

Manfred Bayer
Technical University of Dortmund, Dortmund, Germany

Andreas Becker
University of Colorado, Boulder, CO, United States

S Bhandarkar
Alfred University, Alfred, NY, USA

P Blood
Cardiff University, Cardiff, UK

A Brenier
University of Lyon, Villeurbanne, France

WB Bridges
California Institute of Technology, Pasadena, CA, USA

Lauren E Buchanan
Vanderbilt University, Nashville, TN, United States

Francesca Calegari
Center for Free-Electron Laser Science, Hamburg, Germany; University of Hamburg, Hamburg, Germany; and Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

Giulio Cerullo
Institute of Photonics and Nanotechnology, Milano, Italy

CR Chatwin
University of Sussex, Brighton, UK

JJ Coleman
University of Illinois, Urbana, IL, USA

Kankan Cong
Rice University, Houston, TX, United States

A Costela
Consejo Superior de Investigaciones Cientificas, Madrid, Spain

DW Coutts
University of Oxford, Oxford, UK

Juraj Darmo
Vienna University of Technology, Vienna, Austria

Alberto de la Torre
California Institute of Technology, Pasadena, CA, United States

Sandro De Silvestri
Institute of Photonics and Nanotechnology, Milano, Italy

Jean-Claude Diels
University of New Mexico, Albuquerque, NM, United States

FJ Duarte
Eastman Kodak Company, New York, NY, USA

Thomas Elsaesser
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

JJ Ewing
Ewing Technology Associates, Inc., Bellevue, WA, USA

I Galbraith
Heriot-Watt University, Edinburgh, UK

John Harter
University of California, Santa Barbara, CA, United States

Martin R Hofmann
Ruhr University Bochum, Bochum, Germany

David Hsieh
California Institute of Technology, Pasadena, CA, United States

Rupert Huber
University of Regensburg, Regensburg, Germany

Ulrich Huttner
Philipps University of Marburg, Marburg, Germany

Agnieszka Jaroń-Becker
University of Colorado, Boulder, CO, United States

Yushi Kaneda
The University of Arizona, Tucson, AZ, United States

Mackillo Kira
University of Michigan, Ann Arbor, MI, United States

Stephan W Koch
Philipps University of Marburg, Marburg, Germany

AG Kofman
Weizmann Institute of Science, Rehovot, Israel

Junichiro Kono
Rice University, Houston, TX, United States

G Kurizki
Weizmann Institute of Science, Rehovot, Israel

PT Landsberg
The University of Southampton, Southampton, UK

Mikhail Larionov
Dausinger + Giesen GmbH, Stuttgart, Germany

Alexandre Laurain
University of Arizona, Tucson, AZ, United States

Xiaoqin Li
The University of Texas at Austin, Austin, TX, United States

Bachana Lomsadze
University of Michigan, Ann Arbor, MI, United States

Cristian Manzoni
Institute of Photonics and Nanotechnology, Milano, Italy

SR Meech
University of East Anglia, Norwich, UK

Jerome V Moloney
University of Arizona, Tucson, AZ, United States

Galan Moody
National Institute of Standards & Technology, Boulder, CO, United States

GA Mourou
University of Michigan, Ann Arbor, MI, USA

Gururaj V Naik
Rice University, Houston, TX, United States

G Timothy Noe II
Rice University, Houston, TX, United States

Jennifer P Ogilvie
University of Michigan, Ann Arbor, MI, United States

D Petrosyan
Institute of Electronic Structure and Laser, Heraklion, Greece

Rohit P Prasankumar
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Klaus Reimann
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

Marco Santagiustina
University of Padova, Padova, Italy

Kenneth I. Schepler
University of Central Florida, Orlando, FL, United States

J Shah
Arlington, VA, USA

Mark Siemens
University of Denver, Denver, CO, United States

Yin Song
University of Michigan, Ann Arbor, MI, United States

Antoinette J Taylor
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

James R Taylor
Imperial College London, London, United Kingdom

Darius Torchinsky
Temple University, Philadelphia, PA, United States

Karl Unterrainer
Vienna University of Technology, Vienna, Austria

E Vauthey
University of Geneva, Geneva, Switzerland

Caterina Vozzi
Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

L Wang
Chinese Academy of Sciences, Beijing, China

Michael Woerner
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

Wei Xiong
University of California, San Diego, CA, United States

Dmitry A Yarotski
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Harumasa Yoshida
Mie University, Tsu, Mie, Japan

W Zawadzki
Polish Academy of Sciences, Warsaw, Poland

John J Zayhowski
Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, United States

Liuyan Zhao
University of Michigan, Ann Arbor, MI, United States

Xiushan Zhu
University of Arizona, Tucson AZ, United States

EDITORS IN CHIEF



Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.



Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Transient Holographic Grating Techniques in Chemical Dynamics

E Vauthey, University of Geneva, Geneva, Switzerland

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

C	concentration [mol L ⁻¹]	β	cubic volume expansion coefficient [K ⁻¹]
C_v	heat capacity [J K ⁻¹ kg ⁻¹]	γ	angle between transition dipoles
d	sample thickness [m]	Λ	fringe spacing [m]
D_{th}	thermal diffusivity [m ² s ⁻¹]	Δn_d	variation of refractive index due to density changes
I_{dif}	diffracted intensity [W m ⁻²]	Δn_d^e	Δn_d due to electrostriction
I_{pr}	intensity of the probe pulse [W m ⁻²]	Δn_K	variation of refractive index due to optical Kerr effect
I_{pu}	intensity of a pump pulse [W m ⁻²]	Δn_d^p	Δn_d due to volume changes
k_{ac}	acoustic wavevector [m ⁻¹]	Δn_p	variation of refractive index due to population changes
k_r	rate constant of a heat releasing process [s ⁻¹]	Δn_d^t	Δn_d due to temperature changes
k_{re}	rate constant of reorientation [s ⁻¹]	Δx	peak to null variation of the parameter x
k_{th}	rate constant of thermal diffusion [s ⁻¹]	ε	molar decadic absorption coefficient [cm L mol ⁻¹]
K	attenuation constant	ζ	angle of polarization of the diffracted signal
n	refractive index	η	diffraction efficiency
$\tilde{n}$	complex refractive index	θ_B	Bragg angle (angle of incidence of the probe pulse)
N	number of molecule per unit volume [m ⁻³]	θ_{pu}	angle of incidence of a pump pulse
r	polarization anisotropy	λ_{pr}	probe wavelength [m]
R	fourth rank response tensor [m ² V ⁻²]	λ_{pu}	pump wavelength [m]
v_s	speed of sound [m s ⁻¹]	ρ	density [kg m ⁻³]
V	volume [m ³]	τ_{ac}	acoustic period [s]
α_{ac}	acoustic attenuation constant [m ⁻¹]	v_{ac}	acoustic frequency [s ⁻¹]

Introduction

Over the past two decades, holographic techniques have proved to be valuable tools for investigating the dynamics of chemical processes. The aim of this chapter is to give an overview of the main applications of these techniques for chemical dynamics. The basic principle underlying the formation and the detection of elementary transient holograms, also called transient gratings, is first presented. This is followed by a brief description of a typical experimental setup. The main applications of these techniques to solve chemical problems are then discussed.

Basic Principle

The basic principle of the transient holographic technique is illustrated in Fig. 1. The sample material is excited by two laser pulses at the same wavelength and crossed at an angle $2\theta_{pu}$. If the two pump pulses have the same intensity, I_{pu} , the intensity distribution at the interference region, assuming plane waves, is

$$I(x) = 2I_{pu} \left[1 + \cos\left(\frac{2\pi x}{\Lambda}\right) \right] \quad (1)$$

where $\Lambda = \lambda_{pu}/(2\sin \theta_{pu})$ is the fringe spacing, λ_{pu} is the pump wavelength, and I_{pu} is the intensity of one pump pulse.

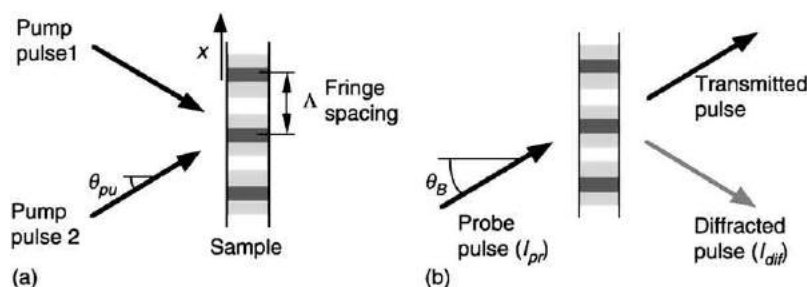


Fig. 1 Principle of the transient grating technique: (a) grating formation (pumping), (b) grating detection (probing).

As discussed below, there can be many types of light-matter interactions that lead to a change in the optical properties of the material. For a dielectric material, they result in a spatial modulation of the optical susceptibility and thus of the complex refractive index, $\tilde{n}$. The latter distribution can be described as a Fourier cosine series:

$$\tilde{n}(x) = \tilde{n}_0 + \sum_{m=1}^{\infty} \tilde{n}_m \cos\left(\frac{m2\pi x}{\Lambda}\right) \quad (2)$$

where $\tilde{n}_0$ is the average value of $\tilde{n}$. In the absence of saturation effects, the spatial modulation of $\tilde{n}$ is harmonic and the Fourier coefficients with $m > 1$ vanish. In this case, the peak to null variation of the complex refractive index, $\Delta\tilde{n}$, is equal to the Fourier coefficient $\tilde{n}_1$. The complex refractive index can be split into its real and imaginary components:

$$\tilde{n} = n + iK \quad (3)$$

where n is the refractive index and K is the attenuation constant.

The hologram created by the interaction of the crossed pump pulses consists in periodic, one-dimensional spatial modulations of n and K . Such distributions are nothing but phase and amplitude gratings, respectively. A third laser beam at the probe wavelength, λ_{pr} , striking these gratings at Bragg angle, $\theta_B = \arcsin(\lambda_{pr}/2\Lambda)$, will thus be partially diffracted (Fig. 1(b)). The diffraction efficiency, η , depends on the modulation amplitude of the optical properties. In the limit of small diffraction efficiency ($\eta < 0.01$), this relationship is given by

$$\eta = \frac{I_{dif}}{I_{pr}} \cong \left[\left(\frac{\ln 10 \Delta A}{4 \cos \theta_B} \right)^2 + \left(\frac{\pi d \Delta n}{\lambda_{pr} \cos \theta_B} \right)^2 \right] \times \exp\left(-\frac{\ln 10 A}{\cos \theta_B}\right) \quad (4)$$

where I_{dif} and I_{pr} are the diffracted and the probe intensity respectively, d is the sample thickness, and $A = 4\pi d K / (\lambda \ln 10)$ is the average absorbance.

The first and the second terms in the square bracket describe the contribution of the amplitude and phase gratings respectively, and the exponential term accounts for the reabsorption of the diffracted beam by the sample.

The main processes responsible for the variation of the optical properties of an isotropic dielectric material are summarized in Fig. 2.

The modulation of the absorbance, ΔA , is essentially due to the photoinduced concentration change, ΔC , of the different chemical species i (excited state, photoproduct, ...):

$$\Delta A(\lambda_{pr}) = \sum_i \varepsilon_i(\lambda_{pr}) \Delta C_i \quad (5)$$

where ε_i is the absorption coefficient of the species i .

The variation of the refractive index, Δn , has several origins and can be expressed as

$$\Delta n = \Delta n_K + \Delta n_p + \Delta n_d \quad (6)$$

Δn_K is the variation of refractive index due to the optical Kerr effect (OKE). This nonresonant interaction results in an electronic polarization (electronic OKE) and/or in a nuclear reorientation of the molecules (nuclear OKE) along the direction of the electric field associated with the pump pulses. As a consequence, a transient birefringence is created in the material. This effect is usually discussed within the framework of nonlinear optics in terms of intensity dependent refractive index or third order nonlinear

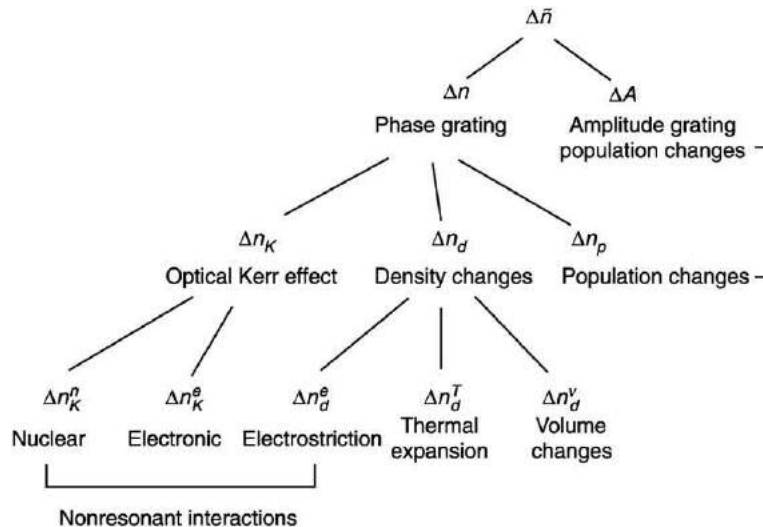


Fig. 2 Classification of the possible contributions to a transient grating signal.

susceptibility. Electronic OKE occurs in any dielectric material under sufficiently high light intensity. On the other hand, nuclear OKE is mostly observed in liquids and gases and depends strongly on the molecular shape.

Δn_p is the change of refractive index related to population changes. Its magnitude and wavelength dependence can be obtained by Kramers–Kronig transformation of $\Delta A(\lambda)$ or $\Delta K(\lambda)$:

$$\Delta n_p(\lambda) = \frac{1}{2\pi^2} \int_0^\infty \frac{\Delta K(\lambda')}{1 - (\lambda'/\lambda)^2} d\lambda' \quad (7)$$

Δn_d is the change of refractive index associated with density changes. Density phase gratings can have essentially three origins:

$$\Delta n_d = \Delta n_d^t + \Delta n_d^v + \Delta n_d^e \quad (8)$$

Δn_d^t is related to the temperature-induced change of density. If a fraction of the excitation energy is converted into heat, through a nonradiative transition or an exothermic process, the temperature becomes spatially modulated. This results in a variation of density, hence to a modulation of refractive index with amplitude Δn_d^t . Most of the temperature dependence of n originates from the density. The temperature-induced variation of n at constant density is much smaller than Δn_d^t .

Δn_d^v is related to the variation of volume upon population changes. This volume comprises not only the reactant and product molecules but also their environment. For example, in the case of a photodissociation, the volume of the product is larger than that of the reactant and a positive volume change can be expected. This will lead to a decrease of the density and to a negative Δn_d^v .

Finally, Δn_d^e is related to electrostriction in the sample by the electric field of the pump pulses. Like OKE, this is a nonresonant process that also contributes to the intensity dependent refractive index. Electrostriction leads to material compression in the regions of high electric field strength. The periodic compression is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm (2\pi/\Lambda)\vec{i}$, where $\vec{i}$ is the unit vector along the modulation axis. The interference of these acoustic waves leads to a temporal modulation of Δn_d^e at the acoustic frequency ν_{ac} , with $2\pi\nu_{ac} = k_{ac}v_s$, v_s being the speed of sound. As Δn_d^e oscillates between negative and positive values, the diffracted intensity, which is proportional to $(\Delta n_d^e)^2$, shows at temporal oscillation at twice the acoustic frequency. In most cases, Δn_d^e is weak and can be neglected if the pump pulses are within an absorption band of the sample.

The modulation amplitudes of absorbance and refractive index are not constant in time and their temporal behavior depends on various dynamic processes in the sample. The whole point of the transient grating techniques is precisely the measurement of the diffracted intensity as a function of time after excitation to deduce dynamic information on the system.

In the following, we will show that the various processes shown in Fig. 2, that give rise to a diffracted signal, can in principle be separated by choosing appropriate experimental parameters, such as the timescale, the probe wavelength, the polarization of the four beams, or the crossing angle.

Experimental Setup

A typical experimental arrangement for pump–probe transient grating measurements is shown in Fig. 3. The laser output pulses are split in three parts. Two parts of equal intensity are used as pump pulses and are crossed in the sample. In order to ensure time coincidence, one pump pulse travels along an adjustable optical delay line. The third part, which is used for probing, can be frequency converted using a dye laser, a nonlinear crystal, a Raman shifter, or white light continuum generation. The probe pulse is sent along a motorized optical delay line before striking the sample at the Bragg angle. There are several possible beam configurations for transient grating and the two most used are illustrated in Fig. 4. When the probe and pump pulses are at different wavelengths, they can be in the same plane of incidence as shown in Fig. 4(a). However, if the pump and probe wavelengths are the same, the folded boxcars geometry shown in Fig. 4(b) has to be used. The transient grating technique is background free and the diffracted signal propagates in a well-defined direction. In a pump–probe experiment, the diffracted signal intensity is measured as a function of the time delay between the pump and probe pulses. Such a setup can be used to probe dynamic processes occurring in timescales going from a few fs to a few ns, the time resolution depending essentially on the duration of the pump and probe pulses. For slower processes, the grating dynamics can be probed in real time with a *cw* laser beam and a fast photodetector.

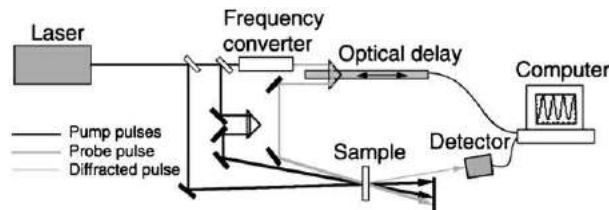


Fig. 3 Schematic of a transient grating setup with pump–probe detection.

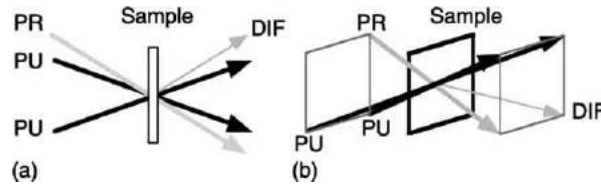


Fig. 4 Beam geometry for transient grating: (a) in plane; (b) boxcars.

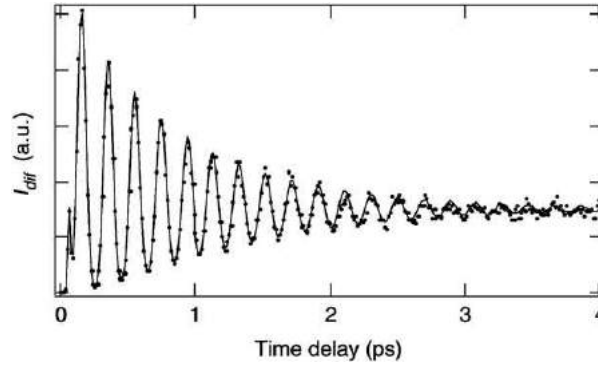


Fig. 5 Time profile of the diffracted intensity measured with a solution of malachite green after excitation with two pulses with close to counterpropagating geometry.

Applications

The Transient Density Phase Grating Technique

If the grating is probed at a wavelength far from any absorption band, the variation of absorbance, $\Delta A(\lambda_{pr})$, is zero and the corresponding change of refractive index, $\Delta n_p(\lambda_{pr})$, is negligibly small. In this case, Eq. (4) simplifies to

$$\eta \cong \left(\frac{\pi d}{\lambda_{pr} \cos \theta_B} \right)^2 \times \Delta n_d^2 \quad (9)$$

In principle, the diffracted signal may also contain contributions from the optical Kerr effect, Δn_K , but we will assume here that this non-resonant and ultrafast response is negligibly small. The density change can originate from both heat releasing processes and volume differences between the products and reactants, the former contribution usually being much larger than the latter. Even if the heat releasing process is instantaneous, the risetime of the density grating is limited by thermal expansion. This expansion is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm (2\pi/\Lambda)\vec{i}$. One can distinguish two density gratings: 1) a diffusive density grating, which reproduces the spatial distribution of temperature and which decays by thermal diffusion; and 2) an acoustic density grating originating from the standing acoustic wave and whose amplitude oscillates at the acoustic frequency ν_{ac} .

The amplitudes of these two gratings are equal but of opposite sign. Consequently, the time dependence of the modulation amplitude of the density phase grating is given by

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V \right) \rho \left(\frac{\partial n}{\partial \rho} \right) R(t) \quad (10a)$$

with

$$R(t) = 1 - \cos(2\pi\nu_{ac}t) \times \exp(-\alpha_{ac}\nu_s t) \quad (10b)$$

where ρ , β , C_v , and α_{ac} are the density, the volume expansion coefficient, the heat capacity, and the acoustic attenuation constant of the medium, respectively, Q is the amount of heat deposited during the photoinduced process, and ΔV is the corresponding volume change. As the standing acoustic wave oscillates, its corresponding density grating interferes with the diffusive density grating. Therefore, the total modulation amplitude of the density and thus Δn_d exhibits the oscillation at ν_{ac} . Fig. 5 shows the time profile of the diffracted intensity measured with a solution of malachite green. After excitation to the S_1 state, this dye relaxes nonradiatively to the ground state in a few ps. For this measurement, the sample solution was excited with two 30 ps laser pulses at 532 nm crossed with an angle close to 180° . The continuous line is the best fit of Eqs. (9) and (10). The damping of the oscillation is due to acoustic attenuation. After complete damping, the remaining diffracted signal is due to the diffusive density grating only.

$R(t)$ can be considered as the response function of the sample to a prompt heat release and/or volume change. If these processes are not instantaneous compared to an acoustic period ($\tau_{ac} = v_{ac}^{-1}$), the acoustic waves are not created impulsively and the time dependence of Δn_d is

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V \right) \rho \left(\frac{\partial n}{\partial \rho} \right) F(t) \quad (11a)$$

with

$$F(t) = \int_{-\infty}^t R(t-t') \cdot f(t') dt' \quad (11b)$$

where $f(t)$ is a normalized function describing the time evolution of the temperature and/or volume change. In many cases, $f(t) = \exp(-k_r t)$, k_r being the rate constant of the process responsible for the change. Fig. 6 shows the time profile of the diffracted intensity calculated with Eqs. (9) and (11) for different values of k_r .

If several processes take place, the total change of refractive index is the sum of the changes due to the individual processes. In this case, Δn_d should be expressed as

$$\Delta n_d(t) = \sum_i \left(\frac{\beta Q_i}{\rho C_v} + \Delta V_i \right) \rho \left(\frac{\partial n}{\partial \rho} \right) F_i(t) \quad (12)$$

The separation of the thermal and volume contributions to the diffracted signal is problematic. Several approaches have been proposed, the most used being the measurement in a series of solvents with different expansion coefficient β . This method requires that the other solvent properties like refractive index, dielectric constant, or viscosity, are constant or that the energetics of the system investigated does not depend on them. This separation is easier when working with water because its β value vanishes at 4 °C. At this temperature, the density variations are due to the volume changes only.

The above equations describe the growth of the density phase grating. However, this grating is not permanent and decays through diffusive processes. The phase grating originating from thermal expansion decays via thermal diffusion with a rate constant k_{th} given by

$$k_{th} = D_{th} \left(\frac{2\pi}{\Lambda} \right)^2 \quad (13)$$

where D_{th} is the thermal diffusivity. Table 1 shows k_{th} values in acetonitrile for different crossing angles.

The decay of the phase grating originating from volume changes depends on the dynamics of the population responsible for ΔV (vide infra).

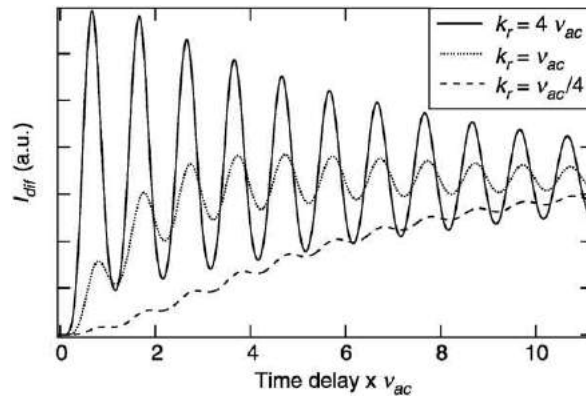


Fig. 6 Time profiles of the diffracted intensity calculated using Eqs. (9) and (11) with various k_r values.

Table 1 Fringe spacing, Λ , acoustic frequency, v_{ac} , thermal diffusion rate constant, k_{th} , for various crossing angles of the pump pulses, $2 \theta_{pu}$, at 355 nm and in acetonitrile

$2 \theta_{pu}$	Λ (μm)	v_{ac} (s^{-1})	k_{th} (s^{-1})
0.5°	40.7	3.2×10^7	4.7×10^3
50°	0.42	3.1×10^9	4.5×10^7
180°	0.13	9.7×10^9	4.7×10^8

Time resolved optical calorimetry

A major application of the density phase grating in chemistry is the investigation of the energetics of photo-induced processes. The great advantage of this technique over other optical calorimetric methods, like the thermal lens or the photoacoustic spectroscopy, is its superior time resolution. The time constant of the fastest heat releasing process that can be time resolved with this technique is of the order of the acoustic period. The shortest acoustic period is achieved when forming the grating with two counter-propagating pump pulses ($2\theta_{pu} = 180^\circ$). In this case the fringe spacing is $\Lambda = \lambda_{pu}/2n$ and, with UV pump pulses, an acoustic period of the order of 100 ps can be obtained in a typical organic solvent.

For example, Fig. 7(a) shows the energy diagram for a photoinduced electron transfer (ET) reaction between benzophenone (BP) and an electron donor (D) in a polar solvent. Upon excitation at 355 nm, $^1\text{BP}^*$ undergoes intersystem-crossing (ISC) to $^3\text{BP}^*$ with a time constant of about 10 ps. After diffusional encounter with the electron donor, ET takes place and a pair of ions is generated. With this system, the whole energy of the 355 nm photon ($E = 3.49$ eV) is converted into heat. The different heat releasing processes can be differentiated according to their timescale. The vibrational relaxation to $^1\text{BP}^*$ and the ensuing ISC to $^3\text{BP}^*$ induces an ultrafast release of 0.49 eV as heat. With donor concentrations of the order of 0.1 M, the heat deposition process due to the electron transfer is typically in the ns range. Finally, the recombination of the ions produces a heat release in the

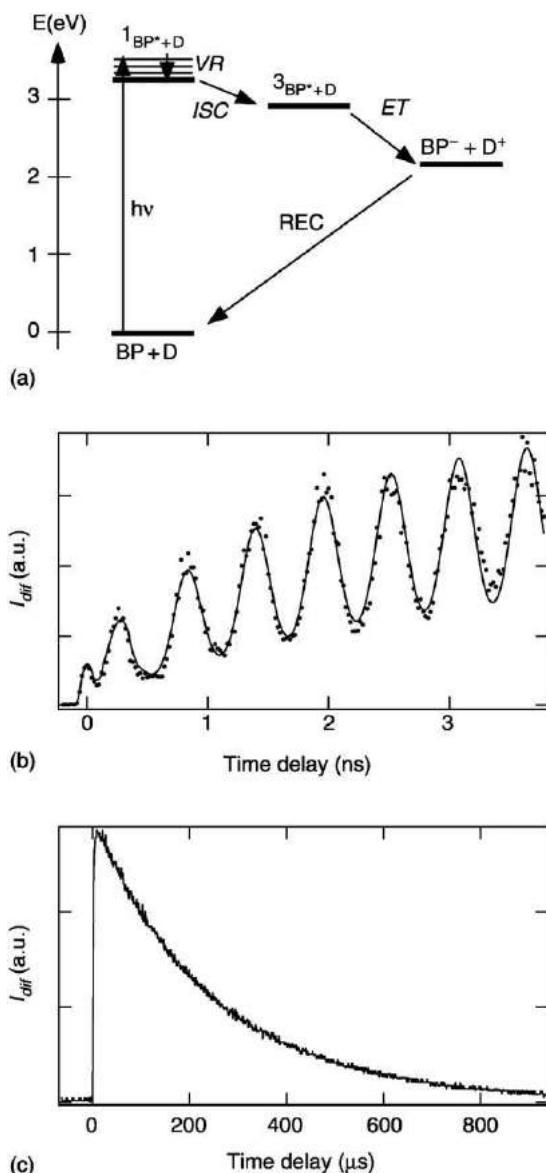


Fig. 7 (a) Energy diagram of the states involved in the photo-induced electron transfer reaction between benzophenone (BP) and an electron donor (D) (VR: vibrational relaxation, ISC: intersystem crossing; ET: electron transfer; REC: charge recombination). (b) Time profile of the diffracted intensity measured at 590 nm after excitation at 355 nm of a solution of BP and 0.05 M D. (c) Same as (b) but measured at 1064 nm with a cw laser.

microsecond timescale. Fig. 7(b) shows the time profile of the diffracted intensity measured after excitation at 355 nm of BP with 0.05 M D in acetonitrile. The 30 ps pump pulses were crossed at 27° ($\tau_{ac}=565$ ps) and the 10 ps probe pulses were at 590 nm. The oscillatory behavior is due to the ultrafast heat released upon formation of $^3BP^*$, while the slow dc rise is caused by the heat dissipated upon ET. As the amount of energy released in the ultrafast process is known, the energy released upon ET can be determined by comparing the amplitudes of the fast and slow components of the time profile. The energetics as well as the dynamics of the photoinduced ET process can thus be determined. Fig. 7(c) shows the decay of the diffracted intensity due to the washing out of the grating by thermal diffusion. As both the acoustic frequency and the thermal diffusion depends on the grating wavevector, the experimental time window in which a heat releasing process can be measured, depends mainly on the crossing angle of the pump pulses as shown in Table 1.

Determination of material properties

Another important application of the transient density phase grating technique is the investigation of material properties. Acoustics waves of various frequencies, depending on the pump wavelength and crossing angle, can be generated without physical contact with the sample. Therefore, acoustic properties, such as the speed of sound and the acoustic attenuation of the material, can be easily obtained from the frequency and the damping of the oscillation of the transient grating signal.

Similarly, the optoelastic constant of a material, $\rho \partial n / \partial \rho$, can be determined from the amplitude of the signal (see Eq. (11)). This is done by comparing the signal amplitude of the material under investigation with that obtained with a known standard.

Finally, the thermal diffusivity, D_{th} , can be easily obtained from the decay of the thermal density phase grating, such as that shown in Fig. 7(c). This technique can be used with a large variety of bulk materials as well as films, surfaces and interfaces.

Investigation of Population Dynamics

The dynamic properties of a photogenerated species can be investigated by using a probe wavelength within its absorption or dispersion spectrum. As shown by Eq. (4), either ΔA or Δn_p has to be different from zero. For this application, Δn_K and Δn_d should ideally be equal to zero. Unless working with ultrashort pulses (< 1 ps) and a weakly absorbing sample, Δn_K can be neglected. Practically, it is almost impossible to find a sample system for which some fraction of the energy absorbed as light is not released as heat. However, by using a sufficiently small crossing angle of the pump pulses, the formation of the density phase grating, which depends on the acoustic period, can take as much as 30 to 40 ns. In this case, Δn_d is negligible during the first few ns after excitation and the diffracted intensity is due to the population grating only:

$$I_{dif}(t) \propto \sum_i \Delta C_i^2(t) \quad (14)$$

where ΔC_i is the modulation amplitude of the concentration of every species i whose absorption and/or dispersion spectrum overlaps with the probe wavelength.

The temporal variation of ΔC_i can be due either to the dynamics of the species i in the illuminated grating fringes or to processes taking place between the fringes, such as translational diffusion, excitation transport, and charge diffusion. In liquids, the decay of the population grating by translational diffusion is slow and occurs in the microsecond to millisecond timescale, depending on the fringe spacing. As thermal diffusion is typically hundred times faster, Eq. (14) is again valid in this long timescale. Therefore if the population dynamics is very slow, the translational diffusion coefficient of a chemical species can be obtained by measuring the decay of the diffracted intensity as a function of the fringe spacing. This procedure has also been used to determine the temperature of flames. In this case however, the decay of the population grating by translational diffusion occurs typically in the sub-ns timescale.

In the condensed phase, these interfringe processes are of minor importance when measuring the diffracted intensity in the short timescale, i.e., before the formation of the density phase grating. In this case, the transient grating technique is similar to transient absorption, and it thus allows the measurement of population dynamics. However, because holographic detection is background free, it is at least a hundred times more sensitive than transient absorption.

The population gratings are usually probed with monochromatic laser pulses. This procedure is well suited for simple photoinduced processes, such as the decay of an excited state population to the ground state. For example, Fig. 8 shows the decay of the diffracted intensity at 532 nm measured after excitation of a cyanine dye at the same wavelength. These dynamics correspond to the ground state recovery of the dye by non-radiative deactivation of the first singlet excited state. Because the diffracted intensity is proportional to the square of the concentration changes (see Eq. (14)), its decay time is twice as short as the ground state recovery time. If the reaction is more complex, or if several intermediates are involved, the population grating has to be probed at several wavelengths. Instead of repeating many single wavelength experiments, it is preferable to perform multiplex transient grating. In this case, the grating is probed by white light pulses generated by focusing high intensity fs or ps laser pulses in a dispersive medium. If the crossing angle of the pump pulses is small enough ($< 1^\circ$), the Bragg angle for probing is almost independent on the wavelength. A transient grating spectrum is obtained by dispersing the diffracted signal in a spectrograph. This spectrum consists in the sum of the square of the transient absorption and transient dispersion spectra. Practically, it is very similar to a transient absorption spectrum, but with a much superior signal to noise ratio. Fig. 9 shows the transient grating spectra measured at various time delays after excitation of a solution of chloranil (CA) and methylnaphthalene (MNA) in acetonitrile. The reaction

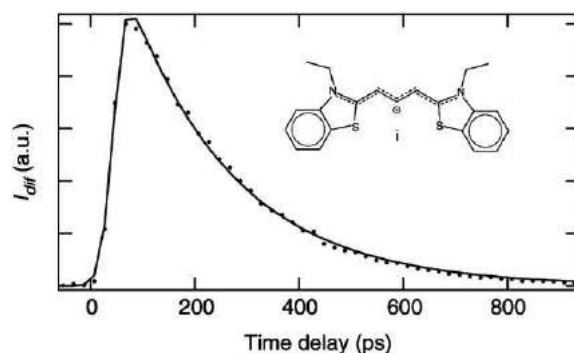


Fig. 8 Time profile of the diffracted intensity at 532 nm measured after excitation at the same wavelength of a cyanine dye (inset) in solution and best fit assuming exponential ground state recovery.

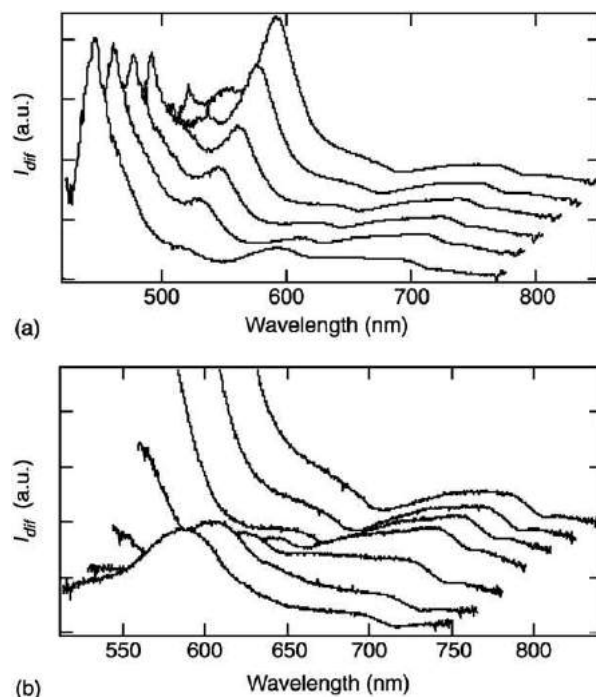
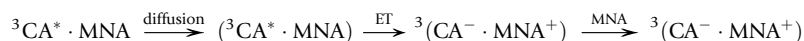


Fig. 9 Transient grating spectra obtained at various time delays after excitation at 355 nm of a solution of chloranil and 0.25 M methylnaphthalene (a): from top to bottom: 60, 100, 180, 260, 330 and 600 ps; (b): from top to bottom: 100, 400, 750, 1100, 1500 ps).

that takes place is



After its formation upon ultrafast ISC, ${}^3\text{CA}^*$, which absorbs around 510 nm, decays upon ET with MNA to generate CA^- (450 nm) and MNA^+ (690 nm). The ion pair reacts further with a second MNA molecule to form the dimer cation (580 nm).

The time profile of the diffracted intensity reflects the population dynamics as long as these populations follow first- or pseudo-first order kinetics. Higher order kinetics leads to inharmonic gratings and in this case the time dependence of the diffracted intensity is no longer a direct measure of the population dynamics.

Polarization Selective Transient Grating

In the above applications, the selection between the different contributions to the diffracted signal was essentially made by choosing the probe wavelength and the crossing angle of the pump pulses. However, this approach is not always sufficient. Another important parameter is the polarization of the four waves involved in a transient grating experiment. For example, when measuring population dynamics, the polarization of the probe beam has to be at magic angle (54.7°) relatively to that of the

pump pulses. This ensures that the observed time profiles are not distorted by the decay of the orientational anisotropy of the species created by the polarized pump pulses.

The magnitude of ΔA and Δn depends on the pump pulses intensity, and therefore the diffracted intensity can be expressed by using the formalism of nonlinear optics:

$$I_{\text{dif}}(t) = C \int_{-\infty}^{+\infty} dt I_{\text{pr}}(t - t'') \left[\int_{-\infty}^t dt' |R_{ijkl}(t'' - t')| I_{\text{pu}}(t') \right]^2 \quad (15)$$

where C is a constant and R_{ijkl} is an element of the fourth rank tensor $\mathbf{R}$ describing the nonlinear response of the material to the applied optical fields. In isotropic media, this tensor has only 21 nonzero elements, which are related as follows:

$$R_{1111} = R_{1122} + R_{1212} + R_{1221} \quad (16)$$

where the subscripts are the Cartesian coordinates. Going from right to left, they design the direction of polarization of the pump, probe, and signal pulses. The remaining elements can be obtained by permutation of these indices ($R_{1122} = R_{1133} = R_{2211} = \dots$). In a conventional transient grating experiment, the two pump pulses are at the same frequency and are time coincident and therefore their indices can be interchanged. In this case, $R_{1212} = R_{1221}$ and the number of independent tensor elements is further reduced:

$$R_{1111} = R_{1122} + 2R_{1212} \quad (17)$$

The tensor $\mathbf{R}$ can be decomposed into four tensors according to the origin of the sample response: population and density changes, electronic and nuclear optical Kerr effects:

$$\mathbf{R} = \mathbf{R}(p) + \mathbf{R}(d) + \mathbf{R}(K, e) + \mathbf{R}(K, n) \quad (18)$$

As they describe different phenomena, these tensors do not have the same symmetry properties. Therefore, the various contributions to the diffracted intensity can, in some cases, be measured selectively by choosing the appropriate polarization of the four waves.

Table 2 shows the relative amplitude of the most important elements of these four tensors.

The tensor $\mathbf{R}(p)$ depends on the polarization anisotropy of the sample, r , created by excitation with polarized pump pulses:

$$r(t) = \frac{N_{\parallel}(t) - N_{\perp}(t)}{N_{\parallel}(t) + 2N_{\perp}(t)} = \frac{2}{5} P_2[\cos(\gamma)] \exp(-k_{re}t) \quad (19)$$

where $N_{\parallel}$ and $N_{\perp}$ are the number of molecules with the transition dipole oriented parallel and perpendicular to the polarization of the probe pulse, respectively, γ is the angle between the transition dipoles involved in the pump and probe processes, P_2 is the second Legendre polynomial, and k_{re} is the rate constant for the reorientation of the transition dipole, for example, by rotational diffusion of the molecule or by energy hopping.

The table shows that $R_{1212}(d) = 0$, i.e., the contribution of the density grating can be eliminated with the set of polarization ($0^\circ, 90^\circ, 0^\circ, 90^\circ$), the so-called crossed grating geometry. In this geometry, the two pump pulses have orthogonal polarization, the polarization of the probe pulse is parallel to that of one pump pulse and the polarization component of the signal that is orthogonal to the polarization of the probe pulse is measured. In this geometry, $R_{1212}(p)$ is nonzero as long as there is some polarization anisotropy, ($r \neq 0$). In this case, the diffracted intensity is

$$I_{\text{dif}}(t) \propto |R_{1212}(p)|^2 \propto [\Delta C(t) \times r(t)]^2 \quad (20)$$

The crossed grating technique can thus be used to investigate the reorientational dynamics of molecules, through $r(t)$, especially when the dynamics of r is faster than that of ΔC . For example, **Fig. 10** shows the time profile of the diffracted intensity measured in the crossed grating geometry with rhodamine 6G in ethanol. The decay is due to the reorientation of the molecule by rotational diffusion, the excited lifetime of rhodamine being about 4 ns.

On the other hand, if the decay of r is slower than that of ΔC , the crossed grating geometry can be used to measure the population dynamics without any interference from the density phase grating. For example, **Fig. 11** shows the time profile of the diffracted intensity after excitation of a suspension of TiO_2 particles in water. The upper profile was measured with the set of

Table 2 Relative value of the most important elements of the response tensors $\mathbf{R}(i)$ and polarization angle ζ of the signal beam, where the contribution of the corresponding process vanishes for the set of polarization ($\zeta, 45^\circ, 0^\circ, 0^\circ$).

Process	R_{1111}	R_{1122}	R_{1212}	ζ
Electronic OKE	1	1/3	1/3	-71.6°
Nuclear OKE	1	$-1/2$	3/4	63.4°
Density	1	1	0	-45°
Population:				
$\gamma = 0^\circ$	1	1/3	1/3	-71.6°
$\gamma = 90^\circ$	1	2	$-1/2$	-26.6°
No correlation: $r=0$	1	1	0	-45°

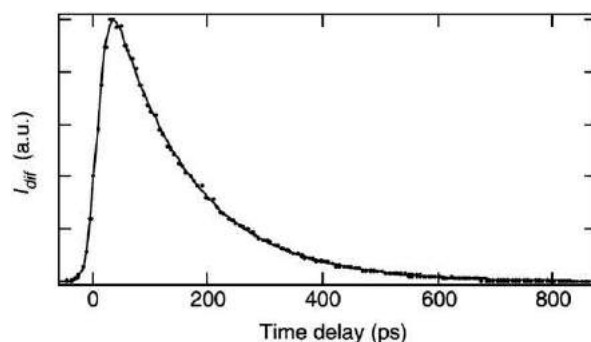


Fig. 10 Time profile of the diffracted intensity measured with a solution of rhodamine 6G with crossed grating geometry.

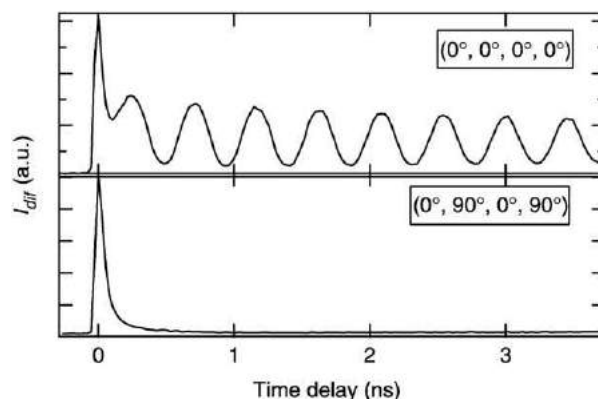


Fig. 11 Time profiles of the diffracted intensity after excitation at 355 nm of a suspension of TiO_2 particles in water and using different set of polarization of the four beams.

polarization $(0^\circ, 0^\circ, 0^\circ, 0^\circ)$ and thus reflects the time dependence of $R_{1111}(p)$ and $R_{1111}(d)$. $R_{1111}(p)$ is due to the trapped electron population, which decays by charge recombination, and $R_{1111}(d)$ is due to the heat dissipated upon both charge separation and recombination. The lower time profile was measured in the crossed grating geometry and thus reflects the time dependence of R_{1212} and in particular that of $R_{1212}(p)$.

Each contribution to the signal can be selectively eliminated by using the set of geometry $(\zeta, 45^\circ, 0^\circ, 0^\circ)$ where the value of the 'magic angle' ζ for each contribution is listed in Table 2. This approach allows for example the nuclear and electronic contributions to the optical Kerr effect to be measured separately.

Concluding Remarks

The transient grating techniques offer a large variety of applications for investigating the dynamics of chemical processes. We have only discussed the cases where the pump pulses are time coincident and at the same wavelength. Excitation with pump pulses at different wavelengths results to a moving grating. The well-known CARS spectroscopy is such a moving grating technique. Finally, the three pulse photon echo can be considered as a special case of transient grating where the sample is excited by two pump pulses, which are at the same wavelength but are not time coincident.

See also: Atomic Physics

Further Reading

- Bräuchle, C., Burland, D.M., 1983. Holographic methods for the investigation of photochemical and photophysical properties of molecules. *Angewandte Chemie International Edition Engl* 22, 582–598.
- Eichler, H.J., Günter, P., Pohl, D.W., 1986. *Laser-Induced Dynamics Gratings*. Berlin: Springer.
- Fleming, G.R., 1986. *Chemical Applications of Ultrafast Spectroscopy*. Oxford: Oxford University Press.
- Fourkas, J.T., Fayer, M.D., 1992. The transient grating: a holographic window to dynamic processes. *Accounts of Chemical Research* 25, 227–233.

- Hall, G., Whitaker, B.J., 1994. Laser-induced grating spectroscopy. *Journal of the Chemistry Society Faraday Trans* 90, 1–16.
- Levenson, M.D., Kano, S.S., 1987. *Introduction to Nonlinear Laser Spectroscopy*, revised edn. Boston: Academic Press.
- Mukamel, S., 1995. *Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.
- Rullière, C. (Ed.), 1998. *Femtosecond Laser Pulses*. Berlin: Springer.
- Terazima, M., 1998. Photothermal studies of photophysical and photochemical processes by the transient grating method. *Advances in Photochemistry* 24, 255–338.

Atomic Physics

G Kurizki and AG Kofman, Weizmann Institute of Science, Rehovot, Israel
D Petrosyan, Institute of Electronic Structure and Laser, Heraklion, Greece

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

$|... \rangle$ Atomic or photonic eigenstates [dimensionless]
 p Complex polarizability [cm^{-1}]
 $\kappa, C^{2/3}, B$ Coupling constant [s^{-1}]
 γ Decay rate [s^{-1}]
 ρ Density of modes [s]
 ω, Δ, s Frequency [s^{-1}]

ϕ Phase [dimensionless]
 α, β, c Probability amplitudes [dimensionless]
 G Reservoir spectral response [s^{-1}]
 a_0 Resonant absorption coefficient [cm^{-1}]
 W Spontaneous emission spectrum [s]
 k Wave vector [m^{-1}]

Introduction

The fabrication and study of dielectric structures whose refractive index is periodically modulated on a micron or submicron scale, known as photonic crystals (PCs), are attracting considerable interest at present. One-, two-, or three-dimensional (1D-, 2D-, or 3D-) periodic PCs exhibit photonic band gaps (PBGs), analogously to electronic band gaps in ordinary crystals, and can incorporate defects, designed to form localized narrow-linewidth (high-Q) modes at PBG frequencies. The advent of such structures opens up new perspectives in atomic physics and quantum optics, since they are expected to allow an unprecedented control over the spectral density of modes (DOM) and the spatial modulation of narrow-linewidth (high-Q) modes, in the microwave, infrared, and optical domains. An interesting situation arises when foreign atoms or ions – dopants – with transition frequencies within the PBG, are implanted in the PC. Then light near one of these frequencies resonantly interacts with the dopants and is concurrently affected by the PBG dispersion. Consequently, highly nonlinear processes with a rich variety of unusual PC-related features are anticipated. Such nonlinear optical processes, involving near-resonant transitions in PCs, undergo basic modifications as compared to the corresponding processes in free space, which are attributed to the strong suppression of the DOM within PBGs, to sharp bandedges and to intra-gap narrow lines associated with high-Q defect modes.

Results are detailed below for spontaneous emission of radiation in PCs and for the resonant interaction of atoms with the field of a single high-Q defect mode. These results stem from the failure of perturbation theory and the onset of strong field-atom coupling near sharp bandedges or narrow defect-mode lines in 2D and 3D PCs. 3D-PBGs are needed, in order to extinguish spontaneous emission in all possible directions of propagation and dipole orientations. If only one polarized atomic transition is involved in the spontaneous emission, 2D-PCs suffice for its suppression. For controlling strictly unidirectional field propagation, it is sufficient to resort to PBGs in 1D-periodic structures (Bragg reflectors or dielectric multi-layer mirrors).

As a further example of field-atom interactions in doped PCs, we will show that two ultraweak electromagnetic fields (or photons), coupled to appropriate transitions of the dopants in a PC, can mutually induce large phase shifts or drastic changes in absorption. Appreciable photon-photon correlations can then be established, which can be utilized in both classical and quantum optical communications.

Radiative Decay and Photon-Atom Binding in a PC

The interaction of a two-level atom with an arbitrary field-mode continuum can be described by the following second-quantized Hamiltonian in the rotating-wave approximation (RWA)

$$H = \hbar\omega_a|e\rangle\langle e| + \hbar\sum_{\vec{k}}\omega_{\vec{k}}a_{\vec{k}}^\dagger a_{\vec{k}} + \hbar\sum_{\vec{k}}\left[\kappa(\omega_{\vec{k}})a_{\vec{k}}^\dagger|g\rangle\langle e| + \text{hermitian constant}\right] \quad (1)$$

Here $|e\rangle$ and $|g\rangle$ are the excited and ground atomic states, respectively, ω_a is the atomic transition frequency, $a_{\vec{k}}^\dagger$ and $a_{\vec{k}}$ are the creation and annihilation operators of a field mode with wavevector $\vec{k}$, and frequency $\omega_{\vec{k}}$, and $\hbar\kappa(\omega_{\vec{k}})$ is the resonant coupling energy of this mode with the atomic dipole. The $\vec{k}$ -summation can be replaced by an integral over the continuum DOM in the frequency domain, $\sum_{\vec{k}} \rightarrow \int_0^\infty d\omega \rho(\omega)$, avoiding, for the sake of simplicity, the study of direction-(angle-)dependent DOM effects.

We consider a two-level atom that is excited at time $t=0$ in an empty, periodic dielectric structure. The wavefunction of the combined system field + atom can be cast in the general form:

$$|\Psi(t)\rangle = \alpha(t)|e, \{0_\omega\}\rangle + \int \beta_\omega(t)|g, 1_\omega\rangle \rho(\omega) d\omega \quad (2)$$

where $\{0_\omega\}$ signifies the completeness of field modes in the vacuum state and $|1\omega\rangle$ denotes single-photon occupation of the ω -mode. The corresponding Schrödinger equation can be solved with the initial condition $|\Psi(0)\rangle = |e, \{0_\omega\}\rangle$ by means of the continuum spectral response:

$$G(\omega) = |\kappa(\omega)|^2 \rho(\omega) \quad (3)$$

The analysis of Eq. (2), with $G(\omega)$ appropriate for photonic bandstructures, is aimed at revealing the prominent features of the atomic excitation decay $\alpha(t)$ and the corresponding emission spectrum.

In what follows, we consider a photonic bandstructure with several PBGs, separated by allowed bands. Each PBG is labeled by index i and has lower and upper cutoff frequencies ω_{Li} and ω_{Ui} respectively. Then $G(\omega) = 0$ for $\omega_{Li} < \omega < \omega_{Ui}$. Naively, one might expect that an excited atom would either be stable against single-photon decay, if ω_a is within a PBG, or decay completely at $t \rightarrow \infty$, if ω_a is anywhere in an allowed band. However, both statements turn out to be inaccurate.

Incomplete decay of $\alpha(t \rightarrow \infty)$ occurs if there is a stable eigenvalue (energy level) $\hbar\omega_i$ of the total (field-atom) Hamiltonian Eq. (1). Such an energy level is possible only if $G(\omega_i) = 0$, i.e., for ω_i in a PBG. It must satisfy:

$$\omega_i = \omega_a + \Delta(\omega_i) \quad (4)$$

$$\Delta(\omega_i) = \int_0^{\omega_{Li}} \frac{G(\omega')}{\omega_i - \omega'} d\omega' + \int_{\omega_{Ui}}^\infty \frac{G(\omega')}{\omega_i - \omega'} d\omega' \quad (5)$$

Here the integrals are the frequency shifts of the atomic resonance $\hbar\omega_a$ induced by the spectral parts of the reservoir situated, respectively, below and above the i th PBG (Fig. 1). If Eq. (4) holds, then the corresponding term in $\alpha(t)$ is proportional to $\exp(-i\omega_i t)$, and does not decay. Physically, such a stable level can be interpreted as representing the binding of the photon to the atom (photon-dressed atom), without the ability to leave the atomic vicinity, due to Bragg reflection in the PBG.

Assuming that ω_i occurs in the i th PBG, we observe that the first shift is positive whereas the second shift is negative, i.e., each part of the continuum repels the atomic level from its PBG edge. Eq. (4) has a solution if ω_a falls between the minimum and maximum values assumed by its left-hand side in the i th PBG, i.e., if

$$\omega_{Li} - \Delta(\omega_{Li}) < \omega_a < \omega_{Ui} - \Delta(\omega_{Ui}) \quad (6)$$

Incomplete decay occurs if this condition holds at least for one PBG. There may be, at most, one stable state in a single PBG (Fig. 1). However, a two-level atom in a periodic structure can have several stable dressed states simultaneously, when conditions hold for several PBGs. It is quite extraordinary that the conditions (Eq. (6)) for incomplete decay can be fulfilled with ω_a , not only in a PBG, but even in an allowed band, e.g., when $\Delta(\omega_{Ui})$ is negative or when $\Delta(\omega_{Li})$ is positive (Fig. 1(a) and (b)).

Equally counterintuitive is the converse possibility, of complete decay for ω_a within a PBG. When there is a single PBG and one of the inequalities in Eq. (6) is violated, complete decay will occur for $\omega_{Li} < \omega_a < \omega_{Li} - \Delta(\omega_{Li})$, if $\Delta(\omega_{Li}) < 0$, or $\omega_{Ui} - \Delta(\omega_{Ui}) < \omega_a < \omega_{Ui}$ if $\Delta(\omega_{Ui}) > 0$.

The possibility that one or several discrete, stable, states exist yields the corresponding wavefunction in the form:

$$|\Psi(t)\rangle = \sum_i \sqrt{c_i} |\psi_i\rangle e^{-i\omega_i t} + |\Psi_c(t)\rangle \quad (7)$$

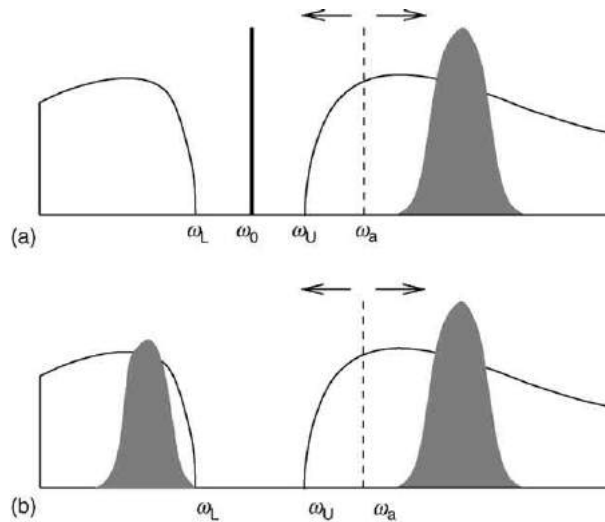


Fig. 1 Frequency shifts of the atomic resonance ω_a (dashed line) due to ‘repulsion’ (arrows) by asymmetric spectral parts of the continuum below and above the PBG: (a) Incomplete decay for ω_a in allowed band. The interaction with the continuum splits the state into a superposition of a stable part in the PBG (solid line) and a decaying part above ω_a (shaded peak). (b) Complete decay for same ω_a , due to larger shifts that split the state into decaying parts in two allowed bands.

Here the summation is over all discrete atomic photon-dressed states with energies $\hbar\omega_i$, $|\psi_i\rangle$ is a discrete-(dressed-)state eigenfunction of the Hamiltonian (Eq. (1)) normalized to unity, and weighted by the amplitude $\sqrt{c_i} = [1 + \int_0^\infty d\omega G(\omega)/(\omega - \omega_i)^2]^{-1/2}$. The dressed state $|\psi_i\rangle$ consists of an excited-state component and a photon-bound ground-state component.

The population of the excited state for long times:

$$|\alpha(t)|^2 = \sum_{i,i'} c_i c_{i'} \cos(\omega_i - \omega_{i'})t \quad \text{for } t \rightarrow \infty \quad (8)$$

is a nonzero time-constant if there is one discrete stable state, whereas in the case of several such states, it undergoes beats at the frequencies corresponding to their energy differences. The splitting of an excited state $|e\rangle$ into superposed stable states, oscillating at bandgap frequencies ω_i , whose amplitudes c_i , and eigenfrequencies ω_i are controllable by the atomic transition detuning from cutoff, constitutes spontaneous coherence control.

If, however, ω_a is far from the cutoff, then Eq. (2) results in an exponentially decaying amplitude:

$$\alpha(t) \approx e^{-(\gamma_a + i\tilde{\omega}_a)t} \quad (9)$$

where $\tilde{\omega}_a = \omega_a + \Delta_a$, Δ_a , and γ_a being the effective spectral shift and width of the decaying atom. This regime holds for a locally smooth $G(\omega \approx \omega_a)$ such that

$$|\gamma'_a|, |\Delta'_a| \ll 1; |\gamma'_a \Delta_a| \ll \gamma_a; |\Delta_a| \ll |\omega_a - \omega_g| \quad (10)$$

where ω_g is the bandedge frequency nearest to ω_a and the primes denote differentiation with respect to ω_a .

We now apply the foregoing general results to a model DOM distribution. This distribution is derived on keeping the lowest term in the Taylor expansion of the dispersion relation $\omega(\vec{k})$ near a photonic bandedge ω_U (the effects of the further PBG edge are neglected). This yields the effective mass approximation:

$$\omega \approx \omega_U + \sum_{i=x,y,z} (k - k_U)_i^2 / m_i \quad (11)$$

with $1/m_i = (1/2)(\partial^2 \omega / \partial k_i^2)|_{\omega=\omega_U}$. In a structure with period L , k_U satisfies the Bragg condition $k_U = \pi/L$. The corresponding DOM in a 3D-periodic structure with an allowed symmetry may be approximated as

$$\rho(\omega) \sim (\omega - \omega_U)^{(1-D)/2} \theta(\omega - \omega_U) \quad (12)$$

where θ is the step function and D is the dimension of the Brillouin-zone surface spanned by bandedge modes with vanishing group velocity. In realistic photonic structures $D \leq 2$, $D=2$ corresponding to completely isotropic dispersion (spherical Brillouin-zone surface) and $D=0$ corresponding to an anisotropic three-dimensional Brillouin zone. Both cases can be represented, respectively, as the limits $\varepsilon \rightarrow 0$ and $\varepsilon \rightarrow \infty$ of the function:

$$G(\omega) = \frac{C}{\pi} \frac{\sqrt{\omega - \omega_U}}{\omega - \omega_U + \varepsilon} \theta(\omega - \omega_U) \quad (13)$$

where ε is the 'cutoff-smoothing' parameter and C is the continuum coupling constant. Depending on whether $C^{2/3}/\varepsilon$ is greater or less than one, the continuum is close to the case $\varepsilon=0$ ($D=2$) or $\varepsilon \rightarrow \infty$ ($D=0$), respectively.

Using the properties of the model DOM (Eq. (13)), we can infer the criteria for the two regimes discussed above:

- (i) The conditions for incomplete decay, Eq. (6), are now (Fig. 2) $\Delta_c \equiv \omega_a - \omega_U < C\varepsilon^{-1/2}$. Hence, the abrupt, singular cutoff of the DOM with $D=2$ ($\varepsilon \rightarrow 0$) implies the existence of a discrete state for any ω_a , either inside or outside the PBG. The energy of the discrete state, $\hbar\omega_0$ which must lie in the PBG, is found to be a real and positive root of the equation $\omega_a - \omega_0 = C/(\sqrt{\omega_U - \omega_0} + \sqrt{\varepsilon})$.
- (ii) The conditions for nearly exponential decay, Eq. (10), can be shown to reduce now to the requirement that ω_a be in the allowed zone sufficiently far from cutoff, $\Delta_c \gg \min\{C^{2/3}, C\varepsilon^{1/2}\}$. Under this condition, the intermediate-time exponential decay of the excited state (Fig. 3) is given to first approximation by Eq. (9), with $\gamma_a = C\sqrt{\Delta_c}/(\varepsilon + \Delta_c)$ and $\Delta_a = -C\sqrt{\varepsilon}/(\varepsilon + \Delta_c)$.

The resulting atomic frequency shift Δ_a is negative in the present model, vanishing for $\varepsilon \rightarrow 0$ ($D=2$). The long-time behavior of $\alpha(t)$ can be shown to exhibit a tail decaying as $t^{-3/2}$ (or $t^{-1/2}$ at $\Delta_c = C\varepsilon^{-1/2}$) and oscillating at the cutoff frequency ω_U (Fig. 3).

With the increase of ε , the smoothing inhibits the decay more strongly for ω_a at cutoff, $\Delta_c=0$, because $G(\omega_a \simeq \omega_U)$ is now weaker and the stable-state probability c_0 (Eq. (7)) is correspondingly larger. By contrast, at a large detuning from cutoff Δ_c the large- ε smoothing allows the decay to become complete, and the corresponding spectrum is entirely Lorentzian.

A defect in the periodic structure can produce a narrow-linewidth local mode in the PBG, whose spectral response is describable by a Lorentzian:

$$G_d(\omega) = \frac{\gamma_d}{\pi} \frac{\Gamma_d^2}{\Gamma_d^2 + (\omega - \omega_d)^2} \quad (14)$$

where γ_d characterizes the coupling strength of the atomic dipole with the defect field, whereas ω_d and Γ_d stand for the line center and width, respectively. The presence of a nonvanishing DOM in the PBG, due to a defect, causes spontaneous emission in the

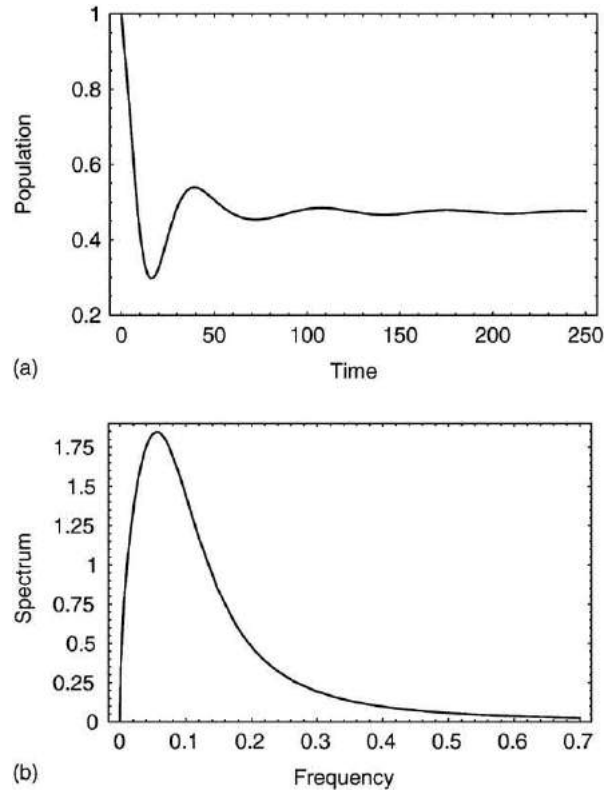


Fig. 2 (a) Incomplete decay of the excited state population as a function of $\gamma_c t$, where $\gamma_c = C\epsilon^{-1/2}$, for $\epsilon = 10^{-3}$, at cutoff $\Delta_c = 0$. The beat period is $2\pi/(\omega_U - \omega_0)$ and the nondecaying probability is $c_0^2 = 4/9$. (b) The corresponding spectrum. The frequency is normalized to γ_c .

PBG spectral range. This broadens the discrete state ω_0 , which becomes metastable (see Fig. 4). The oscillator strength of the line at ω_0 is then proportional to c_0 .

EIT and Cross-Coupling of Photons in Doped PCs

Nonlinear effects, whereby one light beam influences another, require strong fields or else light confinement in a high-Q cavity. The analysis above has shown that by choosing an appropriate detuning of an atomic transition from the PBG cutoff, the spontaneous coherence is established between the states of an initially excited atom. The ability to control this coherence, by varying the detuning, opens interesting perspectives for optical processes in PCs containing multilevel atoms with a resonant transition near a bandedge. From among such processes, we discuss here electromagnetically induced transparency (EIT) and its applicability to nonlinear photon switching or giant cross-phase modulation.

Let us examine the nonlinear coupling of two weak (single-photon) optical fields E_a and E_b with the frequencies ω_a and ω_b , respectively, propagating along the z -axis in a PC dilutely doped with identical four-level atoms (Fig. 5(a)). These fields interact with the atoms via the transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively, while the transition $|2\rangle \rightarrow |3\rangle$ is coupled to the structured mode-continuum $\rho(\omega)$ in the PC (Fig. 5(b)). Initially the atoms are in the ground state $|1\rangle$ and the continuum is in the vacuum state $\{0_\omega\}$. Then the wavefunction of the system at the position z_l of the l th atom reads:

$$|\Psi(z_l, t)\rangle = \alpha_1 |1, \{0_\omega\}\rangle + \alpha_2 |2, \{0_\omega\}\rangle + \int \beta_{3,\omega} |3, 1_\omega\rangle \rho(\omega) d\omega + \int \beta_{4,\omega} |4, 1_\omega\rangle \rho(\omega) d\omega \quad (15)$$

The consecutive terms in Eq. (15) denote the atom being in states $|1\rangle$, $|2\rangle$, $|3\rangle$, or $|4\rangle$, with zero $\{0_\omega\}$ or one $|1_\omega\rangle$ photon in mode ω whose DOM is $\rho(\omega)$, and α_1 , α_2 , $\beta_{3,\omega}$, or $\beta_{4,\omega}$ are the corresponding probability amplitudes. Upon making the weak-field linear-response approximation, we can set $\alpha_1 \simeq 1$ and solve the set of equations for the slowly varying (compared to an optical cycle) probability amplitudes α and β , using the perturbation theory. Under these conditions, we effectively obtain a free-space propagation of the E_b field. By contrast, the evolution of the E_a field in the slowly varying envelope approximation, is given by $E_a(z, t) = E_a(0, t - z/v_g) \exp(ipz)$. We thus see that the real part of the macroscopic complex polarizability p is responsible for the phase shift ϕ_a of the E_a field, $\phi_a = \text{Re}(p)z$, while the probability of the absorption $\mathcal{A}$ of the field depends on the imaginary part of p , $\mathcal{A} = 1 - \exp[-2\text{Im}(p)z]$. The polarizability is expressed by

$$p = a_0 \frac{i\gamma_2/2}{\gamma_2/2 - i\Delta_a + I(\Delta_a)} \quad (16)$$

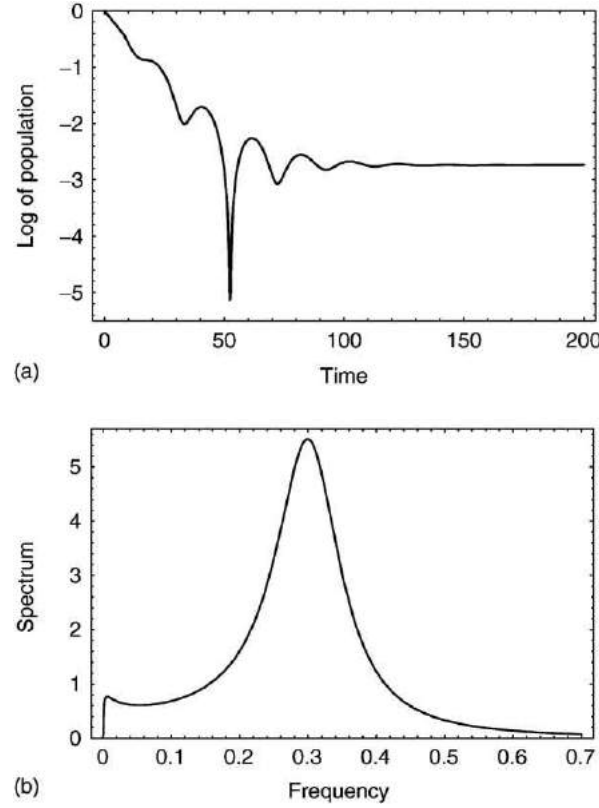


Fig. 3 As in Fig. 2, for a large detuning from the sharp cutoff, $\varepsilon=10^{-3}$, $\Delta_c=0.3$. (a) $\log_{10}|z|/\hbar^2$ is plotted as a function of i . The nearly complete decay ($c_0^2 \sim 10^{-3}$) is modulated by beats with frequency Δ_c . A power tail obtains for $t \gg \gamma_a^{-1} \approx \Delta_c^{1/2}/C$, decreasing as $t^{-1/2}$ for $t \ll \Delta_c^2/C^2$ and as $t^{-3/2}$ for $t \gg \Delta_c^2/C^2$. (b) The nearly Lorentzian spectrum, $w(\omega)$, has a small peak near ω_U , due to the sharply peaked DOM near the cutoff.

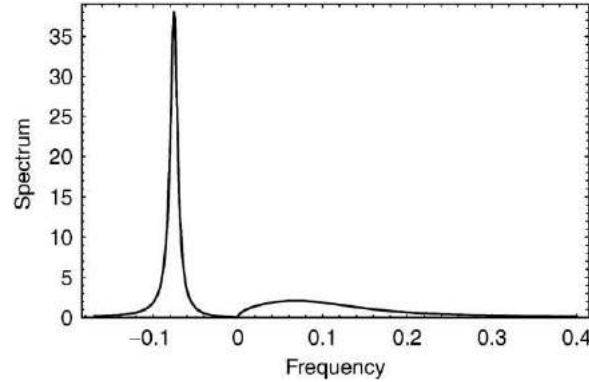


Fig. 4 Spectrum $w(\omega)$ (same units as in Figs. 2 and 3) for ω_a at cutoff, $\Delta_c=0$, in the presence of a defect in the PBG, $\omega_d = -10$, $\gamma=0.1$, $\Gamma_d=3$, $\varepsilon=10^{-3}$. The oscillator strength is now roughly equally shared between the distribution above cutoff (same as in Fig. 2(b)) and the defect peak at ω_0 .

where $a_0 \equiv \sigma_0 N$ is the linear resonant absorption coefficient on the atomic transition $|1\rangle \rightarrow |2\rangle$, with σ_0 the resonant absorption cross-section and N the density of doping atoms, γ_2 is the radiative width of state $|2\rangle$, $\Delta_a = \omega_a - \omega_{21}$ is the detuning from the atomic resonance ω_{21} , and $I(\Delta_a)$ is the integral of the saturation factor over the structured DOM. The group velocity v_g is expressed as $v_g = \partial_k \omega_a = [n_a/c + \partial_{\omega_a} \text{Re}(p)]^{-1}$, where n_a is the (averaged) refraction index at the frequency ω_a .

To calculate $I(\Delta_a)$, we assume the isotropic PBG model, Eq. (13) with $\varepsilon \rightarrow 0$, with the atoms doped at the positions of the local defects in the PC separated by a distance d from each other. These defect modes in the PBG are localized around each atomic site in a volume $V_d \simeq (rL)^3$ of several $(r)^3$ lattice cells L^3 , with $L \simeq \pi c/\omega_{23}$, and serve as effective high-Q cavities, Eq. (14) with $\Gamma_d \ll 1$. Assuming $|\Delta_b| \gg |\Delta|$, $|\Delta_a|$, γ_4 , where $\Delta_b = \omega_b - \omega_{43}$ and $\Delta = \omega - \omega_{23}$, the integration leads to

$$I(\Delta_a) = \frac{B_d^2}{\gamma_{31} - i(\Delta_a - \Delta_d - s_3)} - \frac{B_U^{3/2}}{\sqrt{i\gamma_{31} + (\Delta_a - \Delta_U - s_3)}} \quad (17)$$

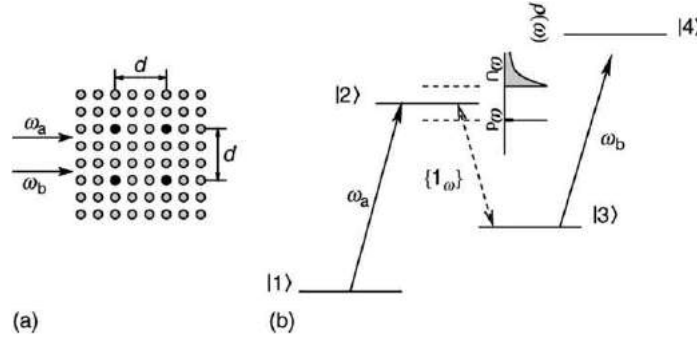


Fig. 5 (a) Photonic crystal dilutely doped with atoms located at black dots. (b) Four-level atom coupled to a structured continuum $\rho(\omega)$ near the band-edge or defect mode frequencies (DOM plotted) via the intermediate transition $|2\rangle \rightarrow |3\rangle$ and interacting with two weak fields E_a and E_b at the sideband transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively.

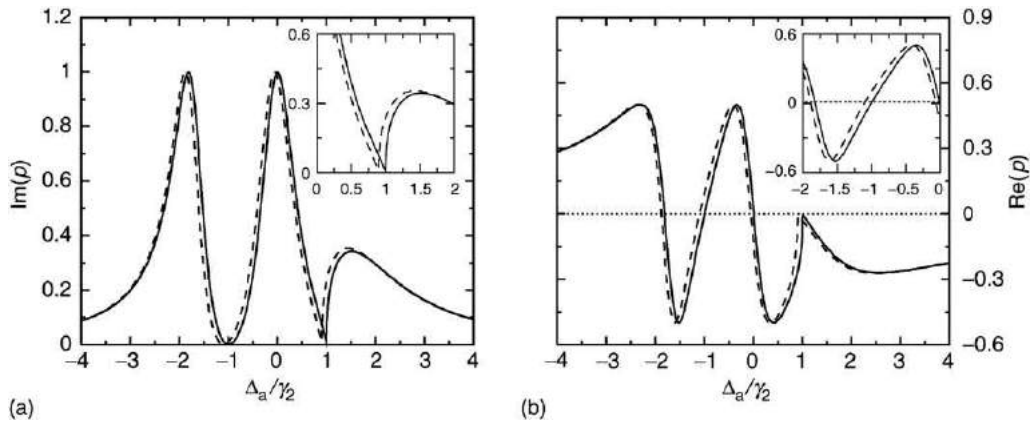


Fig. 6 (a) Imaginary and (b) real part of the complex polarizability ρ as a function of the detuning Δ_a for the case $E_b = 0$ (solid lines) and $E_b \neq 0$ (dashed lines). The parameters (normalized by γ_2) are: $\Delta_d = -1$, $\Delta_U = 1$, $\gamma_{31} = 0.001$, $B_d = B_U = 1$, $s_3 = -0.1$, and $a_0 = 1 \text{ cm}^{-1}$. The insets magnify the important frequency regions.

where $\Delta_{d,U} = \omega_{d,U} - \omega_{23} \ll \omega_{23}$ are the detunings of the defect-mode and PBG-edge frequencies from the atomic resonance ω_{23} , γ_{31} the $|1\rangle \leftrightarrow |3\rangle$ decoherence rate, $s_3 = (\mu_{34}^2 / \hbar^2 \Delta_b) |E_b|^2$ is the E_b field-induced ac Stark shift of level $|3\rangle$ (μ_{ij} is the atomic dipole matrix element on the transition $|i\rangle \rightarrow |j\rangle$), and B_d and B_U are the coupling constants of the atom with the structured reservoir, whose main contributions are near ω_d and ω_U .

To illustrate the results of the foregoing analysis, we plot in **Fig. 6** the imaginary and real parts of the polarizability (Eq. (16)). Consider first the case of one incident field E_a ($E_b = 0$). Clearly, two frequency regions, $\Delta_a \sim \Delta_d$ and $\Delta_a \sim \Delta_U$, where the absorption vanishes and, at the same time, the dispersion slope is steep, are of particular interest. One can see that there is, however, a substantial difference between the spectra in the foregoing frequency regions, for the following physical reasons. First, in the vicinity of Δ_d , the atom interacts with the defect mode as in a high- Q cavity. This strong interaction ‘dresses’ the atomic states $|2\rangle$ and $|3\rangle$, thereby splitting the spectrum around $\Delta_a \simeq \Delta_d$ by the amount equal roughly to $2B_d$. Near the two-photon Raman resonance $\Delta_a = \Delta_d$, the two alternative transition paths $|1\rangle \rightarrow |2\rangle$ (the direct transition) and $|1\rangle \rightarrow |2\rangle \rightarrow |3\rangle \rightarrow |2\rangle$ (transition via state $|3\rangle$) interfere destructively with each other, cancelling thus the absorption of the E_a field and the medium becomes transparent to the radiation. This effect has been widely studied in atomic vapors, where the transition $|2\rangle \rightarrow |3\rangle$ is strongly driven by a coherent laser field, and is called electromagnetically induced transparency (EIT). The transparency window is rather broad and is given by the inverse Lorentzian (see the first term on the right-hand side of Eq. (17)). Due to the steepness of the dispersion curve, the corresponding group velocity is much smaller than the speed of light, $v_g \simeq 2B_d^2 / (\gamma_2 a_0) \ll c$, which leads to a large delay time $T_{\text{del}} = \zeta / v_g$ at the exit $z = \zeta$ from the medium. One has to keep in mind, however, that the absorption-free propagation time is limited by the EIT decoherence time $T_{\text{del}} < \gamma_{31}^{-1}$, which imposes a limitation on the length ζ of the active PC medium. Second, in the vicinity of Δ_U , the strong interaction of the atom with the continuum near the bandedge ω_U causes the Autler–Townes splitting of level $|2\rangle$ into a doublet with a separation equal roughly to B_U . One component of this doublet is shifted out of the PBG, while the other one remains within the gap and forms the photon–atom bound state. Consequently, there is vanishing absorption and rapid variation of the dispersion at $\Delta_a \simeq \Delta_U$. Since the transparency region is very narrow with a width $\delta\omega \sim \gamma_{31} (\ll \gamma_2, B_U)$, for an absorption-free propagation of the E_a pulse, its temporal width τ_a should satisfy the condition $\tau_a > \pi / \delta\omega$. Simultaneously, a small deviation from the condition $\Delta_a = \Delta_U$ will lead to a strong increase in the absorption of the E_a field.

Let us now switch on the E_b field. As seen from Eq. (17), its effect is merely to shift the spectrum by the amount equal to s_3 (Fig. 6). This shift, however, will have different implications in the two frequency regions distinguished above: if $s_3 \ll B_d$, i.e., the Stark shift is smaller than the width of the EIT window at Δ_d , the medium will still remain transparent for an E_a field with the detuning $\Delta_a = \Delta_d$, but its phase will experience an appreciable nonlinear shift ϕ_a , given by $\phi_a = \text{Re}(p)z \simeq -\gamma_2 a_0 s_3 z / (2B_d^2)$. On the other hand, for an E_a field with the detuning $\Delta_a = \Delta_U$, the medium, which is transparent for $E_b = 0$, $\text{Im}(p) \ll a_0$, will become highly absorptive (opaque) even for such a small frequency shift as s_3 (provided $s_3 < 0$ and $|s_3| > \Delta\omega$), $\text{Im}(p) \simeq \gamma_2 a_0 \sqrt{|s_3|} / (2B_U^{3/2})$ and thus acting as an ultrasensitive, effective switch.

The remaining question is how to maximize the interaction between the E_a pulse, which propagates with a small group velocity, and the E_b pulse, which propagates with a velocity close to the speed of light. The interaction between the fields is maximized if: they enter the medium simultaneously; the transverse shapes of their envelopes overlap completely; and the pulse length l_b of the E_b field satisfies the condition $(l_b + \zeta)/c \leq \zeta/v_g$. Then the E_b pulse leaves the medium not later than the E_a pulse. The effective interaction length between the two pulses is, therefore, $z_{\text{eff}} \sim l_b v_g/c \leq \zeta$, after which the two pulses slip apart. Thus, the presence of the E_b (control) field induces either strong absorption or a large phase shift of the E_a (signal) field, depending on the frequency region employed.

The effects surveyed above reveal unusual features of spontaneous emission and photon–atom binding in PCs, atomic interaction with the field of a high-Q defect mode, and nonlinear coupling of two fields via four-level dopant atoms in PCs. These effects are of fundamental interest. In addition, they can serve as the basis for highly efficient optical communications and data processing, in either the classical or the quantum domain, by providing two key elements: ultrasensitive nonlinear phase-shifters and photon switches.

Further Reading

- Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. Photonic Crystals: Molding the Flow of Light. Princeton: Princeton University Press.
- John, S., Wang, J., 1991. Quantum optics of localized light in a photonic band gap. *Physics Review B* 43, 12772–12789.
- Kofman, A.G., Kurizki, G., Sherman, B., 1994. Spontaneous and induced atomic decay in photonic band structures. *Journal of Modern Optics* 41, 353–384.
- Lambropoulos, P., Nikolopoulos, G.M., Nielsen, T.R., Bay, S., 2000. Fundamental quantum optics in structured reservoirs. *Reports on Progress in Physics* 63, 455–503.
- Loudon, R., 1983. *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.
- Meystre, P., Sargent III, M., 1991. *Elements of Quantum Optics*. Berlin: Springer Verlag.
- Petrosyan, D., Kurizki, G., 2001. Photon–photon correlations and entanglement in doped photonic crystals. *Physics Review A* 64.023810-1-6.
- Sakoda, K., 2001. *Optical Properties of Photonic Crystals*, Springer Series in Optical Sciences, 80., Berlin: Springer Verlag.
- Scully, M.O., Zubairy, M.S., 1997. *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Terahertz Physics of Semiconductor Heterostructures

Juraj Darmo and Karl Unterrainer, Vienna University of Technology, Vienna, Austria

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Semiconductor nanostructures are very attractive quantum systems which allow engineering of wavefunctions and transition energies. By changing size and geometry the optical and electronic properties of nanostructures can be designed in a wide range. This makes nanostructures very interesting for possible applications and for fundamental questions about light matter interactions. The quantum confinement leads to the occurrence of quantized states. The optical matrix element for transitions between quantized states can be orders of magnitude larger than that of atoms and the selection rules can be adjusted to meet the requirements of specific experiments. Thus, semiconductor nanostructures are in the center of solid state physics and device related research. Regarding application, optoelectronic devices have been targeted due to the possibility to design the wavelength especially in the infrared and THz spectral region. The THz spectral region is scientifically very interesting since many fundamental resonances (co)exist there: Collective excitations of electrons, phonons, impurity transitions and quantized transitions. Consequently, THz spectroscopy was developed to study these elementary processes. This had been quite difficult due to the lack of sources, but the realization of femtosecond laser generated THz pulses provided an ideal tool to study quantized transitions. Modern THz time-domain spectroscopy (TDS) allows performing time-resolved and non-linear spectroscopy which was previously impossible or restricted to few large scale facilities. Semiconductor nanostructures allow the realization of quantum optical systems, for example, by coupling several different quantum wells which can then be studied by TDS. Important questions are connected to the lifetimes of the electrons of excited states, to dephasing times and to collective effects. In addition, the possibility to externally control the quantized electrons by an electrical field is very attractive for experiments as well as for applications.

The development of THz quantum cascade lasers (Köhler *et al.*, 2002) has shown in a very impressive way that employing quantum structure can lead to conceptual new optoelectronic devices. The targeted spectral range spans from the THz range, through the infrared up to the telecommunication band. In the THz and infrared the quest is for sources and amplifiers while fast modulators are needed for telecommunications. For all these proposals the population dynamics and the dephasing times are the most crucial parameters. Ultrafast spectroscopy has enabled to study this times in a very direct way. For the measurement of carrier dynamics in nanostructures, pulsed excitation is combined with ultrashort THz probe pulses to perform time-resolved intersubband absorption spectroscopy. With the rapid developments in THz-technology a new class of non-equilibrium phenomena in nanostructures can be studied. With the obtained knowledge the realization and further development of monolithic semiconductor THz devices has become possible.

In Section Few Cycle THz Spectroscopy of Semiconductor Quantum Structures of this article, we discuss experiments with semiconductor quantum wells. THz Time-domain spectroscopy is employed for investigating the optical response of an intersubband transition. The study of parabolic quantum wells shows how the intersubband transitions are governed by collective effects and how these collective excitations react to short THz pulses. A detailed study of relaxation times is performed for a coupled quantum well system. High intensity THz pulses allow the observation of non-linear processes in a multi-level quantum well.

In Section Few Cycle THz Spectroscopy of Quantum Cascade Heterostructures, time-resolved THz spectroscopy is employed to study THz Quantum Cascade (QC) heterostructures and lasers with a focus on gain dynamics, dispersion and broadband amplification. In addition, few cycle THz pulses are used to injection seed a broadband THz QCL resulting in very short THz pulses.

Few Cycle THz Spectroscopy of Semiconductor Quantum Structures

There is a large number of studies that use THz pulses for the investigation of semiconductor bulk materials. The complex index of refraction in the THz regime gives insight in the conductivity (i.e., mobility of free carriers) of the material. Free carriers can be either provided by doping or by photoexcitation across the bandgap. There are compelling reasons to use THz pulses to excite and probe carriers in semiconductor quantum structures. The photon energies are comparable to the subband spacing, kinetic energies of carriers, impurity transition and phonon energies. In addition, THz radiation does not generate electron-hole pairs so that the experiments directly probe free carriers rather than excitons.

Linear Response

In an initial experiment (Heyman *et al.*, 1998) THz pulses were used to investigate a modulation-doped GaAs/AlGaAs multiple quantum well ($d=51$ nm) with a transition energy between the two lowest subbands of 1.5 THz and a carrier density of $n_s=2.75 \times 10^{10} \text{ cm}^{-2}$ only populating the lowest subband. The THz pulses are coupled into the cleaved edge of the quantum well sample so that the electric field is parallel to the growth direction which is necessary to excite the intersubband transition. On top

of the sample an aluminum Schottky gate is evaporated and the well is contacted with ohmic contacts. The wells can be depleted of carriers by applying a negative gate voltage to the Schottky gate (−10 V gate bias). By modulating the voltage between 0 and −10 V the response of the intersubband system can be measured. This modulation technique allows separating the intersubband signal from the response of the SI GaAs substrate that causes a substantially chirped THz pulse due to the frequency-dependent index of refraction. The intersubband response is measured with a cross-correlation technique: Femtosecond laser pulses (100 fs) from a Ti:Sapphire laser are split into two pulse trains. The first one generates THz “driving” pulses which are transmitted through the quantum well. The transmitted pulses are then mixed at the detector with short THz “probe” pulses generated in the second path by photoexcitation of a low-temperature grown GaAs chip. The duration of the probe pulses (~100 fs) determines the time resolution and the probe delay controls the time difference. This allows studying the time dependence of amplitude and phase of the transmitted THz pulse. Fig. 1 shows the cross-correlations signal of the QW. The signal from the carriers rises during the first 2 ps in response to the exciting THz pulses and then continues to oscillate at a constant frequency of 1.5 THz. Due to the femtosecond time resolution in this experiment it is possible to observe the intersubband electrons even during the process of excitation. At first the incident THz pulse generates a coherent superposition between the first and second subband. After the THz driving pulse is over (> 2 ps) the carriers continue to oscillate and radiate at the intersubband transition frequency and the signal is damped out by the free induction decay. Since the amplitude and phase of the radiated electric field is recorded, not only the absorption due to the QW electrons but also the phase delay associated with the absorption can be extracted (Fig. 1).

In a subsequent experiment (Kersting *et al.*, 2000) the capability of studying amplitude and phase of THz pulses was used to study a modulation-doped parabolic quantum well. The advantage of parabolic quantum wells is that the transition frequency is independent of the carrier concentration (Kohn’s theorem) or the applied electric field. The transition frequency for the electrons in the parabolic quantum well is given by

$$\omega_0 = \sqrt{\frac{8\Delta}{L^2 m^*}}$$

where Δ is the energetic depth of the well, L is the well width, and m^* the effective mass. The THz radiation was coupled to the intersubband transition via a metallic Schottky grating. Fig. 2(a) and (b) shows the response of the QW electrons to resonant THz radiation. The electrons follow the driving pulse and continue to radiate at the parabolic quantum well frequency. The response to a non-resonant THz pulse is shown in Fig. 2(c)–(e). First, the electrons inside the QW follow the driving pulse but when the driving pulse is over a phase jump occurs and the electrons lock to their eigenfrequency. The electrons oscillate and emit at the parabolic quantum well resonance frequency and slowly lose their coherence. This behavior was explained by using time-dependent perturbation theory for a two level system with the THz pulse as perturbation. For the time-dependent wave function the ansatz $\Psi(t) = a_1(t)\Psi_1 + a_2(t)\Psi_2$ was used. The time-dependent $a_1(t)$ and $a_2(t)$ coefficients were computed using the measured driving THz

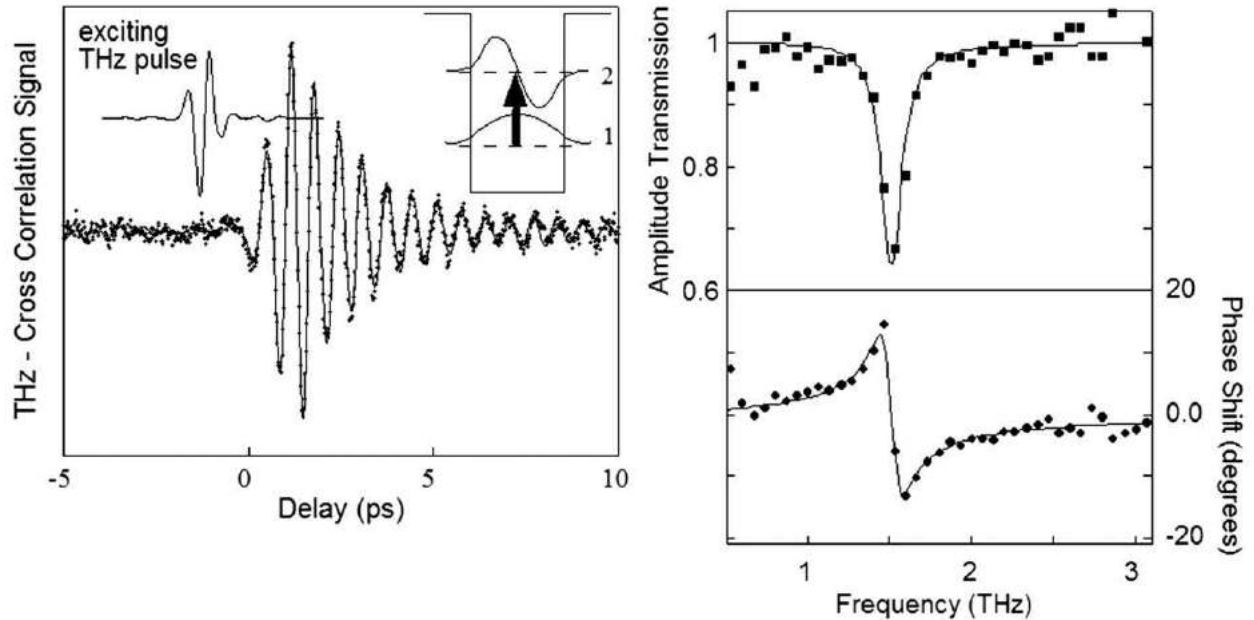


Fig. 1 Left: Measured cross-correlation signal obtained by modulating the charge density in the 51 nm quantum well sample, and recording the change in signal with a lock-in amplifier. The THz pulse excites electrons into a coherent superposition of the $n=1$ and $n=2$ subband states. The resulting oscillating polarization radiates, producing the oscillations visible in the signal. The upper trace shows the signal recorded with no sample (Adopted from Heyman, J., Kersting, R., Unterrainer, K., 1998. Time-domain measurement of intersubband oscillations in a quantum well. Applied Physics Letters 72, 644.). Right: Absorption coefficient and index of refraction in the quantum well sample due to the intersubband transition. Solid points are calculated from the time-domain data.

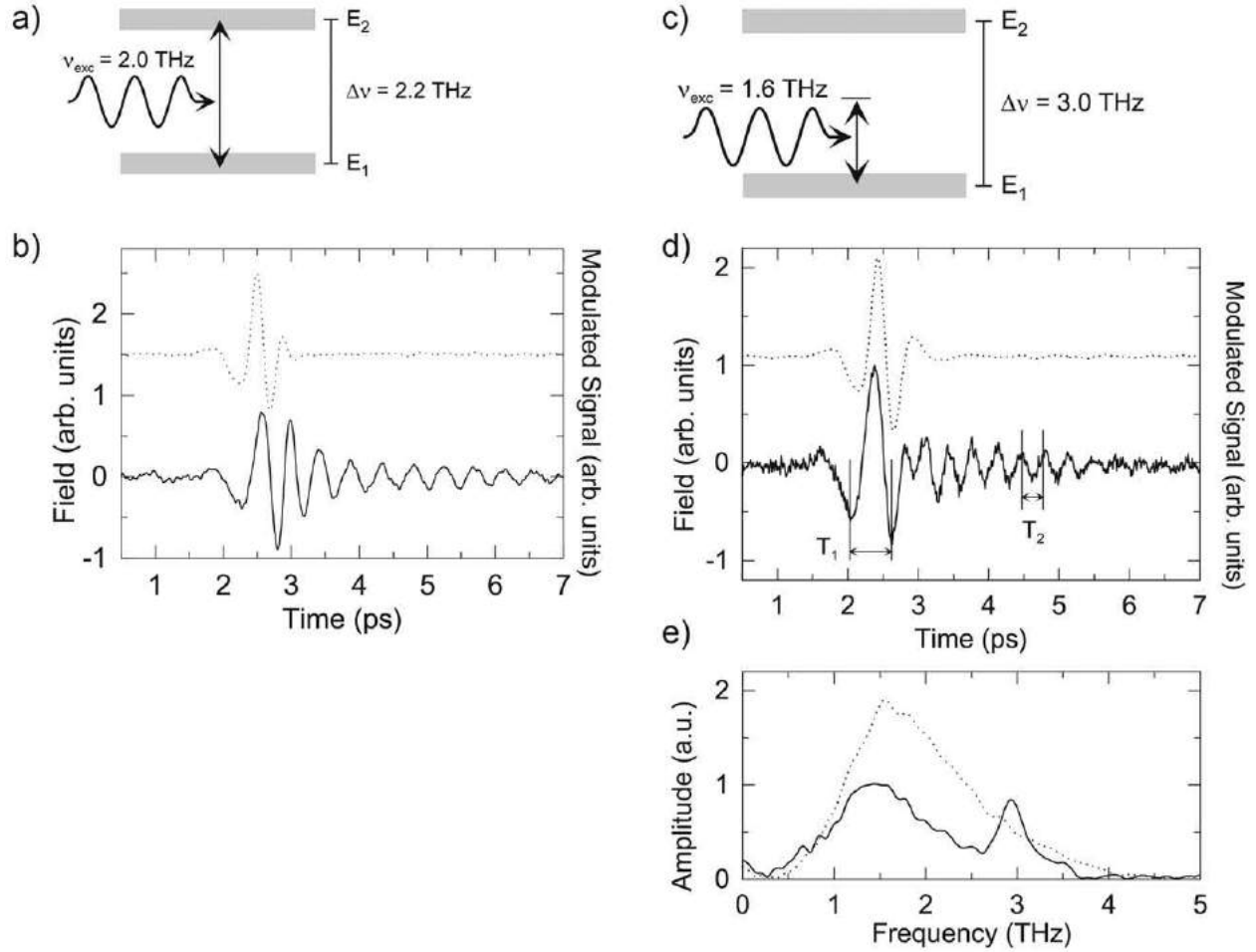


Fig. 2 Resonant and non-resonant excitation of a parabolic quantum well. Adopted from Kersting, R., Bratschitsch, R., Strasser, G., Unterrainer, K., Heyman, J., 2000. Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25, 272.

pulse as input. The comparison to the observed THz transmission is very good and explains the abrupt frequency change as well as the observed phase jump at the end of the driving pulse. This shows that Kohn's theorem even holds for impulsive and broadband excitation.

Carrier Density Dependence

With time-resolved THz spectroscopy we have also the possibility of using synchronized laser pulses with photon energies above the bandgap. These laser pulses can be used to excite electrons into the subbands of undoped quantum wells. The electrons in the subbands can then be studied by THz pulses (Müller *et al.*, 2004) providing a powerful method to access the dynamical processes in semiconductor quantum wells. It was used to investigate the carrier density dependence of the intersubband relaxation in a double quantum well structure. In this experiment an interband pump pulse excites electrons into the first and second subband of a double QW with a $|1\rangle \rightarrow |2\rangle$ level spacing smaller than the LO phonon energy. The time evolution of the electron population in these two subbands is monitored by probing the intersubband absorption to a third (empty) level. Ultrabroadband THz pulses are generated by phase-matched difference frequency mixing in 30 μm -thick GaSe crystal (Huber *et al.*, 2000) and detected with an interferometric cross-correlation setup (Heyman *et al.*, 1998). The resulting time-dependent THz absorption spectrum exhibits two resonances, corresponding to the $|1\rangle \rightarrow |3\rangle$ and $|2\rangle \rightarrow |3\rangle$ intersubband absorption, respectively. On basis of the absorption spectra the carrier density dependence of the relaxation time between states $|2\rangle$ and $|1\rangle$ can be obtained.

Time-resolved absorption spectra, recorded at a photo-excited sheet carrier density of $1 \times 10^{10} \text{ cm}^{-2}$, are shown in Fig. 3. The spectra exhibit absorption peaks at 27 THz and 30 THz, corresponding to the $|2\rangle \rightarrow |3\rangle$ and $|1\rangle \rightarrow |3\rangle$ intersubband transition, respectively. The amplitude of the low-energy peak shows a monotonous decrease with time-delay after excitation due to intersubband relaxation. The amplitude of the second peak, however, rises slightly in the beginning and decays afterwards due to carrier recombination. Since the spectrally integrated absorption is directly proportional to the subband population, it is possible to determine the population dynamics on the basis of the time-resolved absorption spectra. About 40% of the photoexcited electrons

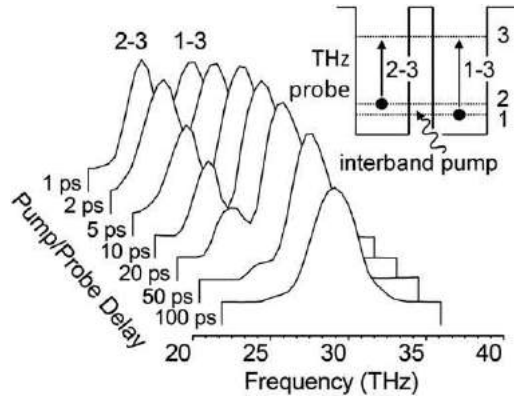


Fig. 3 Intersubband absorption spectra at different time-delays after the interband pump pulse for an excitation density of $n_s = 1 \times 10^{10} \text{ cm}^{-2}$. The inset shows the double quantum well and the two intersubband transitions used for probing the populations. Adopted from Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier-carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. *Physical Review B* 70, 155324.

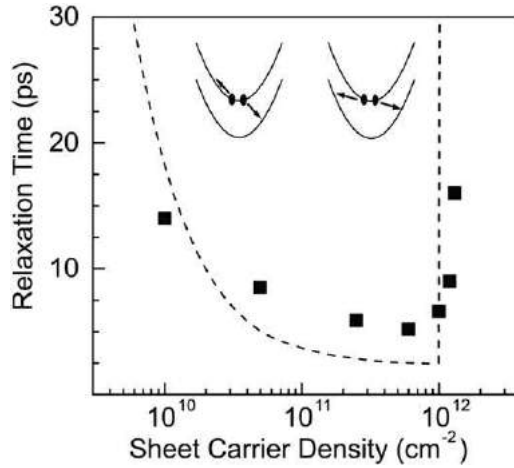


Fig. 4 Experimental intersubband relaxation times (symbols) as function of excitation density. The calculation (dashed line) confirms the experimental results. The inset shows possible two-electron scattering processes. Adopted from Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier-carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. *Physical Review B* 70, 155324.

are injected into the second subband, while the remaining 60% are injected into the first subband at higher k -value. The population in the second subband shows an exponential decay. The electrons which relax down add to the population in the ground level. Subsequently, the population in the ground level drops because of carrier recombination. By fitting the results from a phenomenological rate-equation model to the experimental data, an intersubband relaxation time of 14 ps is obtained for the excitation density of $1 \times 10^{10} \text{ cm}^{-2}$. At a higher excitation density of $3 \times 10^{11} \text{ cm}^{-2}$ the dynamics changes a lot. First, the $|2\rangle \rightarrow |3\rangle$ absorption decays much faster, and a relaxation time of only 5.9 ps is obtained. Second, a blue-shift is observed of the $|1\rangle \rightarrow |3\rangle$ absorption as the time evolves, i.e., as electrons in the second subband relax into the first one. The shift can be understood as follows: The depolarization shift of the $|1\rangle \rightarrow |3\rangle$ transition is proportional to the population in subband $|1\rangle$. Thus, as the population in $|1\rangle$ increases, a shift of the resonance to higher energy occurs.

In Fig. 4 the excitation density dependence of the intersubband relaxation time is shown for varying the excitation sheet carrier density between $1 \times 10^{10} \text{ cm}^{-2}$ and $1.3 \times 10^{12} \text{ cm}^{-2}$. With increasing carrier density a significant shortening of the $|2\rangle \rightarrow |1\rangle$ relaxation time is observed. In the high density regime a sudden increase of the relaxation time occurs. In order to understand the physical mechanism behind this dependence the possible scattering mechanisms have to be considered and numerical estimates of scattering times have to be compared with the experimental results. Since the energy spacing between the two lowest subbands of our quantum well structure is smaller than the LO phonon energy, the electrons in the second subband do not possess sufficient energy to emit directly LO phonons. Relaxation can thus only be due to acoustic phonon emission and/or carrier-carrier scattering. Using the model described in (Ferreira and Bastard, 1989) the acoustic phonon scattering time was estimated to be ~ 300 ps for this structure. This time is too long to explain the experimental findings. Earlier time-resolved photoluminescence experiments have shown that the intersubband carrier-carrier scattering rates can be very high, almost approaching in some circumstances the

intersubband scattering rate due to LO phonon emission. Thus carrier-carrier scattering has to be considered and scattering rates were calculated in the Born-approximation (Smet *et al.*, 1996). The results of this calculation are shown in Fig. 4 as dashed line. At an excitation density of $\sim 1 \times 10^{12} \text{ cm}^{-2}$ the Fermi-level in the ground state approaches the bottom of the second subband and the intersubband relaxation drastically slows down due to Pauli-blocking. This is illustrated by the vertical dashed line. The strong dependence of the intersubband relaxation time on the carrier density and the good agreement with the calculation shows the importance of carrier-carrier scattering for intersubband relaxation. As a result, even for intersubband transitions below the optical phonon energy the relaxation times will be much shorter than the acoustic phonon scattering time which is very important to consider for all device concepts based on intersubband transitions.

Nonlinear Response

Going to the non-linear regime promises additional very interesting insights. So far, the spectroscopy of intersubband excitations in quantum wells subject to intense THz irradiation has been mainly limited to narrowband pulses from free-electron lasers (Craig *et al.*, 1996) or has been based on optical probing of intraband transitions. The direct observation of intersubband dynamics using few-cycle pulses has been either restricted to the linear regime or has been carried out at higher frequencies in the mid-infrared region (Luo *et al.*, 2004; Dynes *et al.*, 2005). The THz spectral range offers unique possibilities to study novel non-linear physics. For instance, the vacuum Rabi splitting can become a significant fraction of the intersubband transition frequency, giving rise to intersubband polaritons and the so-called ultrastrong coupling regime (Günter *et al.*, 2009; Ciuti *et al.*, 2005). In addition, the large bandwidth of single-cycle THz pulses allows the coupling of adjacent intersubband transitions. This opens the road to novel physics beyond the two-level atom model. One very important ingredient for quantum optics and quantum information processing in general is the ability to prepare the system in a certain quantum state. The simplest example would be the controlled transfer of population from the ground state to any of the excited states, for example, to achieve a complete population inversion between two subbands. Apart from the pure scientific interest, such an inverted system may also be useful as a switchable THz gain medium in real world applications, such as THz amplifiers or Q-switches. Dietze *et al.* (2012) reported the first direct observation of intersubband dynamics in a modulation doped multiple quantum well (MQW) sample subject to intense few-cycle THz pulses. Single cycle THz pulses with a bandwidth of 5.5 THz and maximum electric field amplitude of 20 kV/cm are focused on the cleaved facet of the multi quantum well sample (52 nm well width, 10 periods). By applying a bias voltage, the wells can be depleted which allows using an electrical modulation technique to selectively measure the effect of the electronic polarization on the transmitted THz pulses. Fig. 5 shows the level scheme of the QWs. The transition frequencies have been calculated self-consistently including the depolarization shift. At 5 K, all electrons are assumed to be in the ground state. Thus, with the bandwidth of the driving pulse indicated by the blue bar, only the first three states can be accessed. The pump spectrum peaks around 1.4 THz, which is close to the first transition $|1\rangle \rightarrow |2\rangle$. For sufficiently high electric field strengths, Rabi oscillations between the ground and first excited state will lead to a level splitting, which can be probed by the next higher transition $|2\rangle \rightarrow |3\rangle$. Furthermore, the direct transition $|1\rangle \rightarrow |3\rangle$ is dipole forbidden, but state $|3\rangle$ can be populated via two interfering pathways, either through the intermediate state $|2\rangle$ or directly by a 2-photon resonant transition (indicated by red arrows). This quantum interference might lead to an enhanced transmission similar to electromagnetically induced transparency (Luo *et al.*, 2004). In addition, the absorption strength of each transition is dependent on the relative occupation numbers of the involved levels. Thus, for strong driving, we expect to observe saturation of the intersubband transitions.

Thus we expect that the non-linear effects are thereby strongly dependent on the electric field amplitude of the THz driving pulse. Fig. 6 shows the measured modulation signal, ΔE , for different values of the pump power incident on the GaP crystal. For the chosen experimental parameters, the THz peak field scales almost linearly with the optical pulse energy. At the lowest pump energy of 125 μJ , the incident THz peak field is estimated to be 0.8 kV/cm. The modulation signal shown in Fig. 6(a) corresponds to the expected free-induction decay of a weakly driven two-level system (black curve). The oscillations are monochromatic with a center frequency of 1.5 THz and a dephasing time of 2.3 ps. When the THz amplitude is increased, a beating pattern gradually emerges, indicated by the black arrows in Fig. 6(b). This beating is a clear indication for a second frequency component in the spectrum. We attribute this to increased occupation of the first excited level, which activates the next transition $|2\rangle \rightarrow |3\rangle$.

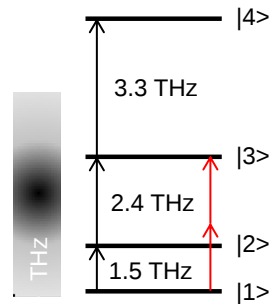


Fig. 5 Level scheme of the quantum well and possible transitions.

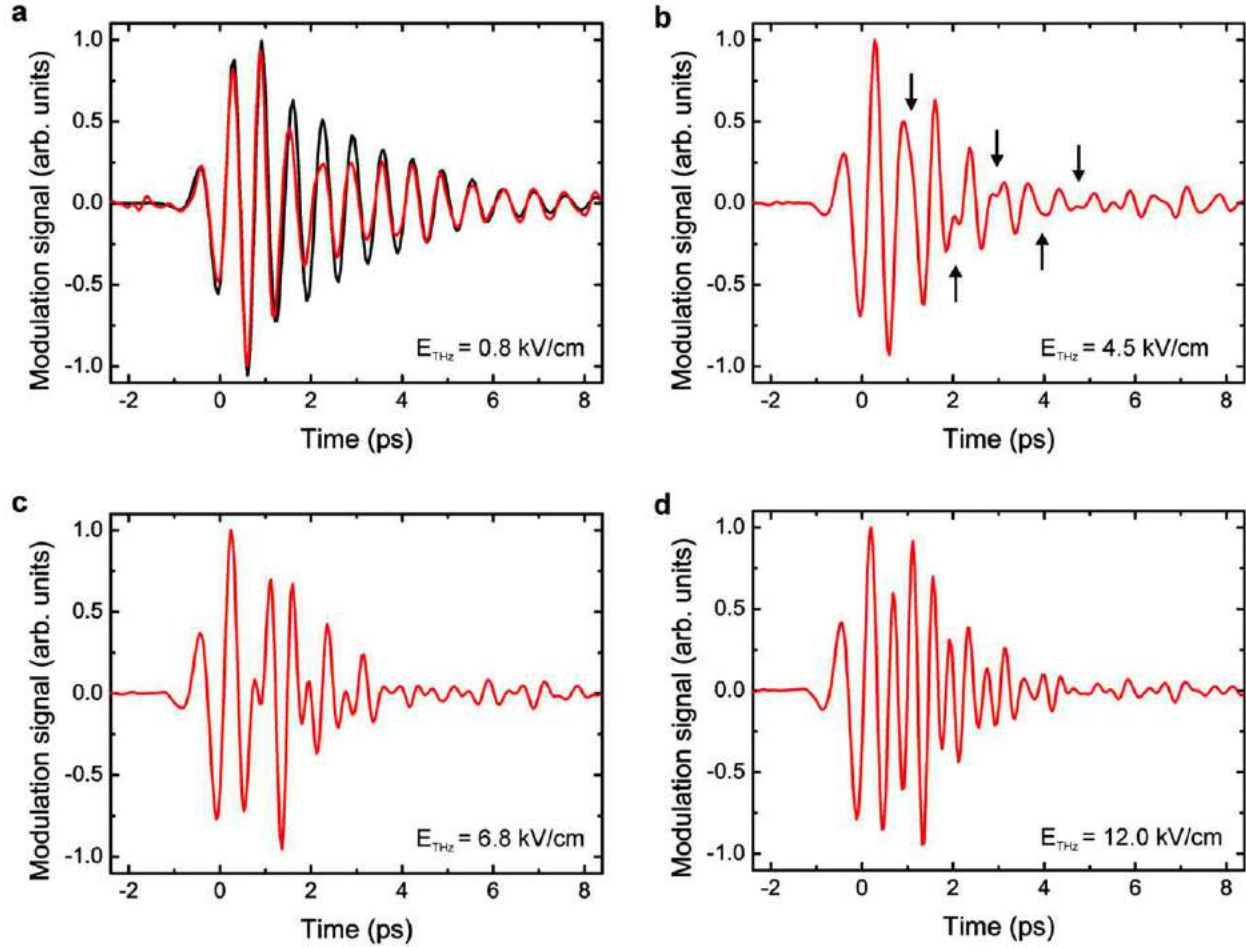


Fig. 6 Modulation signals for different pump powers. (a) For the lowest THz field amplitude, the free induction decay is almost monochromatic (red). The black curve shows the results of FDTD simulations for a two-level system. (b) For 4.5 kV/cm, a beating becomes obvious. The temporal positions of the beat nodes are indicated by the arrows. (c) For higher peak fields, the total pulse becomes shorter and shows echo-like features. (d) Above 12 kV/cm, the free induction decay contains several higher frequencies. Adopted from Dietze, D., Darmon, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. *Optics Express* 20, 23053.

In addition, the modulation signal decays faster indicating that additional dephasing mechanisms occurring compared to the simple two-level model.

Around 6.8 kV/cm (**Fig. 6(c)**), the beating pattern gets more pronounced and resembles the photon echoes produced by the free-induction decay. This pulse-like structure is an indication for a broadened modulation spectrum which contains many frequency components. The character of the modulation signal changes again for very high pump powers (**Fig. 6(d)**). Both the oscillation period and the decay time are much faster than in **Fig. 6(a)**. **Fig. 7** shows the modulation spectra obtained by Fourier transforming the time domain signals. The signal originating from the free induction decay of the three allowed intersubband transitions are marked by 1, 2, and 3. For the lowest driving field, there is only a single feature present around 1.5 THz. For increasing pump powers, electrons are efficiently transferred to higher lying states and the higher transitions start to appear in the spectra: the $|2\rangle \rightarrow |3\rangle$ transition around 2.4 THz and the $|3\rangle \rightarrow |4\rangle$ transition at 3.3 THz. The 2.4 THz peak shows a doublet structure, which vanishes for higher pump powers. The observed behavior is in principle consistent with the scenario of sequential pumping of electrons from the ground state to higher excited states by the broadband THz pulse. For intermediate peak electric fields around 6 kV/cm, an additional feature with a broad frequency content appears centered in between the 1.5 and 2.4 THz transition (marked by asterisk). To determine whether this effect is related to a real electron current (in-phase), we have calculated the (single-pass) absorption according to:

$$\alpha(\omega) \sim \text{Im}(-i\Delta E(\omega)/E_{\text{ref}}(\omega))$$

which is valid for the assumption, that the depleted sample is non-absorbing and $\Delta E \ll E_{\text{ref}}$. **Fig. 7(b)** shows the absorption spectra associated to **Fig. 7(a)**. The shaded area is excluded from the discussion due to the limited signal-to-noise ratio. The broadband spectral feature has disappeared, which indicates that it is indeed related to an in-phase polarization. It resembles the ponderomotive contribution from free carriers in the quantum well plane (Golde *et al.*, 2009). However, the spectrum of the broad

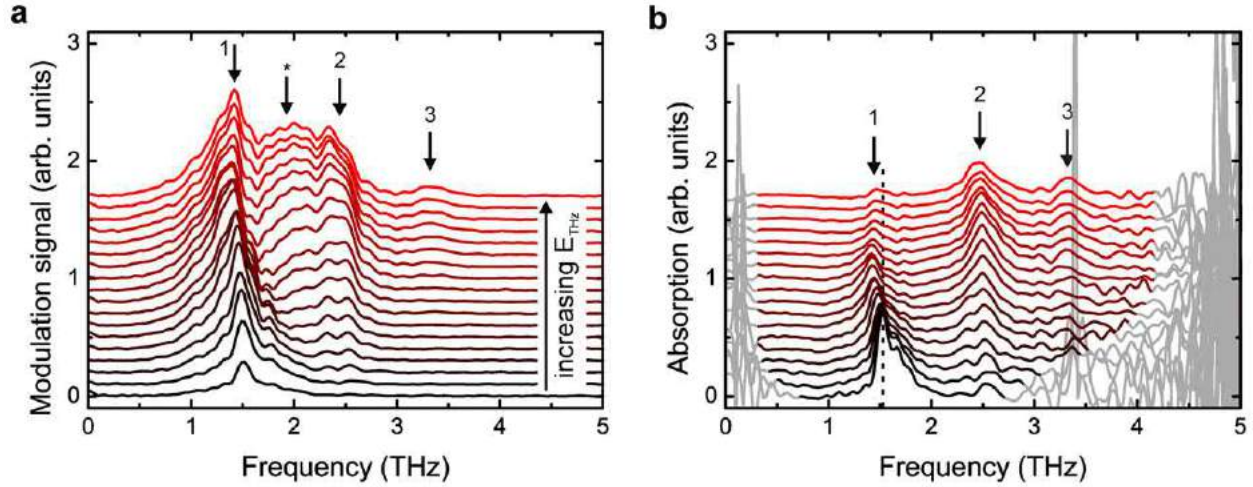


Fig. 7 Modulation and absorption spectra for different pump powers. (a) Modulation spectrum. The three allowed transitions are marked by (1, 2, 3). There appears also a broadband feature for higher pump powers marked by the asterisk. (b) Absorption spectrum. The broadband feature has disappeared. The fundamental transition exhibits a clear red shift for higher pump powers. The curves are vertically offset for clarity. Adopted from Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. *Optics Express* 20, 23053.

feature does not coincide with the spectrum of the driving pulse, and, in addition, the chosen experimental geometry excludes the in-plane ponderomotive current because the THz electric field is oriented perpendicular to the quantum well plane. As an alternative explanation it is suggested that this coherent signal is a consequence of the non-linear mixing of the THz pulse with the induced in-phase polarization of the quantum well (Dietze *et al.*, 2013). The intersubband absorption shows a clear dependence on the pump power. The decrease of the absorption peak associated to the $|1\rangle \rightarrow |2\rangle$ transition is a sign for THz induced saturation of the intersubband transition. In addition, the transition frequency is shifted to lower values for increased amplitudes of the driving field. This effect has been predicted theoretically to be related to an undressing of the collective intersubband excitation which is causing the depolarization shift of the resonance frequency (Zaluzny, 1993). Previously it has only been observed experimentally using intense narrowband THz radiation from a free electron laser with peak powers on the order of 1 kW. The absorption line of the $|2\rangle \rightarrow |3\rangle$ transition shows a doublet-like structure for optical 6 pump energies below 1 mJ, which is reminiscent of the Autler-Townes effect in a three-level system (Zaks *et al.*, 2011). The disappearance of the splitting with increasing field strengths, however, is counter intuitive. A possible explanation could be based on the scaling properties of the vacuum Rabi splitting in an ensemble of two-level atoms: $\Omega_R \sim (n_{2d})^{-1}$, where n_{2d} is the areal density of electrons (Ciuti *et al.*, 2005). In the present case, electrons are shared among all levels. Thus, the level splitting is proportional only to the number of electrons participating in the coherent population transfer between the ground and first excited state. For higher pump field strengths, electrons are efficiently transferred to the higher lying states and are not accessible for the Rabi oscillations. As a consequence, the level splitting disappears.

Few Cycle THz Spectroscopy of Quantum Cascade Heterostructures

Quantum Cascade semiconductor (QC) heterostructures are very powerful structures providing optical gain in the frequency range spanning from near-infrared to far-infrared (Faist *et al.*, 1994; Köhler *et al.*, 2002). Thereby, the optical gain is achieved by the optical transition between appropriately spaced quantized levels of QC heterostructures. Due to the vast space of design and technology parameters, the detailed information on the optical gain value, its bandwidth and profile, the gain recovery time and their dependence on bias and operating temperature, is of great importance. Few-cycle THz spectroscopy is the tool of choice to access this information.

Usually, studying the optical gain using standard detectors recording the power emitted by the laser is limited to operation conditions below the lasing threshold. For the first time, Kröll *et al.* (Kröll *et al.*, 2007) have accessed the internal laser processes in the whole range of operating conditions. For detection of the THz signal electro-optic detection (EOD) (Wu *et al.*, 1996) featuring a sub 100 fs resolution is used. Thereby, the necessary synchronization between the THz signal and the probe is guaranteed by THz photon mutually phase-locked to the probe and injected into the investigated laser structure. Therefore, only the electric field of the original photons and of those generated by a stimulated emission (hence with a timing and phase equal to the injected photons) is detected, while photons generated by amplified spontaneous emission feature a random phase and timing produce a zero averaged EOD signal. This technique originally demonstrated for THz quantum cascade lasers (Kröll *et al.*, 2007), has been later extended for mid-infrared lasers (Parz *et al.*, 2008) and near-infrared coherent sources (Keiber *et al.*, 2016).

Fig. 8 schematically shows the layout of a typical THz time-domain system (TDS) used to investigate THz QC heterostructures. It does not differ from the standard THz-TDS (Smith and Arnold, 2011), except of the signal modulation. In THz-TDS, the lock-in

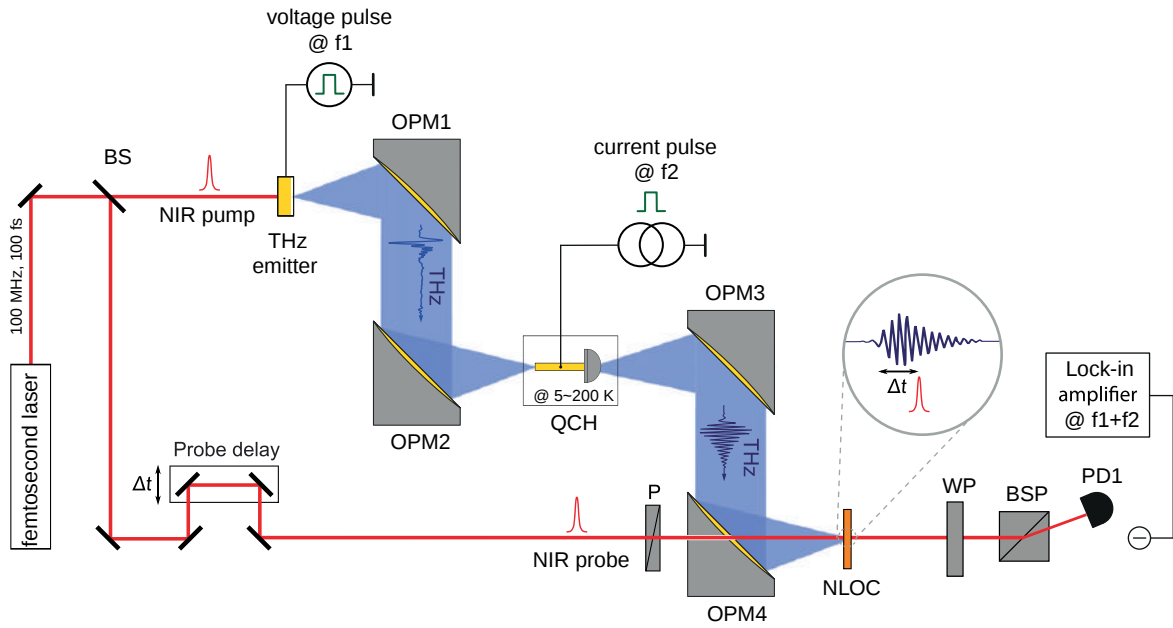


Fig. 8 Time-domain spectroscopy setup for investigation of THz Quantum Cascade heterostructures. BS, beam splitter; NLOC, non-linear optical crystal; OPM, off-axis parabolic mirror; P, NIR polarizer; PBS, polarizing beam splitter; PD, near-infrared photodetector; WP, $\lambda/4$ waveplate.

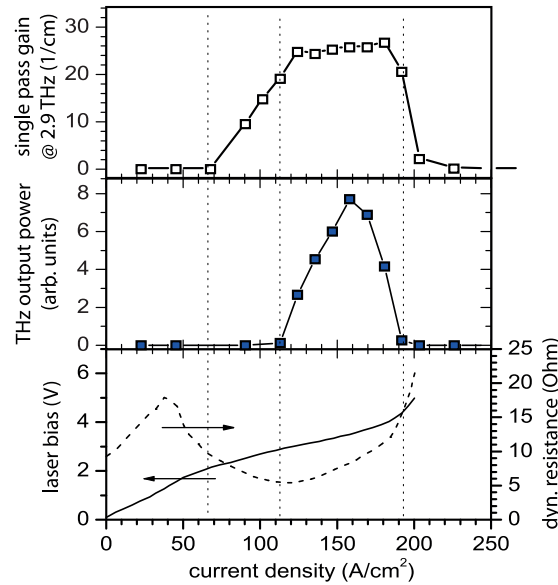


Fig. 9 Basic electrical and optical characteristics of a THz QCL: (bottom panel) the current-voltage dependence of the device (solid line) and corresponding dynamic resistance (dashed line); (middle panel) the current – THz light intensity characteristics; (top panel) the gain observed from the THz injection measurement. Details on THz QCL and method is adopted from Kröll, J., Darmo, J., Dhillon, S.S., *et al.*, 2007. Phase resolved measurements of stimulated emission in a laser. *Nature* 449, 698–702.

technique is typically used to analyze the signal from the THz detector (photocurrent of a THz antenna or of a balanced photodetector) that inherently requires a modulation of the THz signal. When an active device is tested, an additional modulation of the device is required, so a double-modulation technique is employed for the investigation of THz QC heterostructures. Thereby, the THz emitter is chopped at frequency f_1 , the THz QC heterostructure is electrically modulated at frequency f_2 and the signal at the THz electric field detector is demodulated at $f_1 \pm f_2$. Due to the high nonlinearity of both modulations of the detector and the QC heterostructure, the modulation frequencies should preferably be chosen not to be harmonic pairs (Kröll *et al.*, 2007).

Using THz-TDS Kröll *et al.* (Kröll *et al.*, 2007) have investigated a THz QC heterostructure providing a gain at 2.7 THz at different operation conditions (Fig. 9). As expected, the THz gain is a function of the electric current flowing through the device and is correlated with the THz power emitted from the laser and with the current-voltage characteristics. The injected THz photons

get amplified by stimulated emission well below the laser threshold (at a current density of $\sim 110 \text{ A/cm}^2$) and the gain steadily increases as the electron injection into upper lasing level is rising with increasing current through the device. Thereby, the onset of the gain (at a current density of $\sim 110 \text{ A/cm}^2$) appears at the applied voltage necessary to reach the correct QC heterostructure alignment. The drop of the dynamic resistance indicates that this condition is reached. When the THz gain is high enough to compensate waveguide and mirror losses, the lasing threshold is reached. Coherent emission starts and the THz gain gets clamped to the value of the total laser loss. This regime is observed for a device current density between 110 and 180 A/cm^2 , for which the measured THz gain remains almost constant. The residual weak increase of the gain with the device current is attributed to the gradual increasing total loss due to, for example, heating. Lasing stops when the bias applied to the QC heterostructure is too high and the heterostructure alignment is broken. Accordingly, the observed THz gain drops very quickly (for a current density $> 180 \text{ A/cm}^2$).

The issue of the clamped gain has been addressed in a follow-up work by Jukam *et al.* (Jukam *et al.*, 2009). They used ultrafast switching of the QC heterostructure device current in order to achieve effective THz gain switching. When the switching was synchronized to the optical injection, a temporal increase of the gain has been observed (Oustinov *et al.*, 2010). It is well demonstrated for the amplification of the THz probe pulse injected into the laser device (see Fig. 10). If the probe pulse is entering the laser before the gain switching occurs, the pulse experiences a steady-state gain. However, when the gain is switched before the probe pulse is injected, the amplification is significantly increased. The details of the dynamics of the gain rise, persistence and decay depend on the THz pulse length ($\sim 1.5 \text{ ps}$), the transit time ($\sim 18 \text{ ps}$), and the temporal characteristics of the switching pulse itself ($\sim 100 \text{ ps}$). It is worth noting, that after the gain switching event the gain eventually drops to values below the steady-state values, indicating a significant gain depletion due to a temporal burst of the THz emitted power.

When the length of the gain switching pulse is much longer than the QCL cavity round-trip time, a very complex THz pulse dynamics is observed (Bachmann *et al.*, 2015). In Fig. 11 three different operation modes of a THz QCL are compared. First, the THz pulse train formed in the device driven close to threshold is shown. The injected few-cycle THz pulse gets gradually attenuated as at this operation point the total loss still exceeds the gain in the laser device. For this condition, the cavity mirrors loss dominates. When the laser is cw operated at the point far above threshold, a stable train of THz pulses is generated, although the pulse amplitude is small in comparison to the THz seed amplitude due to the small spectral overlap of the THz seed and THz gain, and the small clamped THz gain. The latter limit is readily overcome by employing gain switching. Voltage pulses as short as 600 fs bring the laser from the sub-threshold operation point to the operation for maximum output power (in the CW regime). This switching has also a dramatic impact on the THz pulse train formation. The pulse amplitude exceeds the amplitude of the THz seed in spite of the very different spectral content which indicates significant amplification of the frequency components within the THz gain spectrum. The pulse amplitude closely follows the temporal shape of the voltage pulse with small lag at the rising and falling edges. At the beginning, the THz pulse needs about two cavity round-trips to reach saturation of amplification, while the slow decay of the pulse amplitude is a typical cavity ring-down behavior. Finally, the time domain output of the gain switched QCL gets structured due to the modal dispersion which is discussed later in this section.

The experimental observations mentioned in the previous paragraphs have been compared with a general model of QC heterostructure based lasers. A single Quantum Cascade (Fig. 12(a)) can be considered as a two-level quantum system, where the upper states can be populated by electron injection or by absorption from the occupied ground state. Such a system can be described by a density matrix and its time development is given by the optical Bloch equation (OBE). OBE together with the Maxwell equations form the theoretical base for the complete description of the interaction of the electromagnetic wave with the THz QC heterostructure gain medium. Since the THz time domain data have to be modeled, the features of the THz TDS need to be also considered. In standard THz TDS the changes of the few-cycle THz pulse after interaction with an object are evaluated in terms of the complex permittivity (or permeability) of the object. The injected THz photons, however, can be scattered, absorbed, multiplied or temporarily trapped in the device cavity. Hence, the time domain data have to be evaluated with respect to these

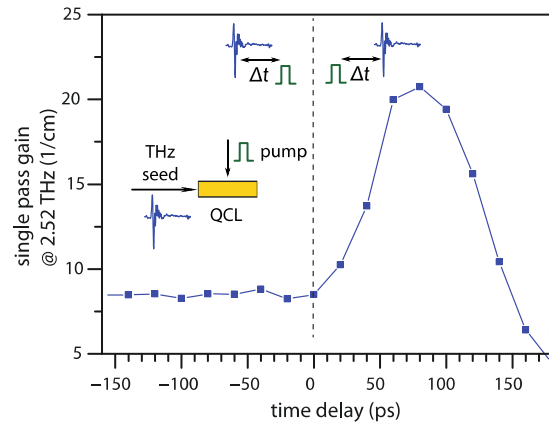


Fig. 10 Gain switching of a THz QC heterostructure with an ultra-fast photoconductive switch. Adopted from Oustinov, D., Jukam, N., Rungsawang, R., *et al.*, 2010. Phase seeding of a terahertz quantum cascade laser. *Nature Communications* 1, 69–75.

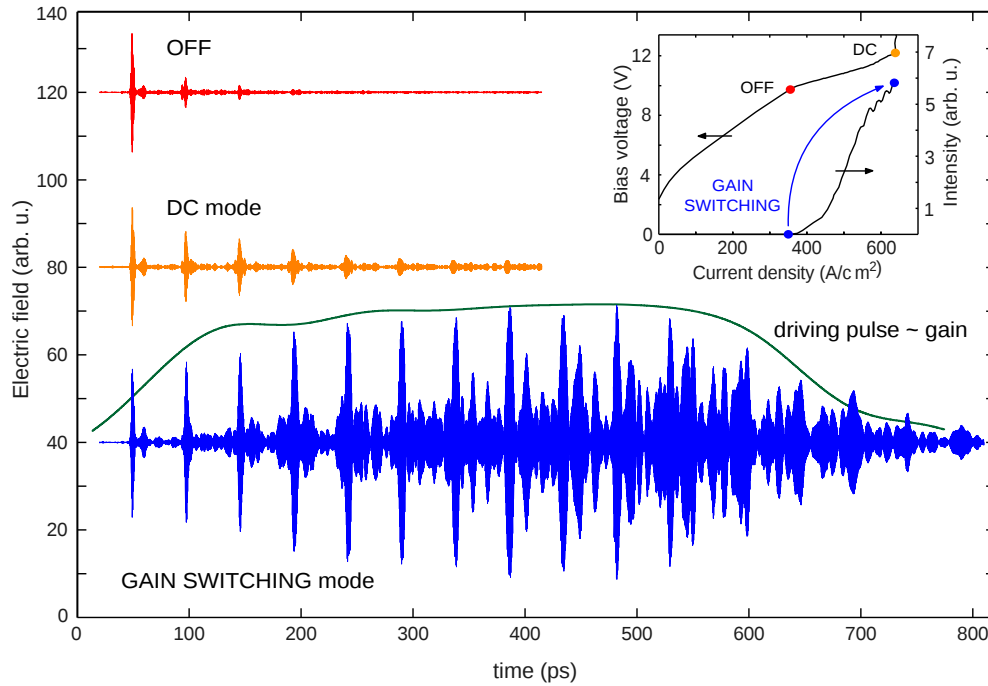


Fig. 11 Dynamics of the gain and THz pulse train formation in a broadband THz QC heterostructure driven with a sub-nanosecond voltage pulse (its envelope indicated with green solid line). Adopted from Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. *Optics Express* 23, 3117–3125.

processes and a detailed model of the device has to be considered. A complete QC heterostructure based laser model has been applied by Kröll *et al.* (Kröll *et al.*, 2007), Darmo *et al.* (Darmo *et al.*, 2006) and Freeman *et al.* (Freeman *et al.*, 2013) to study the THz emission in time-domain with respect to different aspects of gain medium and waveguide. The emission of a driven two-level system is demonstrated in Fig. 12(b), which shows the internal electric field in a QCL waveguide of the length of 2 mm when a THz seed with a bandwidth of 300 GHz is injected. The amplitude of the electric field is not constant and indicates the propagation of a dispersed pulse. This internal amplitude profile corresponds to the externally observed laser emission in the time-domain, hence the internal dynamics of the laser can be monitored. The effect of the laser cavity is demonstrated in Fig. 12(c). Therein the impact of the material and modal dispersion and the modal phase diffusion is shown.

For the purpose of frequency comb generation and the subsequent achievement of pulsed output of the QCL, the knowledge of the dispersion of the QC heterostructure based emitter is crucial. For the THz QCL this issue has been recently addressed by Burghoff *et al.* (Burghoff *et al.*, 2014), Rösch *et al.* (Rösch *et al.*, 2015) and Bachmann *et al.* (Bachmann *et al.*, 2016a). The system dispersion can be assessed from the few cycle THz pulses circulating within the QCL cavity due to partial reflections on the cavity mirrors (Burghoff *et al.*, 2014; Bachmann *et al.*, 2016a). During the cavity round-trip the frequency components of the THz pulse accumulate different phase shifts according to the frequency dependent effective refractive index of the QCL cavity. Bachmann *et al.* (Bachmann *et al.*, 2016a) performed a detailed analysis of this phase shift for a THz QCL with a heterogeneous QC heterostructure (Turčínková *et al.*, 2011). Fig. 13(a) shows the observed spectral dependence of the group velocity dispersion (GVD) that exhibits strong oscillations in the whole gain bandwidth of the QC heterostructure. Bachmann *et al.* (2016a) have looked into the origin of these oscillations considering a model in which the dispersion in the THz QCL consists of contributions from the dispersion of the heterostructure material; the waveguide and from the optical gain of the intersubband transitions. The measured GVD can only be fitted when the optical gain induced dispersion is considered (see Fig. 13(b)). Thereby a real design of the QC heterostructure has been considered (Turčínková *et al.*, 2011; Bachmann *et al.*, 2014) with three different sections. The result demonstrates unambiguously that the QCL cavity dispersion (in terms of the group velocity dispersion GVD) is dominated by the dispersion caused by the optical gain medium (Bachmann *et al.*, 2016a). Therefore, the dispersion is dependent on the operation temperature and on the bias of the QCL (Fig. 13(b)), the fact has to be considered for the correct design of a dispersion compensating structure (Burghoff *et al.*, 2014; Villares *et al.*, 2016).

Injection seeding of a QCL with a few-cycle THz pulse (Kröll *et al.*, 2007) enables also to map the laser cavity modes. The used seeding mechanism (Kröll *et al.*, 2007; Dhillon *et al.*, 2010; Burghoff *et al.*, 2011) does not prefer certain modes, but feeds all possible cavity modes, including higher order lateral modes. Different lateral modes travel at different group velocities leading to an additional dispersion of the injected THz pulse and instead of clean separated pulses a rich dynamic inter-pulse structure is formed (Fig. 14(a)). Bachmann *et al.* (Bachmann *et al.*, 2016b) have demonstrated the impact of lateral mode control on THz pulse formation in a THz QC heterostructure operating in the amplifier regime. When the propagation loss of higher-order lateral modes is compromised (e.g., by an additional loss), almost bandwidth limited THz pulses are generated from the few-cycle THz seed

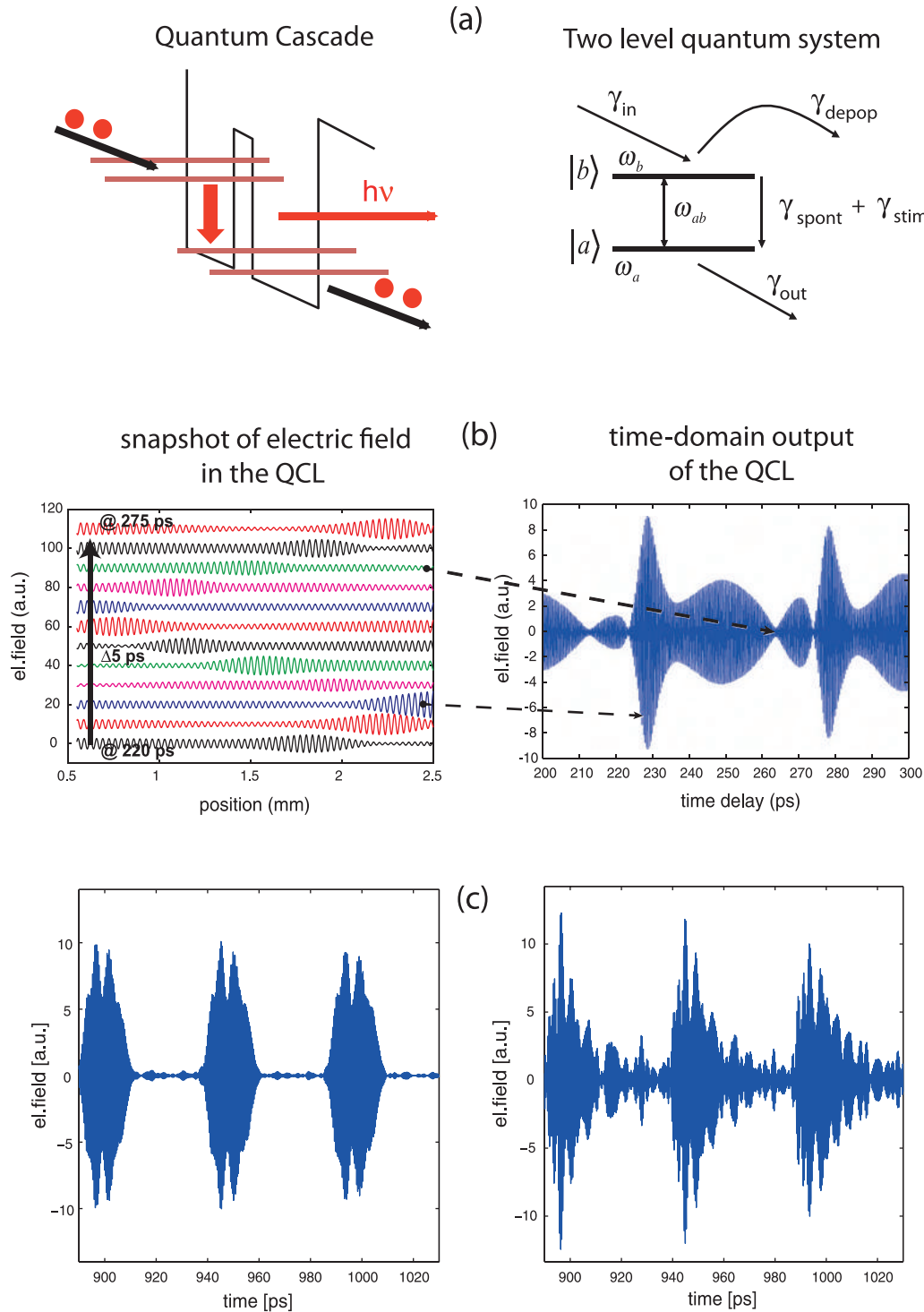


Fig. 12 Modeling the time response of a THz QC heterostructure: (a) schematic drawing of a Quantum Cascade and its two-level quantum-mechanical model; (b) electric field time snapshots in the THz QCL waveguide and their correlation with the laser output waveform; (c) impact of different parameters on the laser output.

(Fig. 14(b)). Thereby, the seeded THz gain bandwidth of ~ 400 GHz guarantees the THz pulse length of ~ 2.5 ps. Such a significant control of the THz pulse form was achieved by using lateral mode control in the form of absorbing boundary conditions on the edges of the QCL cavity ridge (Bachmann *et al.*, 2016b).

In addition to the very successful application as laser device, THz QC heterostructure based systems can also be used as amplifier of the few-cycle THz pulses. This approach is especially appealing for the THz spectroscopy community due to the THz

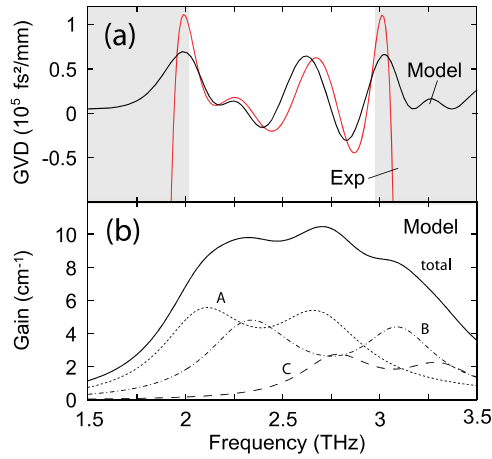


Fig. 13 Dispersion in a THz broadband QCL: (top panel) measured group velocity dispersion GVD (red line) and model GVD (black line); gray shaded areas indicate the spectral range with too low signal-to-noise ratio to extract correct GVD values; (bottom panel) model gain curve assembled from three different QC sections fitting the experimental GVD. Adopted from Bachmann, D., Rösch, M., Scali, G., *et al.*, 2016a. Dispersion in a broadband terahertz quantum cascade laser. *Applied Physics Letters* 109, 221107. doi:10.1063/1.4969065.

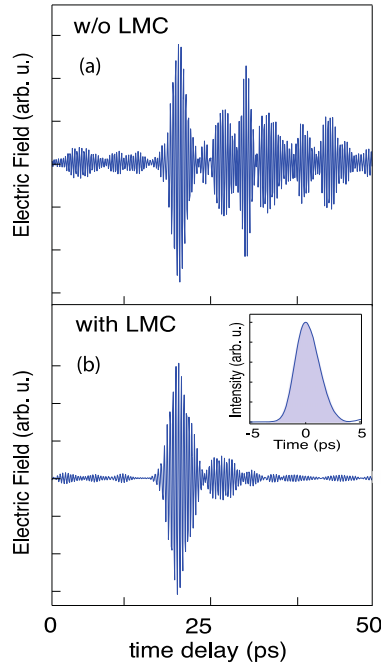


Fig. 14 The electric field of a THz pulse formed in the seeded THz-QCL: (a) device without lateral mode control (LMC); (b) device with lateral mode control. Inset: pulse intensity profile indicating ~ 2.5 ps pulse length. Adopted from Bachmann, D., Rösch, M., Süess, M.J., *et al.*, 2016b. Short pulse generation and mode control of broadband terahertz quantum cascade lasers. *Optica* 3, 1087–1094. doi:10.1364/OPTICA.3.001087.

QCL operating frequency window of 2.0–5.0 THz, i.e., the frequency window in which typical THz sources used for the THz time-domain spectroscopy have a severely compromised spectral brightness. First THz amplification demonstrations have been done with a THz QC heterostructure providing a limited gain bandwidth of < 200 GHz (Jukam *et al.*, 2009; Dhillon *et al.*, 2010), hence with a low potential for applications. The successful development of broadband THz QC heterostructures (Turčinková *et al.*, 2011; Freeman *et al.*, 2011) enabled a THz seed amplification with a bandwidth as large as 600 GHz (Bachmann *et al.*, 2015). Recently, Bachmann *et al.* (Bachmann *et al.*, 2016b) demonstrated the amplification of the THz electric field strength by > 20 dB in the spectral window between 2.1–3.0 THz (Fig. 15) in a mode controlled broadband THz QC heterostructure based Fabry-Perot amplifier. Thereby, a train of ~ 2.5 ps long THz pulses with a total length of ~ 650 ps is generated per single THz seed pulse. The pulse train length is defined by the gain switching time window (Bachmann *et al.*, 2015).

Finally, THz injection locking (Oustinov *et al.*, 2010) or RF injection locking (Barbieri, 2011) synchronized with a coherent sampling femtosecond laser are key approaches to build a THz system for the remote sensing/imaging applications, that is

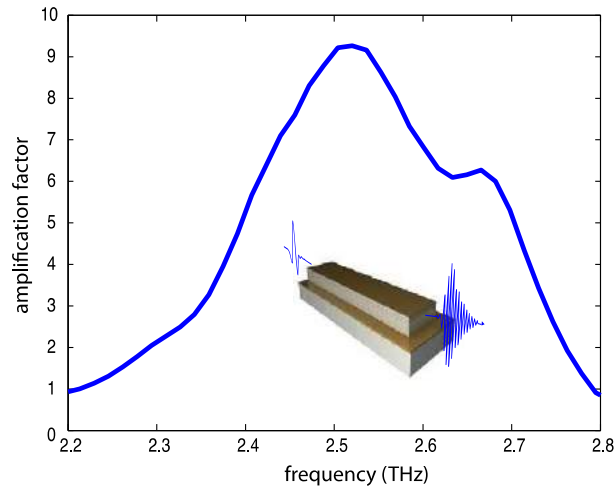


Fig. 15 Amplification factor achieved for the broadband THz QC heterostructure based Fabry-Perot amplifier. The amplification profile follows the gain profile. Adopted from Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. *Optics Express* 23, 3117–3125.

harvesting of the strengths of both THz QC heterostructure technology and THz time-domain spectroscopic technique (Darmo *et al.*, 2010a,b).

Conclusions

Semiconductor nanostructures are very exciting object to study light matter interaction and as a building block for novel optoelectronics devices. The ability to design the energies of the quantized transition, their optical matrix element and the tunnel coupling is unprecedented in solid state physics. Few-cycle THz spectroscopy is a perfectly suited method to study these structures as it provides time-resolution and phase resolution. This allows accessing the full response of the nanostructures to THz radiation. Linear experiments reveal the decay of the intersubband polarization with a decay time of several ps. The carrier lifetime is obtained by optical injection of carriers and monitoring their population dynamics with THz probe pulses. The lifetimes vary as function of their density and range from a few ten ps to a few ps due to carrier-carrier scattering. High intensity THz pulses allow studying the whole field of non-linear optics in semiconductor nanostructures. Higher order transitions, field induced level splitting and complex many body effects are demonstrated. Semiconductor nanostructures enable the realization of Quantum Cascade lasers which are becoming the most powerful compact THz radiation sources. The study of the gain bandwidth, saturation and bias dependence is pursued by few-cycle THz pulses and by coherent electro-optic detection. This unique combination of a designable gain medium and THz time-domain spectroscopy provided for the first time direct insight to the process of stimulated emission and lasing. The formation of spatial hole burning, the influence of the intersubband gain dispersion on the THz propagation, and the saturation of the gain are directly observed. The use of an ultrabroad band QC heterostructure gain medium enables the realization of octave spanning lasers and frequency combs. THz time domain spectroscopy provides a direct measurement of the spectral gain and of the dispersion. The knowledge and control of the dispersion is crucial for frequency comb operation and also for short pulse generation. Injection seeding of ultrabroad band QC heterostructures resulted in the observation of pulses as short as 2.5 ps. This successful development will continue by studying, for example, strong coupling of intersubband transition to metamaterials, modelocking and intersubband transition in novel 2D materials.

See also: Foundations of Coherent Transients in Semiconductors. Semiconductor Lasers. Terahertz Lasers

References

- Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. *Optics Express* 23, 3117–3125.
- Bachmann, D., Rösch, M., Deutsch, C., *et al.*, 2014. Spectral gain profile of a multi-stack terahertz quantum cascade laser. *Applied Physics Letters* 105, 181118.
- Bachmann, D., Rösch, M., Scaleri, G., *et al.*, 2016a. Dispersion in a broadband terahertz quantum cascade laser. *Applied Physics Letters* 109, 221107.
- Bachmann, D., Rösch, M., Süess, M.J., *et al.*, 2016b. Short pulse generation and mode control of broadband terahertz quantum cascade lasers. *Optica* 3, 1087–1094.
- Barbieri, S., 2011. Coherent sampling of active mode-locked terahertz quantum cascade lasers and frequency synthesis. *Nature Photonics* 5, 306–313.
- Burghoff, D., Kao, T.-Y., Ban, D., *et al.*, 2011. A terahertz pulse emitter monolithically integrated with a quantum cascade laser. *Applied Physics Letters* 98, 061112.
- Burghoff, D., Kao, T.-Y., Han, N., *et al.*, 2014. Terahertz laser frequency combs. *Nature Photonics* 8, 462–467.

- Ciuti, C., Bastard, G., Carusotto, I., 2005. Quantum vacuum properties of the intersubband cavity polariton field. *Physical Review B* 72, 115303.
- Craig, K., Galdrikian, B., Heyman, J.N., *et al.*, 1996. Undressing a collective intersubband excitation in a quantum well. *Physical Review Letters* 76, 2382–2385.
- Darmo, J., Bachmann, D., Unterrainer, K., 2010a. Temporal and spectral aspects of quantum cascade heterostructure with a broadband gain. In: ITQW 2017 Conference Proceedings No. 190, TU Wien, Vienna, 2015.
- Darmo, J., Kröll, J., Unterrainer, K., 2006. Theoretical aspects of time-domain spectroscopy applied to semiconductor terahertz gain medium. In: AIP Conference Proceeding No. 893AIP, New York, 2007, pp. 515–518.
- Darmo, J., Maril, M., Dietze, D., Strasser, G., Unterrainer, K., 2010b. Key components of terahertz remote sensing system. IEICE Technical Report ED2010-172, pp. 77 – 81.
- Dhillon, S.S., Sawallich, S., Jukam, N., *et al.*, 2010. Integrated terahertz pulse generation and amplification in quantum cascade lasers. *Applied Physics Letters* 96, 061107.
- Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. *Optics Express* 20, 23053.
- Dietze, D., Darmo, J., Unterrainer, K., 2013. Efficient population transfer in modulation doped single quantum wells by intense few-cycle terahertz pulses. *New Journal of Physics* 15, 065014.
- Dynes, J.F., Frogley, M.D., Beck, M., Faist, J., Phillips, C.C., 2005. A stark splitting and quantum interference with intersubband transitions in quantum wells. *Physical Review Letters* 94, 157403.
- Faist, J., Capasso, F., Sivco, D.J., *et al.*, 1994. Quantum cascade laser. *Science* 264, 553–556.
- Ferreira, R., Bastard, G., 1989. Evaluation of some scattering times for electrons in unbiased and biased single- and multiple-quantum-well structures. *Physical Review B* 40, 1074.
- Freeman, J.R., Brewer, A., Madeo, J., *et al.*, 2011. Broad gain in a bound-to-continuum quantum cascade laser with heterogeneous active region. *Applied Physics Letters* 99, 241108.
- Freeman, J.R., Maysonnave, J., Khanna, S., *et al.*, 2013. Laser seeding dynamics with few-cycle pulses. *Physical Review A* 87, 063817.
- Golde, D., Wagner, M., Stehr, D., *et al.*, 2009. Fano signatures in the intersubband terahertz response of optically excited semiconductor quantum wells. *Physical Review Letters* 102, 127403.
- Günter, G., Anappara, A.A., Hees, J., *et al.*, 2009. Sub-cycle switch-on of ultrastrong light-matter interaction. *Nature* 458, 178.
- Heyman, J., Kersting, R., Unterrainer, K., 1998. Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72, 644.
- Huber, R., Brodschelm, A., Tauser, F., Leitenstorfer, A., 2000. Generation and field-resolved detection of femtosecond electromagnetic pulses tunable up to 41 THz. *Applied Physics Letters* 76, 3191.
- Jukam, N., Dhillon, S.S., Osutinov, D., *et al.*, 2009. Terahertz amplifier based on gain switching in a quantum cascade laser. *Nature Photonics* 3, 715–719.
- Keiber, S., Sederberg, S., Schwarz, A., *et al.*, 2016. Electro-optic sampling of near-infrared waveforms. *Nature Photonics* 10, 159–162.
- Kersting, R., Bratschitsch, R., Strasser, G., Unterrainer, K., Heyman, J., 2000. Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25, 272.
- Köhler, R., Tredicucci, A., Belham, F., *et al.*, 2002. Terahertz semiconductor-heterostructure laser. *Nature* 417, 156–159.
- Kröll, J., Darmo, J., Dhillon, S.S., *et al.*, 2007. Phase resolved measurements of stimulated emission in a laser. *Nature* 449, 698–702.
- Luo, C.W., Reimann, K., Woerner, M., *et al.*, 2004. Phase-resolved nonlinear response of a two-dimensional electron gas under femtosecond intersubband excitation. *Physical Review Letters* 92, 047402.
- Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier–carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. *Phys. Rev. B* 70, 155324.
- Oustinov, D., Jukam, N., Rungsawang, R., *et al.*, 2010. Phase seeding of a terahertz quantum cascade laser. *Nature Communications* 1, 69–75.
- Parz, W., Müller, T., Darmo, J., *et al.*, 2008. Ultrafast probing of light-matter interaction in a mid-infrared quantum cascade laser. *Applied Physics Letters* 93, 091105.
- Rösch, M., Scalari, G., Beck, M., Faist, J., 2015. Octave-spanning semiconductor laser. *Nature Photonics* 9, 42–47.
- Smet, J.H., Fonstad, C.G., Hu, Q., 1996. Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources. *Journal of Applied Physics* 79, 9305.
- Smith, R.M., Arnold, M.A., 2011. Terahertz time-domain spectroscopy of solid samples: principles, applications, and Challenges. *Applied Spectroscopy Reviews* 46, 636–679.
- Turčinková, D., Scalati, G., Castellano, F., *et al.*, 2011. Ultra-broadband heterogeneous terahertz quantum cascade laser emitting from 2.2 to 3.2 THz. *Applied Physics Letters* 99, 191104.
- Villares, G., Riedi, S., Wolf, J., *et al.*, 2016. Dispersion engineering of quantum cascade laser frequency comb. *Optica* 3, 252–258.
- Wu, Q., Litz, M., Zhang, X.-C., 1996. Broadband detection capability of ZnTe electro-optic field detectors. *Applied Physics Letters* 68, 2924–2926.
- Zaks, B., Stehr, D., Truong, T.-A., *et al.*, 2011. THz-driven quantum wells: coulomb interactions and Stark shifts in the ultrastrong coupling regime. *New Journal of Physics* 13, 083009.
- Zaluzny, M., 1993. Influence of the depolarization effect on the nonlinear intersubband absorption spectra of quantum wells. *Physical Review B* 47, 3995.

Further Reading

- Shtrichman, I., Metzner, C., Ehrenfreund, E., *et al.*, 2001. Depolarization shift of the intersubband resonance in a quantum well with an electron-hole plasma. *Physical Review B* 65, 035310.
- Wagner, M., Schneider, H., Stehr, D., *et al.*, 2010. Observation of the intraexcitonic Autler-Townes effect in GaAs/AlGaAs semiconductor quantum wells. *Physical Review Letters* 105, 167401.

Strong-Field Terahertz Excitations in Semiconductors

Ulrich Huttner, Philipps University of Marburg, Marburg, Germany

Rupert Huber, University of Regensburg, Regensburg, Germany

Mackillo Kira, University of Michigan, Ann Arbor, MI, United States

Stephan W Koch, Philipps University of Marburg, Marburg, Germany

© 2018 Elsevier Inc. All rights reserved.

Introduction

Under resonant excitation conditions, already weak electromagnetic fields can induce optical transitions in matter systems. Here, the exciting photon energy matches the energetic separation of two electronic states that have a finite dipole transition matrix element. An example is the hydrogen atom, described by the atomic Coulomb potential schematically shown in Fig. 1(a). Its electronic eigenstates are defined by the Rydberg series with a ground-state energy of $R_y = -13.6$ eV. A light field can only create a transition of an electron from the ground state $n=1$ to a higher state $n=2$ if its photon energy $\hbar\omega_1$ matches the energetic separation of both energy levels. Less energetic light $\hbar\omega_2 < \hbar\omega_1$ cannot be absorbed by the unexcited atom. At the same time, the energetic separation of the electronic states defines the frequency spectrum of the light which can be emitted by the atom, producing, for example, the Balmer series in hydrogen.

The presence of strong electromagnetic fields changes this picture. Such a strong external electric field acts as an electrostatic bias, effectively modifying the atomic potential. This is demonstrated in Fig. 1(b), where an external field with a field strength of 100 MV/cm tilts the hydrogen potential. As a result, the previously non-resonant excitation with photon energy $\hbar\omega_2$ can create a transition to an unbound state in the continuum. Consequently, also non-resonant excitations become possible in the strong-field regime (Boyd, 2008). The required strong electro-magnetic bias can be realized dynamically by the electric field of a light pulse created by a strong laser.

In that case, the oscillating electric field will ionize the atom at peak electric fields and subsequently accelerate the electron in the continuum. Upon a change of the field's polarity, the electron will be accelerated back to the ion such that it can recombine under emission of a high-energy photon. Under these conditions, the emitted radiation contains resonances at integer multiples of the exciting field; this effect is known as *high-harmonic generation* (HHG). Atomic HHG is often explained by the above three-step model (Corkum, 1993). It is routinely used in the creation of ultrashort light sources with pulse durations in the attosecond regime, up to X-ray photon energies (Krausz, 2016). Furthermore, it has been demonstrated that HHG may be used as a spectroscopic tool to study the underlying matter excitations. Recent experiments have applied the same type of strong-field excitations to solids, realizing HHG also in solid-state systems like insulators or semiconductors (Ghimire et al., 2011; Schubert et al., 2014; Luu et al., 2015; Vampa et al., 2015).

Principles

An ideal crystalline semiconductor consists of a periodic arrangement of atoms. In such a periodic lattice, the individual energy levels of the atoms merge into the bandstructure ϵ_k^λ which defines the possible values for the kinetic energy of an electron with crystal momentum $\hbar\mathbf{k}$. Generally, multiple bands λ exist, where the energetically uppermost completely filled band is labelled *valence band* and the energetically lowest fully unoccupied band is the *conduction band*. In semiconductors an energetically forbidden region exists between these bands. Their minimal separation defines the bandgap energy, which is typically in the range of 1–2 eV, corresponding to the photon energy of optical fields (Haug and Koch, 2009). Therefore, the non-resonant excitations

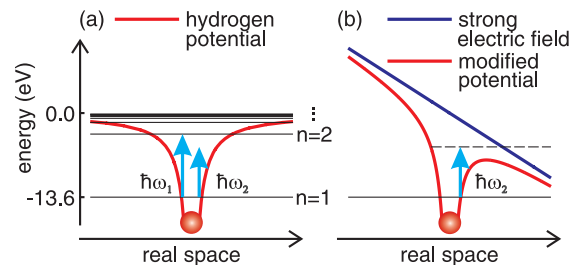


Fig. 1 Resonant vs. non-resonant excitations: (a) Potential energy of a hydrogen atom (red line) together with its electronic eigenstates marked by black lines. A transition of an electron from one energetic state to another can only be realized by a resonant excitation, where the photon energy $\hbar\omega_1$ matches the energetic separation of both states and (b) the presence of a strong electrostatic potential (blue line) creates a modified atomic potential (red line), such that a previously non-resonant excitation with photon energy $\hbar\omega_2$ can create a transition to an unbound state (dashed line).

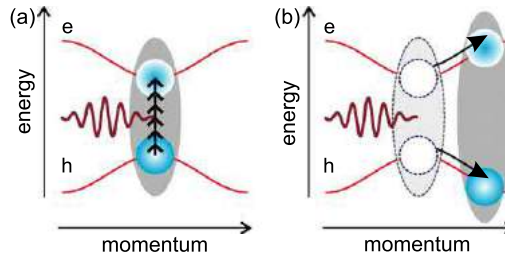


Fig. 2 Fundamental excitations in a semiconductor: (a) Interband excitations induced by a strong THz field (red pulse) lead to the creation of polarization (shaded area) and charge carriers (spheres) also for very non-resonant excitations, where the THz photon energy (black arrows) is much smaller than the bandgap between valence- and conduction band (red lines) and (b) the THz field accelerates charge carriers as well as polarization inside the bands creating intraband currents (black arrows).

needed for HHG can be realized with strong terahertz (THz) or near infrared fields with photon energies of several tens of meV. Since a few years, such strong fields can be experimentally realized (Schubert *et al.*, 2014).

An optical excitation of a semiconductor induces a transition amplitude between a state in the completely filled valence band and the empty conduction band. This microscopic polarization will eventually lead to the creation of charge carriers for strong enough fields, even for very non-resonant excitations (Haug and Koch, 2009). These interband excitations are schematically depicted in Fig. 2(a), where the THz photon energy (black arrows) is much smaller than the bandgap between valence and conduction band (red lines). Simultaneously, the THz field accelerates charge carriers inside the bands as well as the polarization. This leads to the creation of intraband currents, symbolized in Fig. 2(b) by the black arrows. The interplay of interband- and intraband excitations, described by the microscopic polarization and carrier distributions, leads to the emission of high-harmonic radiation (Golde *et al.*, 2011). As HHG is a fully coherent process, the emitted radiation not only depends on the intensity of the excitation, but also its phase (Schubert *et al.*, 2014).

Semiconductor-Bloch Equations

A microscopic quantum many-body theory of HHG is provided by the semiconductor Bloch equations (SBEs) (Haug and Koch, 2009; Kira and Koch, 2006, 2012). The SBEs describe electrons in a periodic lattice, which interact via the Coulomb interaction and with an external light field. In addition, also the coupling to lattice vibrations or other external potentials can be included. The solution of the SBEs yields the time dynamics of the microscopic polarizations and the carrier distributions in their respective bands for arbitrary excitations. As the strong field excitations applied for HHG can couple multiple bands, the SBEs are schematically presented below for a multiband situation with an arbitrary number of bands λ . The polarization between a valence band h and a conduction band e is then given by

$$i\hbar \frac{\partial}{\partial t} p_k^{he} = (\varepsilon_k^{he} + i|e|E(t)\nabla_k) p_k^{he} - \hbar\Omega_k^{eh} (1 - f_k^e - f_k^h) + \sum_{\lambda \neq e,h} (\hbar\Omega_k^{eh} p_k^{\lambda e} - \hbar\Omega_k^{e\lambda} p_k^{h\lambda}) + \Gamma_k^{he} \quad (1)$$

in the electron-hole picture (Haug and Koch, 2009; Schubert *et al.*, 2014). In this description, a missing electron, i.e., a vacancy in the valence band is treated as a separate quasiparticle, a *hole*. The microscopic polarization p_k^{he} depends on the energetic separation ε_k^{he} between bands e and h . It is driven by the electric field via the Rabi frequency Ω_k^{eh} , which contains a product of the electric field $E(t)$ with the dipole matrix element $d_k^{e\lambda}$ between the two bands $\lambda = e$ and $\lambda' = h$. Additionally, p_k^{he} is coupled to all other microscopic polarizations involving either of the two bands e or h via the Rabi frequency. Both, ε_k^{he} and $\hbar\Omega_k^{eh}$ are renormalized by the Coulomb interaction between the charge carriers. The term containing the gradient describes the acceleration of the polarization by the THz field. Higher-order correlations, including interactions between more than two particles, which are created by the Coulomb interaction are included in Γ_k^{he} (Kira and Koch, 2006). In HHG Γ_k^{he} is dominated by scattering terms resulting in a dephasing of the electronic excitations (Golde *et al.*, 2011). The carrier occupation of the conduction band e is determined by

$$\hbar \frac{\partial}{\partial t} f_k^e = -2\text{Im}[\sum_{\lambda \neq e} \hbar\Omega_k^{e\lambda} p_k^{\lambda e}] + |e|E(t)\nabla_k f_k^e + \Gamma_k^e \quad (2)$$

It is driven by the microscopic polarizations between the band e and all other bands, which are connected by dipoles. Similar to the microscopic polarizations, the acceleration of the carrier distributions by the THz field is described by the gradient term, while Γ_k^e contains higher-order correlations. Importantly, Eqs. (1) and (2) are mutually coupled, such that the SBEs form a set of coupled differential equations. This coupling leads to a strong intermixing of interband excitations and intraband currents, both contributing to the high-harmonic emission (Golde *et al.*, 2011). In detail, the emitted radiation has a polarization source $P(t)$, which is defined by a summation over all momentum states of the microscopic polarization between all dipole coupled bands and a current source $J(t)$, which is effectively defined by the carrier distributions. The emitted high-harmonic field $E_{\text{HHG}}(t)$ is then obtained from $E_{\text{HHG}}(t) \propto \frac{\partial}{\partial t} P(t) + J(t)$ in the time domain. Its Fourier transform yields the corresponding frequency spectrum $I_{\text{HHG}}(\omega) \propto |\omega P(\omega) + iJ(\omega)|^2$.

Examples

High-Harmonic Generation in Two Versus Three Bands

In order to characterize the basic properties of high harmonic frequency spectra and to demonstrate the influence of the number of involved electronic bands, $I_{\text{HHG}}(\omega)$ is compared for a two-band system and a three-band system in Fig. 3.

In detail, the SBEs are evaluated for both bandstructure configurations shown in the inset using an identical THz excitation with central frequency of $f_{\text{THz}} = 30$ THz. Using two electronic bands, $I_{\text{HHG}}(\omega)$ in Fig. 3 shows narrow and well-separated resonances at odd multiples of f_{THz} , where the intensity of the harmonics rapidly decreases for higher harmonic orders. Such a behavior is known from atomic HHG, where the inversion symmetry prevents the emission of even harmonic orders (Lewenstein *et al.*, 1994). In a three-band calculation, however, the $I_{\text{HHG}}(\omega)$ spectrum contains resonances at frequencies corresponding to even and odd harmonic orders. Furthermore, a plateau of harmonic orders with comparable intensity is visible. The appearance of a plateau is a typical feature of high-harmonic spectra, which is well-known from atomic HHG and a hallmark of a non-perturbative process. In the non-perturbative regime, which can only be reached by strong excitations, the intensity I_n of the harmonic order n does no longer obey the perturbative scaling law $I_n \sim E^{2n}$.

Non-Perturbative Quantum Interference

The fundamentally different emission characteristics of a two-band configuration in contrast to a system involving at least three bands can be explained by analyzing the microscopic excitation paths. In a two-band system consisting of one conduction band (e) and one valence band (h), as displayed in Fig. 4(a), an excitation between both bands can only proceed directly from one band to another, as symbolized by the gray arrow. Such transitions are denoted as *direct* excitation paths.

The strength of this transition and its contribution to the high-harmonic emission is determined by the polarization $P_{\text{dir}}(t)$ between both bands. Due to the mutual coupling of microscopic polarizations and carrier occupations, $P_{\text{dir}}(t)$ contains a sequence of terms, which depend only on odd powers of the electric field. Carrying out a Fourier transform in order to compute the corresponding emission spectrum will, thus, produce resonances at odd multiples of its driving frequency. Consequently, $I_{\text{HHG}}(\omega)$ computed with two electronic bands shows only odd harmonic orders in Fig. 3. In the presence of strong, non-perturbative excitations, this field dependence of $P_{\text{dir}}(t)$ is then modified into a nonlinear function P_{odd} with odd parity with respect to the driving field $E(t)$, i.e., $P_{\text{odd}}(-E) = -P_{\text{odd}}(E)$.

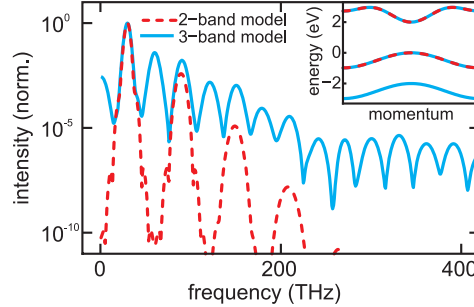


Fig. 3 High harmonic emission spectra: Computed $I_{\text{HHG}}(\omega)$ spectra of a two-band system (red-dashed line) and a three-band system (blue-solid line), which are excited with strong THz light with central frequency $f_{\text{THz}} = 30$ THz. The band dispersion used in both calculations is shown in the inset in corresponding colors. Reproduced from Huttner, U., Schuh, K., Moloney, J.V., Koch, S.W., 2016. Similarities and differences between high-harmonic generation in atoms and solids. *J. Opt. Soc. Am. B* 33, C22–C29.

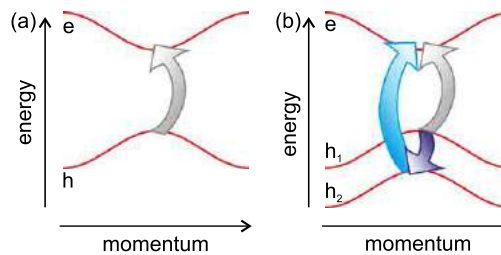


Fig. 4 Direct and indirect transition paths: (a) In a two-band system consisting of one conduction band e and one valence band h, an excitation can only proceed directly from one band to another, symbolized by the gray arrow and (b) considering two valence bands h_1 and h_2 , an excitation from one valence band to the conduction band e may either proceed directly (gray arrow) or via the other valence band forming an indirect transition path, symbolized by the combination of both blue arrows.

In a three-band system, as shown in Fig. 4(b), excitations from one band to another may either proceed directly (symbolized by the gray arrow), or via an intermediate step including the third band (blue arrows). These *indirect* transitions are generated by the coupling of a direct transition from the lower valence band to the conduction band to a transition between the valence bands, mediated by the electric field. As a result, the corresponding polarization $P_{\text{indir}}(t)$ consists of terms which depend only on even powers of the electric field. Therefore, $P_{\text{indir}}(t)$ creates the even harmonic orders in the $I_{\text{HHG}}(\omega)$ spectrum computed with three electronic bands. In analogy to the direct transitions, non-perturbative excitations modify this field dependence of $P_{\text{indir}}(t)$ into a nonlinear function P_{even} with even parity, i.e., $P_{\text{even}}(-E) = P_{\text{even}}(E)$.

The total transition yield P_{tot} is defined by an interference of both transition paths according to

$$P_{\text{tot}}(E) = P_{\text{even}}(E) + P_{\text{odd}}(E). \quad (3)$$

This quantum interference is also present in the non-perturbative regime creating even and odd harmonic orders with comparable strength (Hohenleutner *et al.*, 2015).

Time-Resolved HHG

The interference of different excitation paths becomes most prominent in time-resolved studies (Hohenleutner *et al.*, 2015). Therefore, Fig. 5 shows intensity envelopes $I_{\text{HHG}}(t)$ of computed $E_{\text{HHG}}(t)$ time traces. They are calculated for a strong THz excitation, $E_{\text{THz}}(t)$ using two and three bands, respectively. For two electronic bands, $I_{\text{HHG}}(t)$ shows intensity modulations on a time scale which corresponds to the duration of a THz half-cycle. Furthermore, the emission maxima occur delayed with respect to the field crests of $E_{\text{THz}}(t)$.

In contrast to that, the three-band $I_{\text{HHG}}(t)$ yields high-harmonic radiation as a series of ultrashort bursts, which appear only at positive field cycles of $E_{\text{THz}}(t)$. The emission at negative field cycles is strongly suppressed. Furthermore, the emission bursts are synchronized to the crests of the driving field. Non-perturbative quantum interference can explain this observation. In the two-band computation, the high-harmonic emission is exclusively defined by the direct paths producing the polarization $P_{\text{dir}}(E_{\text{THz}})$. The HHG emission intensity, thus, has only a single source term and is independent of the sign of $E_{\text{THz}}(t)$. In the three-band computation, however, direct and indirect paths contribute to the emission according to Eq. (3) via the sum of their amplitudes $P_{\text{indir}}(E_{\text{THz}}) + P_{\text{dir}}(E_{\text{THz}})$. For positive half cycles of $E_{\text{THz}}(t)$ they interfere constructively. For negative half cycles of $E_{\text{THz}}(t)$, in contrast, a sign flip of the electric field $E_{\text{THz}} \rightarrow -E_{\text{THz}}$ switches from constructive interference to destructive interference $P_{\text{indir}}(-E_{\text{THz}}) + P_{\text{dir}}(-E_{\text{THz}}) = P_{\text{indir}}(E_{\text{THz}}) - P_{\text{dir}}(E_{\text{THz}})$, where the contributions of direct and indirect paths cancel. Consequently, these half cycles lead to strongly suppressed emission. Additionally, it has been shown, that non-perturbative excitations balance the amplitudes of direct and indirect excitation paths producing efficient interference conditions (Hohenleutner *et al.*, 2015).

New, sophisticated experimental techniques based on frequency-resolved optical gating, allow for the monitoring of the high-harmonic emission $I_{\text{HHG}}(t)$ on the same timescale as the exciting THz waveform (Hohenleutner *et al.*, 2015). Fig. 6(a) shows such a measurement in GaSe in comparison to computations performed using five electronic bands in Fig. 6(b). Both, experiment and theory show an identical unipolar high-harmonic emission, as well as a synchronization of the emission bursts to the positive field crests of the driving field.

Dynamical Bloch Oscillations

When a strong field bias is applied to a semiconductor via a static electric field, charge carriers will be accelerated inside the bands leading to an increase of their momentum $\hbar k$. If the field is strong enough, it can accelerate an electron wave packet up to the border of the first Brillouin zone (BZ) and beyond. Upon leaving the first BZ on one side, the wave packet re-enters at the opposite side of the BZ with sign-flipped momentum. This corresponds to oscillations of the electron wave packet in real and momentum space. These oscillations are referred to as *Bloch oscillations*. Remarkably, they are created by a constant electric field bias, however, they are hard to observe experimentally due to the fast scattering of the accelerated electrons.

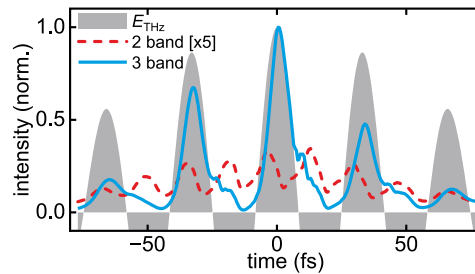


Fig. 5 Time-resolved high-harmonic emission. (a) High-harmonic emission intensity $I_{\text{HHG}}(t)$ as function of time computed for a two-band system (red dashed line) and a three-band system (blue solid line), driven by a strong THz field (shaded area). The emission intensity of the two-band system is magnified by a factor of five. Reproduced from Huttner, U., Schuh, K., Moloney, J.V., Koch, S.W., 2016. Similarities and differences between high-harmonic generation in atoms and solids. J. Opt. Soc. Am. B 33, C22–C29.

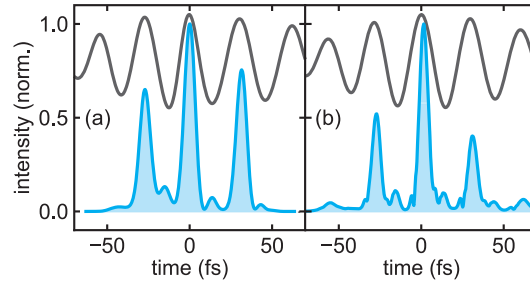


Fig. 6 Theory-experiment comparison of time-resolved high-harmonic emission in GaSe: (a) High-harmonic emission intensity $I_{\text{HHG}}(t)$ (blue) as function of time, measured in GaSe for a THz peak field (black line) of 44 MV/cm with central frequency of 33 THz and (b) corresponding, computed high-harmonic emission using five electronic bands. Adapted from Hohenleutner, M., Langer, F., Schubert, O., *et al.*, 2015. Real-time observation of interfering crystal electrons in high-harmonic generation. *Nature* 523, 572–575.

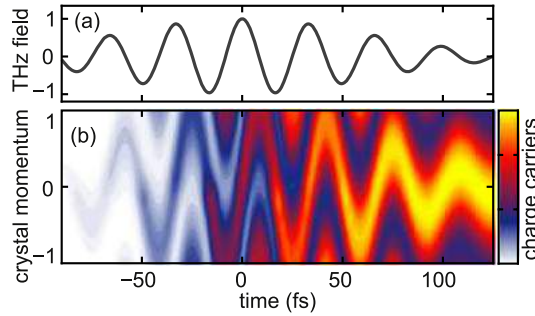


Fig. 7 Dynamical Bloch oscillations: (b) Computed electron wave-packet dynamics in the lowest conduction band of GaSe as function of time and crystal momentum, excited by a strong THz pulse (a) with duration of 100 fs. Adapted from Schubert, O., Hohenleutner, M., Langer, F., *et al.*, 2014. Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nat. Photon.* 8, 119–123.

The situation is different for transient excitation conditions with strong THz fields which can lead to a dynamical version of the Bloch oscillations (Schubert *et al.*, 2014). These oscillations can be visualized in simulations, for example, by monitoring the time-dependent microscopic carrier distribution in a THz excited semiconductor. As an illustration, we show in Fig. 7 the distribution of electrons in the lowest conduction band of GaSe as function of time and crystal momentum, after the excitation of the system by an ultrashort, strong THz waveform depicted in Fig. 7(a). During one half cycle of the THz excitation, the electronic wave packet is accelerated throughout the whole BZ, leaves the BZ on one side, and enters at the other side. Because they create an oscillatory motion of electrons in a non-parabolic bandstructure, these dynamic Bloch oscillations strongly contribute to the emission of high-harmonic radiation (Schubert *et al.*, 2014).

Electron–Hole Recollisions: High-Order Sideband Generation

It is interesting to extend the strong THz excitation scheme by applying an additional resonant light pulse. For example, this light pulse can resonantly create coherent excitons, i.e., an electron-hole (e–h) polarization oscillating with the energy of the excited exciton resonance. Once created, the THz field will then accelerate electrons and holes in opposite directions due to their opposite charge. In a similar fashion as described by the three-step model of atomic HHG, electrons and holes may eventually recollide and recombine under photon emission.

This process of e–h recollisions is the driving mechanism of high-order sideband generation (HSG) (Zaks *et al.*, 2012; Langer *et al.*, 2016). The HSG emission spectrum produces a series of resonances at both sides of the frequency of the optical excitation – called sidebands, which are separated by twice the THz central frequency.

However, in contrast to atomic HHG, the time separation between optical and THz pulse can precisely be adjusted and controlled. Using an ultrashort optical excitation, this time delay defines and controls at which phase of the THz field E_{THz} the coherent excitons are created. Consequently, in analogy to atomic recollision models, this creation time determines the degree to which a recollision between electrons and holes is possible within the ultrashort coherence time of the excitonic excitations. As a result, *good* and *bad* excitation times exist, entailing strong or weak sideband emission, respectively. The quasiparticle dynamics for both excitation times are schematically illustrated by the cartoon in Fig. 8.

At bad excitation times (Fig. 8(a)), coherent excitons are created shortly after a zero-crossing of the THz bias. Consequently, the THz fields separates electrons and holes, such that no recollision occurs. At good excitation times (Fig. 8(b)), however, coherent

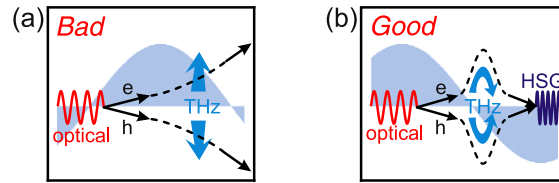


Fig. 8 Schematics of e-h recollisions for good and bad excitation times: An optical excitation (red line) creates coherent e-h pairs, which are subsequently separated and accelerated by a strong THz bias (blue area). For bad excitation times (a), electrons and holes do not recollide within the finite and ultrashort lifetime of the coherent excitons. For good excitation times (b), in contrast, e-h recollisions lead to strong sideband emission (dark blue line). Adapted from Langer, F., Hohenleutner, M.C., Schmid, P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. *Nature* 533, 225–229.

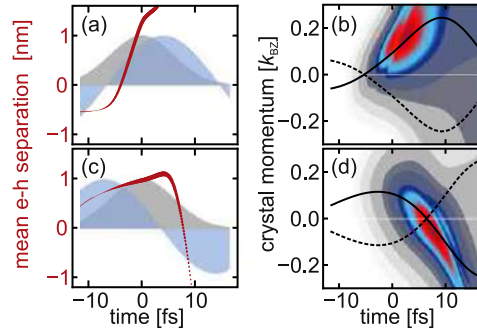


Fig. 9 Microscopic quasi-particle dynamics: (a), (c) Mean e-h separation obtained by evaluating the e-h pair-correlation function for bad (a) and good (c) excitation times. Envelope of the optical excitation (shaded area) and THz waveform (blue area). (b), (d) Electron dynamics in the conduction band as function of time and crystal momentum for bad (b) and good (d) excitation times. Mean electron (solid line) and hole (dashed line) momentum as function of time. Adapted from Langer, F., Hohenleutner, M.C., Schmid, P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. *Nature* 533, 225–229

excitons are created shortly after a field crest of E_{THz} , such that electrons and holes are accelerated and recollided, producing strong sideband emission.

The underlying microscopic quasi-particle dynamics can be monitored in real space, as well as in momentum space. Therefore, **Fig. 9(a)** and **(c)** shows the mean relative distance between an electron and the corresponding hole (red line) in real space for (a) bad and (c) good excitation times. The relative e-h separation $\langle r \rangle$ is defined by means of the e-h pair-correlation function (Kira and Koch, 2006). While $\langle r \rangle$ grows monotonically for bad excitation times, effectively suppressing recollisions, $\langle r \rangle$ is rapidly crossing $r=0$ after an initial increase for good excitation times. Around $r=0$ the electron and hole wavepackets have maximum overlap enabling an efficient recombination, which leads to strong HSG emission, confirming the interpretation of **Fig. 8**.

A similar picture arises in momentum space by monitoring the electron dynamics in the conduction band in **Fig. 9(b)** and **(d)**. For bad excitation times (b), the electron wavepacket is moved away from the center of the BZ, where e-h recombination is most efficient, while the hole wavepacket is moving in the opposite direction. For good excitation times (d), electron and hole wavepackets, which are initially separated from each other, are pushed back towards the BZ center enabling recollisions and producing strong HSG emission.

Consequently, the emitted sideband radiation contains details about the microscopic quasiparticle dynamics and information about the colliding quasiparticles themselves. Monitoring the intensity of the emitted sideband radiation as function of the delay time allows, for example, for an estimate of the excitonic coherence time (Langer *et al.*, 2016).

See also: Coherent Terahertz Sources. Multidimensional Terahertz Spectroscopy. Terahertz Physics of Semiconductor Heterostructures

References

- Boyd, R.W., 2008. *Nonlinear Optics*, third ed. Academic press.
- Corkum, P.B., 1993. Plasma perspective on strong field multiphoton ionization. *Physical Review Letters* 71, 1994.
- Ghimire, S., DiChiara, A.D., Sistrunk, E., *et al.*, 2011. Observation of high-order harmonic generation in a bulk crystal. *Nature Physics* 7, 138.
- Golde, D., Kira, M., Meier, T., Koch, S.W., 2011. Microscopic theory of the extremely nonlinear terahertz response of semiconductors. *Physica Status Solidi (b)* 248, 863–866.
- Haug, H., Koch, S.W., 2009. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. World Scientific.
- Hohenleutner, M., Langer, F., Schubert, O., *et al.*, 2015. Real-time observation of interfering crystal electrons in high-harmonic generation. *Nature* 523, 572.

- Kira, M., Koch, S., 2006. Many-body correlations and excitonic effects in semiconductor spectroscopy. *Progress in Quantum Electronics* 30, 155.
- Kira, M., Koch, S.W., 2012. *Semiconductor Quantum Optics*. Cambridge: Cambridge University Press.
- Krausz, F., 2016. The birth of attosecond physics and its coming of age. *Physica Scripta* 91, 063011.
- Langer, F., Hohenleutner, M., Schmid, C.P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. *Nature* 533, 225.
- Lewenstein, M., Balcou, P., Ivanov, M.Y., LHuillier, A., Corkum, P.B., 1994. Theory of high-harmonic generation by low-frequency laser fields. *Physical Review A* 49, 2117.
- Luu, T.T., Garg, M., Kruchinin, S.Y., *et al.*, 2015. Extreme ultraviolet high-harmonic spectroscopy of solids. *Nature* 521, 498.
- Schubert, O., Hohenleutner, M., Langer, F., *et al.*, 2014. Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nature Photonics* 8, 119.
- Vampa, G., Hammond, T.J., Thire, N., *et al.*, 2015. Linking high harmonics from gases and solids. *Nature* 522, 462.
- Zaks, B., Liu, R.B., Sherwin, M.S., 2012. Experimental observation of electron-hole recollisions. *Nature* 483, 580.

Rydberg States in Semiconductors

Manfred Bayer and Marc Assmann, Technical University of Dortmund, Dortmund, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

In atomic physics, Rydberg atoms are highly excited atoms with at least one valence electron excited to a state with huge principal quantum number n . In the limit of a simple hydrogen-like system, a classical Bohr-like description of the electron is a good approximation of the system. Some peculiar properties of the system can be derived from this correspondence. For circular motion of an electron around a nucleus carrying charge e , the arising Coulomb force acts as a centripetal force on the electron:

$$\frac{e^2}{4\pi\epsilon_0 r^2} = \frac{mv^2}{r} \quad (1)$$

In Bohr-like models angular momentum is quantized, so

$$mvr = n\hbar \quad (2)$$

which yields a direct relation between the electron orbit radius and the principal quantum number:

$$r = \frac{4\pi\epsilon_0 n^2 \hbar^2}{me^2} \quad (3)$$

So the radius of an orbit scales quadratically with the principal quantum number, which means that spatial extensions in the micrometer range can be reached for $n \approx 100$. Many other fundamental properties of Rydberg atoms show similar scaling laws with respect to the principal quantum number (Gallagher, 2005). The dipole moment also shows a quadratic scaling, while the radiative lifetime scales as n^3 and the geometric cross section already follows a n^4 -dependence. Polarizability even scales as n^7 . It is easy to see that such an easily polarizable system, where the largest possible carrier separation is almost equal to the radius of the electron orbit is extremely sensitive to external fields and acts like a miniaturized antenna.

While cold atoms provide an excellent environment for studies of fundamental physics, it is desirable to have a physical system that is more suitable for everyday life when thinking of applications. Semiconductors form the material basis for modern-day electronics and optoelectronics, so one might envisage them as good candidate systems to look for Rydberg physics. The basic excitation of a semiconductor system is the exciton, a pair consisting of a negatively charged conduction band electron and a positively charged valence band hole bound by the attractive Coulomb force. They can be considered as the semiconductor analogues of hydrogen atoms. Indeed highly excited excitons are the most promising candidates for identifying Rydberg physics in a semiconductor setting. However, many characteristic quantities of hydrogen and semiconductor systems differ strongly, which renders the observation of highly excited excitons more complicated. This can already be seen from the binding energy of the states, which is for unperturbed hydrogen-like systems given by $E_n = -\frac{E_{\text{Ryd}}}{n^2}$, where the Rydberg energy E_{Ryd} amounts to:

$$E_{\text{Ryd}} = \frac{\mu e^4}{32\epsilon_0^2 \pi^2 \hbar^2} \quad (4)$$

Differences arise due to the presence of the dielectric environment denoted by ϵ and most importantly due to the difference in effective mass μ . For hydrogen, the proton is about 1836 times heavier than the electron, while electrons and holes have similar mass, which results in the effective mass in atomic and semiconductor systems differing by several orders of magnitude. Therefore, also the Rydberg energy differs strongly. It amounts to 13.6 eV for hydrogen, but is only in the range of meV for typical semiconductors. For most materials such as the prototypical semiconductor GaAs with a Rydberg energy of only 4.2 meV, this means that highly excited states will be spaced so closely that they are not resolvable. Also, the energies of excited states will approach the ionization continuum quickly, so that even at low temperatures the highly excited excitons will be turned into free carriers by thermal activation. On the other hand, many wide band gap materials such as ZnO or GaN might in principle be good candidates for observing Rydberg physics due to the large exciton binding energies in those materials, but they typically suffer from a large number of impurities and inhomogeneous crystals. Monolayers of transition metal dichalcogenides provide another promising candidate system for observing Rydberg physics in a semiconductor setting. They show exciton binding energies of several hundred meV, but currently the linewidths of exciton states in these systems are too large to observe more than the first few excited exciton states.

At the moment, Cuprous Oxide (Cu_2O) shows the most pronounced Rydberg effects in semiconductor systems and the main part of this article will be devoted to Cu_2O . Historically, Cu_2O was the material, in which excitons were observed first by Gross and Karryjew (1952) and Hayashi and Katsuki (1952). It features a moderately large exciton binding energy of about 90 meV and is a direct band gap semiconductor with a band gap of about 2.172 eV. The highest valence and the lowest conduction band are both formed from copper states, namely the 3d and 4s orbitals. Therefore, in contrast to most semiconductor systems, absorption of a photon results in creation of an electron-hole pair at the same atom. The excitons belonging to the transition between these bands are termed the yellow exciton series because of the wavelengths of these transitions are in the yellow range of the electromagnetic spectrum around 590 nm. Other exciton series at higher energies exist and are called the green, blue and violet series, respectively,

but we will primarily focus on the yellow series. As both the conduction and the valence band states have the same parity, direct dipole transitions between band states are forbidden. However, for excitons, the relative motion of electron and hole opens up another degree of freedom. The total symmetry of the exciton is thus given by the direct product of the symmetries of the bands and the envelope. Accordingly, excitons with a P-envelope are dipole allowed in optical transitions (Agekyan, 1977), while those with an S-envelope are not.

Rydberg Excitons in Cu₂O

The series of P-excitons has been investigated already in the early days of spectroscopic studies on Cu₂O, where principal quantum numbers up to $n=9$ were identified (Gross and Karryjew, 1952; Gross, 1956). The series could be extended up to $n=12$ over the years (Matsumoto *et al.*, 1996). Going to even higher n results in entering the Rydberg exciton regime, but there are several limiting factors to the highest principal quantum number that can possibly be observed in a sample. Besides temperature, which will result in thermal ionization of states with small binding energy, the sample quality has the biggest influence on the observable exciton states. As the spatial extension of excitons grows with their principal quantum number, impurities, strain or sample imperfections within this spatial region may prevent the formation of large exciton states. While usually in semiconductor physics high quality crystals are created by means of artificial fabrication, artificial Cu₂O crystals are at current still inferior in quality compared to their natural counterparts. Accordingly, the highest principal quantum numbers observed so far for excitons in Cu₂O have been observed in natural crystals cut and polished from a rock mined at the Tsumeb mine in Namibia. The P-exciton absorption spectrum of a 34 μm thick slab mounted free of strain and held at a temperature of 1.2 K is shown in Fig. 1 (Kazimierczuk *et al.*, 2014). The spectrum shows a large number of lines, so the lower panels show the high-energy parts of the spectrum with increasing resolution. Exciton lines are labeled in terms of their principal quantum number. In total, states up to $n=25$ can be identified. For such large n , it is instructive to estimate the spatial extent of the exciton wavefunction. In analogy to hydrogen, the average radius of an orbital for a given combination of l and n can be calculated as

$$\langle r_{n,l} \rangle = \frac{1}{2} a_B (3n^2 - l(l+1)) \quad (5)$$

where a_B denotes the Bohr radius, which is 1.11 nm for P-excitons in Cu₂O (Kavoulakis and Chang, 1997). For the highest principal quantum number seen in the spectrum one gets $\langle r_{25,1} \rangle = 1.04 \mu\text{m}$, which corresponds to a spatial extension of more than 2 μm . This is equivalent to more than ten times the wavelength of the light in the material or several billion unit cells of the crystal.

A first test, whether Rydberg excitons can be considered akin to Rydberg atoms is to check the n^{-2} scaling law of the binding energies. However, the shape of the exciton resonances shows a characteristic asymmetric Fano-type line shape. This is a consequence of the interference of discrete exciton states with a continuum of states arising due to optical-phonon assisted absorption of the 1S state. The main consequence of this asymmetric line is that the position of the peak maximum does not necessarily coincide with the position of the resonance. The deviation may be larger than 1 meV. Accordingly, it is mandatory to fit the asymmetric line shape given by (Tyoazawa, 1964)

$$\alpha_n(E) = C_n \frac{\frac{\Gamma_n}{2} + 2q_n(E - E_n)}{\left(\frac{\Gamma_n}{2}\right)^2 + (E - E_n)^2} \quad (6)$$

to the experimental data in order to extract the real positions of the resonances. Here E_n is the real position of the n th resonance, the amplitude C_n is proportional to the oscillator strength, q_n describes the asymmetry of the line and Γ_n is the spectral width of the resonance. These energies are shown in Fig. 2. They match the Rydberg formula $E_n = E_G - \frac{E_{Ryd}}{n^2}$ rather well. As the data shows, the band gap energy E_G and E_{Ryd} can be determined to 2.17208 eV and 90–92 meV, respectively.

A slight deviation from the ideal formula is apparent, but can be explained in terms of a quantum defect as will be discussed in detail later. The linewidths for the different n are shown in Fig. 3. For larger n they decrease down to few μeV . Up to about $n=10$, where one can assume that the lines are only homogeneously broadened, the linewidths follow the inverse cubic law expected for Rydberg states. For even larger principal quantum numbers, one encounters some additional broadening that can arise due to crystal imperfections, the influence of the other exciton series or power broadening. Still, the linewidth is as low as 3 μeV for the largest exciton states, which corresponds to lifetimes in the nanosecond range. At first, such long lifetimes might seem surprising, as it might seem that the large spatial extension of these excitons might make it susceptible to scattering. The main recombination pathways besides direct radiative recombination are carrier-carrier scattering, which is negligible at low excitation densities, relaxation into lower exciton states by spontaneous emission of infrared photons, which is unlikely due to the small spontaneous emission rate for low-energy photons and relaxation by emission of an optical phonon, which also shows an inverse cube law scaling as seen in the experiments.

Rydberg Blockade

The final scaling law of interest is the exciton oscillator strength, which is supposed to scale as n^{-3} . The oscillator strengths are proportional to the peak areas of the resonances, which are shown in Fig. 4 for a laser intensity of 6 $\mu\text{W mm}^{-2}$.

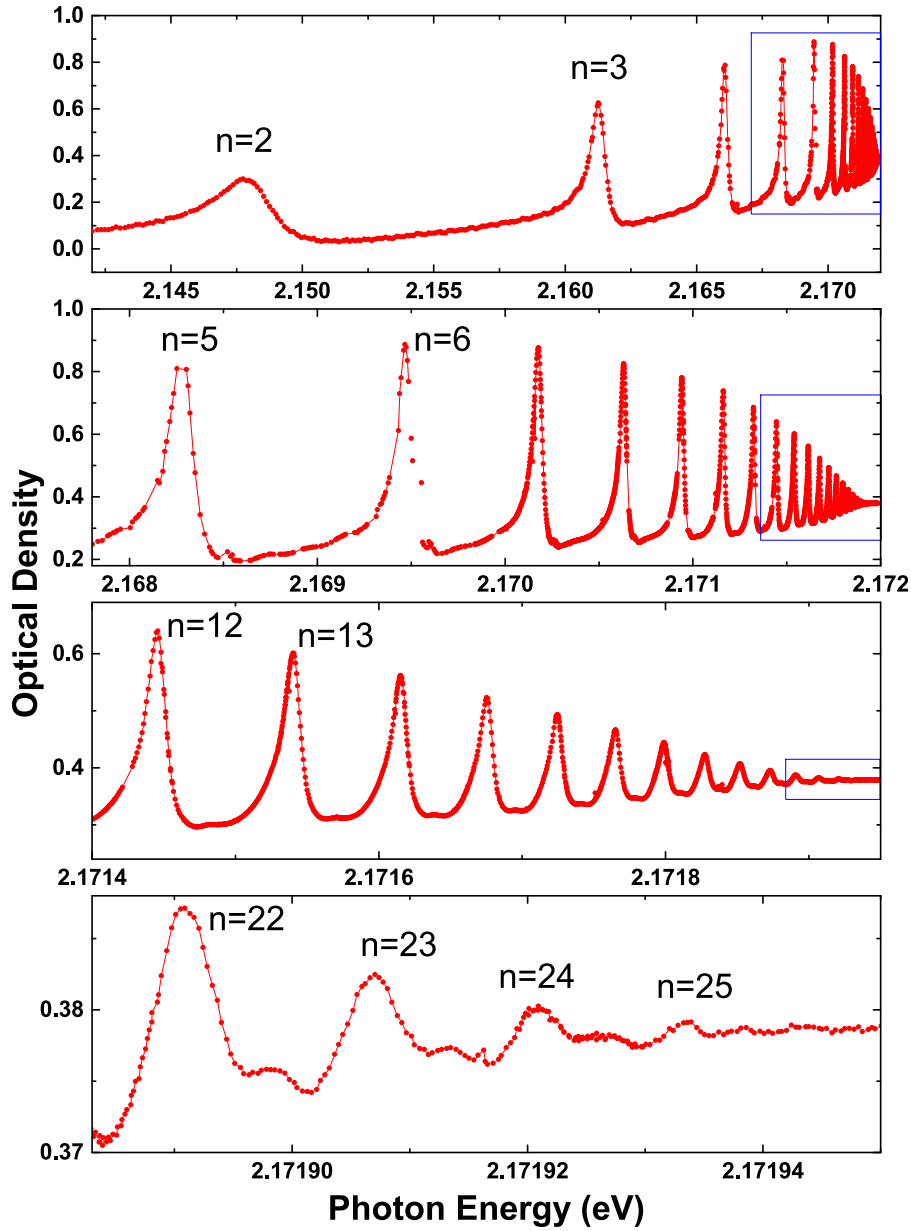


Fig. 1 High-resolution spectrum of the yellow exciton series in Cu_2O . Peaks correspond to P-excitons with principal quantum number n . Blue rectangles mark the region shown in close-up view in the next panel.

The expected inverse cubic law dependence is seen up to $n=17$. Beyond that point, the oscillator strength of larger excitons is significantly reduced. This reduction shows a pronounced intensity dependence as shown in Fig. 5. With increased laser power, the resonances start to vanish continuously with the higher-energy states being effected prior to and stronger than the states of lower n . The areas of the resonance peaks are plotted in Fig. 6 (Kazimierzczuk *et al.*, 2014). One can clearly see that the excitation power, at which the reduction of the peak area sets in, is shifted to lower values with increasing n .

This behavior implies that the reduced absorption can be traced back to exciton interaction effects akin to the dipole blockade effect known from Rydberg atoms (Urban *et al.*, 2009; Gaetan *et al.*, 2009). The blockade arises due to dipole-dipole interactions between Rydberg excitons, which strongly depend on their separation and spatial extension. Upon creation of an exciton, the required energy to create another one will be shifted by the dipole interaction energy. This energy shift will depend on the distance to the first exciton and within a certain blockade volume V_b it will be larger than the linewidth of the laser used for excitation. Accordingly, it will not be possible to create a second Rydberg exciton within V_b . In most cases in a real crystal, the illuminated volume will be larger than V_b . The experimental signature of dipole blockade will thus be a reduction of the absorption α when Rydberg excitons with density ρ_X are present. For a fixed laser power P_L , the absorption of the crystal at a certain laser energy will be reduced from the empty crystal absorption $\alpha_0(E)$ by a factor of $(1 - \rho_X V_b)$. ρ_X will in turn depend on P_L , α and the exciton lifetime τ_n .

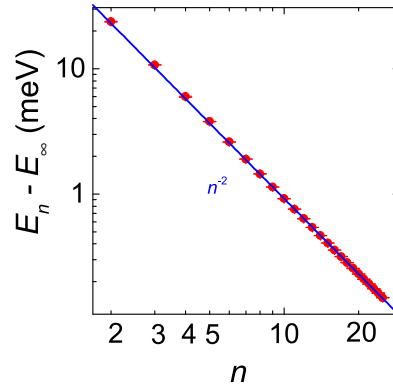


Fig. 2 Rydberg exciton binding energies. Symbols represent the measured energies. The solid line depicts the theoretically expected binding behavior assuming a binding energy of 92 meV. The experimental values show excellent agreement with the expected inverse square law.

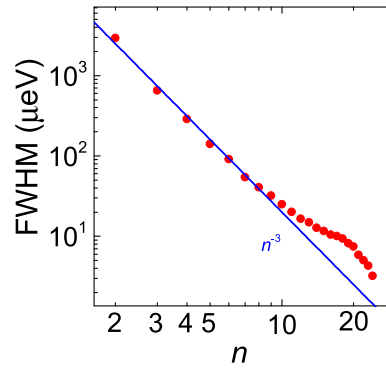


Fig. 3 Spectral width of the Rydberg exciton series. Symbols represent the experimentally determined values. The solid line represents the expected n^{-3} -behavior.

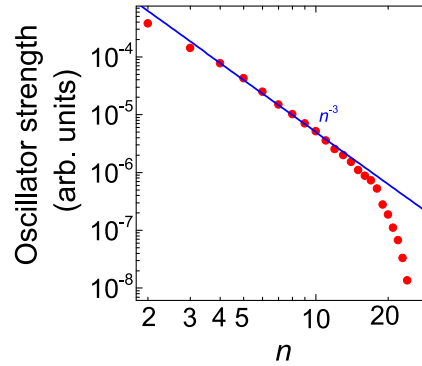


Fig. 4 Oscillator strength of Rydberg excitons. Symbols represent experimental data, while the solid line corresponds to the inverse cubic law expected in the absence of blockade effects.

Taking all of these effects into account, one can find the following simple scaling law for the absorption:

$$\alpha(P_L, E) = \frac{\alpha_0(E)}{1 + S_n P_L} \quad (7)$$

Here S_n denotes an effective efficiency with which the presence of excitons with principal quantum number n blocks further absorption of photons at the same energetic position. **Fig. 6** shows fits of Eq. (7) to the peak areas of the different exciton peaks seen in the experiment. The agreement is reasonable and S_n can be deduced from the fits. For large energies, it changes drastically with the principal quantum number and shows variations of several orders of magnitude. However, this strong effect helps one to

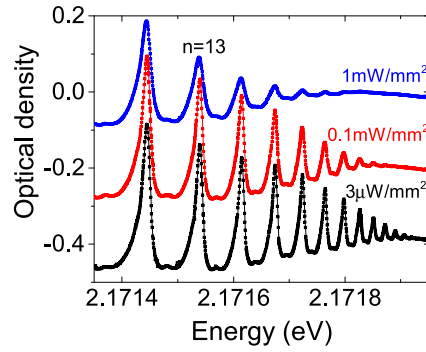


Fig. 5 Rydberg series absorption spectra measured with different laser intensities. The resonances at large n vanish at higher laser powers.

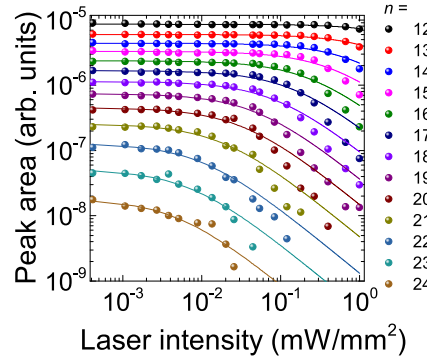


Fig. 6 Oscillator strength of Rydberg excitons for different laser powers. The solid lines represent fits according to Eq. (7). Reproduced from Feldmaier, M., Main, J., Schweiner, F., Cartarius, H., Wunner, G., 2016. Rydberg systems in parallel electric and magnetic fields: An improved method for finding exceptional points. *Journal of Physics B: Atomic, Molecular and Optical Physics* 49 (14), 144002.

identify the nature of the underlying effect causing the blockade. S_n shows a tenth power scaling law behavior with n . Possible dipole-type interaction mechanisms that could explain the blockade are either van-der-Waals interactions, where the interaction energy of two excitons separated by a distance R is given by

$$E_{VDW}(n) = -\frac{C_6}{R^6} \quad (8)$$

or Förster type interactions with a typical interaction energy of

$$E_F(n) = -\frac{C_3(n)}{R^3} \quad (9)$$

Theory shows that C_6 scales as n^{11} , while C_3 scales as n^4 . In the limit of a spectrally narrow laser, the blockade radius R_b can be estimated by the dipole interaction becoming larger than the linewidth of the absorption:

$$R_B(n) = \sqrt[4]{\frac{C_q(n)}{\Gamma_n}} \quad (10)$$

Assuming that the blockade efficiency is proportional to the product of $V_B = \frac{4}{3}\pi R_B^3$ and the exciton lifetime $\tau_n = \frac{\hbar}{\Gamma_n}$, one surprisingly finds that both types of interaction result in a blockade efficiency that scales as n^{10} in agreement with the experimental data, which is a good indicator that dipole blockade is indeed taking place in the exciton system (Fig. 7).

Symmetry Considerations

While it is instructive to study the similarities between hydrogen and excitons, especially for high-symmetry cases such as bulk semiconductors with cubic symmetry, also deviations from the hydrogen model, which arise due to reduced symmetry in other Rydberg systems, are highly interesting. For hydrogen, the spatial symmetry is governed by the continuous rotation group $SO(3)$, which results in the square of the orbital momentum $L^2 = l(l+1)\hbar^2$ and its z -component $L_z = m\hbar$ being conserved. Also, this rotational symmetry assures that energy levels are degenerate with respect to the magnetic quantum number m . The hydrogen problem is also a Kepler-type inverse-square law central force problem. The bound Kepler problem may be mapped to a particle

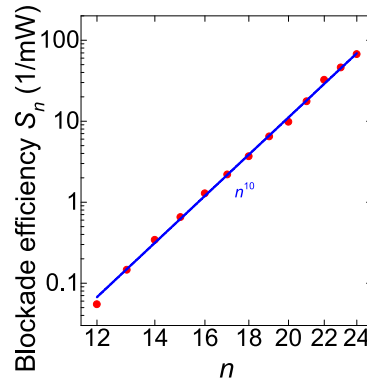


Fig. 7 Blockade efficiency for the different principal quantum numbers of the Rydberg exciton series. The solid line is a fit corresponding to the n^{10} -dependence.

confined on a three-dimensional sphere in four dimensions. Therefore, the problem is also symmetric with respect to rotations in four dimensions and the symmetry group of this problem is $SO(4)$, which results in degeneracy of the energy levels with respect to the azimuthal quantum number l . One may now consider effects arising due to deviations of other Rydberg systems from the highly symmetric hydrogen problem. First, the $SO(4)$ symmetry is broken if the central force shows deviations from the inverse-square law behavior. The most prominent cause for a modified central force is non-perfect screening, which is present in semiconductor Rydberg systems, but also in non-hydrogenic Rydberg atoms, where the inner electron shells only partially screen the nuclear Coulomb potential. For Rydberg atoms, this effect has been considered in terms of a quantum defect δ_l , which results in a modified binding energy. The probability to find an electron close to the nucleus varies for different l , which results in a pronounced dependence of the quantum defect on the angular momentum. However, screening by fully occupied electron shells leaves the $SO(3)$ symmetry intact. The rotational symmetry is also lifted in semiconductor systems due to the presence of the periodic crystal lattice. This results in the reduction of the symmetry from a continuous rotation group towards discrete groups. For cubic bulk crystals these are O_h or T_d , depending on whether inversion symmetry is present or not. These still represent moderately high symmetry. Accordingly, it is in most cases sufficient to capture the deviations from full rotational symmetry in terms of the lattice-periodic Bloch functions of electrons and holes, while still treating the envelope wave function as hydrogen-like. The breakdown of this approximation results in the degeneracy of states with the same l being lifted and also in mixing of states with different l . In the following, we will discuss both symmetry reductions in terms of the equivalent of a quantum defect and the mixing of angular momentum states due to broken rotational symmetry for Rydberg excitons.

Mixing of Angular Momentum States

Due to its large Rydberg energy and small linewidths, Cu_2O is a promising candidate system for observing deviations from perfect rotational symmetry. Indeed these deviations have been observed in high resolution absorption studies. First, it is instructive to consider, which states one expects to see in experiments by means of group theory arguments. The valence and conduction bands associated with excitons of the yellow series correspond to the irreducible representations $\mathcal{D}_h = \Gamma_7^+$ and $\mathcal{D}_e = \Gamma_6^+$, respectively. The excitons may also be categorized in analogy to hydrogen in terms of the symmetry of the relative motion of electron and hole $\mathcal{D}_l$, which may or may not be irreducible. In order for a state to be optically active in one-photon transitions from the crystal ground state, the total symmetry representation of the exciton $\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_l$ must include the representation of the dipole operator corresponding to Γ_4^- . The representations of the most common exciton states (neglecting degeneracies) are listed in Table 1. As one can immediately see, P- and F-excitons are dipole-allowed, while S- and D-excitons are dipole-forbidden, but quadrupole-allowed. Also, one can readily see that only the representations of S- and P-excitons are irreducible, while the other angular momentum states consist of several representations and are thus reducible. This is already a signature of angular momentum not being a good quantum number anymore. Still, one can keep the classification of excitons in terms of angular momentum for simplicity, but should keep in mind that admixtures from other states are present (Schweiner et al., 2016a). Therefore, the appearance of dipole-allowed F-exciton states in optical transmission experiments would be a good indicator for conservation of angular momentum being violated. A high-resolution transmission spectrum of a 30 μm crystal slab of Cu_2O held at 1.2 K in the region up to $n=9$ is shown in Fig. 8 (Thewes et al., 2015). While the P-excitons are most dominant in the spectrum, on their high energy side several weak signatures can be identified for excitons with a principal quantum number larger than 3. It should be noted that the highest possible angular momentum state that can occur for a fixed principal quantum number n is $n-1$. Accordingly, 4 is the lowest n , for which F-excitons can exist. Each of these structures consists of a triplet of lines, where the width of each line is in the low μeV range. Both the splitting between the lines of the triplet and the width of the individual lines decrease with increasing n . This triplet of states corresponds to the F-excitons. Starting from $n=6$ onwards another even weaker feature appears on the high energy side of

Table 1 Basic symmetries of Cu_2O

Object	Symbol	Representation
Electric dipole		Γ_4^-
Electric quadrupole		$\Gamma_3^+ \oplus \Gamma_5^+$
Magnetic dipole		Γ_4^+
Holes	$\mathcal{D}_h$	Γ_7^+
Electrons	$\mathcal{D}_e$	Γ_6^+
S-envelope	$\mathcal{D}_S$	Γ_1^+
P-envelope	$\mathcal{D}_P$	Γ_4^-
D-envelope	$\mathcal{D}_D$	$\Gamma_3^+ \oplus \Gamma_5^+$
F-envelope	$\mathcal{D}_F$	$\Gamma_2^- \oplus \Gamma_4^- \oplus \Gamma_5^-$
Excitons	$\mathcal{D}_h \otimes \mathcal{D}_e$	$\Gamma_2^+ \oplus \Gamma_5^+$
S-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_S$	$\Gamma_2^+ \oplus \Gamma_5^+$
P-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_P$	$\Gamma_2^- \oplus \Gamma_3^- \oplus \Gamma_4^- \oplus 2\Gamma_5^-$
D-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_D$	$\Gamma_1^+ \oplus 2\Gamma_3^+ \oplus 3\Gamma_4^+ \oplus 2\Gamma_5^+$
F-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_F$	$2\Gamma_1^- \oplus \Gamma_2^- \oplus 2\Gamma_3^- \oplus 4\Gamma_4^- \oplus 3\Gamma_5^-$

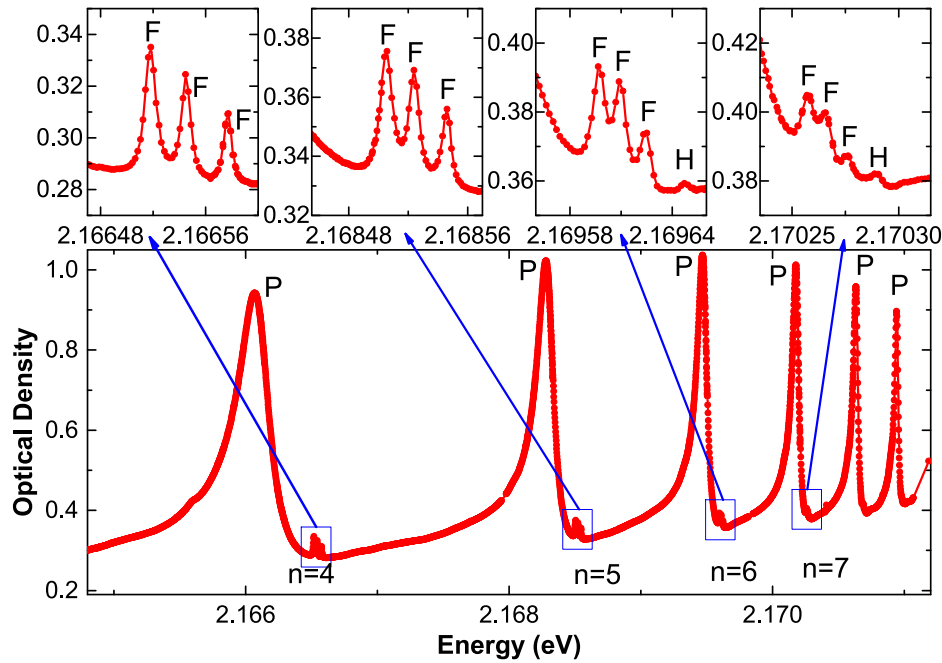


Fig. 8 Optical density of excitons in the spectral region between $n=4$ and $n=9$. On the high-energy side of the P-excitons additional resonances occur, which can be identified as F- and H-excitons. Insets show these resonances in more detail. The F-excitons show a threefold splitting, while a splitting is not resolvable for H-excitons.

the triplet. This state approaches the triplet with increasing n and can be identified as the H-exciton. It may consist of up to 5 lines, which are, however, not resolvable in the spectrum shown.

Quantum Defect and Non-Parabolic Bands

Historically, the first series of discrete emission lines in atomic spectra successfully classified, was the Balmer series of hydrogen (Jakob Balmer, 1885), which revealed the n^{-2} -scaling discussed earlier. However, already slightly more complicated atoms like alkali required modifications to the energy level scheme. The reason for that is that for all hydrogen-like atoms besides hydrogen the positive charges in the nucleus are not completely screened by the inner electrons. If the outer shell electron has a non-zero probability to be found close to the nucleus, the substructure of the nucleus will result in a slightly modified Coulomb potential. Accordingly, the largest deviations occur for s-states, which have a moderately large probability of being close to the nucleus. Rydberg (1890) showed that these deviations can be accounted for by adding an empirical correction to the binding energy term. He replaced

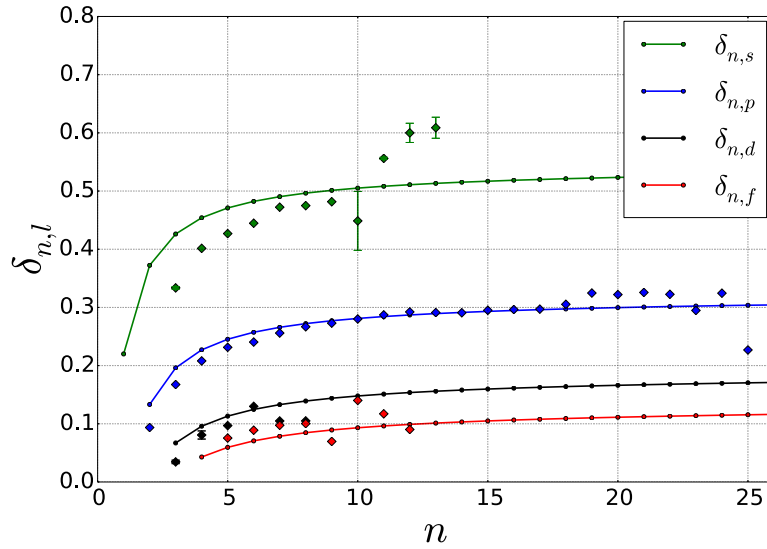


Fig. 9 Comparison of quantum defects determined experimentally (symbols) and via ab initio theory (lines) for the yellow exciton series. For S- and D-excitons, the experimental data has been obtained using an external electric field and interpolation to the zero-field limit. For lines showing a splitting, the center of gravity of the lines has been used. Adapted from Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016a. Coupled valence band dispersions and the quantum defect of excitons in Cu_2O . *Journal of Physics B: Atomic, Molecular and Optical Physics* 49 (13), 134003.

the principal quantum number n by an effective non-integer quantum number $n^* = n - \delta_l$, where δ_l is the quantum defect mentioned before. Although this approach seems simplistic at first sight, it has been hugely successful because the quantum defect directly translates into a phase shift of the wave function in the region outside the core and thus simplifies calculations drastically.

Rydberg excitons in Cu_2O also show deviations from the ideal Rydberg series. These deviations can also be treated in terms of a quantum defect as the quantum defect determined by doing so converges to a constant value for fixed angular momentum and large principal quantum numbers (Schöne *et al.*, 2016b). Of course one has to keep the differences between atom and semiconductor physics in mind: Most calculations in atom physics are performed in real space, while Rydberg excitons are usually treated in momentum space and the physical origin of the deviations is vastly different. In fact, several effects could play a role. Central cell corrections (Washington *et al.*, 1977) corresponding to short range deviations of the electron-hole interaction potential from a pure Coulomb potential might arise due to coupling to LO-phonons (Schweiner *et al.*, 2016b), Coulomb screening, a strongly frequency dependent permittivity of the material or exchange interactions. However, it could be shown by fitting the results of spin-DFT calculations to the band Hamiltonian of Suzuki and Hensel (1974), that the non-parabolicity of the valence band is the dominant contribution to the quantum defect in Cu_2O (Schöne *et al.*, 2016a). A comparison of these results, which have been calculated without using experimentally determined fitting parameters, to the experimental data is shown in Fig. 9. The experimental data were obtained via absorption spectroscopy for P- and F-states. For S- and D-states an additional electrical field was applied in order to mix S-excitons with P-excitons and D-excitons with P- and F-excitons in first order. The results were then extrapolated to zero external field. The overall agreement between theory and experiment is convincing. Similar to the situation in atom physics, the quantum defect becomes largest for states of low angular momentum. The theory value saturates at a value slightly above 0.5 for S-excitons, while it amounts to about 0.3 for P-excitons. For D- and F-excitons the quantum defect is between 0.1 and 0.2, which is a rather large value compared to cold atoms, where the quantum defect approaches zero quickly for large angular momentum states. However, for the S-series one still observes systematic differences between experiment and theory for low n . In this spectral range, the background permittivity $\epsilon(k, \omega)$ varies strongly (Dawson *et al.*, 1973) and the binding energy for $n=2$ is in the close vicinity of a phonon resonance. Most likely the deviations can be traced back to these effects.

Coherent Features

For principal quantum numbers larger than 12, the transmission spectrum of Rydberg excitons shown in Fig. 1 shows some additional resonances between two P-exciton states. Even assuming a huge quantum defect, these additional states are too far away from the P-states to be excitons with larger angular momentum. They appear almost, but not exactly, in the middle between two P-states and are only visible in laser transmission experiments, but not using a white light sources and vanish for large laser powers. It was shown that these resonances can be traced back to coherent coupling between P-excitons of different n (Grünwald *et al.*, 2016) and therefore to off-diagonal elements of the system density matrix in the Fock basis. This coherent effect is a good testbed for estimating whether Rydberg excitons are a suitable candidate for investigating more complex quantum coherent effects. One may gain deeper insights into coherences and dephasing in the system. Especially the latter is usually very strong in bulk systems. A close-up view of some of these intermediate resonances is shown in Fig. 10. Their existence cannot be explained by a theory just

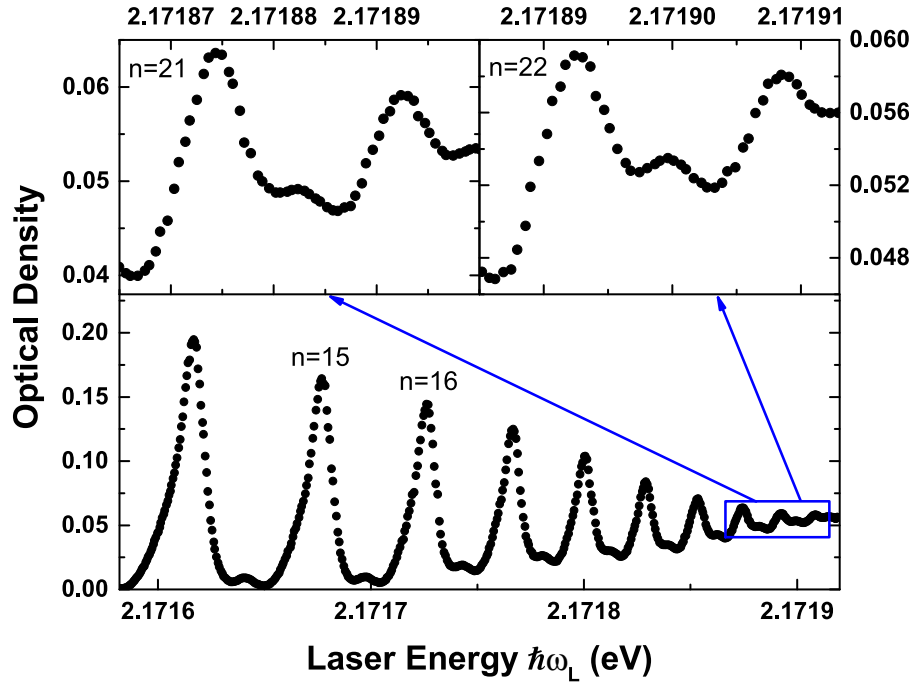


Fig. 10 Experimental absorption spectra of the Rydberg exciton series showing additional resonances between isolated resonances.

assuming the bare absorption spectrum of the system according to the Toyozawa line shape of resonances shown in Eq. (6). However, the presence of a strong light field driving a system may change its spectral properties by introducing incoherent and coherent scattering contributions. These result in resonance fluorescence contributions known, for example, from well-studied effects like the Mollow triplet (Mollow, 1969) and also in modified absorption properties (Mollow, 1972). Accordingly, the total spectral response of the system can be decomposed into a contribution due to the bare absorption $a(E)$ of the system and a function $I_X(E, E_L)$ that describes the incoherent redistribution towards other energies when driving the system at a well defined energy E_L . The total absorption of the system $A(E_L)$ will be given by a convolution of both:

$$A(E_L) = \int_{-\infty}^{\infty} dE a(E) I_X(E, E_L) \quad (11)$$

In the limit of small light-matter coupling strength, the so-called Heitler regime (Matthiesen et al., 2012), I_X will correspond to a delta distribution. In that case, the basic Toyozawa lineshape is recovered. For larger light intensities, one can recover the full spectrum for an isolated resonance using a master equation approach that yields:

$$A(E_L) = g^2 \Omega_n^2 C_n \frac{\frac{\Gamma_n}{2} - 2q_n(E_n - E_L)}{\left[\left(\frac{\Gamma_n}{2} \right)^2 + 2\Omega_n^2 + (E_n - E_L)^2 \right]^2} \quad (12)$$

where g denotes the exciton dipole transition moment and Ω_n represents the Rabi energy. While there are differences between this corrected lineshape and the Toyozawa lineshape, for isolated resonances these can only be clearly distinguished far away from the peak and the resonances are too close to allow for a trivial identification of the more appropriate lineshape. However, there are other effects like power broadening not covered by the Toyozawa lineshape, which can be used to identify the difference in a systematic measurement at different laser powers. Also, for the range of states where the correction matters, the spacing between states of consecutive principal quantum numbers is roughly of the same order of magnitude as the Rabi energy. The energy separation of Rydberg exciton states with $n=15$ and $n=16$ amounts to about $50 \mu\text{eV}$. Accordingly, the resonances rather form a V-type system, where the states $|n\rangle$ and $|n+1\rangle$ are the excited states coupled to the excitonic vacuum $|0\rangle$ and a laser placed in the middle between the two resonances can drive both of them at a moderate detuning. The level scheme is sketched schematically in Fig. 11. The excited states are not coupled to each other and there is only one exciton present in the system. This system can be modeled using a master equation approach (Grünwald et al., 2016). One finds that off-diagonal terms in the density matrix occur, which correspond to an effective quadrupolar coherent coupling of the two excited states via the common ground state. The transition rate directly depends on the two Rabi frequencies. The amplitude of this intermediate resonance will depend strongly on the incoherent redistribution spectrum $S_X = \frac{I_X}{g^2}$ of the excitons, which is shown in Fig. 12 for pumping exactly in the middle between the resonances. When increasing the Rabi energy, the resonance at the laser position starts to build, which corresponds to a dressed state of the two quadrupole-coupled exciton states. The total absorption spectrum is now given by the convolution of I_X and the bare Toyozawa absorption lineshape. A comparison of experimental results and theoretical results obtained by

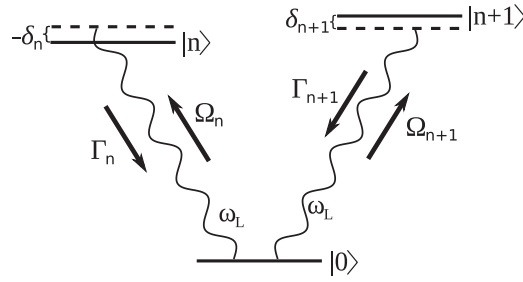


Fig. 11 Model of Rydberg excitons as a V-type three level system. Two Rydberg excitons with principal quantum numbers n and $n+1$ are coupled to the excitonic ground state via a single light field at different detunings.

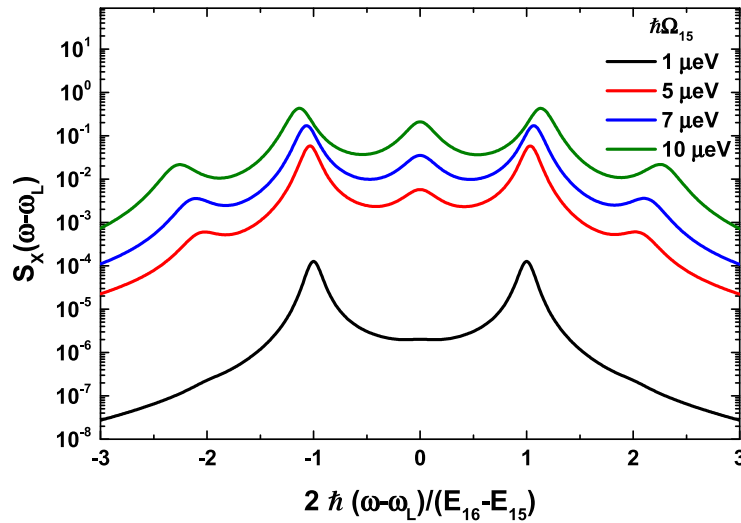


Fig. 12 Incoherent exciton spectral redistribution function for a laser placed exactly in the middle between resonances with principal quantum number $n=15$ and $n=16$.

incorporating the quadrupolar coupling is shown in [Fig. 13](#). Both the asymmetry of the intermediate resonance and its appearance at moderate and disappearance at huge pump powers are reproduced qualitatively. The former is a consequence of the asymmetry of the underlying Toyozawa lineshape, while the latter is caused by the onset of saturation. The onset of Rydberg blockade may also add to this effect, but is not incorporated into the theoretical model. It should be noted that the presence of pure dephasing should destroy these coherences quickly, so the results imply that the pure dephasing rate for Rydberg excitons is surprisingly small, which makes them a good candidate to look for other coherent effects and implement coherent control in a bulk system.

Summary and Outlook

This article has placed a special emphasis on Rydberg states in Cu_2O , but this material system is of course not the only possible candidate to observe Rydberg physics in semiconductor systems. Any system with large exciton binding energy and moderately narrow linewidths might show similar effects. Low-lying levels of the Rydberg series have been observed, for example, in transition metal dichalcogenides ([Chemikov et al., 2014](#)) or impurities in Si ([Vinh et al., 2008](#)), but so far either the linewidths or the quality of the structures prohibit observation of states with large n . This article has placed some emphasis on differences to Rydberg atoms, which arise due to the semiconductor setting. Along these lines, it has further been suggested that the deviations between excitons and the case of hydrogen should also provide deeper insights into the classical-to-quantum transition in large external fields. When placed in moderate magnetic fields, it could already be shown that the Rydberg exciton system shows quantum chaos of a non-trivial kind ([Aßmann et al., 2016](#)). In fact, the breaking of all anti-unitary symmetries results in a drastic change of the level spacing statistics, which in turn leads to a quadratic suppression of small level spacings. As the Rydberg energy of exciton systems is much smaller compared to atom physics, also the regime of high external fields is reached much easier. Typical field strengths available in a laboratory setting are sufficient. Therefore, one may expect to find further specific effects of quantum chaos, which are hard to observe in atomic physics like exceptional points ([Feldmaier et al., 2016](#)), which have been observed so far only in non-Hermitian photonic billiards ([Gao et al., 2015](#)).

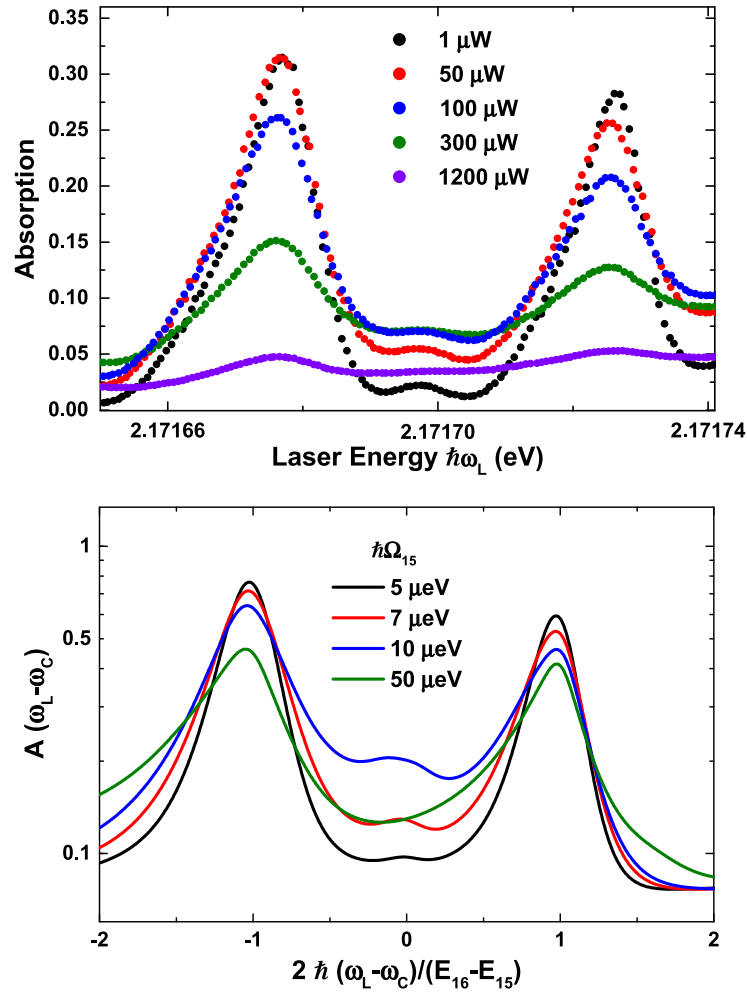


Fig. 13 Upper panel: Experimental absorption spectra between the resonances with $n=15$ and $n=16$ for the laser powers shown in the legend. Bottom: Theoretical calculation of the absorption spectrum in the same spectral region.

For Rydberg excitons in Cu_2O one may envisage that future research will also develop along the lines of what has already been done for cold atoms (Saffman *et al.*, 2010). Either by going to higher principal quantum numbers or thinner samples, it should become possible to create a blockade radius comparable to the thickness of the sample, resulting in an effective two-dimensional Rydberg system, where each blockade volume is occupied by a single exciton. Using tailored pulses matched to the dephasing time and the spectral width of the resonances, it may become possible to drive Rabi oscillations in such a system (Dudin *et al.*, 2012; Johnson *et al.*, 2008), which would pave the way towards transferring the broad range of quantum technologies already realized with Rydberg atoms to a semiconductor system.

Acknowledgements

This article owes much to our collaboration with active researchers in the field. It is our pleasure to express our gratitude to Dietmar Fröhlich, Tomasz Kazimierzuk, Stefan Scheel, Heinrich Stolz, Jörg Main, Harald Gießen and Mikhail Glazov. We also gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft in the frame of SPP 1929 GiRyd and ICRC TRR 160, the Russian Foundation for Basic Research in the frame of ICRC TRR 160 and the support from the Russian Ministry of Science and Education (contract number 14. Z50.31.0021).

References

- Aßmann, M., Thewes, J., Fröhlich, D., Bayer, M., 2016. Quantum chaos and breaking of all anti-unitary symmetries in Rydberg excitons. *Nature Materials* 15, 741.
 Agekyan, V.T., 1977. Spectroscopic properties of semiconductor crystals with direct forbidden energy gap. *Physica Status Solidi (a)* 43 (1), 11–42.

- Chernikov, A., Berkelbach, T.C., Hill, H.M., *et al.*, 2014. Exciton binding energy and non-hydrogenic Rydberg series in monolayer WS_2 . *Physical Review Letters* 113 (7), 076802.
- Dawson, P., Hargreave, M.M., Wilkinson, G.R., 1973. The dielectric and lattice vibrational spectrum of cuprous oxide. *Journal of Physics and Chemistry of Solids* 34 (12), 2201–2208.
- Dudin, Y.O., Li, L., Bariani, F., Kuzmich, A., 2012. Observation of coherent many-body Rabi oscillations. *Nature Physics* 8 (11), 790–794.
- Feldmaier, M., Main, J., Schweiner, F., Cartarius, H., Wunner, G., 2016. Rydberg systems in parallel electric and magnetic fields: An improved method for finding exceptional points. *Journal of Physics B: Atomic, Molecular and Optical Physics* 49 (14), 144002.
- Gaetan, A., Miroshnychenko, Y., Wilk, T., *et al.*, 2009. Observation of collective excitation of two individual atoms in the Rydberg blockade regime. *Nature Physics* 5 (2), 115–118.
- Gallagher, T.F., 2005. *Rydberg Atoms*, vol. 3. Cambridge University Press.
- Gao, T., Estrecho, E., Bliokh, K.Y., *et al.*, 2015. Observation of non-hermitian degeneracies in a chaotic exciton-polariton billiard. *Nature* 526 (7574), 554–558.
- Gross, E.F., 1956. Optical spectrum of excitons in the crystal lattice. *Il Nuovo Cimento* (1955–1965) 3, 672–701.
- Gross, E.F., Karyjew, I.A., 1952. The optical spectrum of the exciton. *Doklady Akademii Nauk SSSR* 84, 471–474.
- Grünwald, P., Alßmann, M., Heckötter, J., *et al.*, 2016. Signatures of quantum coherences in Rydberg excitons. *Physical Review Letters* 117, 133003.
- Hayashi, M., Katsuki, K., 1952. Hydrogen-like absorption spectrum of cuprous oxide. *Journal of the Physical Society of Japan* 7, 589.
- Jakob Balmer, J., 1885. Notiz über die spectrallinien des wasserstoffs. *Annalen der Physik* 261 (5), 80–87.
- Johnson, T.A., Urban, E., Henage, T., *et al.*, 2008. Rabi oscillations between ground and Rydberg states with dipole–dipole atomic interactions. *Physical Review Letters* 100, 113003.
- Kavoulakis, G.M., Chang, Y.-C., Baym, G., 1997. Fine structure of excitons in Cu_2O . *Physical Review B* 55 (12), 7593.
- Kazimierczuk, T., Fröhlich, D., Scheel, S., Stolz, H., Bayer, M., 2014. Giant rydberg excitons in the copper oxide Cu_2O . *Nature* 514 (7522), 343–347.
- Matsumoto, H., Saito, K., Hasuo, K., Kono, S., Nagasawa, N., 1996. Revived interest on yellow-exciton series in Cu_2O : An experimental aspect. *Solid State Communications* 97, 125–129.
- Matthiesen, C., Vamivakas, A.N., Atatüre, M., 2012. Subnatural linewidth single photons from a quantum dot. *Physical Review Letters* 108, 093602.
- Mollow, B.R., 1969. Power spectrum of light scattered by two-level systems. *Physical Review* 188, 1969–1975.
- Mollow, B.R., 1972. Stimulated emission and absorption near resonance for driven systems. *Physical Review A* 5, 2217–2222.
- Rydberg, J.R., 1890. XXXIV. On the structure of the line-spectra of the chemical elements. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 29 (179), 331–337.
- Saffman, M., Walker, T.G., Mølmer, K., 2010. Quantum information with Rydberg atoms. *Reviews of Modern Physics* 82, 2313–2363.
- Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016a. Coupled valence band dispersions and the quantum defect of excitons in Cu_2O . *Journal of Physics B: Atomic, Molecular and Optical Physics* 49 (13), 134003.
- Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016b. Deviations of the exciton level spectrum in Cu_2O from the hydrogen series. *Physical Review B* 93, 075203.
- Schweiner, F., Main, J., Feldmaier, M., Wunner, G., Uihlein, C., 2016a. Impact of the valence band structure of Cu_2O on excitonic spectra. *Physical Review B* 93, 195203.
- Schweiner, F., Main, J., Wunner, G., 2016b. Linewidths in excitonic absorption spectra of cuprous oxide. *Physical Review B* 93, 085203.
- Suzuki, K., Hensel, J.C., 1974. Quantum resonances in the valence bands of germanium. I. Theoretical considerations. *Physical Review B* 9 (10), 4184.
- Thewes, J., Heckötter, J., Kazimierczuk, T., *et al.*, 2015. Observation of high angular momentum excitons in cuprous oxide. *Physical Review Letters* 115, 027402.
- Toyozawa, Y., 1964. Interband effect of lattice vibrations in the exciton absorption spectra. *Journal of Physics and Chemistry of Solids* 25 (1), 59–71.
- Urban, E., Johnson, T.A., Henage, T., *et al.*, 2009. Observation of Rydberg blockade between two atoms. *Nature Physics* 5 (2), 110–114.
- Vinh, N.Q., Greenland, P.T., Litvinenko, K., *et al.*, 2008. Silicon as a model ion trap: Time domain measurements of donor Rydberg states. *Proceedings of the National Academy of Sciences* 105 (31), 10649–10653.
- Washington, M.A., Genack, A.Z., Cummins, H.Z., *et al.*, 1977. Spectroscopy of excited yellow exciton states in Cu_2O by forbidden resonant Raman scattering. *Physical Review B* 15 (4), 2145.

Two-Dimensional Coherent Spectroscopy of Transition Metal Dichalcogenides

Galan Moody, National Institute of Standards & Technology, Boulder, CO, United States

Published by Elsevier Ltd.

Introduction

Coherent optical spectroscopy is a powerful technique for understanding light–matter interaction in condensed matter, atomic, and molecular systems. The simplest conceptual experiment one can perform is to excite a system with a weak optical field (compared to the internal fields of the system) and measure the optical absorption. The incident light generates a coherent superposition between the ground and excited states, which is described macroscopically as a polarization that is linearly proportional to the incident field (Fig. 1). The linear absorption spectrum, arising from interference between the incident light and re-radiated light from the polarization, provides information about the optical transition frequencies, linewidths, and dipole moments.

Despite its utility, linear absorption spectroscopy has several limitations, particularly when used to study heterogeneous ensembles. An individual emitter in the ensemble – a semiconductor nanostructure, atom, or molecule – experiences scattering processes that alter its phase, resulting in free induction decay, or dephasing, of the optical polarization (or equivalently, *homogeneous broadening* γ of the absorption linewidth). Each emitter is also coupled to the surrounding environment, which leads to stochastic fluctuations of its optical and electronic properties. For example, random disorder potentials from imperfections in semiconductor nanostructures introduce a distribution of transition frequencies to the system (*inhomogeneous broadening* σ), which results in polarization decay as individual emitters in the ensemble oscillate out of phase. In linear spectroscopy, inhomogeneous broadening masks the homogeneous optical response as shown in Fig. 1.

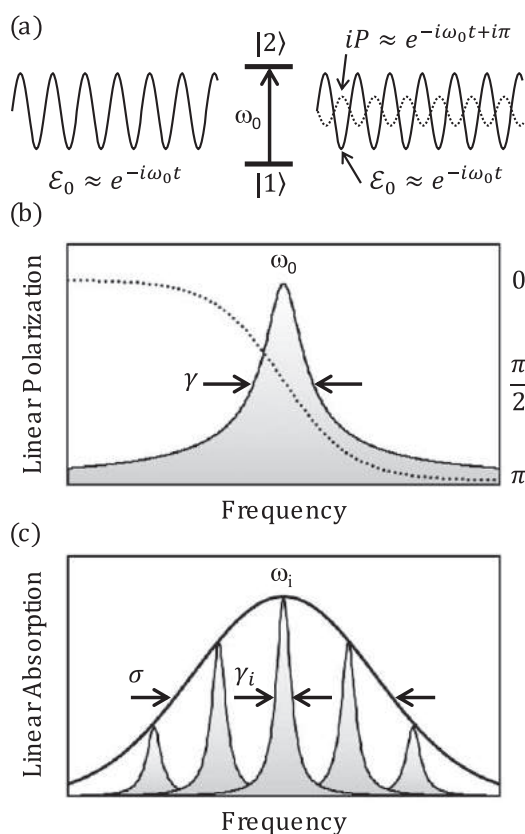


Fig. 1 (a) An incident optical field on resonance with a two-level transition drives a macroscopic polarization, which radiates out of phase with the incident field. (b) The full-width at half-maximum of the polarization frequency response is equal to the homogeneous dephasing rate $\gamma = \hbar/T_2$. On resonance, the polarization is $\pi/2$ out of phase with the incident field (dashed line, right axis). (c) The linear absorption spectrum of an ensemble of two-level systems, where the i th transition has frequency ω_i and homogeneous linewidth γ_i . The distribution of transition frequencies leads to inhomogeneous broadening σ of the absorption linewidth, masking the individual transitions.

This ambiguity can be circumvented with coherent nonlinear optical spectroscopy techniques, which are performed with sufficiently strong optical fields to induce a nonlinear polarization that is proportional to higher-order contributions of the electric field. If the incident field comprises a series of pulses with variable delays, the dynamical evolution of the system's optical response can be measured. One of the most widely-used nonlinear spectroscopy techniques is two- or three-pulse transient four-wave mixing (FWM). The FWM field – re-radiated light from the nonlinear polarization – contains rich information about the photo-excited states in the system. Examples include spectrally resolved transient absorption measurements that provide excited-state lifetimes and photon-echo experiments that reveal the average homogeneous dephasing rate of an ensemble even in the presence of disorder.

An enhanced version of three-pulse FWM, multi-dimensional coherent spectroscopy (MDCS) provides a more comprehensive picture of a system's optical and electronic properties. Whereas one-dimensional nonlinear spectroscopy yields the average system response, the appeal of MDCS lies in its ability to interrogate different emitters in an ensemble. In MDCS, the phase evolution of each oscillator is correlated across two (2DCS) or three (3DCS) time periods. Fourier transformation of time-domain measurements spreads the optical response into a multi-dimensional spectrum, elucidating both the individual and ensemble-averaged properties of the system. In principle, each emitter can be isolated and studied one-at-a-time with one-dimensional techniques; however, this approach does not provide details of interactions or coupling between them.

The Fourier-transform methods of optical MDCS were originally developed for nuclear magnetic resonance spectroscopy of spin systems. With the advent of femtosecond laser technologies, multi-dimensional Fourier-transform techniques emerged in the infrared and visible regimes to study vibrational and electronic coherences in molecules, energy transfer and many-body interactions in semiconductor nanostructures, and dipolar interactions in atomic vapors. More recently, MDCS experiments on a new class of atomically thin semiconductors – namely, monolayer transition metal dichalcogenides – have demonstrated the capabilities of the technique for disentangling complex one-dimensional spectra, elucidating the Coulomb interaction strengths between photo-excitations, and revealing new types of quasiparticles not easily accessible with other techniques.

In this article, the basic principles of multi-dimensional coherent spectroscopy are introduced with a focus on optical 2DCS. A typical experimental implementation and a theoretical description of light-matter interaction in semiconductors based on the optical Bloch equations are introduced. The capabilities of 2DCS in identifying and characterizing fundamental optical excitations in disordered media are illustrated with examples from recent experiments on monolayer semiconductors.

Optical Two-Dimensional Coherent Spectroscopy

Coherent Light-Matter Interaction

Nonlinear optical spectroscopy techniques rely on sufficiently strong optical fields to drive the polarization beyond the linear regime, i.e.

$$P = \chi^{(1)} \mathcal{E} + \chi^{(2)} \mathcal{E}^2 + \chi^{(3)} \mathcal{E}^3 + \dots \quad (1)$$

where $\mathcal{E}$ is the incident electric field, $\chi^{(n)}$ is the n^{th} -order susceptibility, and P is the induced polarization (the vector and tensor notation has been omitted). For weak fields, only the linear response is measured, which contains information about the refractive index and linear absorption. Second-order $\chi^{(2)}$ nonlinearities, which encompass three-wave mixing effects including sum- and difference-frequency generation, are absent in materials with spatial inversion symmetry. Thus, the lowest-order nonlinear optical response that exists in all materials arises from $\chi^{(3)}$ processes.

The $\chi^{(3)}$ susceptibility gives rise to a host of nonlinear optical phenomena, including Kerr-lens mode-locking, self-phase modulation, and four-wave mixing (FWM), which is the basis for optical 2DCS techniques. In 2DCS experiments, each optical pulse from the output of a mode-locked oscillator or parametric amplifier is split into four phase-coherent replicas, three of which generate an optical polarization in the system. The incident field intensity is sufficiently strong to drive the system into the nonlinear regime but is kept weak enough to avoid generating higher-order nonlinearities beyond $\chi^{(3)}$, i.e. FWM spectroscopy is a perturbative technique. The FWM signal field is measured independently from other linear and nonlinear contributions to the polarization through spatial or temporal pulse modulation schemes, discussed in more detail in the following section.

The FWM excitation pulse sequence for optical 2DCS experiments is shown in Fig. 2. The electric field of each pulse can be written as $\mathcal{E}_i(\mathbf{r}, t) = \hat{\mathcal{E}}_i e^{i(\mathbf{k}_i \mathbf{r} - \omega_i t)} + \text{c.c.}$, where $\mathbf{k}_i$, ω_i , and $\hat{\mathcal{E}}_i$ are the wavevector, angular frequency, and field envelope of the i th pulse. The total electric field is a sum of the three incident fields: $\mathcal{E}(\mathbf{r}, t) = \sum_i \mathcal{E}_i(\mathbf{r}_i, \omega_i, t - t_i)$, where the i th pulse arrives at time t_i . Only considering contributions to the third-order polarization term, $P^{(3)} = \chi^{(3)} \mathcal{E}^3$, that are a product of all three fields, the polarization can be written as

$$P^{(3)}(\mathbf{r}, t_1, t_2, t_3) = \int_0^\infty \mathcal{R}^{(3)}(t'_1, t'_2, t'_3) \mathcal{E}_1(\mathbf{r}, t'_1 - t_1) \mathcal{E}_2(\mathbf{r}, t'_2 - t_2) \mathcal{E}_3(\mathbf{r}, t'_3 - t_3) dt'_1 dt'_2 dt'_3, \quad (2)$$

where $\mathcal{R}^{(3)}$ is the third-order time-dependent response function that contains details of the system's transition frequencies, dipole moments, excited-state dynamics, couplings, etc. The n th-order response function $\mathcal{R}^{(n)}$ can be calculated using time-dependent perturbation theory; in general, a sum-over-states expression for $\mathcal{R}$ can be derived for an arbitrary system. $P^{(3)}$ obtained from Eq. (2) can be inserted into Maxwell's equations as a source term to describe radiation and propagation of the FWM signal field.

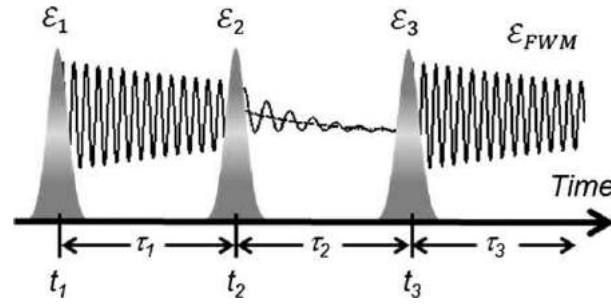


Fig. 2 Time-ordering for three-pulse four-wave mixing (FWM). The FWM signal field can be measured as a function of delays τ_1 , τ_2 , and τ_3 . Fourier transformation with respect to any two delays generates a two-dimensional coherent spectrum.

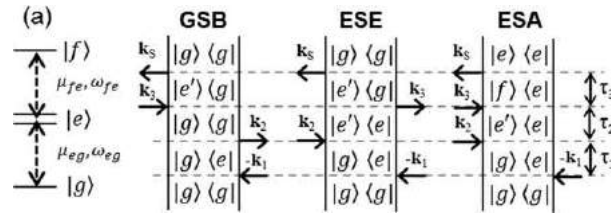


Fig. 3 Double-sided Feynman diagrams representing the coherent quantum pathways for a 2DCS experiment in the rephasing time-ordering (conjugate pulse arrives at the sample first). The singly-excited state manifold $|e\rangle$ with transition frequency ω_{eg} and dipole moment μ_{eg} is accessed through ground-state bleaching (GSB) and excited-state emission (ESE) pathways; the doubly-excited state manifold $|f\rangle$ with transition frequency ω_{fe} and dipole moment μ_{fe} is accessed through excited-state absorption (ESA) pathways.

A convenient approach for generating $\mathcal{R}^{(3)}$ is through the density matrix formalism, which is useful for describing a statistical ensemble of quantum states. The equations of motion for the density operator $\rho = \sum_n p_n |\psi_n(t)\rangle\langle\psi_n(t)|$ are written as $\dot{\rho} = -i/\hbar[H, \rho]$, where $|\psi_n(t)\rangle$ is the wave function of state n with occupation probability p_n . The light-matter interaction with the system can be described by the Hamiltonian $H = H_0 + H_I$, where the free-particle eigenenergies are given by H_0 and the light-matter interaction takes on the form $H_I = -\boldsymbol{\mu} \cdot \boldsymbol{\mathcal{E}}(\mathbf{r}, t)$ in the dipole approximation. The equations of motion for ρ , known as the optical Bloch equations (OBEs), describe the time evolution of the density matrix elements, which as expressed do not contain any relaxation or dephasing parameters. These can be included phenomenologically, resulting in modified OBEs with matrix elements described by

$$\dot{\rho}_{ij} = \frac{-i}{\hbar} \sum_k \left(H_{ik} \rho_{kj} - \rho_{ik} H_{kj} \right) - \gamma_{ij} \rho_{ij}, \quad (3)$$

where $\gamma_{ij} = (\Gamma_i + \Gamma_j)/2 + \gamma_{ij}^{ph}$, Γ_i (Γ_j) is the decay rate of state i (j), and γ_{ij}^{ph} ($i \neq j$) is the elastic pure dephasing rate corresponding to scattering processes that affect coherence without altering the population state of the system. The OBEs are useful for modeling strong light-matter interaction that gives rise to non-perturbative phenomena such as Rabi flopping and coherent control of quantum systems.

For FWM experiments that are performed in the weak-field limit, the light-matter interaction can be treated perturbatively by expanding the OBEs in terms of the Rabi frequency $\Omega \equiv \mu \hat{\mathcal{E}}/\hbar$. The n th-order equation of motion for the density matrix (in the rotating-wave approximation) is expressed as:

$$\dot{\rho}_{ij}^{(n)} = -i\Delta_{ij}\rho_{ij}^{(n)} + \frac{i\Omega}{2}\rho_{kl}^{(n-1)}, \quad (4)$$

where n is the perturbation order and $\Delta_{ij} \equiv \omega_i - \omega_j - i\gamma_{ij}$. From Eq. (4), it is evident that the light-matter interaction increases the perturbation order from $n-1$ to n . To first-order, the optical field drives the system from a ground state population $\rho_{11} = |1\rangle\langle 1|$ (diagonal element of the density matrix) to an optical coherence $\rho_{12} = |1\rangle\langle 2|$ or $\rho_{21} = |2\rangle\langle 1|$ (off-diagonal elements). The action of the second field drives the coherence to a second-order population (ρ_{11} or ρ_{22}). Thus, odd-orders of the field generate coherences while even-orders generate populations.

Four-wave mixing experiments can be modeled by expanding Eq. (4) up to $\rho^{(3)}$, the components of which can be conveniently represented through doubled-sided Feynman diagrams that depict the quantum pathways connecting the initial and final states. Representative Feynman diagrams for a three-level ladder system are shown in Fig. 3, where the singly excited-state manifold $|e\rangle$ is connected to the ground state and doubly excited-state manifold $|f\rangle$ through the electric dipole interaction. The Feynman diagrams are shown for a “rephasing”, or “photon echo”, time ordering of the pulse sequence in which the first pulse incident on the sample acts as a conjugate field. In an inhomogeneous ensemble, the field $\mathcal{E}_1$ drives a first-order coherence that evolves with opposite phase compared to the third-order coherence generated by subsequent excitation with fields $\mathcal{E}_2$ and $\mathcal{E}_3$. In an inhomogeneous

ensemble, polarization decay due to inhomogeneous dephasing of the different oscillators during τ_1 is reversed during the delay τ_3 , resulting in constructive interference of the polarization and a photon echo signal.

The three Feynman diagrams in Fig. 3 represent ground state bleaching (GSB), excited-state emission (ESE), and excited state absorption (ESA) quantum pathways. Feynman diagrams are extremely useful for developing an intuitive understanding of FWM spectroscopy experiments. For example, $\rho^{(3)}$ for the ESA pathway is given by

$$\rho_{32}^{(3)} = \frac{i\mu_{fe}\mu_{eg}\mu_{ge}}{8\hbar^3} e^{ik_s r} \hat{\mathcal{E}}_1^* \hat{\mathcal{E}}_2 \hat{\mathcal{E}}_3 \Theta(\tau_1) \Theta(\tau_2) \Theta(\tau_3) \times e^{-(\gamma_{ge} + i\omega_{ge})\tau_1} e^{-\gamma_{ee}\tau_2} e^{-(\gamma_{fe} + i\omega_{fe})\tau_3}, \quad (5)$$

for $|e\rangle=|e'\rangle$ and delta-function pulse envelopes. In Eq. (5), μ_{ij} and $\omega_{ij}=\omega_i-\omega_j$ are the transition dipole moment and frequency between states i and j , respectively, Θ is the Heaviside step function, $k_s = -k_1 + k_2 + k_3$ is the signal wavevector, and γ_{ij} is the relaxation rate defined in Eq. (3). This expression can be adapted for other quantum pathways by replacing the variables with the appropriate values in the corresponding Feynman diagram. The macroscopic third-order polarization, $P^{(3)}$, is obtained by taking the trace of $\rho^{(3)}$ and μ .

Time-domain expressions like Eq. (5) for all quantum pathways for a specific level system can be summed to obtain the total third-order optical response. A two-dimensional rephasing spectrum is then generated by taking a Fourier transform with respect to τ_1 and τ_3 . For instance, Fig. 4 illustrates how the different quantum pathways contribute to the 2D rephasing spectrum for different types of level systems. The vertical and horizontal axes are the frequency-domain representation of the first-order (excitation energy, $\hbar\omega_1$) and third-order (emission energy, $\hbar\omega_3$) coherences during the delays τ_1 and τ_3 , respectively, while the delay τ_2 is fixed at zero. Because the first field $\mathcal{E}_1$ acts as a conjugate pulse for this sequence, the excitation energies are shown as negative.

For a three-level V-system, the 2D spectrum features four peaks. The two peaks on the diagonal dashed line arise from excitation and emission of the two transitions described by GSB and ESE pathways in Fig. 3, where $|e\rangle=|e'\rangle$. The off-diagonal peaks indicate quantum mechanical coupling between the transitions, which is expected because they share a common ground state. These peaks are described by GSB and ESE quantum pathways for which $|e\rangle \neq |e'\rangle$. The first two pulses drive the system into a non-radiative Raman-like coherence that evolves at the difference frequency between the transitions, which appears as peaks off of the diagonal line in the 2D spectrum; in the time-domain, these low-frequency oscillations are responsible for coherent quantum beats between the transitions.

For two-independent two-level systems, the absence of a shared state eliminates the off-diagonal coupling peaks (right panel of Fig. 4) and only two peaks appear on the diagonal; however, off-diagonal peaks can appear if incoherent population relaxation, or energy transfer, is present between the transitions. In the example in Fig. 4, energy relaxation from state B to state A appears as a below-diagonal peak. The dynamics of this process is described by $\rho_{BB} \rightarrow \rho_{AA}$, which can be mapped by generating 2D spectra for various delays τ_2 .

Examples of simple, homogeneous systems have been presented to illustrate the ability of 2DCS to discern different types of coupling between transitions. Inhomogeneous broadening in disordered media can also be modeled by integrating expressions like Eq. (5) over a two-dimensional frequency distribution function with characteristic width σ (see Fig. 5). In a rephasing 2D spectrum, inhomogeneity appears as broadening of a peak along the diagonal dashed line. In the inhomogeneous limit ($\sigma \gg \gamma$), the diagonal and anti-diagonal lineshapes provide a measure of the inhomogeneous and homogeneous linewidths, respectively. Thus, the homogeneous linewidth for an ensemble can be obtained for all resonance energies by taking anti-diagonal slices across the inhomogeneous distribution. For an arbitrary amount of inhomogeneous broadening described by a Gaussian distribution

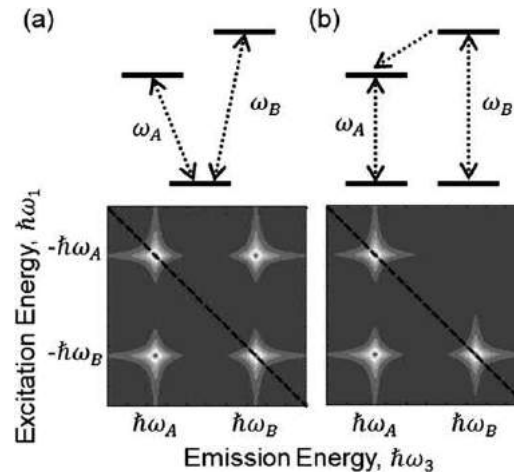


Fig. 4 (a) A three-level V-system and the corresponding rephasing 2D spectrum. (b) Two-independent two-level systems coupled by incoherent energy relaxation (population transfer) from state B to state A and the corresponding 2D spectrum.

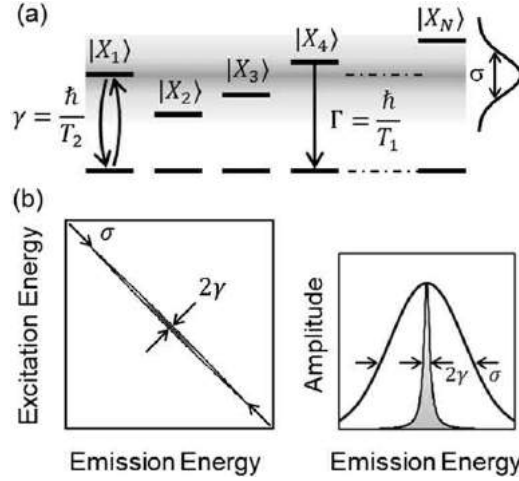


Fig. 5 (a) A heterogeneous ensemble of two-level systems with a distribution of transition energies given by σ ; each transition $|X_i\rangle$ has a characteristic homogeneous linewidth $\gamma_i = \hbar/T_{2,i}$ and excited state lifetime $\Gamma_i = \hbar/T_{1,i}$. (b) Calculated 2D rephasing spectrum for a Gaussian inhomogeneous distribution function. In the strong inhomogeneous limit, the lineshapes along the anti-diagonal and diagonal directions, shown in the right panel, provide a measure of the homogeneous (γ) and inhomogeneous (σ) linewidths, respectively.

function, the projection-slice theorem of Fourier transforms enables the diagonal (S_D) and anti-diagonal (S_{AD}) lineshapes to be expressed as

$$S_D = \frac{\sqrt{2\pi}}{\gamma} \text{Voigt}(\gamma, \sigma), \quad (6)$$

$$S_{AD} = \frac{e^{-(\gamma-i\omega)^2/2\sigma^2} \text{Erfc}\left(\frac{\gamma-i\omega}{\sqrt{2}\sigma}\right)}{\sigma(\gamma-i\omega)} \quad (7)$$

where the *Voigt* profile is a convolution of Gaussian and Lorentzian functions and *Erfc* is the complementary error function. In the limit of strong inhomogeneous broadening, these expressions take on the form of Gaussian and square-root Lorentzian lineshapes, respectively.

The density matrix formalism presented in this section provides a background for intuitively interpreting and modeling a 2D spectrum in the rephasing pulse sequence. The rephasing spectrum reveals the individual and collective properties of disordered media generally not accessible with other techniques. For example, projecting the 2D spectrum onto the emission energy axis is analogous to acquiring an pump-probe spectrum; in this case, the homogeneous and inhomogeneous linewidths cannot be separated, and diagonal peaks spectrally overlap with off-diagonal peaks, masking evidence of coherent coupling in the system. Several different types of 2D spectra not discussed here are also accessible by employing different pulse time-ordering sequences and scanning different delays. These techniques can provide additional insight into the nature of many-body interactions, lifetime dynamics, and non-radiative coherences.

Experimental Implementation of 2DCS

2DCS techniques are implemented in several different configurations, several of which are depicted in Fig. 6. In one commonly used variant, the pulses propagate in a non-collinear geometry (the so-called “box” geometry). For three pulses $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ incident on the system (in that order) with wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$, respectively, the FWM signal field is radiated in the phase-matched direction determined by conservation of momentum, which separates the signal from other linear and nonlinear contributions to the polarization. Different types of 2D spectra can be generated by detecting the FWM signal along different directions: a rephasing or “photon echo” spectrum (S_I) along $\mathbf{k}_s = -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$; a non-rephasing spectrum (S_{II}) along $\mathbf{k}_s = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$; and a two-quantum spectrum (S_{III}) along $\mathbf{k}_s = \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$. In practice, the signal is usually detected along only one of these directions (see Fig. 6) and the different types of 2D signals are measured by changing the arrival time of the conjugate pulse $\mathcal{E}_1^*$. Other pulse geometries include a two-pulse pump-probe configuration and a collinear geometry, in which case the 2D signals are isolated through dynamic phase cycling techniques instead of wavevector selection.

An advantage of 2DCS compared to conventional pump-probe and linear spectroscopy techniques is its ability to measure both the FWM field and phase. Phase-sensitive measurements are performed through heterodyne interferometry with a fourth phase-stabilized local oscillator (LO) field $\mathcal{E}_{LO}$ with a known spectral phase. In a 2DCS experiment, the FWM/LO interference $\hat{\mathcal{E}}_s e^{i\phi} \hat{\mathcal{E}}_{LO}^* e^{-i\phi_{LO}} + c.c.$ is recorded as a function of two delays. For example, a 2D rephasing spectrum is obtained by recording the FWM signal while delays τ_1 and τ_3 are stepped by scanning pulses $\mathcal{E}_1^*$ and $\mathcal{E}_{LO}$. The FWM signal is then numerically Fourier

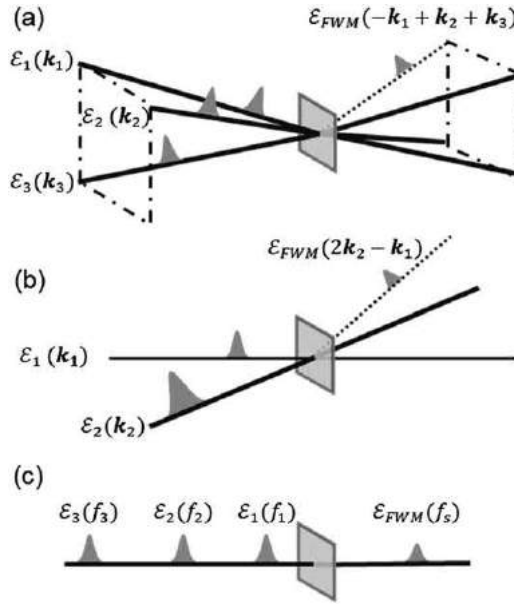


Fig. 6 Different implementations for MDCS experiments: (a) Three-pulse “box” geometry where the FWM signal is detected along $\mathbf{k}_s = -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$; (b) Two-pulse pump–probe geometry, where the strong pump field $\mathcal{E}_1$ acts twice, resulting in $\mathbf{k}_s = 2\mathbf{k}_2 - \mathbf{k}_1$; (c) Collinear geometry in which the FWM signal is detected through phase-cycling techniques.

transformed with respect to these delays to generate the excitation and emission energies $\hbar\omega_1$ and $\hbar\omega_3$ of the 2D spectrum, respectively. Alternatively, the emission energies can be accessed by spectrally resolving the FWM/LO interference.

The FWM pulse sequence shown in Fig. 6 is obtained by splitting the output of a mode-locked laser into a series of variable-delay replicas using Michelson interferometers, Mach–Zehnder interferometers, or diffraction optics and pulse shapers. With any approach, the relative phase of the pulses must be precisely controlled with accuracy to much less than the wavelength to reliably perform a multi-dimensional Fourier transform of the FWM signal. Phase stability to better than $\lambda/100$ and timing-stepping resolution of ~ 1 fs are possible using actively-stabilized interferometers or common path (passively-stabilized) optics. The advantage of the interferometer-based approach is the long maximum scan duration ($\sim$ nanoseconds compared to $\sim$ picoseconds for pulse shapers), but at the expense of greater experimental complexity.

In the rephasing time-ordering, 2DCS is an extremely powerful technique for characterizing disordered media owing to the separation and simultaneous determination of homogeneous and inhomogeneous dephasing. This capability makes 2DCS an ideal technique to investigate layered van der Waals semiconductors including transition metal dichalcogenides, which are sensitive to disorder and the surrounding environment. In the following section, examples of 2DCS experiments that have elucidated the intrinsic optical properties of 2D semiconductors are presented.

Optical Properties of Monolayer Transition Metal Dichalcogenides

Monolayer transition metal dichalcogenides (TMDs) are a recently discovered class of layered van der Waals semiconductors with the chemical formula MX_2 , where $M = \text{Mo, W}$ and $X = \text{S, Se, and Te}$. In the bulk form, TMDs are an indirect bandgap semiconductor with a honeycomb lattice. When reduced to a single monolayer (three atoms thick), TMDs become a direct bandgap semiconductor with conduction and valence band extrema at the K and K' points, or valleys, in the first Brillouin zone in momentum space (see Fig. 7). The combination of large spin-orbit splitting and time-reversal symmetry between the valleys introduces valley-contrasting properties; charges residing at the K and K' points possess opposite orbital magnetic moment, spin, and Berry curvature. This link between the electronic properties and the valley index leads to valley-dependent transition dipole selection rules. Specifically, photo-excitation of electron–hole pairs occurs in the K (K') valley for left (right) circularly polarized optical excitation.

One consequence of the atomic thickness of monolayer TMDs is a reduction in dielectric screening of the Coulomb interaction between band-edge electrons and holes. Combined with the large carrier effective masses, photo-excited electron–hole pairs form tightly bound hydrogenic states (excitons) with a binding energy on the order of 500 meV – nearly two orders of magnitude larger than conventional semiconductors such as GaAs. The exciton binding energy is significantly larger than k_bT at room temperature, which has implications for practical opto-electronic, photonic, and spin/valleytronic devices. In addition to excitons, higher-order few-body configurations exist, including charged excitons (trions) in doped TMDs, bound two-excitons (biexcitons), and charged biexcitons.

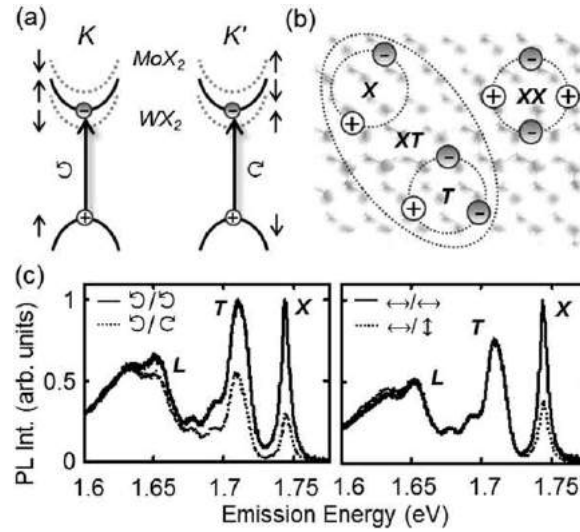


Fig. 7 (a) Optical transitions in monolayer TMDs at the K and K' points at the edge of the first Brillouin zone. Electron-hole pairs are excited in the K (K') valley for left (right) circularly polarized light. The optical transition to the lowest- (highest-) energy conduction band is dark in tungsten (molybdenum) based materials, indicated by the dashed bands. (b) Strong Coulomb interactions lead to the formation of tightly bound electron-hole pairs (excitons, X), charged excitons (trions, T), biexcitons (XX), and charged biexcitons (XT). (c) Photoluminescence spectra of monolayer WSe₂ at 10 K for circularly (linearly) polarized excitation and detection in the left (right) panel. The peaks are associated with excitons (X), trions (T), and localized (L) states. (Part (c) is adapted from Jones *et al.* 2013. Optical generation of excitonic valley coherence in monolayer WSe₂. *Nature Nanotechnology* 8: 634–638. Copyright (2013) by the Nature Publishing Group.)

Owing to their large binding energies, excitons and trions dominate the band-edge optical response of TMDs. Shown in **Fig. 7**, the low temperature photoluminescence spectrum of monolayer WSe₂ features several peaks that are attributed to excitons (X), trions (T), and localized states (L) that arise from impurities, defects, and many-body interactions with background charge carriers in doped samples. Several important aspects of TMDs can be gleaned from this spectrum alone: (1) the trion binding energy (difference between the exciton and trion resonance energies) is ~ 30 meV; (2) the optical linewidths are on the order of ~ 10 meV; and (3) after photo-excitation with left circularly polarized light, the exciton and trion emission is primarily left polarized. This last point reveals that photo-excitation and recombination of excitons and trions primarily occurs from the same valley (valley polarization). Furthermore, using linearly polarized excitation, the emission polarization axes for the exciton is parallel with the excitation, which is evidence of a coherent superposition of an exciton between the K and K' valleys (valley coherence).

The dynamics of valley depolarization and decoherence are sensitive to the type of TMD examined; for example, the lowest-energy transition in WX₂ (MoX₂) is optically dark (bright), which can affect the valley and lifetime dynamics. Despite the utility of photoluminescence (and other linear spectroscopies) in identifying excitonic states in TMDs, more advanced spectroscopy techniques are required to fully characterize the incoherent (lifetime) and coherent (dephasing time) dynamics.

2DCS of Monolayer TMDs

Homogeneous and Inhomogeneous Broadening

A representative 2D rephasing spectrum of monolayer WSe₂ on a transparent sapphire substrate is shown in **Fig. 8** for co-circular excitation and detection of the signal. The spectrum features a single peak centered on the diagonal dashed line associated with the exciton. The peak lineshape asymmetry indicates that the exciton is inhomogeneously broadened, which is attributed to a spatially varying disorder potential from impurities and defects that weakly localize excitons in the transverse plane of the sample.

A square-root Lorentzian fit to the anti-diagonal lineshape yields a homogeneous linewidth $\gamma = 2.7$ meV corresponding to a dephasing time $T_2 = 250$ fs (in this case, $\sigma \gg \gamma$). Compared to T_1 measurements of the exciton lifetime, T_2 measurements are particularly sensitive to many-body interactions that perturb the phase evolution of optical coherences. Excitation-density dependent measurements of 2D spectra, obtained by increasing the average power of the excitation fields, reveal that the homogeneous linewidth increases by more than a factor of two for an order-of-magnitude increase in the exciton density (see **Fig. 9**). The density dependence of γ is a clear indication of excitation-induced dephasing (EID) due to exciton-exciton interactions. A comparison of the slope in **Fig. 9** to other material systems, such as GaAs and CdTe bulk and quantum well nanostructures, reveals nearly an order of magnitude stronger exciton-exciton interaction. Pronounced EID effects in TMDs are consistent with strong Coulomb interactions due to reduced dielectric screening in two dimensions.

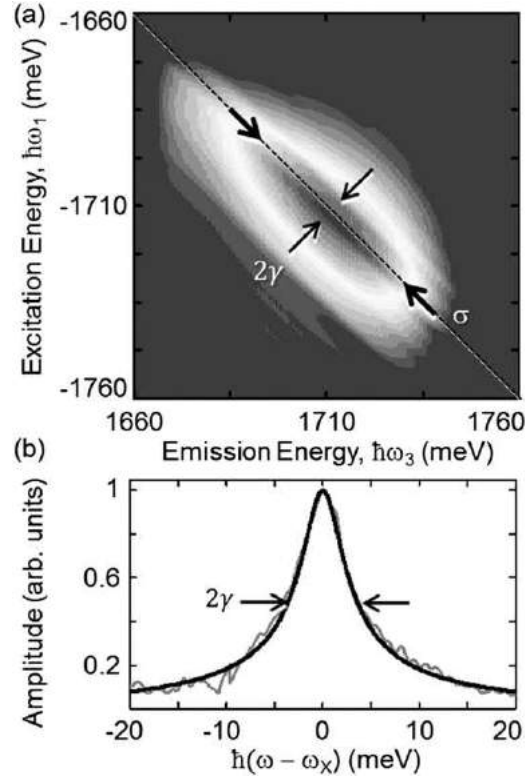


Fig. 8 (a) 2D rephasing spectrum amplitude of excitons in WSe₂ at 10 K. The linewidth along the diagonal dashed line provides a measure of the inhomogeneous broadening (σ) due to static disorder; the linewidth along the anti-diagonal direction provides a measure of the homogeneous linewidth γ . (b) The homogeneous lineshape is fit with a square-root Lorentzian function with a full-width at half-maximum equal to 2γ .

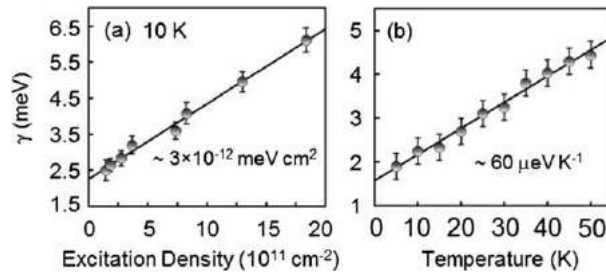


Fig. 9 The homogeneous linewidth of excitons in monolayer WSe₂ measured as a function of excitation density (a) and temperature (b). Extrapolating the linear fits (solid lines) reveals an intrinsic, lifetime-limited linewidth at zero temperature and density of $\gamma = 1.6$ meV ($T_2 = 0.4$ ps).

The role of exciton scattering with phonons is also apparent from the homogeneous broadening with increasing temperature (Fig. 9). The linear increase in γ with temperature T is indicative of exciton dephasing from acoustic phonons with energy much smaller than $k_b T$. Extrapolation of the linewidth to zero temperature and density reveals a residual dephasing rate $\gamma = 1.6$ meV ($T_2 = 400$ fs), which is more than an order of magnitude smaller than the inhomogeneous linewidth $\sigma \approx 50$ meV.

Additional 2DCS experiments performed by scanning the delay τ_2 instead of τ_1 , which are sensitive to the exciton population lifetime (T_1) dynamics, reveal that $T_2 = 2T_1$, i.e. no additional pure dephasing exists in the material at low temperature and density ($\gamma_{ph} = 0$). Thus, exciton dephasing is limited only by the recombination lifetime (sub-picosecond). Short T_2 and T_1 times are consistent with theoretical models of the radiative lifetime, which is predicted to be on the order of ~ 100 fs due to the large exciton oscillator strength.

Coherent and Incoherent Energy Transfer

In doped TMD monolayers, the linear optical spectrum features a peak associated with the trion in addition to the exciton (Fig. 7); however, 2DCS reveals a more complex scenario than just two independent transitions. Fig. 10 shows a time series of the 2D

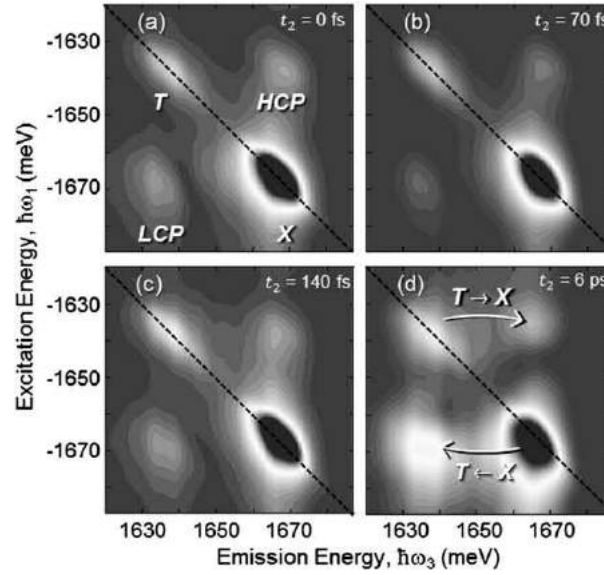


Fig. 10 Normalized 2D rephasing spectrum (amplitude) of monolayer MoSe₂ for co-circular polarization at 10 K versus delay τ_2 equal to (a) 0 fs, (b) 70 fs, (c) 140 fs, and (d) 6 ps. At short delays (τ_2 less than a few hundred femtoseconds), oscillations of the off-diagonal coupling peaks *HCP* and *LCP* indicate coherent coupling between excitons (*X*) and trions (*T*). At longer delays ($\tau_2 > 1$ ps), an increase in the relative magnitude of *LCP* and *HCP* compared to *X* and *T* indicate incoherent energy and charge transfer between excitons and trions due to trion formation (*LCP*) and dissociation (*HCP*).

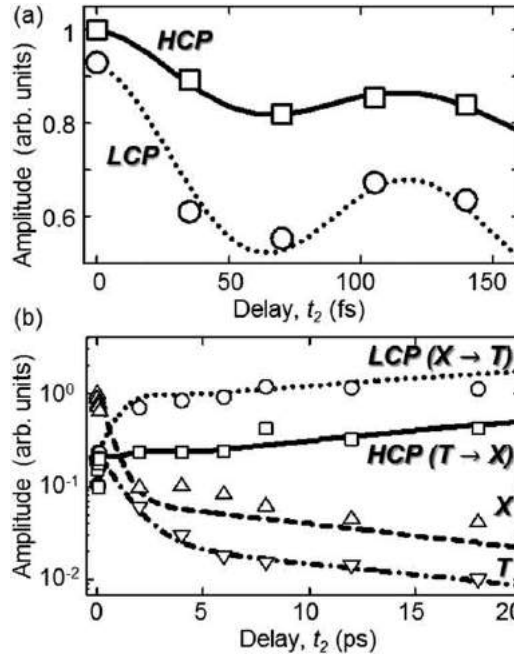


Fig. 11 (a) Oscillations of the cross peaks *HCP* and *LCP* versus delay τ_2 arise from quantum beats between the exciton and trion. The data are modeled with the optical Bloch equations for a four-level diamond system, revealing the dephasing time $\tau_c = 250$ fs and the oscillation period $\tau_{XT} = 130$ fs (corresponding to the exciton–trion energy difference of ~ 31 meV). (b) On longer timescales, the exciton (*X*) and trion (*T*) peaks exhibit biexponential relaxation dynamics due to radiative and nonradiative recombination and exciton–trion energy transfer. The increase in relative magnitude of the *LCP* and *HCP* (compared to *X* and *T*) indicate exciton-to-trion and trion-to-exciton energy transfer, respectively.

rephasing spectrum from a doped (*n*-type) monolayer MoSe₂ at 10 K. At $\tau_2 = 0$ fs, the spectrum features two peaks on the diagonal dashed line that correspond to GSB and ESE of the exciton (*X*) and trion (*T*). The 2D spectrum reveals a new aspect of TMDs not apparent from the linear spectrum – the appearance of off-diagonal peaks (*LCP* and *HCP*), which are evidence of coherent coupling between excitons and trions. Furthermore, the off-diagonal peak amplitudes oscillate as the delay τ_2 is scanned up to a

few hundred femtoseconds. These oscillations can be understood from the fact that fields $\mathcal{E}_1^*$ and $\mathcal{E}_2$ drive the system into a second-order coherence described by $\rho_{XT}=|X\rangle\langle T|$ (HCP) and $\rho_{TX}=|T\rangle\langle X|$ (LCP). These non-radiative Raman-like coherences oscillate during the second delay τ_2 at the exciton–trion difference frequency (Fig. 11). Modeling the 2D spectra with the OBEs for a four-level diamond system, which enables exciton–trion interactions to be included phenomenologically, yields excellent agreement with the data (solid lines in Fig. 11). An exciton–trion coherence dephasing rate $\gamma_c=2.6$ meV ($\tau_c=250$ fs) and an oscillation period $\tau_{XT}=130$ fs (corresponding to the exciton–trion difference energy of ~ 31 meV) are obtained from the fits.

On longer timescales after the coherences have decayed ($\tau_2 > 1$ ps), the off-diagonal peaks decay more slowly than the diagonal exciton and trion peaks. The persistence of the coupling peaks is a characteristic signature of incoherent energy transfer between the two states. For example, exciton-to-trion relaxation (equivalently, trion formation) can be described by quantum pathways in which the first two pulses $\mathcal{E}_1^*$ and $\mathcal{E}_2$ drive the system into an excited-state exciton population $\rho_{XX}=|X\rangle\langle X|$. During τ_2 , relaxation occurs through $\rho_{XX} \rightarrow \rho_{TT}$, which is monitored by the interaction with the third pulse $\mathcal{E}_3$. Whereas the off-diagonal peaks exhibit a relative enhancement in their amplitudes, the diagonal peaks decay with a biexponential behavior (see Fig. 11).

A global fit to the exciton and trion population relaxation and transfer dynamics is performed using a rate equation model that considers radiative and non-radiative recombination of bright and dark excitons and trions and incoherent energy transfer between them. Results from the model (solid lines in Fig. 11) reveal the dominant population relaxation processes in this sample: (1) $\sim$ picosecond radiative recombination and bright-to-dark state scattering for X and T; (2) exciton-to-trion down-conversion (trion formation) via the exciton binding to a residual electron in ~ 2.5 ps; and (3) trion-to-exciton up-conversion (trion dissociation) in ~ 8 ps.

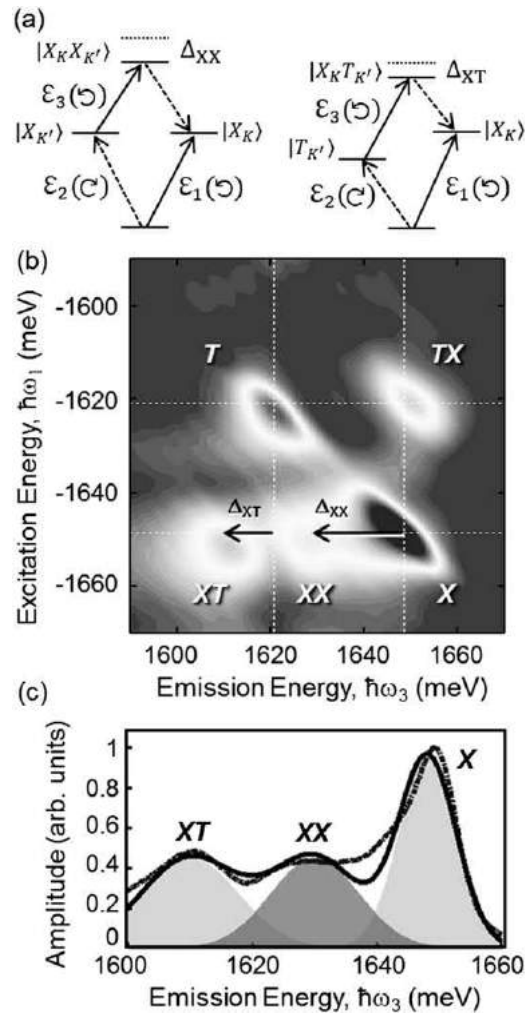


Fig. 12 Neutral and charged biexcitons appear in the 2D spectrum of monolayer MoSe₂ at 10 K when using a cross-circularly polarized excitation sequence. The quantum pathways giving rise to the neutral and charged biexciton resonances are shown in the left and right panels of (a), respectively. (b) A normalized 2D rephasing spectrum (magnitude) for the cross-circular polarization sequence. Neutral (XX) and charged (XT) biexcitons appear red-shifted along the emission energy axis from the exciton and trion emission energies by their corresponding binding energies Δ_{XX} and Δ_{XT} , respectively. (c). A slice along the emission energy axis at the excitation energy of the exciton showing the exciton, neutral biexciton, and charged biexciton. The biexciton binding energies are obtained from Gaussian fit functions.

Intervalley Biexcitons

The strong interactions responsible for tightly bound excitons lead to enhanced interactions between them. Higher-order correlated states, such as a bound biexciton resulting from four-particle correlations between two electrons and two holes, are expected to be stable in TMDs. These states are accessible with one-dimensional techniques such as photoluminescence and pump probe spectroscopy; however, signatures of biexcitons are difficult to distinguish from other phenomena, such as excitons localized by impurities or defects, exciton–electron interactions, and exciton renormalization, which can all spectrally overlap in one-dimensional spectra and exhibit a similar optical response.

Optical 2DCS is particularly sensitive to exciton many-body interactions, since congested one-dimensional spectra are unraveled onto two frequency axes. Signatures of biexcitons appear in the 2D rephasing spectrum of monolayer MoSe₂ (Fig. 12) when using a cross-circular polarization scheme, i.e. fields $\mathcal{E}_1^*$ and $\mathcal{E}_3$ ($\mathcal{E}_2$ and $\mathcal{E}_s$) are left (right) circularly polarized. Compared to the co-circular spectra shown in Fig. 10, the new peaks XX and XT are ascribed to the neutral and charged biexciton, respectively. For example, the quantum pathway associated with the biexciton peak XX can be interpreted from a four-level diamond system with ground state $|g\rangle$, singly excited states $|K\rangle$ and $|K'\rangle$, and doubly excited state $|XX\rangle$: field $\mathcal{E}_1^*$ drives a coherence $\rho_{gk}=|g\rangle\langle K|$ between the ground and K-valley exciton states during τ_1 ; field $\mathcal{E}_2$ drives the system into an excited-state $\rho_{K'K}=|K'\rangle\langle K|$; field $\mathcal{E}_3$ generates a third-order coherence $\rho_{XX,K}=|XX\rangle\langle K|$ during τ_3 , which radiates as the FWM signal. The system evolves during τ_1 and τ_3 at energies $\hbar\omega_X$ and $\hbar\omega_X - \Delta_{XX}$, where Δ_{XX} is the biexciton binding energy. In the Fourier domain, this pathway appears as peak XX; a similar pathway describes the bound charged biexciton state with binding energy Δ_{XT} .

A horizontal slice from the spectrum along the emission energy axis at the exciton excitation energy is shown in Fig. 12. The data are fit with a triple Gaussian function, which provides approximate binding energies of $\Delta_{XX}=20$ meV and $\Delta_{XT}=5$ meV. These values are consistent with current predictions from microscopic calculations. It is worth noting that information about these many-body states are masked in a 1D pump-probe spectrum by the trion resonance (projection of the 2D spectrum onto the emission axis); thus, while all the details of incoherent interactions in the sample are accessible with pump-probe spectroscopy, numerous quantum pathways spectrally overlap making analysis and interpretation difficult.

Summary

This article provides an overview of the basic principles of two-dimensional coherent spectroscopy and its implementation to study monolayer TMDs. In just the last couple of years, 2DCS has enhanced our understanding of the optical and electronic properties of 2D semiconductors, including: (1) the intrinsic homogeneous linewidth is on the order of 1 meV and is lifetime-limited for the exciton, whereas significant pure dephasing exists for the trion; (2) the exciton, trion, and background free carriers in doped TMDs form a coherently coupled system; (3) energy transfer between states is an important non-radiative relaxation channel; (4) high-order correlated states are stable in the monolayer TMD MoSe₂ and only exist between excitons/trions in opposite valleys. Compared to other semiconductor systems, TMDs offer novel opportunities for exploring quantum phenomena including condensation of exciton–polaritons, ultrathin biexciton lasers, and polarization-entangled photon sources.

The field of two-dimensional materials is extremely active with new insights reported almost daily. The ability to isolate and stack different 2D materials with precision, transfer them to nanopatterned substrates, and incorporate them with photonic microcavities and plasmonic nanostructures provide unprecedented possibilities for tailoring their physical properties for novel opto-electronic, photonic, and coherent valley/spintronic applications. Optical 2DCS will likely play a role in understanding and characterizing TMD-based heterostructures and nanophotonics in the same way that it has impacted our understanding of optical phenomena in monolayers.

See also: Two-Dimensional Electronic Spectroscopy

Further Reading

- Berkelbach, T.C., Hybertsen, H.S., Reichman, D.R., 2013. Theory of neutral and charged excitons in monolayer transition metal dichalcogenides. *Physical Review B* 88, 045318.
- Cundiff, S.T., Mukamel, S., 2013. Optical multidimensional coherent spectroscopy. *Physics Today* 66, 44.
- Czech, K.J., Thompson, B.J., Kain, S., *et al.*, 2015. Measurement of ultrafast excitonic dynamics of few-layer MoS₂ using state-selective coherent multidimensional spectroscopy. *ACS Nano* 9, 12146–12157.
- Dey, P., Paul, J., Wang, Z., *et al.*, 2016. Optical coherence in atomic-monolayer transition-metal dichalcogenides limited by electron–phonon interactions. *Physical Review Letters* 116, 127402.
- Ernst, R.R., Bodenhausen, G., Wokaun, A., 1987. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford: Oxford University Press.
- Hao, K., Xu, L., Nagler, P., *et al.*, 2016. Coherent and incoherent coupling dynamics between neutral and charged excitons in monolayer MoSe₂. *Nano Letters* 16, 5109–5113.
- Hao, K., Xu, L., Specht, J.F., *et al.*, 2017. Neutral and charged intervalley-biexcitons in monolayer MoSe₂. *Nature Communications* 8, 15552.
- Jakubczyk, T., Delmonte, V., Koperski, M., *et al.*, 2016. Radiatively limited dephasing and exciton dynamics in MoSe₂ monolayers revealed with four-wave mixing microscopy. *Nano Letters* 16, 5333–5339.
- Moody, G., Dass, C.K., Hao, K., *et al.*, 2015. Intrinsic homogeneous linewidth and broadening mechanisms of excitons in monolayer transition metal dichalcogenides. *Nature Communications* 6, 8315.
- Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. *Annual Review of Physical Chemistry* 51, 691–729.
- Xu, X., Yao, W., Xiao, X., Heinz, T.F., 2014. Spins and pseudospins in layered transition metal dichalcogenides. *Nature Physics* 10, 343–350.

Excitons in Magnetic Fields

Kankan Cong, G Timothy Noe II, and Junichiro Kono, Rice University, Houston, TX, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

When a photon of energy greater than the band gap is absorbed by a semiconductor, a negatively charged electron is excited from the valence band into the conduction band, leaving behind a positively charged hole. The electron can be attracted to the hole via the Coulomb interaction, lowering the energy of the electron-hole (e - h) pair by a characteristic binding energy, E_b . The bound e - h pair is referred to as an exciton, and it is analogous to the hydrogen atom, but with a larger Bohr radius and a smaller binding energy, ranging from 1 to 100 meV, due to the small reduced mass of the exciton and screening of the Coulomb interaction by the dielectric environment.

Like the hydrogen atom, there exists a series of excitonic bound states, which modify the near-band-edge optical response of semiconductors, especially when the binding energy is greater than the thermal energy and any relevant scattering rates. When the e - h pair has an energy greater than the binding energy, the electron and hole are no longer bound to one another (ionized), although they are still correlated. The nature of the optical transitions for both excitons and unbound e - h pairs depends on the dimensionality of the e - h system. Furthermore, an exciton is a composite boson having integer spin that obeys Bose-Einstein statistics rather than fermions that obey Fermi-Dirac statistics as in the case of either the electrons or holes by themselves. One of the most interesting consequences of the bosonic nature of excitons is the possibility of having many that occupy the same quantum state as opposed to fermions, which obey the Pauli exclusion principle.

An applied magnetic field, $\vec{B}$, will quantize the energies and angular momenta of charged particles and thus change the electronic state of an exciton. In general, the excitonic wavefunction shrinks in the direction perpendicular to $\vec{B}$, and the exciton binding energy increases with $B = |\vec{B}|$. How the energy of each exciton state varies with B depends on basic exciton parameters, and, therefore, the study of excitons in B provides a tool to characterize various exciton properties, such as the reduced mass, binding energy, Bohr radius, and g -factor. Importantly, the nature of the dependence of exciton energy on B depends on the strength of B . For example, at low B , where the effect of the Coulomb interaction is larger than that of B and thus B is treated as a perturbation, excitons retain a hydrogenic nature: diamagnetic shifts and Zeeman splittings modify the energies of excitonic states. Alternatively, in the high B limit where the effect of B dominates and the Coulomb interaction is treated as a perturbation, the energies of exciton states show Landau-level-like, linear- B dependence, similar to free e - h pairs.

For the hydrogen atom, the binding energy, 13.6 eV, is much higher than those for excitons in semiconductors, and an extremely high B (over 2×10^5 T) is required to enter the high B regime. The much smaller binding energies of excitons in semiconductors allow one to enter the high- B regime readily, which, together with the bosonic nature of excitons, makes excitons in high B attractive for the investigation of many-body phenomena in condensed matter in a highly controllable manner. A large body of theoretical work has thus been devoted in the last several decades to e - h systems in strong B , and a fascinating array of possible physical phenomena/states has been predicted – the excitonic insulator phase, a gas-liquid-type phase transition, Bose-Einstein condensation of magnetoexcitons, and quantum chaos. Recently, attention has been paid to spatially separated two-dimensional (2D) electrons and holes in layered structures, since the pioneering work of Lozovik and co-workers, who considered the pairing of electrons and holes across an interface of two-media, and the superfluidity of such pairs. A very interesting but complicated many-particle situation arises, in the presence of strong perpendicular B . In such a situation, spatially separated 2D electrons and holes, both Landau quantized, are present, and there is a Coulomb attraction between them as well as electron-electron (e - e) and hole-hole (h - h) correlations. However, all these interactions are expected to cancel each other in the high- B limit, where the “hidden symmetry” of 2D magnetoexcitons protects them from ionization, leading to the absence of an excitonic Mott transition. Furthermore, excitonic gain enhancement occurs in a high-density e - h system in high B , which leads to cooperative and coherent light emission (known as superfluorescence) from the highest-occupied magnetoexciton state.

Here, we will review various observations related to excitons in magnetic fields. First, we present the theory of excitons in magnetic fields and show how the exciton energy's magnetic field dependence relates to fundamental exciton parameters. We will explore how different dimensionalities of excitons and the strength of magnetic fields affect the exciton properties. Then, interband and intraband magneto-optical absorption observations are discussed in different semiconductor systems. Finally, some exotic phenomena unique to high-density excitons in high magnetic fields are presented.

Exciton States in Magnetic Fields

In this section, we describe the states of excitons in semiconductors in different dimensions and B regimes. We introduce the basic Hamiltonians of excitons, discuss the eigenenergies in the low- B and high- B limits, and define some basic parameters of excitons and various quantities to be used throughout the article.

For an electron with effective mass m_e^* located at $\vec{r}_e$ and a hole with effective mass m_h^* located at $\vec{r}_h$, interacting through the Coulomb potential $U(\vec{r}) = \frac{e^2}{4\pi\epsilon r}$, the Hamiltonian at zero B is given by

$$\hat{H} = -\frac{\hbar^2}{2M^*} \nabla_R^2 - \frac{\hbar^2}{2\mu^*} \nabla_r^2 + U(\vec{r}) \quad (1)$$

where $M^* = m_e^* + m_h^*$ is the total mass, $\mu^* = ((m_e^*)^{-1} + (m_h^*)^{-1})^{-1}$ is the reduced mass, the center-of-mass coordinate $\vec{R} = (m_e^* \vec{r}_e + m_h^* \vec{r}_h)/M^*$, the relative coordinate $\vec{r} = \vec{r}_e - \vec{r}_h$, ∇_R^2 (∇_r^2) is the Laplacian expressed in the center-of-mass (relative) coordinates, $\hbar$ is the reduced Planck constant, e is the electronic charge, $\epsilon = \epsilon_r \epsilon_0$ is the permittivity, ϵ_r is the relative permittivity (dielectric constant), and ϵ_0 is the vacuum permittivity.

Three-Dimensional Case

In a three-dimensional (3D) system at zero B , the Schrödinger equation for the relative motion is given by

$$\left(-\frac{\hbar^2}{2\mu^*} \nabla_r^2 - \frac{e^2}{4\pi\epsilon r} \right) \psi(\vec{r}) = E \psi(\vec{r}) \quad (2)$$

where $\Psi(\vec{r})$ is the exciton wavefunction and E is the energy of the exciton. In spherical coordinates, the wavefunction can be expressed as $\Psi(r, \theta, \phi) = R_{nl}(r) Y_{lm}(\theta, \phi)$, where $R_{nl}(r)$ and $Y_{lm}(\theta, \phi)$ are the radial and angular wavefunctions, respectively, n ($= 1, 2, 3, \dots$) is the principal quantum number, l ($= 0, 1, 2, \dots, n-1$) is the azimuthal quantum number, and m ($= 0, \pm 1, \dots, \pm l$) is the magnetic quantum number. The eigenenergies are

$$E_n = -\frac{R_y^*}{n^2}, \quad n = 1, 2, 3, \dots \quad (3)$$

where R_y^* is the effective Rydberg constant defined as

$$R_y^* = \frac{\mu^* e^4}{32\pi^2 \epsilon^2 \hbar^2} = \frac{\mu^* / m_e}{\epsilon_r^2} \times 13.6 \text{ eV} \quad (4)$$

where $m_e = 9.11 \times 10^{-31} \text{ kg}$ is the mass of a free electron in vacuum. Therefore, the binding (ionization) energy of a 3D exciton is $E_{b,3D}^* = E_\infty - E_1 = R_y^*$. Using the effective Bohr radius

$$a_{B,3D}^* = \frac{4\pi\epsilon\hbar^2}{\mu^* e^2} = \frac{\epsilon_r}{\mu^* / m_e} \times 0.0529 \text{ nm} \quad (5)$$

the binding energy can be expressed as

$$E_{b,3D}^* = \frac{\hbar^2}{2\mu^*} \frac{1}{(a_{B,3D}^*)^2} \quad (6)$$

and the 1s state wavefunction is given by

$$\Psi_{1s,3D} = \Psi_{100} = \frac{1}{\sqrt{\pi}} \frac{1}{(a_{B,3D}^*)^{3/2}} \exp\left(\frac{-r}{a_{B,3D}^*}\right) \quad (7)$$

Now that we have defined the basic quantities for excitons, let us consider the effect of a uniform and constant magnetic field, $\vec{B} = (0, 0, B)$. The excitonic eigenenergies and wavefunctions are modified through the inclusion of vector potential $\vec{A}$ in the Hamiltonian, for which $\vec{B} = \nabla \times \vec{A}$. The Hamiltonian of the system becomes (neglecting electron spin)

$$\begin{aligned} \hat{H} &= \frac{1}{2m_e^*} \left(\vec{p}_e + e\vec{A}(\vec{r}_e) \right)^2 + \frac{1}{2m_h^*} \left(\vec{p}_h - e\vec{A}(\vec{r}_h) \right)^2 - \frac{e^2}{4\pi\epsilon r} \\ &= -\frac{\hbar^2}{2m_e^*} \nabla_e^2 - \frac{\hbar^2}{2m_h^*} \nabla_h^2 - \frac{e^2}{4\pi\epsilon r} - \frac{ie\hbar}{m_e^*} \vec{A}(\vec{r}_e) \cdot \vec{\nabla}_e \\ &\quad + \frac{ie\hbar}{m_h^*} \vec{A}(\vec{r}_h) \cdot \vec{\nabla}_h + \frac{e^2}{2m_e^*} A^2(\vec{r}_e) + \frac{e^2}{2m_h^*} A^2(\vec{r}_h) \end{aligned} \quad (8)$$

where $\vec{p}_e$ and $\vec{p}_h$ are the momentum operators of the electron and hole, respectively.

Following Gor'kov and Dzyaloshinskii, we use a canonical transformation to obtain a new wavefunction $\Psi'(\vec{R}, \vec{r})$

$$\Psi'(\vec{R}, \vec{r}) = \exp\left(i\left[\vec{K} - \frac{e}{\hbar} \vec{A}(\vec{r})\right] \cdot \vec{R}\right) F(\vec{r}) \quad (9)$$

where $F(\vec{r})$ is an arbitrary function of $\vec{r}$, and $\vec{K} = \vec{k}_e - \vec{k}_h$. After the transformation, the Schrödinger equation becomes

$$\hat{H}' F(\vec{r}) = \left[E - \frac{\hbar^2 K^2}{2M^*} \right] F(\vec{r}) \quad (10)$$

with the new Hamiltonian expressed as

$$\begin{aligned}\hat{H}' &= \left(-\frac{\hbar^2}{2\mu^*} \nabla_r^2 - \frac{e^2}{4\pi\epsilon r} \right) - ie\hbar \left(\frac{1}{m_e^*} - \frac{1}{m_h^*} \right) \vec{A}(\vec{r}) \cdot \vec{\nabla}_r \\ &\quad + \frac{e^2}{2\mu^*} A^2(\vec{r}) - \frac{2e\hbar}{M^*} \vec{A}(\vec{r}) \cdot \vec{K} \\ &= \hat{H}_{B=0} + \frac{\omega_{c,e} - \omega_{c,h}}{2} \hat{L}_z + \frac{e^2 B^2}{8\mu^*} (x^2 + y^2) \\ &\quad - \frac{eB\hbar}{M^*} (xK_y - yK_x)\end{aligned}\quad (11)$$

$$- \frac{eB\hbar}{M^*} (xK_y - yK_x) \quad (12)$$

where $\omega_{c,i} = eB/m_i^*$ ($i=e,h$) is the cyclotron frequency, $\hat{L}_z = -i\hbar(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x})$ is the z -component of the orbital angular momentum operator, and the symmetric gauge, $\vec{A} = (1/2)(-By, Bx, 0)$, was used in the final step. The first term of $\hat{H}'$ is the Hamiltonian at zero B . The second term is referred to as the (orbital) Zeeman term, which leads to shifts and splittings of exciton states, depending on the value and degeneracy of the orbital angular momentum of a given state. The third term is the Langevin diamagnetism term, which is proportional to the spatial extent of the exciton wavefunction in a plane perpendicular to $\vec{B}$ and increases quadratically with increasing B . The fourth term can be rewritten as $-e(\vec{V}_R \times \vec{B}) \cdot \vec{r}$, where $\vec{V}_R$ is the center-of-mass velocity. Since $\vec{r}$ is the relative coordinate, the fourth term implies that the relative and center-of-mass degrees of freedom are coupled in the presence of a B .

It should be pointed out that one cannot obtain an exact solution to the Schrödinger equation corresponding to Eq. (11), which is related to quantum chaos. This is due to the fact that the number of conserved quantities (two), i.e., E and L_z , is smaller than the number of degrees of freedom (three), a situation known to be *non-integrable*, for which the classical trajectories of energy states in phase space are chaotic. Correspondingly, energy levels exhibit fluctuations; with increasing energy, the statistics of levels evolve from Poissonian to Gaussian. Approximate solutions can be obtained only for low-energy levels either in the low- B limit or the high- B limit, and a one-to-one correspondence of states between these two limits does not exist, in general, because of the different symmetries the magnetic field and the Coulomb potential possess: the cylindrical symmetry (for the former) and the spherical symmetry (for the latter).

To define the low- B and high- B limits, we introduce a dimensionless parameter

$$\gamma \doteq \frac{\hbar\omega_c}{2}/R_\gamma^* = \left(\frac{a_B^*}{l_B} \right)^2 = \frac{16\pi^2 \hbar^3 e^2}{\mu^{*2} e^3} B \quad (13)$$

where $l_B = \sqrt{\hbar/eB}$ is the magnetic length and $\omega_c = eB/\mu^*$ is the reduced cyclotron frequency. The value of γ is a measure of whether the energy of the exciton is dominated by the influence of the magnetic field or the Coulomb interaction. In the low- B limit ($\gamma \ll 1$), the system is nearly hydrogenic, and the Zeeman and diamagnetic terms can be treated as a perturbation in Eq. (11). (Note that the fourth term in $\hat{H}'$ is usually negligible because of the smallness of K relevant to optical transitions.) For the $1s$ [or $(n,l,m)=(100)$] state, $\langle L_z \rangle = m\hbar = 0$, and therefore, the Zeeman term is zero, and the first-order correction is the diamagnetic term:

$$\begin{aligned}\Delta E_{1s,3D} &= \left\langle \frac{e^2 B^2}{8\mu^*} (x^2 + y^2) \right\rangle_{1s,3D} \\ &= \frac{e^2 B^2}{8\mu^*} \langle \Psi_{1s,3D} | (x^2 + y^2) | \Psi_{1s,3D} \rangle \\ &= \frac{e^2 B^2}{8\mu^*} \cdot 2 \left(a_{B,3D}^* \right)^2 = \sigma_{3D}^* B^2\end{aligned}\quad (14)$$

$$\sigma_{3D}^* \doteq \frac{e^2}{4\mu^*} \left(a_{B,3D}^* \right)^2 = \frac{4\pi^2 e^2 \hbar^4}{e^2 (\mu^*)^3} \quad (15)$$

Here, σ_{3D}^* is the 3D diamagnetic-shift coefficient, which depends on μ^* and $a_{B,3D}^*$. This fact offers a straightforward experimental method for determining $a_{B,3D}^*$ (or, more generally, the spatial extent of the exciton wavefunction) once the reduced mass is known. In addition, measuring σ_{3D}^* as the direction of $\vec{B}$ is varied allows one to map out the wavefunction anisotropy. Similarly, the diamagnetic shift for the ns exciton state can be calculated to be

$$\Delta E_{ns,3D} = \frac{n^2(5n^2 + 1)}{6} \sigma_{3D}^* B^2 \quad (16)$$

For calculating the energy for a p -like state, one must include the Zeeman term. For example, the $2p$ state splits into three states in a B with $\langle L_z \rangle = 0, \pm 1$, so the orbital Zeeman energy is

$$\Delta E_{\text{Zeeman}} = 0, \pm \frac{\hbar\omega_c}{2} \quad (17)$$

In the high- B limit ($\gamma \gg 1$), where the effect of the magnetic field is much larger than the Coulomb interaction, the Coulomb interaction is treated as a small perturbation. Under this condition, the energies and wavefunctions of excitons can be separated into different components for the x - y plane and the z -direction (known as the 'adiabatic' approximation). In the x - y plane, only the effect of the magnetic field has to be considered, while in the z -direction, only the Coulomb interaction contributes, because a magnetic field applied in the z -direction exerts a Lorentz force only in the x - y plane. The wavefunction can then be expressed as

$$\Psi_{NMi} = \Phi_{NM}(x, y) f_{NMi}(z) \quad (18)$$

where $N=0, 1, 2, \dots$ is the Landau quantum number, $M=N, N-1, \dots, -\infty$ is the azimuthal quantum number, and $i=0, 1, 2, \dots$ is the quantum number associated with the motion in the z -direction. The wavefunction $\Phi_{NM}(x, y)$ describes the in-plane motion and is proportional to the associated Laguerre function $L_{N+|M|}^{(|M|)}$.

The total energy is characterized by three quantum numbers (N, M, v^+) for $i=2v$ ($v=0, 1, 2, \dots$), and (N, M, v^-) for $i=2v-1$ ($v=1, 2, \dots$):

$$E_{NM(v)} = \left(N + \frac{1}{2}\right) \hbar \omega_c + \Delta E_{NMv}(z) \quad (19)$$

Hence, in the high- B limit, the states become more Landau-level-like, exhibiting a linear B dependence (the first term) with the Coulombic correction, $\Delta E_{NMv}(z)$, which is a red-shift for each eigenenergy at a fixed value of z . The wavefunction $f_{NMv^+}(z)$ is an even function, while $f_{NMv^-}(z)$ is an odd function with respect to z .

Because of the reason stated earlier, there is no general correspondence between the low-field notation (n, l, m) and high-field notation (N, M, v) for a 3D exciton. However, using the principle of conservation of the number of nodal surfaces, Shinada *et al.* made the following correspondence for the lowest-lying exciton states:

$$\begin{aligned} 1s &\leftrightarrow (0, 0, 0^+) \\ 2s &\leftrightarrow (0, 0, 1^+) \\ 2p_0 &\leftrightarrow (0, 0, 1^-) \\ 2p_+ &\leftrightarrow (1, 1, 0^+) \\ 2p_- &\leftrightarrow (0, -1, 0^+) \\ 3d_0 &\leftrightarrow (1, 0, 0^+) \end{aligned} \quad (20)$$

Two-Dimensional Case

For 2D excitons, we use a polar coordinate system,

$$\vec{\rho} = (x, y) = (\rho \cos \phi, \rho \sin \phi) \quad (21)$$

At zero B , the Hamiltonian for the relative motion for a 2D e - h pair is written as

$$\begin{aligned} \hat{H} &= -\frac{\hbar^2}{2\mu^*} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) - \frac{e^2}{4\pi\epsilon \sqrt{x^2 + y^2}} \\ &= -\frac{\hbar^2}{2\mu^*} \left(\frac{\partial^2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} \right) - \frac{e^2}{4\pi\epsilon \rho} \end{aligned} \quad (22)$$

By solving the Schrödinger equation

$$\hat{H}\Psi(\rho, \phi) = E\Psi(\rho, \phi) \quad (23)$$

we obtain the eigenenergies as

$$E_n = \frac{\mu^* e^4}{32\pi^2 \epsilon^2 \hbar^2 (n - \frac{1}{2})^2} = -\frac{R_y^*}{(n - \frac{1}{2})^2}, \quad n = 1, 2, 3, \dots \quad (24)$$

The binding energy is then calculated to be

$$E_{b,2D}^* = E_\infty - E_1 = 4R_y^* = 4E_{b,3D}^* \quad (25)$$

which is four times larger than that of the 3D case for the same semiconductor. The eigenstates have the form

$$\Psi_{nm} = R_{nm}(\rho) e^{im\phi}$$

where $n (= 1, 2, 3, \dots)$ is the principal quantum number, and $m (= -n+1, \dots, -2, -1, 0, 1, 2, \dots, n-1 = \dots, d_-, p_-, s, p_+, d_+, \dots)$ is the angular momentum quantum number. The $1s$ exciton wavefunction is written as

$$\Psi_{1s,2D} = \Psi_{10} = R_{10} = \frac{1}{\sqrt{2\pi}} \frac{2}{a_{B,2D}^*} \exp\left(\frac{-\rho}{a_{B,2D}^*}\right) \quad (26)$$

with the effective Bohr radius,

$$a_{B,2D}^* = \frac{2\pi\epsilon\hbar^2}{\mu^*e^2} = \frac{a_{B,3D}^*}{2} \quad (27)$$

which is only half of that of a 3D exciton in the same semiconductor material. From Eqs. (4), (25) and (27), we have

$$E_{b,2D}^* = \frac{\hbar^2}{2\mu^*} \frac{1}{(a_{B,2D}^*)^2} \quad (28)$$

In a similar manner to the 3D case, the Hamiltonian of the 2D system at a finite B after the Gor'kov-Dzyaloshinskii transformation can be written as

$$\begin{aligned} \hat{H}' = & \left(-\frac{\hbar^2}{2\mu^*} \nabla_\rho^2 - \frac{e^2}{4\pi\epsilon\rho} \right) - ie\hbar \left(\frac{1}{m_e^*} - \frac{1}{m_h^*} \right) \vec{A}(\vec{\rho}) \cdot \vec{\nabla}_\rho \\ & + \frac{e^2}{2\mu^*} A^2(\vec{\rho}) - \frac{2e\hbar}{M} \vec{A}(\vec{\rho}) \cdot \vec{K} \end{aligned} \quad (29)$$

$$\begin{aligned} = & \hat{H}_{B=0} + \frac{\omega_{c,e} - \omega_{c,h}}{2} \hat{L}_z + \frac{e^2 B^2}{8\mu^*} (x^2 + y^2) \\ & - \frac{eB\hbar}{M} (xK_y - yK_x) \end{aligned} \quad (30)$$

where the symmetric gauge is used. Also, similar to the 3D case, the diamagnetic shift for the $1s$ state can be calculated to be

$$\begin{aligned} \Delta E_{1s,2D} &= \frac{e^2 B^2}{8\mu^*} \langle \rho^2 \rangle_{1s,2D} = \frac{e^2 B^2}{8\mu^*} \frac{3}{8} (a_{B,3D}^*)^2 \\ &= \sigma_{2D}^* B^2 \end{aligned} \quad (31)$$

$$\begin{aligned} \sigma_{2D}^* &= \frac{3e^2}{64\mu^*} (a_{B,3D}^*)^2 = \frac{3e^2}{16\mu^*} (a_{B,2D}^*)^2 \\ &= \frac{3\pi^2 \epsilon^3 \hbar^4}{4e^2 (\mu^*)^3} = \frac{3}{16} \sigma_{3D}^* \end{aligned} \quad (32)$$

$$a_{B,2D}^* = \sqrt{\langle \rho^2 \rangle_{1s,2D}} = \sqrt{8\mu^* \sigma_{2D}^*} / e \quad (33)$$

Here, σ_{2D}^* is the 2D diamagnetic-shift coefficient, and $a_{B,2D}^*$ is equal to the root mean squared (r.m.s.) radius of the $1s$ exciton. So, the diamagnetic-shift coefficient of a 2D exciton is much smaller than the corresponding 3D excitons in the same semiconductor material, i.e., a larger Coulomb interaction makes the wavefunction more compact, leading to a smaller diamagnetic shift.

In a 2D system, both the B and the Coulomb energy have circular symmetry, and thus, there is a straightforward, one-to-one correspondence between the low- B notation (n,m) and the high- B notation (N,M) , as summarized in Table 1. The correspondence rules are:

$$m = N - M \quad (34)$$

$$n = N + 1 \text{ (for } N \geq M) \quad (35)$$

$$n = M + 1 \text{ (for } M \geq N) \quad (36)$$

Here, n and m are the principle and angular momentum quantum numbers, respectively, in the low- B hydrogenic notation, while M and N are the Landau indices of the electron and hole, respectively, in the high- B Landau-level notation. For example, if $N=0$ and $M=0$, we get $n=1$ and $m=0$, i.e., $(N,M)=(0,0)$ corresponds to the $1s$ state. Also, if $N=3$ and $M=2$, we obtain $n=4$ and $m=+1$, and so $(N,M)=(3,2)$ corresponds to the $4p^+$ state.

Table 1 Correspondence between the low- and high- B labels $(n,m) \leftrightarrow (N,M)$ for 2D magnetoexcitons, where $m=N-M$ and $n=N+1$ for $N \geq M$ and $n=M+1$ for $M \geq N$

M	0	1	2	3
N				
0	$1s$	$2p^-$	$3d^-$	$4f^-$
1	$2p^+$	$2s$	$3p^-$	$4d^-$
2	$3d^+$	$3p^+$	$3s$	$4p^-$
3	$4f^+$	$4d^+$	$4p^+$	$4s$

Source: Reproduced (adapted) with permission from MacDonald, A.H., Ritchie, D.S., 1986. Hydrogenic energy levels in two dimensions at arbitrary magnetic fields. Physical Review B 33, 8336–8344. Copyright 1986 by the American Physical Society.

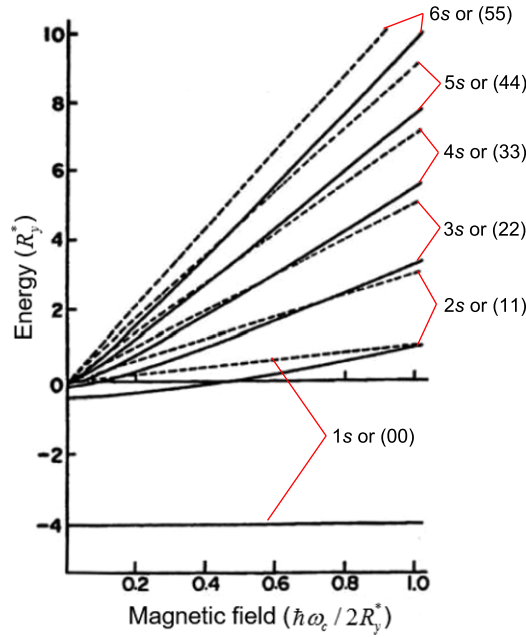


Fig. 1 (color online). Calculated eigenenergies of 2D excitons as a function of dimensionless magnetic field, $\gamma = \hbar\omega_c/2Ry^*$, with (solid lines) and without (dashed lines) the Coulomb interaction. Reproduced (adapted) with permission from Akimoto, O., Hasegawa, H., 1967. Interband optical transitions in extremely anisotropic semiconductors. II. Coexistence of exciton and the Landau levels. Journal of the Physical Society of Japan 22, 181–191. Copyright 1967 by the Physical Society of Japan.

Fig. 1 plots the eigenenergies of the lowest six 2D exciton states as a function of dimensionless magnetic field γ , calculated by Akimoto and Hasegawa. The solid and dashed lines represent the energies of the exciton states with and without including the Coulomb interaction in the Hamiltonian, respectively. At low B , the excitonic feature dominates the low-index energy states. The high-index energy states and the states at high fields show a more Landau-level-like behavior, more appropriate for free electrons and holes. The lowest-energy state or the 1s [or the (00)] state remains constant with increasing γ up to ~ 1 , but in the higher γ range it curves upward to cross the $E=0$ line and eventually runs parallel to the linear Landau level line at $\gamma \sim 12.5$.

Interband Magneto-optical Absorption

Interband magneto-optical absorption can provide direct information on the energy levels of optically accessible (or ‘bright’) exciton states: ns states (1s, 2s, 3s, ...) in the low- B notation, or $(N,M)=(00), (11), (22), \dots$ states in the high- B notation. To be clear, these are simply notations, and there is a one-to-one correspondence between the two (see **Table 1**). In the low- B limit, the 1s energy state shows a hydrogenic character with the energy quadratically dependent on B . The effective diamagnetic-shift coefficient, σ^* , can be determined from the B^2 fitting of the exciton energy. If the reduced mass of excitons is known, the Bohr radius can be obtained from **Eq. (33)**. In addition, the binding energy can be determined using **Eq. (28)**, assuming that a hydrogenic model applies to the semiconductor under study. The reduced effective mass can be obtained from the cyclotron frequency $\omega_c = eB/\mu^*$, which is deduced from the slope versus B of exciton energies in the high- B limit. Furthermore, by extrapolating the B -dependent energy lines of excited exciton states in the high- B limit to the zero- B value, it is found that they nearly converge to a single point, which gives the band gap. The difference between the band gap and the 1s exciton peak at zero B then gives the exciton binding energy.

Quantum Wells

The first magneto-optical absorption experiment in a semiconductor quantum well system was performed by Tarucha *et al.* in 1984. They used undoped GaAs/AlAs multiple-quantum wells grown on a (100) GaAs substrate by molecular beam epitaxy. Optical absorption spectra were taken at 4.2 K in pulsed B up to 30 T applied perpendicular to the sample. **Fig. 2(a)** shows absorption spectra at different B . At 0 T, the absorption spectrum shows two distinct peaks, corresponding to the 1s heavy-hole ($E_{hh}(1s)$) and 1s light-hole ($E_{lh}(1s)$) excitons, respectively. With increasing B , the 1s exciton peaks were observed to blue-shift, and additional peaks emerged at higher energies, corresponding to the ns states (2s, 3s, ...) in the low- B notation, or $(N,M)=(11), (22), \dots$ states in the high- B notation. These higher-energy peaks blue-shifted more rapidly with increasing B .

Fig. 2(b) shows the photon energies of the absorption peaks as a function of B . The 1s heavy-hole and light-hole excitons are represented by solid and open circles, respectively, and the peaks of higher energy states are plotted with crosses. In the low- B

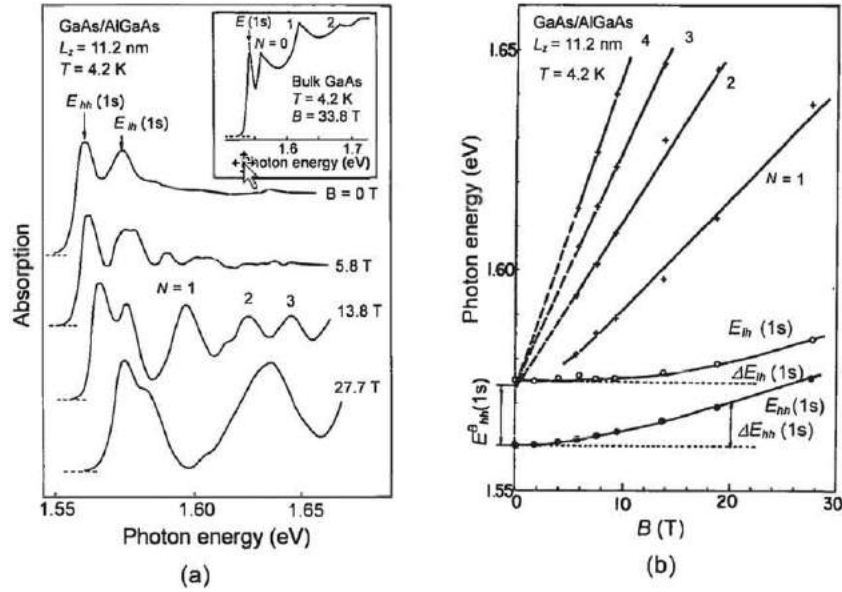


Fig. 2 (color online). (a) Optical absorption spectra at different magnetic fields for a GaAs/AlGaAs multiple-quantum-well sample. (b) Magnetic field dependence of the photon energies of absorption peaks. Reproduced (adapted) with permission from Tarucha, S., Okamoto, H., Iwasa, Y., Miura, N., 1984. Exciton binding energy in GaAs quantum wells deduced from magneto-optical absorption measurement. *Solid State Communications* 52, 815–819. Copyright 1984 by Elsevier.

region, the $1s$ exciton energy shows a quadratic field dependence, in agreement with the expected diamagnetic shift in energy for excitons. In addition, the change of the heavy-hole transition energy at a certain field $\Delta E_{hh}(1s)$ is observed to be larger than the change of the light-hole transition energy $\Delta E_{lh}(1s)$ at the same field, indicating a smaller reduced mass for the $1s$ heavy-hole excitons than that of the $1s$ light-hole excitons, since the diamagnetic shift is proportional to $1/\mu^*$; see Eq. (31). The difference in the reduced masses of excitons can be explained by the anisotropic nature of the kinetic energy expression in the diagonal term of the Luttinger-Kohn Hamiltonian. The diagonal term gives the larger reduced mass for the light-hole excitons associated with $J_z(\pm 3/2)$ bands than that for the heavy-hole excitons associated with $J_z(\pm 1/2)$ bands.

In contrast, the absorption peaks for higher-energy heavy-hole states show a linear B dependence. The effect of the Coulomb interaction can only be seen on the $2s$ state, since its energy shift shows a small curvature in the low- B region. By extrapolating the $3s$, $4s$, and $5s$ absorption peaks from the high- B region to the low- B region with straight lines, it was found that the lines converge at a photon energy within an error of less than 0.5 meV. The difference between the energy at the convergent point and the zero field value of $E_{hh}(1s)$ gives the binding energy of the $1s$ heavy-hole exciton, which was ~ 15 meV in this case. The exciton binding energy of the $1s$ heavy-hole exciton increases with decreasing well size, since the quasi-2D nature is a closer approximation for smaller quantum well sizes. On the contrary, the diamagnetic shift increases with increasing well size, indicating a larger Bohr radius in a higher-dimensionality system.

Transition Metal Dichalcogenide

Atomically thin transition-metal dichalcogenides (TMDs), a new class of 2D materials, have recently attracted much attention for magneto-optical studies due to their extremely large exciton binding energies. For example, Stier *et al.* performed low-temperature polarized reflection measurements on atomically thin WS_2 and MoS_2 in high magnetic fields up to 65 T, observing a strong Zeeman splitting and a small diamagnetic shift for the 2D excitons. Based on their measurements, some basic parameters of excitons in these TMDs, such as the g -factor, exciton size, and binding energy, were determined.

As shown in Fig. 3(a), 2D TMDs have a direct band gap located at the degenerate K and K' valleys of their hexagonal Brillouin zones. The strong spin-orbit coupling of the valence band lifts the degeneracy of the spin-up and spin-down components, leading to well-separated A and B excitons. Due to the valley-specific optical selection rules, the interband transitions in the K valley only couple to right circularly polarized light, σ^+ , for both the A and B excitons, but left circularly polarized light, σ^- , for the K' valley.

Fig. 3(b) plots the B -dependent energy shifts of the conduction and valence bands in the K valley (σ^+ polarized light) with $B \parallel \pm z$, and demonstrates different contributions to the Zeeman splitting. For a specific valley in a B , the Zeeman shift is $\Delta E_Z = -(\mu_c - \mu_v) \cdot B$. In addition, the time-reversal symmetry in TMDs leads to an equal-but-opposite total magnetic moment ($\mu_K^{c,v} = -\mu_{K'}^{c,v}$) in the K and K' valleys. However, this time-reversal symmetry will be broken by B , and the time-reversed pairs in K and K' valleys will shift to opposite directions, i.e., the Zeeman splitting for excitons is doubled considering the broken time-reversal symmetry. Generally, the total magnetic moment μ comes from three sources: spin (μ_s), atomic orbital (μ_l), and valley

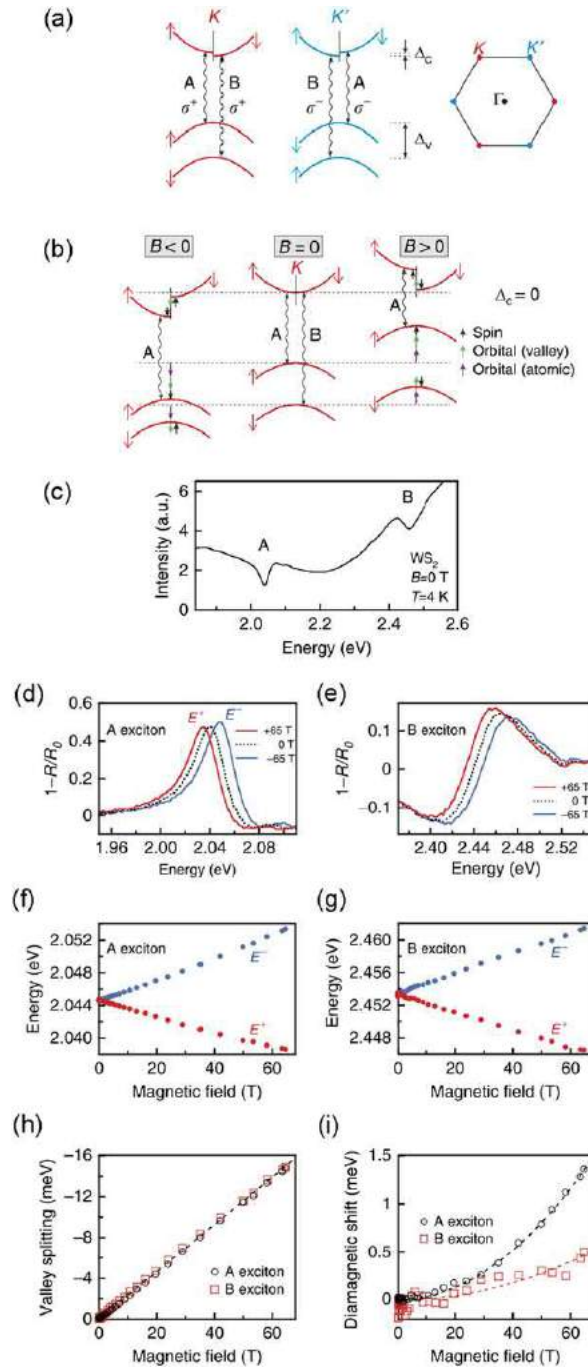


Fig. 3 (color online). (a) Excitonic transitions in monolayer transition metal dichalcogenides. (b) Zeeman shifts in different B fields. (c) Reflection spectra of monolayer WS_2 at $B=0$ T and $T=4$ K. (d) Normalized reflection spectra of the A exciton resonance in different B . (e) Same, but for the B exciton. (f) Energy of the field-split A exciton versus magnetic field. (g) Same, but for the B exciton. (h) The measured valley Zeeman splitting versus B , for both the A and B excitons. (i) The diamagnetic shifts for both the A and B excitons. Reproduced (adapted) with permission from Stier, A.V., McCreary, K.M., Jonker, B.T., Kono, J., Crooker, S., 2016. Exciton diamagnetic shifts and valley Zeeman effects in monolayer WS_2 and MoS_2 to 65 Tesla. *Nature Communications* 7, 10643. Copyright 2016 by Nature Publishing Group.

orbital (μ_k) (Berry curvature). Here, μ_s is zero, since the optically allowed transitions couple the conduction and valence bands with the same spin. μ_l is non-zero, since the conduction and valence bands have different atomic orbitals. The d_{z^2} orbitals of the conduction band have $L_z=0$ ($\mu_l^c=0$), while the hybridized $d_{x^2-y^2} \pm id_{xy}$ orbitals for the valence bands have $L_z=\pm 2\hbar$ ($\mu_l^v=\pm 2\mu_B$) in the K and K' valleys, generating a Zeeman shift of $\mp 2\mu_B B$ for the K and K' excitons, and leading to a total exciton splitting of $-4\mu_B B$. For the valley orbital contribution to the conduction and valence bands, $\mu_k^c=\pm(m_e/m_e^*)\mu_B$ and $\mu_k^v=\pm(m_e/m_h^*)\mu_B$ in the K and K' valleys. Assuming a simple two-band tight-binding model where $m_e^*=m_h^*$, it is easy to see that $\mu_k^c=\mu_k^v$, i.e., the

contribution of the valley orbital to the Zeeman splitting is zero. In total, the Zeeman splitting for excitons is expected to be $\Delta E_Z = -4\mu_B B$ assuming $m_e^* = m_h^*$.

Fig. 3(c) shows the reflection spectrum of monolayer WS_2 at 4 K and 0 T, from which the A and B excitonic transitions can be seen. The normalized reflection spectra at 0 T and ± 65 T of the A exciton are shown in **Fig. 3(d)**, indicating a clear Zeeman splitting of ~ 15 meV. The valley splitting of the B exciton can also be observed, as shown in **Fig. 3(e)**. For both excitons, the exciton transition energy in a positive B (called E^+) shifts to a lower energy, while a shift to a higher energy is observed for the exciton transition energy in a negative B (called E^-). **Fig. 3(f)** and **(g)** show the transition energies, E^+ and E^- , as a function of B , for the A and B excitons, respectively. The splittings between the two valleys, $E^+ - E^-$, are shown in **Fig. 3(h)** for both the A and B excitons. The measured Zeeman splitting energies are negative but increase linearly with B with almost identical rates of $-228 \pm 2 \mu\text{eV T}^{-1}$ for the A exciton and $-231 \pm 2 \mu\text{eV T}^{-1}$ for the B exciton. Such rates correspond to g -factors of -3.94 ± 0.04 and -3.99 ± 0.04 , respectively, very close to the expected value (-4).

Fig. 3(i) shows the diamagnetic shifts for the A and B excitons in monolayer WS_2 . As we saw in Eq. (33), $a_{B,2D}^*$ can be calculated if μ^* and σ_{2D}^* are known.

Through fits to the data, $\sigma_A^* = 0.32 \pm 0.02 \mu\text{eV T}^{-2}$ and $\sigma_B^* = 0.11 \pm 0.02 \mu\text{eV T}^{-2}$ were obtained for the A and B excitons, respectively. Using the theoretically estimated reduced mass of the A exciton, ranging from 0.15 to $0.22m_e$, $a_{B,A}^* = 1.48 - 1.79$ nm is obtained.

Furthermore, the exciton binding energy and wavefunction can be determined by numerically solving the 2D Schrödinger equation with the values σ and a_B determined in these measurements. For 2D TMDs, a modified Coulomb potential $U(\rho)$ was used in the calculation,

$$U(\rho) = -\frac{e^2}{8\epsilon_0\rho_0} \left[H_0\left(\frac{\rho}{\rho_0}\right) - Y_0\left(\frac{\rho}{\rho_0}\right) \right] \quad (37)$$

where H_0 and Y_0 are the Struve function and Bessel function of the second kind, respectively, and the characteristic screening length $\rho_0 = 2\pi\chi_{2D}$, where χ_{2D} is the 2D polarizability of the monolayer material. Such a potential follows a $1/\rho$ Coulomb-like potential for large electron-hole separations $\rho \gg \rho_0$, but diverges weakly as $\log(\rho)$ for small separations $\rho \ll \rho_0$, leading to a different Rydberg series of exciton states with modified wavefunctions and binding energies that cannot be described within a hydrogen-like model. By doing this, some characteristic parameters are obtained for the A and B excitons in monolayer WS_2 . For example, with $\mu_A^* = 0.16m_0$ and $a_{B,A}^* = 1.53$ nm, the binding energy of the A exciton is calculated as $E_{b,A}^* = 410$ meV. For the B exciton with $\mu_B^* = 0.27m_0$ and $a_{B,B}^* = 1.16$ nm, $E_{b,B}^* = 470$ meV.

Perovskites

Solar cells based on organic-inorganic tri-halide perovskites show significantly improved performance, and therefore, it is very important to know their fundamental optical properties, such as the exciton binding energy and the reduced mass, in order to better understand the functions of solar cells. In addition, these materials have structural phase transitions from cubic ($T > 350$ K) to tetragonal ($T > 145$ K) to orthorhombic (low T), which leads to different band structures and band gaps in the different phases. By performing transmission measurements on a ~ 300 -nm-thick polycrystalline film of $\text{CH}_3\text{NH}_3\text{PbI}_3$ in a high B up to 150 T and at different temperatures, Miyata *et al.* obtained accurate values for the exciton binding energy and reduced mass.

As shown in **Fig. 4**, at 2 K, the transition energies are observed to depend quadratically on B for the lowest two states (labeled “ $N=0$ ”), as well as the “ $N=1$ ” transition at $B < 50$ T, indicating an excitonic feature for these low-energy states. On the other hand, the transition energies depend linearly on B for the $N=1$ transition at $B > 50$ T, and the higher energy states. The authors used full numerical calculations to fit the B -dependent transition energy data and obtained an accurate value for the reduced mass $\mu^* = 0.104 \pm 0.003m_e$. They also obtained the exciton binding energy $E_b^* = 16 \pm 2$ meV, which is over three times smaller than previously assumed. At room temperature, in the tetragonal structural phase, the exciton binding energy decreases to only a few meV, due to a frequency dependent dielectric constant, which reveals the free-carrier like nature of the photovoltaic device performance at room temperature.

Quantum Wires and Dots

Similar to 2D excitons in quantum wells, 1D excitons are observed in quantum wires, showing combined effects of magnetic fields and quantum potential in high magnetic fields. Nagamune *et al.* measured photoluminescence (PL) spectra of a GaAs quantum wire in fields up to 40 T, with three orthogonal B configurations. **Fig. 5(a)** displays the PL peak position as a function of B in different configurations, for the GaAs quantum wire and a bulk GaAs sample. The bulk PL peak shifts are almost independent of the magnetic field configuration, in contrast with a clear magnetic anisotropy observed in the quantum wire sample.

In the low- B limit, the PL peaks for the quantum wire show a quadratic increase with B , from which the diamagnetic-shift coefficients were determined as $16.3 \mu\text{eV T}^{-2}$, $7.0 \mu\text{eV T}^{-2}$, and $4.5 \mu\text{eV T}^{-2}$ for the $\vec{W} \perp \vec{B} \parallel \vec{k}$, $\vec{W} \perp \vec{B} \perp \vec{k}$ and $\vec{W} \parallel \vec{B} \perp \vec{k}$ configurations, respectively, where $\vec{W}$ is the direction along the quantum wire, and $\vec{k}$ the wavevector of the excitation beam. These values are smaller than $103.9 \mu\text{eV T}^{-2}$ for the bulk sample, indicating increased confinement in a reduced dimensionality.

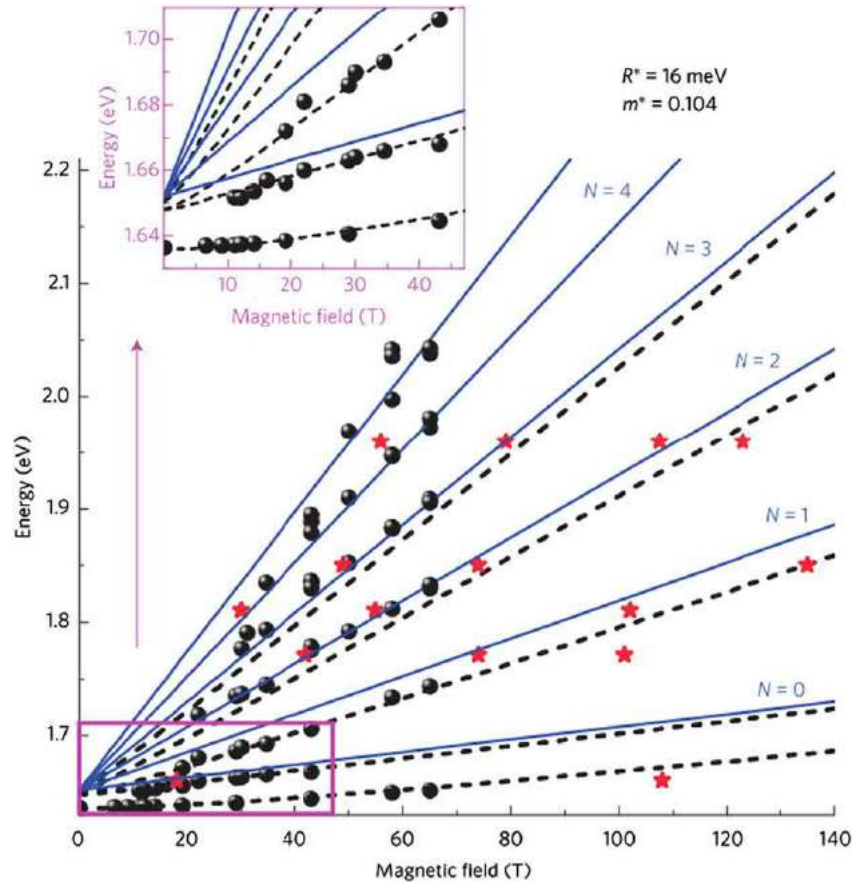


Fig. 4 (color online). Magnetic-field-dependent transition peaks of the perovskite $\text{CH}_3\text{NH}_3\text{PbI}_3$ at 2 K. The calculated transition energies are shown for the free-electron and hole levels (solid lines) and the excitonic transitions (dashed lines). Inset: measured and calculated transition energies at low fields. Reproduced (adapted) with permission from Miyata, A., Mitoglu, A., Plochocka, P., *et al.*, 2015. Direct measurement of the exciton binding energy and effective masses for charge carriers in organic-inorganic tri-halide perovskites. *Nature Physics* 11, 582–588. Copyright 2015 by Nature Publishing Group.

In the high- B limit, the PL peak shifts increase roughly linearly with increasing B . In particular, the PL energy shift under $\vec{W} \perp \vec{B} \parallel \vec{k}$ is almost parallel to that of a bulk sample, which is ascribed to the exciton radius becoming much smaller than the width of the confinement potential in a high B .

In quantum dots, electrons and holes are confined in all three directions, and quasi-zero-dimensional (0D) excitons are observed. Hayden *et al.* performed magneto-PL measurements on self-assembled InGaAs quantum dots in B up to 42 T. The PL peak shifts as a function of B are shown in Fig. 5(b) for different samples and B directions. Similar to the case of quantum wires, the low- B data is characterized by a diamagnetic shift, and the high- B data indicates a Landau-level like behavior. In addition, the diamagnetic shift is always smaller when B is applied perpendicular to the growth direction rather than parallel to it, revealing that the 0D excitons are better confined in the plane perpendicular to B .

Carbon Nanotubes

Single-wall carbon nanotubes (SWCNTs) are another representative 1D semiconductor, which are tubular crystals composed of sp^2 -bonded carbon atoms. Unlike quantum wires, the basic properties of SWCNTs are determined by a pair of integers, or chirality, (n, m) . For tubes with $n = m$, they are metals, called “armchair” tubes; for tubes with $n - m = 3j$ (j is nonzero integer), they are small-gap semiconductors; and all others are medium-gap semiconductors. In an applied B where the magnetic flux ϕ passes through the axial direction of the SWCNTs (a tube threading flux), the quantum states of electrons in SWCNTs acquire an Aharonov-Bohm phase $2\pi\phi_0/\phi_0$ ($\phi_0 = h/e$ is the magnetic flux quantum), causing the band structure of SWCNTs to change with ϕ/ϕ_0 . For $0 < \phi/\phi_0 < 1/6$, the Aharonov-Bohm phase can cause a band gap change of $6E_g\phi/\phi_0$ (E_g is the band gap) in a semiconducting SWCNT, which can be observed in interband absorption and PL measurements.

In a semiconducting SWCNT with time-reversal symmetry, at $B = 0$ T, as shown in Fig. 6(a), the bonding-like and anti-bonding-like linear combinations of two equivalent valleys, K and K' , result in two lowest energy exciton states: optically active (or “bright”) and inactive (or “dark”) states. The energies of the two states are split by Δ_X , which is determined by the Coulomb interaction, the

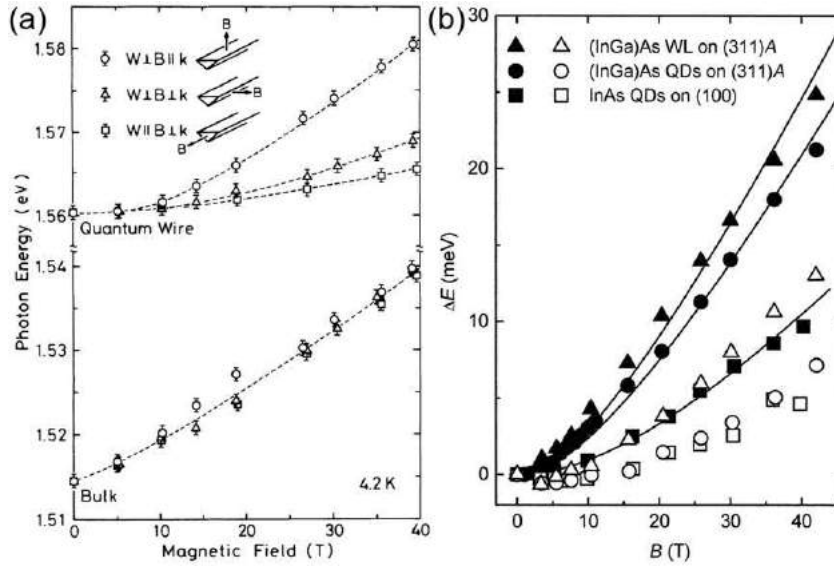


Fig. 5 (color online). (a) PL peak positions for a GaAs quantum wire and bulk GaAs samples as a function of B under various B configurations. Reproduced (adapted) with permission from Nagamune, Y., Arakawa, Y., Tsukamoto, S., Nishioka, M., 1992. Photoluminescence spectra and anisotropic energy shift of GaAs quantum wires in high magnetic fields. *Physical Review Letters* 69, 2963–2966. Copyright 1992 by the American Physical Society. (b) PL peak shifts as a function of B applied parallel to the growth axis (filled symbols) and perpendicular to the growth axis (open symbols), for quantum dots and wetting layers on different substrates. Reproduced (adapted) with permission from Hayden, R.K., Uchida, K., Miura, N., *et al.*, 1998. High field magnetoluminescence spectroscopy of self-assembled (InGa)As quantum dots on high index planes. *Physica B: Condensed Matter* 246, 93–96. Copyright 1998 by Elsevier.

tube diameter, and the dielectric constant of the environment. However, an applied B will break the time-reversal symmetry, and lift the $K-K'$ degeneracy, so that the dark state becomes optically active when the Aharonov-Bohm-induced splitting is larger than the Coulomb-induced splitting ($\Delta_{AB} > \Delta_X$). As a result, magnetic brightening of the dark exciton state can be observed in interband absorption or PL.

Fig. 6(b) shows absorption spectra of semiconducting SWCNTs at different B up to 71 T in a Voigt geometry, i.e., the light propagation vector is perpendicular to $\vec{B}$. The absorption increases with B , since the suspended carbon nanotubes progressively align with B more strongly due to the anisotropy of the magnetic susceptibility. The three main absorption peaks at $B=0$ T are labeled as A, B, and C, and each absorption peak splits into two well-resolved peaks at $B > 55$ T, due to the Aharonov-Bohm-induced splitting. This phenomenon can only be observed when the light polarization is parallel with B , as shown in **Fig. 6(c)**, as a result of magnetic anisotropy in SWCNTs.

Fig. 6(d) shows a PL contour map of a partially aligned carbon nanotube film in pulsed fields up to 56 T at 5 K. The PL intensity increases with B due to magnetic brightening. Furthermore, the Aharonov-Bohm effect causes a red-shift of each PL emission peak, since only the formerly dark state is populated due to a splitting much larger than the thermal energy. At low temperatures, the broken time-reversal symmetry increases the PL quantum yield by as much as a factor of 6, as shown in **Fig. 6(e)**. The magnetic brightening becomes weaker with increasing temperature; for instance, at 200 K the integrated PL intensity only increases by ~ 2 times.

Magnetic brightening has also been observed in single-nanotube PL. In this case, inhomogeneous broadening due to environmental effects are negligible, especially at low temperatures, and the dark-bright splitting, Δ_X , can be larger than the linewidth. Then, the dark state can become bright with an applied magnetic field. **Fig. 6(f)** shows B -dependent single-tube PL spectra at 5 K with B applied along the tube axis. At $B=0$ T, only a single, sharp PL peak can be observed, due to the bright exciton state. With increasing B , a second PL peak appears on the low energy side, grows rapidly in intensity and finally dominates the emission spectra at $B > 3$ T. The splitting between the two PL peaks increases with B . These behaviors were universally observed for more than 50 tubes of different chiralities, and the dark-bright splitting was 1–4 meV for tube diameters 1.0–1.3 nm.

Intraexciton Magneto-optical Absorption

Photoluminescence and absorption measurements can reveal information on interband transitions of s -like states of excitons. However, in order to explore the internal structure of excitons, intraexciton transitions such as $1s \rightarrow np$ ($n=2,3,\dots$) transitions have to be studied using far-infrared (FIR) or terahertz (THz) radiation, whose photon energy is comparable to or smaller than the binding energy of excitons.

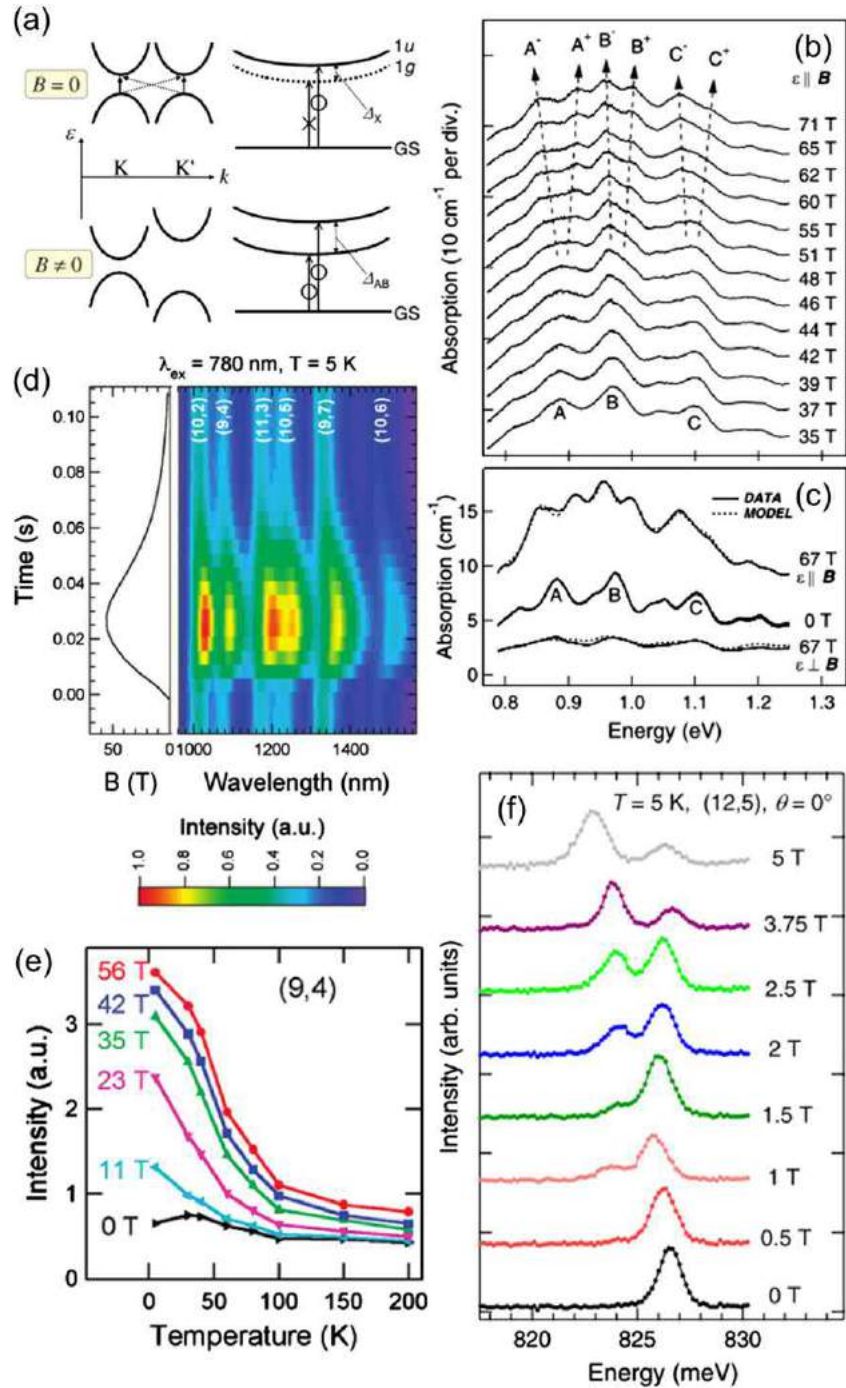


Fig. 6 (color online). (a) The expected B evolution of $K - K'$ intervalley mixing and splitting in a single-particle picture (left) and an excitonic picture (right). The solid (dashed) line represents a bright (dark) exciton state. Δ_X : Coulomb-induced splitting; Δ_{AB} : Aharonov-Bohm-induced splitting. Near band-edge absorption in semiconducting SWCNTs in high B for polarization (b) parallel to B (traces are offset) and (c) both polarizations (no intentional offset). Reproduced (adapted) with permission from Zaric, S., Ostojic, G. N., Shaver, J., *et al.*, 2006. Excitons in carbon nanotubes with broken time-reversal symmetry. *Physical Review Letters* 96, 016406. Copyright 2006 by the American Physical Society. (d) Magnetic brightening in SWCNTs in pulsed B . (e) Temperature dependence of magnetic brightening for (9,4) SWCNTs. Reproduced (adapted) with permission from Shaver, J., Kono, J., Portugall, O., *et al.*, 2007. Magnetic brightening of carbon nanotube photoluminescence through symmetry breaking. *Nano Letters* 7, 1851–1855. Copyright 2007 by the American Chemical Society. (f) B -dependent PL spectra for a single tube, showing the appearance of a dark exciton peak at a lower energy with respect to the main bright emission peak when B is applied parallel to the tube axis. Reproduced (adapted) with permission from Srivastava, A., Htoon, H., Klimov, V.I., Kono, J., 2008. Direct observation of dark excitons in individual carbon nanotubes: inhomogeneity in the exchange splitting. *Physical Review Letters* 101, 087402. Copyright 2008 by the American Physical Society.

FIR Magnetoabsorption in Photoexcited Bulk Semiconductors

Observation of intraexciton magneto-absorption in bulk semiconductors was initiated by Gershenzon *et al.* on the indirect bandgap semiconductor germanium in the microwave frequency range. Subsequently, Muro *et al.* and Ohyaama observed FIR magnetoabsorption of indirect excitons in germanium and silicon, respectively.

Fig. 7(a) shows FIR magnetoabsorption spectra for undoped germanium induced by interband photoexcitation at 4.2 K for various wavelengths. The observed spectra consist of exciton lines (E_X) and cyclotron resonance (CR) lines (C_e and C_h). Since the valence band is four-fold degenerate at the Γ point in germanium, the indirect free exciton states are described by four coupled effective mass equations, leading to the observation of four types of excitonic peaks in the spectra, E_{X1} , E_{X2} , E_{X3} , and E_{X4} . Up to the highest photoexcitation intensity used, these exciton peaks were stable against the formation of e - h droplets. The splitting of C_e' comes from the slight tilt of B from the [100]-crystal axis. C_{h1} , C_{h2} , C_{h3} , and C_{h4} correspond to different cyclotron transitions related to the light-hole and heavy-hole bands.

Fig. 7(b) shows the transition energies of intraexciton lines and CR lines as a function of B . An indirect exciton in germanium is composed of an electron at the L -point and a hole at the Γ -point bound by the Coulomb interaction, giving a hydrogenic structure. At zero B , the exciton transition lines converge to a point, indicating a finite transition energy of indirect excitons, around 3 meV. At high B , where the cyclotron energy is much larger than the Coulomb interaction, the intraexciton energies are expected to increase linearly with B , although it is not observed in this experiment since the highest B (7 T) was still not large enough to satisfy the adiabatic condition (required for the decoupling of the x - y and z degrees of freedom). These results demonstrate that FIR magnetoabsorption measurements can not only reveal the CR of electrons and holes but also provide detailed information on the intraexciton transitions in indirect semiconductors with complicated band structures.

Optically Detected FIR Resonance Spectroscopy of Semiconductor Quantum Wells

In optically detected resonance (ODR) measurements, two lasers illuminate the sample at the same time. One laser with near-infrared (NIR) or visible radiation is used to create excitons, while the other with FIR (or THz) radiation manipulates them, with the frequency tuned to the resonance at a given B . By monitoring the band-edge PL, information about intraexcitonic transition energy and strength can be obtained. Fig. 8(a) shows the ratio of PL amplitudes with and without FIR irradiation as a function of B for a GaAs/AlGaAs quantum well sample, measured by Černe *et al.* Two types of transitions are observed: electron CR and the $1s \rightarrow 2p^+$ excitonic transition. At a low FIR frequency, such as 20 cm^{-1} , only electron CR can be observed at a B . With increasing FIR frequency, the electron CR shifts to a higher B , and at the same time the $1s \rightarrow 2p^+$ excitonic transition appears at a relatively low B , along with a possible $1s \rightarrow 3p^+$ excitonic transition. At a high FIR frequency, such as 130 cm^{-1} , the electron CR energy does not become equal to the FIR photon energy within the measured B range, and only excitonic transitions can be observed.

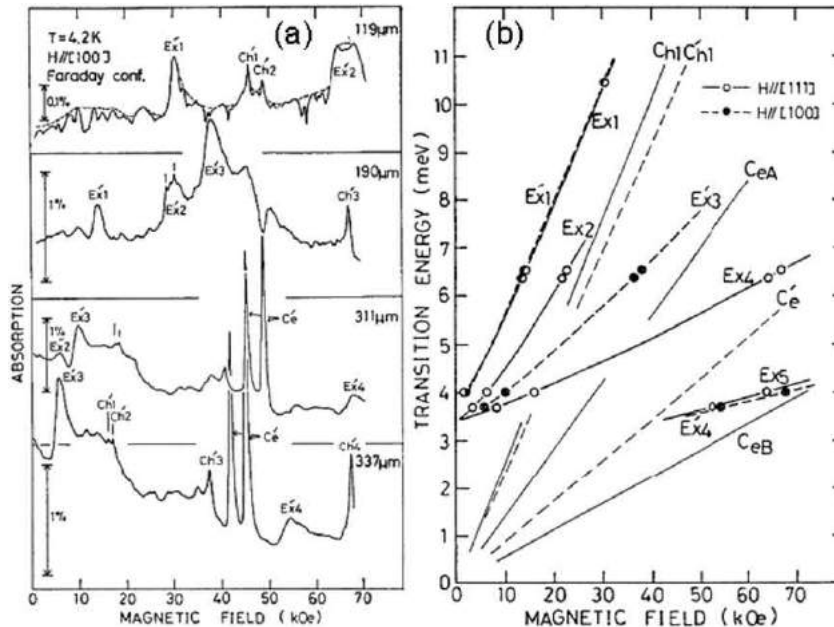


Fig. 7 (color online). (a) Magneto-absorption spectra for undoped germanium at 4.2 K with $H||[100]$ under optical excitation at varied wavelengths. (b) Transition energies of main exciton Zeeman lines and cyclotron resonance lines as a function of magnetic field. Reproduced (adapted) with permission from Muro, K., Nisida, Y., 1976. Far-infrared magneto-absorptions in photo-excited germanium. *Journal of the Physical Society of Japan* 40, 1069–1077. Copyright 1976 by the Physical Society of Japan.

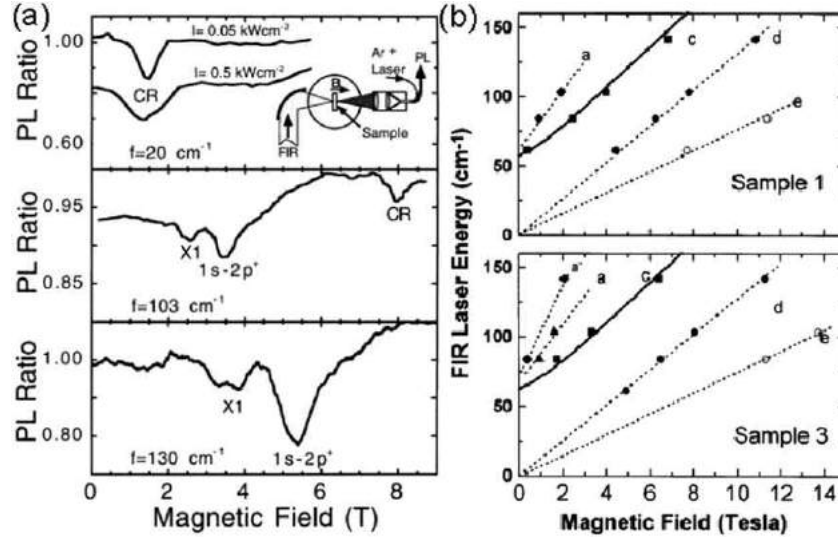


Fig. 8 (color online). Optically detected resonance spectroscopy. (a) The ratio of the PL amplitude with and without FIR irradiation as a function of B at three FIR frequencies. The inset shows the experimental setup. Reproduced (adapted) with permission from Černe, J., Kono, J., Sherwin, M. S., *et al.*, 1996. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. *Physical Review Letters* 77, 1131–1134. Copyright 1996 by the American Physical Society. (b) Energies of the optically detected resonances as a function of B at 4.2 K. Sample 1: an undoped (12.5 nm well/15 nm barriers)₃₀ structure. Sample 3: a (8 nm well/15 nm barrier)₄₅ structure, doped with Si donors ($n = 1 \times 10^{16} \text{ cm}^{-3}$) throughout the central one-third of the wells. Reproduced (adapted) with permission from Salib, M.S., Nickel, H.A., Herold, G.S., Petrou, A., McComber, B.D., 1996. Observation of internal transitions of confined excitons in GaAs/AlGaAs quantum wells. *Physical Review Letters* 77, 1135–1138. Copyright 1996 by the American Physical Society.

Salib *et al.* also performed similar ODR measurements on GaAs/AlGaAs quantum wells to observe intraexciton transitions. **Fig. 8(b)** plots the energies of the ODR features for different quantum wells as a function of B . In both samples, electron and hole CR lines (Features d and e) are observed, with their frequencies linearly dependent on B . In addition, intraexcitonic transitions, $1s \rightarrow 2p^+$ (Feature c) and $1s \rightarrow 3p^+$ (Feature a), can also be measured. In Sample 3, even the $1s \rightarrow 4p^+$ transition (Feature a') can be well resolved. These ODR measurements provide a method to explore the internal $1s \rightarrow np$ ($n = 2, 3, \dots$) transitions of photoexcited excitons, directly determining the energies of p -like exciton states.

In the nonlinear optics version of ODR, the sample is irradiated simultaneously by a NIR beam and an intense THz beam in the presence of B . Then very strong NIR emission lines, or sidebands, appear at frequencies $\omega_{\text{NIR}} \pm 2n\omega_{\text{THz}}$, where ω_{NIR} (ω_{THz}) is the frequency of the NIR (THz) beam and n is an integer. This process is highly resonant, and thus, changes sensitively with B , providing a novel method for intraexciton spectroscopy.

Fig. 9(a) and **(b)** show typical sideband spectra at 10 T, with the THz frequency $\omega_{\text{THz}} = 115 \text{ cm}^{-1}$. **Fig. 9(a)** shows the down-conversion effect: a narrow emission line, resulting from the creation of $2s$ heavy-hole excitons, is observed at 12745 cm^{-1} , and the -2ω sideband appears at exactly $12515 \text{ cm}^{-1} [= 12745 - (2 \times 115) \text{ cm}^{-1}]$ only under the THz irradiation. **Fig. 9(b)** shows the up-conversion behavior, with $+2\omega$ and $+4\omega$ sidebands in the higher energy region, through the resonant creation of $1s$ excitons at 12531 cm^{-1} . The typical intensities of the -2ω , $+2\omega$ and $+4\omega$ sidebands are 0.05%, 0.15% and 0.0015%, respectively, of the incident intensity of the NIR excitation. The intensity of the $+2\omega$ sideband increases quadratically with the THz power under a constant NIR intensity, and increases linearly with the NIR power for a constant THz intensity. Therefore, the 2ω THz sideband generation process is a third-order nonlinear optical process, involving one NIR photon and two THz photons.

Fig. 9(c) shows the $+2\omega$ sideband intensity as a function of B with $\omega_{\text{THz}} = 115 \text{ cm}^{-1}$ and $\omega_{\text{NIR}} = \omega_{1s}$, where ω_{1s} is the frequency for $1s$ heavy-hole excitons. Three distinct resonances can be observed, located at 4.5 T, 9.5 T and 11.5 T. These resonances occur when $\omega_{\text{THz}} = \omega_{2p^+} - \omega_{1s}$, $2\omega_{\text{THz}} = \omega_{2s} - \omega_{1s}$, and $\omega_{\text{THz}} = \omega_{2p^-} - \omega_{1s}$, as shown in **Fig. 9(d)**. According to calculations using perturbation theory, the third-order nonlinear optical susceptibility $\chi^{(3)}$ is resonantly enhanced, when (A) $\omega_{\text{NIR}} = \omega_{ns}$, (B) $\omega_{\text{NIR}} + \omega_{\text{THz}} = \omega_{n'pm}$ ($m = \pm 1$), and (C) $\omega_{\text{NIR}} + 2\omega_{\text{THz}} = \omega_{n''s}$. In all three cases in **Fig. 9(d)**, two of the three conditions (A)-(C) are satisfied: (left) $\omega_{\text{NIR}} = \omega_{1s}$, $\omega_{\text{NIR}} + \omega_{\text{THz}} = \omega_{2p^+}$, (middle) $\omega_{\text{NIR}} = \omega_{1s}$, $\omega_{\text{NIR}} + 2\omega_{\text{THz}} = \omega_{2s}$, and (right) $\omega_{\text{NIR}} = \omega_{1s}$, $\omega_{\text{NIR}} + \omega_{\text{THz}} = \omega_{2p^-}$. In other words, when ω_{THz} coincides with magnetically tuned transitions of excitons, the intensity of sidebands is dramatically enhanced, which provides a highly sensitive way to explore the internal structure of excitons.

Optical-Pump/THz-Probe Magnetospectroscopy of Semiconductor Quantum Wells

In order to obtain information on the dynamics of intraexcitonic transitions, time-resolved THz spectroscopy can be a powerful technique. Zhang *et al.* performed time-resolved THz absorption measurements on photoexcited e - h pairs in undoped GaAs quantum wells at various magnetic fields, temperatures and pump intensities, investigating both CR and intraexciton resonance features through resonant and nonresonant excitations of the heavy-hole $1s$ excitons.

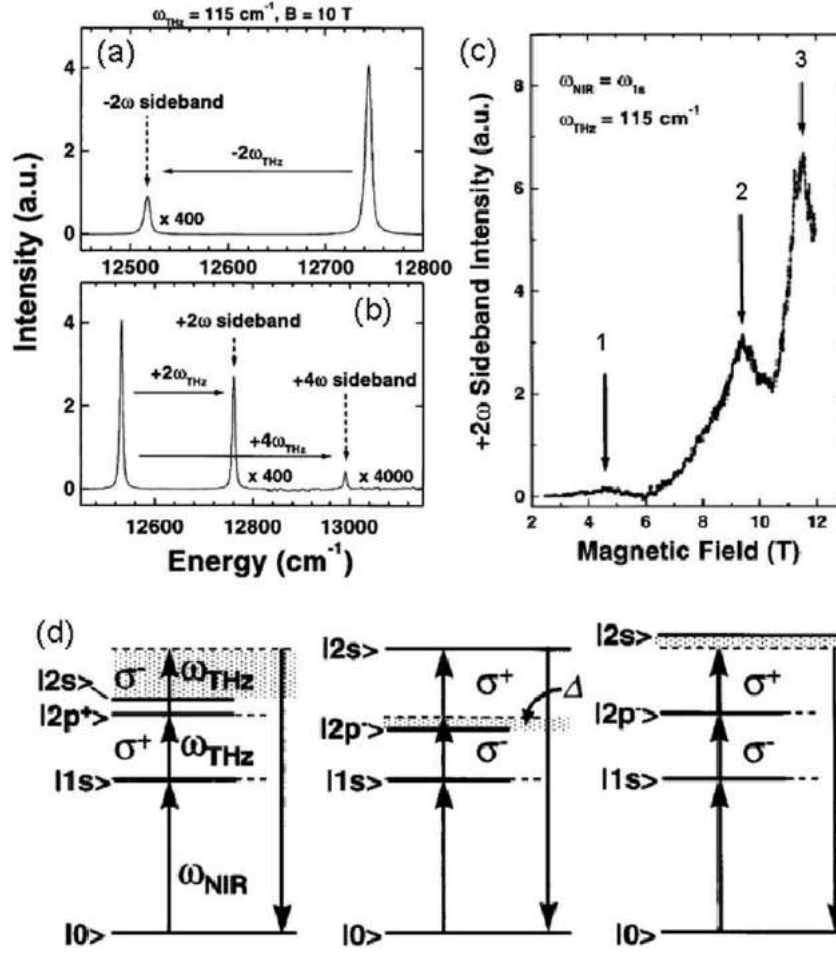


Fig. 9 (color online). Nonlinear THz ODR with $\omega_{\text{THz}} = 115 \text{ cm}^{-1}$. (a) Down conversion: $\omega_{\text{NIR}} = \omega_{2s} = 12745 \text{ cm}^{-1}$ and $\omega_{\text{sideband}} = \omega_{\text{NIR}} - 2\omega_{\text{THz}} = 12515 \text{ cm}^{-1}$. (b) Up conversion: $\omega_{\text{NIR}} = \omega_{1s} = 12515 \text{ cm}^{-1}$, $\omega_{\text{sideband}} = \omega_{\text{NIR}} + 2\omega_{\text{THz}} = 12745 \text{ cm}^{-1}$, and $\omega_{\text{NIR}} + 4\omega_{\text{THz}} = 12975 \text{ cm}^{-1}$. (c) The B dependence of the $+2\omega$ sideband intensity for $\omega_{\text{NIR}} = \omega_{1s}$ at all B . The data demonstrate nonlinear THz ODR. Pronounced resonances occur when 1: $\omega_{\text{THz}} = \omega_{2p^+} - \omega_{1s}$, 2: $2\omega_{\text{THz}} = \omega_{2s} - \omega_{1s}$, and 3: $\omega_{\text{THz}} = \omega_{2p^-} - \omega_{1s}$. (d) Diagrammatic representation of the three resonances in (c). The dashed lines represent virtual levels, whereas the solid lines represent real magnetoexcitonic levels. Resonances occur when real levels coincide with virtual levels. The shaded area represents the magnitude of the detuning, Δ . Reproduced (adapted) with permission from Kono, J., Su, M.Y., Inoshita, T., *et al.*, 1997. Resonant terahertz optical sideband generation from confined magnetoexcitons. *Physical Review Letters* 79, 1758–1761. Copyright 1997 by the American Physical Society.

Fig. 10(a) and **(b)** show the photoinduced change in conductivity, $\text{Re}\Delta\sigma(\omega)$, as a function of B and energy, at different time delays after nonresonant excitation at 5 K. At 15 ps, two distinct CR features due to unbound electrons and holes are observed, from which the effective masses can be determined: $m_e^* = 0.070m_e$ and $m_h^* = 0.54m_e$. At 1 ns, the intraexcitonic transitions, $1s \rightarrow 2p^+$ and $1s \rightarrow 2p^-$, are also observed, indicating the formation of excitons.

Fig. 10(c) and **(d)** present the time evolution of the change in conductivity at 9 T and 6 K with a pump fluence of 200 nJcm^{-2} for nonresonant and resonant excitation, respectively. Under nonresonant excitation, hole CR is observed at early time delays, such as 5 ps and 100 ps, but then decays with time and eventually the $1s \rightarrow 2p^-$ intraexciton transition appears after 100 ps. However, in the case of resonant excitation, only the $1s \rightarrow 2p^-$ transition can be observed throughout the entire time delay range, due to the direct creation of excitons. The change in conductivity gradually decreases with time through interband recombination. Through optical pump/THz-probe spectroscopy of magneto-excitons, in addition to exploring the internal structure of excitons, we are able to explore the dynamics of their formation and transitions from bound excitons to unbound electron-hole pairs.

Many-Body Effects of High-Density Excitons in High Magnetic Fields

With increasing e - h pair density, excitons can ionize to form an e - h plasma, where the electrons and holes are no longer bound to one another. This transition from excitons to an e - h plasma is referred to as the excitonic Mott transition. Here, we will focus on this high-density regime but in high B , where many-body effects can lead to many exotic behaviors, which are not observed in

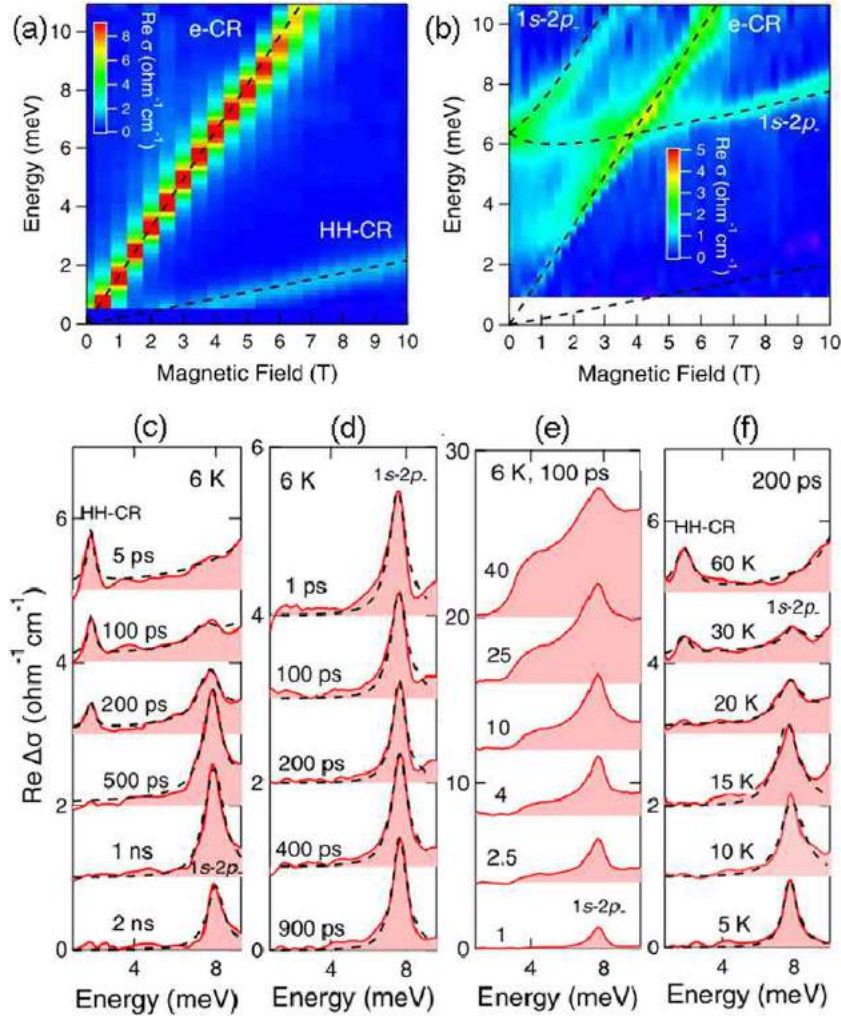


Fig. 10 (color online). Photoinduced change in conductivity, $\text{Re}\Delta\sigma(\omega)$, as a function of B , at a time delay of (a) 15 ps and (b) 1 ns after nonresonant excitation of undoped GaAs quantum wells. Conductivity spectral evolution after (c) nonresonant and (d) resonant excitations with a fluence of 200 nJcm^{-2} at 9 T. (e) Pump fluence dependent conductivity spectra at 9 T and 200 ps after resonant excitation. Reproduced (adapted) with permission from Zhang, Q., Wang, Y., Gao, W., *et al.*, 2016. Stability of high-density two-dimensional excitons against a Mott transition in high magnetic fields probed by coherent terahertz spectroscopy. *Physical Review Letters* 117, 207402. Copyright 2016 by the American Physical Society.

low-density interband and intraband absorption measurements. For example, in the high-density regime of a magnetoplasma, a macroscopic coherent dipole polarization can form initially from quantum fluctuations and finally release coherent, intense superfluorescent pulses of radiation. Another interesting phenomenon that occurs in the high- B limit of 2D magnetoexcitons is that they become ultrastable against the Mott transition due to the presence of a “hidden symmetry” that cancels the interactions between excitons.

Hidden Symmetry of 2D Magnetoexcitons

The hidden symmetry is unique to 2D magnetoexcitons and can be observed under the following conditions: the Coulomb interaction is charge-symmetric, and mixing of Landau levels can be neglected. Experimentally, the hidden symmetry can be observed when the filling factors of electrons and holes satisfy $0 \leq \nu_e, \nu_h \leq 2$, i.e., only the lowest Landau level is occupied. Under these conditions, the Coulomb interaction between excitons vanishes, and the 2D magnetoexcitons become ultrastable.

One example for the existence of hidden symmetry in 2D magnetoexcitons is shown in Fig. 10(e). Under 9 T, 5 K, and 100 ps after a resonant excitation of undoped GaAs quantum wells, both the intensity and linewidth of the $1s \rightarrow 2p^-$ peak increase with the pump fluence. Under the highest pump fluence ($40 \times 200 \text{ nJcm}^{-2}$), corresponding to a pair density of 10^{11} cm^{-2} per quantum well and a filling factor of $\nu_e < 2$, the intraexcitonic feature still remains and no CR is observed. Such a pair density is almost comparable to the Mott density at 0 T, so the absence of the Mott transition does confirm that the 2D magnetoexcitons are very

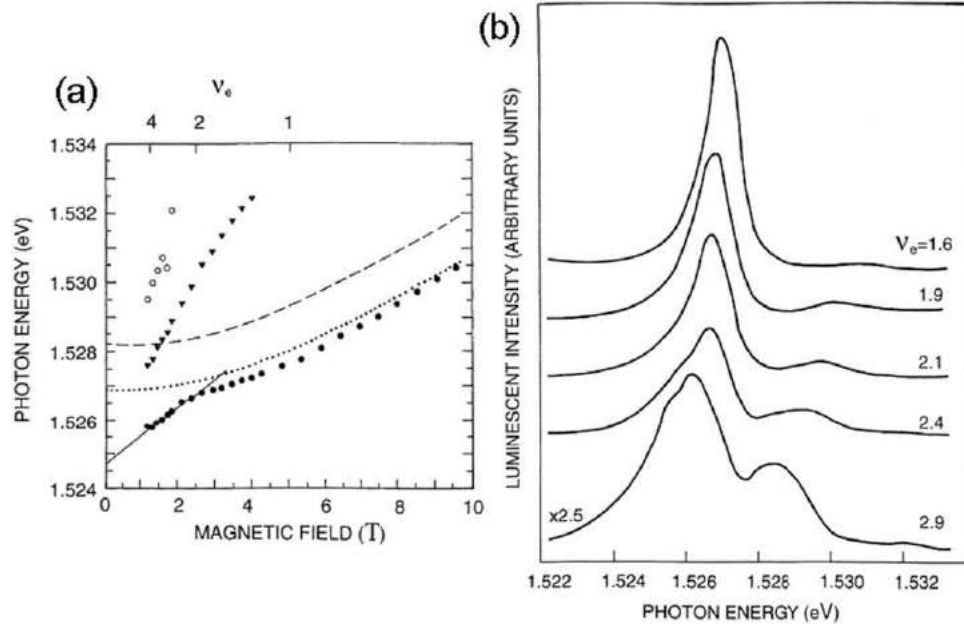


Fig. 11 (color online). (a) Photon energies of PL peak as a function of B at 4.2 K of a symmetric 20 nm quantum well with an electron density of $1.2 \times 10^{11} \text{ cm}^{-2}$. The solid line is the best fit to the linear shift of the lowest Landau level at low fields. The dashed line is the B dependence of the exciton at a low electron density, and the dotted line is that of the trion. (b) Photoluminescence spectra in the vicinity of $\nu_e = 2$. Reproduced (adapted) with permission from Rashba, E.I., Sturge, M.D., Yoon, H.W., Puffer, L.N., 2000. Hidden symmetry and the magnetically induced “Mott transition” in quantum wells containing an electron gas. *Solid State Communications* 114, 593–596. Copyright (2000) by Elsevier.

stable against the density-driven dissociation. However, the 2D magnetoexcitons dissociate under thermal excitation, as shown in Fig. 10(f). At a low temperature, the $1s \rightarrow 2p^-$ peak dominates the spectrum, and disappears with increasing temperature. At a high temperature, only the CR feature can be observed.

The hidden symmetry was also observed in quantum wells containing an electron gas. As shown in Fig. 11(a), at a certain B , the photon energy of the lowest transition changes from a linear dependence to a quadratic dependence on B . Above this critical B , the emission energy is very close to that of the singlet trion transition observed in the same quantum well at a low electron density. Different from a Mott transition, which is induced by the changes in screening and band-filling, this magnetically induced transition at a certain field results from the changes of the dynamic symmetry of the system, or hidden symmetry. Therefore, it is called a symmetry-driven transition.

Fig. 11(b) shows the effect of hidden symmetry in the vicinity of $\nu_e = 2$. At a relatively large filling factor for electrons, such as $\nu_e = 2.9$, the PL spectrum consists of three peaks, from the lowest Landau level transition, the blue-shifted cyclotron resonance, and a red-shifted satellite, possibly due to a phonon-assisted transition. With decreasing ν_e , the shape of the spectrum changes and the linewidth becomes sharp. That is because, under a critical B where $\nu_e < 2$, the frequency of the optical transition is not affected by the presence of electrons or holes, and the transition remains sharp.

Superfluorescence from High-Density 2D Magnetoexcitons

The concept of superradiance (SR) was proposed by Dicke in 1954, where he studied the radiative decay of an ensemble of N_{atom} incoherently excited two-level atoms, as shown in Fig. 12(a). If all of the atoms are confined in a region smaller than λ , where λ is the wavelength corresponding to the level separation, the emission behavior dramatically depends on the density of atoms. At a low density, the atoms do not interact with each other, and they radiate spontaneously with an intensity $\propto N_{\text{atom}}$ and a decay rate T_1^{-1} , where T_1 is the spontaneous radiative decay time of an isolated atom. However, at high density, all of the atoms lock in phase by photon exchange, and a giant dipole $P \sim N_{\text{atom}}d$ (d : individual atomic dipole) develops during a decay time τd . The macroscopic dipole leads to an intense coherent burst of radiation, with a decay rate $\Gamma \sim N_{\text{atom}}T_1^{-1}$, N_{atom} times faster than that of an individual atom, and a short pulse duration, $\tau_p \propto 1/N_{\text{atom}}$. Therefore, the intensity scales as $\propto N_{\text{atom}}/\tau_p \propto N_{\text{atom}}^2$, a hallmark of coherent emission.

Different from SR, where the polarization is generated by an external laser field, superfluorescence (SF) is a process in which a macroscopic polarization spontaneously develops from quantum fluctuations, which has been widely observed in atomic, molecular and solid systems. Very recently, SF was observed in InGaAs/GaAs quantum wells in quantizing B , in a setup schematically shown in Fig. 12(b). Interband transitions were studied through time-resolved PL, time-integrated PL and pump-probe spectroscopy. Since optical gain exists only for electromagnetic waves propagating along the quantum well plane, SF can be observed in the edge collection while ordinary spontaneous emission (SE) is collected in the center fiber.

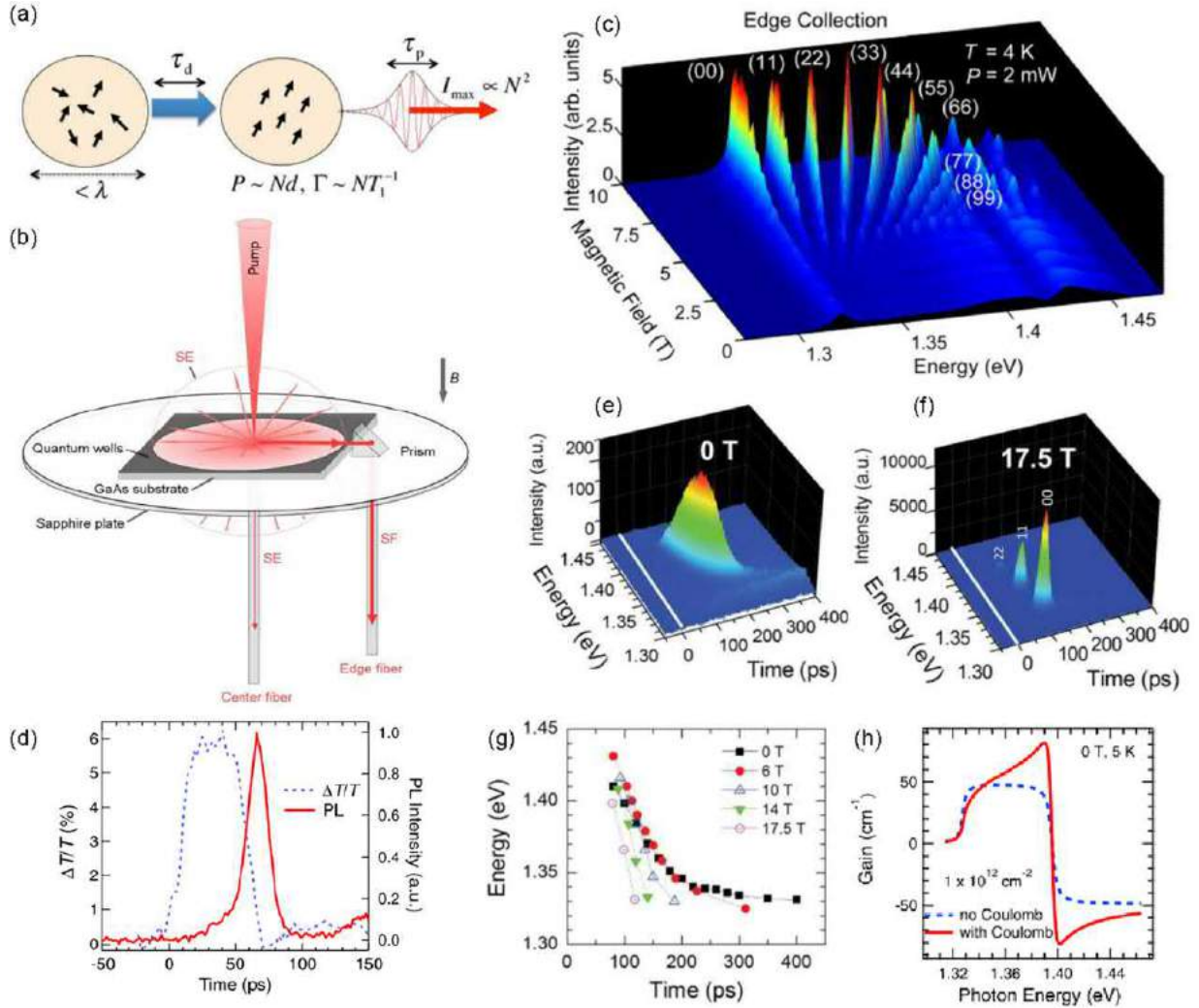


Fig. 12 (color online). (a) Basic processes and characteristics of SF. (b) Schematic diagram of the experimental geometry for SF observation from photoexcited semiconductor quantum wells in a B . (c) B dependence of time-integrated PL collected with the edge fiber at 4 K with an average excitation laser power of 2 mW. Reproduced (adapted) with permission from Cong, K., Wang, Y., Kim, J.H., *et al.*, 2015. Superfluorescence from photoexcited semiconductor quantum wells: Magnetic field, temperature, and excitation power dependence. *Physical Review B* 91, 235448. Copyright 2015 by the American Physical Society. (d) Simultaneously taken pump-probe and time-resolved PL data for the (22) transition at 17 T and 5 K, showing direct evidence of SF in the GaAs/InGaAs quantum well system. (a) and (d) reproduced (adapted) with permission from Cong, K., Zhang, Q., Wang, Y., *et al.*, 2016. Dicke superradiance in solids [Invited]. *Journal of the Optical Society of America B* 33, C80-C101. Copyright (2016) by the Optical Society of America. Time-resolved PL spectra at (e) 0 T and (f) 17.5 T with an excitation power of 2 mW at 5 K. (g) Peak shift of emission as a function of time at different magnetic fields. (h) Theoretical calculations of Coulomb-induced many-body enhancement of gain at the Fermi energy at 0 T. Reproduced (adapted) with permission from Kim, J.H., Noe II, G.T., McGill, S.A., *et al.*, 2013. Fermi-edge superfluorescence from a quantum-degenerate electron-hole gas. *Scientific Reports* 3, 3283. Copyright (2013) by Nature Publishing Group.

Fig. 12(c) shows B -dependent time-integrated SF emission at a pump intensity of 2 mW and a temperature of 4 K. At a low B (< 4 T), SF is characterized by two peaks at around 1.32 eV and 1.43 eV, corresponding to the heavy-hole and light-hole transitions, respectively. With increasing B , emission peaks due to the $(N,M) = (00), (11), \dots$, heavy-hole transitions, are observed, with stronger intensities and narrower widths.

In order to obtain direct evidence for SF emission, time-resolved pump-probe and PL measurements were performed. **Fig. 12(d)** shows pump-probe differential transmission and time-resolved PL data for the (22) transition at 17 T and 5 K. The pump creates the population inversion, as seen from the differential transmission, but at a time delay around 70 ps, it suddenly drops to zero, and at the same time a strong pulse of emission appears, as indicated by the time-resolved PL data.

The SF emission at 0 T is characterized by a continuous burst after a certain time delay, as shown in **Fig. 12(e)**, while the effects of quantization are obvious at a high B , as seen in **Fig. 12(f)**. By analyzing the emission time for different energies at various magnetic fields, as shown in **Fig. 12(g)**, a sequential manner of SF emission is found: SF from the highest occupied states is emitted

first, which is followed by emission from lower energy states. In addition, it can be seen that the SF emission occurs at shorter delay times with increasing B .

These results can be explained by the excitonic enhancement of gain near the Fermi energy in a high-density electron-hole system, as shown in Fig. 12(h). After relaxation and thermalization, degenerate Fermi gases form inside the conduction and valence bands each with quasi-Fermi energies. The recombination gain just below the quasi-Fermi energy is enhanced due to Coulomb interactions among carriers, which causes a SF burst to form at the Fermi edge. After the SF emission, the population is depleted, leading to a decreased Fermi energy. Thus, as time progresses, the Fermi level moves toward the band edge continuously. As a result, a continuous SF emission is observed at zero field and a series of sequential SF bursts appear in a magnetic field.

Summary

In this review, we have shown that the presence of excitons significantly modifies the optical response of semiconductor materials. Magneto-optical spectroscopy provides a powerful experimental means to determine the basic parameters of excitons such as the reduced mass, binding energy, Bohr radius, g -factor, internal structure, and formation dynamics. Depending on the dimensionality of excitons, these parameters can vary greatly, and we have presented experimental observations in magnetic fields for varied systems, including bulk semiconductors, 2D semiconductor quantum wells, monolayer TMDs, 1D quantum wires, SWCNTs, and semiconductor quantum dots. In addition to characterizing exciton parameters, magneto-optical spectroscopy of semiconductors can reveal exotic phenomena such as the “hidden symmetry” of 2D excitons in high magnetic fields and superfluorescence from a semiconductor magnetoplasma. The powerful experimental tools we have discussed will surely continue to reveal new phenomena related to excitons in semiconductors, including intriguing features related to the bosonic nature of excitons.

Further Reading

- Akimoto, O., Hasegawa, H., 1967. Interband optical transitions in extremely anisotropic semiconductors. II. Coexistence of exciton and the Landau levels. *Journal of the Physical Society of Japan* 22, 181–191.
- Černe, J., Kono, J., Sherwin, M.S., *et al.*, 1996. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. *Physical Review Letters* 77, 1131–1134.
- Cong, K., Wang, Y., Kim, J.H., *et al.*, 2015. Superfluorescence from photoexcited semiconductor quantum wells: magnetic field, temperature, and excitation power dependence. *Physical Review B* 91, 235448.
- Cong, K., Zhang, Q., Wang, Y., *et al.*, 2016. Dicke superradiance in solids [Invited]. *Journal of the Optical Society of America B* 33, C80–C101.
- Hayden, R.K., Uchida, K., Miura, N., *et al.*, 1998. High field magnetoluminescence spectroscopy of self-assembled (InGa)As quantum dots on high index planes. *Physica B: Condensed Matter* 246, 93–96.
- Kim, J.H., Noe II, G.T., McGill, S.A., *et al.*, 2013. Fermi-edge superfluorescence from a quantum-degenerate electron-hole gas. *Scientific Reports* 3, 3283.
- Kono, J., Su, M.Y., Inoshita, T., *et al.*, 1997. Resonant terahertz optical sideband generation from confined magnetoexcitons. *Physical Review Letters* 79, 1758–1761.
- MacDonald, A.H., Ritchie, D.S., 1986. Hydrogenic energy levels in two dimensions at arbitrary magnetic fields. *Physical Review B* 33, 8336–8344.
- Miyata, A., Mitioglu, A., Plochocka, P., *et al.*, 2015. Direct measurement of the exciton binding energy and effective masses for charge carriers in organic-inorganic tri-halide perovskites. *Nature Physics* 11, 582–588.
- Muro, K., Nisida, Y., 1976. Far-infrared magneto-absorptions in photo-excited germanium. *Journal of the Physical Society of Japan* 40, 1069–1077.
- Nagamune, Y., Arakawa, Y., Tsukamoto, S., Nishioka, M., 1992. Photoluminescence spectra and anisotropic energy shift of GaAs quantum wires in high magnetic fields. *Physical Review Letters* 69, 2963–2966.
- Rashba, E.I., Sturge, M.D., Yoon, H.W., Pfeffer, L.N., 2000. Hidden symmetry and the magnetically induced “Mott transition” in quantum wells containing an electron gas. *Solid State Communications* 114, 593–596.
- Salib, M.S., Nickel, H.A., Herold, G.S., Petrou, A., McComber, B.D., 1996. Observation of internal transitions of confined excitons in GaAs/AlGaAs quantum wells. *Physical Review Letters* 77, 1135–1138.
- Shaver, J., Kono, J., Portugall, O., *et al.*, 2007. Magnetic brightening of carbon nanotube photoluminescence through symmetry breaking. *Nano Letters* 7, 1851–1855.
- Srivastava, A., Htoon, H., Klimov, V.I., Kono, J., 2008. Direct observation of dark excitons in individual carbon nanotubes: inhomogeneity in the exchange splitting. *Physical Review Letters* 101, 087402.
- Stier, A.V., McCreary, K.M., Jonker, B.T., Kono, J., Crooker, S., 2016. Exciton diamagnetic shifts and valley Zeeman effects in monolayer WS₂ and MoS₂ to 65 Tesla. *Nature Communications* 7, 10643.
- Tarucha, S., Okamoto, H., Iwasa, Y., Miura, N., 1984. Exciton binding energy in GaAs quantum wells deduced from magneto-optical absorption measurement. *Solid State Communications* 52, 815–819.
- Zaric, S., Ostojic, G.N., Shaver, J., *et al.*, 2006. Excitons in carbon nanotubes with broken time-reversal symmetry. *Physical Review Letters* 96, 016406.
- Zhang, Q., Wang, Y., Gao, W., *et al.*, 2016. Stability of high-density two-dimensional excitons against a Mott transition in high magnetic fields probed by coherent terahertz spectroscopy. *Physical Review Letters* 117, 207402.

Introduction

Linear optical spectroscopy, using the techniques of absorption, transmission, reflection and light scattering, has provided invaluable information about the electronic and vibrational properties of atoms, molecules, and solids. Optical techniques also possess some additional unique strengths: the ability to

1. generate nonequilibrium distributions functions of electrons, holes, excitons, phonons, etc. in solids;
2. determine the distribution functions by optical spectroscopy;
3. determine the nonequilibrium distribution functions on femtosecond timescales;
4. determine femtosecond dynamics of carrier and exciton transport and tunneling; and
5. investigate interactions between various elementary excitations as well as many-body processes in semiconductors.

Researchers have exploited these unique strengths to gain fundamental new insights into nonequilibrium, nonlinear, and transport physics of semiconductors and their nanostructures over the past four decades.

A major focus of these efforts has been devoted to understanding how a semiconductor in thermodynamic equilibrium, excited by an ultrashort optical pulse, returns to the state of thermodynamic equilibrium. Four distinct regimes can be identified:

- (a) the coherent regime in which a well-defined phase relationship exists between the excitation created by the optical pulse and the electromagnetic (optical) field creating it;
- (b) the nonthermal regime in which the coherence (well-defined phase relationships) has been destroyed by various collision and interference processes but the distribution function of excitations is nonthermal (i.e., cannot be described by a Maxwell–Boltzmann distribution or its quantum (degenerate) counterparts);
- (c) the hot carrier regime in which the distributions for various excitations are thermal, but with different characteristic temperatures for different excitations and thermal bath; and
- (d) the isothermal regime in which the excitations and the thermal bath are at the same temperature but there is an excess of excitation (e.g., electron–hole pairs) compared to thermodynamic equilibrium.

Various physical processes take the semiconductor from one regime to the next, and provide information not only about the fundamental physics of semiconductors but also about the physics and ultimate performance limits of electronic, optoelectronic, and photonic devices.

In addition to coherent and relaxation dynamics, ultrafast studies of semiconductors provide new insights into tunneling and transport of carriers, and demonstrate novel quantum mechanical phenomena and coherent control in semiconductors.

The techniques used for such studies have also advanced considerably over the last four decades. Ultrafast lasers with pulse widths corresponding to only a few optical cycles (1 optical cycle ≈ 2.7 fs at 800 nm) have been developed. The response of the semiconductor to an ultrafast pulse, or a multiple of phase-locked pulses, can be investigated by measuring the dynamics of light emitted by the semiconductor using a streak camera (sub-ps time resolution) or a nonlinear technique such as luminescence up-conversion (time resolution determined by the laser pulse width for an appropriate nonlinear crystal). The response of the semiconductor can also be investigated using a number of two or three beam pump-probe techniques such as transmission, reflection, light scattering or four-wave mixing (FWM). Both the amplitude and the phase of the emitted radiation or the probe can be measured by phase-sensitive techniques such as spectral interferometry. The lateral transport of excitation can be investigated by using time-resolved spatial imaging and the vertical transport and tunneling of carriers can be investigated using the technique of ‘optical markers’. Electro-optic sampling can be used for transmission/reflection studies or for measuring THz response of semiconductors. More details on these techniques, and indeed many topics discussed in this brief article, can be found in the Further Reading.

Coherent Dynamics

The coherent response of atoms and molecules is generally analyzed for an ensemble of independent (noninteracting) two-level systems. The statistical properties of the ensemble are described in terms of the density matrix operator whose diagonal components relate to population of the eigenstates and off-diagonal components relate to coherence of the superposition state. The time evolution of the density matrix is governed by the Liouville variant of the Schrödinger equation $i\hbar\dot{\rho} = [H, \rho]$ where the system Hamiltonian $H = H_0 + H_{\text{int}} + H_R$ is the sum of the unperturbed, interaction (between the radiation field and the two-level system), and relaxation Hamiltonians, respectively. Using a number of different assumptions and approximations, this equation of motion is transformed into the optical Bloch equations (OBE), which are then used to predict the coherent response of the system, often using iterative procedures, and analyze experimental results.

Semiconductors are considerably more complex. Coulomb interaction profoundly modifies the response of semiconductors, not only in the linear regime (e.g., strong exciton resonance in the absorption spectrum) but also in the coherent and nonlinear regime. The influence of Coulomb interaction may be introduced by renormalizing the electron and hole energies (i.e., introducing excitons), and by renormalizing the field–matter interaction strength by introducing a renormalized Rabi frequency. These changes and the relaxation time approximation for the relaxation Hamiltonian lead to an analog of the optical Bloch equations, the semiconductor Bloch equations which have been very successful in analyzing the coherent response of semiconductors.

In spite of this success, it must be stressed that the relaxation time approximation, and the Boltzmann kinetic approach on which it is based, are not valid under all conditions. The Boltzmann kinetic approach is based on the assumption that the duration of the collision is much shorter than the interval between the collisions; i.e., the collisions are instantaneous. In the non-Markovian regime where these assumptions are not valid, each collision does not strictly conserve the energy and momentum and the quantum kinetic approach becomes more appropriate. This is true for photo-excited semiconductors as well as for the quantum transport regime in semiconductors. Also, semiconductor Bloch equations have been further extended by including four-particle correlations. Finally, the most general approach to the description of the nonequilibrium and coherent response of semiconductors following excitation by an ultrashort laser pulse is based on nonequilibrium Green's functions.

Experimental studies on the coherent response of semiconductors can be broadly divided into three categories:

1. investigation of novel aspects of coherent response not found in other simpler systems such as atoms;
2. investigation of how the initial coherence is lost, to gain insights into dephasing and decoherence processes; and
3. coherent control of processes in semiconductors using phase-locked pulses.

Numerous elegant studies have been reported. This section presents some observations on the trends in this field.

Historically, the initial experiments focused on the decay of the coherent response (time-integrated FWM as a function of delays between the pulses), analyzed them in terms of the independent two-level model, and obtained very useful information about various collision processes and rates (exciton–exciton, exciton–carrier, carrier–carrier, exciton–phonon, etc.). It was soon realized, however, that the noninteracting two-level model is not appropriate for semiconductors because of the strong influence of Coulomb interactions. Elegant techniques were then developed to explore the nature of coherent response of semiconductors. These included investigation of time- and spectrally resolved coherent response, and of both the amplitude and phase of the response to complement intensity measurements. These studies provided fundamental new insights into the nature of semiconductors and many-body processes and interactions in semiconductors. Many of these observations were well explained by the semiconductor Bloch equations. When lasers with < 10 fs pulse widths became laboratory tools, the dynamics was explored on a time-scale much shorter than the characteristic phonon oscillation period (≈ 115 fs for GaAs longitudinal optical phonons), the plasma frequency (≈ 150 fs for a carrier density of $5 \times 10^{17} \text{ cm}^{-3}$), and the typical electron–phonon collision interval (≈ 200 fs in GaAs). Some remarkable features of quantum kinetics, such as reversal of the electron–phonon collision, memory effects, and energy nonconservation, were demonstrated. More recent experiments have demonstrated the influence of four-particle correlations on coherent nonlinear response of semiconductors such as GaAs in the presence of a strong magnetic field.

The ability to generate phase-locked pulses with controllably variable separation between them provides an exciting opportunity to manipulate a variety of excitations and processes within a semiconductor. If the separation between the two phase-locked pulses is adjusted to be less than the dephasing time of the system under investigation, then the amplitudes of the excitations produced by the two phase-locked pulses can interfere either constructively or destructively depending on the relative phase of the carrier waves in the two pulses. The nature of interference changes as the second pulse is delayed over an optical period. A number of elegant experiments have demonstrated coherent control of exciton population, spin orientation, resonant emission, and electrical current. Phase-sensitive detection of the linear or nonlinear emission provides additional insights into different aspects of semiconductor physics.

These experimental and theoretical studies of ultrafast coherent response of semiconductors have led to fundamental new insights into the physics of semiconductors and their coherence properties. Discussion of novel coherent phenomena is given in a subsequent section. Coherent spectroscopy of semiconductors continues to be a vibrant research field.

Incoherent Relaxation Dynamics

The qualitative picture that emerges from investigation of coherent dynamics can be summarized as follows. Consider a direct gap semiconductor excited by an ultrashort laser pulse of duration τ_L (with a spectral width $\hbar\Delta\nu_L$) centered at photon energy $\hbar\nu_L$ larger than the semiconductor bandgap E_g . At the beginning of the pulse the semiconductor does not know the pulse duration so that coherent polarization is created over a spectral region much larger than $\hbar\Delta\nu_L$. If there are no phase-destroying events during the pulse (dephasing rate $1/\tau_D \ll 1/\tau_L$), then a destructive interference destroys the coherent polarization away from $\hbar\nu_L$ with increasing time during the excitation pulse. Thus, the coherent polarization exists only over the spectral region $\hbar\Delta\nu_L$ at the end of the pulse. This coherence will be eventually destroyed by collisions and the semiconductor will be left in an incoherent (off-diagonal elements of the density matrix are 0) nonequilibrium state with peaked electron and hole population distributions whose central energies and energy widths are determined by $\hbar\nu_L$, $\hbar\Delta\nu_L$ and E_g , if the energy and momentum relaxation rates ($1/\tau_E$, $1/\tau_M$) are much smaller than the dephasing rates.

The simplifying assumptions that neatly separate the excitation, dephasing and relaxation regimes are obviously not realistic. A combination of all these (and some other) physical parameters will determine the state of a realistic system at the end of the exciting pulse. However, after times $\sim \tau_D$ following the laser pulse, the semiconductor can be described in terms of electron and hole populations whose distribution functions are not Maxwell-Boltzmann or Fermi-Direct type, but nonthermal in a large majority of cases in which the energy and momentum relaxation rates are much smaller than the dephasing rates. This section discusses the dynamics of this incoherent state.

Extensive theoretical work has been devoted to quantitatively understand these initial athermal distributions and their further temporal evolution. These include Monte Carlo simulations based on the Boltzmann kinetic equation approach (typically appropriate for time-scales longer than ~ 100 fs, the carrier-phonon interaction time) as well as the quantum kinetic approach (for times typically less than ~ 100 fs) discussed above. We present below some observations on this vast field of research.

Nonthermal Regime

A large number of physical processes determines the evolution of the nonthermal electron and hole populations generated by the optical pulse. These include electron-electron, hole-hole, and electron-hole collisions, including plasma effects if the density is high enough, intervalley scattering in the conduction band and intervalence band scattering in valence bands, intersub-band scattering in quantum wells, electron-phonon, and hole-phonon scattering processes. The complexity of the problem is evident when one considers that many of these processes occur on the same time-scale. Of these myriad processes, only carrier-phonon interactions and electron-hole recombinations (generally much slower) can alter the total energy in electronic systems. Under typical experimental conditions, the redistribution of energy takes place before substantial transfer of energy to the phonon system. Thus, the nonthermal distributions first become thermal distributions with the same total energy. Investigation of the dynamics of how the nonthermal distribution evolves into a thermal distribution provides valuable information about the nature of various scattering processes (other than carrier-phonon scattering) described above.

This discussion of separating the processes that conserve energy in the electronic system and those that transfer energy to other systems is obviously too simplistic and depends strongly on the nature of the problem one is considering. The challenge for the experimentalist is to devise experiments that can isolate these phenomena so each can be studied separately. If the laser energy is such that $h\nu_L - E_g < \hbar\omega_{LO}$, the optical phonon energy, then majority of the photo-excited electrons and holes do not have sufficient energy to emit an optical phonon, the fastest of the carrier-phonon interaction processes. The initial nonthermal distribution is then modified primarily by processes other than phonon scattering and can be studied experimentally without the strong influence of the carrier-phonon interactions. Such experiments have indeed been performed both in bulk and quantum well semiconductors. These experiments have exhibited spectral hole burning in the pump-probe transmission spectra and thus demonstrated that the initial carrier distributions are indeed nonthermal and evolve to thermal distributions. Such experiments have provided quantitative information about various carrier-carrier scattering rates as a function of carrier density. In addition, experiments with $h\nu_L - E_g > \hbar\omega_{LO}$ and $h\nu_L - E_g > \text{intervalley separation}$ have provided valuable information about intravalley as well as intervalley electron-phonon scattering rates. Experiments have been performed in a variety of different semiconductors, including bulk semiconductors and undoped and modulation-doped quantum wells. The latter provide insights into intersub-band scattering processes and quantitative information about the rates.

These measurements have shown that under typical experimental conditions, the initial nonthermal distribution evolves into a thermal distribution in times $< \text{or} \sim 1$ ps. However, the characteristic temperatures of the thermal distributions can be different for electrons and holes. Furthermore, different phonons may also have different characteristic temperatures, and may even be nonthermal even when the electronic distributions are thermal. This is the hot carrier regime that is discussed in the next subsection.

Hot Carrier Regime

The hot carriers, with characteristic temperatures T_c (T_e and T_h for electrons and holes, respectively), have high energy tails in the distribution functions that extend several kT_c higher than the respective Fermi energies. Some of these carriers have sufficient energy to emit an optical phonon. Since the carrier-optical phonon interaction rates and the phonon energies are rather large, this may be the most effective energy loss mechanism in many cases even though the number of such carriers is rather small. The emitted optical phonons have relatively small wavevectors and nonequilibrium populations larger than expected for the lattice temperature. These optical phonons are often referred to as hot phonons although nonequilibrium phonons may be a better term since they probably have nonthermal distributions. The dynamics of hot phonons will be discussed in the next subsection, but we mention here that in case of a large population of hot phonons, one must consider not only phonon emission but also phonon absorption. Acoustic phonons come into the picture at lower temperatures when the fraction of carriers that can emit optical phonons is extremely small or negligible, and also in the case of quantum wells with sub-band energy separation smaller than optical phonons.

The dynamics of hot carriers have been studied extensively to obtain a deeper understanding of carrier-phonon interactions and to obtain quantitative measures of the carrier-phonon interaction rates. Time resolved luminescence and pump-probe transmission spectroscopy are the primary tools used for such studies. For semiconductors like GaAs, the results indicate that polar optical phonon scattering (longitudinal optical phonons) dominates for electrons whereas polar and nonpolar optical phonon

scattering contribute for holes. Electron–hole scattering is sufficiently strong in most cases to maintain a common electron and hole temperature for times $> \sim 1$ ps. Since many of these experiments are performed at relatively high densities, they provide important information about many-body effects such as screening. Comparison of bulk and quasi-two-dimensional semiconductors (quantum wells) has been a subject of considerable interest. The consensus appears to be that, in spite of significant differences in the nature of electronic states and phonons, similar processes with similar scattering rates dominate both types of semiconductors. These studies reveal that the energy loss rates are influenced by many factors such as the Pauli exclusion principle, degenerate (Fermi–Dirac) statistics, hot phonons, and screening and many-body aspects.

Hot carrier distribution can be generated not only by photo-excitation, but also by applying an electric field to a semiconductor. Although the process of creating the hot distributions is different, the processes by which such hot carriers cool to the lattice temperature are the same in two cases. Therefore, understanding obtained through one technique can be applied to the other case. In particular, the physical insights obtained from the optical excitation case can be extremely valuable for many electronic devices that operate at high electric fields, thus in the regime of hot carriers. Another important aspect is that electrons and holes can be investigated separately by using different doping. Furthermore, the energy loss rates can be determined quantitatively because the energy transferred from the electric field to the electrons or holes can be accurately determined by electrical measurements.

Isothermal Regime

Following the processes discussed above, the excitations in the semiconductor reach equilibrium with each other and the thermal bath. Recombination processes then return the semiconductor to its thermodynamic equilibrium. Radiative recombination typically occurs over a longer time-scale but there are some notable exceptions such as radiative recombination of excitons in quantum wells occurring on picosecond time-scales.

Hot Phonons

As discussed above, a large population of nonequilibrium optical phonons is created if a large density of hot carriers has sufficient energy to emit optical phonons. Understanding the dynamics of these nonequilibrium optical phonons is of intrinsic interest.

At low lattice temperatures, the equilibrium population of optical phonons is insignificant. The optical phonons created as a result of photo-excitation occupy a relatively narrow region of wavevector (k) space near $k=0$ ($\sim 1\%$ of the Brillouin zone). This nonequilibrium phonon distribution can spread over a larger k -space by various processes, or anharmonically decay into multiple large-wavevector acoustic phonons. These acoustic phonons then scatter or decay into small-wavevector, low-energy acoustic phonons that are eventually thermalized.

Pump-probe Raman scattering provides the best means of investigating the dynamics of nonequilibrium optical phonons. Such studies, for bulk and quantum well semiconductors, have provided invaluable information about femtosecond dynamics of phonons. In particular, they have provided the rate at which the nonequilibrium phonons decay. However, this information is for one very small value of the phonon wavevector so the phonon distribution function within the optical phonon branch is not investigated. The large-wavevector acoustic phonons are even less accessible to experimental techniques. Measurements of thermal phonon propagation provide some information on this subject. Finally, the nature of phonons is considerably more complex in quantum wells, and some interesting results have been obtained on the dynamics of these phonons.

Tunneling and Transport Dynamics

Ultrafast studies have also made important contributions to tunneling and transport dynamics in semiconductors. Time-dependent imaging of the luminescing region of a sample excited by an ultrashort pulse provides information about spatial transport of carriers. Such a technique was used, for example, to demonstrate negative absolute mobility of electrons in p-modulation-doped quantum wells. The availability of near-field microscopes enhances the resolution of such techniques to measure lateral transport to the subwavelength case.

A different technique of ‘optical markers’ has been developed to investigate the dynamics of transport and tunneling in a direction perpendicular to the surface (vertical transport). This technique relies heavily on fabrication of semiconductor nanostructures with desired properties. The basic idea is to fabricate a sample in which different spatial regions have different spectral signatures. Thus if the carriers are created in one spatial region and are transported to another spatial region under the influence of an applied electric field, diffusion, or other processes, the transmission, reflection, and/or luminescence spectra of the sample change dynamically as the carriers are transported.

This technique has provided new insights into the physics of transport and tunneling. Investigation of perpendicular transport in graded-gap superlattices showed remarkable time-dependent changes in the spectra, and analysis of the results provided new insight into transport in a semiconductor of intermediate disorder. Dynamics of carrier capture in quantum wells from the barriers provided information not only about the fundamentals of capture dynamics, but also about how such dynamics affects the performance of semiconductor lasers with quantum well active regions.

The technique of optical markers has been applied successfully to investigate tunneling between the two quantum wells in an a-DQW (asymmetric double quantum well) structure in which two quantum wells of different widths (and hence different

energy levels) are separated by a barrier. The separation between the energy levels of the system can be varied by applying an electric field perpendicular to the wells. In the absence of resonance between any energy levels, the wavefunction for a given energy level is localized in one well or the other. In this nonresonant case, it is possible to generate carriers in a selected quantum well by proper optical excitation. Dynamics of transfer of carriers to the other quantum well by tunneling can then be determined by measuring dynamic changes in the spectra. Such measurements have provided much insight into the nature and rates of tunneling, and demonstrated phonon resonances. Resonant tunneling, i.e., tunneling when two electronic levels are in resonance and are split due to strong coupling, has also been investigated extensively for both electrons and holes. One interesting insight obtained from such studies is that tunneling and relaxation must be considered in a unified framework, and not as sequential events, to explain the observations.

Novel Coherent Phenomena

Ultrafast spectroscopy of semiconductors has provided valuable insights into many other areas. We conclude this article by discussing some examples of how such techniques have demonstrated novel physical phenomena.

The first example once again considers an a-DQW biased in such a way that the lowest electron energy level in the wide quantum well is brought into resonance with the first excited electron level in the narrow quantum well. The corresponding hole energy levels are not in resonance. By choosing the optical excitation photon energy appropriately, the hole level is excited only in the wide well and the resonant electronic levels are excited in a linear superposition state such that the electrons at $t=0$ also occupy only the wide quantum well. Since the two electron eigenstates have slightly different energies, their temporal evolutions are different with the result that the electron wavepacket oscillates back and forth between the two quantum wells in the absence of damping. The period of oscillation is determined by the splitting between the two resonant levels and can be controlled by an external electric field. Coherent oscillations of electronic wavepackets have indeed been experimentally observed by using the coherent nonlinear technique of four-wave mixing. Semiconductor quantum wells provide an excellent flexible system for such investigations.

Another example is the observation of Bloch oscillations in semiconductor superlattices. In 1928 Bloch demonstrated theoretically that an electron wavepacket composed of a superposition of states from a single energy band in a solid undergoes a periodic oscillation in energy and momentum space under certain conditions. An experimental demonstration of Bloch oscillations became possible only recently by applying ultrafast optical techniques to semiconductor superlattices. The large period of a superlattice (compared to the atomic period in a solid) makes it possible to satisfy the assumptions underlying Bloch's prediction. The experimental demonstration of Bloch oscillations was also performed using four-wave mixing techniques. This provides another prime example of the power of ultrafast optical studies.

Summary

Ultrafast spectroscopy of semiconductors and their nanostructures is an exciting field of research that has provided fundamental insights into important physical processes in semiconductors. This article has attempted to convey the breadth of this field and the diversity of physical processes addressed by this field. Many exciting developments in the field have been omitted out of necessity. The author hopes that this brief article inspires the readers to explore some of the Further Reading.

Further Reading

- Allen, L., Eberly, J.H., 1975. *Optical Resonance and Two-Level Atoms*. New York: Wiley Interscience.
- Chemla, D.S., 1999. In: Garmire, E., Kost, A.R. (Eds.), *Nonlinear Optics in Semiconductors*. New York: Academic Press, pp. 175–256 (Chapter 3).
- Chemla, D.S., Shah, J., 2001. Many-body and correlation effects in semiconductors. *Nature* 411, 549–557.
- Haug, H., 2001. Quantum kinetics for femtosecond spectroscopy in semiconductors. In: Willardson RK and Weber ER (eds) *Semiconductors and Semimetals*, vol. 67, pp. 205–229. London: Academic Press.
- Haug, H., Jauho, A.P., 1996. *Quantum Kinetics in Transport and Optics of Semiconductors*. Berlin: Springer-Verlag.
- Haug, H., Koch, S.W., 1993. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.
- Kash, J.A., Tsang, J.C., 1992. In: Shank, C.V., Zakharchenya, B.P. (Eds.), *Spectroscopy of Nonequilibrium Electrons and Phonons*. Amsterdam: Elsevier, pp. 113–168.
- Mukamel, S., 1995. *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.
- Shah, J., 1992. *Hot Carriers in Semiconductor Nanostructures: Physics and Applications*. Boston: Academic Press.
- Shah, J., 1999. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*, 2nd edn. Heidelberg, Germany: Springer-Verlag.
- Steel, D.G., Wang, H., Jiang, M., Ferrio, K.B., Cundiff, S.T., 1994. In: Phillips, R.T. (Ed.), *Coherent Optical Interactions in Solids*. New York: Plenum Press, pp. 157–180.
- Wegener, M., Schaefer, W., 2002. *Semiconductor Optics and Transport*. Heidelberg, Germany: Springer-Verlag.
- Zimmermann, R., 1988. *Many-Particle Theory of Highly Excited Semiconductors*. Leipzig, Germany: Teubner.

Band Structure and Optical Properties

W Zawadzki, Polish Academy of Sciences, Warsaw, Poland

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The purpose of this article is to describe the influence of the electronic energy band structure of semiconductors on their optical properties. By the band structure one understands the energy–wavevector dispersion relation $\varepsilon(\mathbf{k})$ for the electron energy bands and the relative positions of band maxima and minima in the Brillouin zone.

The main problem in calculating the band structure of a semiconductor is that of the periodic potential of the lattice in which the electrons move. It is, in fact, a self-consistent problem since the moving electrons partly screen the potential. Different approximations have been developed to deal with the question, and in all of them the symmetry of the lattice is of fundamental importance. Thus, in the so-called empty-lattice approximation one deals exclusively with the symmetry and periodicity of the lattice without specifying the potential. This can give qualitative ordering and symmetry of the bands but no quantitative results. In the opposite limit one uses the tight binding approximation, in which the bands are constructed from the atomic states of separate atoms. This method gives quite a good description of lower (valence) bands but poor results for higher (conduction) bands. A powerful way to describe real energy bands is obtained by various forms of pseudopotential calculation, where one approximates the potential near actual atoms by simple parametrized forms and then adjusts the parameters to fit experimental (mostly optical) data. The pseudopotential methods give a good overall description of various bands in the entire Brillouin zone, but they do not provide sufficiently precise results near the band extrema. A semi-empirical way to serve the latter purpose is called $\mathbf{k}\cdot\mathbf{p}$ theory which we describe below. Various methods of band structure calculation are reviewed in books on solid state physics as outlined in the suggestions for Further Reading at the end of this article.

In the majority of important semiconductor materials (Si, Ge, many III–V and II–VI compounds) the maxima of the valence bands are at the center of the Brillouin zone (Γ point). The minima of the conduction bands in Si and Ge are not at the Γ point. This means that the fundamental optical interband transitions in these materials are indirect (in the ε – $\mathbf{k}$ diagram), i.e., they require phonon assistance. On the other hand, the minima of conduction bands in important III–V and II–VI compounds are at the Γ point, so that the interband optical transitions are direct and they do not require phonon assistance. Both systems are utilized in opto-electronic devices, particularly for detectors. However, for emitters (light-emitting diodes and lasers) we are mostly concerned with the second case, on which we concentrate here.

Bloch States

The Bloch theorem states that if the potential $V(\mathbf{r})$ in which the electron moves is periodic with the periodicity of the lattice, then the solutions $\Psi(\mathbf{r})$ of the Schrödinger wave equation

$$\left[\frac{\mathbf{p}^2}{2m_0} + V(\mathbf{r}) \right] \Psi(\mathbf{r}) = \varepsilon \Psi(\mathbf{r}) \quad (1)$$

are of the form $\Psi(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r})u_{\mathbf{k}}(\mathbf{r})$, where $u_{\mathbf{k}}(\mathbf{r})$ is periodic with the periodicity of the direct lattice, and $\mathbf{k}$ is the wavevector ($\hbar\mathbf{k}$ is the pseudomomentum). The proof of this theorem can be found, for example, in undergraduate textbooks on solid state physics.

$\mathbf{k}\cdot\mathbf{p}$ Theory

The underlying idea of semi-empirical $\mathbf{k}\cdot\mathbf{p}$ theory is to describe the energy bands in the vicinity of a given point of the Brillouin zone (usually near a band extremum). Symmetry properties are used to minimize the number of unknown band parameters (effective mass, energy gap, etc.) which are then determined by experiment. The initial Schrödinger equation for an electron in the periodic potential $V(\mathbf{r})$ of the crystal lattice reads

$$[\mathbf{p}^2/2m_0 + V(\mathbf{r}) + H_{\text{so}}]\Psi = \varepsilon\Psi \quad (2)$$

where m_0 is the free electron mass and H_{so} is the spin–orbit interaction. Since $H_{\text{so}}(\mathbf{r})$ is also periodic with the lattice periodicity, the solutions of Eq. (2) are given by the Bloch states above.

We use $\mathbf{k}\cdot\mathbf{p}$ perturbation theory, as follows, to describe the energy bands near $\mathbf{k}=0$ (i.e., the so-called Γ point of the Brillouin zone). Since we are interested in small values of $\mathbf{k}$, we expand our unknown cell periodic functions, $u_{\mathbf{k}}(\mathbf{r})$, in terms of the set of Γ -point functions, $u_{\mathbf{0}}(\mathbf{r})$, whose symmetry we know. One defines a representation $\Phi_{\mathbf{k}} = \exp(i\mathbf{k}\cdot\mathbf{r})u_{\mathbf{0}}(\mathbf{r})$, where $u_{\mathbf{0}}(\mathbf{r})$ is the periodic

$\mathbf{k}$ -independent function satisfying the equation

$$[\mathbf{p}^2/2m_0 + V + H_{so}]u_{l0} = \varepsilon_{l0}u_{l0} \quad (3)$$

whose solution we know in terms of the band-edge energies, ε_{l0} . This representation is sometimes referred to as the Luttinger and Kohn (LK) representation. The u_{l0} functions are in general mixtures of spin up and spin down states because of the spin-orbit interaction. One looks for the solutions of Eq. (2) in the form

$$\Phi_{n\mathbf{k}} = \exp(i\mathbf{k} \cdot \mathbf{r}) \sum_l c_l^{(n)}(\mathbf{k}) u_{l0}(\mathbf{r}) \quad (4)$$

where the sum is over all bands and $c_l^{(n)}(\mathbf{k})$ are the coefficients to be determined. Inserting the above form into Eq. (2), performing the operation $\mathbf{p}^2$ (i.e., operating twice on the Bloch product function with the operator $\mathbf{p} = -i\hbar\nabla$) and using the property (3) one eliminates the unknown potential $V(\mathbf{r})$. Multiplying the resulting equation from the left by $u_{l'0}^*$, integrating over the unit cell Ω and using the orthonormality of u_{l0} , one finally obtains

$$\sum_l \left[(\varepsilon_{l0} - \varepsilon') \delta_{l'l} + \frac{\hbar}{m_0} \mathbf{k} \cdot \mathbf{p}_{l'l} \right] c_l^{(n)}(\mathbf{k}) = 0 \quad (5)$$

for $l' = 1, 2, 3, \dots$. Here $\delta_{l'l}$ is the Kronecker delta function, $\varepsilon' = \varepsilon - \hbar^2 k^2 / 2m_0$, and $\mathbf{p}_{l'l}$ is the so-called matrix element of the momentum given by $\mathbf{p}_{l'l} = (1/\Omega) \int_{\Omega} u_{l'0}^* \mathbf{p} u_{l0} d\mathbf{r}$. The index l' runs over all energy bands.

Eq. (5) formulates the famous $\mathbf{k} \cdot \mathbf{p}$ theory. The nondiagonal part $\mathbf{k} \cdot \mathbf{p}_{l'l}$ can be eliminated by the perturbation procedure, and the method is referred to as $\mathbf{k} \cdot \mathbf{p}$ perturbation theory. In the second order of perturbation the bands are separated and in each band the carriers are described by a dispersion relation $\varepsilon_l = (\hbar^2/2)\mathbf{k} \cdot \left(1/m_{l0}^*\right)\mathbf{k}$ where $\left(1/m_{l0}^*\right)$ is an inverse effective mass tensor at the band edge in question. The inverse mass is generally a 3×3 tensor, but for cubic materials in the center of the Brillouin zone it is a scalar, so that $\varepsilon_l(\mathbf{k}) = \hbar^2 k^2 / 2m_{l0}^*$ (where $1/m_{l0}^* \equiv 1/m_0 + (2/m_0^2) \sum_{l' \neq l} (p_{ll'}^2) / [\varepsilon_{l0} - \varepsilon_{l'0}]$). We then say that the band is spherical and parabolic. The second order of perturbation is a good approximation if the energy ε (as counted from a nondegenerate band edge) is small compared to forbidden energy gaps. For the triply degenerate p-like valence band one has to do degenerate perturbation theory and treat the bands together as a 3×3 matrix equation.

In the same approximation the wavefunction for a given band is: $\Psi_{l\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k} \cdot \mathbf{r}) u_{l0}(\mathbf{r})$. In the absence of spin-orbit coupling, H_{so} , the u_{l0} function would just have the symmetry of the parent band (i.e., labeled S for the s-like conduction band, and X, Y or Z for the triply degenerate p-like valence band – each with a single spin up or spin down component). In the presence of H_{so} , the triply degenerate valence band states become mixtures of X, Y and Z with mixed spin character and part of the degeneracy is raised. In atomic notation the presence of H_{so} has transformed the basis from the (l, s) to the (J, m_J) representation. This results in the well-known light and heavy hole bands (degenerate at $k=0$) and the spin-orbit split-off band.

Under the influence of a radiation field, of vector potential $\mathbf{A}'$ and frequency ω , the optical transition probability for an electron to be raised from state Ψ_i (initial) to Ψ_f (final) is proportional to the square of $M = (e/mc) \langle \Psi_f | \mathbf{A}' \cdot \mathbf{p} | \Psi_i \rangle$ i.e., it is determined by the same momentum matrix element, p_{fi} , which governs the effective mass. For interband transitions one immediately gets the selection rule $\Delta k = 0$ (i.e., direct transitions). The polarization selection rules are determined by the $u_{l0}(\mathbf{r})$ components of the initial and final states, through the momentum matrix elements.

Narrow-Gap Semiconductors

Semiconductors such as InSb and $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ (with $x < 0.3$) have small energy gaps and are referred to as narrow gap semiconductors (NGSs). They are commonly used in opto-electronic applications which may be accessing energies in the conduction or valence band which are a significant fraction of the energy gap. Under these circumstances it is not valid to cut off the expansion to order k^2 , and one has to deal with so-called nonparabolic bands. Thus, in NGSs (or indeed in any situation where the energy becomes an appreciable fraction of ε_g) the above procedure is not valid and an alternative approximation is in order. Following semidegenerate perturbation theory one includes exactly in Eq. (5) a finite number of bands (near each other in energy) and neglects distant bands. This is referred to as the Kane model, and the energy bands near $k=0$ are shown for InSb-type III-V compounds and for HgTe-type II-VI compounds in Fig. 1. In this case we include only the conduction and valence band in our set of states (Eq. (5)). Including spin and degeneracy of the Γ_8 symmetry the set (5) then has eight equations. They can be solved analytically and the resulting energies are given by

$$\varepsilon'(\varepsilon' + \varepsilon'_g)(\varepsilon' + \varepsilon'_g + \Delta) - \left(\varepsilon' + \varepsilon'_g + \frac{2\Delta}{3} \right) \kappa^2 k^2 = 0 \quad (6)$$

where $\varepsilon' = \varepsilon - \hbar^2 k^2 / 2m_0$. We see that for this restricted set of states the only unknown parameters are the energy gap, ε_g , the spin-orbit splitting energy, Δ , and $\kappa = -i(\hbar/m_0) \langle S | p_z | Z \rangle$ (note that this is the only non-vanishing matrix element of the momentum). The fourth energy is $\varepsilon' = -\varepsilon_g$. All four energies are spin degenerate. In NGSs one may neglect the free electron term, i.e., set $\varepsilon' \approx \varepsilon$. For $\varepsilon \ll \varepsilon_g + (2/3)\Delta$ the resulting quadratic equation for the conduction and the light-hole bands is

$$\varepsilon(1 + \varepsilon/\varepsilon_g) = \hbar^2 k^2 / 2m_0^* \quad (7)$$

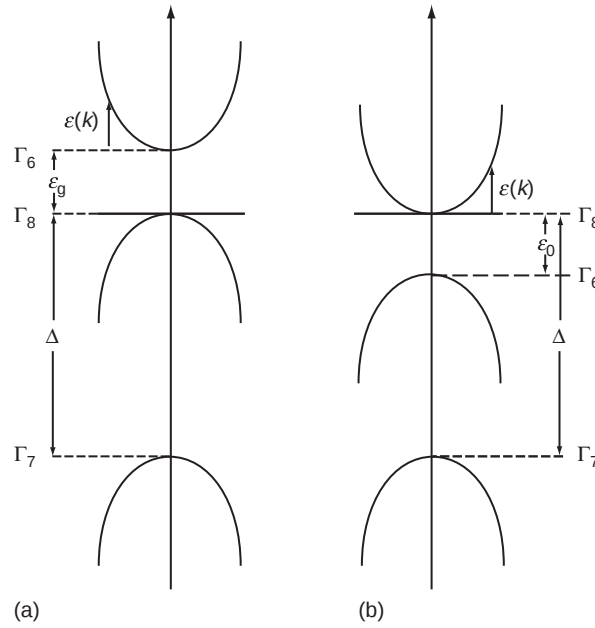


Fig. 1 Three-level model of band structure near the Γ point of the Brillouin zone: (a) InSb-type semiconductors; (b) HgTe-type semiconductors.

where $1/m_0^* = (4\kappa^2/3\hbar^2\varepsilon_g)(\varepsilon_g + 2\Delta/3)/(\varepsilon_g + \Delta)$ defines the effective mass at the band edge. The root for the heavy holes is $\varepsilon = -\varepsilon_g$ i.e., this band is not correctly described within the three-level model. The bands given by Eq. (6) are spherical but nonparabolic. for $\varepsilon \ll \varepsilon_g$ one recovers the standard dispersion, $\varepsilon = \hbar^2 k^2/2m_0^*$.

For HgTe-type materials, setting the zero of energy at the Γ_8 edge and replacing ε_g by $-\varepsilon_0$ (cf. Fig. 1), one obtains

$$\varepsilon'(\varepsilon' + \varepsilon_0)(\varepsilon' + \Delta) - \left(\varepsilon' + \frac{2\Delta}{3}\right)\kappa^2 k^2 = 0 \quad (8)$$

For $\varepsilon' \ll (2/3)\Delta$ the dispersion (7) is valid with $1/m_0^* = 4\kappa^2/3\hbar^2\varepsilon_0$. In this case the conduction and the heavy hole bands ($\varepsilon \equiv 0$) are degenerate at $k=0$, i.e., the thermal gap is zero.

The important property of the above model is that it holds also in the limit of $\varepsilon_g \rightarrow 0$ (i.e., for mixed $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ crystals). The effective mass in Eq. (7) tends to zero, but the dispersion described by Eq. (6) is valid and it gives $\varepsilon = (2/3)\kappa k$, i.e., a linear band.

The electron and hole wavefunctions in NGSs resulting from semidegenerate perturbation theory are given by Eq. (4), in which the sum runs over all the included states. Thus they involve truly $\mathbf{k}$ -dependent amplitudes of the Bloch states, cf. Eq. (1). In addition, these functions are mixtures of spin-up and spin-down states. These features have important consequences for optical and dc transport phenomena.

If the conduction band minimum is not at the Γ point (Ge and Si), there are at least two different matrix elements of momentum and the resulting band is ellipsoidal.

Effective Mass

The nonparabolic dispersions $\varepsilon(k)$ in NGSs require generalizations of important relations. The momentum mass $\widehat{m}^*$ is introduced as a connection between the pseudomomentum and the velocity: $\hbar\mathbf{k} = \widehat{m}^*\mathbf{v}$. Since $v_i = \delta\varepsilon/\delta(\hbar k_i)$, one obtains for a spherical band

$$\frac{1}{m^*} = \frac{1}{\hbar^2 k} \frac{d\varepsilon}{dk} \quad (9)$$

The above mass is a scalar, but it depends on the energy (or momentum). This mass is more useful than the one usually defined in textbooks, relating force to acceleration. The latter is not a scalar even for a spherical band. However, for the standard parabolic band, $\varepsilon = \hbar^2 k^2/2m_0^*$, both masses are the same and are equal to m_0^* . The mass (9) is related via velocity to the electric current, so that it enters naturally in transport phenomena (also free carrier optics). In particular, this mass enters the definition of mobility $\mu = q\tau/m^*$, where q is the charge and τ is the relaxation time. It enters the density of states (see below). Finally, it is the momentum mass (9) which is measured in cyclotron resonance.

For the dispersion (7) (the so-called two-level model), the mass (9) in the conduction band is

$$m^*(\varepsilon) = m_0^* \left(1 + \frac{2\varepsilon}{\varepsilon_g}\right) \quad (10)$$

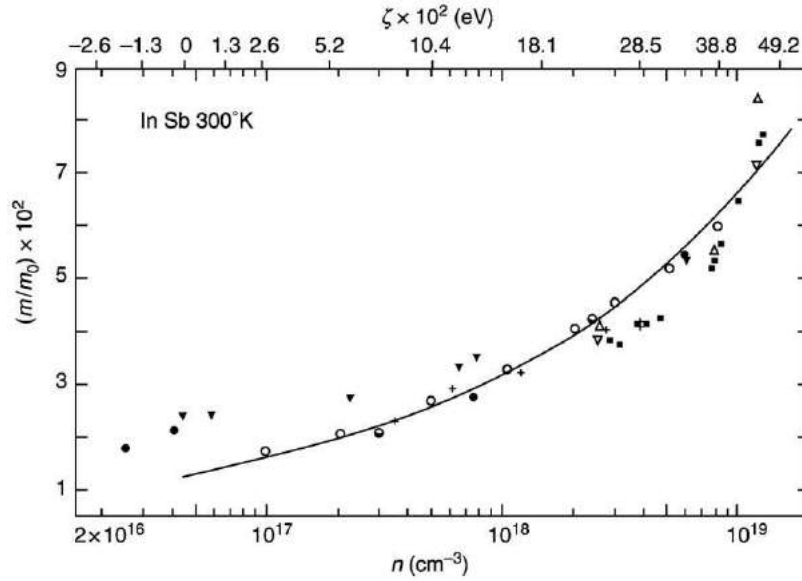


Fig. 2 Electron effective mass in InSb at room temperature versus free electron concentration. The solid line, calculated for the two-level model, represents the mass value at the Fermi energy, as indicated on the upper abscissa. The experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

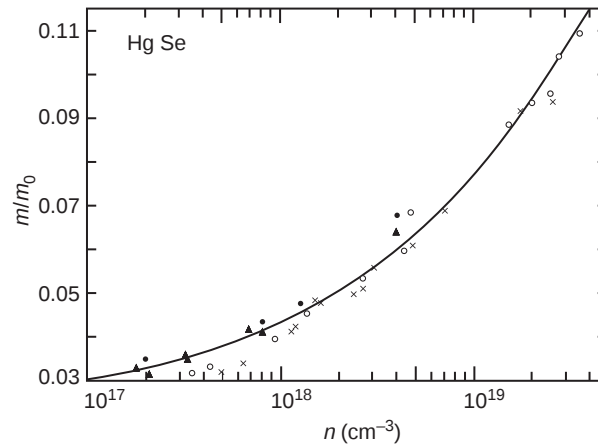


Fig. 3 Electron effective mass in zero-gap material HgSe versus free electron concentration. The solid line is calculated for the three-level model. Experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

Again, for $\varepsilon/\varepsilon_g \ll 1$ the energy dependence of the mass is negligible, which represents the standard regime. **Figs. 2** and **3** illustrate energy dependences of the electron masses in InSb and HgSe, as measured by various optical and transport methods. The increase of the mass with energy is very strong in both cases.

Density of States

The density of states in energy space is calculated beginning with pseudomomentum space. For the spherical band one obtains

$$\rho(\varepsilon) = \frac{k^2}{\pi^2} \frac{dk}{d\varepsilon} = \frac{m^* k}{\pi^2 \hbar^2} \quad (11)$$

The momentum mass (9) enters naturally into this important quantity. For the parabolic case $m^* = \text{const.}$ and $k \sim \varepsilon^{1/2}$, so that the standard result is recovered. In the general case one should use **Eq. (10)** for $m^*(\varepsilon)$ and determine k from **Eqs. (6)** or **(7)**.

Electron–photon Interaction

The electron–photon interaction can be introduced into the theory on two levels. If one replaces in the initial Eq. (2) the momentum $\mathbf{p}$ by $\mathbf{p} + (e/c)\mathbf{A}'$, the interaction term is obtained in the scalar form $H_{\text{int}} = (e/m_0c) \mathbf{A}' \cdot \mathbf{p}$, where $\mathbf{A}'$ is the vector potential of radiation. The wavefunctions to be used for the calculation of matrix elements are of the form (4), i.e., they include the periodic LK amplitudes. In fact, the matrix elements of momentum for optical transitions are identical to those of the $\mathbf{k} \cdot \mathbf{p}$ theory, as pointed out earlier. Since (as noted above) the wavefunctions for all bands have the same form (4), differing only by the coefficients $c_l^{(n)}(\mathbf{k})$, the matrix elements for interband and intraband optical transitions have the same form.

The electron–photon interaction can also be introduced directly to the $\mathbf{k} \cdot \mathbf{p}$ theory of Eq. (5) replacing $\hbar\mathbf{k}$ by $\hbar\mathbf{k} + (e/c)\mathbf{A}'$. If the free electron term $\hbar^2 k^2/2m_0$ is neglected, the interaction matrix involving $\mathbf{A}'$ terms is a number matrix. In this procedure the wavefunctions for initial and final states are simply columns and rows of $c_l^{(n)}$ coefficients and LK amplitudes no longer come into play. Here the $\mathbf{p}_{ll}$ elements of $\mathbf{k} \cdot \mathbf{p}$ theory determine directly the electron–photon interaction. Both procedures described above give the same results.

Quantum Wells

If charge carriers are placed in a quantum well described by a potential $U(z)$, the motion in the z -direction is quantized while the motion in the x – y plane remains free. One takes the same LK basis (cf. Eq. (3)), but the general form of solution is given by

$$\Psi = \sum_l f_l(\mathbf{r}) u_{l0}(\mathbf{r}) \quad (12)$$

since, in the presence of an external potential, $\mathbf{k}$ is not a good quantum number and the envelope functions $f_l(\mathbf{r})$ are unknown. If $U(z)$ is a slowly varying function within the unit cell the potential appears only on the diagonal of the set (5) and $\hbar\mathbf{k}$ is replaced by $\mathbf{p}$. One obtains

$$\sum_l \left[\left(\varepsilon_{l0} + \frac{\mathbf{p}^2}{2m_0} + U - \varepsilon \right) \delta_{ll} + \frac{\mathbf{p}_{ll} \cdot \mathbf{p}}{m_0} \right] f_l(\mathbf{r}) = 0 \quad (13)$$

One can now eliminate the nondiagonal terms applying the standard Luttinger and Kohn scheme of second-order perturbation theory and arrive at the decoupled band equations with effective masses. This results in the standard quantization due to electric (and magnetic, if present) fields, and wavefunctions are of the form $\Psi_{lm} = f_m u_{l0}$ for the m -th sub-band of the l -th band.

For NGSs this procedure is a poor approximation and one should proceed as before including exactly a finite number of bands (cf. Eq. (12)). This results in a set of coupled differential equations for the envelope functions $f_l(\mathbf{r})$. In some important cases one can find the eigenenergies without finding the functions by using the semiclassical approximation (the so-called WKB method).

We shall omit the calculations with coupled differential equations by using a semiclassical procedure also in another sense. Namely, we shall generalize the nonparabolic dispersion relation (7) to include the presence of external fields. To this end we observe that, including the potential $U(\mathbf{r})$ on the diagonal of Eq. (13), one replaces $-\varepsilon$ by $-\varepsilon + U$. Further, if a magnetic field is introduced to the $\mathbf{k} \cdot \mathbf{p}$ theory one replaces $\hbar\mathbf{k}$ by $\mathbf{P} = \mathbf{p} + (e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential of the magnetic field. Thus the semiclassical equation resulting from Eq. (7) is

$$\left[\frac{\mathbf{P}^2}{2m_0^*} - (\varepsilon - U) \left(1 + \frac{\varepsilon - U}{\varepsilon_g} \right) \right] \Psi = 0 \quad (14)$$

It can be seen that for $\varepsilon - U \ll \varepsilon_g$ one recovers the standard one-band approximation mentioned above. However, below we consider the general situation described by Eq. (14).

Let us first consider the case of no magnetic field, i.e., $\mathbf{P} = \mathbf{p}$. For $U(\mathbf{r}) = U(z)$ one can separate the variables by looking for solutions in the form $\Psi = \exp(ik_x x + ik_y y) \Phi(z)$. One can now transform Eq. (14) into

$$\left[\frac{1}{2m_0^*} p_z^2 - \frac{1}{\varepsilon_g} (\varepsilon - \varepsilon_{\perp} - U)(\varepsilon_g + \varepsilon + \varepsilon_{\perp} - U) \right] \Phi(z) = 0 \quad (15)$$

where

$$\varepsilon_{\perp} = \frac{\varepsilon_g}{2} + \left[\left(\frac{\varepsilon_g}{2} \right)^2 + \varepsilon_g \frac{\hbar^2 k_{\perp}^2}{2m_0^*} \right]^{1/2} \quad (16)$$

in which $k_{\perp}^2 = k_x^2 + k_y^2$. Eq. (15) is suitable for the WKB semiclassical quantization with which one determines the eigenenergies ε by one integration.

For a magnetic field applied along the z -direction (parallel electric and magnetic fields) the effects of both fields are separated and the form of Eq. (15) is valid, with $\hbar^2 k_{\perp}^2/2m_0^*$ in Eq. (18) replaced by $\hbar\omega_c(n+1/2) \pm (1/2)g_0^*\mu_B B$. Here $\omega_c = eB/m_0^*c$ is the cyclotron frequency, $n=0, 1, 2, \dots$ is the Landau quantum number, g_0^* is the spin splitting factor, and μ_B is the Bohr magneton. This results in nonequal spacing of Landau levels for a fixed value of B and in nonlinear B dependence of the Landau energies as functions of B (similarly but not identically to the bulk case). These features are illustrated by the cyclotron resonance experiment and theory shown in Fig. 4.

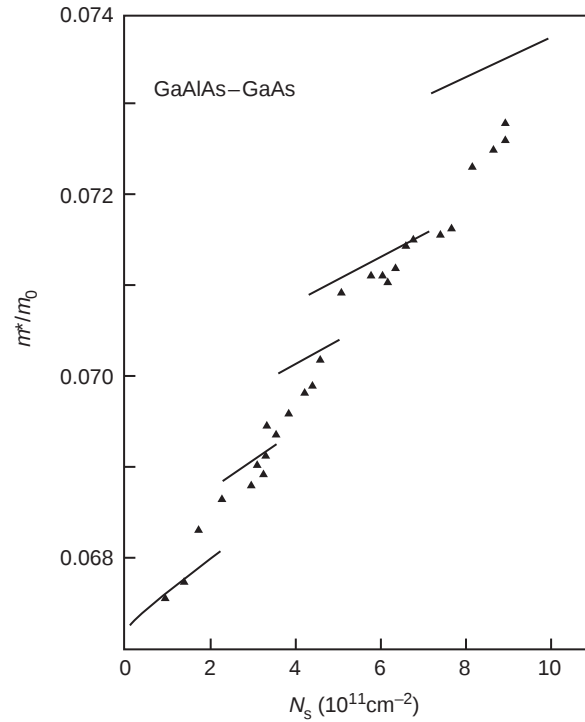


Fig. 4 Electron effective mass versus two-dimensional electron density in GaAs/GaAlAs heterostructures, as measured by far infrared cyclotron resonance. The solid lines are calculated using the effective two-level model for GaAs. Reproduced from Zawadzki W, *et al.* (1994) Cyclotron emission study of electron masses in GaAs-GaAlAs heterostructures. *Semiconductor Science and Technology* 9: 320–328.

Selection Rules for Intersub-band Optical Transitions

For the one-band effective mass approximation the electron–photon interaction is $H_{\text{int}} \sim \mathbf{A}' \cdot \mathbf{p}$ (cf. above), and the wavefunctions for the initial and final sub-bands are $\Psi_m = \exp(ik_x x + ik_y y) \Phi_m(z)$, in which the $\Phi_m(z)$ must be orthogonal to each other since they describe different energies. It is then clear that the matrix element $\langle f | \mathbf{A}' \cdot \mathbf{p} | i \rangle$ is finite only if $\mathbf{A}'$ is polarized along the z -direction (also called the growth direction) since for $\mathbf{A}'$ parallel to x or y the integral over z is $\langle \Phi_m | \Phi_m \rangle = 0$ giving a vanishing matrix element. Experimentally this is a serious problem for spectroscopic applications involving intraband (i.e., intersub-band) transitions in n -type semiconductors, since it is not possible to access these using normal incidence radiation.

This selection rule is somewhat relaxed for p -type semiconductors as a result of the mixed spin states of the complex valence band. In addition the narrow-gap band structure introduces interesting new possibilities into the selection rules. The wavefunctions have the form (12) and for the light polarizations A'_x or A'_y the momentum operators p_x or p_y act also on the periodic amplitudes $u_{j0}(\mathbf{r})$, which leads to a nonvanishing transition probability for intersub-band transitions. This has indeed been observed in narrow-gap materials.

Further Reading

- Ashcroft, N.W., Mermin, N.D., 1976. *Solid State Physics*. Holt, Rinehart and Winston.
- Kane, E.O., 1980. Band structure of narrow gap semiconductors. In: Zawadzki, W. (Ed.), *Narrow Gap Semiconductors. Physics and Applications*. Berlin: Springer, pp. 13–31.
- Luttinger, J.M., Kohn, W., 1955. Motion of electrons and holes in perturbed periodic fields. *Physical Review* 97, 869.
- Zawadzki, W., 1974. Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23, 435.
- Zawadzki, W., 1984. Electric and magnetic quantisation of two-dimensional systems – elementary theory. In: Bauer, G., *et al.* (Eds.), *Two-Dimensional Systems, Heterostructures, and Superlattices*. Berlin: Springer, p. 2.
- Zawadzki, W., 1991. Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr, G., Rashba, E.I. (Eds.), *Landau Level Spectroscopy*. Amsterdam: North-Holland.

Excitons

I Galbraith, Heriot-Watt University, Edinburgh, UK

© 2005 Elsevier Ltd. All rights reserved.

In semiconductors electron–hole pairs can lower their energy by correlating their motion in an exciton state. Such states dominate the bandedge optical properties of direct-gap semiconductors. The resultant absorption peaks are affected by quantum confinement, electric, and magnetic fields and carrier scattering.

Introduction

In the simplest, single-particle picture, the ground state of a semiconductor consists of a full valence band and an empty conduction band. The lowest lying excitation above the ground state in such a scheme is produced by excitation of one electron from the valence band to the conduction band. To achieve this optically requires a photon energy greater than the bandgap, E_g . However in many semiconductor materials this simple picture is unable to explain the observed absorption spectrum in the neighborhood of the fundamental bandgap. The origin of the discrepancy lies in the neglect of the Coulomb interaction between the negatively charged excited electron and the hole which is left in the valence band.

The term exciton refers to a two-particle excitation, consisting of a bound electron and hole. Such excitations dominate the bandedge optical spectra of most direct gap semiconductors, especially at low temperatures. In particular there exists a series of hydrogen-like bound states lying below the bandedge. These states are bound by energies

$$E_b^{3D} = \frac{\mu e^4}{8h^2 \epsilon_r^2 \epsilon_0^2 n^2} \quad (1)$$

where $\mu = m_e m_h / (m_e + m_h)$, is the reduced mass, $m_{e(h)}$ the electron (hole) effective mass, e is the electronic charge, h is Planck's constant, ϵ_0 is the permittivity of free space, ϵ_r the relative permittivity of the material and $n = 1, 2, 3, \dots \infty$ is the principal quantum number.

The binding energy of the lowest lying exciton state varies considerably from one semiconductor to another being, e.g., 0.5 meV in InSb and over 60 meV in ZnO. Similarly the spatial extent of the 1 s electron-hole relative wave function or effective Bohr radius is given by

$$a_0^{3D} = \frac{\hbar}{\sqrt{2\mu E_b^{3D}}} \quad (2)$$

and ranges from 750 μm in InSb to ≈ 2 nm in ZnO. Typically, in semiconductors the exciton spatial extent is much larger than the lattice constant of around 0.6 nm, and such excitons are classed Wannier excitons. At the opposite limit, which occurs in many molecular materials, the exciton is localized around a particular atomic site and is termed a Frenkel exciton. Wannier excitons are characterized by the hydrogen-like quantum numbers of their relative motion and an overall center of mass momentum which describes the wave-like motion of the bound entity through the crystal.

Another type of exciton is often referred to as a bound exciton. This consists of an electron–hole pair bound to a neutral impurity center. Such excitons are localized around the impurity and cannot move through the crystal in the same way as a free exciton.

Optical Properties

Excitons manifest themselves primarily as strong modifications to the bandedge optical properties of semiconductors. In particular the bound states give rise to discrete absorption lines at lower energy than the bandgap energy. Several peaks corresponding to the different hydrogen-like bound states (1s, 2s, etc.) can be seen provided the lines are sufficiently narrow. A theoretical calculation of this is shown in Fig. 1.

Associated with these absorption changes are concomitant changes in the refractive index. Contributions to the spectral width of the absorption peaks are characterized as either homogeneous or inhomogeneous. Homogeneous broadening refers to broadening associated with the finite lifetime of a particular state, e.g., due to phonon scattering. Inhomogeneous broadening is due to non-uniformity in the system, e.g., in a sample having a spatially non-uniform strain field.

There is also enhancement of the absorption at photon energies above the bandgap due to the residual influence of the Coulomb attraction on the electron-hole scattering states. Although not bound, electrons and holes in these states still have an enhanced probability of being found close together. This is called the Sommerfeld Enhancement Factor (see Fig. 1).

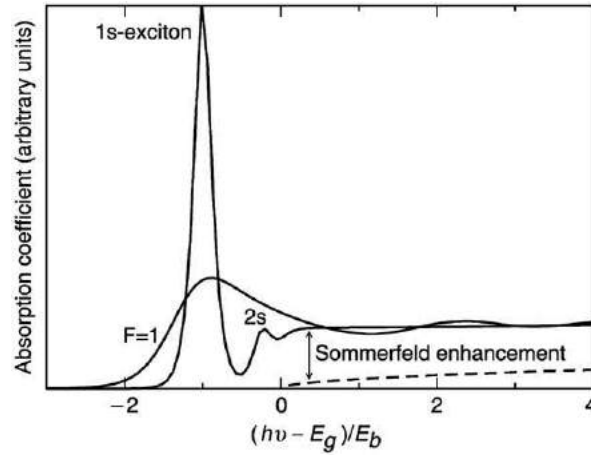


Fig. 1 Calculated bulk absorption spectra including (solid-line) and neglecting (dashed line) the Coulomb interaction. The Sommerfeld enhancement is seen as photon energies above the bandgap. Also shown is the spectrum with an applied electric field of one exciton Rydberg per Bohr radius.

It is important to note that although there may be an excitonic peak in the absorption spectrum this does not imply that excitons form a stable population in such a sample. Often, especially at room temperature, the lifetime of an optically generated exciton is determined by the scattering time with optical phonons which may be very short (< 1 ps). In this case the excitons are quickly ionized and the quasi-thermal equilibrium which is formed is dominated by essentially plasma-like behavior. In photoluminescence experiments, where carriers are generated high in the band and the spectrum of the resulting emitted luminescence is measured, peaks associated with the free exciton are often observed and taken as a signature of the presence of excitons. This view has been challenged recently by theoretical calculations which show that a photoluminescence peak at the exciton energy can be also be produced from an uncorrelated plasma if proper account of the Coulomb interaction is taken. This issue remains controversial.

In bulk samples, experiments such as absorption and reflectivity are complicated by the existence of polariton effects. An exciton-polariton is a quantum mixture of the propagating photon inside the semiconductor sample and the material excitation. Where the photon dispersion and the exciton energy dispersion meet there is an anti-crossing and this is manifested in, for example, the appearance of two lines in the reflectivity spectrum.

It was hoped during the 1970s that laser action in wide bandgap material such as ZnSe would be possible based on excitonic lasing. In this process a quasi-equilibrium exciton population would form the injected excitation and stimulated recombination would occur with the associated emission of a scattering partner which would take up the necessary momentum to ensure overall momentum conservation. This leads to a light emission wavelength below the absorption edge which is advantageous for minimizing losses. In fact, this process has only been seen at low temperature for scattering with LO-phonons, electrons and other excitons. Each mechanism has its own characteristic lasing threshold and temperature dependence.

Influence of Quantum Confinement

The advent of Molecular Beam Epitaxy enabled the growth of semiconductor layers of thickness similar to the electronic de Broglie wavelength. By sandwiching a low bandgap semiconductor between two high bandgap layers, one can make quantum well structures in which the electronic motion is effectively confined to the plane of the layer. In such structures the excitonic properties are also radically altered by this confinement. An example of the absorption spectrum of a ZnCdSe quantum well embedded in ZnSe is shown in Fig. 2. The heavy and light hole excitons correspond to different spin quantum numbers for the electron and hole making up the exciton.

The exciton binding energy is enhanced by a factor of 4 compared to the bulk case and the spatial extent of the 1s wavefunction is also reduced by half:

$$E_b^{2D} = \frac{\mu e^4}{8\hbar^2 \epsilon_r^2 \epsilon_0^2} \frac{1}{(n - \frac{1}{2})^2} \quad n = 1, 2, 3 \dots \quad (3)$$

This enhancement can be understood since the particles are confined to lie in the plane and are hence closer together on average than in the three-dimensional case. In practice, a four-fold enhancement in the binding energy is never observed since the quantum wells are neither infinitely deep nor infinitely thin. Both these limitations lead to a reduction in the actual enhancement from 4 to about 2, depending on the structure. For example, the binding energy in a realistic quantum well corresponds to the bulk value of the well material for very wide wells and to the bulk value of the barrier material for very narrow wells. In between the

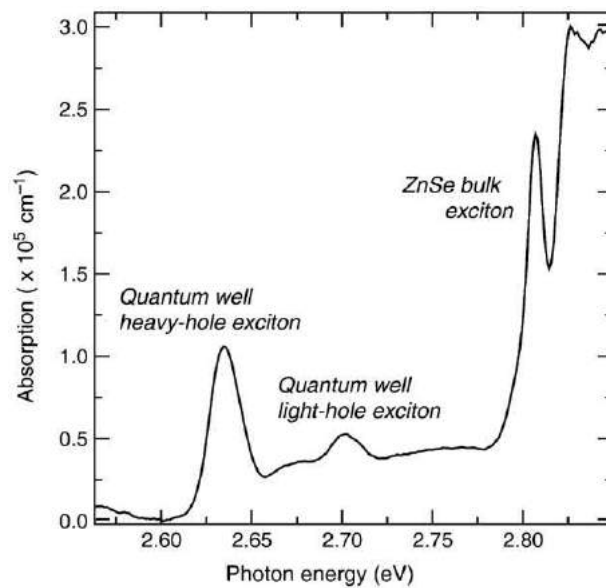


Fig. 2 Heavy and light hole exciton absorption spectrum for ZnCdSe quantum wells embedded between ZnSe barriers. Also shown is the bulk exciton of the ZnSe barriers.

value is enhanced by the quantum confinement. The maximum enhancement typically occurs when the well width is around half the bulk Bohr radius for the well material.

Accurate calculation of the exciton binding energy is complicated by the fact that the quantum confinement influences the band structure, making it strongly anisotropic and splitting heavy and light holes. A complete inclusion of these effects is at the limit of current computational power due to the need to calculate the electron-hole interaction matrix elements. Approximate schemes, including variational approaches, have proved successful in fitting a variety of samples, including GaAs and ZnSe based quantum wells. The limitation to such calculations is often the requirement for accurate values of the material parameters such as effective masses and energy band offsets.

The Sommerfeld enhancement of the absorption into the continuum states of quantum wells leads to a doubling of the absorption at the bandedge. This enhancement reduces only slowly as the photon energy is increased.

If multiple quantum wells with very thin barrier layers are used, the electronic states in the neighboring wells couple together to form a superlattice. The superlattice dispersion in the growth direction behaves like a heavy mass and the resulting exciton is three-dimensional but strongly anisotropic.

More extreme quantum confinement is possible making quantum wires, by etching or selective growth of quantum well samples. In these quasi-one dimensional structures, as in quantum wells, the exciton binding energy is again increased further. The Coulomb problem in one-dimension is pathological in that the ground-state binding energy diverges logarithmically, and higher states are pair-wise degenerate. However, accounting for the finite cross-section of realistic wires eliminates this problem and recovers a finite binding energy. One consequence of this is that there is no simple enhancement factor which corresponds to the four-fold enhancement in going from bulk to two-dimensional excitons. The density of states in one-dimensional systems diverges at the bandedge, implying the existence of an infinite absorption in a perfect quantum wire. However, the influence of excitonic correlations cancels this divergence, resulting in an enhanced, but not divergent, bandedge absorption feature.

The ultimate limit in quantum confinement can be achieved in quantum dots, where the carriers are confined in all three directions. In this case the term exciton should be used with some care as there are no bandedge states to compare with. The discrete energy levels dictated by the confinement are affected by the Coulomb interaction and the energy is renormalized – in some cases substantially. The absorption spectrum should consist of a series of sharp atomic-like transitions corresponding to s-s, p-p, d-d, ... transitions and indeed luminescence spectra obtained from single quantum dots do display such fine structure. Even in the best controlled material systems, such as InAs quantum dots, the spectra are strongly inhomogeneously broadened due to variations in the dot size, shape, and alloy composition.

In large quantum wells, where the width is larger than the exciton Bohr radius the center-of-mass part of the wavefunction can be confined by the barriers. This leads to a quantized motion for the exciton as a whole, rather than for the individual electron and hole separately. One implication of this is that for such excitons the center-of-mass momentum is no longer a good quantum number.

The above effects are a controlled case of disorder induced effects in semiconductors. This is an area which is only now beginning to be understood. The presence of local potential fluctuations will influence the energy spectrum of excitons in that vicinity. Clearly both the depth and spatial extent of these fluctuations will be important – very short range (less than Bohr radius),

shallow potential fluctuations will not have any influence on the exciton. While deep potentials can lead to the center of mass confinement above.

Influence of Electric and Magnetic Fields

Just as an externally applied electric field affects the absorption edge via the Franz–Keldysh Effect, broadening the edge itself and inducing oscillations in the continuum part of the spectrum, so, in the case of an excitonic bandedge, there is substantial modification due to the presence of the field. At low fields, when the Coulomb interaction is strong compared to the potential drop across the exciton diameter, the exciton will be little affected by the applied field. At higher fields the exciton will become gradually more polarized by the field until the exciton is field ionized. This ionization process is manifested in the field broadening of the exciton line (see Fig. 1).

Of more technological importance is the influence of an electric field applied perpendicular to the plane of a quantum well. The field forces the electron and hole in opposite directions leading to a reduction in the exciton binding energy. This would lead to a blue shift in the exciton absorption peak were it not for an even larger reduction in the single particle electron and hole energy levels within the quantum wells. Overall a strong red-shift of the excitonic absorption peak is observed with increasing field. This is accompanied by a reduction in the oscillator strength due to the reduced electron-hole overlap. This is termed the Quantum Confined Stark Effect and is the basis for the operation of some semiconductor electro-optic modulators. An electric field applied in the plane of the quantum well gives rise to Franz–Keldysh oscillations analogous to the bulk case.

The influence of magnetic fields on excitons is more subtle and richer than the electric field. This is because the magnetic field seeks to induce a circular cyclotron orbit motion on the carriers, but this is influenced by their mutual Coulomb interaction. Two regimes exist where either the cyclotron energy:

$$\hbar\omega_c = \hbar(eB/\mu) \quad (4)$$

or the exciton binding energy dominates. In the former strong field case, one can treat the Coulomb interaction as a small perturbation on the electron and hole states in the presence of the magnetic field. In exciton states, which have a finite magnetic moment at zero field, we find a linear Zeeman shift in the magneto-exciton energy. In the other case we must use the magnetic field as the perturbation which produces a mixing of the zero-field exciton states. For example, for the ground state the magnetic field induces an admixture of the p-like excited state with the s-like ground state. The total angular momentum of the mixed state is proportional to B and as the energy of a magnetic dipole in a magnetic field is also proportional to B the shift of the magneto-exciton is proportional to B^2 . This distinction between linear and quadratic shifts has been used to identify the nature of carrier populations (excitons or unbound free carriers) in a variety of samples. Clearly there exists an intermediate regime where both the magnetic and Coulomb energies are comparable and in this case the precise energy shifts need to be evaluated numerically.

In quantum well samples the orientation of the magnetic field with respect to the confinement direction is crucial in determining the physics. When the field is perpendicular to the confinement plane the cyclotron orbits exist as before and essentially the same phenomenon as in the bulk emerge. For other orientations, the behavior is much more complex and beyond the scope of this article.

Exciton Scattering

An electron and hole in an excitonic bound state execute a correlated motion which can be disturbed by scattering of either partner. This may lead to the ionization of the exciton and the destruction of the correlation or, alternatively, it may change the center of mass momentum of the exciton as a whole. Almost all of the important scattering mechanisms for excitons arise from the charged nature of the electron and hole. Via the Coulomb interaction, excitons can scatter with lattice phonons, other excitons, free carriers, and impurities. Each of these scattering processes has its own regime of dominance, dependent on, for example, temperature or material quality.

When the scattering rate with other particles is much larger than the recombination rate for electrons and holes, a quasi-equilibrium state is reached. The detailed nature of such an interacting electron/hole plasma remains a long-standing open question. The reasons for this can be traced to interplay between the intrinsic Coulomb interactions within the plasma, which give rise to both bound and scattering states, and Pauli exclusion arising from the fermionic nature of the electrons and holes. This leads to a complex phase diagram which encompasses, e.g., electron-hole droplets, the ionized electron-hole plasma, biexciton phase, and the exciton gas which, some suggest, may undergo Bose–Einstein condensation. The parameters of this phase diagram are carrier density, temperature, and the semiconductor material parameters. Interest in this essentially fundamental question has remained high, stimulated by the stream of technological benefits that even partial answers have brought.

In this context it is worth mentioning that there have been two theoretical approaches to exciton physics which each have their advantages and problems. One approach is to treat excitons as bosonic quasi-particles and derive results from Hamiltonians formulated using bosonic operators. This approach has the appeal of simplicity and produces acceptable results in the low-density regime. It does, however, omit a key feature of the constituent particles making up an exciton namely the fermionic nature of

electrons and holes. A more rigorous approach based around electron and hole operators has been followed but this is considerably more complex and numerically involved.

Excitonic Molecules

Just as two hydrogen atoms can lower their total energy by forming a bound hydrogen molecule, so two excitons can make a bound complex which has lower energy than the two isolated excitons. Such complexes are called biexcitons and are mostly important in high quality material and at low temperatures. Their binding energy is less than that of the exciton by a factor of about 0.1–0.3, depending on the ratio of electron to hole effective masses.

Three particle complexes have also been observed consisting of two electrons bound to a hole. These are termed trions or charged excitons. Their binding energy lies intermediate between excitons and biexcitons.

See also: Foundations of Coherent Transients in Semiconductors

Further Reading

- Bastard, G., 1998. *Wave Mechanics Applied to Semiconductor Heterostructures*. New York: Halstead John Wiley.
- Haug, H., Koch, S.W., 1993. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.
- Klingshirn, C.F., 1997. *Semiconductor Optics*. Berlin: Springer.
- Rashba, E.I., Sturge, M.D. (Eds.), 1982. *Excitons*. Amsterdam: North Holland.
- Schmitt-Rink, S., Chemla, D.S., Miller, D.A.B., 1989. Linear and nonlinear optical properties of semiconductor quantum wells. *Advances in Physics* 38, 89–188.
- Ueta, M., Kanzaki, H., Kobayashi, K., Toyozawa, Y., Hanamura, E., 1986. *Excitonic Processes in Solids*. Berlin: Springer.
- Yu, P.Y., Cardona, M., 1999. *Fundamentals of Semiconductors*. Berlin: Springer.

Quantum Wells and GaAs-Based Structures

P Blood, Cardiff University, Cardiff, UK

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

E	energy	$\mathbf{r}$	position vector
E_f	Fermi energy	t	time
E_g	bandgap	T_N	fraction of light transmitted at normal incidence through N quantum wells
F	envelope wavefunction	$\langle \mathbf{x} \rangle$	unit vector along x -direction, similarly for y, z directions
$f(E)$	Fermi–Dirac distribution function	α	optical absorption coefficient in a bulk material, per unit length
g_{2D}	two-dimensional density of states $[\text{energy}]^{-1} [\text{area}]^{-1}$	γ	fraction of light absorbed by a quantum well at normal incidence to the plane of the well
$\hbar$	Planck's constant divided by 2π	λ	wavelength
I_{ph}	optical field intensity	Φ	photon flux (photons per unit time crossing unit area)
$\mathbf{k}$	wavevector	Ψ	electronic wavefunction
k_B	Boltzmann's constant	ν	light frequency
m	electron mass		
n	carrier density, expressed per unit area in quantum well structures		
Q_c, Q_v	heterostructure band offset ratios		

Introduction

A quantum well is a potential minimum within a semiconductor structure which is sufficiently thin to localize charge carriers on a length-scale similar to their de Broglie wavelength which, for an electron in GaAs at room temperature, is about 30 nm. When electrons are localized in this way their electronic and optical properties are determined by quantum mechanical aspects of their behavior which are not apparent in larger-scale structures. The potential well is usually formed by a sandwich of a thin layer of narrow-gap semiconductor between two layers of a wider gap material as depicted in Fig. 1. Typical quantum wells have widths of about 5 nm and the key to their routine production has been the development of advanced epitaxial semiconductor crystal growth techniques such as metalorganic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE). The most significant

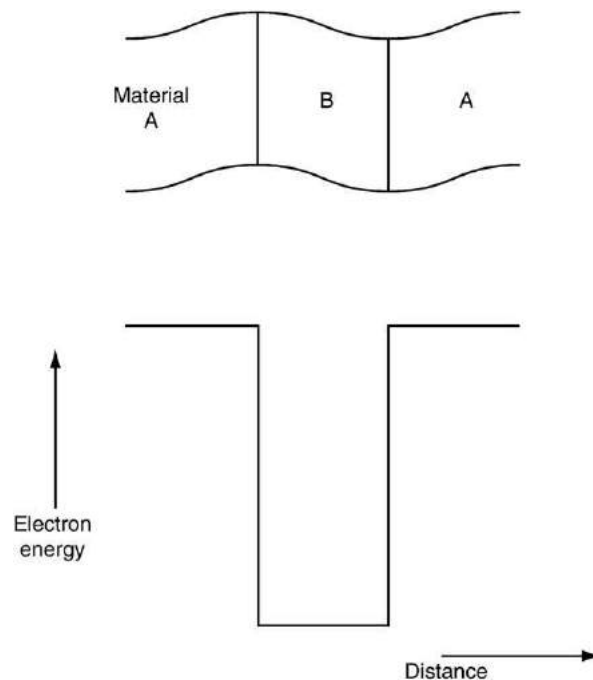


Fig. 1 Electron energy diagram illustrating the formation of a potential well by sandwiching a layer of narrow gap semiconductor material (B) between two wider gap layers (A).

application of these structures has been in opto-electronic devices, especially laser diodes, and in this article I describe how electrons behave in quantum wells with particular reference to optical properties.

The optical properties of semiconductors are usually considered in terms of transitions of electrons between quantum mechanical energy levels, as in the Bohr atom. An alternative description is to consider the spatial displacement of electrons in the oscillating electric field of the electromagnetic radiation. These two descriptions are linked through Schrödinger's equation. For each energy state there is an associated wavefunction which determines the spatial probability distribution of electrons in the state. Consequently, a change in quantum state implies a change in spatial distribution and energy of the electron system. Both approaches are used: here I adopt the viewpoint of transitions between energy states.

We begin with a quantum mechanical description of electrons in a 'bulk' material, as appears in many textbooks. Here electrons are constrained by a three-dimensional potential well of extent given by the sample dimensions, typically millimeters or centimeters and therefore large compared with the de Broglie wavelength. Then we examine what happens when the size of the sample is reduced in one direction: in the limiting case of a quantum well electrons are able to move in only two directions. Quantum confinement is also possible in two or three dimensions to produce a quantum wire or quantum dot, respectively, and it is easy to use the concepts developed in this article to determine the properties of such systems.

Electron States in Bulk Material

We first consider the behavior of an electron in one direction (the x -direction) within a potential well with large dimensions (side length L), then generalize this to three directions (see Fig. 2). The wavefunction ψ_n and energy E_n (measured with respect to the bottom of the well) are given by solutions of Schrödinger's equation, which within the well is:

$$\frac{\hbar^2}{2m} \frac{d^2 \psi_n}{dx^2} = E_n \quad (1)$$

It is assumed that the potential is infinitely deep compared with the energy of the electrons. Within the sample the electrons are described by plane waves of the form $\psi_n = A_n \sin(k_n x)$, where k_n is a wavevector ($= 2\pi/\lambda_n$, where λ_n is the wavelength), and substitution into Eq. (1) gives the corresponding energy eigenvalues as

$$E_n = \frac{\hbar^2 k_n^2}{2m} \quad (2)$$

where m is the mass of the electron. [Substituting $E_n = k_B T$ at room temperature (0.025 eV) and using an effective mass of $0.067m_0$ (for GaAs) in Eq. (2) gives $k_n = 2.1 \times 10^8 \text{ m}^{-1}$ and $\lambda = 30 \text{ nm}$.] Applying cyclic boundary conditions which permit traveling-wave solutions of Eq. (1) the wavelength must satisfy the condition $\lambda_n = nL$, i.e., $k_n = 2\pi n/L$, as illustrated in Fig. 2, where n is an integer. The dimensions of a typical sample are much greater than the 'size' of an electron so n can be a very large number. The energy eigenvalues are therefore given by

$$E_n = \frac{\hbar^2}{2m} \left(\frac{2\pi n}{L} \right)^2 \quad (3)$$

This treatment can be extended by solving Eq. (1) for motion in three orthogonal directions, x , y , and z . In this case electron motion is described by a wavevector $\mathbf{k}$ comprising components k_x , k_y , k_z , each of which satisfies the cyclic boundary condition in the respective direction. We take the sample to be of dimension L in each direction with no loss of generality. Thus (representing unit vectors by $\langle x \rangle$, etc.)

$$\begin{aligned} \mathbf{k} &= k_x \langle x \rangle + k_y \langle y \rangle + k_z \langle z \rangle \\ &= \frac{2\pi n_x}{L} \langle x \rangle + \frac{2\pi n_y}{L} \langle y \rangle + \frac{2\pi n_z}{L} \langle z \rangle \end{aligned} \quad (4)$$

the energy is

$$E = \frac{\hbar^2}{2m} \{k_x^2 + k_y^2 + k_z^2\} \quad (5)$$

and the wavefunction takes the vector form

$$\psi_k(\mathbf{r}) = A_k \sin(\mathbf{k} \cdot \mathbf{r}) \quad (6)$$

$\psi_k^2(\mathbf{r})$ represents the probability of an electron being at the location $\mathbf{r}$. Each allowed electron state obtained by solution of Schrödinger's equation is specified by a unique combination of the numbers (n_x, n_y, n_z) , which take both positive and negative values corresponding to plane waves traveling in positive and negative directions. The Pauli exclusion principle states that only one electron of each spin can occupy a given quantum state so the amplitude of each wavefunction is normalized such that

$$\int \psi_k^2(\mathbf{r}) d\mathbf{r} = \int [A_k \sin(\mathbf{k} \cdot \mathbf{r})]^2 d\mathbf{r} = 1 \quad (7)$$

where the integrals are evaluated over the volume of the sample. Since the amplitude of these solutions is constant throughout the sample electrons may be anywhere in the sample with equal probability. The motion of an *individual* electron of momentum $\hbar \mathbf{k}$ is represented by a wavepacket formed by combination of a number of wavefunctions having similar values of $\mathbf{k}$.

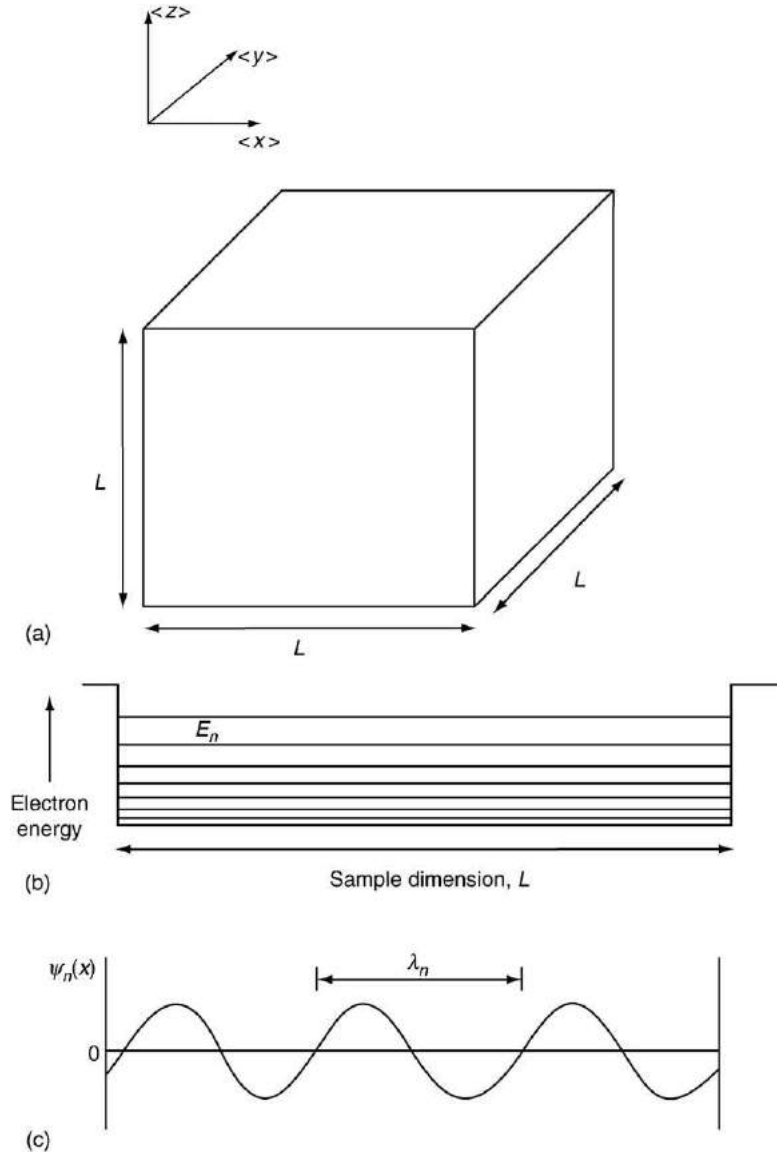


Fig. 2 The properties of electrons in a large cubic sample (a) can be obtained by solving Schrödinger's equation for the potential well formed by the sample in each direction (b). This potential well keeps the electrons within the material. The wavefunctions $\psi(x)$ shown in part (c) satisfy cyclic boundary conditions as described in the text, leading to a series of electron energy levels, E_n , given by Eq. (3), shown in (b) of the figure.

Because the sample dimensions are large the energies of the states allowed by Schrödinger's equation take a series of very closely spaced values for increasing integer values of n as shown in Fig. 2(b). Substituting a typical sample size of $L=1$ into Eq. (3), the $n=1$ and 2 states are separated by only 7×10^{-11} eV in GaAs and these values are very small compared with thermal energies: at room temperature, $k_B T = 0.025$ eV. The allowed values of k and E are discrete and there is a finite but very large number of allowed states in the sample. Since these are very closely spaced we can regard k and E as continuous variables and the allowed energy states form a continuum. We can calculate the number of allowed energy states dN in a small energy interval, dE , and hence determine the density of states in energy $g = (dN/dE)$ (the number per unit energy interval) at any value of energy.

Electron States in Quantum Wells

Now imagine a sample where the length of one side is reduced in the z -direction (L_z), as illustrated in Fig. 3. Gradually the allowed values of k_z become more widely spaced as a consequence of the boundary conditions (Eq. (4)). Eventually the width of the potential well becomes similar to the wavelength of the plane-wave state corresponding to the lowest energy level in the well. In these circumstances electron motion is not possible in the z -direction and the wavefunctions become standing waves with boundary conditions $n(\lambda_n/2) = L_z$ in an infinitely deep potential (where n_z takes only positive values) and the amplitude of the

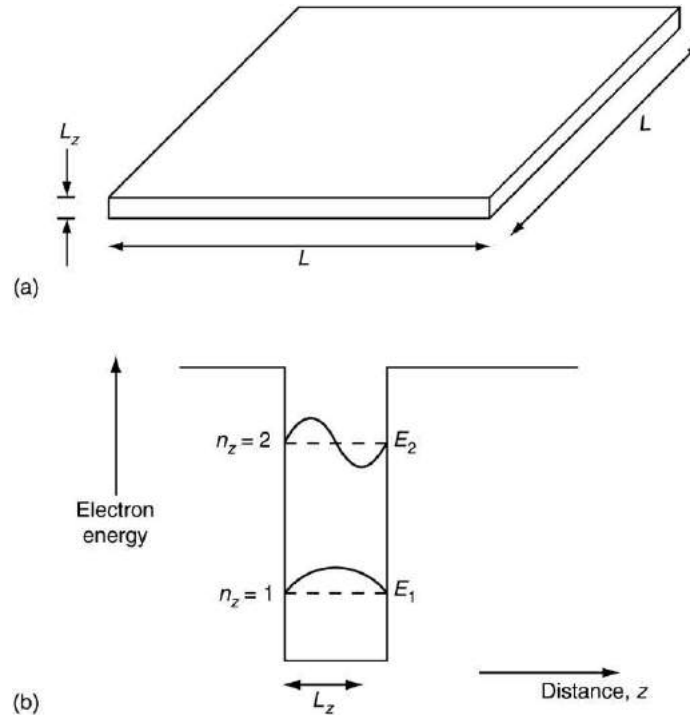


Fig. 3 Electron energy diagram (b) for a sample (a) in which one dimension, L_z , is very small. The electron motion in the z -direction is constrained and the associated energy levels become widely spaced, defined by the integers $n_z=1,2,\dots$ in Eq. (8). The wavefunctions in the z -direction, $\psi(z)$, are also illustrated for the first two electron levels.

wavefunction outside the well must be zero. The lowest energy state, given by Eq. (5) with $n_x=n_y=n_z=1$, no longer lies near the bottom of the potential well. Since the energies associated with motion in the x and y directions remain very small (of order 10^{-10} eV derived earlier) because L_x and L_y are large, the energy of the lowest state is effectively determined by the z -dimension because it is very small, so Eq. (5) gives

$$E_{n_z} = \frac{\hbar^2}{2m} \left(\frac{n_z \pi}{L_z} \right)^2 \quad (8)$$

for the lowest energy state in the well. For $L_z=5$ nm (and $m=0.067$) E_1 is 0.22 eV above the bottom of the well. Eq. (8) shows that this energy can be changed by choice of the well thickness. A simple interpretation of this behavior is as follows. When L_z becomes very small it is no longer possible for the z -component of the wavefunction of the lowest energy state to satisfy the boundary condition of zero amplitude at opposite sides of the sample. The condition can only be satisfied by wavefunctions which have a smaller wavelength, and consequently a higher energy.

Electrons can occupy states defined by all values of (n_x, n_y, n_z) , thus for $n_z=1$ there is a continuum of allowed states at increasing energies corresponding to increasing values of (n_x, n_y) and corresponding to motion in the (x, y) plane (Fig. 4(a)). Since L_x and L_y are both large these states are closely spaced and form a continuum. There is a similar continuum of energy states above the energy levels for $n_z=2, 3$, etc. Thus the energy level diagram is a series of sub-bands, each defined by n_z and with a continuum of states associated with motion in two dimensions in the (x, y) plane, as shown in Fig. 4. The number of continuum energy states in a given sub-band in a small energy interval gives the density of states for both spin directions

$$g_{2D}(E) = \frac{m}{\pi \hbar^2} \quad (9)$$

per unit energy interval per unit sample area.

The density of states is independent of n_z and therefore the same for all sub-bands, and is independent of energy. The density of states function is therefore a series of steps of height given by Eq. (9) at each sub-band energy as illustrated in Fig. 4(b). The density of states per unit area within a given sub-band is independent of L_z as a consequence of dealing with a two-dimensional system.

In real samples the quantum well is not infinitely deep: typically the well depth is in the range 100–400 meV, consequently electrons are not totally confined to the well but are able to penetrate into the barrier material by quantum mechanical tunneling. Solution of Schrödinger's equation for a finite well (treated in many quantum mechanics textbooks) again yields a series of sub-bands, finite in number and with an energy spacing smaller than that given by Eq. (8). In a typical GaAs well the lowest energy state may be about 100 meV from the bottom of the well. The density of states in each sub-band, which is due to the (x, y) motion, remains unchanged. The wavefunction in the z -direction $F(z)$ comprises a sine wave-like standing wave within the well and an

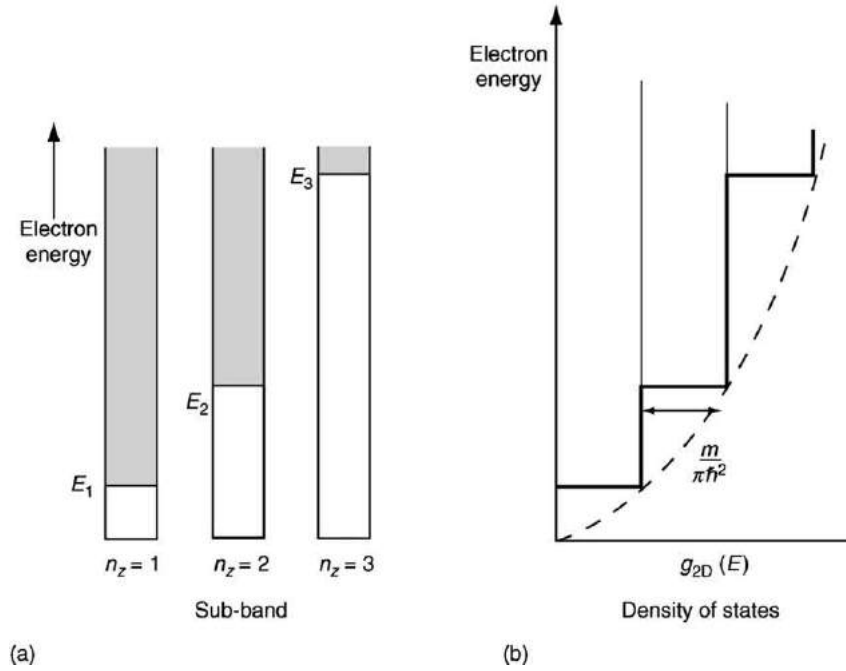


Fig. 4 (a) Electron energy diagram showing the energy states associated with unconstrained motion in the (x,y) plane for each sub-band formed by localizing the electrons in the z -direction, defined by $n_z=1,2,3$, etc. When all these allowed states are summed at any energy and expressed as a number of states per unit energy interval we obtain the density of states function $g_{2D}(E)$ shown in (b). This has a series of steps corresponding to each sub-band edge.

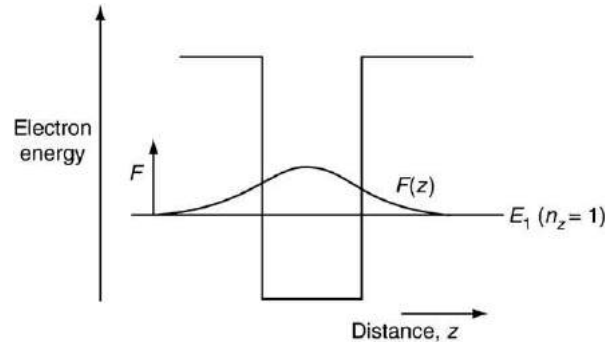


Fig. 5 Illustration of the envelope wavefunction $F(z)$ in the z -direction for an electron localized in the $n_z=1$ sub-band of a well of finite depth. The electron is able to penetrate the barrier by tunneling and this is represented by the decaying part of the wavefunction outside the well.

exponentially decaying wave in the barrier representing tunneling (Fig. 5). The wavefunction is therefore made up of (x,y) plane-wave components and the localized z component and is of the form

$$\Psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}_{xy} \cdot \mathbf{r}_{xy}) F(z) \quad (10)$$

where $\mathbf{k}_{xy}$ and $\mathbf{r}_{xy}$ are the wavevector and position vector in the (x,y) plane. This wavefunction is again normalized according to Eq. (7) such that each state is occupied by one electron of a given spin direction. We see from Fig. 5 that we cannot regard the electron as being localized in the well: the electron distribution in the z -direction is specified by the probability distribution $F^2(z)$.

Occupancy of States in Quantum Wells

The density of states function describes the energy distribution of allowed states in the quantum well. Whether or not particular states are occupied by electrons is determined by the electron concentration and the temperature through the thermodynamic properties of the system. In most cases electron–electron scattering is sufficiently rapid to bring the electron population into an internal equilibrium. Furthermore the interaction between the electrons and the crystal lattice is able to redistribute energy

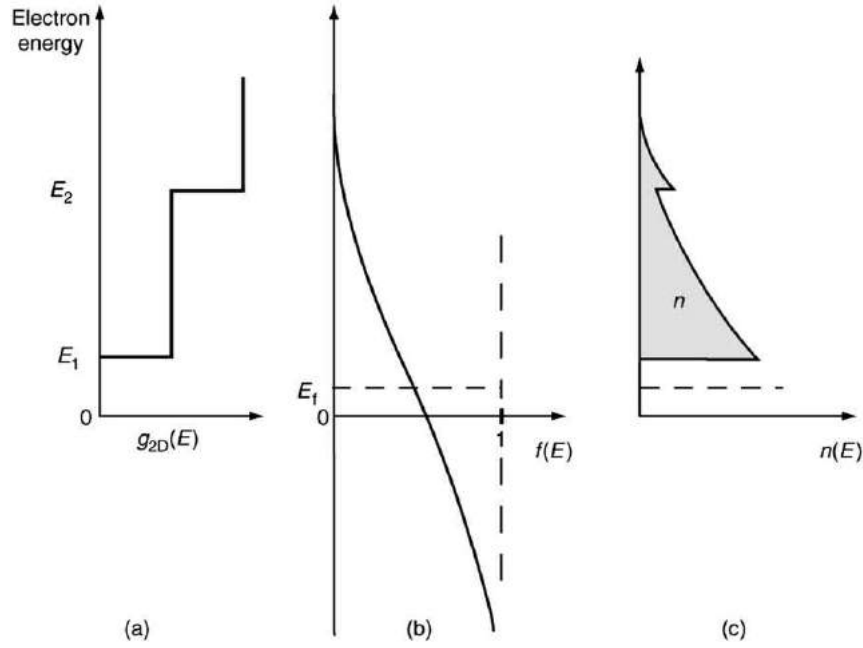


Fig. 6 (a) The density of states function for a quantum well, with the value of $g_{2D}(E)$ on the horizontal axis indicating how the density of states varies with increasing electron energy, (b) is an illustration of the probability that a state is occupied by an electron, showing how $f(E)$ varies with increasing electron energy, given by Eq. (11). The actual density of electrons in a given energy interval at any energy $n(E)$ is given by the product of the density of available states and the probability that the state is occupied as shown in (c). The integral of $n(E)$ over energy gives the total density of electrons given for each sub-band by Eq. (12).

between the two systems so that the lattice and the electron distribution have the same temperature. The electron distribution can therefore be described by Fermi–Dirac statistics for which the probability of a state at energy E being occupied with an electron is

$$f(E) = \frac{1}{1 + \exp\left\{\frac{E - E_f}{k_B T}\right\}} \quad (11)$$

where E_f is the chemical potential (which can be equated with the Fermi level for electrons), T is the lattice temperature and all energies are positive quantities measured with respect to the same arbitrary zero. The energy distribution of electrons in the well, $n(E)$, is then the product of the density of states function, $g_{2D}(E)$, and the occupation probability, $f(E)$, as depicted in Fig. 6. The total number of electrons, n , is the integral over energy. Combining Eqs. (9) and (11), for a single sub-band gives:

$$n_a = \frac{mk_B T}{\pi \hbar^2} \ln \left\{ 1 + \exp\left(-\frac{E_n - E_f}{k_B T}\right) \right\} \quad (12)$$

per unit area, single sub-band.

We cannot specify the carrier density in the well ‘per unit volume’ since the extent of the distribution in the z -direction is not defined because electrons may tunnel into the barrier material. The electron distribution is properly specified as a number per unit area having a probability distribution in the z -direction given by the function $F(z)$. Where more than one sub-band is populated the total electron density is obtained by adding the contributions from each sub-band each given by Eq. (12) with the same Fermi energy and the appropriate sub-band energy. We always use Fermi factors to specify the probability of occupation of a state by an electron. The probability that a state is empty is $(1 - f)$.

Formation of Quantum Wells

A quantum well is formed by sandwiching a narrow gap semiconductor layer (E_{g1}) between two layers of wider gap material (E_{g2}). To avoid the formation of defects and dislocations, which have a deleterious effect on many properties of the structure, the constituent materials must have the same lattice parameter. Fig. 7(a) is an energy band diagram for an n-type double heterostructure with layers of micron dimensions. At each interface there are discontinuities in the conduction and valence bands, ΔE_c and ΔE_v , respectively, which account for the difference in the band gaps of the two materials, ΔE_g . While the value of ΔE_g is known for any system, the manner in which this difference is apportioned between the two band edges is not known a priori and the values of ΔE_c and ΔE_v must be obtained from experiment. The discontinuities are often expressed as fractions Q_c , Q_v of the band gap difference:

$$\Delta E_c = Q_c \Delta E_g, \quad \Delta E_v = Q_v \Delta E_g \quad (13)$$

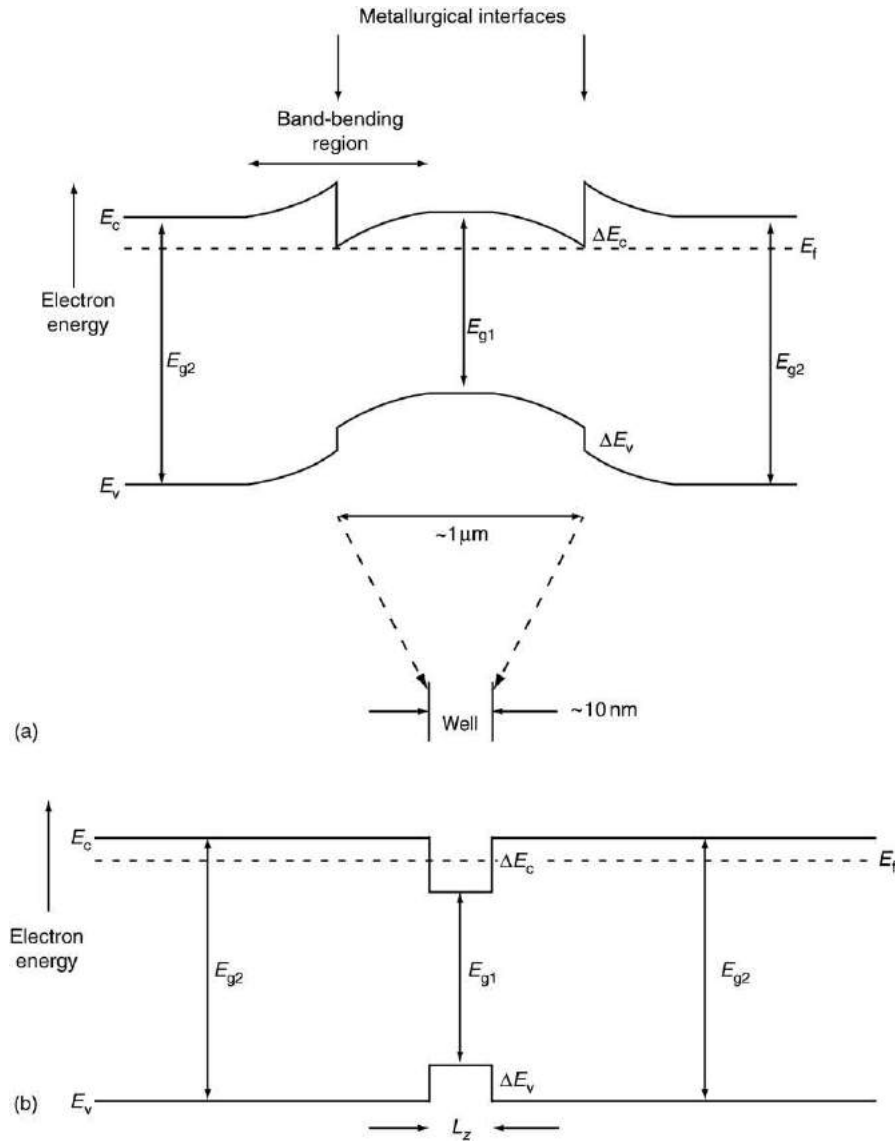


Fig. 7 Electron energy diagrams showing the formation of a quantum well as a very thin double heterostructure made up of materials of band gaps E_{g1} and E_{g2} . When the layers are thick the energy diagram is influenced by band bending at the interfaces as shown in (a). When the narrow gap layer is made thinner than the band-bending distances a rectangular well is formed in the conduction and valence bands with depths determined by the respective band offset as shown in (b). On the small distance scale of the well the bands are flat in each region of the structure.

In equilibrium the Fermi energy is constant across the diagram and there is a band-bending region on each side of the discontinuity such that the conduction band edges return to their equilibrium energy values at large distances from the interface. The extent of these regions is determined by the discontinuity and the doping level and is typically 200–1000 nm. At each single interface a triangular well is formed which localizes electrons and whose precise shape is determined by the doping density, the carrier distribution and discontinuity through simultaneous solution of Schrödinger's equation and Poisson's equation. Single heterointerfaces only localize the majority carrier. Quantum confinement of both types of carrier can be achieved by a very thin narrow gap semiconductor layer.

When the layer of narrow gap material is made thinner than the band-bending regions, the conduction band edge in this layer is not able to regain its bulk equilibrium position and the two triangular wells coalesce as illustrated in Fig. 7(b). When the well is sufficiently thin, say L_z less than 20 nm, the band-bending across this layer is negligibly small and the well becomes effectively rectangular. This well accumulates electrons from the surrounding barrier material and its potential relative to the Fermi level rises such that there is band bending in the barrier material on each side of the well inhibiting further accumulation. Since the band bending in the barrier occurs on a distance of hundreds of nanometers (being determined by the doping density) it is not apparent over the distance scale of the diagram of the quantum well, which is on the order of 10 nm. Consequently, as shown in Fig. 7(b), the well can be drawn as a rectangle to a good degree of approximation. This thin, narrow gap layer localizes both electrons and holes. The form of the well is determined by the respective band discontinuities, ΔE_c and ΔE_v , and the thickness of the layer.

Potential wells formed in this way are the basis for most quantum-confined device structures. Other forms are possible. In certain material combinations ΔE_c or ΔE_v may be negative, meaning that, at the interface, both discontinuities are in the same sense and only one carrier type is confined in the thin narrow gap material. These are known as 'type II' structures. If there is an electric field across the well due to doping or strain (piezoelectric) effects the well becomes triangular. Wells can also be engineered with steps in the potential profile and multiple well systems can be grown in which the states are quantum mechanically coupled. All these variants provide opportunities to engineer the properties of the structure.

One important development of the rectangular quantum well deserves mention. It is possible for the well material to have a different lattice parameter to that of the surrounding barrier material provided that the strain energy can be accommodated elastically within the structure. This means that the strain energy in the layer must be below the energy necessary for formation of dislocations and this translates into an upper limit on the layer thickness (the critical thickness) for a given mismatch. Incorporation of elastic strain is significant because relaxing the lattice match requirement widens the choice of well and barrier materials and because strain modifies the properties of the electronic states in the quantum well and these effects can be used to advantage in the design of devices.

Here we concentrate on the rectangular potential well. Such wells are widely used, particularly in opto-electronic devices, and this structure provides the basic concepts which lie at the heart of all quantum-confined systems. We consider the 'model' lattice-matched quantum well system: a GaAs well bounded by AlGaAs barriers.

GaAs/AlGaAs Quantum Well Structures

As Al is substituted for Ga in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ alloy system the direct band gap increases from 1.424 eV in GaAs ($x=0$) to 3.018 eV in AlAs ($x=1$) (Fig. 8) while the lattice parameter remains unchanged. A quantum well is formed using a narrow GaAs layer sandwiched between barrier layers of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ having a composition chosen to give the desired well depth. The whole structure is grown on a GaAs substrate. The band discontinuity ratio between GaAs and $\text{Al}_x\text{Ga}_{1-x}\text{As}$ is independent of alloy composition (ΔE_c and ΔE_v are constant fractions of ΔE_g as x is varied) with $Q_c=0.66$ and $Q_v=0.33$. For a typical barrier composition of $x=0.3$, $\Delta E_g=0.374$ eV and ΔE_c and ΔE_v are 0.247 eV and 0.127 eV, respectively.

Optical Properties of AlGaAs/GaAs Quantum Wells

General Principles

Fig. 9 illustrates a transition of an electron between an occupied state at E_1 in the $n_z=1$ sub-band in the conduction band and an empty state (i.e., a hole) at E_2 in the $n_z=1$ sub-band in the valence band, resulting in emission of a photon of energy $h\nu=E_1-E_2$ as required by energy conservation. This is the process of luminescence (or spontaneous emission), which requires external excitation such as illumination (photoluminescence) or biasing a p-n junction (electroluminescence, as in a light-emitting diode).

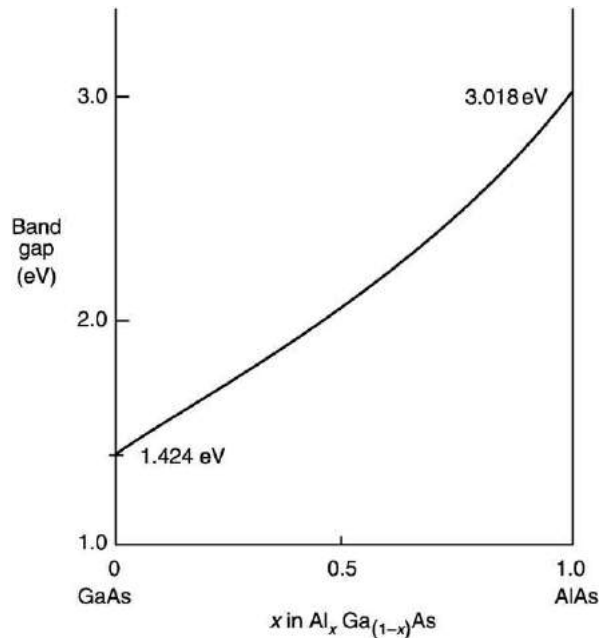


Fig. 8 Variation of the direct band gap of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ with Al content (x).

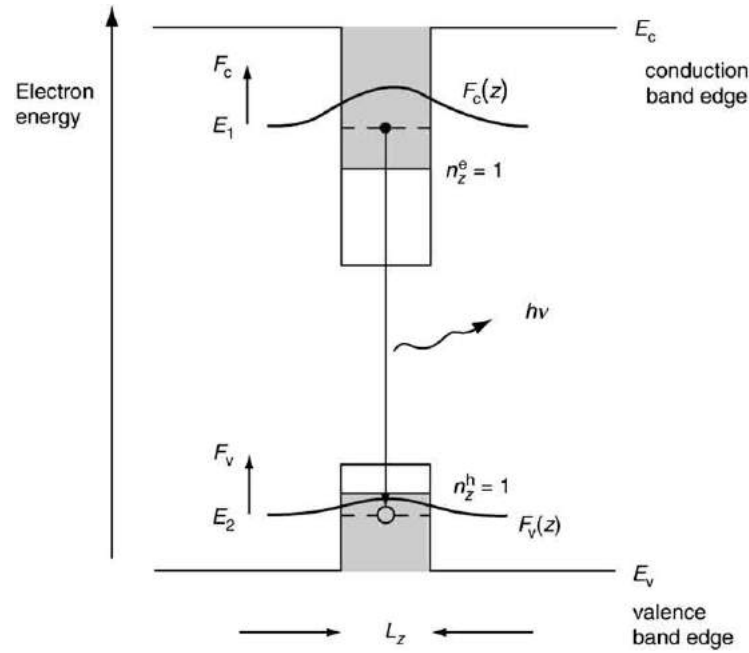


Fig. 9 Illustration of a downward optical transition of an electron from a state in the $n_z=1$ conduction sub-band to an empty state (hole) in the $n_z=1$ valence sub-band resulting in emission of a photon of energy $h\nu=E_1-E_2$. The spatial probability distribution of the electron and hole is specified by their respective envelope functions $F_c(z)$ and $F_v(z)$ and the probability of the transition occurring is proportional to their overlap integral, Eq. (14).

Alternatively, incident light may promote an electron from an occupied state in the valence band to an empty state in the conduction band with the consumption of the energy of a photon. This is the process of optical absorption.

At typical temperatures the electrons and holes are at energies near to their respective band edges so the photon energy $h\nu$ is close to the separation of the $n_z=1$ sub-bands. Since these energies are determined by the well width it is possible to change the photon energy for emission or absorption by changing the well width. As the well width is reduced the energy separation increases and the wavelength of light emitted by the structure is reduced. This ability to engineer the emission wavelength by simply adjusting the well width, without changing the chemical composition of the constituent materials, is one of the major attractions of quantum-confined structures. The shortest wavelength possible for a given material combination is determined by the band gap of the wide gap barrier material alongside the well.

The strength of the luminescence signal or the absorption is determined by the rates at which the electronic transitions occur. In general these rates are determined by the following intrinsic factors.

1. *The quantum mechanical probability of an electron making a transition between a full and an empty state.* This depends upon the 'matrix element' for the transition, which is a fundamental property of the well material and the overlap of the envelope functions. The electron and hole in the initial and final states must be in the same region of space and if the spatial distributions of electrons and holes, determined by F_c and F_v , do not overlap the probability for an electron to make a transition to the hole state is very low. The overlap is almost complete in a rectangular well (see Fig. 9), but is only partial in a triangular well. The transition probability is proportional to

$$I_{\text{overlap}} = \left\{ \int F_v(z) F_c(z) dz \right\}^2 \quad (14)$$

2. *The probability of occupation of initial and final states by electrons, specified by the Fermi factors (Eq. (11)).* The initial state must be occupied and the Pauli exclusion principle prohibits an electron entering a final state which is already occupied. Thus for a downward transition resulting in emission of light the rate is proportional to $f_c(1-f_v)$.
3. *The density of initial and final states (Eq. (9)).* The more states there are within a given small energy interval, the greater the emission rate per unit energy.
4. *Photon distribution.* For processes such as optical absorption, which are 'induced' by the presence of a photon, the transition rate is also proportional to the photon density in the region of the well. Because the electrons are not wholly localized in the well, and because the spatial extent of the optical field is larger than that of the carriers, the coupling between the photon field (intensity $I_{\text{ph}}(z)$) and the carrier probability distributions is expressed as

$$I_{\text{coupling}} = \frac{\left\{ \int F_v(z) [I_{\text{ph}} z]^{1/2} F_c(z) dz \right\}^2}{\int I(z) dz} \quad (15)$$

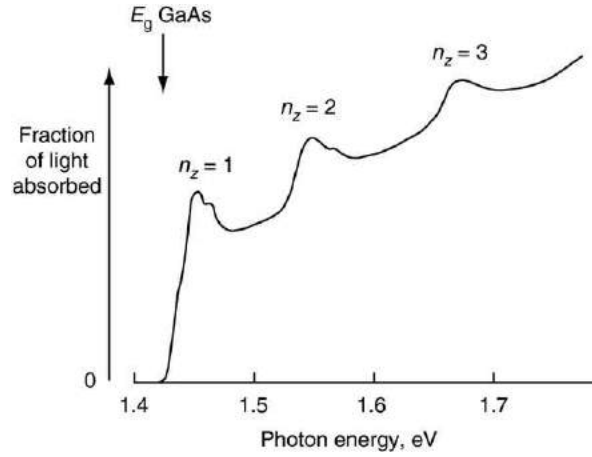


Fig. 10 Absorption spectrum of a multiple quantum well structure having 10 nm GaAs wells measured at room temperature. The absorption edges corresponding to transitions between $n_z=1, 2$, and 3 pairs of sub-bands can be identified. The peaks are due to absorption by formation of excitons.

5. *Optical mode density.* The emission rate also depends upon the density of optical modes available to receive the recombination radiation. In most situations the emission occurs into the cavity of free space which is very large compared with the wavelength of the radiation so the modes are very closely spaced and the cavity can accept radiation at any transition energy (this is the well-known Rayleigh-Jeans theory of cavity radiation). However, in some special situations the cavity has dimensions comparable with the wavelength of the radiation (microcavities) and then the modes are widely spaced. In such a microcavity, the optical cavity effectively controls the electronic transitions.

Optical Absorption

Fig. 10 is an optical absorption spectrum for a GaAs quantum well measured with light incident normal to the plane of the quantum wells. A series of steps can be seen which correspond to absorption at increasing photon energy between increasing orders of sub-bands, $E_{1e} \rightarrow E_{1h}$, $E_{2e} \rightarrow E_{2h}$, etc. Normally $n_z=1$ to $n_z=2$ transitions are forbidden in rectangular wells. At each step there is a pair of peaks (not always resolved) and above each step the absorption is roughly constant, following the step-like density of states function sketched in **Fig. 4**. The peaks are due to the formation of excitons which are weakly bound electron-hole pairs. In bulk materials excitons are only observed at low temperatures whereas in quantum wells the localization increases the binding energy (E_{bx} , typically 6–8 meV) such that they are observed at room temperature. The excitonic peak is at an energy E_{bx} below the sub-band separation. There are in fact two different kinds of electrons in the valence band with different masses so there are pairs of $n_z=1, 2, 3, \dots$ sub-bands due to the different confinement energies for the different valence electrons. Spectra of this kind provided the first evidence for the distinctive sub-band structure of quantum well systems produced by carrier confinement. Absorption measurements provide data on the energy spacing of the sub-bands in the well, and it is possible to determine the well width and depths, from which the band offsets can also be determined.

The strength of the absorption is also of interest because it provides a measure of the optical matrix element if the Fermi factors f_c and f_v are known. If the incident optical beam is very weak such that the absorption process excites very few electrons then the system remains close to thermal equilibrium so the upper states are empty ($f_c=0$) and the lower states are full ($f_v=1$).

In *bulk material* the change in photon flux over a small distance Δx is proportional to the flux Φ and the distance traveled

$$\Delta\Phi = -\Phi\alpha\Delta x \quad (16)$$

where α is the absorption coefficient which has units L^{-1} . This results in an exponential decrease of photon flux with distance traveled through the material

$$\Phi(x) = \Phi_0 \exp(-\alpha x) \quad (17)$$

For a quantum well structure with light incident in the z -direction perpendicular to the plane of the quantum well, the fraction of light absorbed by a single well by transitions between a single sub-band pair (e.g., transitions between the $n_z=1$ conduction and valence sub-bands) is independent of the thickness of the well because the density of states for a single sub-band is independent of well thickness. The strength of the absorption cannot be expressed by an absorption coefficient but by the fraction of light absorbed per well:

$$\Delta\Phi = -\gamma_w \Phi \quad (18)$$

This arises because the quantum well is effectively a sheet in the z -direction: changing its thickness does not change the density of states per unit area (**Eq. (9)**) and we cannot specify the electron concentration along the z -direction. This behavior also occurs because the well is much thinner than the wavelength of the light and it is not possible to specify the variation of photon flux on a

distance scale smaller than the wavelength. The barrier material surrounding the well has a wider band gap and does not absorb at the photon energies absorbed by the well. The well thickness does affect the sub-band spacing and at a fixed photon energy the absorption increases as the well width is decreased as light is absorbed by transitions between more sub-bands. Eq. (18) applies individually to each sub-band because the density of states is the same for each sub-band.

In a multiple well system, the fraction of light transmitted through N independent wells is

$$T_N = \{1 - \gamma_w\}^N \quad (19)$$

so the fraction of light absorbed by such a system is

$$\frac{\Delta\Phi}{\Phi} = - \left[1 - \{1 - \gamma_w\}^N \right] = -N\gamma_w \quad (20)$$

which is proportional to the number of wells when $\gamma_w \ll 1$.

The fraction of light absorbed by transitions between a single pair of sub-bands for a rectangular GaAs quantum well is predicted to be about 0.01, and a similar value has been obtained from absorption measurements. In a rectangular well the envelope function overlap is complete; however, in a triangular well of the same material the absorption is reduced by the smaller overlap of the envelope functions.

Photoluminescence and Spontaneous Emission

We turn now to discuss the optical emission from quantum wells as a result of external excitation. An external light source with photon energy above the effective band gap produces excess electron–hole pairs by excitation of electrons from the valence band. These rapidly lose their excess energy and thermalize to the sub-band edges to take up Fermi energy distributions. The electrons subsequently recombine to vacant valence band states by spontaneous emission. The low energy edge of this spectrum corresponds to the sub-band separation and the shape of the spectrum at higher energies corresponds to the thermal distribution of carriers in the bands. At very low excitation intensities where the excess electron population is small the luminescence may be due to recombination of electrons within an exciton and occur at an energy E_{bX} below the sub-band separation. Generally the sub-band separation is several $k_B T$ so only the lowest sub-band is populated and photoluminescence is not seen from higher sub-bands. In this respect absorption measurements provide a more complete characterization of the structure.

While the energy position of features in the photoluminescence spectrum provides useful information, it is difficult to use the strength of the photoluminescence signal to provide quantitative information about the rate of optical transitions within the material because the excitation power and volume within the sample must be known and a known fraction of emitted light must be collected and measured with a calibrated detector. The relative intensity of photoluminescence under standard conditions does provide comparative information and the linewidth of the spectrum at low intensity provides a qualitative indication of ‘quality’ (e.g., well width variations in the structure). A quantitative determination of the internal recombination rate can be obtained by measuring the decay of the luminescence signal following short-pulse excitation. Fuller discussion can be found in books dealing with the characterization of semiconductor structures.

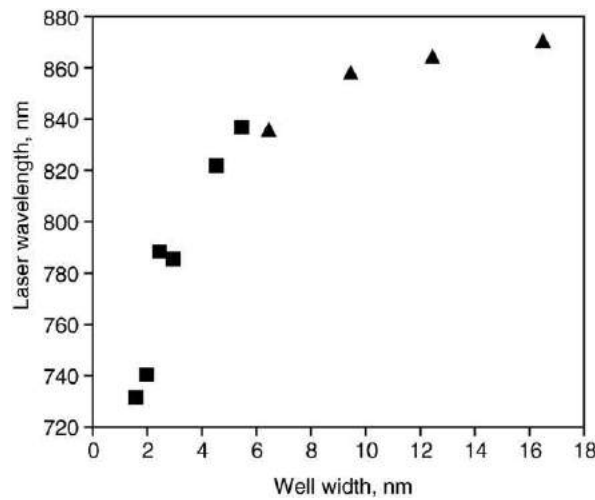


Fig. 11 Measured values of the emission wavelength of GaAs/AlGaAs quantum well lasers as a function of well width, showing how the wavelength can be engineered by choice of the well width. Data from: squares: Woodbridge K, Blood P, Fletcher ED and Hulyer PJ (1984) Short wavelength (visible) GaAs quantum well lasers grown by molecular beam epitaxy. *Applied Physics Letters* 45: 16–18; and triangles: Chen HZ, Ghaffari A, Morkoç H and Yariv A (1987) Effect of substrate tilting on molecular beam epitaxial grown AlGaAs/GaAs lasers having very low threshold current densities. *Applied Physics Letters* 51: 2094–2096.

Concluding Remarks

We have shown that an understanding of the electronic structure of quantum wells follows naturally from quantum mechanical treatments of electrons in bulk materials. When one of the dimensions of the structure is reduced to be similar to the de Broglie wavelength, quantization of the electron motion in this direction becomes apparent. The sub-bands which are formed have a profound effect on the electronic and optical properties when their spacing exceeds thermal energies such that only one band is significantly populated with carriers. The density of states function is a series of steps, each corresponding to a sub-band, and this produces characteristic step features in the optical absorption spectrum. The effective band gap of the structure is controlled by the width of the well. Quantum confinement also increases the binding energy of excitons compared with bulk materials and under low excitation conditions absorption and luminescence spectra are dominated by excitonic features even at room temperature.

One of the most successful application areas of quantum well structures has been in laser diodes. Fig. 11 shows experimental data for laser emission wavelengths as a function of well width, showing the dramatic reduction which can be achieved simply by changing the width of the quantum well. Furthermore the step density of states function, characteristic of quantum wells, is one of the fundamental reasons for the reduction in threshold current of quantum well lasers relative to bulk GaAs devices.

We have considered GaAs because it is a model material system for production of quantum well structures. The concepts developed in this chapter are, however, quite general and can be applied to other material systems, other forms of quantum well and quantum wires and dots.

See also: Cavity QED in Semiconductors

Further Reading

- Bastard, G., 1988. *Wave Mechanics Applied to Semiconductor Heterostructures*. New York: Les Editions de Physique, Halstead Press.
- Bimberg, D., Grundmann, M., Ledentsov, N.N., 1999. *Quantum Dot Heterostructures*. Chichester, UK: Academic Press.
- Blood, P., 1991. Heterostructures in lasers. In: Morgan, D.V., Williams, R.H. (Eds.), *Physics and Technology of Heterojunction Devices*. Stevenage, UK: Peter Peregrinus.
- Blood, P., 2000. On the dimensionality of optical absorption, gain and recombination in quantum-confined structures. *Journal of Quantum Electronics* 36, 354–362.
- Blood, P., Orton, J.W., 1992. *The Electrical Characterisation of Semiconductors: Majority Carriers and Electron States*. London: Academic Press.
- Casey Jr, H.C., Panish, M.B., 1978. *Heterostructure Lasers*. San Diego, CA: Academic Press.
- Dingle, R., Gossard, A.C., Weigmann, W., 1975. Direct observation of superlattice formation in semiconductor heterostructures. *Physical Review Letters* 34, 1327–1330.
- Eisberg, R., Resnick, R., 1985. *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, second ed. New York: John Wiley.
- Hook, J.R., Hall, H.E., 1991. *Solid State Physics*. Chichester, UK: John Wiley.
- Kelly, M.J., 1995. *Low-Dimensional Semiconductors*. Oxford, UK: Clarendon Press.
- Orton, J.W., Blood, P., 1990. *The Electrical Characterisation of Semiconductors: Measurement of Minority Carrier Properties*. London: Academic Press.
- Perkowitz, S., 1993. *Optical Characterisation of Semiconductors*. London: Academic Press.
- Wolfe, C.M., Holonyak Jr, N., Stillman, G.E., 1989. *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice Hall.

Recombination Processes

PT Landsberg, The University of Southampton, Southampton, UK

© 2005 Elsevier Ltd. All rights reserved.

Introduction

In studies of recombination there occur considerable complications; many are associated with interactions between electrons, electrons and phonons, excitons, bi-excitons, impurity centers, etc. Most of these are not required in the present exposition. Modern work normally assumes that the basics of recombination physics are understood and the present exposition offers an appropriate outline.

Basic Assumptions

Electron–Electron Interactions for Electrons in Bands

The recombination problem in semiconductors is greatly complicated by the interaction of the electrons with each other. This allows one to speak only of the quantum states of the semiconductor crystal as a whole. However, as in metals, so also in semiconductors, a simplified picture is successful. In this, the electron interactions, and other interactions, are first neglected, but are later taken into account as a perturbation. Thus, the electrons are first treated as if they can move through each other; the fact that they collide and deflect each other by virtue of their Coulomb interaction is treated as a perturbation. The transitions are still described within the framework of the single-particle states of the unperturbed problem.

The Effect of Electron–Boson Interactions

Our first approximation is to neglect most (but not all) electron interactions. Later we take into account the two-electron transitions that arise. This two-particle recombination process is referred to as Auger recombination and its inverse as the generation of current carriers by impact ionization. In addition, electrons interact with the radiation and lattice fields and emit or absorb photons or phonons. These electron–boson interactions result in transitions of single electrons in an energy band scheme. We shall attach a superfix ‘S’ (for single-electron) to the recombination coefficients for such processes. These are then denoted by B^S or T^S depending whether only bands are involved or whether traps are also involved. They are illustrated in Fig. 1, where the solid horizontal lines represent the conduction and valence band edges and the dashed line refers to traps. Note that the two-electron transitions do not have the superfix ‘S’, while n and p refer to the electron and hole concentrations; N_0 and N_1 are the concentration of unoccupied and occupied trap states, respectively.

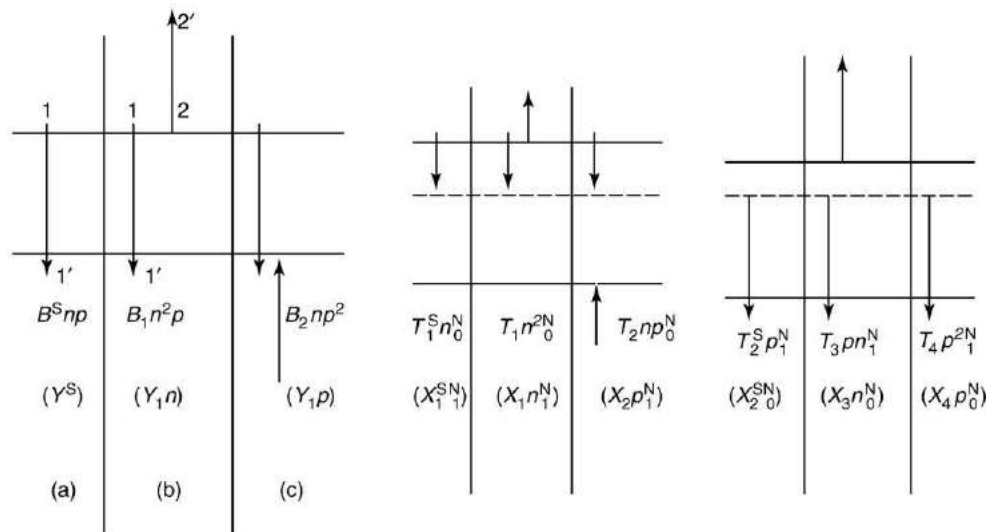


Fig. 1 Definition of recombination coefficients. Transition rates per unit volume are stated with each process, and, in brackets, for the reverse process. Thus B_1 , B_2 , T_1 – T_4 refer to Auger processes; B^S , T_1^S , T_2^S refer to single-electron recombination; Y^S , X_1^S , X_2^S refer to carrier generation processes; Y_1 , Y_2 , X_1 – X_4 refer to impact ionization processes. Arrows indicate transitions made by electrons.

Electron–Electron Interaction in Traps

Electron–electron interactions can at least formally be taken into account in connection with electrons trapped in a center: the spectrum is a function of the number, $r (= 1, 2, 3, \dots, M)$, of electrons captured. The ‘irremovable’ electrons can be included with the ion core. Given that the center captured r electrons (say), it can still be in a variety of quantum states, and they will be denoted by the symbol ℓ , yielding a set of quantum states (ℓ, r) for a center. The energy of such a state, divided by kT to make it dimensionless, is denoted by $\eta(\ell, r)$, yielding the canonical partition function

$$Z_r = \sum_{\ell} \exp[-\eta(\ell, r)] \quad (1)$$

The location of a center in space will here be considered to be of no significance.

Centers can capture several electrons. This brings in a need for the chemical potentials (or Fermi levels) for r -electron centers. They are here denoted by γ (when divided by kT) and the suffix ‘eq’ denotes an equilibrium value. Thus, the equilibrium probability of finding an r -electron center in state ℓ is given by:

$$P(\ell, r)_{\text{eq}} = \frac{\exp[r\gamma_{\text{eq}} - \eta(\ell, r)]}{\sum_{s=0}^M (\exp s\gamma_{\text{eq}}) Z_s} \quad (2)$$

One can identify the dominant energies from optical or electrical experiments.

To understand the properties of these centers, suppose we can arrange for the equilibrium Fermi level to rise from the valence band to the conduction band. At first practically no electrons are captured ($r=0$). As the Fermi level rises the states corresponding to $r=1$ begin to appear. The states ℓ of the center which correspond to the ground states in equilibrium are always more highly populated than the states ℓ corresponding to excited states, so that we can often confine attention to them. As the Fermi level rises, the ground states for $r=2$ become more important, and there may now be hardly any centers which have captured fewer electrons. If a larger number of electrons cannot be captured, by the time the Fermi level reaches the conduction band, then states of the center with $r>2$ can (normally) be neglected as unstable. That this is a satisfactory picture is our second approximation.

Assumptions for Nonequilibrium Statistics

The now much simplified recombination problem is still complex because of the many states available to electrons in bands and centers. A key simplification arises from the fact that it is often possible to talk about a small number of groups of quantum states: let $I (= 1, 2, \dots)$ label quantum states in a group i , let $J (= 1, 2, \dots)$ label quantum states in a group j , etc. Within each group it is supposed that the transitions are much more rapid than they are between groups. In that sense electrons in each group are in equilibrium among themselves. Their quasi-equilibrium is then characterized by what is called a quasi-Fermi level. The dimensionless version, obtained by dividing it by kT , is denoted by γ_i, γ_j , etc., for groups i, j , etc. That this is reasonable is our third assumption. Thus γ_c, γ_v are quasi-Fermi levels which refer to the quasi-equilibrium of the conduction and valence band, respectively, and γ_i to an i -electron center. Groups of states with different quasi-Fermi levels are not in equilibrium with each other.

Recombination problems can now be discussed by neglecting transitions within one (quasi-equilibrium) group, because they proceed at exactly the same rate as their reverse transitions. The number of transition types to be considered is thereby greatly reduced. Away from equilibrium we shall adopt Eq. (3) instead of (2) in order to allow for distinct quasi-Fermi levels:

$$P(\ell, r) = \frac{\exp[r\gamma_r - \eta(\ell, r)]}{\sum_{s=0}^M (\exp s\gamma_s) Z_s} \quad (3)$$

Eq. (3) is not always correct, but is an approximation arising from the quasi-Fermi level assumption. Among the improved theories are, for example, the ‘cascade’ theories.

The assumption of a quasi-Fermi level for each state of charge of a center implies that all the excited states of an r -electron center are populated according to Eq. (2).

The Main Recombination Rates

General Results and The Two-Band Case

Recombination and its converse, generation, consist of a transition of an electron from one state to another. The observed rate is the net rate of recombination and is the algebraic sum of the recombination and generation processes. During these processes both total energy and total momentum must be conserved. This is achieved by creation or absorption of photons, or phonons, excitation of secondary electrons, etc.

The transition probability per unit time, S_{IJ} , for a single-electron transition from state I in a band to state J requires that state I should be occupied, with probability p_I say, and state J should be vacant with probability q_J say. The general expression for the average rate of the transition from I to J then takes the form $p_I S_{IJ} q_J$. For the reverse process electron state J has to be occupied and

state I vacant. Thus the rate of this reverse process is $p_I S_{JI} q_I$. The net recombination rate per unit volume of the process I to J can then be written as:

$$u_{IJ} = (p_I S_{IJ} q_J - p_J S_{JI} q_I) V^{-1} \quad (4)$$

By the principle of detailed balance, this expression vanishes at equilibrium. If one puts $X_{JI} \equiv S_{IJ} p_I q_J / S_{JI} p_J q_I$, one has $u_{IJ} = p_J S_{JI} q_I (X_{JI} - 1) V^{-1}$. In equilibrium $X_{JI} \rightarrow (X_{JI})_{\text{eq}} = 1$, and the recombination rate is zero.

One can say a little more if one assumes that states I and J are in conduction or valence bands, each with a quasi-Fermi level (divided by kT to make it dimensionless). If these are denoted by γ_e and γ_h , then $p_I = [1 + \exp(\eta_I - \gamma_e)]^{-1}$ for a conduction band. For the total recombination rate per unit volume between the bands i and j one finds:

$$u_{ij} = \left[\sum_{I \in i} \sum_{J \in j} p_I S_{JI} q_I \right] \left[\exp\left(\frac{|e|\phi}{kT}\right) - 1 \right] V^{-1} \quad (5)$$

where ϕ is essentially the difference between the quasi-Fermi levels: $\gamma_e - \gamma_h$. The first factor depends on the bands involved and it has been assumed that the transition probability ratio S_{IJ}/S_{JI} is independent of the excitation.

A transition rate, when multiplied by the charge of current carriers, is a current, and when divided by the area of the surface involved, is a current density. If i and j denote the states of the conduction and valence bands of a semiconductor, excitation independence may often be assumed, and one then expects a current density proportional to $\exp(|e|\phi/kT) - 1$. This is characteristic of the current through pn junctions, metal semiconductor junctions, etc. In these configurations the Fermi level difference between the ends of the device determines the voltage across it. When radiation is involved, however, excitation dependence of some of the parameters introduced above (e.g. S_{JI}) tends to spoil this simple story.

The Case of Defects

So far we have considered only two bands. When a trap is involved matters are rather different. Because of the interactions among the electrons on a center it is not possible to talk of the same level being occupied or vacant. Consequently, identification of forward and reverse processes in terms of levels becomes impossible. Instead one deals with a center, say an r -electron center, as a whole; we then need the probability that a given center is an r -electron center. For example, the capture of an electron converts an $(r-1)$ -electron center into an r -electron center. Thus the p 's and q 's must be replaced by more complicated expressions. It is convenient to denote u_{ij} by u_{cv} in the simple two-band case and its structure is given (with a sign change) by a recombination coefficient:

$$u_{cv} = B^S n p [1 - \exp(\gamma_h - \gamma_e)] \equiv B^S n p - Y^S \quad (6)$$

using Fig. 1(a). Here, n and p are the electron and hole concentrations. Auger effects in Figs. 1(b) and 1(c) can also be included, at least formally. Then B^S has to be replaced in Eq. (6) by $B^S + B_1 n + B_2 p$. Analogous replacements are found for recombination processes involving defects.

In the simplest case of one type of localized defect one finds a steady-state recombination rate per unit volume of the form:

$$u = \frac{np - (np)_0}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \quad (7)$$

Here τ_{n0} , τ_{p0} are parameters with the dimension of time and p_1 , n_1 are parameters with the dimension of concentration. The recombination increases with the defect concentration, which is actually in the denominators of τ_{n0} and τ_{p0} . This is an old and much used result associated with the names of W Shockley and W T Read.

Radiative Transitions

Spontaneous and Stimulated Emission

Quantum theory was initiated by Planck's law for black-body radiation at temperature T . This gives the spatial energy density as a function of frequency ν in this system as

$$f(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp(h\nu/kT) - 1} \quad (8)$$

For low frequencies one finds the Rayleigh-Jeans law which makes (8) proportional to kT in agreement with the classical equipartition theorem. For high temperatures (8) diverges, as required.

The result (8) may also be obtained by writing

$$N = [A + Bf(\nu, T)] N_u \quad (9)$$

for the emission rate of photons by atoms, where N_u is the number of atoms in the upper of two states which are separated by the energy $h\nu$. The first term on the right-hand side is due to the normal decay of an excited state ('spontaneous emission'). The second term refers to additional emission ('stimulated emission') induced by the radiation of frequency $h\nu$ itself.

The first factor in (8) is due to the density of states and the second factor is due to the fact that the energy is considered. If one considers the number of photons of energy $h\nu$ at temperature T one comes up with

$$N_v = \left[\exp \frac{h\nu - \mu}{kT} - 1 \right]^{-1} \quad (10)$$

(Bose–Einstein distribution)

Here μ is a possible chemical potential of the radiation which is non-zero only in nonequilibrium situations such as in a semiconductor laser.

In a pn junction the two bulk materials several diffusion lengths away from the junction are approximately in equilibrium even if a modest current is flowing. On one side one has then just one quasi-Fermi level, say μ_i , and on the other side one has just one quasi-Fermi level, say μ_j . Then the difference

$$\mu \equiv \mu_i - \mu_j = q\phi \quad (11)$$

corresponds to the applied voltage ϕ . This is in agreement with Eq. (10), in that $\phi=0$ implies no current and hence thermal equilibrium is possible.

Statistical mechanics teaches one the rule that, in equilibrium, occupation probabilities of individual quantum states are always less for states of higher energies. This ensures that the total energy of a quantum system converges, the occupation probability p being a kind of convergence factor. However, if one is away from equilibrium, the above rule can be suspended and one can have population inversion.

For $h\nu = \mu$ there is trouble with (10) because the steady-state photon occupation diverges. This does not correspond to a ‘death ray’, but is the result of imperfect modeling; for example, the leakage of photons from the cavity may have been neglected. A formula of type (10) is also needed in connection with solar cells.

Donor–Acceptor Pair Recombination

A striking demonstration of radiative donor–acceptor transitions in GaP at low temperature (1.6 K) was revealed by sharp lines (Fig. 2) at photon energies $h\nu_i$ given by

$$h\nu_i = E_G - E_A - E_D + e^2/\epsilon R_i \quad [-E] \quad (12)$$

where R_i is the distance between the i th-order nearest neighbors.

The discrete nature of the peaks is due to the fact that the impurities involved settle in general on lattice sites so that only definite separations R_i are possible. The energy gap is E_G ; the value of $E_A + E_D$ can be inferred from the experimental lines by extrapolation to $R_i \rightarrow \infty$. The energy term E is sometimes neglected.

The ZnS phosphors were the first materials in which donor–acceptor radiation was hypothesized. However, it is hard to control its stoichiometry and its impurity content, and it has relatively large carrier mass and hence relatively large impurity activation energies. In such cases spectra such as those shown in Fig. 2 are hard or impossible to obtain. This applies in general to many II–VI compounds.

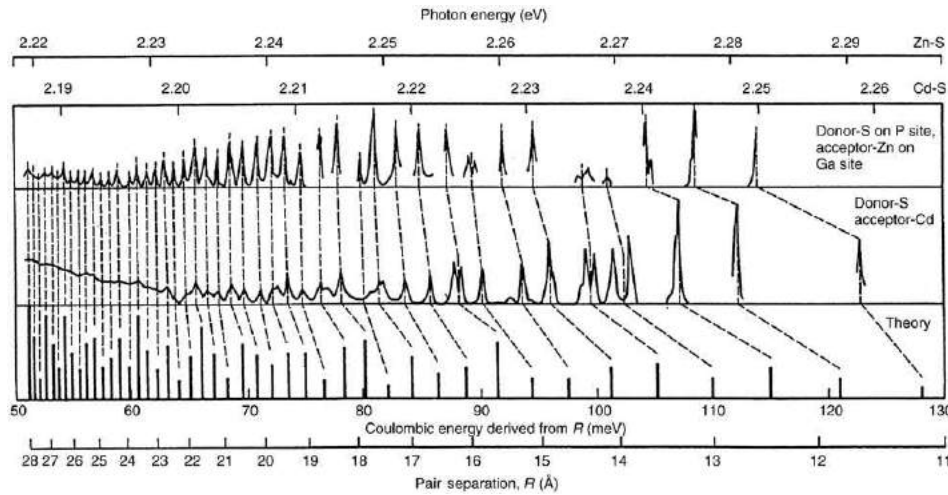


Fig. 2 Comparison of the positions and intensities of the sharp line spectra at 1.6 K corresponding to both ZnS and CdS acceptor–donor pairs with the predicted pair distribution. The lower scales show the pair separation (R) and the Coulombic energy derived from R . The emission energy scales for the two measured spectra are shown at the top. Reproduced with Gershenson M, Logan RA, Nelson DF, *et al.* (1968) *Proceedings of the International Conference on Luminescence*. Budapest, Hungary: Akademia Kiada.

Quantum Efficiency

The quantum efficiency is the radiative recombination as a fraction of the total recombination. It can also be regarded as the average number of electrons produced per incident photon. Consider an intrinsic semiconductor, i.e., one in which the electron and hole concentrations are equal. Then with the notation of Fig. 1 the radiative band-band recombination rate is $B^s n^2$. The nonradiative rate is $(B_1 + B_2)n^3$. We shall add another nonradiative rate, An say, proportional to the injected carrier density $n \sim p$. The quantum efficiency is then

$$\eta = \frac{B^s n^2}{An + (B_1 + B_2)n^3 + B^s n^2} \quad (13)$$

Traps are here neglected, and one finds a maximum

$$\eta_{\max} = \left\{ 1 + \frac{2}{B^s} A^{1/2} (B_1 + B_2)^{1/2} \right\}^{-1} \quad (14)$$

For a high-quality epitaxial AlGaAs/GaAs double heterostructure of great purity we may take

$$\begin{aligned} A &\sim 0.5 \times 10^6 \text{ s}^{-1}, B \sim 10^{-10} \text{ cm}^3 \text{ s}^{-1}, \\ B_1 + B_2 &\sim 10^{-29} \text{ cm}^6 \text{ s}^{-1} \end{aligned} \quad (15)$$

whence $n_1 \sim 2 \times 10^{17} \text{ cm}^{-3}$, $\eta_{\max} \sim 0.96$.

Detailed Balance

It is clear that the radiative recombination rate from a material should be obtainable in terms of its optical 'constants', the absorption coefficient $\alpha(\lambda)$ and the refractive index $\mu(\lambda)$. Both are functions of the wavelength. This connection can be formalized by comparing the optical absorption process with the emission process and this can be done by using the principle of detailed balance. This equates the rate of disappearance by absorption of photons with the rate of production of photons by radiative recombination. The radiative recombination rate can thus be obtained as an integral over the optical functions. This relationship, pioneered by van Roosbroeck and Shockley in 1954, has been used a great deal because the optical quantities are often known with some accuracy.

This result corresponds to balancing the rate $B^s np$ and the rate Y^s in Fig. 1(a). Analogous detailed balance results are obtainable for the other pairs of processes in Fig. 1. Thus, the Auger recombination rate can be related to an integral over the impact ionization rate. However, this connection is less useful since impact ionization data are generally less well known than the optical absorption information.

Nonradiative Processes

Auger Effects

The Auger effect was discovered in 1925 in gases by Pierre Auger. An atom is ionized in an inner shell. An electron drops into the vacancy from a higher orbit and a second electron takes up the energy which is used to eject it from the atom. In solids the effect is roughly analogous. One of its characteristics is that it is not radiative. Since radiation is what is seen in many experiments, the Auger effect is suspected if and when there is less radiation than expected in the first place. It is rather harder to investigate than radiative transitions.

The effect has proved to be important as it limits the performance of semiconductor lasers, light-emitting diodes and solar cells, and it can be crucial in transistors and similar devices whose performance is governed by lifetimes. When heavy doping is required, as it is in the drive towards microminiaturization, its importance tends to increase, since the Auger recombination rate behaves roughly as $n^2 p$ or $p^2 n$ (see Fig. 1) compared with the radiative rate which behaves more like np . The inverse process is impact ionization, and is important in the photodiode, the impact avalanche transit time (IMPATT) diode, and hot electron devices.

Impact ionization can be regarded as an autocatalytic reaction of order one:



i.e., one extra particle is produced of the type present in the first place. This is a key feature for impact-induced nonequilibrium phase transitions in semiconductors.

In the theory one needs wavefunctions for four states of two electrons which are relevant after the many-electron problem has been reduced to a two-electron problem. In a matrix element calculation the four wavevectors imply a 12-fold integration in k -space to cover all possible states of the two electrons. However, momentum and energy conservation usually reduce this to an eight-fold integration. Such a calculation is hard, for it requires (1) good wavefunctions and (2) accurate integrations.

Here we shall merely concentrate on the broad principles. In addition, the many-electron nature of the problem implies that another approximation is often inherent in the treatment apart from the use of perturbation theory. This is clear if one considers that the electron interactions are screened twice: once by an exponential screening factor and a second time by the dielectric constant. So there is some double counting, and the treatment of the effect as uncorrelated electronic transitions mediated by

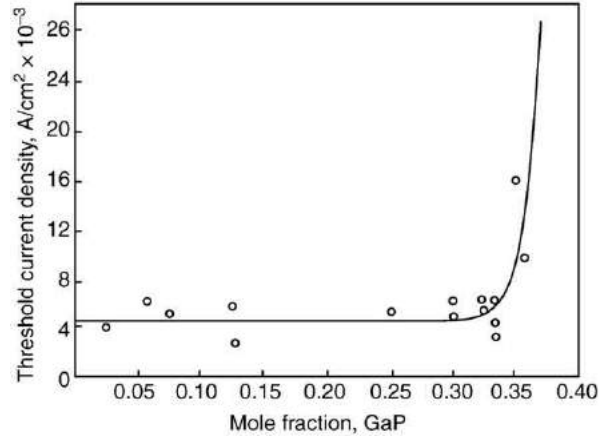


Fig. 3 Lowest values of laser threshold current density at 77 K as a function of mole fraction of GaP for $\text{Ga}(\text{As}_{1-x}\text{P}_x)$ junction lasers. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. *Solid-State Electronics* 9: 735–742, with permission from Elsevier.

screened Coulomb interactions is another approximation. The collective effects that enter require more sophisticated field theoretic methods which take care of electron–hole correlations, plasmon effects and the effect of free excitons, the so-called excitonic Auger effect. These calculations broadly confirm the results obtained by the simpler methods used for energy gaps large compared with the plasmon energies, provided the high-frequency dielectric constant is used and doping is not too heavy. Significant corrections are required for the narrow gap lead compounds, for example, and a considerable computational effort is required.

Let us now consider lifetimes for high carrier concentrations, as limited by band–band processes. They are best studied by looking at the emitted radiation. In this connection we recall the striking rise of the threshold current density J in a semiconductor laser as the material is changed from a direct material like GaAs to an indirect material, by mixing it with a compound such as GaP. **Fig. 3** shows this spectacular rise for $\text{GaAs}_{1-x}\text{P}_x$ at 77 K near $x=0.38$. It is due to the drop in the radiative transition probabilities, as the substance becomes indirect, thus requiring a higher current density for threshold. So we should start with band–band processes in direct materials as most favorable for the experimentally accessible radiative transitions.

The essential point here is that the injected carrier density behaves as:

$$n_{\text{inj}} = \frac{J\tau}{qd} \sim \frac{(5 \times 10^3 \text{ A cm}^{-2})(10^{-9} \text{ s})}{(1.6 \times 10^{-19} \text{ C})(10^{-4} \text{ cm})} \sim 10^{17} \text{ cm}^{-3} \quad (17)$$

where J/q is the particle current density at threshold, τ is the lifetime of the carriers, and d is the thickness of the active layer. As active layers are made thinner n_{inj} increases to 10^{19} cm^{-3} or so, far in excess of the defect concentration in the (undoped) material. This brings in band–band Auger effects. These have also been invoked to explain the undesirable increase in J with temperature. Thus the interest in direct band–band Auger effects in III–V compounds is fuelled by the need for better and smaller optoelectronic devices which work at long wavelengths (1.3–1.5 μm). It matches the interest in indirect band–band Auger and impurity Auger effects in silicon due to the importance of heavy doping in VLSI (very large scale integration) and transistor technology. **Fig. 4** shows a typical Auger process, called CHHS, as the conduction band (C) and the split-off band (S) are involved. Two relevant states are in the heavy hole band (H). State 2' is referred to as the Auger carrier, as it has more kinetic energy than the others.

Impact Ionization

The CHHS and CHCC processes and their inverses can be written as

$$e_C + 2h_H \leftrightarrow h_S \quad 2e_C + h_H \leftrightarrow e_C \quad (18)$$

where the suffix refers to the band. Viewed from left to right these are Auger processes. Viewed from right to left they are autocatalytic and impact ionizations.

Such processes are important for the impact-induced nonequilibrium phase transition in semiconductors. Also reaction rates with autocatalytic elements imply nonlinear equations in the concentrations and this gives rise to much interesting behavior as regards stability and bifurcation phenomena.

Momentum and energy conservation are very restrictive conditions on the four states involved in (18), and it is easily seen that they cannot normally be satisfied if in a direct band semiconductor the recombining electron drops from the band minimum to the valence band maximum. This effect raises all the kinetic energies of possible processes. The Auger electron (on the right-hand sides of (18)) must also have a minimum energy in order that the impact-ionization process can proceed. This leads to an activation energy for the process. For **Fig. 4**, for example, and in the case of nondegeneracy, the energetic hole has a kinetic energy

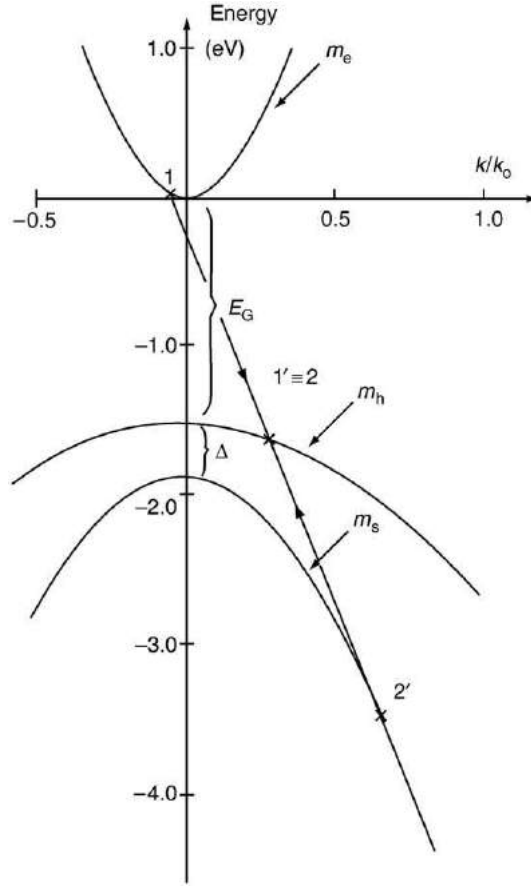


Fig. 4 Conduction band, heavy hole and split-off valence bands of GaAs, all treated as parabolic with $\hbar^2 k_0^2 / 2m_h \equiv E_G - \Delta$. x denotes a quadruplet of states for a most probable transition. The two states in the heavy hole band are not shown separately. The arrows indicate electron transitions. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. Solid-State Electronics 9: 735–742, with permission from Elsevier.

at threshold of

$$E_{th} = \frac{m_e + 2m_h}{m_e + 2m_h - m_s} (E_G - \Delta) \quad (\text{CHSS}) \quad (19)$$

As the band gap increases we see that E_{th} goes up and so the Auger rate for simple parabolic bands decreases. Values of E_{th} for other transitions can be obtained from Eq. (19).

The total kinetic or threshold energy E_{th} can be converted to an activation energy by subtracting the basic energy which must under all conditions be part of the Auger particle energy. In the case of Eq. (19) one finds for $E_G > \Delta$:

$$\begin{aligned} E_a &= E_{th} - (E_G - \Delta) \\ &= \frac{m_s}{m_e + 2m_h - m_s} (E_G - \Delta) \end{aligned} \quad (20)$$

This can result in a strong temperature dependence in accordance with an Arrhenius factor $\exp(-E_{th}/kT)$. However, this is again lost, and the direct band–band Auger effect will certainly be important and only weakly temperature dependent, if $E_G \sim \Delta$. This can occur, for example, for InAs, GaSb and their solid solutions.

The connection between Auger processes and impact ionization has also been formalized. Of all Auger quadruplets of states in a set of nondegenerate bands, the most probable Auger transition involves the quadruplet, which yields the threshold for impact ionization. The three crosses in Fig. 4 indicate such a quadruplet (the central cross represents two states).

Identification of Auger Effects

How does one know that an Auger effect has occurred? In pure but highly excited materials there is the original solid-state Auger effect which leads to a lifetime broadening of the electronic states at the band edge. For a semiconductor this effect causes fuzziness in the band edge which is a contributory factor to the overall bandgap shrinkage. The mechanism is illustrated in Fig. 5, which

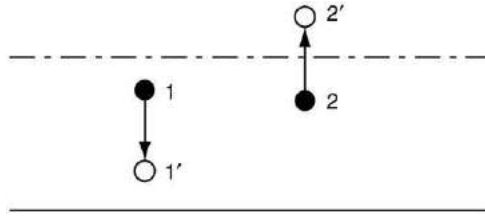


Fig. 5 The filling of a state k by the Auger effect. The arrows indicate electron transitions.

shows how the lifetime of a vacant electron state near the band edge is shortened by the Auger processes within this band. Under normal conditions this effect is small in a semiconductor. But under degenerate conditions the effect leads by lifetime broadening to a low-energy tail in emission.

The blanket term 'Auger effect' for all Coulombically excited two-particle transitions has been used here. In semiconductor device work the 'Auger effect' has become associated with transitions in which one electron bridges an energy gap. But there is no need to limit the concept in this way. Its competition with radiative effects means that it is often detrimental and a full discussion of means of suppressing it has recently been given by Pidgeon *et al.*

More spectacular and more convincing is the *detection* of the energetic Auger electron or hole. For this purpose one may look for the weak luminescence emitted when this carrier recombines radiatively. This has been done for the band–band process in Si and for the band–impurity process in GaAs. It leads to so-called $2E_G$ -emission. Its rate-limiting step is the rate of the Auger process per unit volume, $B_1 n^2 p$, which populates the high-energy level. The radiative process then proceeds at a rate $B'_1 n^2 p^2$ per unit volume with $B'_1 \sim 10^{-54} - 10^{-51} \text{ cm}^9 \text{ s}^{-1}$.

More usual is the identification of the band–band Auger recombination mechanism from the minority carrier lifetime τ , which behaves as

$$\frac{1}{\tau_p} = B_1 n^2 \quad \text{or} \quad \frac{1}{\tau_n} = B_2 p^2 \quad (21)$$

where B_1 and B_2 are Auger coefficients when the Auger particle is an electron or hole, respectively. A more complete expression allows also for trapping processes.

The departure from parabolic bands plays an important part for the energetic Auger particle (in state $2'$). Theory shows that when this effect is taken into account, it tends to lower the band–band Auger coefficient. In fact in GaAs the CHCC process is ruled out altogether at 0 K for nonparabolic bands, suggesting that it should be an ideal material for radiative transitions. However, the possibility of phonon participation brings the effect back again, though it is still comparatively weak. The difficulty of conserving electron energy and momentum, which gives rise to the activation energies (Eq. (20)) is greatly alleviated if phonons can take up some of the momentum in a direct gap semiconductor.

We have seen that the activation energy barrier against the band–band Auger effect can be overcome by a suitable disposition ($E_G = \Delta$) of the three direct bands and by phonon participation. It can also be overcome if there is an indirect minimum near a Brillouin zone edge at about half the direct energy gap.

Acknowledgments

The author wishes to acknowledge support from NATO under their grant PST.CLG 975758.

Further Reading

- Fraser, D.A., 1986. *The Physics of Semiconductor Devices*, fourth ed. Oxford, UK: Clarendon Press.
- Hangleiter, A., 1985. Experimental proof of impurity Auger recombination in silicon. *Physical Review Letter* 55, 2976–2978.
- Harrison, D., Abram, R.A., Brand, S., 1999. Characteristics of impact ionization rates in direct and indirect gap semiconductors. *Journal of Applied Physics* 85, 8186.
- Landsberg, P.T., 1991. *Recombination in Semiconductors*. Cambridge: Cambridge University Press.
- Nimtz, G., 1980. Recombination in narrow gap semiconductors. *Physics Reports* 63, 265.
- Pidgeon, C.R., Ciesla, C.M., Murrin, B.N., 1997. Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21, 361–419.

Coherent Terahertz Sources

L Wang, Chinese Academy of Sciences, Beijing, China

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Terahertz (THz) radiation is usually referred to as an electromagnetic wave with frequencies ranging from 1 to a few terahertz (1 THz = 10^{12} Hz). In the electromagnetic spectrum, terahertz radiation lies between the infrared and microwave, as shown in Fig. 1. The techniques for generation and detection of THz radiation bridge photonics and electronics in the sense of the concepts and the techniques discussed below.

Physical quantities corresponding to 1 THz are listed as follows;

- Frequency, 1 THz
- Wavelength, 300 μm
- Wavenumber, 33 cm^{-1}
- Energy, 4.1 meV
- Temperature, 48 K

In this spectral region, there exist many rich physical, chemical, and biological phenomena. For example, many phonon resonance and other elementary excitations in condensed matter fall in this region; many molecular vibrations and rotations occur at THz frequency; conformation-related collective vibrational modes in macromolecules, especially in bio-molecules such as proteins and DNA, also lie in THz frequency. However, the development of THz technique lags far behind that of photonics and electronics, due to lack of feasible, reliable, and economic coherent THz sources. The situation has been rapidly changing since the 1990s. Benefiting from advanced material science and the ultrafast laser techniques, various coherent THz sources are being developed and commercialization is soon to follow. We will introduce these coherent THz sources and their main features, and discuss their generation mechanisms. More details can be found in the listed books and articles in the Further Reading section at the end of this article.

Coherent THz Radiation

When a time varying polarization, $\mathbf{P}(\mathbf{r}, t)$, is generated in a medium and oscillates at THz frequency, it will radiate a THz wave, $\mathbf{E}(\mathbf{r}, t)$, according to Maxwell's equation:

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}, t) + \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial^2}{\partial t^2} \mathbf{P}(\mathbf{r}, t) \quad (1)$$

or

$$\nabla \times \nabla \times \mathbf{E}(\omega) - \frac{\omega^2}{c^2} \mathbf{E}(\omega) = \omega^2 \mu_0 \mathbf{P}(\omega) \quad (2)$$

in frequency domain, for a monochromatic wave or a Fourier component of a wave with a finite bandwidth. The properties of THz radiation are determined by the polarization source $\mathbf{P}(\mathbf{r}, t)$. Incoherent THz sources exist in blackbody radiation, and are also detected in astronomy observations. If the phase evolution of the THz wave keeps in step, both spatially and temporally, it is called coherent radiation. A coherent THz source is more powerful and useful in spectroscopy, material characterization, imaging, and many other applications. It is clear that a coherent polarization source is first required for generation of a coherent THz wave. In practice, the working media for THz emitting can be free electrons, quasi-free electrons in solids, or bound electrons and other charge oscillations, such as plasmon oscillation, lattice vibration, etc. Classified by the excitation properties, the emitting process can be of resonance or nonresonance. Coherent polarization sources with different features can be created using various excitation techniques under different principles and device configurations, which lead to different coherent THz sources. Here we focus on coherent THz sources, primarily based on photonic techniques, and working in the spectral range from 1 THz to a few THz.

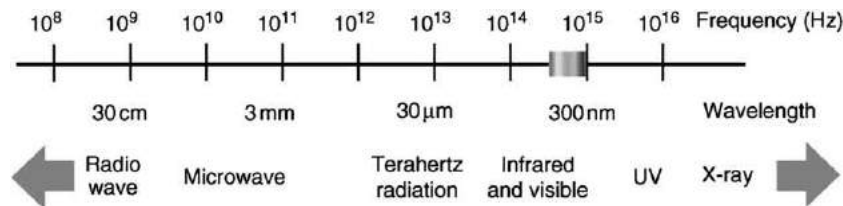


Fig. 1 The electromagnetic spectrum with THz radiation sitting between the microwave and the infrared.

Gas THz Laser

Molecules have many rotational and vibrational energy levels spaced at THz frequency. For polar molecules, the radiative transition between these levels can radiate THz wave. Following the standard laser technology, molecules pumped either optically or electronically can be used as laser gain media for coherent THz wave generation. For THz gas lasers, the very low THz photon energy poses a problem for lasing, because the coherent radiative transition can be disrupted by thermal depopulation and dephasing. An effective population inversion is difficult to set up and maintain. Therefore, it is important to carefully choose working medium and energy levels, as well as the way to excite the molecules. A number of media has been used for THz laser, such as methanol and HCN vapor, working in both continuous wave (CW) and pulsed modes. For example, methanol lasers operate under excitation of the rocking and asymmetric deformation modes of methanol excited by a CO₂ infrared laser. HCN laser is pumped electrically, which needs a large voltage and current to excite the molecules, a long cavity to obtain sufficient gain, and dedicated control of the temperature in the cavity for stable operation.

In general, THz gas lasers can be tuned discretely in thousands of lasing lines, ranging from a few tens to a few hundreds of micrometers, with power output from μW to mW. THz gas lasers have existed since the early 1970s, and are the only commercialized THz lasers to date. They are still subject to development, for less bulky volume, enhanced tunability and efficiency, and reduced costs.

Free Electron THz Laser

A free electron laser (FEL) uses free electrons as working medium, rather than bound atomic or molecular states in a conventional laser. A typical FEL consists of three parts: (i) an electron source that generates an electron beam with high current; (ii) an accelerator to raise the electron energy; and (iii) a lasing cavity. After the electron beam is generated, electrons are first accelerated to a relativistic velocity, typically with energy of hundreds of MeV, and then enter the cavity consisting of end mirrors and a spatially periodic magnets array called a wiggler (Fig. 2). Due to the Lorentzian force imposed by the periodic magnetic field, the high-energy electrons move in the cavity along a sinusoidal path, and emit coherent radiation with a wavelength determined by the spacing between magnets and the electron velocity, as well as the magnetic induction. With the feedback provided by the cavity mirrors, electrons are accelerated or decelerated continuously by the optical field, and are bunched via the resonant interaction. The collective motion of the electron bunches radiates powerful coherent synchrotron radiation. In a standard configuration, the wavelength of the radiation can be expressed as:

$$\lambda = \frac{\Lambda}{2\gamma^2} [1 + (k\Lambda B_0)^2/2] \quad (3)$$

where Λ is the spatial period of the wiggler magnets, γ the Lorentz factor of the electron beam, B_0 the peak magnetic induction, and k a constant. With a suitable arrangement of the magnet array, as well as other related factors, the facility can work in THz spectral region.

An FEL generates tunable, coherent, high-power THz radiation. It is the most powerful coherent source of THz radiation available. Up to 50 W of average THz power has been generated in the Jefferson Lab recently. The free electron THz laser is becoming an important tools for studying THz radiation and its interactions with condensed matter and biological materials. On the other hand, the FEL is a complicated and expensive facility, and needs dedicated maintenance.

Semiconductor THz Laser

In the twentieth century, semiconductor devices have achieved tremendous success, both in electronic and photonic regime, such as transistors and diode lasers. The semiconductor industry has resulted in revolutionary changes to our world and everyday life. To fill the THz gap between electronics and photonics, great efforts have been made to meet the demand in compact, economic, tunable, and highly efficient THz sources. Encouraging advances have been made since we stepped into the new millennium.

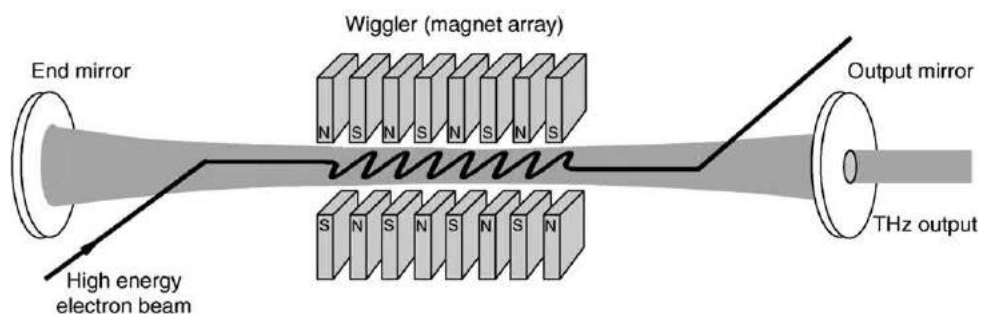


Fig. 2 A schematic of THz generation in a free electron laser.

In a semiconductor quantum well, the energy difference between a pair of sublevels can be artificially designed at THz frequency. As early as 1971, the concept of far infrared lasers based on intersub-band transition in semiconductor superlattices was proposed. Although the sublevels can be populated by electrical or optical pumping, the low photon energy at THz frequency causes serious difficulties for lasing, because the two levels sit so close to each other that many thermal processes could destroy the population inversion. Various designs and configurations have been proposed to tackle the problem. An attractive idea is so-called quantum cascade heterostructure. In this kind of device, the electrons are injected into a series of coupled quantum wells electrically. When the electrons are driven by the biased voltage, they can make radiative transition between a pair of sub-bands in a well. Subsequently, electrons enter the next well through resonant tunneling, and so on. However, this kind of device needs the materials of very high quality and suffers severely from fast depopulation of the excited states, even at very low temperatures.

As the advance in material science and unremitting efforts in pursuing the dedicated design of the device structures continued, the real breakthrough came in 2001, when Köhler *et al.* at Pisa demonstrated THz lasing in a quantum cascade heterostructure. Many improvements have been made since then; for example, the idea of using phonon resonance to selectively deplete the population of the lower sub-band has been successful, so that much stabler and robust population inversion can be set up. The operation temperature has been raised to 136 K, well above the liquid nitrogen temperature, which opens the way for semiconductor THz lasers stepping into many more daily applications. These semiconductor devices are pumped with low voltages and currents at mA levels, and generate narrow bandwidth radiation with mW output at a frequency of a few THz. It is also probable that a four-level lasing system could even working at room temperature.

Another promising device is the p-germanium (Ge) laser developed recently. The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is formed by polishing the surfaces of the Ge crystal. These lasers could run at 5% duty cycles, with several mW output power. The output wavelength can be continuously tuned from 1 to 4 THz. At the moment, the devices have to operate at 20–30 K and consume about 10–20 W for refrigeration.

It is safe to predict that the feasibility, low cost, and compact semiconductor THz laser will emerge in the near future.

Beside the THz lasers, coherent THz radiation can also be generated by coherently exciting suitable media in other ways. Since the mid-1990s, the development of ultrafast laser techniques has led to two important pulsed coherent THz sources, i.e., photoconductive antenna and optical rectification THz sources. A unique feature of the pulsed coherent sources is broad bandwidth, which is useful in spectroscopic applications. Fig. 3 is a typical setup for pulsed THz generation and spectroscopic study, using photoconductive antenna or optical rectification as THz sources.

THz Photoconductive Antenna (PCA)

Different from a conventional antenna working in radio and microwave frequency that are driven directly by electrical current, a photoconductive antenna for THz wave generation works with ultrafast pulsed lasers. When the photoconductive gap is irradiated by an ultrafast laser pulse, photocarriers are generated in the semiconductor. These photocarriers are driven by a DC voltage bias and form a current transient. The antenna couples the electromagnetic field associated with the time varying current into free space, and produces a THz pulse. The THz pulse usually has a duration of a few picoseconds, with a broad frequency bandwidth centered at THz frequency. The output THz power scales with both the optical pulse power and the DC bias field:

$$E_{\text{THz}} \propto \frac{\partial J}{\partial t} \propto \frac{\partial^2 n_c}{\partial t^2} E_{\text{bias}} \quad (4)$$

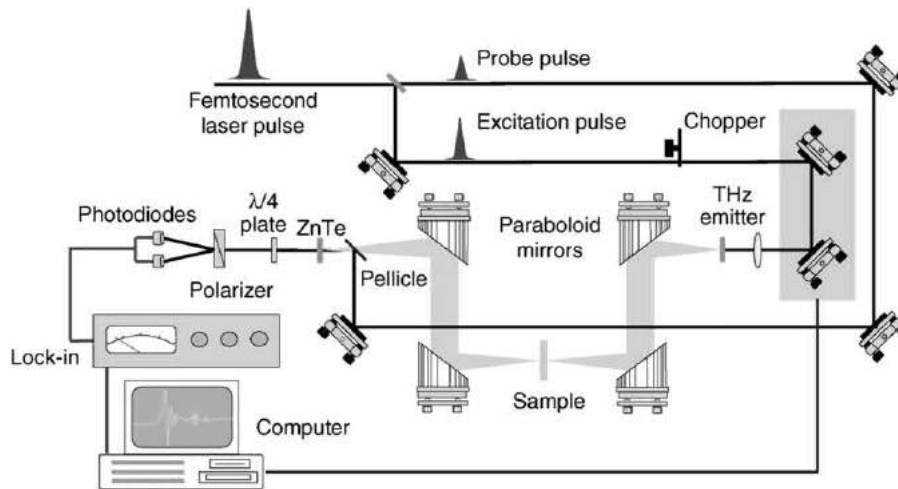


Fig. 3 A typical setup for broadband pulsed THz generation and spectroscopic study with a femtosecond laser as excitation source.

The output power, frequency bandwidth, and polarization of the THz radiation depend on the photoconductive features of the semiconductor layer and the geometric structure of the antenna. For an optimal performance, the substrate should have a suitable bandgap for optical interband excitation, a high electron mobility and short carrier lifetime to support an ultrafast current transient, and high dielectric breakdown threshold and to sustain high bias voltage. High-power PCAs have been demonstrated with semi-insulating GaAs, ion-implanted GaAs, and In-GaAs. In the aspects of antenna structures, coplanar strip lines and large aperture emitters are the most effective. Average output power from a good system can reach 30–40 μW .

A broad bandwidth is a special characteristic of a THz source, which is particularly desirable in the spectroscopic study as a coherent probe source. From the property of Fourier transform, we know that the shorter the THz pulse duration, the broader the spectral band. PCAs made of GaAs typically have a useful bandwidth extending from ~ 100 GHz to 3 THz, which can be extended to 4 THz by tuning the pumping laser wavelength close to the bandedge. Up to 6 THz bandwidths have been reported for PCAs. The pulse duration and, therefore the achievable bandwidth, are ultimately limited by carrier mobility and TO phonon resonant absorption which is around 8.3 THz in GaAs.

In many applications, THz radiation is coupled to free space from the antenna using a closely attached silicon hemispherical lens. This practice increases the system's output and also provides control of the radiation pattern. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field. In the far-field, the THz beam has a Gaussian cross-section with high-frequency components concentrated in the center.

As the first practical pulsed coherent THz source, PCA has been widely used for many applications in scientific research and material characterization. Combined with commercial ultrafast optical fiber lasers, full functional and self-contained THz spectroscopy systems with PCAs as emitters, have been demonstrated and are under commercialization.

Optical Rectification

Optical rectification is a second-order nonlinear optical effect originating from susceptibility $\chi^{(2)}$, and was first observed in 1962 using two intense monochromatic nanosecond laser pulses. In this process, the interaction between laser and nonlinear optical crystal produces a frequency difference polarization:

$$P(0) = \chi^{(2)} E(\omega_1) E(-\omega_1) \quad (5)$$

When a nonlinear optical crystal is irradiated by an ultrafast laser pulse, the different frequency components in the laser pulse will be coupled with each other by $\chi^{(2)}$, to induce frequency difference polarizations as:

$$P(\Omega) = \chi^{(2)} (\omega_1 - \omega_2 = \Omega) E(\omega_1) E(-\omega_2) \quad (6)$$

A femtosecond laser pulse usually has a spectral linewidth of at least a few nm, the induced polarization $P(\Omega)$ has a finite frequency distribution centered at THz frequency in the far field, and results in THz radiation $E_{\text{THz}} \propto \partial^2 P(t) / \partial t^2$.

THz radiation in free space by optical rectification was obtained in the early 1990s. One of the differences from a PCA device is that the THz output by optical rectification scales with both the optical pulse power and the DC bias field, while the THz radiation generated by optical rectification solely results from the incident laser. The optimal performance depends on several factors. The generation efficiency first depends on the magnitude of the $\chi^{(2)}$ and phase-matching condition between the optical and THz pulses. The effective magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation. Up to now, the most popular material is ZnTe, for its physical durability and excellent phase matching, when a Ti:sapphire femtosecond laser is used as the excitation source. With a moderately focused optical pump beam and a ZnTe crystal, nW T-ray average power can be generated by optical rectification. In general, the generation efficiency may be increased by increasing the laser power, but there are two factors limiting the effect. First, the incident laser power cannot exceed the damage threshold. Second, other second-order nonlinear processes may compete with the optical rectification, such as second-harmonic generation of the pump beam at high optical flux.

For THz radiation generated by optical rectification, the ultimate bandwidth is primarily determined only by the linewidth of the pump laser pulse. Because it is a nonresonance process in most cases, optical rectification can reach broader bandwidths than PCA. Using ultrashort optical pulses and thin nonlinear crystal, THz pulses have been generated with spectra extending to the mid-infrared, well beyond the phonon band of the materials. Like the photoconductive antenna, optical rectification has become a popular technique to generate coherent broadband THz radiation in various applications. Beside the conventional bulk inorganic crystals, many different media have been used for THz generation, such as organic crystal DAST, biased quantum wells, periodically poled LiNbO₃ (PPLN), poled polymers, super-conducting thin films, etc.

Other Coherent THz Sources

When two light sources with slightly different wavelengths interact together with a nonlinear optical crystal, a beating polarization will be generated and will radiate a THz wave. It is called difference frequency generation, or three-wave mixing. Optical parametric processing is another way to generate a THz wave. In this process, a near-infrared pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation can be generated from the beating of the pump and the idler. The merit of this technique is the continuous frequency tunability of the THz output, which is valuable in spectroscopic applications, and relatively cheaper

nanosecond lasers can be used. If the nonlinear crystal is further placed in a cavity and the idler beam is amplified by feedback from end mirrors, a THz optical parametric oscillator is formed. A number of improvements have been made since the mid-1990s with this kind of device. More recently, THz parametric generation with an injection seeded idler beam has shown a reduced linewidth ($\Delta\nu/\nu \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4 ns pulses. A narrow linewidth combined with reasonable tunability make this kind of THz source attractive in spectroscopic studies.

Besides using photonic techniques, coherent THz radiation can also be generated by electronic techniques through increasing the output frequency of the microwave devices. For example, an electrically driven microwave generator, backward wave oscillator, can generate CW THz radiation up to 2 THz. A backward wave oscillator, running under optimal conditions, can provide up to 300 mW of polarized, narrowband radiation with a limited tunability about 30% of its central frequency.

See also: Strong-Field Terahertz Excitations in Semiconductors. Terahertz Physics of Semiconductor Heterostructures

Further Reading

- Grüner, G., 1998. Vol. 74 of Topics in Applied Physics. Berlin: Springer-Verlag.
 Kaiser, W., 1988. Vol. 60 of Topics in Applied Physics. Berlin: Springer-Verlag.
 Khurgin, J.B., 1994. Optical rectification and terahertz emission in semiconductors excited above the bandgap. *Journal of the Optical Society of America B* 11, 2492–2501.
 Rullière, C., 1988. Femtosecond Laser Pulses. Berlin: Springer-Verlag.
 Saldin, E.L., Schneidmiller, E.A., Yurkov, M.V., 2000. Advanced Texts in Physics. Berlin: Springer-Verlag.
 Siegel, P.H., 2002. Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques* 50, 910–928.
 Tredicucci, A., Gmachl, C., Capasso, F., *et al.*, 2001. Novel quantum cascade devices for long wavelength IR emission. *Optics Materials* 17, 211–217.
 Tsen, K.T., 2001. Vol. 67 of Semiconductors and Semimetals. New York: Academic Press.

Using Ultrafast Optical Spectroscopy to Unravel the Properties of Correlated Electron Materials

Rohit P Prasankumar, Dmitry A Yarotski, and Antoinette J Taylor, Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Published by Elsevier Ltd.

Introduction

Conventional metals, semiconductors, and insulators form the backbone of modern technology. This is possible because the properties of these materials are well understood and can largely be described in a single electron picture, where interactions between electrons do not play a significant role. For metals in particular, this is quite surprising, since one would think that their typical carrier densities of $\sim 10^{23} \text{ cm}^{-3}$ (~ 1 electron/unit cell) would make it difficult to ignore the Coulomb interactions between electrons. Nevertheless, band theory can accurately predict the properties of these conventional materials in a quasi-free-electron picture, facilitating their use in a wide range of applications.

Despite the many technological successes achieved with these materials, researchers are continually searching for new materials with novel functionalities. Correlated electron materials (CEM) represent one of the most promising opportunities, since the electron–electron interactions governing their properties give rise to a host of unique phenomena, including magnetism, ferroelectricity, metal–insulator transitions (MIT), heavy fermion behavior, and superconductivity (SC). In addition, relatively small external perturbations (e.g., temperature, magnetic field, doping) can tune CEM between competing phases (e.g., SC and antiferromagnetic (AFM) order), further increasing their appeal. Intense interest in the fundamental properties of these materials, along with their great potential for applications, has thus driven decades of research into their properties.

The driving force for the unique properties of most CEMs is the competition between electron itinerancy and localization. For example, in conventional metals (with valence electrons in *s* or *p* orbitals), the electron hopping amplitude, *t*, is large, resulting in itinerant electrons with large kinetic energy that can travel freely throughout the material. However, when the potential due to Coulomb interactions between electrons (*U*) becomes comparable to or larger than *t*, electrons can be localized, even when band theory would predict them to be itinerant. This strongly influences the properties of CEM, which tend to have valence electrons in narrow bandwidth, low kinetic energy ($< 1 \text{ eV}$) *d* or *f* orbitals that are more susceptible to localization. The ratio between *U* and *t* is thus one of the first factors to consider when determining whether a material is strongly correlated.

The relatively low energies of electrons in CEM also make interactions with other degrees of freedom (DOF) (e.g., spin, charge, orbital and lattice) more significant, since they have comparable energy scales. This makes CEM extremely sensitive to small changes in external parameters, such as temperature, pressure, and magnetic field, due to these multiple competing energy scales; it also makes them remarkably flexible, since it is relatively easy to tune a given material between different ordered states.

A typical phase diagram for a representative CEM is shown in Fig. 1. This shows the different phases exhibited as a function of doping (*x*) and temperature (*T*) (although similar diagrams would appear when varying other parameters, such as magnetic field

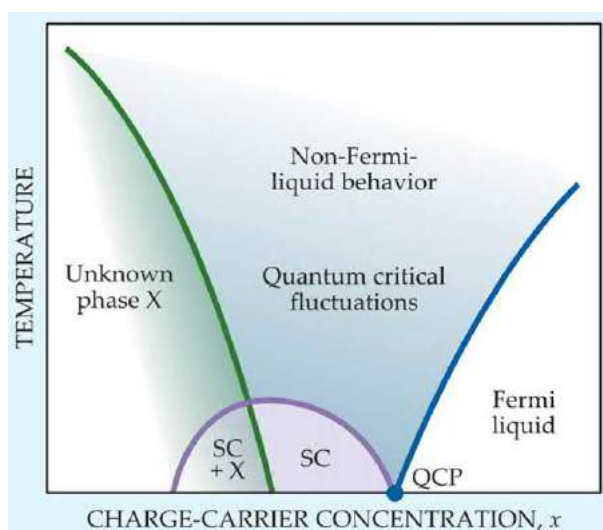


Fig. 1 Typical phase diagram for a hypothetical CEM. Reprinted with permission from Orenstein, J., 2012. Ultrafast spectroscopy of quantum materials. *Physics Today* 65, 44.

or pressure). It is clear from this schematic that small changes in T or x can result in transitions between different phases. In fact, the interplay between various DOF often leads to nanoscale phase separation, where different phases can simultaneously exist at a given temperature, although that is beyond the scope of this article.

Much insight into CEM can be obtained from optical spectroscopy (both time-integrated and time-resolved), which has several advantages over other experimental techniques for probing their properties. First, the energy scales relevant for CEM, ranging from meV (low energy excitations such as magnons and phonons) to eV (chemical bonds) are readily accessible using optics. In addition, optical techniques are non-contact and are therefore unlikely to introduce spurious effects in a given measurement. Furthermore, signatures of different ordered phases are present in the optical conductivity spectrum $\sigma(\omega)$; for example, the imaginary part ($\sigma_{\text{imag}}(\omega)$) in a superconductor has a $1/\omega$ dependence on the frequency (ω), while low energy excitations show up as peaks in the real part, $\sigma_{\text{real}}(\omega)$.

The addition of femtosecond (fs) temporal resolution to optical measurements offers additional benefits when probing CEM, particularly in unraveling the coexistence and interactions between different DOF. In many materials, various DOF relax on different timescales; for example, electrons typically lose their energy to phonons within ~ 1 picosecond (ps) and the phonons subsequently interact with spins (in a magnetic material) on a timescale of tens of ps. Time-resolved measurements of these relaxation processes (typically using pump–probe techniques) can therefore be used to extract fundamental material parameters, such as the electron–phonon and spin–lattice coupling constants. More recently, ultrafast optical techniques including time-resolved second harmonic generation (SHG), Faraday/Kerr rotation, and X-ray diffraction have been used to *selectively* probe ferroelectric, magnetic, and structural dynamics, respectively. This is an important advance, because one can selectively photoexcite one DOF with a femtosecond optical pulse and probe the resulting transient response of another DOF to unravel the interactions between them, in a manner that simply cannot be done using conventional, static electronic, magnetic and structural characterization techniques.

In this article, our goal is to describe the ability of ultrafast optical spectroscopic techniques to provide deep insight into correlated electron materials that often cannot be attained using any other technique. We will begin with an overview of the essential properties and phenomena exhibited by various classes of CEM, underlining the many similarities in the phenomena exhibited by different materials. A brief description of the use of linear optical spectroscopy to study CEM will then be given, focusing on the Drude and Lorentz models as well as the various low energy excitations present in optical conductivity spectra.

This will set the stage for the main focus of the article, ultrafast optical spectroscopy (UOS) of CEM. We will begin that section by briefly describing the results from all-optical pump–probe spectroscopy, which was the first ultrafast optical technique used to study CEM in the early 1990s and is often still the first UOS technique used to study a new class of materials. We will then move on to ultrafast spectroscopy at mid-to-far-infrared (IR) frequencies, which has been particularly fruitful due to the rich spectrum of low energy excitations in this frequency range, and has recently garnered particular attention with the advent of intense terahertz (THz) and mid-IR pulses that can be used to drive these low energy excitations. UOS experiments at high frequencies (ultraviolet (UV) to X-rays) are somewhat less common, but have still provided much insight into magnetic and structural dynamics, and will thus be described in the following subsection.

The final section will give an overview of other ultrafast optical techniques that have been used to study CEM, going beyond conventional pump–probe techniques to directly examine different DOF. These include time-resolved SHG, scanning near-field optical microscopy, angle-resolved photoemission spectroscopy, and X-ray/electron diffraction. We will conclude with a future outlook, describing directions that we feel will be particularly promising as new UOS techniques and correlated materials are discovered.

We note that it is impossible for us to cover all of the exciting work done in this important field, and we apologize in advance to any of our colleagues whose work we have overlooked. We have focused here on a few specific classes of CEM with which we have direct experience; however, there are other interesting and important classes of CEM (e.g., charge/spin density wave materials), as well as other materials showing hallmarks of electronic correlations (e.g., carbon-based materials), where UOS techniques have made important contributions. Some of these areas are covered in the reading list at the end, which both overlap with this article and cover areas that we have neglected. Our hope here is that the reader will gain an appreciation of the many advantages of ultrafast optical techniques for studying CEM, and the deep insight they can provide into the unique properties and phenomena exhibited by these fascinating materials.

Classes of Correlated Electron Materials

Electron–electron correlations are important in many different classes of materials, ranging from simple metals like Fe and Pb (exhibiting ferromagnetic (FM) order and superconductivity, respectively) to exotic f -electron systems such as URu_2Si_2 , which exhibits a variety of known and hidden phases. Here, we will focus on CEM whose properties are strongly influenced by multiple interacting degrees of freedom (unlike, e.g., simple ferromagnets) and thus generally display a rich and complex phase diagram similar to that in Fig. 1. We will begin this section by describing the properties of transition metal oxides (TMOs), arguably the most extensively studied class of CEM, and one that displays nearly all of the phenomena manifested in these materials (e.g., Mott physics, high- T_c superconductivity, metal–insulator phase transitions, etc.). This will be followed by a description of heavy fermion materials. In this section, our primary goal is to provide a basic understanding of the electronic, structural, and magnetic properties of CEMs, as this is critical for interpreting the results of ultrafast optical experiments. However, an in depth discussion of CEM properties is beyond the scope of this article, and we encourage the interested reader to peruse the reading list.

Transition Metal Oxides

Oxides are ubiquitous in our daily lives, as this class of compounds includes rust, sand, glass, precious gemstones such as ruby and sapphire, and the zinc oxide found in sunscreen. When transition metals are added to these materials, however, a multitude of novel and extremely interesting phenomena can appear, as mentioned above. Research on these materials has been ongoing since at least 1937, when Peierls and Mott pointed out that crystals with partially filled d orbitals (such as NiO) should be metallic based on band theory, but instead are insulators. We know now (in no small part due to the efforts of Peierls and Mott themselves) that electron–electron correlations are the origin of their insulating behavior. However, for a long time TMOs were neglected, partially due to the difficulty in fabricating them as compared to metals and semiconductors. This changed in 1986 upon the discovery of high- T_c superconductivity in layered copper oxides (where T_c is the critical temperature below which superconductivity exists), stimulating an enormous amount of effort in this area and leading to the discovery of additional effects, including colossal magnetoresistance (CMR), magnetoelectric coupling, and interfacial two-dimensional electron gases (2DEGs) in other TMOs.

The parent compounds of many TMOs (including the most well known examples, high- T_c superconductors (HTSC) and CMR manganites) are antiferromagnetic Mott insulators, with the chemical formulas ABO_3 or A_2BO_4 (where A is a rare earth ion and B is a transition metal (TM) ion). These materials possess an odd number of electrons/site and thus are expected to be metals from band theory, but Coulomb repulsion impedes electron transport, making them insulators (i.e., $U > t$ in the terminology used above).

This competition between electron itinerancy and localization in CEMs occurs at relatively small energy scales, as described in the previous section, amplifying the influence of spin, orbital, charge, and lattice degrees of freedom on their physics. In TMOs, this can readily be seen from the simple example of a transition metal ion placed in the cubic ABO_3 perovskite crystal structure (probably the most common TMO structure) (Fig. 2(a)). In this structure, the TM (e.g., Mn or Fe) occupies the corners of the unit cell. The cubic crystal field splits the 5-fold degenerate $3d$ energy levels of the TM ion, causing two e_g levels to go up in energy and three t_{2g} levels to go down in energy. This occurs because the TM ions are surrounded by oxygen octahedra; the e_g orbitals are

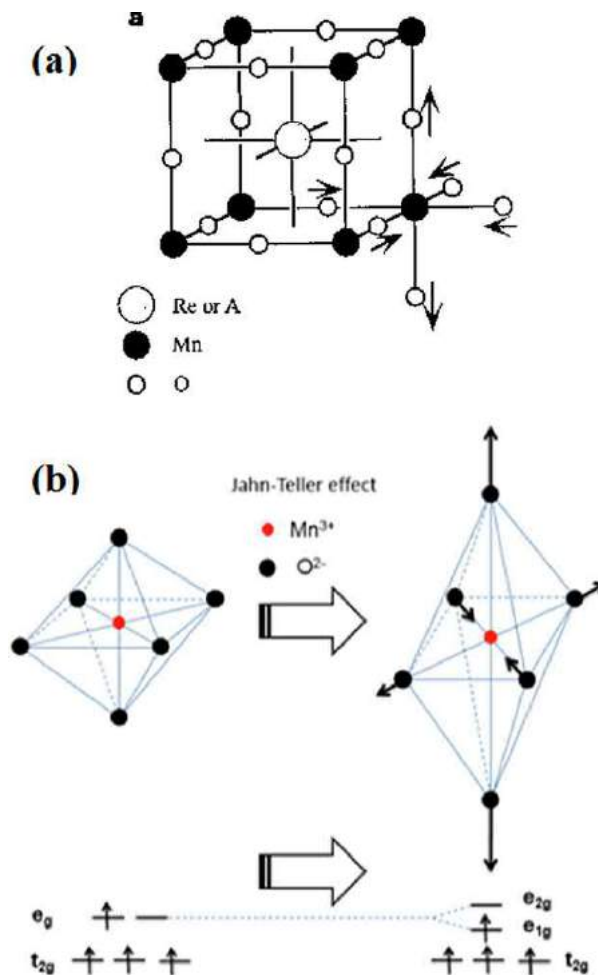


Fig. 2 Schematic of (a) the perovskite crystal structure (reprinted with permission from Millis, A.J., 1998. Nature 392, 147) and (b) the $Mn^{3+}O_6$ octahedron distortion and associated energy level splitting due to the JT effect.

directed towards the negatively charged oxygen ions and thus increase in energy (due to Coulomb repulsion), while the t_{2g} orbitals have lobes pointing in between the oxygen ions and decrease in energy. This illustrates the importance of the orbital DOF.

The lattice DOF also plays an important role, since the oxygen octahedra can elongate or compress along different axes to stabilize a given occupied orbital. If these distortions are strong enough, they can form a potential minimum that traps carriers, known as a lattice polaron. The most common distortion found in TMOs is due to the Jahn–Teller (JT) effect, where four oxygen ions move in one direction and two move in the other, creating a JT polaron that further splits the e_g energy levels (Fig. 2(b)). This has a significant influence on electron transport, as will be shown in the following sections. Finally, the spins of the localized electrons interact with one another through virtual hopping to their nearest neighbors (known as superexchange), leading to AFM spin order that typically alternates from site to site (or from plane to plane along a given crystal axis).

The most interesting physics occurs when TMO parent compounds are doped, typically by substituting a fraction of the rare earth A atoms with another alkaline ion that adds holes to the system. This tunes the carrier concentration (x) through the phase diagram of Fig. 1, initially suppressing AFM order and leading to SC or FM phases for $\sim 0.1 < x < \sim 0.4$ in the cuprates and manganites, respectively. A variety of other ordered phases can also emerge, often coexisting and competing with one another, as described in more detail in the following sections for specific materials. However, it is clear from this brief overview of TMO physics that the interplay between the charge, spin, lattice, and orbital DOF drives the distinctive phenomena observed in these materials.

High- T_c superconductors

The discovery of high- T_c superconducting cuprates revolutionized condensed matter physics and materials science, stimulating a worldwide effort to understand and optimize their properties that continues to this day. HTSC cuprates ($A_{2-x}B_xCuO_4$) have a layered tetragonal structure, with CuO_2 planes separated by (A,B)O planes. The CuO_2 planes are responsible for the superconducting behavior; the interleaved (A,B)O planes essentially serve to add electrons or more often, holes to the CuO_2 planes, which suppresses the AFM order of the parent A_2CuO_4 compound. The first HTSC cuprate to be discovered was $La_{2-x}Ba_xCuO_4$, with $T_c = 29$ K, followed within a year by $YBa_2Cu_3O_7$ (YBCO) ($T_c \sim 93$ K). The latter discovery was particularly significant since T_c was above the temperature of liquid nitrogen (77 K), making practical applications more feasible.

Two of the most hotly investigated topics after the discovery of HTSC cuprates included the symmetry of the pairing wave function and the mechanism underlying pairing in HTSC. In conventional superconductors, such as Pb, the paired electrons (“Cooper pairs”) have an isotropic s -wave spin singlet wave function, resulting in a superconducting energy gap (Δ) proportional to T_c with equal magnitude throughout momentum (k) space. Pairing in these materials is mediated by phonons. In contrast, the Cooper pair wave function in HTSC cuprates was found to have d -wave symmetry, leading to an energy gap along certain directions in k space and gapless “nodes” along other directions. This, in combination with the complex phase diagram and proximity to AFM order, implied that phonons may not be responsible for electron pairing in the cuprates, and indeed, calculations show that the maximum T_c for phonon-mediated pairing is well below the experimentally observed T_c . Other possibilities for the pairing “glue” include AFM spin fluctuations or electron–electron correlations themselves. However, the nature of the pairing mechanism in HTSC cuprates remains one of the biggest unsolved problems in condensed matter physics to this day.

It is worth mentioning that in the last decade, other unconventional HTSC have also been discovered that are not TMOs. In 2008, iron-based superconductors (known as “pnictides”), with T_c up to 56 K, were first reported. This was quite surprising, since superconductivity was not believed to be possible in materials with strong magnetic ions like iron. These materials have attracted much attention because of their similarities and differences with the cuprates; i.e., they exhibit similar phase diagrams in which SC order emerges when doping an AFM parent compound. However, the parent compounds are metallic (in contrast with cuprates) and the pairing wave function has a specific type of s -wave symmetry. Despite the relatively recent discovery of these materials (along with the related iron chalcogenides, which have a T_c possibly as high as 109 K), the extensive experimental and theoretical expertise that has been developed from studying the cuprates for the past three decades has provided much insight into their physics. In fact, more is arguably known about their underlying physics than that of the cuprates; for example, the pairing mechanism, although still in question, is believed to be either due to spin/orbital fluctuations or electron correlations.

Finally, conventional (phonon-mediated) superconductivity was unexpectedly discovered in sulfur hydride (H_2S) at high pressures > 160 GPa in 2015. Although the implications of this discovery are still being explored, it raises our hopes for finding a room-temperature superconductor at ambient pressure in the near future.

Vanadium dioxide and related materials

Vanadium dioxide (VO_2) is a canonical metal–insulator transition (MIT) material that has been studied using nearly every ultrafast optical technique, due to the longstanding controversy about the underlying origin of the MIT. At high temperatures, VO_2 is a metal with a rutile structure; upon cooling below $T_c = 340$ K, it becomes an insulator, with a concomitant change to a monoclinic structure that has a unit cell twice the size of that above T_c . The high temperature of the MIT has made VO_2 very attractive for potential applications. Importantly, the structural and metal–insulator transitions, though occurring at the same temperature, are not necessarily linked, and the question of whether the structural transition drives the electronic MIT or they occur independently has been the subject of much contention for decades.

Vanadium atoms in the high temperature rutile phase are arranged in a tetragonal lattice and surrounded by oxygen octahedra. Upon cooling below T_c , the vanadium atoms form dimers that tilt away from the c -axis, effectively doubling the unit cell; this structural change is known as the Peierls distortion, and has been suggested to directly lead to the insulating bandgap. However,

Mott believed that electron–electron correlations were essential to the MIT; as the system is cooled below T_c , the electron hopping parameter t decreases below the Coulomb interaction potential U , causing the system to become a Mott insulator.

Many researchers have used UOS techniques to address this problem, under the premise that the structural and electronic transitions could be disentangled in the time domain after ultrafast photoexcitation. However, to date the results have been mixed, as will be shown below, and the origin of the MIT is thus still an open question.

Magnetically ordered oxides

The discovery of colossal magnetoresistance in Mn-based oxides was perhaps the most significant offshoot of the intense interest devoted to HTSC cuprates. These materials, known as manganites (with the chemical formula $A_{1-x}B_x\text{MnO}_3$, where A and B are rare and alkaline earth ions, respectively), exhibit not only CMR, but also a variety of AFM, FM, orbital and charge-ordered phases depending on temperature and doping, with a phase diagram similar to that in Fig. 1 and crystal structures similar to Fig. 2. The CMR effect occurs when a magnetic field is applied near the Curie temperature T_c , which separates the high temperature paramagnetic (PM) phase from the low temperature FM phase. With no applied field, the PM to FM transition is typically accompanied by an insulator-to-metal transition, although this depends on the specific material. However, if a magnetic field of a few Tesla (T) is applied near T_c , changes of $> 1000\%$ in resistance can be observed (Fig. 3). This stimulated a substantial amount of effort aimed at both understanding the origin of this effect, as well as harnessing it for applications.

In the simplest picture, the CMR effect occurs when the magnetic field aligns the Mn spins between different sites. This influences the conduction of electrons in e_g levels, since their site-to-site-hopping depends on the alignment of their spins with that of the localized t_{2g} electrons; this is known as the double exchange mechanism. However, it was shown that double exchange cannot fully account for the dramatic decrease in resistivity observed in these materials; the strong electron–phonon coupling also plays a significant role by localizing carriers into polaronic states (particularly due to the JT effect described earlier). The interplay between these effects is believed to be the primary origin of CMR in the manganites, although nanoscale phase inhomogeneities are also thought to play an important role, perhaps even dominating the physics in certain parameter regimes.

The full spectrum of phases exhibited by the manganites, as with TMOs in general, depends on the ratio of U and t . As with many other TMOs, their parent compounds are AFM insulators, with the formula AMnO_3 ; doping at the A site gives the chemical formula $A_{1-x}B_x\text{MnO}_3$, which introduces holes into the system that can conduct (leading to CMR for $x \sim 0.15 - 0.4$). More specifically, manganites are often classified based on their electronic bandwidth, W (directly proportional to t), which depends on the Mn–O–Mn bond angle θ . This angle is determined by the ionic size of the A and B ions; the larger those ions are, the closer θ is to 180° (and the higher the bandwidth). Low bandwidth manganites, such as $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$ (PCMO), do not exhibit a metallic phase at any temperature without an applied magnetic field. Furthermore, their low bandwidth makes the interactions with other DOF more significant, leading to complex charge and orbital ordered (CO/OO) or magnetically ordered states for specific temperature and x ranges. In contrast, large bandwidth manganites, such as $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ (LSMO), exhibit a FM metallic phase above room temperature and are much less sensitive to interactions with other DOF. Intermediate bandwidth manganites like $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ (LCMO) exhibit the largest variety of phases and the greatest sensitivity to external parameters.

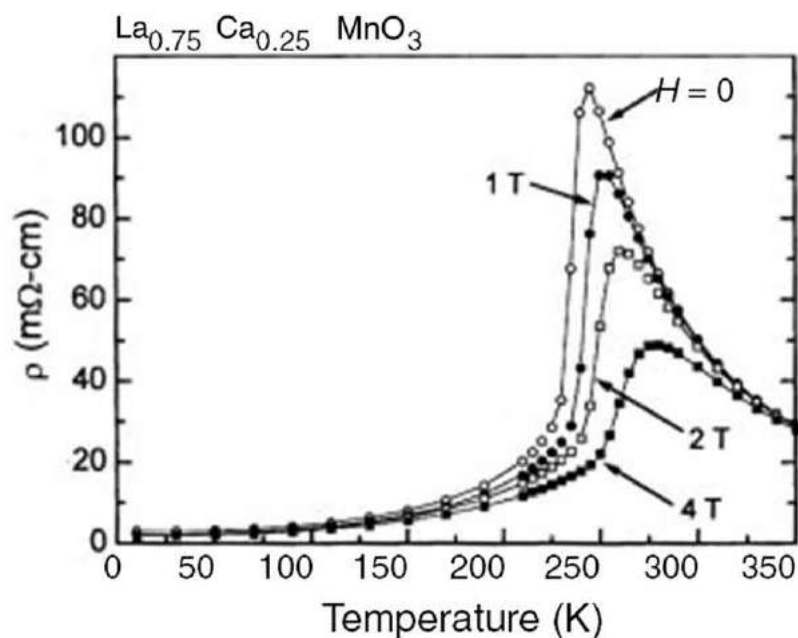


Fig. 3 Colossal magnetoresistance in LCMO. Reprinted with permission from Millis, A.J., 1998. Nature 392, 147.

Although the majority of manganites have a cubic perovskite structure, they can also adopt a layered structure (much like that of the cuprates), in which one or more conducting Mn oxide layers are separated by insulating layers. Layered manganites show many of the same complex magnetic and CO/OO phases as the cubic manganites, with the important distinction that the electronic and magnetic order is quasi-two-dimensional. This leads to a variety of interesting phases in which the system may exhibit magnetic and/or charge/orbital order in 2D and/or 3D.

Finally, although we have focused on the manganites here, it is important to note that other magnetically ordered oxides, such as the iron-based ferrites and the nickel-based nickelates, have similar structures (including layered phases) and display a similarly wide range of complex ordered phases originating from the interplay between charge, spin, lattice, and orbital DOF. These materials, particularly the ferrites, have been extensively studied in their own right using both conventional and ultrafast optical techniques; some examples will be given in the following sections.

Ferroelectrics

In analogy to magnetically ordered oxides, ferroelectric (FE) materials have a spontaneous electric polarization (P) below a critical temperature (T_c) that can be reversed by applying an external electric field (E). This property makes them useful for a wide variety of applications, including electronic memory, field-effect transistors, and capacitors, with a particular drive towards non-lead-based ferroelectrics in recent years. All ferroelectrics are non-centrosymmetric, piezoelectric, and pyroelectric in the FE phase. As with magnetic materials, FE materials exhibit domains (in which the orientation of P changes from domain to domain) and hysteresis (where P does not necessarily return to the same value as E is swept from positive to negative values and back). This enables much of the same theoretical machinery to be used to understand and model these materials.

Oxides with the perovskite structure (e.g., BaTiO_3 (BTO)) are perhaps the most common FE materials and will be our focus here. These oxides are a broad class of materials and thus exhibit different mechanisms for generating FE order. The most common mechanism is ionic displacement, where positive metallic ions (at the B site in the perovskite structure of Fig. 2) are displaced relative to the surrounding negatively charged oxygen octahedra (Fig. 4(a)); this occurs in the most well-known ferroelectrics, BTO and $\text{Pb}(\text{ZrTi})\text{O}_3$ (PZT), resulting in an infrared-active soft mode phonon that can be observed in the optical response. Other mechanisms include charge ordering and octahedral tilting, although this is more rare. Finally, as will be discussed in the next section, magnetic order can also generate FE order in certain materials.

Magnetoelectric multiferroics

Magnetoelectric (ME) multiferroics, materials that simultaneously possess both magnetic and electric order, have garnered substantial interest since they can exhibit ME coupling between these two order parameters; notably, this could enable control of the FE polarization with a magnetic field or magnetism with an electric field. This would facilitate the exploration of novel physical phenomena in these materials as well as the creation of new multifunctional devices (e.g., multiferroic elements in which information is stored in either the magnetization or the polarization and could be written and read by both electric and magnetic fields). However, despite substantial effort in this area over the past decade, strong ME coupling in single-phase materials has only been realized at low temperatures, primarily due to a lack of knowledge regarding the underlying mechanisms. Obtaining a deep understanding of ME coupling in different multiferroics has thus been a focus of research in this burgeoning field.

Multiferroics are generally divided into two groups, depending on the nature of the coupling between magnetic and electric order. In type-I multiferroics, magnetic and electric orders originate from different mechanisms, and therefore they tend to be weakly coupled. However, both ferroelectricity and magnetism often appear well above room temperature, and the FE polarization can reach fairly high values ($\sim 10\text{--}100\text{ C/cm}^2$). BiFeO_3 (BFO) is one of the most well known examples of this class of multiferroics, as FE order, appearing below $T_c = 1100\text{ K}$, originates from the ionic displacement mechanism discussed above (Fig. 4(a)) (and is further stabilized by the “lone pair” valence electrons of the Bi ion), while AFM order with a spiral spin structure arises below the Néel temperature of $T_N = 640\text{ K}$. Growth of BFO in thin film form causes a significant increase in the FE polarization, likely due to strain induced by the lattice mismatch between the film and substrate; this large polarization and room temperature coexistence of FE and AFM order has made BFO arguably the most heavily studied multiferroic material.

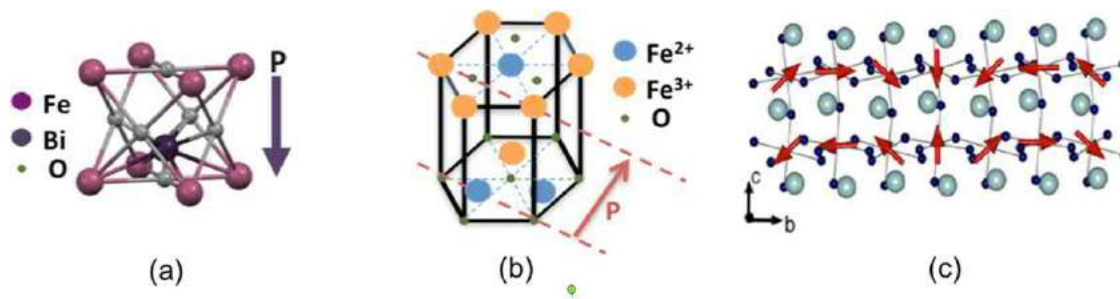


Fig. 4 (a) Structure of BFO, showing how the displacement of the Bi ion contributes to the FE polarization. (b) Charge ordering and FE polarization in LuFe_2O_4 . (c) Spin spiral order in TbMnO_3 . Reprinted from Kimura, T., 2007. Annual Review of Materials Research 37, 387.

Type-I multiferroicity can also occur through charge ordering; if a material contains transition metal ions with different valences, then FE order can emerge if the ions are ordered such that they can sustain a macroscopic polarization below the charge ordering temperature T_{CO} . LuFe_2O_4 (LFO) is perhaps the best known example of this class of multiferroics, with charge ordering below $T_{CO}=320$ K and AFM order below $T_N=240$ K that originates from its bilayered structure, in which charge ordering causes the FE polarization to appear in the direction between the two layers (Fig. 4(b)). However, it is worth noting that recent studies have indicated that LuFe_2O_4 may have nonpolar charge order, with no macroscopic FE polarization, calling the multiferroic nature of this material into question.

In contrast, in type-II multiferroics, ferroelectricity originates from magnetic order (through the Dzyaloshinskii–Moriya interaction), and therefore the two parameters are strongly coupled. However, the FE polarization is typically smaller than in type-I multiferroics, and strong ME coupling only occurs at low temperatures. The canonical example of a type-II multiferroic is TbMnO_3 , which is AFM ordered below $T_N=41$ K, with ferroelectricity emerging below $T_c=28$ K. In this material, Mn moments orient along the crystallographic b axis in a sinusoidal pattern at T_N and in a cycloidal spiral pattern at T_c (where the spins rotate around the a axis) (Fig. 4(c)). Furthermore, application of a strong magnetic field along the b axis can cause the FE polarization to “flop” from the c axis to the a axis, demonstrating the strong link between the FE polarization and spiral magnetic order in TbMnO_3 .

Finally, multiferroics display a number of low energy excitations that appear in the optical response, including IR-active phonons and magnons. Perhaps the most interesting excitations are electromagnons, magnetic excitations driven by the electric field of light that typically appear at low temperatures when magnetism and ferroelectricity are strongly coupled. These low energy (~ 1 – 10 meV) excitations were initially observed in TbMnO_3 and thus were thought to be intimately related to the mechanism underlying ME coupling. Electromagnons have since been observed in other materials, such as BFO and the non-multiferroic material $\text{Ba}_2\text{Mg}_2\text{Fe}_{12}\text{O}_{22}$. This last demonstration significantly expands their prospective impact, as one could potentially use light to control magnetic and/or FE order in any material supporting EMs (not only multiferroics), making them the subject of much interest in recent years.

Oxide heterostructures

Despite the many novel phenomena displayed by CEM in general and TMOs in particular, their discovery has often relied on serendipity, and the inability to fabricate crystalline TMO films with quality comparable to that of semiconductors has limited their applicability. However, this is rapidly changing, as TMO heterostructures can now be fabricated with a level of control and purity that is approaching that of semiconductor heterostructures. In these hybrid structures, two or more materials with a particular set of competing orders are interfaced in 1D, 2D and more recently, 3D patterns in order to tease apart the multiple DOF involved in the collective behavior that produces material functionality. Moreover, the interfaces often exhibit emergent properties that cannot be obtained in the individual constituents. For example, FM states in superlattices composed of AFM insulators have been observed, as well as extremely high charge mobility and superconductivity at the interface between insulating compounds. Interfaces also enable or enhance coupling between different order parameters that could not be achieved in single-phase materials, for example, strain-mediated ME coupling in multiferroic nanocomposites of magnetostrictive and electrostrictive materials. This has made the development and characterization of TMO heterostructures an extremely active field in current materials science and condensed matter physics.

One of the first discoveries in this field was that of a high mobility, superconducting 2DEG at the interface between insulating LaAlO_3 and SrTiO_3 . This stimulated a substantial amount of effort aimed at understanding the origin of this phenomenon and developing heterostructures composed of materials with different functionalities. In this vein, an important focus has been the development of multiferroic heterostructures, where ME coupling can be engineered by proper choice of the interface geometry and constituent materials. For example, a magnetostrictive/electrostrictive interface enables tuning the magnetization of FM CoFe_2O_4 through the strain generated by an electric field in FE BTO. Alternatively, several groups have grown heterostructures consisting of multiferroic BFO and a FM material such as LSMO or CoFe (Fig. 5(a)), although the microscopic mechanism governing the coupling across the BFO/FM interface was not unambiguously identified. Finally, heterostructures composed of superconducting (e.g., YBCO) and FM or AFM (e.g., various manganites) layers have displayed novel interfacial phenomena, including orbital reconstruction and suppressed superconductivity.

More recently, researchers have gone beyond TMO bilayers to develop novel superlattice and vertically aligned 3D heterostructures that offer additional control over functionality (Fig. 5(b)). As will be seen in Section “Time-Resolved Optical Spectroscopy of Correlated Electron Materials,” ultrafast optical spectroscopy is an essential tool for disentangling the interactions between the different DOF in these heterostructures through selectively photoexciting and probing their response in the time domain.

Heavy Fermion Materials

Heavy fermion (HF) materials display many of the phases encountered in other correlated electron materials, including SC, FM, and AFM order, and are thus an excellent testing ground for exploring these phenomena. In HF systems, a coherent heavy electron state forms below a characteristic temperature T^* , due to hybridization of the conduction electrons with localized f -electrons. This gives rise to a coherent “Kondo liquid” state in which the electron mass is enhanced and the resistivity drastically reduced. A wide spectrum of states, such as magnetism and superconductivity, can then emerge at low temperatures, depending on the interactions between the coherent heavy electrons and spins related to the localized f -electrons. These materials thus possess a complex phase diagram (Fig. 6), with phases that are both common to other CEM and unique to HF materials.

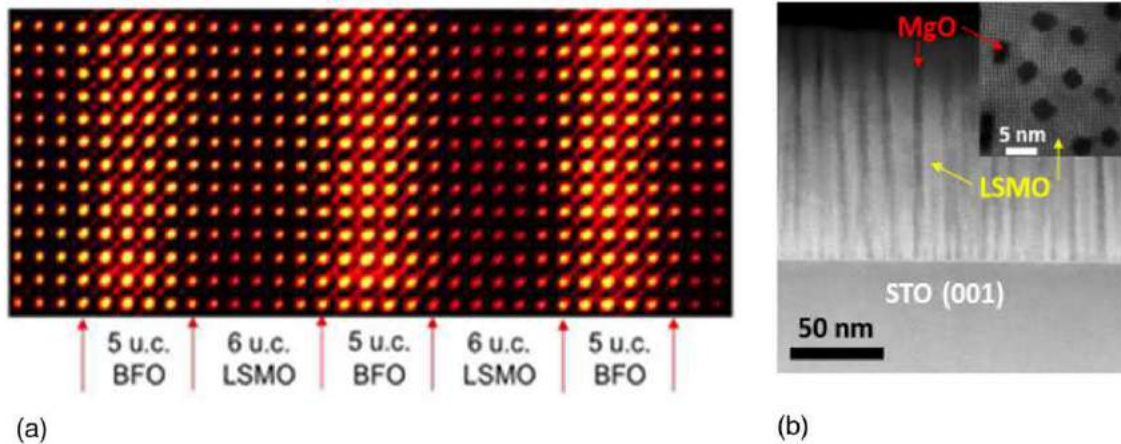


Fig. 5 Atomically sharp interfaces in (a) 1D superlattices (reprinted with permission from Singh, S., Haraldsen, J.T., Xiong J., *et al.*, 2014. *Physical Review Letters* 113, 047204 (supplementary info)) and (b) 2D TMO heterostructures used to control meso-to-macroscale functionality (reprinted with permission from Chen, A., Hu, J.M., Lu, P., *et al.*, 2016. *Science Advances* 2, e1600245).

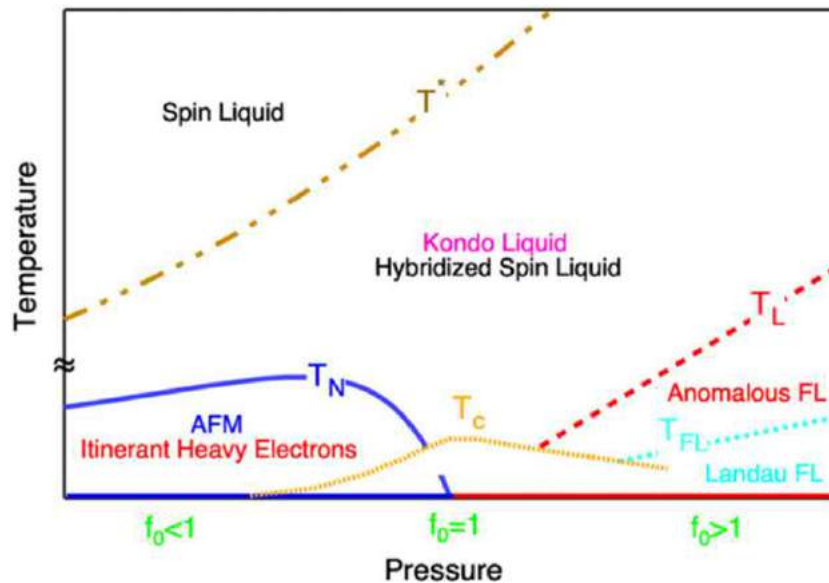


Fig. 6 Generic phase diagram for heavy fermion materials. Reprinted with permission from Yang, Y.-F., Pines, D., 2012. *Proceedings of the National Academy of Sciences* 109, E3060.

For most purposes, these materials may be treated as being primarily governed by the interactions between localized and itinerant electrons, making them a clean system for exploring strong correlations. However, the microscopic factors that govern the hybridization strength, V , and how it determines the resulting ground state are not well known. Optical measurements can shed light on this, since hybridization causes a gap (Δ) to appear in the electronic structure. This is manifested in the optical conductivity as a Drude peak centered at zero frequency, due to the reduction of scattering in the coherent state, as well as an IR peak due to transitions across the gap. Both time-integrated and time-resolved optical spectroscopy have thus been important in providing new insight into heavy fermion materials, as will be shown below, although the application of advanced UOS techniques (beyond all-optical pump-probe spectroscopy) to these systems is still relatively unexplored.

Time-Integrated Optical Spectroscopy of Correlated Electron Materials

Optical spectroscopy has been an invaluable tool for understanding the properties of correlated electron materials. Many of the extraordinary phenomena displayed by these materials exhibit unique signatures in the optical response, particularly at lower infrared photon energies ($\hbar\omega < \sim 1$ eV, where ω is the frequency). For example, optically active phonons and magnons have resonant peaks at mid-to-far-IR frequencies that can be used to probe their intrinsic properties. Similarly, metals exhibit a

characteristic increase in the conductivity at low frequencies (described by the Drude model), and semiconductors and insulators have a gap in their absorption at photon energies ranging from ~ 0.3 to 10 eV that leads to interband transitions in the optical response. Superconductors and heavy fermion materials also exhibit gaps at much smaller (< 100 meV) energies. Importantly, most optical techniques do not require contacts to the sample, making them relatively immune to the experimental artifacts that can affect data obtained by other characterization techniques.

In this section, we will give an overview of the optical response of CEMs. We will begin by describing the information contained within a generic optical conductivity spectrum for a CEM. We will then briefly describe the Drude-Lorentz model, arguably the most common starting point when fitting these spectra. It is worth noting that we will focus on information obtained from Fourier transform infrared (FTIR) and ultraviolet-visible (UV-visible) spectroscopy, and will not discuss other time-integrated optical techniques, such as Raman, photoluminescence (PL), and photoconductivity spectroscopy, although those techniques have often provided valuable insight into CEM as well. The discussion in this section will set the stage for understanding the results from UOS experiments discussed in the remainder of this article; as before, this is a general overview, and more detail is given in the reading list.

General Features of the Optical Response in Correlated Electron Materials

Fig. 7 displays a generic optical conductivity spectrum for correlated electron materials. Most CEM will exhibit some, if not all, of the features in this spectrum, making it worth a general description. At low photon energies (far-IR), materials that are metallic or semi-metallic display a Drude response, i.e., an increasing real part of the optical conductivity with decreasing frequency; this will be quantitatively described in the next section. Increasing the photon energy reveals low energy optically active magnon, phonon, and electromagnon excitations, typically modeled with a Lorentzian response (also described in the next section). These quasiparticle excitations are often obscured by the Drude response in metals, but can usually be seen in semiconductors and insulators. At still higher photon energies, interband transitions between occupied and unoccupied bands appear.

The optical conductivity can be measured as a function of various parameters, including temperature, pressure, and magnetic field. This enables researchers to track the evolution of CEM properties as a function of these parameters, giving insight into their properties. For example, tracking the Drude response as a function of temperature can indicate the appearance of a metal-insulator transition. Similarly, magnon excitations can appear when a system goes from paramagnetic to FM or AFM order, and soft mode phonons can emerge upon the onset of FE order. These measurements are also sensitive to the onset of polarons in, for example, CMR manganites, manifested through suppression of the Drude response at low frequencies and enhancement of the mid-IR resonant polaron absorption (i.e., spectral weight transfer). This further emphasizes that optical techniques are an important, non-contact probe of CEMs that can often give quantitative insight into their properties.

The Drude-Lorentz Model

The complex optical conductivity is given by $\sigma(\omega) = \sigma_1(\omega) + i\sigma_2(\omega)$. Experimentally, this is measured using FTIR spectroscopy (typically from low to intermediate frequencies (~ 0.01 – 5 eV)) and UV-visible spectroscopy at higher frequencies (~ 0.1 – 10 eV). For metallic systems (or any system with non-interacting carriers in a partially filled parabolic band), the Drude model is typically

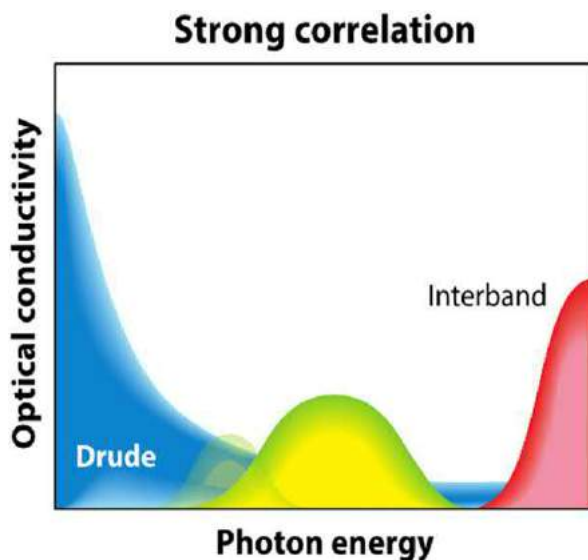


Fig. 7 Generic optical conductivity spectrum for CEM. Reprinted with permission from Zhang, J., Averitt, R.D., 2014. Annual Review of Materials Research 44, 19.

the first model used to fit the low frequency optical conductivity. This is obtained by classically considering a free charge driven by an electric field in a system with a scattering time, τ , between carrier–carrier collisions. At zero frequency, the DC conductivity is given by $\sigma_0 = \frac{Ne^2\tau}{m}$, where N is the carrier density and m is their mass. This can be generalized to finite frequencies:

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau}$$

This model yields a dependence of the optical conductivity on frequency that resembles the low frequency portion of Fig. 7. Although this is a classical model, it often provides a good approximation to the conductivity for a wide range of metallic systems. Furthermore, the Drude conductivity is often expressed in terms of the plasma frequency, $\omega_p = \sqrt{\frac{4\pi Ne^2}{m}}$:

$$\sigma(\omega) = \frac{\omega_p^2}{4\pi} \frac{\tau}{1 - i\omega\tau}$$

The Drude response is thus completely determined by the two frequencies ω_p and $1/\tau$, and by considering the ratio between these quantities one can often quickly get insight into material properties.

At higher frequencies, quasiparticle excitations (such as phonons and magnons) and interband transitions can be classically fit using a Lorentzian oscillator with a frequency ω_0 . Combining this with the Drude model gives the Drude–Lorentz conductivity,

$$\sigma(\omega) = \frac{Ne^2}{m} \frac{\omega}{i(\omega_0^2 - \omega^2) + \omega/\tau}$$

Again, although this is a classical model, it is surprisingly useful in modeling the response of a broad range of materials, particularly when electron–electron interactions are not significant (and even in some cases where they are, as will be shown below). Finally, it is worth noting that multiple oscillators can be used in situations when there are several quasiparticle excitations contained within the optical conductivity spectrum, making this approach quite general for fitting optical conductivity spectra at low to intermediate photon energies.

Time-Resolved Optical Spectroscopy of Correlated Electron Materials

Ultrafast optical spectroscopy provides the ability to temporally resolve phenomena at the fundamental timescales of atomic and electronic motion. From a materials science perspective, 10–100 fs temporal resolution combined with spectral selectivity enables detailed studies of electronic, spin, and lattice dynamics, and crucially, the coupling between these degrees of freedom, an especially important aspect for correlated electron materials. In this sense, UOS complements both “static” measurement techniques, such as time-integrated optical spectroscopy, and dynamic techniques, such as inelastic neutron scattering. Furthermore, UOS offers exciting possibilities to investigate non-equilibrium phenomena, such as selective photo-doping, to create a metastable state that would not, for a given material, be thermally accessible. It is now possible to routinely generate and detect 10–100 fs pulses in the visible, mid-, and far-IR portions of the electromagnetic spectrum, enabling novel time-resolved spectroscopic investigations. Further, the development of time-gated detection techniques (especially at THz frequencies) has enabled direct measurement of the electric field, which permits the determination of the complex conductivity $\sigma(\omega)$ over the spectral bandwidth of the probe pulse. In the context of CEM, spectral selectivity from approximately 0.001 to 4.0 eV is especially important since many relevant quasiparticle excitations lie in this range, including electron–phonon coupling, charge ordering, hybridization phenomena, phonon and polaron dynamics, and condensate relaxation and recovery.

Fig. 8(a) depicts, in general terms, how this is accomplished. An incident pump pulse induces a change in the optical properties of a sample on a 10–100-fs time scale. Subsequently, this perturbation is probed with a second ultrashort pulse, which can, depending on the wavelength and experimental setup, measure the pump-induced change in the reflectivity, transmission, conductivity, or polarization. By changing the relative delay of the pump and probe pulses it is possible to temporally measure the subsequent relaxation dynamics. The ensuing photoinduced changes can be described in terms of the complex dielectric function ($\varepsilon(\omega) = \varepsilon_1 + i\varepsilon_2$) of the sample. For example, the pump induced change in reflectivity (R), typically denoted as $\Delta R/R$, can be related to ε as follows:

$$\frac{\Delta R}{R} = \frac{\partial(\ln R)}{\partial \varepsilon_1} \frac{\partial \varepsilon_1}{\partial \kappa} \Delta \kappa + \frac{\partial(\ln R)}{\partial \varepsilon_2} \frac{\partial \varepsilon_2}{\partial \kappa} \Delta \kappa$$

where κ is a parameter that is modulated by the pump pulse and could be, for example, the temperature (T) or carrier density (N). Thus, the dynamics can be related to the complex permittivity (or equivalently the complex conductivity), which in turn is related to specific microscopic models. There are, of course, variations and more advanced approaches, but the ultimate goal is to relate dynamic changes in the dielectric function to specific microscopic phenomena. Approaches for modeling these dynamics range from simple rate equations, to density matrices, to Boltzmann transport equations.

Most often, the pump pulse creates an initially non-thermal electron distribution (Fig. 8(b) 1 → 2) fast enough that (to first order) there is no coupling to other degrees of freedom. During the first ~ 100 fs, the non-thermal distribution relaxes primarily by electron–electron scattering (Fig. 8(b), 2 → 3). Coherence effects due to the impulsive nature of the photoexcitation can also lead to a phase-coherent collective response, resulting in coherent magnons or phonons. Subsequently, the hot Fermi–Dirac distribution thermalizes through coupling to the other DOF (3 → 1) through, for example, electron–phonon coupling, phonon or magnon

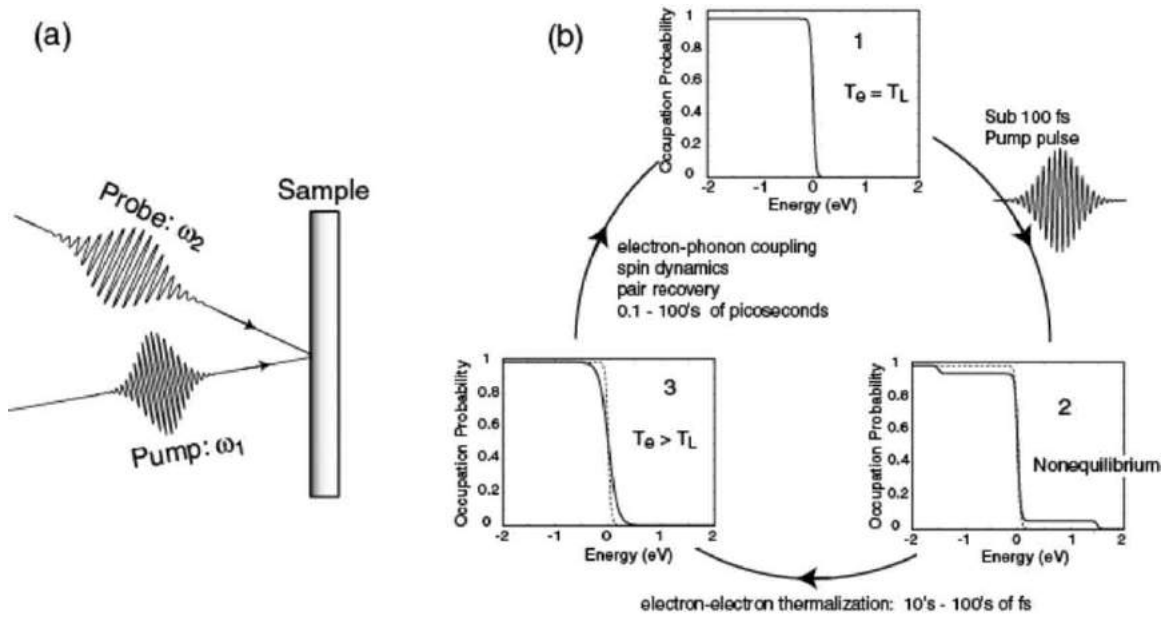


Fig. 8 (a) General pump-probe experimental setup. (b) Simple description of a pump-probe experiment in terms of changes in the occupation probability. 1. All DOF are initially in equilibrium. 2. Following ultrashort pulsed excitation, a non-equilibrium electron distribution is created which, primarily through electron-electron-scattering, relaxes to a Fermi-Dirac distribution, albeit at an elevated temperature in comparison to other DOF. 3. The quasiparticles subsequently relax through various (potentially competing) pathways in CEM, including electron-phonon relaxation, recombination/pair-recovery, or spin wave emission. Reprinted with permission from Averitt, R.D., Taylor, A.J., 2002. *Journal of Physics: Condensed Matter* 14, R1357.

emission, or recombination/pair-recovery. In some variations of UOS, a probe pulse is not required, since the pump pulse creates a nonlinear polarization in the material, leading to emission of radiation at THz frequencies that can be subsequently detected.

All-Optical Pump-Probe Spectroscopy of Correlated Electron Materials

Illustrating the dynamics in Fig. 8(b), electron phonon (e-ph) relaxation in conventional metals can be revealed through single color optical-pump, optical-probe (OPOP) experiments. Early experiments in noble metals, such as gold or silver, showed that relaxation occurs on a timescale of ~ 1 ps and is relatively temperature independent, as described by the two-temperature model. To explain these experiments, P.B. Allen proposed that electron energy relaxation occurs through their coupling to the lattice bath, connecting the measured relaxation time to the dimensionless e-ph coupling parameter λ , an important parameter in the BCS theory of superconductivity, with $\lambda \sim 1$ indicative of a potential superconductor. Early OPOP experiments at room temperature thus enabled the determination of λ , revealing $\lambda = 0.08$, 0.13 and 0.13 for the non-superconducting metals Cu, Au and Cr, respectively, while for the superconducting metals Nb, Pb, and NbN $\lambda = 1.16$, 1.45, and 0.95, respectively.

Following these measurements of e-ph relaxation and coupling in normal metals, OPOP experiments were used to characterize quasiparticle recombination in cuprate superconducting single crystals, with a view towards understanding pairing mechanisms in these unconventional superconductors. A systematic study performed on $\text{Y}_{1-x}\text{Ca}_x\text{Ba}_2\text{Cu}_3\text{O}_{7-\delta}$ as a function of temperature and doping using OPOP spectroscopy at 1.5 eV revealed i) a temperature-dependent slow (2–6 ps) relaxation process, corresponding to superconducting pair recovery across the superconducting gap, with pair breaking by phonons increasing the recovery time and ii) a temperature-independent fast (0.5 ps) component associated with the pseudogap. Two-color pump-probe measurements with a mid-IR probe (1.5 eV pump, 0.06–0.2 eV probe) on $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ (YBCO) thin films, in addition to confirming this behavior, observed that the amplitude of the signal followed the temperature dependence of the antiferromagnetic 41 meV peak observed in neutron scattering, supporting coupling between charge and spin excitations in these materials.

Detailed OPOP measurements (as well as optical-pump, THz-probe measurements, to be discussed in the next section) were also performed as a function of doping on films and single crystals of $\text{Y}_{1-x}\text{Ba}_x\text{Cu}_3\text{O}_{7-x}$ and $\text{Bi}_2\text{Sr}_2\text{Ca}_{1-y}\text{Dy}_y\text{Cu}_2\text{O}_{8+\delta}$, revealing the importance of bimolecular recombination kinetics, consistent with the interaction of opposite spin quasiparticles as they recombine into Cooper pairs. A sharp transition in the quasiparticle dynamics was observed at optimal doping in $\text{Bi}_2\text{Sr}_2\text{Ca}_{1-y}\text{Dy}_y\text{Cu}_2\text{O}_{8+\delta}$, with underdoped samples revealing bimolecular kinetics (recombination rate proportional to quasiparticle density) and with overdoped samples exhibiting a faster decay that is independent of excitation density.

More generally, the original motivation behind studying the recovery dynamics in superconductors on an ultrafast timescale was to understand the fundamental process of quasiparticle recombination (i.e., how free quasiparticles form into Cooper pairs). However, this process is complicated by the excitation of high energy ($> 2\Delta$) phonons during the photoexcitation process, since

these phonons can also break Cooper pairs. The Rothwarf-Taylor (RT) rate equations comprise a simple model describing many of the features of quasiparticle recombination in superconductors:

$$\begin{aligned} dn/dt &= \beta N - Rn^2 + I_0 \\ dN/dt &= \frac{1}{2} [Rn^2 - \beta N] - [N - N_T]/\tau_\gamma \end{aligned}$$

where n is the quasiparticle density, N is the excess density of high energy ($> 2\Delta$) phonons, τ_γ is the phonon relaxation time, R is the bare quasiparticle recombination rate, N_T is the equilibrium number of phonons at a temperature T , and β is the pair breaking coefficient for $> 2\Delta$ phonons. When N_T is small, the n^2 term dominates the recovery, but in many experiments the phonon decay significantly affects the recovery dynamics, a phenomenon described as the “phonon bottleneck.” The RT model can account for the intensity and temperature dependence of both the photoinduced quasiparticle density and the pair-breaking/superconducting state recovery dynamics in conventional as well as cuprate superconductors.

For example, an OPOP study of the photoexcited quasiparticle dynamics in the HTSC $\text{Ti}_2\text{Ba}_2\text{Ca}_2\text{Cu}_3\text{O}_y$ revealed that, deep into the superconducting state (below 40 K), a dramatic change was observed in the temporal dynamics, associated with photoexcited quasiparticles rejoining the condensate. An analysis based on the RT equations suggested that these dynamics resulted from the entry into a coexistence phase which opens a gap in the density of states (in addition to the superconducting gap) that competes with superconductivity, resulting in a depression of the superconducting gap. Similarly, competing orders and phase separation were revealed through a RT analysis of the quasiparticle relaxation dynamics in the underdoped HTSC $(\text{Ba,K})\text{Fe}_2\text{As}_2$, obtained through all-optical ultrafast pump-probe experiments. Specifically, it was observed that spin density wave (SDW) order forms at 85 K, followed by superconductivity at 28 K. Additionally, a normal-state order emerges, suppressing the SDW around ~ 60 K, which was hypothesized to constitute a precursor to superconductivity.

Adaptations of the Rothwarf-Taylor equations have been used to interpret ultrafast pump-probe experiments in other gapped CEMs. For example, the ultrafast carrier dynamics in heavy fermion systems exhibits dramatic nonlinear behavior as a function of temperature and fluence. Fig. 9 depicts the relaxation time τ obtained using OPOP spectroscopy at 1.5 eV in the prototypical HF compound YbAgCu_4 , which has a Kondo temperature $T_K \sim 100$ K and an electronic specific heat coefficient $\gamma = 210$ mJ/molK². Surprisingly, below T_K , τ increases by more than two orders of magnitude (to ~ 100 ps), in dramatic contrast to the nonmagnetic analog LuAgCu_4 , where $\tau < 1$ ps at all temperatures. This behavior of the relaxation dynamics is consistently observed in various HF metals and Kondo insulators. The temperature and fluence dependence of the observed dynamics in all of these systems can be modeled using a RT framework, where the dynamics are governed by the presence of the hybridization gap (see Section “Heavy Fermion Materials”). In addition, OPOP spectroscopy, combined with a RT analysis, was utilized to investigate carrier dynamics in the “hidden order” HF compound URu_2Si_2 (discussed in more detail below in Section “Time and Angle-Resolved Photoemission Spectroscopy”). The amplitude and decay time of the photoinduced reflectivity were found to increase in the vicinity of the coherence temperature $T^* \sim 57$ K, consistent with the presence of a 10 meV hybridization gap. However, at 25 K, a crossover regime is manifested as a new feature in the carrier dynamics that saturates below the hidden-order transition temperature of 17.5 K, indicating the formation of a 5 meV pseudogap that separates the normal Kondo lattice state from a hidden order phase.

Finally, in the itinerant HF antiferromagnet UNiGa_5 , OPOP measurements revealed a sharp increase in τ at the Neel temperature, $T_N = 85$ K, and exhibited a temperature dependence of τ below T_N that is consistent with the opening of a SDW gap

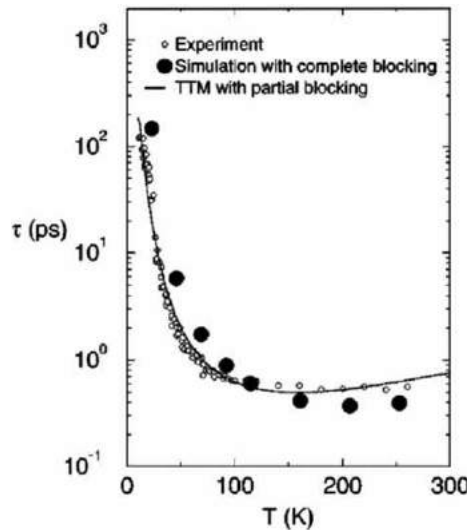


Fig. 9 e-ph relaxation time in YbAgCu_4 versus temperature. The solid line is a fit using the phenomenological two-temperature model, while the filled black circles are calculations using a semiclassical Boltzmann approach. Reprinted with permission from Demsar, J., Averitt, R.D., Ahn, K.H., *et al.*, 2003. Physical Review Letters 91, 027401.

leading to a quasiparticle recombination bottleneck, as revealed by the RT model. Following these results, ultrafast dynamics in the SDW phase of the similar itinerant HF antiferromagnet UPtGa₅ were also studied, revealing two relaxation components: (a) a slow component whose amplitude appears below T_N and relaxation time τ_{slow} exhibits an upturn near T_N , and (b) a fast component (τ_{fast}) that persists at all temperatures, also exhibiting an upturn near T_N . The differences with the dynamics observed in UNiGa₅ were explained using a RT framework and were primarily due to UPtGa₅ having A-type (rather than G-type) AFM order, resulting in partial Fermi surface nesting, partial gapping, and consequently a finite density of states at the Fermi surface.

Ultrafast Infrared Spectroscopy: THz to Mid-IR

Dynamic investigations of correlated electron materials in the mid-to-far-infrared spectral range are of particular interest, because the multitude of low energy excitations in this frequency range (such as charge, spin or superconducting gaps, polaronic excitations, and magnon or phonon modes) can be directly probed with ultrafast pump-probe techniques. Low energy photons also enable dynamical measurements of the optical conductivity, relevant, for example, to understanding the competing degrees of freedom involved in metal-insulator phase transitions. Recent advances in femtosecond pulse generation and amplification have enabled the efficient generation of intense ultrashort pulses from ultraviolet to mid-IR (50–500 meV) wavelengths using nonlinear optical parametric amplification and difference frequency generation processes. These pulses can be used as both pump and probe in ultrafast optical experiments, and their phase and amplitude can be manipulated to either coherently drive low-energy modes or monitor the evolution of relevant degrees of freedom following photoexcitation.

Ultrafast mid-IR spectroscopy has been applied to a broad range of CEMs, including high-temperature superconductors and colossal magnetoresistive oxides. As mentioned earlier (see Section “Transition Metal Oxides”), nanoscale inhomogeneities and phase coexistence play an important role in determining the electronic properties and macroscopic response to external electric or magnetic fields in CMR manganites. Ultrafast optical spectroscopy is sensitive to these inhomogeneities; for example, optical excitation can undress JT polarons by transferring electrons between Mn^{3+} and Mn^{4+} ions (Fig. 2(b)), which subsequently reform on a sub-picosecond timescale. Therefore, mid-IR spectroscopy, which can probe the polaron absorption peak observed in the optical conductivity, provides a powerful approach to study nanoscale phase separation in manganites by determining the evolution of the polaronic response with temperature. In particular, the persistence of specific polaronic signatures as the temperature is reduced below T_c should strongly indicate multiple phase coexistence.

Ultrafast mid-IR measurements of polaron dynamics were performed on the manganite Nd_{0.5}Sr_{0.5}MnO₃ (NSMO), revealing a mixed phase response. NSMO undergoes multiple phase transitions, from PM to FM ($T_c \sim 250$ K) to charge ordered (CO) states ($T_{CO} \sim 160$ K), accompanied by AFM ordering below T_{CO} (Fig. 10). Upon cooling towards T_c , correlations develop between polarons, resulting in the formation of nanoscale charge-orbital ordered clusters. The amplitude of the fast response, due to polarons, decreased as the temperature was lowered below T_c but remained relatively large all the way down to T_{CO} (Fig. 10). This indicated coexistence of the semiconducting PM phase with the FM metallic one, in agreement with both X-ray and neutron scattering measurements. Furthermore, at lower temperatures the amplitude of the fast polaronic response scaled with the volume fraction of the CO phase, further demonstrating that UOS can be a sensitive probe of nanoscale inhomogeneities in CEMs.

A contrasting example is provided by magnetoresistive pyrochlores, such as Tl₂Mn₂O₇, which lack JT-active Mn^{3+} sites (unlike CMR manganites), so that charge-lattice interactions (either polaronic or thermally induced) do not play a major role in the

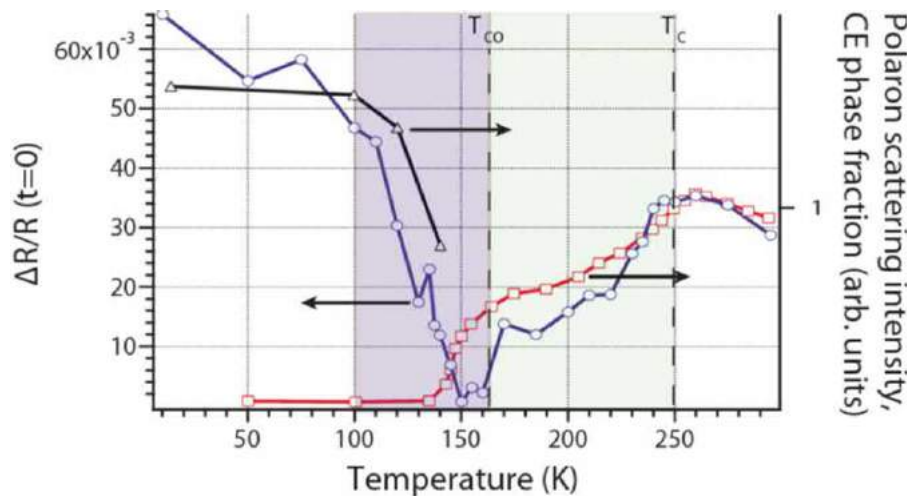


Fig. 10 Amplitude of the polaronic response at a probe wavelength of 13 μm at zero delay (blue line/circles) compared to single-polaron diffuse scattering measured by X-ray diffraction (red line/squares) and the antiferromagnetic volume phase fraction (black line/triangles) measured by neutron scattering. The shaded areas indicate temperature ranges in which minority phases compete with the FM (green) and COO (blue) majority phases. Reproduced from Prasankumar, R.P., Zvyagin, S., Kamenev, K.V., *et al.*, 2007. Physical Review B 76, 020402(R).

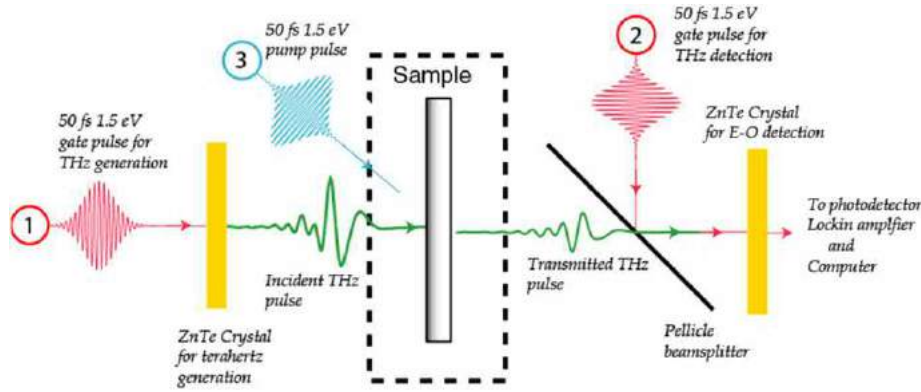


Fig. 11 Schematic of the typical experimental arrangement for THz-TDS and OPTP spectroscopies. Beam 1 generates THz generation via difference frequency mixing in nonlinear crystals (e.g., ZnTe) or photoconductive antennas. The THz radiation transmitted through the sample is measured through the electro-optic effect by overlapping it with beam 2 in the second crystal or antenna. By adding beam 3, the sample can be photoexcited and the photoinduced changes in the THz transmission measured as a function of the pump–probe delay.

emergence of CMR. Nevertheless, the use of ultrafast mid-IR spectroscopy to track photoexcited charge dynamics in $\text{Ti}_2\text{Mn}_2\text{O}_7$ demonstrated that charge localization in the insulating phase is caused by spin, rather than lattice, fluctuations. This implied that the CMR response in this material is determined primarily by the nanoscale inhomogeneities in spin ordering on Mn sites, unlike the double exchange and JT polaron mechanisms believed to govern the CMR manganites.

These studies of low-energy properties and materials dynamics can be further extended into the far-IR using sub-ps, broadband pulses with frequencies of $\sim 0.1\text{--}20$ THz that are generated and measured through photoconductive or electro-optic techniques (Fig. 11). Importantly, both the amplitude and phase of the THz electric field can be measured in the time domain, providing direct access to both the real and imaginary parts of the dielectric function or conductivity at THz frequencies through relatively simple Fourier transform routines. These THz time-domain spectroscopy (THz-TDS) experiments can be performed with extremely high signal-to-noise ratio without the need for cooled detectors, such as bolometers, constituting a major advantage of THz spectroscopy over conventional optical conductivity measurements using FTIR (see Section “Time-Integrated Optical Spectroscopy of Correlated Electron Materials”), which require complex Kramers–Kronig transforms to extract $\sigma(\omega)$. In addition, the same temporally coherent femtosecond pulses that generate and detect THz bursts can also be used for photoexcitation, enabling optical-pump THz-probe (OPTP) studies of photoinduced changes in $\sigma(\omega)$ with sub-ps time resolution.

THz spectroscopy is particularly well suited for investigations of metal-insulator phase transitions, where multiple charge scattering and localization processes compete to determine the ultimate conductivity. For example, OPTP has been used to unveil the relative contributions of spin and lattice fluctuations to the increasing resistivity of the FM metallic phase in CMR manganites. Systematic studies performed on optimally doped manganites using OPOP spectroscopy have shown that the photoinduced dynamics were dominated by two components: (1) a fast (few ps) temperature-independent decay associated with e-ph relaxation, similar to metals (see Section “All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials”), and (2) a slow (hundreds of ps) component, with the time constant peaking at the ferromagnetic T_c . The latter process corresponds to gradual spin disordering caused by energy transfer from the hot electrons to the lattice to the spin subsystem, and can be described by the three-temperature model (including electrons, phonons, and spins). This model was used to extract the electron–lattice, electron–spin, and spin–lattice energy transfer rates in various manganites, but could not quantify the role of particular subsystems in the degradation of the conductivity with increasing temperature.

OPTP provided a means for differentiating the effects of spin and lattice fluctuations on charge transport by following the photoinduced conductivity changes in the time domain. Fig. 12 shows the results of OPTP spectroscopy performed on films of the CMR manganites LCMO and LSMO. These measurements revealed the ability of THz experiments to obtain conductivity values both with and without optical excitation in a contactless manner. In addition, Fig. 12(a) demonstrates that time-resolved changes in the conductivity at various temperatures following photoexcitation with a 100 fs pulse at 1.55 eV evolve on two characteristic timescales. The fast component is related to electron-phonon relaxation, while the slow component is related to spin-lattice thermalization, as in the OPOP results discussed above. An analysis of the relative amplitudes of the fast and slow components observed in the OPTP response enabled the authors to determine the time-dependent spin and lattice temperatures T_{spin} and T_{phonon} , and quantify the relative importance of spin disorder ($\partial\sigma/\partial T_{\text{spin}}$) versus thermally disordered phonons ($\partial\sigma/\partial T_{\text{phonon}}$) in limiting the conductivity below T_c . Fig. 12(b) shows that the total $\partial\sigma/\partial T$ is primarily determined by thermally disordered phonons, while spin fluctuations dominate near T_c , in accordance with double-exchange mechanisms governing charge transport in CMR manganites.

OPTP has also been used to reveal the dynamics of phase separation in VO_2 , caused by the complex interplay between electron correlations and lattice re-arrangements. In this material, the MIT, as well as the accompanying structural dynamics, can be induced non-thermally using impulsive photoexcitation above the ~ 7 mJ/cm² fluence threshold. For this reason, the MIT in VO_2 has been extensively investigated by ultrafast techniques in an attempt to reveal whether electron localization derives from lattice distortions (band insulator) or on-site Coulomb repulsion (Mott insulator). OPOP measurements while varying the laser pulse duration

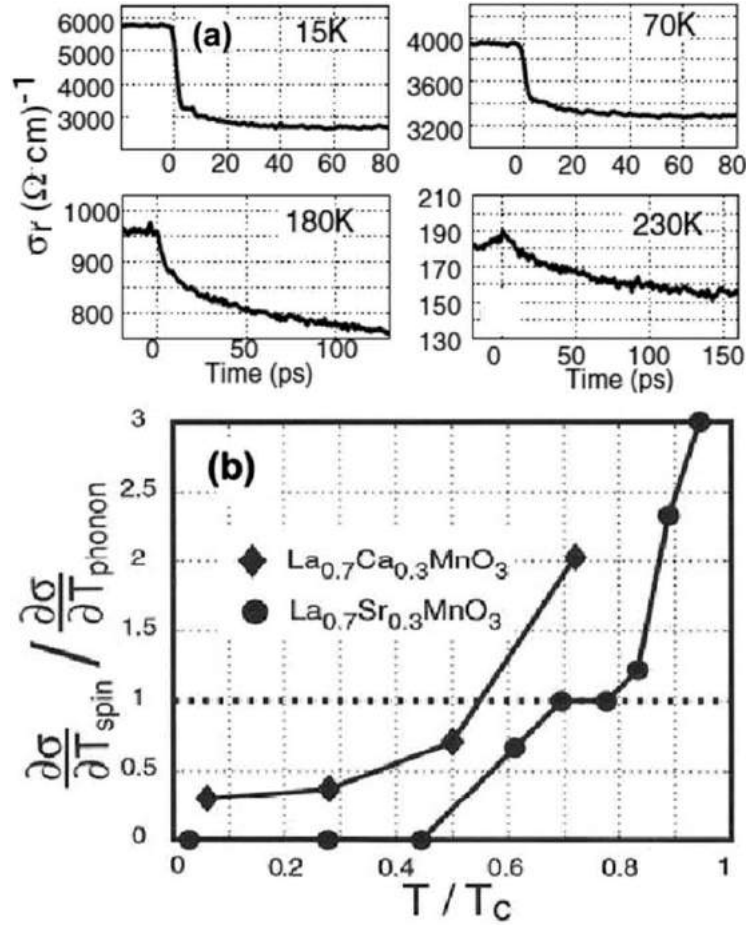


Fig. 12 (a) Two-component decrease in the real conductivity of photoexcited epitaxial LCMO films. (b) Relative contributions of the spin and phonon fluctuations to the conductivity changes as a function of temperature. Reprinted with permission from Averitt, R.D., Lobad, A.I., Kwon, C., *et al.*, 2001. Physical Review Letters 87, 017401.

showed that the intrinsic MIT timescale saturates at ~ 80 fs. This is too fast for thermally driven transitions that occur within e-ph thermalization times of a few ps, but is too slow for the almost instantaneous electronic dynamics expected to occur within the excitation pulsewidth. Therefore, the observed transition rates and concomitant optical phonon dynamics implied that the metallic phase in both the photoinduced and thermally driven MIT emerges primarily from atomic structural motion, rather than the time-varying electron correlations. Moreover, the strong dependence of the MIT rate on the electron excitation density also suggested that the metallic phase gradually grows from the confines of the optical absorption depth into the bulk. The dynamics of phase separation and domain formation near the MIT threshold are reflected in the time-dependent conductivity, which can be directly accessed by OTP. As illustrated in Fig. 13(a), the observed photoinduced conductivity changes for a given excitation fluence are significantly enhanced (up to $\sigma = 1000 \Omega^{-1} \text{cm}^{-1}$ in the metallic phase) as the temperature is increased toward $T_c = 340$ K. This response indicates a “softening” of the insulating state that leads to a reduction of the deposited energy required to drive the MIT in VO_2 (Fig. 13(b)). A simple effective medium model attributed this reduction to an increasing volume fraction of metallic precursors in the insulating bulk of the material as it approaches the MIT. These metallic seeds facilitate the growth and coalescence of a metallic conducting phase, initiated by absorption of the optical pulse. Inhomogeneity and phase coexistence are ubiquitous in CEM, and OTP provides new insight into the mechanisms governing the delicate balance between multiple competing interactions that leads to this behavior.

OTP has also been particularly useful for studying superconductors, since the superconducting condensate density is proportional to a $1/\omega$ component in the imaginary conductivity; these measurements therefore enable direct tracking of the condensate recovery dynamics. The first OTP measurements on superconductors were performed on underdoped and near optimally doped $\text{Y}_{1-x}\text{Ba}_2\text{Cu}_3\text{O}_{7-x}$ and $\text{Y}_{1-x}\text{Pr}_x\text{Ba}_2\text{Cu}_3\text{O}_7$ films. The optimally doped film exhibited a temperature-dependent recovery increasing from ~ 1.7 ps at 10 K to > 4 ps below the superconducting transition temperature (T_c), indicative of dynamics following the appearance of the superconducting gap. The lifetime decreased to ~ 2 ps above T_c , due to e-ph thermalization in the normal state. The underdoped films revealed a temperature-independent recovery time of ~ 3.5 ps even above T_c , indicative of dynamics influenced by a pseudogap.

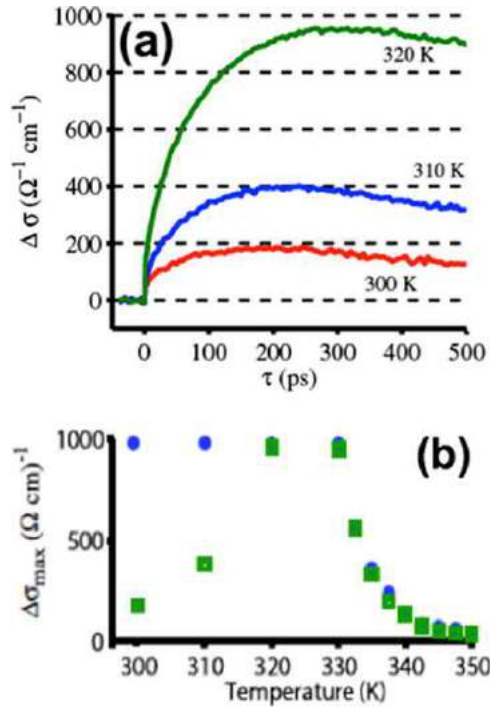


Fig. 13 (a) Induced conductivity change as a function of time for various initial temperatures. (b) Magnitudes of the photoinduced conductivity changes (green squares) measured with the same fluence at different temperatures and the maximum possible conductivity changes that can bring its magnitude to that of the metallic phase (blue circles). Reproduced from Hilton, D.J., Prasankumar, R.P., Fourmaux, S., *et al.*, 2007. Physical Review Letters 99, 226401.

Detailed OPTP studies of the recovery dynamics in the BCS superconductor MgB_2 as a function of photoexcitation fluence and temperature were analyzed using the RT equations (see Section “All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials”), enabling the determination of β , τ_r , and R . Remarkably, the pair breaking process was found to be quite slow, occurring on a timescale of ~ 10 ps, in contrast to the ~ 1 ps timescale observed for Pb, another BCS superconductor, as well as for the cuprate superconductors. This “anomalous” pair breaking dynamics is attributed to energy relaxation to high frequency phonons following photoexcitation instead of, as commonly assumed, electron-electron thermalization.

Apart from photoinduced changes in the conductivity, changes in the resonant absorption of THz radiation by low energy excitations (e.g., magnons, phonons, and polarons) manifest the emergence and disappearance of relevant order parameters as a CEM undergoes multiple phase transitions. Fig. 14(a) shows an example of this in the AFM multiferroic HoMnO_3 , which has a spin wave (magnon) resonance in the THz transmission spectrum when the material becomes magnetically ordered below $T_N = 42$ K. In general, AFM order is difficult to detect with optical techniques due to its net zero magnetization, which nullifies the magneto-optical Kerr effect. THz spectroscopy provides an alternative way of probing AFM order using the associated magnon resonances, and OPTP can unveil the coupling of spins to other degrees of freedom by separating them in the time domain.

The dynamics of spin-lattice relaxation in HoMnO_3 following photoexcitation is shown in Fig. 14(b). Importantly, not only the peak amplitude, but the entire spectrum of the THz pulse can be detected as a function of pump–probe delay, making the temporal evolution of the resonance lineshape readily available for analysis (insets in Fig. 14(b)). Photoinduced changes in the spectra of the transmitted THz E-field measured at different sample temperatures and pump fluences match well with how the magnon mode spectral shape would change with increasing sample temperature, as shown in Fig. 14(a). This implies that the ~ 4 – 12 ps rise time dynamics observed in the OPTP signal corresponds to the transfer of energy from the lattice (heated by much faster e-ph thermalization) to the spins, similar to that observed in FM CMR manganites. However, unlike in CMR manganites, spin-lattice relaxation in HoMnO_3 is much faster, indicating more efficient magnon-phonon scattering in AFMs. This represents a fundamental difference in spin dynamics in AFM and FM materials – the requirement for total spin conservation allows lattice vibrations to directly heat (or disorder) AFM spins, while spin-lattice thermalization in FMs relies on smaller, less efficient terms, like spin-orbit coupling.

Finally, THz emission spectroscopy, where THz fields are generated directly from the sample by optical photoexcitation, has emerged as a promising tool for investigating dynamic phenomena in CEM and other materials. For example, photoexcitation of an iron thin film with an intense optical pulse resulted in the emission of coherent far-IR radiation. This is surprising, since metals are generally good reflectors. Subsequent investigations of the emission efficiency as a function of the sample azimuthal angle (i.e., crystal symmetry) demonstrated that THz radiation is produced by ultrafast demagnetization on a ~ 2 ps timescale. The changing magnetic field necessarily results in a time-varying electric field, yielding a radiated signal, but these timescales are too short to be

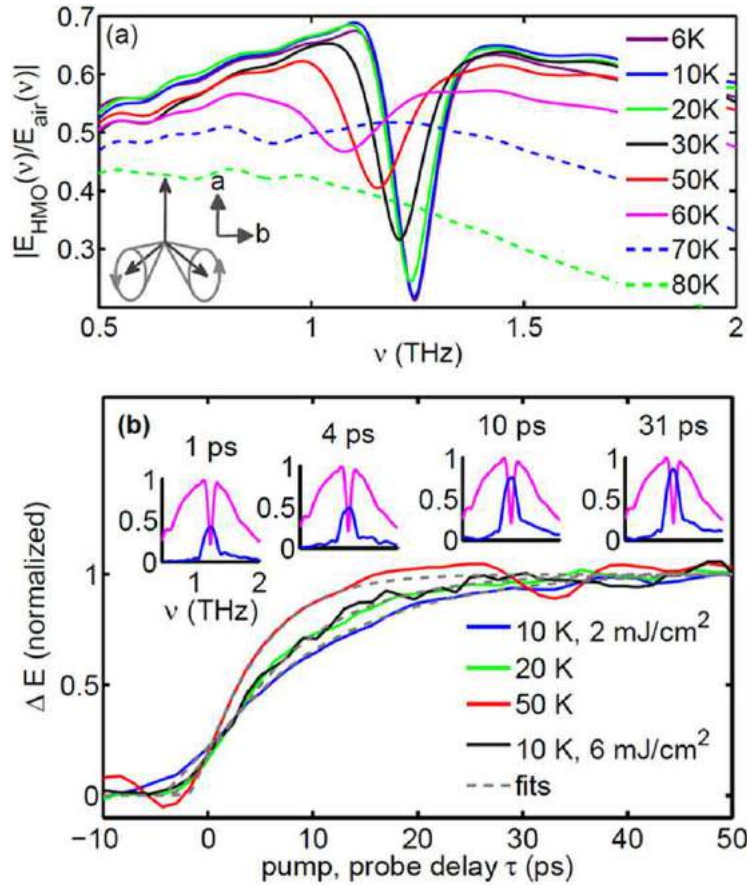


Fig. 14 (a) Temperature dependence of the magnon resonance in the THz transmission spectrum of HoMnO_3 . The magnon mode is illustrated at the bottom left. (b) Normalized OPTP signals taken at different temperatures and excitation fluences. The top insets show photoinduced changes in the THz transmission spectra for several different pump-probe delays (blue), overlapped with the THz transmission spectrum without photoexcitation. Reproduced from Bowlan, P., Trugman, S.A., Bowlan, J., *et al.*, 21016. *Physical Review B (Rapid Communications)* 94, 100404(R).

consistent with a scenario where spin waves are emitted following electron-lattice-spin energy transfer. It is instead likely that ultrafast demagnetization is a non-thermal process, caused either by spin currents or localized spin-orbit interactions. The strong dependence of the emitted THz signal on the magnetization and crystal symmetry might find numerous applications in studies of thermally and photoinduced phase transitions, with strong synergy between structural motion and spin ordering dynamics.

Photoinduced Phase Transitions and Mode-Selective Excitation

The subtle balance between competing interactions in CEMs leads to the coexistence of multiple, almost degenerate states that can serve as an ultimate ground state under proper perturbation. Doping, temperature variation or application of pressure and fields may alter the balance in favor of a desired order parameter and steer the material to assume the corresponding ground state. This multistability makes CEM an attractive choice for both technological applications and studies of complex phase transformation phenomena. However, the energy barriers separating the ground states can be rather large for conventional thermal or pressure drivers to initiate the phase transition (Fig. 15). Ultrafast spectroscopy offers an alternative way to steer the system between phase stability regions on the multidimensional potential energy surface by using (i) electronic excited states (e.g., photodoping) or (ii) coherent excitation of low-energy modes corresponding to relevant DOF (e.g., phonons or magnons), as illustrated by the arrows (i, ii) in Fig. 15. Although the original balance between orders will eventually be re-established by thermal fluctuations, important information can be obtained from these measurements about short-lived metastable states and recovery of the associated competing interactions.

Many investigations of photoinduced phase transitions (PIPT) in CEM used visible or near-IR photons to photodope the system (process (i) in Fig. 15) and broadband probes to follow the ensuing non-equilibrium dynamics. The most extensively studied PIPT is in VO_2 , where femtosecond photodoping was shown to induce a structural shift from monoclinic to rutile, accompanied by collapse of the gap and charge delocalization within ~ 100 fs. In addition, ultrafast electron diffraction was used to reveal a new metastable state of unidentified nature in a cuprate superconductor driven by 1.5 eV photons (see Section “Ultrafast Electron Diffraction”). Finally, ultrafast X-ray probes allowed deciphering of the dynamic coupling among charge order,

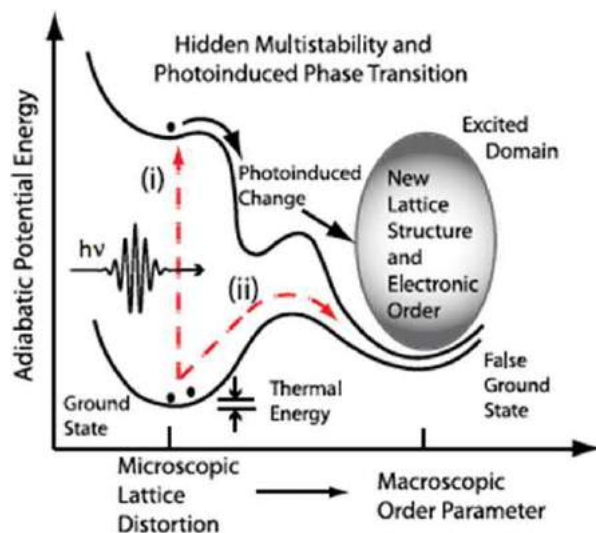


Fig. 15 Schematic representation of the potential energy dependence on a structural or other parameter. Nearly degenerate ground states are separated by an energy barrier, which can be overcome either through thermal fluctuations or by driving the system to the new phase through the excited (i) or ground (ii) states using photons of appropriate energy. Reproduced from Nasu, K., Ping, H., Mizouchi, H., 2001. *Journal of Physics: Condensed Matter* 13, R693.

orbital order and atomic structure during light-induced melting in a charge-and-orbital ordered manganite using a single time-dependent order parameter (see Section “Ultrafast Extreme UV and X-Ray Diffraction and Spectroscopy”).

These experiments provided much insight into the mechanisms governing PIPT, but have shown that interpretation of the time-resolved signals is very complicated and requires simultaneous observation of the dynamics of more than one DOF (e.g., electronic, lattice and spin). A powerful approach to resolving this issue would be to directly excite specific low-energy modes and observe their effect on a given order parameter as it evolves across multiple stability regions without leaving the system ground state (process (ii) in Fig. 15). Coherent lattice motion can be induced in several ways at optical frequencies. Rapid thermal expansion, produced by electron-phonon thermalization of photoexcited carriers, can launch an acoustic phonon wavepacket and modulate the local strain state within the pulse duration. Ionic rearrangement to accommodate photoinduced charge density changes can generate optical phonon oscillations. Alternatively, impulsive stimulated Raman scattering of a tailored train of non-resonant femtosecond pulses can drive lattice oscillations to the anharmonic regime. On the other hand, coherent induction of collective spin oscillations (magnons) can be accomplished through the inverse Faraday effect with circularly polarized femtosecond pulses. Nevertheless, all-optical excitation of low-energy DOF (and PIPT) often introduces a large thermal load on the sample, due to the dissipation of the excess photon energy through electron-phonon scattering, which interferes with non-thermal PIPT mechanisms and steers the order parameter in an undesired direction. Moreover, the number of optically accessible modes is limited by the symmetry of the material, leaving many relevant DOF inaccessible through optical processes.

The above shortcomings of *optically* exciting low energy modes are remedied by the use of intense mid-IR and THz pulses to *directly* excite IR-active modes. Modern optical parametric amplifiers and difference frequency generators supply broadly tunable, high energy mid-IR pulses, while the frequency mixing and tilted pulse-front techniques produce single-cycle THz pulses with electric and magnetic field amplitudes of ~ 1 MV/cm and 0.3 T or more, respectively. These pulses can now be used to directly photoexcite and probe specific low energy modes, improving upon previous all-optical techniques that only indirectly coupled to these excitations. Here, we describe several representative applications of these techniques to mode-selective excitation of correlated materials, while more detail on intense THz generation methods and investigations of a broader class of materials can be found in the reading list.

In the first application to correlated materials, mid-IR mode-selective excitation converted a manganite from insulating to metallic on sub-ps timescales. $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$ is the only perovskite manganite that remains insulating for all doping levels x . However, it was shown to have a thermally inaccessible metallic phase, which could be induced by a magnetic field or optical excitation. In perovskite manganites, the O-Mn-O bond angle is closely related to the inter-site electron hopping amplitude (t) and determines the ratio between electron bandwidth and on-site Coulomb repulsion (U) (see Section “Introduction”), hence determining the conductivity. Modulation of the O-Mn-O angle by resonant excitation of an IR-active Mn-O stretching mode (Fig. 16(a)) at ~ 71 meV was shown to change the reflectivity in the same manner as interband absorption of optical pulses (Fig. 16(b)). Simultaneous transport measurements exhibited a five order of magnitude increase in the DC conductivity, indicating a vibrationally induced MIT. The dominant role of Mn-O oscillations as a PIPT driver was further confirmed by matching the mid-IR excitation spectrum to the respective phonon resonance in optical conductivity (Fig. 16(c)). In subsequent studies of other manganites, resonant Mn-O excitations were demonstrated to induce ultrafast demagnetization and efficient orbital order melting,

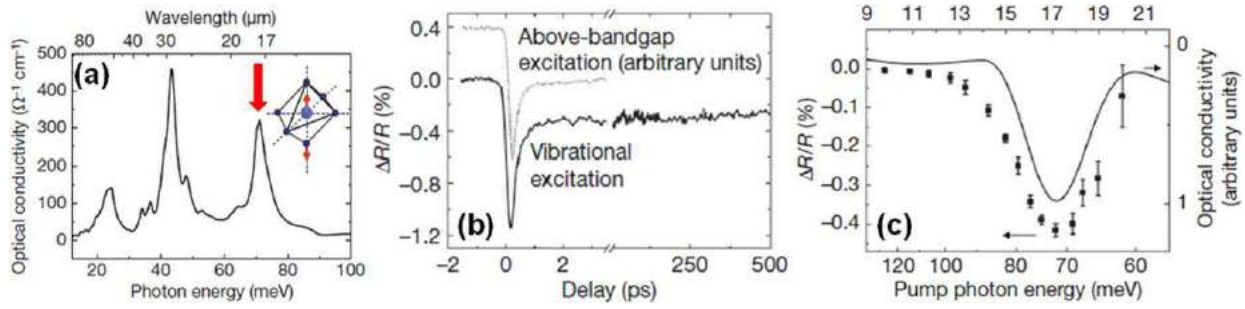


Fig. 16 Mode-selective excitation of the MIT in $\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$. (a) Mid-IR absorption spectrum showing the Mg-O stretching mode (inset) at ~ 70 meV (red arrow). (b) Comparison between the transient reflectivity changes at an 800 nm probe wavelength produced by mid-IR (solid line) and 800 nm (dotted line) excitation. The similarity between these dynamics implies that mid-IR pulses can drive the MIT. (c) Mid-IR static optical conductivity (solid line) and the amplitude of transient reflectivity changes as a function of pump wavelength (squares). Reproduced from Rini, M., Tobey, R., Dean, N., *et al.*, 2007. *Nature* 449, 72.

with absorbed energies of less than 1% of that necessary for thermally induced transitions. Interestingly, both electronic and magnetic orders in manganite heterostructures are amenable to external coherent strain modulation, caused by resonant vibrational excitation of stretching modes in the adjacent substrate.

Mode-selective excitation has yielded even more intriguing results in high- T_c superconductors. In these materials, the search for room temperature superconductivity has inspired intensive inquiry into the mechanisms responsible for converting parent AFM Mott insulators into cuprate HTSC by increasing the hole or electron doping. Across a large portion of the phase diagram, superconductivity in these compounds is suppressed by a rich variety of different order parameters. Mode-selective excitation can unveil the competing states that are hidden from other probes, and help in understanding the HTSC suppression and emergence mechanisms. For example, $\text{La}_{1.675}\text{Eu}_{0.2}\text{Sr}_{0.125}\text{CuO}_4$ belongs to the family of underdoped compounds, where HTSC is suppressed by a striped charge ordered state. Resonant mid-IR excitation of an in-plane Cu-O stretching mode at 80 meV was able to alter the balance between these competing orders and produce a broadband THz response, reminiscent of the equilibrium superconducting phase at higher doping. This transient superconducting state lasted for nanoseconds and persisted to temperatures well above the normal T_c for this doping level. This behavior was attributed to possible formation of an energy barrier that stabilized the new superconducting state against thermal fluctuations.

Intense THz pulses were also used to drive ionic motion into an anharmonic regime and map the potential energy of a resonant phonon mode in SrTiO_3 . The large ferroelectric polarization of SrTiO_3 is associated with a soft mode phonon, in which Ti^{4+} ions oscillate between O^{2-} ions on the opposite face centers. This mode resonantly absorbs 1.5 THz radiation when a low amplitude THz field is used. However, when the field was increased to 80 kV/cm, the phonon resonance shifted to higher frequencies due to dynamic stiffening of the confinement potential. This was an indication that the vibrational amplitude reached the anharmonic (quartic) part of the potential, with the photoinduced ion displacements exceeding 1 picometer (pm) – a magnitude comparable to the static ionic displacements in the FE state. Thus, intense coherent THz excitation can potentially be used to switch the macroscopic FE polarization on picosecond timescales.

Ultrafast control of magnetization dynamics promises numerous applications in magnetic switching devices. Manipulation of FM order can be achieved with polarized optical pulses or the magnetic field of intense THz pulses, but control over potentially faster AFM switching is much harder to implement. However, the large magnetic fields available with intense THz pulses enable resonant excitation and coherent manipulation of spin dynamics in AFM materials. In the AFM oxide NiO , the broadband magnetic field of a single-cycle THz pulse triggered long-lived precession of antiferromagnetically aligned spins, with a resonant magnon frequency of 1 THz. A sequence of THz pulses, spaced by 6.5 precession periods, was then used to coherently quench coherent spin oscillations. Similarly, as described in Section “Magnetoelectric multiferroics,” multiferroics feature new elementary excitations, electromagnons, which are expected to mediate and enable ultrafast switching of AFM order in multiferroics using transient electric fields. Kubacka *et al.* used resonant THz fields to excite electromagnon oscillations in the archetypical multiferroic TbMnO_3 . Resonant X-ray scattering provided direct insight into the induced spin dynamics, demonstrating that the THz field caused synchronous rotation of the spin cycloid plane, with a 4° amplitude. Even larger rotational motion can be achieved by increasing the THz field, reaching $\sim 90^\circ$ for 1–2 MV/cm.

Apart from driving specific modes and observing their effect on various order parameters, the utility of intense THz pulses lies in their ability to induce and monitor dynamic phase transitions, similar to the mid-IR radiation discussed above. As with many other techniques, the MIT in VO_2 provides a fertile ground for exploring the capabilities of coherent THz excitation. Liu *et al.* demonstrated that intense THz excitation with peak fields of 300 kV/cm was insufficient to induce the metallic phase in VO_2 . In order to determine the feasibility and threshold of such an MIT, they created resonant metal structures on the sample surface, where the electric field of the THz pulse could be enhanced by a factor of ~ 30 in the gap between closely spaced metal stripes. This enhancement enabled them to break the MIT threshold, transforming VO_2 into a metallic state within a few ps of excitation. Moreover, they found that fields in excess of ~ 4 MV/cm caused pronounced damage to the material due to electric breakdown.

Ultrafast Extreme UV and X-Ray Diffraction and Spectroscopy

The exotic properties of correlated electron materials emerge from a labyrinthine pattern of complex interactions among charge, lattice, spin and orbital degrees of freedom. Therefore, our understanding of the microscopic mechanisms underlying both static and transient material behavior would greatly benefit from the ability to selectively probe the dynamics and coupling between relevant excitations and order parameters. However, UOS in the visible/UV range often relies on the indirect coupling of microscopic processes to observable variations in the macroscopic dielectric function, and is limited to a material response averaged over large spatial areas and multiple concurrent dynamics. In Sections “Ultrafast Infrared Spectroscopy: THz to Mid-IR” and “Photoinduced Phase Transitions and Mode-Selective Excitation,” we described how the expansion of UOS to mid-IR/THz frequencies enabled specificity in probing and driving low-energy modes. Similar opportunities exist on the other side of the spectrum, where a wealth of ultrafast X-ray techniques provide sub-nanometer spatial resolution, selectivity between charge, spin and lattice dynamics, and high element specificity.

Several different schemes have been implemented for the generation of femtosecond X-ray pulses. Tabletop plasma sources can produce hard (6–10 keV) X-ray bursts by focusing terawatt 1.5 eV femtosecond laser pulses onto a moving metal wire. Alternatively, high harmonic generation (HHG) can convert near and mid-IR radiation to coherent soft-X-ray pulses of variable polarization, with up to keV photon energies. Furthermore, at large facilities, laser slicing of pulsed synchrotron radiation can reduce the X-ray pulse duration to femtoseconds while still preserving the wide energy tunability of the emitted photons. Finally, in the past decade, accelerator-based X-ray free electron lasers (XFEL), such as the Linac Coherent Light Source (LCLS) at Stanford, have appeared on the scene, with the promise to open new areas in ultrafast materials science by producing femtosecond pulses with very high brightness over a wide photon energy range.

One of the earliest applications of ultrafast X-ray spectroscopies focused on correlating the lattice dynamics and electronic structure changes during the photoinduced MIT in VO₂. As illustrated in Fig. 17(a), X-ray diffraction (XRD) measurements on VO₂ below the transition temperature at negative delays (i.e., on an unperturbed sample) show a broad peak around 13.9° that corresponds to the low-temperature insulating monoclinic crystalline structure. Following above-threshold photoexcitation, the structure changes to rutile, as indicated by the appearance of the corresponding diffraction peak around 13.8°. This transition occurs within 100 fs and correlates well with the dynamics of rutile metallic phase formation observed in OPOP experiments. The direct insight into atomic motion given by XRD was complemented by X-ray absorption measurements, which revealed concomitant dynamics in the electronic structure during the MIT. Near-edge X-ray absorption spectroscopy (XAS) probes the transitions from core atomic levels to the valence or conduction band, and is thus very sensitive to variations in the electronic structure, level population and ionic environments. Fig. 17(b) and (c) show time-resolved XAS at the V_{2p}–V_{3d} and O_{1s}–V_{3d} absorption edges, respectively. The photoinduced absorption increase at the vanadium V_{2p} resonance can be explained by rapid band gap collapse,

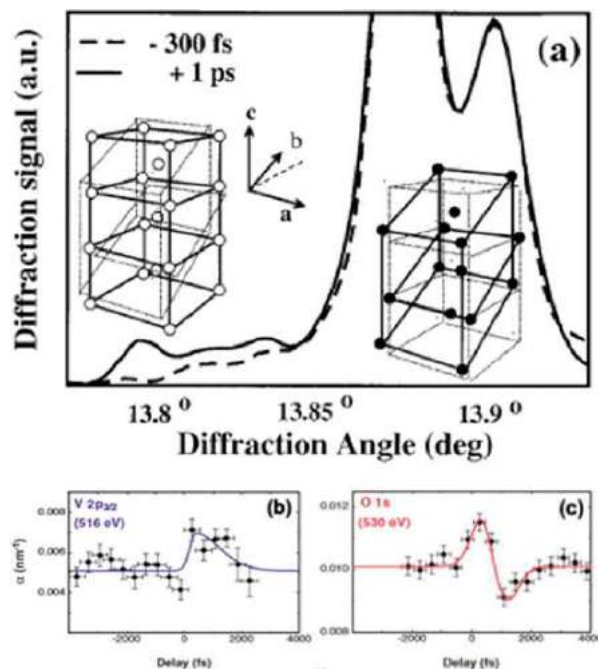


Fig. 17 (a) X-ray diffraction signal measured before (dashed) and after (solid) excitation of VO₂. for negative and positive time delays. The increase in diffraction around the 13.8° rutile peak at positive delay corresponds to the growing volume of the metallic phase during the photoinduced MIT (Reproduced from Cavalleri, A., Toth, Cs., Siders, C.W., *et al.*, 2001. *Physical Review Letters* 87, 237401.). (b, c) X-ray absorption dynamics, measured at the V_{2p} and O_{1s} resonances, is indicative of a gap opening in the electronic structure as the MIT develops in time. Reproduced from Cavalleri, A., Rini, M., Chong, H.H.W., *et al.*, 2005. *Physical Review Letters* 95, 067405.

with subsequent rapid (~ 1 ps) electron-phonon thermalization within the conduction band. This confirms the non-thermal nature of the photoinduced MIT in VO_2 , because the energy gap disappears before the lattice is heated above T_c . However, the positive to negative XAS dynamics measured at the O_{1s} absorption edge (Fig. 17(c)) cannot be explained by gap collapse alone, and indicates core level shifts induced by photodoping and dynamic valence changes of the O^{2-} and V^{4+} ions.

These studies demonstrated the ability of ultrafast X-ray spectroscopy to separate atomic motion and electronic dynamics within bands with different orbital symmetries. In other experiments, sensitivity of the hard XRD efficiency to the lattice, orbital, and charge periodicities, and resonant enhancement of the soft X-ray scattering amplitude at the ionic absorption edges were exploited to selectively probe the dynamics of co-existing charge, orbital and magnetic order in correlated materials. Studies of phase transitions induced by coherent lattice, spin and charge excitations in bulk and heterostructured materials are particularly interesting, even essential, for both fundamental understanding of materials behavior and technological applications.

It is worth mentioning an important recent development in ultrafast X-ray science – the generation of attosecond (as) pulses. Because of the large frequency bandwidth (~ 18 eV for 100 as) required to support such a short pulsewidth, HHG or other nonlinear techniques that produce radiation in the extreme UV or soft X-ray portion of the electromagnetic spectrum have to be used. Attosecond spectroscopy should provide unprecedented insight into electron motion, which happens on an extremely fast timescale. For instance, electron-electron thermalization occurs within 10–100 fs, while other fundamental processes such as screening, charge imaging, and non-thermal electron generation occur on an attosecond timescale. Initial experiments in atomic gases and molecules thus show the great promise of attosecond spectroscopy for resolving electron dynamics; however, very few experiments have been performed in solid state systems, and thus far only to time the delay (~ 100 as) between electron emission from the core and delocalized energy states. Further advances in source reliability and detection techniques are thus necessary to harness attosecond pulses for materials science.

Ultrafast Electron Diffraction

Ultrafast electron diffraction (UED) offers an attractive alternative to X-rays for exploring ultrafast structural dynamics in correlated electron materials, especially low dimensional ones, for example, films or monolayers. The wavelength of electrons is shorter than typical X-ray photons (~ 20 pm for 30 keV electrons versus ~ 120 pm for 10 keV photons), which reduces the diffraction angles and allows more Bragg peaks to appear within the same detector area. This significantly increases the amount of structural information that can be extracted from UED experiments. Furthermore, electrons interact with matter much more strongly than X-rays and are more sensitive to small crystals or microscale sample regions. This is also a disadvantage due to the importance of multiple scattering events, which restricts the electron penetration depth ($\sim$ tens of nm for 10 keV electrons) and might complicate structural determination from the analysis of ED patterns from bulk crystals. Nevertheless, UED is gaining increasing attention, with a growing number of applications to CEM.

UED is a pump-probe technique in which femtosecond UV probe pulses are used to generate electron packets by photoemission from a cathode material. The electrons are then accelerated towards the sample to produce diffraction patterns upon transmission or grazing reflection from the photoexcited material. In table-top systems, electrostatic fields are usually used to reach electron energies up to 100 keV. In order to probe a larger region of reciprocal space or increase the penetration depth, 2–6 MeV electrons can be generated with radio-frequency driven linear accelerators. The temporal resolution of UED is primarily determined by the electron bunch duration, which is limited by Coulomb repulsion to 200–300 fs. Different concepts have been developed to address this issue, extending all the way to producing single-electron pulses with high repetition rates to reach the 100 fs limit. Geometrical broadening due to different pump and probe incidence angles and other propagation effects can also be corrected by tilting the front of the optical excitation pulse. Finally, similar to ultrafast X-ray techniques, efforts have been made to combine UED with spectroscopic measurements in order to correlate structural dynamics to the evolution of electronic and other properties.

The majority of previous UED studies on CEM focused on photoinduced phase transitions, with the goal of separating structural dynamics from other DOF. In the high-temperature superconductor $\text{La}_2\text{CuO}_{4+\delta}$, the emergence of a non-equilibrium structural shift within 27 ps after photon absorption was observed with UED. As previously discussed, undoped HTSC are antiferromagnetic Mott insulators, and hole or electron doping induces superconductivity below the critical temperature (T_c) with a metallic state above it. A photoinduced structural change occurred only when the photodoping reached ~ 0.12 photons per copper site, which was very close to the optimal chemical doping level of 0.16 holes per copper site. The nature of the PIPT in this material was not identified, but possible light-induced charge redistribution might be responsible for driving the system through the inverse Mott transition.

CDW materials feature strong coupling between periodic charge and lattice modulations. The trajectories and cooperative character of charge and atomic motion can be independently followed by simultaneously observing the dynamics of lattice Bragg peaks and CDW satellite peaks using UED. In these experiments, melting of the charge order within hundreds of fs was identified, accompanied by collective lattice oscillations. The ensuing charge and lattice motions demonstrate well-defined separation of the coupled electron-lattice and CDW order parameters, leading to a commensurate-to-incommensurate CDW transition.

Finally, VO_2 was a natural candidate for UED studies, due to the unresolved question of whether the photoinduced MIT is structurally or charge driven. The ability to simultaneously detect multiple diffraction peaks (hence, crystallographic directions) as a function of time provided unprecedented insight into atomic motion during the MIT. Fig. 18(a) shows three characteristic timescales at which the diffraction intensity of Bragg peaks evolves following excitation with a 1.5 eV pulse. Immediately after

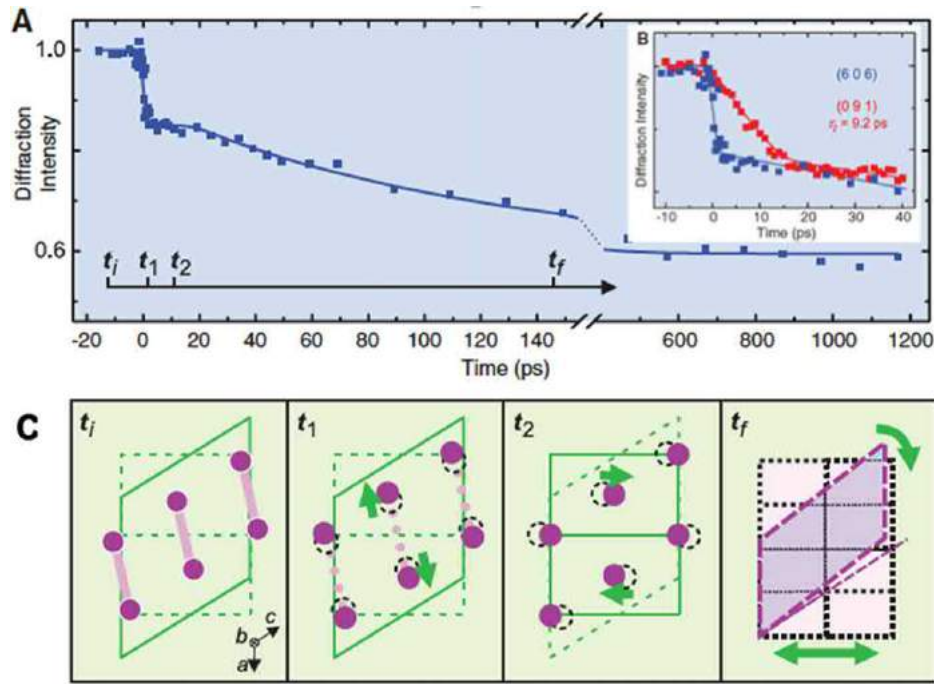


Fig. 18 (a) Intensity change of the (606) diffraction peak, associated with ionic motion along the V–V bond. Three characteristic timescales are also shown. (b) The blue curve describes the dynamics of V–V bond elongation, as measured by the (606) peak, and the red curve shows the translation of V ions normal to the V–V bond (091 peak) dynamics. (c) Schematics of the step-wise structural changes inside the unit cell and, on larger length scales, of shear motion following photoexcitation of VO₂. Reproduced from Baum, P., Yang, D.S., Zewail, A.H., 2007. Science 318, 788.

photon absorption, the peak intensity drops within t_1 and remains almost constant till t_2 . Interestingly, t_1 had different values of 0.3 ps and 10 ps for two groups of UED peaks, which were associated with the atomic motion of vanadium ions along and across the V–V bond, respectively (Fig. 18(b)). The eventual saturation at t_2 was succeeded by much slower decay dynamics, on a timescale of $t_3 \sim 100$ ps. This kinetic behavior (and especially the discreteness of t_1) implied that the structural phase transition proceeded in the stepwise manner depicted in Fig. 18(c). At a time t_i , photon absorption transfers the charge into a vanadium d -band with antibonding character, which induces repulsion between V atoms and their motion away from each other along the bond direction. This motion occurs at $t_1 \sim 0.3$ ps and affects only the corresponding Bragg peak intensities. Subsequent rotation of the V–V bond and unit cell transformation toward the metallic rutile configuration within $t_2 \sim 10$ ps can only be observed in another set of Bragg peaks. Finally, continued change of the peak intensity at $t_3 \sim 100$ ps indicates a shear that propagates into the bulk of the material and converts the crystalline structure to rutile. This process agrees well with the X-ray diffraction measurements in Fig. 17(a). Moreover, comparison with OPOP and OTP results indicated that metallic states appear while the material is still undergoing the monoclinic to rutile transition, i.e., V–V motion at t_1 . This confirmed the presence of an intermediate monoclinic metal phase, similar to the strongly correlated metal state observed by IR microscopy during the thermally-driven MIT.

Time and Angle-Resolved Photoemission Spectroscopy

Knowledge of the electronic energy structure in momentum space and its evolution under external stimuli are essential for understanding the (non)equilibrium properties of correlated electron materials. While powerful, most UOS techniques are intrinsically momentum independent and cannot be used to track carrier relaxation dynamics across momentum space. On the other hand, static angle-resolved photoemission spectroscopy (ARPES) is capable of mapping occupied electronic states in momentum space. ARPES significantly advanced our understanding of correlated systems by revealing the dispersion of quasi-particles and signatures of many-body correlations in reciprocal space. Adding ultrafast time resolution to ARPES (trARPES) gives completely new insight into electronic structure dynamics, both below and above the Fermi level, and provides a complete set of information on the momentum-, energy-, and time-dependent excitations responsible for the properties of CEM.

ARPES is based on electron ejection from a crystalline material after illumination by UV photons with an energy exceeding the material work function (4–5 eV). The energy and momentum of the electrons before and after photoemission are related by conservation laws and can be measured with a proper spectrometer to create energy-momentum dispersion maps. In trARPES, ionizing radiation is usually produced either by upconverting the energy of 1.5 eV laser photons to 6–9 eV in a nonlinear crystal or by generating 10–60 eV photons through HHG in gases. HHG-based trARPES allows probing a larger volume within momentum space due to its higher photon energies, albeit at the cost of lower energy resolution (70–200 meV versus ~ 20 meV obtained with

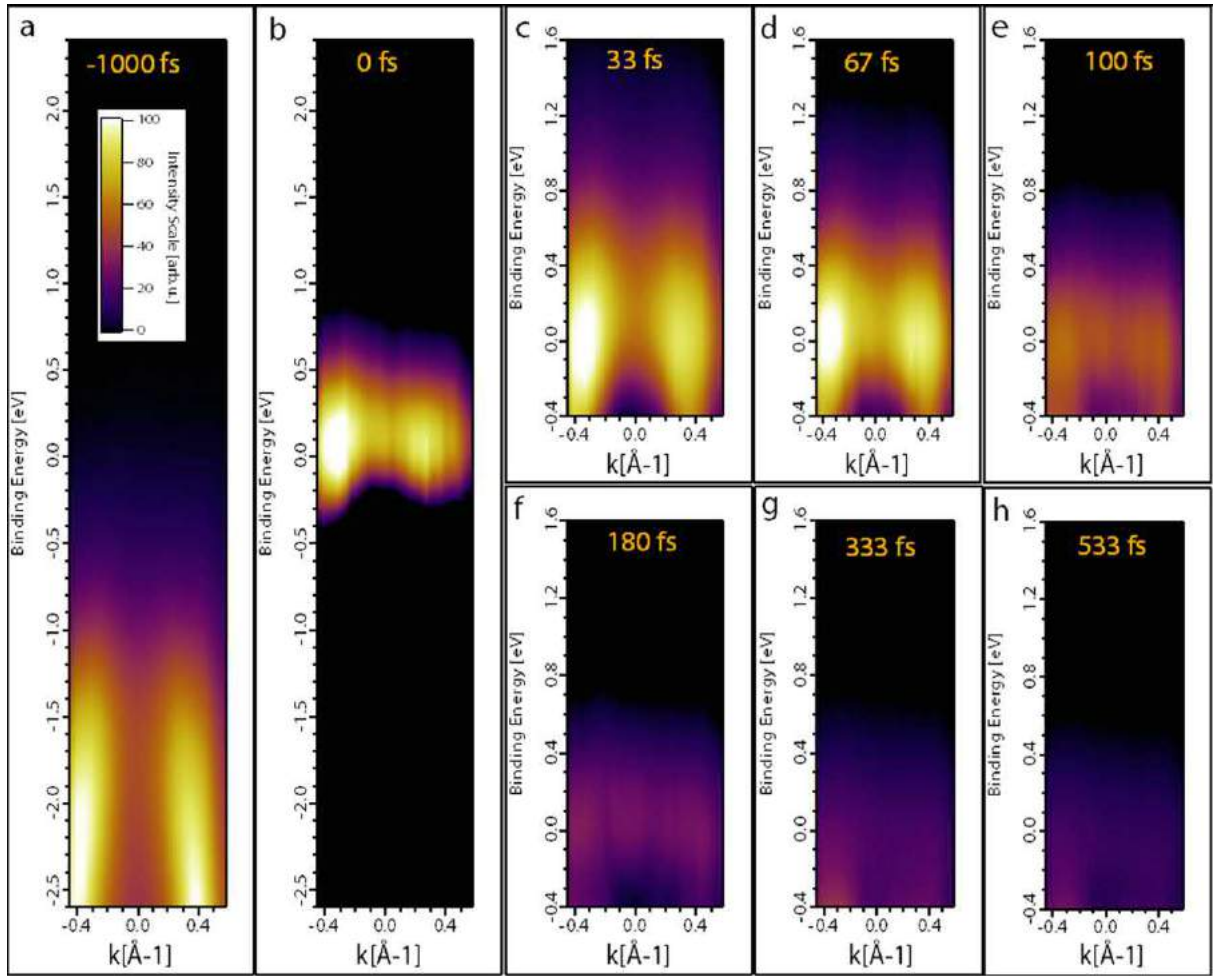


Fig. 19 Time-resolved ARPES study of URu_2Si_2 in the hidden order state. (a) Energy structure before the pump pulse arrives. (b) The double structure above the Fermi level appears after photoexcitation and corresponds to the new, long-lived quasiparticles, decaying in density as shown in (c–h) with a characteristic lifetime of 213 fs. Reproduced from Dakovski, G.L., Li, Y., Gilbertson, S.M., *et al.*, 2011. *Physical Review B* 84, 161103(R).

upconversion techniques). In the past decade, ultrafast trARPES has been extensively used to explore the non-equilibrium behavior of a wide range of materials. Among correlated electron materials, dynamics in HTSC, charge density wave compounds and heavy fermion actinides have been closely investigated.

An excellent example of trARPES capabilities is the investigation of band renormalization processes during the phase transition in URu_2Si_2 , one of the most heavily studied heavy fermion compounds due to the peculiar hidden order (HO) state occurring below 17.5 K. Numerous theories have been proposed over the past two decades to describe this state, but the exact nature of the HO order parameter still remains unclear. As described in Section “All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials,” ultrafast OPOP experiments identified a pseudogap region ($17.5 \text{ K} < T < 25 \text{ K}$) separating the normal Kondo-lattice state, with a 10 meV hybridization gap, from the HO phase, with a 5 meV gap. trARPES was then applied to determine the relative positions of both gaps in energy and momentum space, because optical excitation can populate the states above E_F , enabling their observation with a temporally delayed ARPES probe (compare the static and pumped ARPES results in Fig. 19(a) and (b), respectively). Time-resolved measurements of the quasiparticle population and decay dynamics revealed a multi-peak structure at the upper edge of the HO gap (Fig. 19(b–h)), where quasiparticles accumulate for hundreds of fs due to the phonon bottleneck (see the discussion of the RT modeling of OPOP results in Section “All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials,”). These results suggested that the HO gap forms as a momentum-dependent modification of the momentum-independent hybridization gap, which might explain the contradictory measurements of the transport properties and electronic structure in this compound by other time-averaged techniques.

Ultrafast Second Harmonic Generation Spectroscopy

Nonlinear frequency conversion methods play an important role in optics and laser physics. The conversion efficiency is very sensitive to the properties of the nonlinear medium, since it involves mixing of multiple electromagnetic waves with different

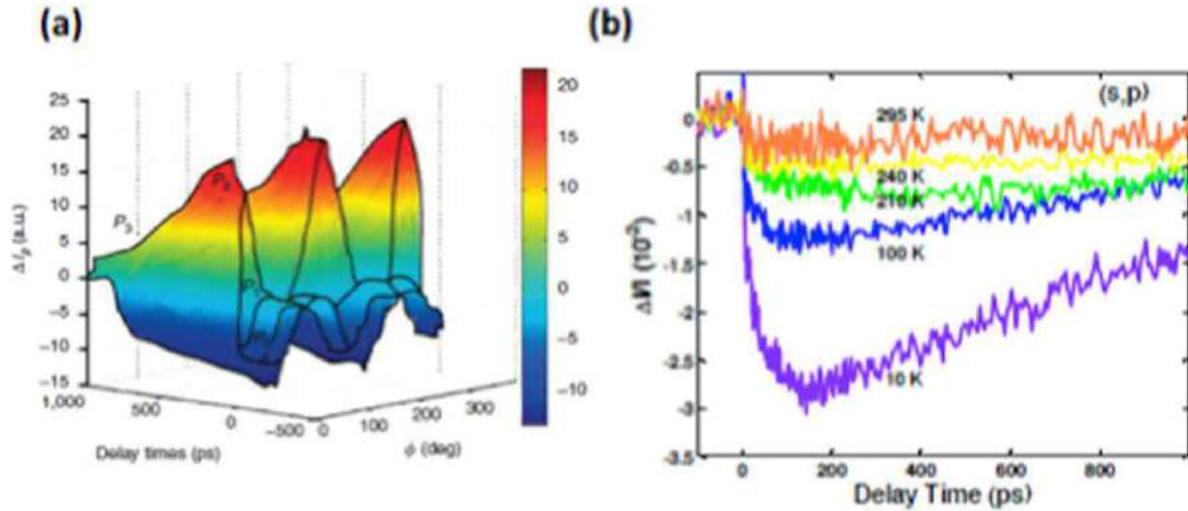


Fig. 20 trSHG signals from a BSTO/LCMO heterostructure as a function of (a) the incident light polarization at 10 K; and (b) temperature, demonstrating ultrafast interfacial ME coupling. Reproduced from Sheu, Y.M., Trugman, S.A., Yan, L., *et al.*, 2014. *Nature Communications* 5, 5832.

polarizations and propagation directions. In second harmonic generation (SHG), two light fields E_j and E_k of the same fundamental frequency ω are coupled through the nonlinear susceptibility $\chi_{ijk}^{(2)}$, producing a nonlinear polarization $P_i(2\omega) = \chi_{ijk}^{(2)} E_j(\omega) E_k(\omega)$ with the ensuing electric fields oscillating at 2ω . The tensor $\chi_{ijk}^{(2)}$ is extremely sensitive to the lattice, electronic and magnetic symmetries of a crystals and disappears when spatial inversion symmetry is present, making SHG a powerful probe of broken symmetry in a wide range of systems.

SHG spectroscopy has been widely used for the characterization of ferroelectric phase transitions, where inversion symmetry breaking by the FE polarization leads to an enhancement of the SHG signal (typically as the temperature is reduced). Moreover, SHG may also depend on magnetic symmetries through higher order nonlinearities or multipole contributions. Extension of SHG methods to the time domain is straightforward and can be done by photoexciting the sample before probing the SHG response, thus providing access to ultrafast dynamics of the electronic, vibrational, magnetic, and crystallographic structure of materials. Time-resolved SHG (trSHG) has been used to characterize non-equilibrium magnetic, orbital, and magnetoelectric responses of CEM. Arguably, the most promising applications of trSHG are as a simple and readily accessible alternative to X-ray scattering techniques for probing structural material dynamics during PIPT.

One example is shown in **Fig. 20**, which illustrates the ability of trSHG to track ferroelectric polarization dynamics caused by concurrent thermal expansion and magnetostrictive coupling across the interface between FM and FE layers in a multiferroic $\text{Ba}_{0.1}\text{Sr}_{0.9}\text{TiO}_3$ (BSTO)/LCMO heterostructure. In these experiments, femtosecond optical pulses were used to selectively excite the LCMO layer, which then underwent e-ph and spin-lattice relaxation within tens of ps (as in the OPTP results described in **Fig. 12** of Section “Ultrafast Infrared Spectroscopy: THz to Mid-IR”). trSHG from the top BSTO layer (**Fig. 20(a)**) exhibited an enhancement of the FE polarization on fast (~ 1 ps) and slow (~ 10 – 100 ps) timescales (**Fig. 20(b)**). Interestingly, the amplitude of the large slow component decreased with increasing spin disorder in LCMO as the temperature increased towards the FM T_c . These dynamics were explained as a strain-induced enhancement of the BSTO polarization mediated by (1) rapid thermal expansion of the photoexcited LCMO layer, caused by e-ph energy transfer and (2) slower expansion of the LCMO layer through changes in the magnetostriction that were induced by spin-lattice relaxation and the ensuing spin disorder. This demonstrates that ultrafast interfacial ME coupling in such heterostructures is dominated by elastic coupling across the interface, not the slower heat diffusion from the LCMO to the BSTO layer, and more generally, provides a nice example of the power of trSHG for shedding light on the electronic and structural dynamics of correlated electron materials.

Ultrafast Microscopy

Strong interactions between various degrees of freedom in CEM produce a delicate balance between competing states with very different natures, such as charge or orbital ordered, magnetic, ferroelectric and superconducting. This balance is easily perturbed by intrinsic fluctuations or extrinsic stimuli, leading to electronic, magnetic and structural inhomogeneities at the nano-to-micrometer scales. Often, such non-uniformity serves as a marker of competing interactions or phase coexistence, for example, near phase transitions, and indicates emergent functionality in materials systems. Over the years, many microscopic techniques have been developed to reveal the intimate connection between these nano-to-microscale spatial effects and the materials properties. However, in correlated materials spatial inhomogeneities also lead to nontrivial temporal dynamics (and vice versa), thus demanding tools that directly measure the properties at both relevant length and time scales to identify the role of intrinsic versus extrinsic effects and the extent to which such inhomogeneities determine functionality.

Here we give a brief, and by no means exhaustive, overview of the microscopic techniques that have already been demonstrated to combine ultrafast time-resolved imaging and spectroscopy. Achieving simultaneous sub-picosecond and sub-micron spatial resolution is a serious technical challenge: these 4D (space and time) experiments are usually performed by averaging the signals of a large number of pump-probe pulse pairs to improve the sensitivity, which requires high reproducibility of the photoinduced spatial phase separation from pulse to pulse. The same mechanisms that are responsible for the intrinsic inhomogeneity of correlated materials also make phase separation a very random process, and most of the previous effort focused on 4D methodology development using rather simple test systems, such as metal or semiconductor nanostructures. Nevertheless, initial encouraging applications to CEM (e.g., the canonical MIT in VO_2) have just begun to appear, and the field is bound for rapid growth as more 4D instruments become available to the research community.

The simplest spatially resolved optical technique of all is conventional optical microscopy, which can be extended to both the UV and IR spectral ranges with appropriate focusing optics. The spatial resolution of this technique is restricted by the diffraction limit, thus constraining its applicability to length scales exceeding the probe wavelength ($\sim 1 \mu\text{m}$ in near-IR region). Broadband reflection microscopy was used to investigate the peculiar striped dimerization patterns in a Mott-Hubbard insulator induced by a local current-driven MIT and metal-insulator phase separation in a correlated superconductor. Optical microscopy is also capable of other spectroscopic modalities besides absorption or reflection measurements. Collection of the SHG signal revealed the ferroelectric domain structure and its correlation with antiferromagnetic order in multiferroic samples. Integration with a DC magnetic field and polarization diagnostics provided a way to measure the effects of large-scale inhomogeneities in the crystalline structure on the ferromagnetic properties of manganites using magneto-optical Kerr effect microscopy (MOKE). Most of these modalities are compatible with time-resolved pump-probe geometries; for example, OPOP microscopy of VO_2 revealed that the timescales governing the photoinduced MIT decrease at higher excitation fluences, as measured from the peak of the focused beam profile to its tail. Ultrafast MOKE microscopy was also used to track the photoinduced magnetization changes in TbFeCo with better than 200 fs/200 nm combined resolution. In this work, it was shown that dipole-dipole interactions at the boundary of the illuminated region enhance the effective magnetic field and lead to local magnetization reversal on sub-nanosecond timescales.

Tuning the probe wavelength to the hard X-ray region brings the spatial resolution down to the sub-micron level when performing X-ray diffraction or spectroscopic measurements. In one approach to X-ray microscopy, the beam is focused to a 10–30 nm spot by various methods, for example, using a zone plate or Kirkpatrick-Baez mirror pair, and raster scanned across the area of interest while the diffraction patterns or scattering spectra are detected at each position. For example, scanning hard X-ray diffraction with 30 nm resolution was exploited to follow the monoclinic to rutile crystalline structure conversion during the MIT in VO_2 . Comparison between the temperature variation of the monoclinic phase fraction and insulating phase fraction measured by scanning near-field infrared microscopy (see below) showed that the electronic MIT proceeds monotonically, while the structure tends to oscillate around T_c . In the soft X-ray region, raster imaging was combined with magnetic spectroscopy to visualize switching of AFM order with electric fields in multiferroic BiFeO_3 . These results demonstrated that AFM order only exists in FE polarized regions of the sample, whereas unpoled areas remain magnetically disordered. Extending these X-ray measurements to the time domain might be challenging, due to photon flux constraints and shot-to-shot instability of the photoinduced crystallographic and spectroscopic inhomogeneities. However, with the development of brighter ultrafast XFEL sources and novel, resonant coherent diffractive (lensless) imaging approaches, it might be possible to perform wide field, rather than raster scanned, measurements of structural and magnetic dynamics in correlated materials, as was recently shown for a non-correlated ferromagnetic GdFeCo alloy. In that study, coherent X-ray spectromicroscopy of photoexcited GdFeCo revealed an initially inhomogeneous transient distribution of the angular spin momentum, after which the subsequent flow of angular momentum proceeded from the Fe-rich nanoregions to the Gd-rich ones on sub-ps timescales.

Another promising ultrafast wide field microscopy approach is photoemission electron microscopy (PEEM). PEEM records electrons emitted from a sample following the absorption of one X-ray or two visible photons. The electrons are accelerated towards the lens and projected to create a magnified image on the detector, similar to electron microscopes. A spatial resolution of $\sim 10 \text{ nm}$ can be routinely achieved, and spin sensitivity added by using linearly or circularly polarized photons. PEEM has been extensively exploited to study the structural, magnetic and electronic properties of CEM. For example, in multiferroic BiFeO_3 , the combination of PEEM and linear X-ray magnetic absorption dichroism (XMLD) spectroscopy was used to image the AFM domain distribution as a function of electric field applied to the sample. The variable electric field was found to modify the FE polarization and FE domain structure, which was measured with piezoelectric force microscopy (PFM). Comparison between results from XMLD-PEEM and PFM on the same sample area verified the strong coupling between FE and AFM orders in BFO and demonstrated AFM switching with an electric field. Temporal resolution can be added to PEEM by incorporating femtosecond pulses from either femtosecond XFEL or tabletop visible/near-IR laser systems. Ultrafast PEEM has been widely used to image the dynamics within plasmonic nanostructures with better than 10 fs temporal and 50 nm spatial resolution, but has never been applied to characterize CEM, to the best of our knowledge.

In the three decades following their invention, scanning probe microscopies have also provided much insight into the nanoscale properties of correlated materials. The highest spatial resolution can be achieved with scanning tunneling microscopy (STM), which is based on quantum tunneling of electrons through the energy barrier formed by the sharp metallic tip and a sample placed in the tip vicinity. The strong dependence of the current on the tip-sample spacing enables surface topography variations to be mapped with sub-angstrom resolution. The major virtue of STM lies in its ability to probe the local density of states of the sample in a narrow energy band around the Fermi level. For instance, the combination of high-resolution topographic and spectroscopic imaging revealed the non-uniform distribution of the superconducting gap size on the surface of a cuprate

HTSC. The gap inhomogeneity was attributed to oxygen impurity distributions, which suppressed the superconducting phase around them. STM has also demonstrated the percolative nature of the MIT in manganites, where a metallic phase gradually grows within the insulating matrix as the temperature is lowered across the MIT. To add time resolution, one can take advantage of the fact that the tunneling junction is a nonlinear region, where femtosecond electromagnetic pulses can be mixed to produce ultrafast transients in the tunneling current. The amplitude and dynamics of these transients depends strongly on the local material properties, making them a good probe of charge transfer and capture processes in metal and semiconductor nanostructures with sub-ps resolution. Modulation of the voltage bias with single-cycle THz pulses also provides sensitivity to the local electronic structure and molecular resonances while allowing tracking of molecular motion and charge dynamics in semiconductors on sub-nm/ps scales. However, ultrafast STM can only be used on conducting surfaces, due to its reliance on detection of the tunneling current.

Scanning near-field optical microscopies (SNOM) have recently risen to prominence for time-averaged and time-resolved exploration of nanoscale properties of CEM. Sub-wavelength imaging in SNOM is achieved either by using a small waveguide aperture for illumination/collection or by scattering the light from a sharp metallic tip. In the former, both the spatial resolution and spectral bandwidth are severely limited by the waveguide characteristics. On the other hand, the scattering-based technique (sSNOM) relies on propagation of the scattered light through free space, and its resolution is determined by the tip curvature, which can be below 5 nm. Furthermore, sSNOM can take full advantage of the recent breakthroughs in ultrafast broadband source development and perform nanoscale spectroscopy in a spectral range spanning visible to THz frequencies. As described earlier (see Section "Ultrafast Infrared Spectroscopy: THz to Mid-IR"), THz and mid-IR radiation encompass the energies of low-energy modes, energy gaps and transport phenomena in correlated materials and are thus of most interest for sSNOM applications. In VO_2 , collapse of the Mott gap during the MIT causes a significant spectral weight shift in the optical conductivity below 3000 cm^{-1} , due to the emergence of the Drude response in the metallic state. This shift was used as a contrast mechanism for sSNOM imaging of

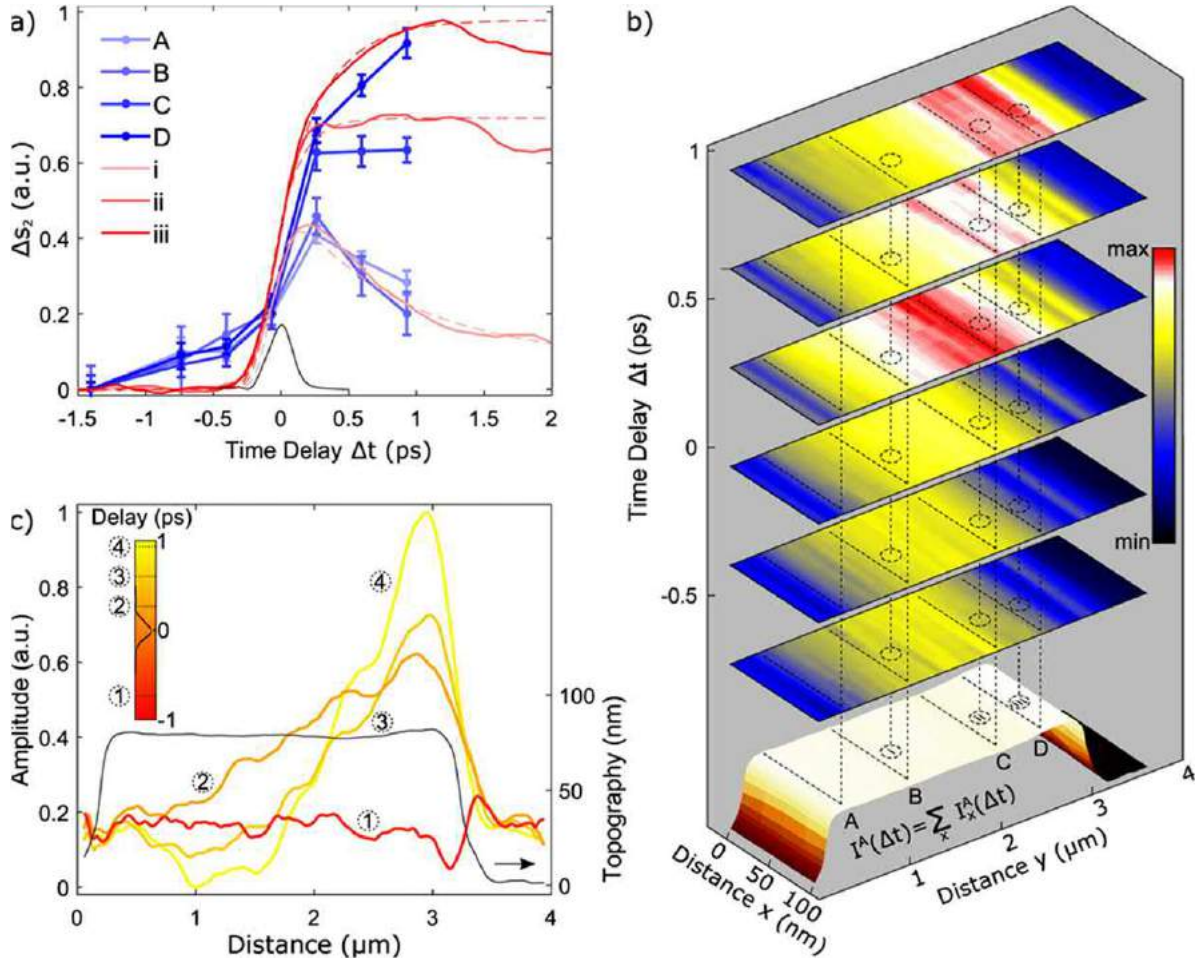


Fig. 21 (a) Time-resolved sSNOM of the photoinduced MIT in VO_2 , acquired with a sub-threshold fluence of 2.9 mJ/cm^2 and 1032 nm pump and $4.7 \mu\text{m}$ probe wavelengths. The red curves were taken at positions (i–iii) marked by circles in (b). The blue time points were extracted from the image series in (b) at the points (A–D). (b) Spatio-temporal images of the metallic phase (red) growth dynamics in a VO_2 microcrystal. (c) Cross-sections of the photoinduced scattering changes in (b) for four time delays across the microcrystal. Reproduced from Donges, S.A., Khatib, O., O’Callahan, B.T., et al., 2016. Nano Letters 16, 3029.

metal–insulator phase inhomogeneity during the MIT. Similar to STM on manganites, sSNOM measurements at 930 cm^{-1} revealed phase coexistence well below and above T_c . Moreover, sSNOM enabled local spectroscopy within an individual metallic island and demonstrated that the emerging metal phase close to T_c is different from the rutile one at higher temperatures, instead representing a strongly correlated metal with a pseudogap-like response in the IR conductivity.

sSNOM is an optical technique and is, therefore, very amenable to incorporating pump–probe measurements with ultrafast lasers. Indeed, the first time-resolved near-field optical microscopic measurements on CEM were recently demonstrated using sSNOM on nanostructured VO_2 . VO_2 was excited above its bandgap with 1032 nm pulses of 190 fs duration, and the amplitude of a $4.7\text{ }\mu\text{m}$, 200 fs probe pulse scattered from the tip was measured as a function of position and time (**Fig. 21(b)**). Although the excitation fluence of $2.9\text{ mJ}/\text{cm}^2$ was below the threshold for the photoinduced MIT, in some areas on the sample the amplitude of the scattering signal did not decay right away, as expected for sub-threshold carrier relaxation. On the contrary, the scattering signal continued to rise until saturating within a few ps, indicating the formation of metallic regions with enhanced scattering (curves (ii–iii) and red regions in **Fig. 21(a)** and **(b)**, respectively). Interestingly, this incredible variety of dynamics, ranging from sub-threshold carrier excitation to a saturated MIT phase transition, was observed just a few hundreds of nm apart. Previous OPTP results (see Section “Ultrafast Infrared Spectroscopy: THz to Mid-IR”) revealed softening of the photoinduced MIT, due to an increasing number of metallic nucleation sites as the temperature approached T_c . Similarly, the localized sub-threshold induction of MIT revealed by sSNOM could be attributed to the presence of a metal-rich phase in these regions, caused by inhomogeneous doping or strain in nominally homogeneous VO_2 crystals. These results demonstrate the feasibility of time-resolved microscopic studies of CEM, even in a multi-shot-averaging geometry, thus encouraging further advances in the field.

Further Reading

Time-integrated and time-resolved optical experiments on correlated electron materials:

- Averitt, R.D., Taylor, A.J., 2002. Ultrafast optical and far-infrared quasiparticle dynamics in correlated electron materials. *J. Phys. Condens. Matter* 14, R1357.
- Basov, D.N., Averitt, R.D., van der Marel, D., Dressel, M., Haule, K., 2011. Electrodynamics of correlated electron materials. *Rev. Mod. Phys.* 83, 471.
- Demsar, J., Sarrao, J.L., Taylor, A.J., 2006. Dynamics of photoexcited quasiparticles in heavy electron compounds. *J. Phys. Condens. Matter* 18, R281.
- Giannetti, C., Capone, M., Fausti, D., *et al.*, 2016. Ultrafast optical spectroscopy of strongly correlated materials and high-temperature superconductors: A non-equilibrium approach. *Adv. Phys.* 65 (2), 58.
- Hilton, D.J., Prasankumar, R.P., Trugman, S.A., Taylor, A.J., Averitt, R.D., 2006. On photo-induced phenomena in complex materials: Probing quasiparticle dynamics using infrared and far-infrared pulses. *J. Phys. Soc. Jpn.* 75, 011006.
- Mankowsky, R., Forst, M., Cavalleri, A., 2016. Non-equilibrium control of complex solids by nonlinear phononics. *Rep. Prog. Phys.* 79, 064503.
- Nasu, K., Ping, H., Mizouchi, H., 2001. Photoinduced structural phase transitions and their dynamics. *J. Phys. Condens. Matter* 13, R693.
- Orenstein, J., 2012. Ultrafast spectroscopy of quantum materials. *Physics Today* 65, 44.
- Prasankumar, R.P., Taylor, A.J. (Eds.), 2011. *Optical Techniques for Solid-State Materials Characterization*. San Francisco, CA: Taylor & Francis.
- Wegkamp, D., Stahler, J., 2015. Ultrafast dynamics during the photoinduced phase transition in VO_2 . *Prog. Surf. Sci.* 90, 464.
- Zhang, J., Averitt, R.D., 2014. Dynamics and control in complex transition metal oxides. *Annu. Rev. Mater. Res.* 44, 19.

General background on correlated electron materials:

- Chakhalian, J., Freeland, J.W., Millis, A.J., Panagopoulos, C., Rondinelli, J.M., 2014. Emergent properties in plane view: Strong correlations at oxide interfaces. *Rev. Mod. Phys.* 86, 1189.
- Dagotto, E., 2003. *Nanoscale Phase Separation and Colossal Magnetoresistance*. Berlin: Springer.
- Degiori, L., 1999. The electrodynamic response of heavy-electron compounds. *Rev. Mod. Phys.* 71, 687.
- Imada, M., Fujimori, A., Tokura, Y., 1998. Metal–insulator transitions. *Rev. Mod. Phys.* 70, 1039.
- Keimer, B., Kivelson, S.A., Norman, M.R., Uchida, S., Zaanen, J., 2015. From quantum matter to high-temperature superconductivity in copper oxides. *Nature* 518, 179.
- Khomskii, D., 2001. Electronic structure, exchange, and magnetism in oxides. In: Thornton, M.J., Ziese, M. (Eds.), *Lecture Notes in Physics*, vol. 569. Berlin: Springer-Verlag, p. 89.
- Scott, J.F., 2007. Applications of modern ferroelectrics. *Science* 315, 954.
- Shuvaev, A.M., Mukhin, A.A., Pimenov, A., 2011. Magnetic and magnetoelectric excitations in multiferroic manganites. *J. Phys. Condens. Matter* 23, 113201.
- Spaldin, N.A., Cheong, S.-W., Ramesh, R., 2010. Multiferroics: Past, present, and future. *Physics Today* 63, 38.
- Tokura, Y., 2003. Correlated-electron physics in transition-metal oxides. *Physics Today* 56, 50.
- Tokura, Y., 2000. *Colossal Magnetoresistive Oxides*. CRC Press.

Tutorial on Multidimensional Coherent Spectroscopy

Mark Siemens, University of Denver, Denver, CO, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction: The Power of 2DCS

Multidimensional spectroscopy measures the coherent response of a system and displays it as a function of two (or more) frequencies, corresponding to resonances such as electron absorption or vibrational excitation in the system. This multidimensional map of the frequency response is extremely helpful for disentangling congested spectra, revealing transport pathways, and measuring quantum coherences in heterogeneous systems. Even better, the fundamental understanding enabled by this multidimensional approach is not limited to a specific physical system, but can provide a comparison with, or in some cases a measurement of, the Hamiltonian for interaction dynamics for very different resonances. For example, in the last 25 years coherent multidimensional spectroscopy has spread across the electromagnetic spectrum:

- ~1970–present: At radio frequencies, multidimensional nuclear magnetic resonance (NMR) measures nuclear spin couplings in molecules to reveal the structure of proteins as they perform their physiological functions;
- 1993–present: In the infrared, two-dimensional infrared (2DIR) is sensitive to molecular vibrations, allowing direct measurement of hydrogen-bond dynamics in water (Fecko *et al.*, 2003) and vibrational couplings in DNA (Krummel *et al.*, 2003);
- 2003–present: Visible and near-infrared light enables multidimensional coherent spectroscopy of electronic resonances, revealing quantum transport pathways in photosynthetic complexes (Collini and Scholes, 2009; Brixner *et al.*, 2005) and many-body dynamics in semiconductor nanostructures (Li *et al.*, 2006; Singh *et al.*, 2016); and
- 2011–present: Terahertz two-dimensional spectroscopy directly measures solid-state lattice vibrations (phonons) and couplings to electrons.

While three-dimensional or even higher-order spectroscopies have been demonstrated, two-dimensional coherent spectroscopy (2DCS) provides enormous new physical insight compared to one-dimensional spectroscopies, and it is by far the most common implementation of coherent multidimensional spectroscopy. Higher-order spectroscopies have the same theoretical underpinnings and use similar, though more complicated, experimental setups, so this discussion of 2DCS is relevant to those as well.

Fig. 1 provides a schematic comparison between 2D and one-dimensional (1D) spectroscopies that highlights why 2DCS is such a powerful tool. A traditional 1D absorption spectrum measurement, as shown schematically in Fig. 1(a), can resolve the number and center frequencies of resonances, but cannot determine the nature of coupling, broadening, or higher-order states within the resonances. In contrast, the 2D spectra in Fig. 1(b)–(e) show significantly more information and provide quick identification of the physics and active transport pathways in a system.

A 2D spectrum provides a 2D response function, showing how the likelihood that excitation at a certain input/absorption frequency (labeled by the vertical axis) will lead to emission at a particular frequency. For isolated resonances, as shown in Fig. 1(b), all peaks lie along the diagonal line (highlighted in green) that specifies equal input and output frequency. If local environmental fluctuations affect the resonance frequencies, inhomogeneous broadening leads to a significant elongation of the signal along the diagonal of the 2D spectrum, as shown in Fig. 1(c). Off-diagonal peaks indicate coupling, as shown in Fig. 1(d), and careful analysis can determine the coherence and strength of the coupling. Fig. 1(e) shows that 2DCS is also well-suited to identifying and measuring multiply-excited state dynamics such as resonance anharmonicities, which show up as a negative peak (negative because of the different quantum pathways sampled) shifted to the left of the primary absorption resonance.

2D coherent spectra provide a direct identification and visualization of quantum pathways, providing quick identification of the physics in a system and even quantitative measurement of essential system parameters. A detailed guide to interpreting the features in 2D spectra is provided in Section Interpreting 2D Spectra.

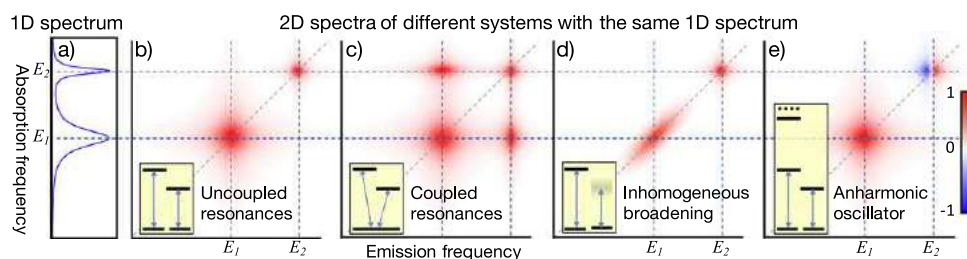


Fig. 1 Schematic comparison of one-dimensional (1D) and 2D spectroscopy. The 1D absorption spectrum shown in panel (a) could correspond to many different physical situations; 2D spectra on the right can resolve these. Yellow insets show the energy level diagrams corresponding to the unique 2D spectrum shown: (b) Uncoupled resonances, (c) inhomogeneous broadening, (d) coupled resonances, and (e) anharmonic oscillator. Interpretation of 2D spectra and implications for each of these physical situations will be discussed in detail in Section Interpreting 2D Spectra. Real part of 2D absorption spectra are shown, scaled by the colorbar on the right.

Theory of Coherent Multidimensional Spectroscopy

The development of 2D spectroscopy draws not only from nonlinear optics and coherent spectroscopy basics, but also from the historical context of the different physical systems studied. Chemists studying molecular vibrations and conformations, physicists studying electron dynamics in semiconductor nanostructures, and biologists studying exciton transport in photosynthetic complexes have differences in presentation and notation, and the relative importance of different features in a 2D spectrum can vary based on the nature of the resonance (electron, molecular vibration, spin, etc.) being studied. However, the underlying theoretical understanding of 2DCS experiments is based on the same physical principles.

Fourier Transform Spectroscopy

The most obvious way to collect a 2D coherent spectrum would probably be a “double-resonance experiment” performed by sweeping the frequency of two excitation beams and recording the signal amplitude as a function of the frequency of both beams: $S(\omega_1, \omega_2)$. However, the sensitivity of this approach is very low because of the small field amplitudes involved in a nonlinear optical process, so it has been superseded by time-domain methods using a series of intense laser pulses. If these pulses have controllably-stepped delays between them, then the time-domain signal can be Fourier-transformed to yield a spectrum. This “Fourier transform spectroscopy” approach for multidimensional spectroscopy was first proposed for NMR in 1971 by Jeener before being implemented by Ernst (1992), and is now widely used for 2DCS experiments throughout the electromagnetic spectrum.

The principle of Fourier transform spectroscopy is that a spectrum can be obtained by Fourier transforming a coherent signal measured in the time domain. The simplest example of an experiment that does this is a time-resolved double-pump-probe measurement, as shown in Fig. 2(a), in which three laser pulses are overlapped on a sample with a variable time delay τ between them. If the complex (both amplitude and phase) coherent signal is measured as a function of the delay time τ between pump pulses (Fig. 2(b)), then the discrete time-domain response can be Fourier transformed to yield the 1D spectrum (Fig. 2(c)) – which is identical to what a spectrometer would measure.

One-dimensional Fourier transform spectroscopy can be used to measure spectra with very high resolution, at wavelengths for which conventional spectrometers are unavailable, and with higher signal-to-noise through time-domain noise suppression techniques such as modulation and lock-in amplification. However, the exciting advance that Ernst pushed forward in the 1970s for NMR (Ernst *et al.*, 1988) is that the method is not limited to one dimension; extended pulse sequences can measure coherent signals from two or even more time dimensions that can be independently Fourier transformed to yield multidimensional spectral maps.

Double-Sided Feynman Diagrams

One of the strengths of 2DCS is that different pulse sequences can be used to study different quantum pathways in a system. In fact, the measured third-order signal from a particular pulse sequence usually results from multiple quantum pathways, and so accurate interpretation of 2D spectra relies on understanding exactly which quantum pathways are contributing. There are various bookkeeping tools available to write down and survey the possible quantum pathways (including ladder diagrams, which are often used in describing Raman spectra), but one that has gained wide usage across the range of 2DCS experiments is the double-sided Feynman diagram. Double-sided Feynman diagrams are particularly powerful because they allow the state coherence to be tracked over time, including changes resulting from photon absorption and emission (Boyd, 2003).

Here is a brief guide to interpreting double-sided Feynman diagrams (FDs). For the following discussion, we will assume a system described by a three-level energy ladder – perhaps corresponding to the first three states of an anharmonic oscillator – with ground-state $|0\rangle$ and excited states $|1\rangle$ and $|2\rangle$, as shown in the gray box in Fig. 3. The following discussion points match the panels in Fig. 3 that illustrate the basics of double-sided Feynman diagrams graphically.

1. The time evolution of the system is tracked by the parallel vertical lines, with the left side indicating the “ket” side of the density matrix ρ (Eq. 1) and the right side representing the “bra.” The initial state of the system is at the bottom, and time increases upward. As is standard in the “bra-ket” notation, a population is indicated when the “bra” and “ket” are the same state (e.g., $|a\rangle\langle a|$), otherwise a coherent superposition between states is specified (such as $|a\rangle\langle b|$). Both populations and coherences can evolve with time under the system Hamiltonian, until another optical pulse arrives and changes the state.

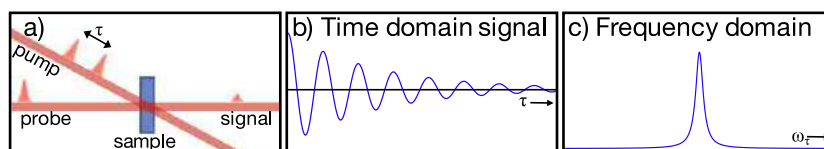


Fig. 2 (a) Schematic depiction of 1D pump-probe Fourier transform spectroscopy. (b) Real part of the coherent signal in the time domain. (c) Real part of the signal in the frequency domain, as measured by a spectrometer or a Fourier transform of the time-domain signal.

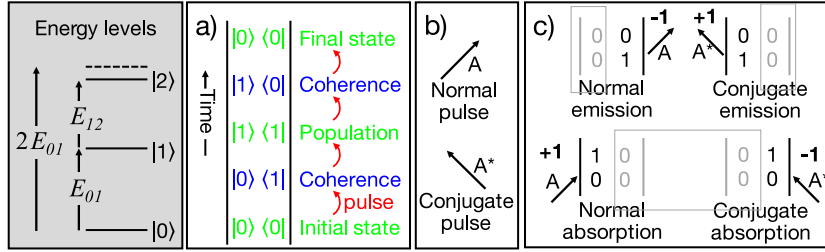


Fig. 3 An introduction to double-sided Feynman diagrams, see details in the text. Gray panel: Schematic 3-level energy ladder. Depending on the system, $E_{01} \neq E_{12}$. (a) Example double-sided Feynman diagram, showing that time increases upward. (b) The direction of normal and conjugate-acting pulsed electro-magnetic fields. (c) How to track absorption and emission of normal and conjugate pulses in a double-sided Feynman diagram. The bold $+1$ or -1 beside each figure tracks the sign of the interaction, which is determined by whether the field interacts with the “bra” or the “ket.” Reproduced from Boyd, R.W., 2003. Nonlinear Optics. Academic Press.

- Photon interactions are represented by diagonal arrows whose orientation indicates the action of the “jth” radiation field on the system: tipped-left indicates a “conjugate” field interaction contributing as $E_j^* = \epsilon_j^* \exp[i\omega_j t - i\mathbf{k}_j \cdot \mathbf{r}]$, while tipped-right means a “normal” field contribution of $E_j = \epsilon_j \exp[-i\omega_j t + i\mathbf{k}_j \cdot \mathbf{r}]$. This distinction is necessary because the light-matter interaction Hamiltonian is $\hat{H}(t) - \hat{\mu}E$, where $E = E_j + E_j^*$, and either the conjugate or non-conjugate portion of the field could act on the system. In the literature, these arrows are labeled in a variety of ways: by photon momentum, photon energy, frequency, or just pulse number.
- Since an arrow on the left (right) means that the field interacts with the ket (bra) of the density operator, an arrow's pointing towards or away from the FD shows how a photon is transferring energy to the system: the system absorbs a photon if the arrow points towards the diagram, and an arrow pointing away from the diagram describes photon emission. Photon absorption acts as a raising operator and photon emission acts as a lowering operator on the corresponding side of the density matrix, so the ket or bra state of the system changes accordingly in response to a photon interaction, and is recorded on the next line up in the diagram.

Careful use of Feynman diagrams allows for all possible quantum pathways to be accounted for and for relevant coherences to be tracked, enabling a direct way to determine expected peak locations in a 2D spectrum. This is essential for proper interpretation of the physical meaning of measured coherent spectra.

Theoretical Signal Origin

2DCS experiments measure the third-order response of a system of dipoles to applied electromagnetic fields, and an understanding of the theoretical origin of the signal is important for interpreting 2D spectra. Why is third-order coherence studied so heavily? The first-order (linear) response describes traditional (linear) spectroscopies like absorption and photoluminescence and the second-order response function is zero in an isotropic medium due to symmetry (Boyd, 2003), so third-order is the dominant term in a nonlinear power-series expansion of the polarization with the incident field. A third-order polarization will result from the system response to three excitation fields (Boyd, 2003), which could all be from the same pulse, but a three-pulse experiment will provide the most control over the state dynamics of a system.

While quantum mechanical descriptions of light interacting with a pure system wavefunction ψ can be sufficient to describe gas-phase coherence, condensed-phase systems involve statistical ensembles of molecules rather than pure states, and so the density matrix of an ensemble ρ is used instead of ψ :

$$\rho = \sum_s p_s |\psi_s\rangle\langle\psi_s| = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{pmatrix} \quad (1)$$

with p_s being the probability of a system being in state $|\psi_s\rangle$. The second half of Eq. (1) assumes a two-level system with states $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, in which case ρ_{00} is the average population in state $|0\rangle$, ρ_{11} is average the population in state $|1\rangle$, and $\rho_{10} = \rho_{01}^*$ is the coherence between states $|0\rangle$ and $|1\rangle$.

The evolution of a system $\rho(t)$ during and after optical excitation can be calculated with the standard optical Bloch equations (OBEs). At this point, different approaches are pursued depending on the system and type of excitation. For example, strong “ π -pulse” excitation in NMR experiments generates pure populations or coherences, while electronic and vibrational spectroscopies in the visible and infrared are safely in the “weak-excitation” limit and so perturbation theory should be applied (Boyd, 2003). Using perturbation theory and assuming the short-pulse (delta-function assumption) limit, the system evolution can be tracked over time as a time-ordered series of independent steps consisting of three different types:

- optically-induced transitions between population/coherence states in the system, which raise or lower the “bra” (“ket”) index in the density operator according to the Feynman diagram rules, and multiply the signal by $+(-)i\mu_{01}$;

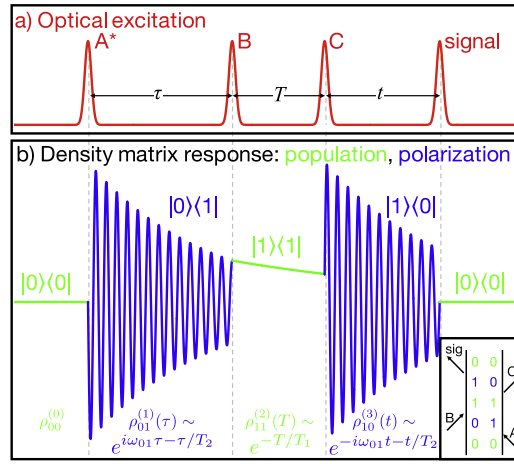


Fig. 4 Tracking population and coherence for a particular quantum pathway. (a) Optical excitation sequence. (b) Coherent system response, with each order of perturbation labeled and the resulting signals tracked. This quantum pathway corresponds to the double-sided Feynman diagram in the inset at lower-right.

- a state $|i\rangle\langle i|$ will evolve with time as an exponentially decaying population e^{-T/T_1} , where T_1 is the population decay time;
- the coherence $|i\rangle\langle j|$ leads to oscillation at a frequency of ω_{ij} and decay at the dephasing rate γ_{ij} , yielding a signal contribution of $e^{-i\omega_{ij}t} e^{-\gamma_{ij}t}$. The inverse of the dephasing rate is the characteristic homogeneous decay time $T_2 \equiv 1/\gamma$, which is frequently used in the literature.

For each “order” of perturbation theory, there will be an additional field interaction introduced in the process, and the constituent step responses can be multiplied to yield the total signal, as we will see shortly.

Double-sided Feynman diagrams can be incorporated with the results from perturbation theory, to calculate the expected signal for each quantum pathway. This is illustrated in **Fig. 4**, which illustrates the populations and polarizations involved in the quantum pathway shown in **Fig. 3(a)**. This quantum pathway leads to a third-order density matrix $\rho_{\text{signal}}^{(3)}(t)$:

$$\rho_{\text{signal}}^{(3)}(t) = \underbrace{(-i\mu_{01})_{A^*} \left(e^{-i\omega_{01}\tau} e^{-\gamma_{01}\tau} \right)}_{\rho_{01}^{(1)}(\tau)} \underbrace{(i\mu_{01})_B \left(e^{-T/T_1} \right)}_{\rho_{11}^{(2)}(\tau+T)} \underbrace{(-i\mu_{01})_C \left(e^{-i\omega_{01}t} e^{-\gamma_{01}t} \right)}_{\rho_{10}^{(3)}(\tau+T+t)} \quad (2)$$

In this equation and in **Fig. 4**, optical interactions from pulses A, B, and C are colored red, coherent oscillations and decays are blue, population dynamics are in green, and the first, second, and third-order perturbation theory steps building to this result are specified with superscript numbers in parenthesis.

The density matrix ρ is a powerful tool for tracking the coherence and population states in a system; and it can be connected to the macroscopic polarization of the system $P(t)$, which radiates to generate the measured signal. Specifically, the macroscopic polarization can be calculated by taking the expectation value of the density matrix multiplied by the dipole operator $\hat{\mu}$:

$$P(t) = \langle \rho_{10}(t) \rangle = \text{Tr}[\hat{\mu}\rho(t)] \xrightarrow{\text{3rd-order}} P^{(3)}(t) = \langle \rho_{10}^{(3)}(t) \rangle \quad (3)$$

where the right-hand side is specialized the third-order polarization resulting from perturbation theory.

While three-pulse experiments provide tremendous control over the quantum pathways probed by providing the three fields that will be applied, quantum pathways corresponding to a given pulse acting more than once are also possible. However, these alternative quantum pathways can be eliminated from the measured signal through quantum pathway selection.

Quantum Pathway Selection in 2D Coherent Spectra

There is inevitably a huge number of possible quantum pathways available to a multi-level system, and measuring all of them at once would only confuse and congest the measured spectrum. Polarization and pulse sequence manipulation provide enhanced control over which pathways are allowed, which is one of the strongest aspects of 2DCS. Selection is possible because each quantum pathway specifies a particular ordering of quantum states that is only possible with the appropriate fields. In addition to this quantum pathway control, experimental techniques such as phase-matching, frequency-tagging, and phase-cycling can

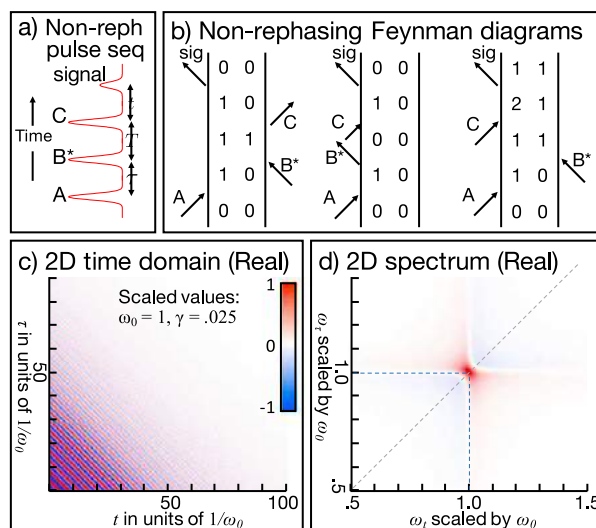


Fig. 5 Summary of non-rephasing 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum pathways, (c) real part of the 2D time domain signal, (d) 2D spectrum.

provide even more pathway discrimination by further narrowing which pathways are detected (even though all allowed pathways are active).

Polarization Control of Quantum Pathways in 2DCS Measurements

Control over the polarization of each of the incident pulses in a 2DCS experiment is often used to enhance or suppress certain quantum pathways. Polarization affects different kinds of resonant systems in different ways – for example, it couples with the relative angles of transition dipoles in molecules (Zanni *et al.*, 2001b), but it is related to optical selection rules in electron spectroscopy (Singh *et al.*, 2016). Controlling the polarization of excitation pulses can enhance or diminish the prevalence of each quantum pathway, and it can even be used to fully eliminate specific pathways; this technique is most often applied to eliminate the strong diagonal peaks from a 2D spectrum, which allows for weaker cross-peaks that are related to interesting transport and coherence dynamics to be more clearly resolved (Zanni *et al.*, 2001b).

Pulse Sequencing to Control Quantum Pathways in 2DCS Measurements

Different pulse sequences can also excite fundamentally-different quantum pathways in systems, and allow for studying different sorts of dynamics. Here we discuss the three most common types of 2D spectra in this article: Rephasing, Non-rephasing, and Double quantum.

Pulse sequence I: Rephasing 2D spectra

Rephasing 2D spectroscopy, also known as photon-echo spectroscopy, is the most common form of 2DCS. In this configuration, the pulses arrive at the sample with the conjugate pulse acting first: A^*-B-C . This is the well-known “photon echo” scheme from transient four-wave mixing, so-called because an inhomogeneously-broadened system will emit a signal at a time t equal to the A^*-B delay time τ , as will be discussed in more detail in Section Interpreting 2D Spectra of Inhomogeneously-Broadened Systems. We will see that rephasing spectroscopy is very powerful because the inhomogeneous and homogeneous broadening act along the $(t + \tau)$ and $(t - \tau)$ dimensions respectively, which are orthogonal directions in the 2D time domain. This orthogonality of the homogeneous and inhomogeneous broadening is also evident in the 2D spectra: the linewidth along the diagonal is dominated by inhomogeneous broadening, while the cross-diagonal linewidth indicates the homogeneous broadening.

Some authors plot rephasing spectra with the ω_t axis negative and pointing down (toward larger negative numbers). This convention is used to emphasize that the vertical spectral dimension corresponds to a conjugate pulse. In this case, the “diagonal” direction on the 2D spectrum is down and to the right.

Pulse sequence II: Nonrephasing 2D spectra

In a non-rephasing 2D spectroscopy experiment, the second pulse B is designated as the conjugate pulse, so the pulse sequence on the sample is $A-B^*-C$. As shown in Fig. 5(b), this excites fundamentally-different quantum pathways than the rephasing scheme. Scanning A before B^* is equivalent to a “ $-\tau$ ” scan, i.e., continuing a rephasing scan to “negative times” with the non-conjugate pulse now first. Homogeneous and inhomogeneous broadening act along the same $(t + \tau)$ direction in the non-rephasing

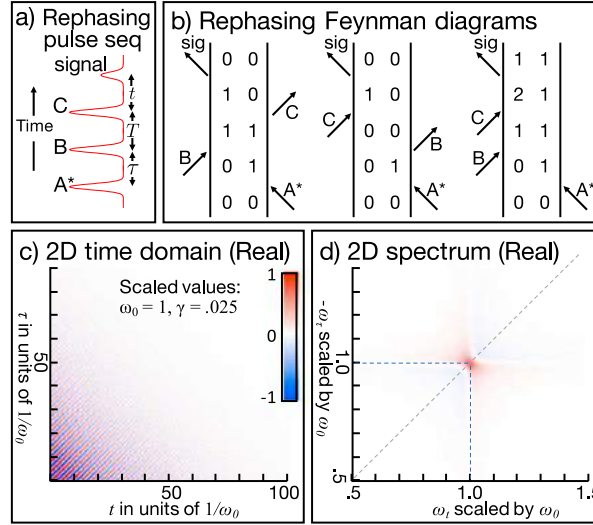


Fig. 6 Summary of rephasing 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum pathways, (c) real part of the 2D time domain signal, (d) 2D spectrum.

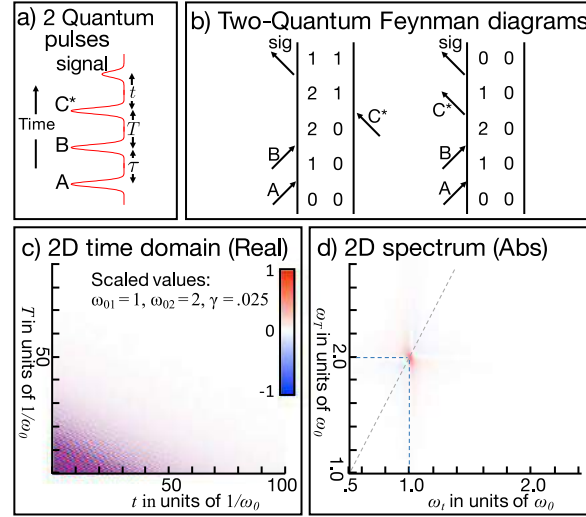


Fig. 7 Summary of double quantum 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum.

configuration, so the two causes of broadening are not separable in a 2D non-rephasing spectrum. For this reason, stand-alone non-rephasing 2D coherent spectroscopy is much less common than rephasing.

While non-rephasing spectra are not common on their own, they are very powerful when combined with rephasing spectra. Rephasing and non-rephasing results can be collected together by scanning τ through both positive and negative values, which results in a combined 2D time picture, which could be seen here by flipping the 2D time plot in Fig. 5(c) and placing it below (Fig. 6(c)). Fourier transforming the sum of rephasing and non-rephasing spectra leads to “absorptive lineshapes,” which are important for two reasons: first, the resonances are narrower and so are easier to resolve in congested spectra, and second, the real part is positive for a single resonance lineshape (Khalil *et al.*, 2003).

Pulse sequence III: Double quantum 2D spectra

Assuming that the system starts in the ground state, rephasing and non-rephasing spectra measure dephasing dynamics between the ground and first excited states, as well as the coherence between the first and second excited states. These spectroscopies are therefore measuring the dynamics of transitions excited by a single photon, i.e., “single-quantum.” Coherences between states that require multiple excitations are accessed and probed if the rephasing pulse is the last of the three excitation pulses to arrive, i.e., the pulse sequence is A–B–C*. In this case, shown in Fig. 7(a), if the time T between pulses B and C is scanned and Fourier transformed, the coherence studied is $|\langle c | \chi | a \rangle|$. This is the non-radiative coherence between the ground state and the doubly-excited state, and so these 2D spectra are often referred to as “double quantum.”

Quantum Pathways in 3DCS and Other Higher-Order Spectroscopies

While two-dimensional coherent spectroscopy provides significant advantages over one-dimensional spectra, recent work has shown that going to three-dimensional coherent spectroscopy (3DCS) or even higher dimensions provides even more separation of quantum pathways (Li *et al.*, 2013; Cundiff, 2014; Hamm, 2006). By analogy with 2DCS, which requires measurements of a third-order coherent signal over *two* time delays in which the system is in a coherent superposition state, 3DCS usually requires *three* time delays over coherence times (When there is a Raman coherence between states (Li *et al.*, 2013), the coherent oscillation between states during T between pulses A and B can be measured, scanned in time, and subsequently Fourier transformed. In this case, 3DCS can be performed with a third-order signal.). A glance at the Feynman diagrams shows that this cannot occur for three-pulse excitation, so a fifth-order process must be used. However, 2D experimental setups equipped with three delay-controlled pulses can still perform 3DCS measurements, as long as the signal is detected in a direction that accounts for fifth-order quantum pathways that are allowed when each pulse acts more than once.

Experimental Implementation: 3-Pulse Transient Four-Wave Mixing

While 2D spectra are reported in the frequency domain, they are usually recorded by excitation of a sample with a carefully-controlled sequence of pulses, measurement of a nonlinear coherent signal, and Fourier transformation of the signal with respect to pulse delay times to obtain the frequency response. Therefore, the usual techniques of nonlinear optics and coherent spectroscopy are very applicable.

2DCS is an extension of a time-resolved four-wave mixing (FWM) experiment, in which excitation wave pulses A , B , and C act on a sample to produce a signal wave pulse. The direction of the outgoing signal is set by momentum conservation and any of these pulses can also act as a conjugate pulse (indicated with a star $*$) on the opposite side of the density matrix, so that $k_{\text{signal}} = \pm k_A \pm k_B \pm k_C$. While quantum pathways corresponding to all possible combinations of normal and conjugate pulse sequences occur, having the pulses incident from different directions allows measurement selection of a particular set of quantum pathways by looking at the signal in a particular direction. A similar result can be obtained by frequency-tagging each excitation pulse and extracting a signal at the beat note corresponding to the desired quantum pathways (Tekavec *et al.*, 2007).

Usually, one pulse (say, pulse “ A ,” but it could be any pulse) is designated as the conjugate pulse, leading to a signal generated in the direction $k_{\text{signal}} = -k_A^* + k_B + k_C$ (There is nothing unique about the light in the conjugate-designated pulse, but by selecting only the signal in the momentum-matched direction for that pulse being conjugated allows selection of just the quantum pathways in which the designated pulse (and only that pulse) acts as a conjugate.). Incidence directions of pulses A , B , and C can be chosen such that the signal is generated in a direction that selects quantum pathways of interest and is easy to align to. One common scheme is to use a “box geometry” as shown in Fig. 8, in which A , B , C , and a reference beam are aligned to propagate parallel to each other at the corners of a box. If these beams are then focused onto the sample with a centered lens, the signal will be generated in the same direction as the “reference,” which provides a guide for alignment (Bristow *et al.*, 2009). Another common technique uses the “pump-probe” geometry, in which the first two pulses are incident from the same direction, i.e., $k_A = k_B$ (Shim and Zanni, 2009), so the signal is generated along the direction of the probe pulse C : $k_{\text{signal}} = -k_A^* + k_B + k_C = k_C$. While this “pump-probe” implementation of 2DCS is much simpler to set up (especially if adaptive optics are used to control the time delay τ between pulses A and B with intrinsic phase stability), this wavevector degeneracy for pulses A and B provides less selectivity over quantum pathways. On the other hand this non-selection can actually be an advantage: $k_A = k_B$ means that rephasing and non-rephasing spectra are both phase-matched along the k_C direction and so the coherent signal in this direction is the sum of rephasing and non-rephasing pathways; this means that absorption spectra are automatically collected (Shim and Zanni, 2009).

In a time-resolved four-wave mixing experiment, the signal is measured as a function of the delay between two of the pulses, usually the first two pulses A and B . One could imagine that time-resolving an additional time between pulses and then

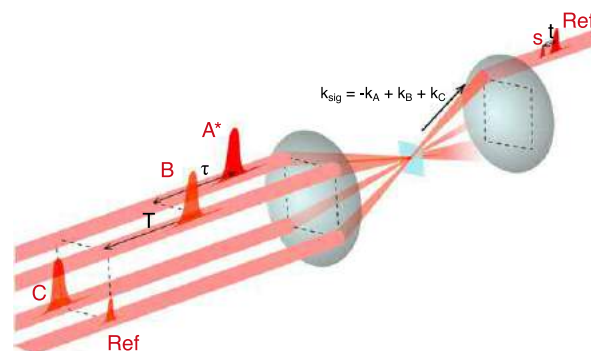


Fig. 8 Common setup for 2DCS that allows full control over all three time delays τ (between pulses A and B), T (between pulses B and C) and t (between pulse C and the output signal). The output can be sent to a spectrometer to directly resolve the output frequency ω_t while τ or T are scanned, or it can be measured with a photodiode while both t and τ are scanned.

performing a 2D Fourier transform would enable 2D coherent spectroscopy, but two additional enhancements are required to enable robust Fourier transformation of pulse delays:

1. Full phase information of the signal is obtained by enforcing interferometric stability between the pulses. Small thermal drifts or acoustic vibrations in optic mounts lead to changes in the signal phase during the course of a time-scanning measurement that cause a distortion of the measured 2D signal after Fourier transformation. A number of phase stabilization methods have been demonstrated, which are suitable for different cases based on various advantages and disadvantages. One of the first methods used common optics for each pulse, and delays were provided by changing the amount of glass in one beampath via translating glass prisms, but this technique was limited to pulse delays ≤ 5 picoseconds (Brixner *et al.*, 2005). Slightly longer delays are possible with intrinsically-phase stable pulsetrains generated by a computer-controlled phase ramp across the spectrum of an incident pulse, and propagated through common optics (Shim and Zanni, 2009). Active methods use mirrors that can be moved by piezoelectric stages to lock each pulse delay to the desired phase; these systems have the advantage of allowing very long delays (few nanoseconds), at the expense of a more complicated setup (Bristow *et al.*, 2009). In a recently-developed technique, the phase is not actively locked, but phase fluctuations are measured and removed from the signal by frequency-tagging each of the excitation pulses (Tekavec *et al.*, 2007).

2. The full complex signal must be recorded as a function of time delays between the excitation pulses. For a time-resolved output signal, this can be done with a lock-in amplifier that can resolve the real and imaginary parts of the signal. Alternatively, the complex signal can be spectrally-resolved in a spectrometer by interfering the signal with a known-phase reference beam with a technique called spectral interferometry (Lepetit *et al.*, 1995); the resulting interference fringes as a function of frequency allow the complex signal to be recovered. Either of these techniques provides the correct relative phase for the signal at all delay times and is sufficient for measuring amplitude 2D spectra, but obtaining the correct absolute phase to determine the real and imaginary parts of a 2D spectrum requires a separate measurement of the absolute phase.

Under these conditions, the phase evolution during different times can be correlated by taking a multidimensional Fourier transform.

Resolution in 2D Spectroscopy

As with any spectroscopic technique, resolution is an important experimental consideration in 2DCS. For a 2D coherent spectrum, different factors can determine the resolution along each spectral dimension. If a spectrometer is used to spectrally-resolve the emission energies (ω_i), then the resolution is determined by the spectrometers slit width, grating constant, and length as usual for spectrometer-resolved measurements. When frequency information is obtained by time-resolving and Fourier-transforming the signal, as is usual for obtaining the absorption spectrum axis (ω_τ) but can be used for resolving emission energies as well, the spectral resolution is determined by the range of the time delay $\tau_{\max} \cdot \Delta\omega_{\tau, \text{res}} = 1/\tau_{\max}$.

Running long scans is both technically-challenging and time-consuming, so the optimal length of the time delay is determined by the resolution required by the goals of the experiment. For example, for signals with a long-lived coherence such as a tightly-confined exciton in a quantum dot, a short and fast scan would suffice for resolving resonance energies or observing relaxation or quantum coherence between energetically-separated states. On the other hand, a scan over very long delays will be required to enable lineshape analysis and coherence linewidth measurements that are not resolution-limited. This, combined with the limited delay range of most 2DCS experimental techniques, limits the spectral resolution; for example, a 1 nanosecond delay yields a spectral resolution of 1 GHz and is already near the limit of what is possible with macroscopic delay-stage movement.

Interpreting 2D Spectra

Not only do 2DCS experiments provide a powerfully-controlled mapping and visualization of specific quantum pathways, they also provide measurements of fundamental system properties such as dephasing rates and inhomogeneity. Here we will provide a guide to understanding and interpreting 2D spectra. As discussed previously, the choice of pulse ordering determines the quantum pathways probed, and will result in different physical quantities being measured. Unless otherwise stated, we will consider the case of rephasing 2D spectra below.

Interpreting Homogeneous Lineshapes

Peaks located along the diagonal of a 2D spectrum are “degenerate,” that is, they indicate that the absorption and emission energies are the same. This indicates that the same coherence is probed during times τ and t , and thus the lineshape of a diagonal peak can be used to characterize the coherent dynamics of that state.

The homogeneous dephasing of a system is the natural linewidth of a pure and isolated resonance, determined by the dephasing dynamics in the system (e.g., for electronic excitations, radiative recombination and scattering with other carriers such as phonons can dephase the electron coherence). Nearly-pure homogeneous responses can be observed in warm, sparse atomic vapors because of the limited interactions between atoms (Li *et al.*, 2013). Polaritonic systems also exhibit homogeneous responses because of the strong coupling to a single photon mode in a cavity. In these cases, the lineshapes have a distinctive “star” or “cross” shape as expected by a direct Fourier transform of the homogeneous signal, and as seen in Fig. 5.

Interpreting 2D Spectra of Coupled Systems

Perhaps the most exciting application of 2DCS is the way in which couplings between resonance can be revealed, characterized, and quantified. This is enabled because a 2D spectrum is effectively a transfer matrix, directly showing the output resonances that can be excited for a given absorption frequency. This reveals powerful transient structural information, such as enabling direct observation of which electronic transfer pathways are active and dominant in the complex quantum networks involved in photosynthetic light-harvesting proteins (Brixner *et al.*, 2005) and confirming surprising coherent transfer in synthetic polymers (Collini and Scholes, 2009).

More than revealing that coupling exists, a dynamic series of 2D coherent spectra can also clearly define the incoherent vs coherent nature of the coupling and can even measure important coupling parameters. This is illustrated in Figs. 9 and 10, and discussed below.

Fig. 9 shows the case of coherent coupling between states. The energy diagram for this case is an upside-down “A” system, in which two excited states $|1\rangle$ and $|1'\rangle$ have a common ground state $|0\rangle$. The possible quantum pathways for a rephasing pulse sequence are shown with double-sided Feynman diagrams in Fig. 9(b). These pathways can be grouped by their input and output resonance energy: diagonal peaks in the 2D spectrum are formed by diagrams R_1 and R_2 at $(\omega_{01}, \omega_{01})$ and R_5 and R_6 at $(\omega_{01'}, \omega_{01'})$, but the other diagrams lead to cross-peaks as shown in Fig. 9(c). Taking the upper-left cross-peak (light blue) as an example, both R_7 and R_8 contribute on the 2D spectrum at $(\omega_\tau = \omega_{01'}, \omega_t = \omega_{01})$, but the dynamics during the time T between pulses B and C are very different for these two pathways: in R_7 , there is a coherence between states $|1\rangle$ and $|1'\rangle$, but in R_8 there is only a population in the ground state. This coherence between excited states is non-radiative, and can be directly measured by measuring the cross-peak amplitude in a series of (ω_τ, ω_t) spectra for different values of T . This is illustrated in Fig. 9(d). The contributions of each pathway are illustrated schematically in this figure, but a real 2DCS experiment could not resolve them. However, the difference in oscillation frequency with T means that a 3DCS experiment could completely separate the individual quantum pathways that make up the cross-peaks in a coherent coupling case.

If there is no coherence between states, there could still be population relaxation between states, as shown in Fig. 10. In this case, the possible quantum pathways for a rephasing pulse sequence are simpler: diagonal peaks in the 2D spectrum are formed by the same diagrams as in the coherent coupling case (because diagonal resonances in a 2D spectrum indicate no coupling), but there is only one coupling diagram that describes the relaxation. This diagram, labeled R_9 , is different from the double-sided Feynman diagrams we have drawn so far because it has an extra line during time T , between pulses B and C. When pulse B generates a population in the $|1'\rangle$ state, diagram R_9 shows that it is possible for that population to relax to the $|1\rangle$ state without optical excitation. This relaxation process leads to an off-diagonal cross-peak just as in the coherent case, but this peak only appears above the diagonal and it has very different time dependence. Fig. 10(d) shows that as a function of T , the relaxation cross-peak grows for time T_{relax} before decaying with the rest of the system's population with the population relaxation time T_2 .

If a system of uncoupled resonances is at a high enough temperature such that the thermal energy is on the order of or greater than the difference between the excited state energies, then activation from low-to-high energy excited states can also be achieved.

Spectra are shown here for pure homogeneous resonances, but this analysis works equally-well for coupled resonances with inhomogeneous broadening and/or anharmonicity.

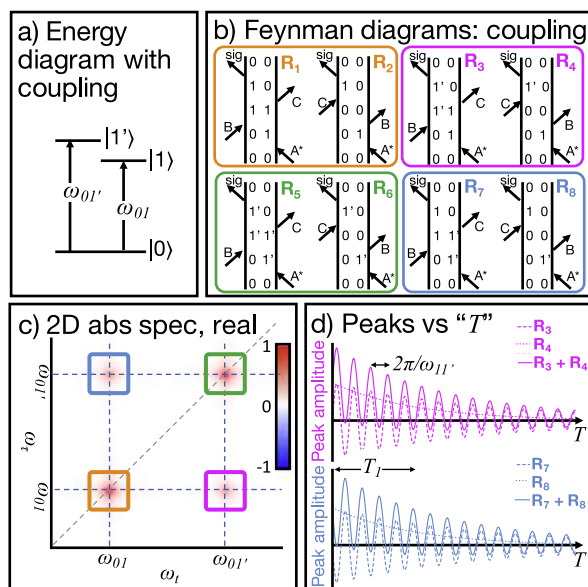


Fig. 9 2D spectra with cross-peaks due to coherent coupling between resonances. (a) energy diagram with coupling, (b) Feynman diagrams: coupling, (c) 2D abs spec, real, (d) Peaks vs. “ T ”

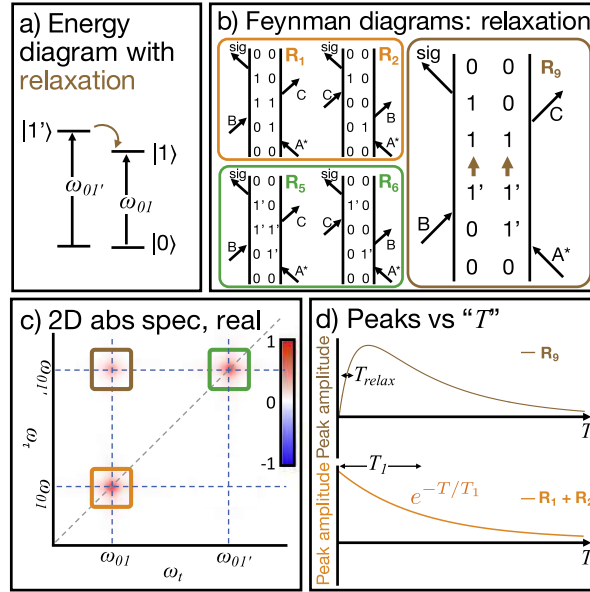


Fig. 10 2D spectra with cross-peaks due to incoherent population relaxation between states. (a) energy diagram with relaxation, (b) Feynman diagrams: relaxation, (c) 2D abs spec, real, (d) Peaks vs. “T”

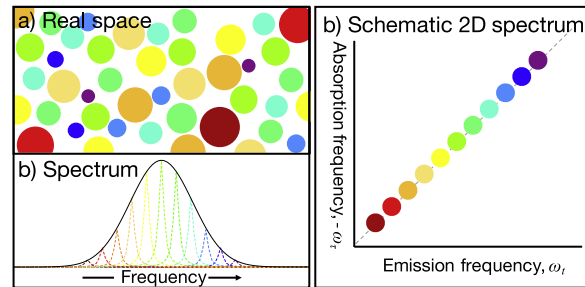


Fig. 11 Schematic depiction of how heterogeneity in a system affects 1D and 2D spectra. (a) Real-space representation of a heterogeneous mixture of nanostructures whose resonance energy depends strongly on the nanostructure size. Color indicates resonance energy. (b) Linear 1D spectrum (absorption or photoluminescence) for the system shown in (a). The fundamental resonance linewidth is obscured by inhomogeneous broadening. (c) Schematic 2D spectrum, illustrating the effect of inhomogeneity to stretch the signal along the diagonal. Color in this schematic 2D spectrum is labeling the shifted resonance frequency and is not related to the amplitude of the response.

Interpreting 2D Spectra of Inhomogeneously-Broadened Systems

Inhomogeneity is unavoidable in condensed matter and molecular systems because of local non-uniformities in the environment, such as fluctuations in the electronic confinement potential that are intrinsic to heterogeneous solid-state systems and conformational variations in molecular systems. In 1D spectra, this leads to a broadening of the fundamental resonance lineshape, as illustrated in Fig. 11(a) and (b) for a cartoon quantum-confined system such as quantum dots where the resonance energy scales inversely with the size, which varies. This inhomogeneous broadening often dominates 1D absorption or photoluminescence spectra, and there is no way to recover the homogeneous linewidth with linear 1D spectroscopies.

In 2D spectra, inhomogeneous broadening is manifest as an elongation of the signal along the diagonal. Inhomogeneous broadening causes this along-the-diagonal broadening because individual resonators experience different shifts in resonance energy $\omega_{0,i} = \omega_0 + \Delta\omega_i$, the effect of which can be seen from the Fourier shift theorem (Ernst *et al.*, 1988):

$$s(\omega_0\tau')\exp[i\Delta\omega\tau'] = S((\omega_0 + \Delta\omega)\tau') \quad (4)$$

where $s(\tau')$ and $S(\omega)$ are the Fourier-pair signals in the time and frequency domains, respectively. In other words, a change of the resonance frequency in the time domain (left side of Eq. (4)), is equivalent to a shift of the signal in the Fourier (spectral) domain. In the 2D time domain shown in Fig. 11(c), the signal is resonant along the diagonal ($\tau = t - \tau$) direction; application of the Fourier shift theorem implies a shift of $\Delta\omega$ along the diagonal ($\omega_\tau = \omega_t - \omega$ direction) of a 2D spectrum. Inhomogeneous broadening can

be thought of as simply a distribution of homogeneous systems with slightly shifted resonance energies, leading to a signal spread along the diagonal in a 2D spectrum, as shown in Fig. 11(c).

Measuring homogeneous and inhomogeneous linewidths

In the midst of inhomogeneous broadening, 2DCS allows for clear separation and quantitative measurement of homogeneous and inhomogeneous linewidths. One of the great strengths of rephasing 2DCS is shown in Fig. 12(b), which clearly shows that the inhomogeneous response is isolated along the diagonal ($\omega_t = \omega_i + \omega_r$ direction) of a 2D spectrum, and the homogeneous response along the cross-diagonal ($\omega_t = \omega_i - \omega_r$). Quantitative measurement of both homogeneous and inhomogeneous line-shapes can be performed by fitting cross-diagonal and diagonal slices from a 2D spectrum (Siemens *et al.*, 2010), but accurate measurements require careful consideration of the quantum pathways involved and the effects of homogeneous and inhomogeneous broadening on the lineshapes.

Spectral diffusion

Because 2D spectroscopy can resolve resonance lineshapes along two orthogonal axes, it is an excellent tool for observing and measuring lineshape dynamics. One example in which lineshape dynamics are extremely important is spectral diffusion. Spectral diffusion describes the dynamic and random redistribution of populated states, such that, given sufficient time, the observed resonance energy is equally-likely to be in any possible state within an inhomogeneous distribution, regardless of absorption frequency. For the schematic system shown in Fig. 11(a), spectral diffusion corresponds to site-to-site hopping such as an excited electron hopping between nanostructures or polymers. The physics governing this hopping depends on the system being studied, and possibilities include including quantum tunneling and phonon scattering. In contrast to this interpretation for spectral diffusion of electrons, spectral diffusion of molecular vibrational resonances in a fluid is usually understood directly in the spectral domain as spectral changes of individual resonators. This is because molecules in a fluid are free to move, so vibrational excitations do not “hop” between molecules; rather, the resonance energy of a specific molecular vibration evolves over time as it moves and so experiences different environments and conformations.

In 2DCS measurements of spectral diffusion, a series of 2D scans is performed for steps of T , and the lineshapes are compared. This is illustrated schematically in Fig. 13, where the population diffusion with increasing T is clearly observable as a spreading of the originally-tight lineshape along the cross-diagonal direction. This long- T “spectrally-diffuse” limit in 2D spectra arises because for any absorbed excitation ω_{τ_0} , there is sufficient time for population mixing between local states that the signal could be emitted at any possible ω_t . The emission frequency of the signal will therefore be equally likely to be emitted across the distribution of inhomogeneously-broadened states, leading to no correlation between absorption and emission frequencies. 2DCS is the ideal tool for measuring the timescale of this process, which is an important parameter for understanding the physics governing inter-site population transfer in various systems.

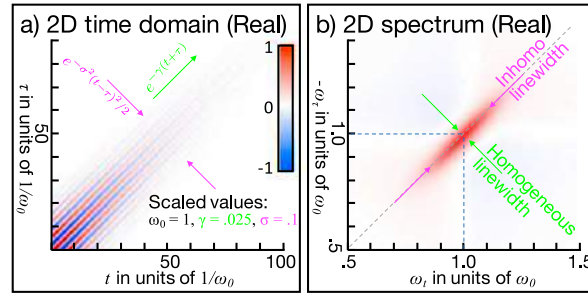


Fig. 12 2D coherent spectroscopy of strong inhomogeneous broadening. (a) Coherent signal in the 2D time domain. (b) 2D spectrum, showing separation of the inhomogeneous and homogeneous linewidths along the diagonal and cross-diagonal directions.

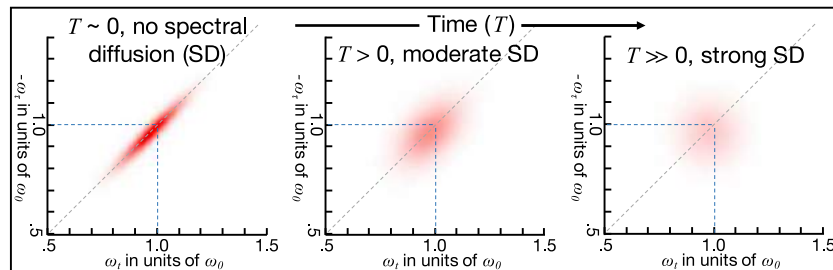


Fig. 13 Spectral diffusion (SD) in 2D spectra. Schematic 2D spectra for increasing waiting time (T , fixed delay between pulses B and C). The cross-diagonal linewidth is only equal to the homogeneous (T_2) linewidth for short waiting time; the cross-diagonal broadening at longer waiting time is due to spectral diffusion, and the lineshape approaches a 2D Gaussian.

Interpreting Anharmonic Oscillator Spectra

The harmonic oscillator is a common theoretical starting point in describing resonant systems near equilibrium, such as vibrational modes in a molecule or solid. However, the perfectly-parabolic potential well required for a pure harmonic response is rare, and so higher-order corrections are usually required. This anharmonicity of resonances can be difficult to identify, much less measure, in congested or inhomogeneous 1D spectra, but 2DCS provides a clear visualization and means of direct measurement of the important anharmonic oscillator parameters. These measurements can be combined with theoretical models to provide information about bond strength, length, and orientation in molecular systems, and many-body interactions in electronic systems.

Consider the anharmonic potential well shown in Fig. 14(a). Assuming that the system starts in the ground state, the double-sided Feynman diagrams for rephasing spectra are shown in Fig. 14(b). Diagrams R_1 and R_2 both oscillate in both τ and t at a frequency of ω_{01} . However, pathway R_3 oscillates in t at frequency ω_{12} , and for an anharmonic oscillator $\omega_{12} \neq \omega_{01}$, so this quantum pathway will lead to quantum beating in the 2D time domain and it will be shifted in the 2D spectral domain, as shown in Fig. 14(c) and (d).

Absorption 2D spectroscopy of an anharmonic oscillator provides two important advantages over rephasing 2D spectra: enhanced resolution because of tighter peak shapes and a clear presentation since the real part of the spectrum is absorptive. From this we can clearly see that pathways R_1 and R_2 have a positive contribution to the absorption spectrum, while the R_3 pathway contributes negatively. This sign flip in R_3 arises from having an odd number of pulse interactions on the right-hand side of the Feynman diagram (Boyd, 2003).

The anharmonicity $\Delta = \omega_{01} - \omega_{12}$ can be approximated directly from a 2D spectrum of an anharmonic oscillator by measuring the frequency separating the positive and negative peaks. When the anharmonicity is larger than the linewidths, these peaks are well-resolved and the peak separation is a good measurement of anharmonicity (Zanni *et al.*, 2001a). But more often, the linewidths are larger than the anharmonicity and so the peaks interfere and partially cancel. In this case, careful fitting and lineshape recovery is needed in order to accurately measure the anharmonicity (Zanni *et al.*, 2001a).

The example shown here was for a single, homogeneous resonance, but the same signals interpretation is valid for ensembles with inhomogeneity, making 2DCS a powerful tool for fully characterizing anharmonicity of vibrations in molecular systems (Zanni *et al.*, 2001a). Additionally, this anharmonic approach has recently been applied to reveal the bosonic nature of excitons and the importance of many-body interactions in a semiconductor quantum well (Singh *et al.*, 2016).

Connecting 2D and 1D Spectra

Along with the interpretation of various 2D spectra, it is important to recognize the connection between 2D and 1D spectra. This important connection can be seen with the “projection-slice theorem” of Fourier Transforms (Ernst *et al.*, 1988):

$$\mathcal{F}\left[\int s(x, y)dx\right] = S(\omega_y) \quad (5)$$

which says that Fourier transformation of a projection (integration of all data perpendicular to an axis) onto an axis is equivalent to a slice in the Fourier-pair domain. This is illustrated by the insets of Fig. 15, each row shows the projection-slice theorem implemented for the time domain and the Fourier-pair frequency domain.

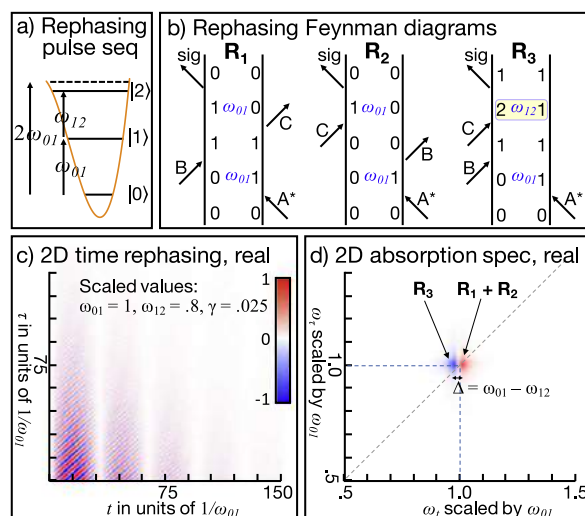


Fig. 14 2DCS of an anharmonic oscillator. (a) Energy level structure. (b) Rephasing Feynman diagrams, with coherence frequencies marked in blue; note the third diagram which has a different oscillation frequency during time t between pulses C and sig . (c) 2D time domain signal (real part of rephasing signal). (d) 2D coherent spectrum of an anharmonic oscillator (real part of absorption spectrum, i.e., rephasing + nonrephasing).

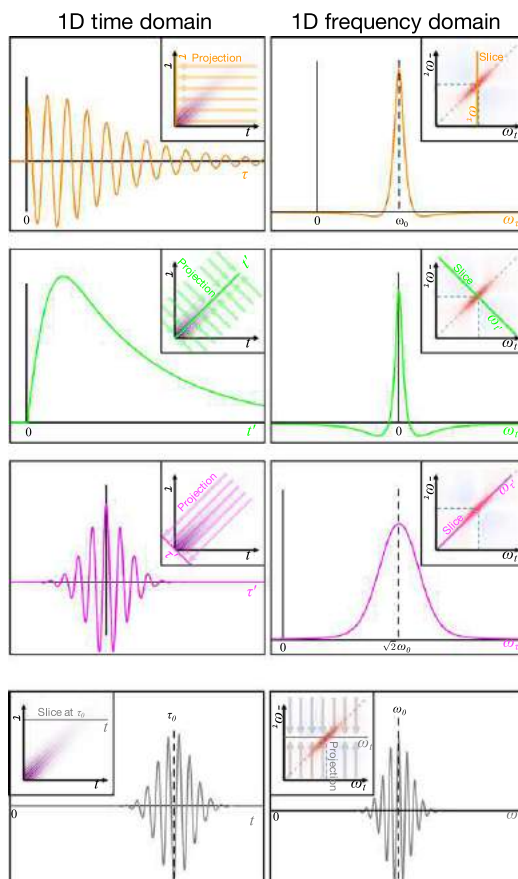


Fig. 15 How to convert from 2D to 1D spectroscopy with the projection-slice theorem. The first three rows show 1D time domain (left column) and frequency domain (right column) signals corresponding to slices along different directions of a 2D spectrum. Insets show the corresponding projection (slice) in the time (frequency) domain. The first row is equivalent to a time-integrated four-wave mixing experiment, but second and third rows do not have simple 1D analogs. In the last row, a slice (projection) in the time (frequency) domain is shown, clearly showing the “photon echo” in which the time-domain signal occurs when the time delays t and τ are equal.

In the case of 2D spectra, projecting onto a given axis in the 2D time domain and then Fourier transforming yields the equivalent 1D spectrum slice along the same direction, in accordance with the projection-slice theorem. Conversely, projection onto a particular axis of the 2D spectrum is equivalent to a slice along the same axis direction in the 2D time domain, as shown for a “photon echo” experiment in the last row of Fig. 15. In this way, 2D scans can be easily compared to their 1D analogs, even along directions different from the acquisition such as the diagonal and cross-diagonal. For example, lineshapes along diagonal and cross-diagonal directions of can be immediately recovered from a 2D spectrum, allowing for new insight into homogeneous and inhomogeneous linewidths (Siemens *et al.*, 2010).

One important caveat should be noted: because 2DCS measures the third-order response of a system, a corollary 1D spectrum obtained with the projection-slice theorem will be a 1D coherent spectrum of the third-order response. This 1D spectrum will be equal to that obtained from a time-resolved pump-probe experiment (which measures the third-order response), but will *not* be the same as a linear absorption or photoluminescence spectrum measuring the linear response of the system.

Conclusion and Outlook

2DCS is an active field of research, and the future is very bright. Higher-order spectroscopies, such as 3D, are one area of active growth in multidimensional coherent spectroscopy research, and show promise for further-enhanced resolution of quantum transport pathways, including the possibility of fully-characterizing the Hamiltonian of a system. The recent development of widely-tunable pulsed laser sources and the expansion of 2DCS from the infrared into both the visible and the terahertz spectral regions is a powerful combination: as new materials or molecules are developed for various applications, 2DCS can be quickly deployed to provide deep insight into fundamental coherence dynamics and quantum transport pathways available to the electronic, vibrational, and phononic energy carriers in the system. In the not-so-distant future, it is possible that 2DCS will be a standard analysis tool that can be applied not just for end-of-cycle characterization, but as an essential part of the development cycle of new materials.

While the future is bright, the present is also an exciting time for 2DCS research. The ability to directly supersede one-dimensional linear and coherent spectroscopies with a multidimensional frequency response map is allowing direct measurements of physical properties and tracking of dynamics with a resolution and precision that would have been unimaginable only 25 years ago. In that time, 2DCS has moved from a niche test-bed experiment to become an essential tool in spectroscopy, which promises further excitement for the next quarter-century and beyond.

References

- Boyd, R.W., 2003. *Nonlinear Optics*. Academic Press.
- Bristow, A.D., Karaiskaj, D., Dai, X., *et al.*, 2009. A versatile ultrastable platform for optical multidimensional fourier-transform spectroscopy. *Review of Scientific Instruments* 80, 073108.
- Brixner, T., Stenger, J., Vaswani, H., *et al.*, 2005. Two-dimensional spectroscopy of electronic couplings in photosynthesis. *Nature* 434, 625–628.
- Collini, E., Scholes, G.D., 2009. Coherent intrachain energy migration in a conjugated polymer at room temperature. *Science* 323, 369–373.
- Cundiff, S.T., 2014. Optical three dimensional coherent spectroscopy. *Physical Chemistry Chemical Physics* 16, 8193–8200.
- Ernst, R.R., 1992. Nuclear magnetic resonance Fourier transform spectroscopy. In: Malmström, B.G. (Ed.), *Nobel Lectures*. Singapore: World Scientific Publishing Co.
- Ernst, R.R., Bodenhausen, G., Wokaun, A., 1988. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford: Clarendon Press.
- Fecko, C., Eaves, J., Loparo, J., Tokmakoff, A., Geissler, P., 2003. Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water. *Science* 301, 1698–1702.
- Hamm, P., 2006. Three-dimensional-ir spectroscopy: Beyond the two-point frequency fluctuation correlation function. *The Journal of chemical physics* 124, 124506.
- Khalil, M., Demirdöven, N., Tokmakoff, A., 2003. Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra. *Physical Review Letters* 90, 047401.
- Krummel, A.T., Mukherjee, P., Zanni, M.T., 2003. Inter and intrastrand vibrational coupling in dna studied with heterodyned 2D-IR spectroscopy. *The Journal of Physical Chemistry B* 107, 9165–9169.
- Lepetit, L., Cheriaux, G., Joffre, M., 1995. Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *Journal of the Optical Society of America B: Optical Physics* 12, 2467–2474.
- Li, H., Bristow, A.D., Siemens, M.E., Moody, G., Cundiff, S.T., 2013. Unraveling quantum pathways using optical 3d fourier-transform spectroscopy. *Nature Communications* 4, 1390.
- Li, X., Zhang, T., Borca, C.N., Cundiff, S.T., 2006. Many-body interactions in semiconductors probed by optical two-dimensional fourier transform spectroscopy. *Physical Review Letters* 96, 1–4.
- Shim, S.-H., Zanni, M.T., 2009. How to turn your pump–probe instrument into a multidimensional spectrometer: 2D IR and vis spectroscopies via pulse shaping. *Physical Chemistry Chemical Physics* 11, 748–761.
- Siemens, M.E., Moody, G., Li, H., Bristow, A.D., Cundiff, S.T., 2010. Resonance lineshapes in two-dimensional fourier transform spectroscopy. *Optics Express* 18, 17699–17708.
- Singh, R., Suzuki, T., Autry, T.M., *et al.*, 2016. Polarization-dependent exciton linewidth in semiconductor quantum wells: a consequence of bosonic nature of excitons. *Physical Review B* 94, 081304.
- Tekavec, P., Lott, G., Marcus, A., 2007. Fluorescence-detected two-dimensional electronic coherence spectroscopy by acousto-optic phase modulation. *The Journal of Chemical Physics* 127, 214307.
- Zanni, M.T., Asplund, M.C., Hochstrasser, R.M., 2001a. Two-dimensional heterodyned and stimulated infrared photon echoes of N-methylacetamide-D. *The Journal of Chemical Physics* 114, 4579–4590.
- Zanni, M.T., Ge, N.-H., Kim, Y.S., Hochstrasser, R.M., 2001b. Two-dimensional ir spectroscopy can be designed to eliminate the diagonal peaks and expose only the crosspeaks needed for structure determination. *Proceedings of the National Academy of Sciences* 98, 11265–11270.

Two-Dimensional Infrared (2D IR) Spectroscopy

Lauren E Buchanan, Vanderbilt University, Nashville, TN, United States

Wei Xiong, University of California, San Diego, CA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Coherent two-dimensional infrared (2D IR) spectroscopy has flourished in its application to problems in chemistry, biophysics and materials science since its first demonstration in 1998 (Hamm *et al.*, 1998). As an IR analogue of 2D nuclear magnetic resonance (NMR) spectroscopy, 2D IR spectroscopy (Hamm and Zanni, 2011; Cho, 2009; Fayer, 2013, 2009; Zheng *et al.*, 2007; Kurochkin *et al.*, 2007; Wright, 2011; Hunt, 2009; Buchanan *et al.*, 2012; Sueur *et al.*, 2015; Mukamel, 2000; Bakulin *et al.*, 2009) probes the time-dependent third-order nonlinear response of molecular vibrational fingerprints, which gives it unique advantages over traditional vibrational spectroscopies, such as linear IR and Raman spectroscopy. These advantages include its ability to: better resolve congested spectral peaks, which enables the determination of molecular structures that are otherwise difficult to obtain; distinguish homogenous and inhomogeneous spectral lineshapes, which allows measurements of local environments; measure spectral diffusions that are related to picosecond molecular dynamics; and observe cross peaks, which reveals structural rearrangements, couplings between vibrational groups, and energy transfer from one site to another. In this article, we will briefly introduce the basic theory of 2D IR spectroscopy and discuss advances in experimental design. Then, we will highlight a few scientific fields to which 2D IR spectroscopy has made seminal contributions. Last, we will present some new developments that are on the horizon which could further bring in transformative capabilities in the fields of non-linear laser spectroscopy.

Brief Theory

2D IR spectroscopy is a coherent non-linear optical process which measures time-dependent vibrational coherences. All 2D IR spectrometers implement the pulse sequence illustrated in Fig. 1(a). IR₁ generates vibrational coherences, which are converted into population states by IR₂. After t_2 , IR₃ creates a second set of vibrational coherences that could be from the same or different vibrational modes as the first set; subsequently, an IR signal is emitted and heterodyne detected by a reference IR beam known as local oscillator (LO). The two vibrational coherences are characterized by scanning t_1 and t_3 in time domain. Typically the spectra are visualized in the frequency domain by Fourier transforming the time domain data along both the t_1 and t_3 axes into the ω_1 and ω_3 axes. The measurement of the two molecular vibrational coherences created by IR₁ and IR₃ allows 2D IR to follow molecular dynamics that occur during the measurement time frame ($t_1 + t_2 + t_3$) and also differentiate homogenous and inhomogeneous lineshape broadenings, as we explain next.

A cartoon of a 2D IR spectrum is shown in Fig. 1(b). Here, the plot follows the convention in 2D NMR, which plots the ω_1 frequency along the vertical axis and ω_3 frequency along horizontal axis. Another plot convention in which the horizontal and vertical axes are flipped is also widely used. The spectral peak pairs (1, 2 and 3) along the diagonal line are referred as diagonal peaks and the off-diagonal peaks (highlighted with dashed boxes) are called cross peaks. The diagonal peaks result from the same vibrational coherences being excited before and after waiting time t_2 . Cross peaks appear if two different vibrational coherences can be coherently excited by IR₁ and IR₃, which results from vibrational coupling (Khalil *et al.*, 2003a), energy transfer (Xiong *et al.*, 2009; Rubtsova and Rubtsov, 2013, 2015; Lin and Rubtsov, 2012; Rubtsova *et al.*, 2015), or chemical exchange (Kim and Hochstrasser, 2005; Anna *et al.*, 2009; Zheng *et al.*, 2005). Typically, a pair of peaks with opposite phase appear together along the same ω_1 frequency. This is because 2D IR spectroscopy measures both the fundamental and the overtone, or combination bands of vibrational modes. The frequency separation between fundamental and overtone peaks along ω_3 is the anharmonic shift and can be used to determine the anharmonicity, and the peak shifts of cross peaks can be used to extract coupling strength of molecular potential energy surfaces (Golonzka *et al.*, 2015). Besides the anharmonic shift, there are many unique molecular insights that can be extracted from 2D IR spectra.

One such property is the local environment and dynamics, which are encoded in the lineshape of the diagonal peaks. In liquid and solid state systems, the local environments can cause both fast pure vibrational dephasing and vibrational energy relaxation, which is referred to as homogenous broadening, and slow but heterogeneous dynamics, known as inhomogeneous broadening. In linear IR spectroscopy, the homogeneous broadening is characterized by a Lorentzian lineshape while inhomogeneous broadening is often Gaussian. It is straightforward to differentiate Lorentzian from Gaussian lineshapes, but more often the local environments create both fast and slow dynamics, which causes ambiguities in lineshape analysis of linear IR spectra. This ambiguity is overcome in 2D IR spectra. As shown in Fig. 1(b), different local environments can lead to distinct lineshapes. For homogeneous dynamics, the local environments around the probed molecules fluctuate on the time scale of molecular vibrations (~ 100 fs) and cause the molecules to lose their vibrational memory. As a result, the t_1 and t_3 vibrational coherences remain uncorrelated and the 2D IR peaks have round lineshapes (peak 2). For inhomogeneous dynamics, the local molecular environments remain static but heterogeneous on the time scale of molecular vibrations. Thus, there is little fluctuation in the vibrational frequency of the probed

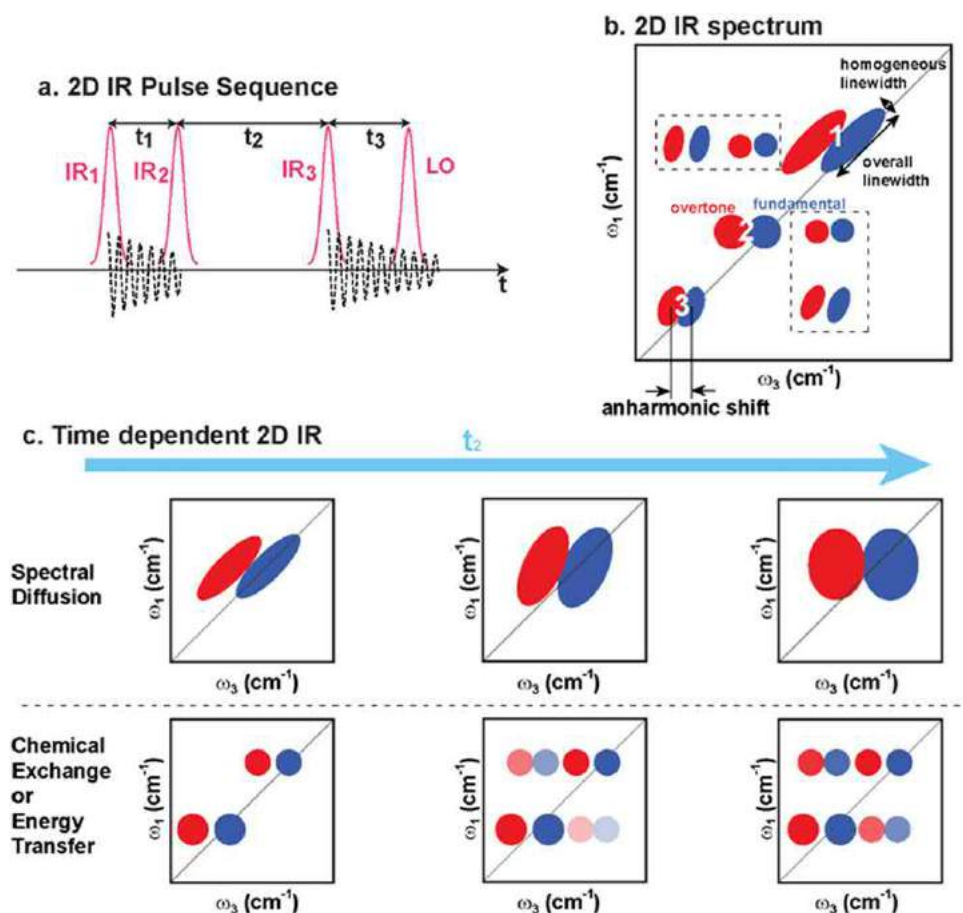


Fig. 1 Schematic 2D IR pulse sequence and spectra. (a) Pulse sequence used to generate two vibrational coherences. IR₁ creates an initial set of vibrational coherences, which are converted into population states by IR₂. IR₃ relaunches the coherences, which emit an IR signal that is heterodyned by LO. The two coherences are characterized by scanning t_1 and t_3 . The time domain scan is Fourier transformed to obtain a 2D IR spectrum in the frequency domain. (b) A schematic 2D IR spectrum highlighting spectral features such as cross peaks (marked with dashed boxes), lineshapes and the anharmonic shifts. (c) Schematic 2D IR spectra demonstrating time-dependent features. Spectral diffusion occurs as vibrational modes lose vibrational memory. As a result, tilted and elongated 2D IR peaks become round and symmetric. Cross peaks can grow in as a result of chemical exchange or energy transfer. The rise time of the two sets of cross peaks can differ, which reflects the rates of forward and backward reactions at equilibrium.

molecules – every vibrational mode maintains its vibrational memory and collectively forms a peak elongated along the diagonal (peak 1). For more complicated scenarios in which a mix of homogeneous and inhomogeneous environments exist around probed molecules, the lineshape is neither round nor elongated, but a convolution of these two (peak 3). Still it is easy to distinguish this lineshape from purely homogeneous or inhomogeneous lineshapes and thus qualitatively determine the nature of the local environments.

To obtain quantitative information, the homogeneous linewidth can be determined from the antidiagonal linewidth, whereas the diagonal linewidth corresponds to the overall linewidth (Fig. 1(b); Hamm and Zanni, 2011; Cho, 2009; Khalil *et al.*, 2003b). For molecular systems that involve dynamics in the intermediate range – dynamics that are slower than homogeneous broadening, but not as slow as inhomogeneous broadening – spectral diffusion measurements are necessary. Spectral diffusion is measured by scanning t_2 and monitoring the change of spectral lineshape. As t_2 increases and the vibrational modes lose their vibrational memory, the lineshapes evolve from elongated and tilted shapes to round shapes (Fig. 1(c), top). Various parameters have been developed to quantify lineshape changes during spectral diffusion measurements (Hybl *et al.*, 2002; Eaves *et al.*, 2005; Demirdöven *et al.*, 2002; Asbury *et al.*, 2004; Kwac and Cho, 2003b; Hamm, 2006; Fayer, 2009), making it a powerful method to characterize dynamics on the picosecond time scale. Overall, lineshape analysis of 2D IR spectra can provide many previously unavailable insights on local molecular environments and dynamics.

A second unique and useful feature of 2D IR spectra is cross peaks. Since cross peaks involve two different vibrational coherences, studying cross peaks can reveal coupling and dynamics between vibrational modes. The intensity of the cross peaks changes as a function of waiting time t_2 and can indicate chemical exchange or energy transfer (Fig. 1(c), bottom). For instance, in a molecular system with two molecular conformations in equilibrium, there will be two vibrational peaks corresponding to these

conformations along the diagonal. If chemical exchange or energy transfer occurs, cross peaks corresponding to the two conformations involved in the exchange or transfer processes should appear (Zheng *et al.*, 2005; Kim and Hochstrasser, 2005; Rubtsova and Rubtsov, 2013, 2015; Lin and Rubtsov, 2012; Rubtsova *et al.*, 2015). The rise time of the cross peaks reflects the rate of chemical exchange or energy transfer – the faster the cross peaks grow, the quicker the chemical exchange or energy transfer occurs. Furthermore, relative angles between vibrational modes can be calculated from the cross peak intensity ratio between 2D IR spectra obtained using parallel versus perpendicular pump/probe polarization schemes (Hochstrasser, 2001).

By measuring third-order vibrational response functions, 2D IR spectroscopy opens up many novel pathways to reveal molecular structure and dynamics in the liquid and solid phase. This section is only a brief introduction for the theoretical framework that prepares the audience for the discussion of scientific applications of 2D IR. There have been many comprehensive theoretical developments and computer simulations in 2D IR research by Skinner, Mukamel, Tanimura, Cho, Jansen, and others, to which we refer the audience for further reading (Auer *et al.*, 2007; Woys *et al.*, 2010; Paarmann *et al.*, 2009; Sanda and Mukamel, 2006; Ishizaki and Tanimura, 2006, 2007; Hasegawa and Tanimura, 2008; Park and Cho, 1998; Kwac *et al.*, 2004; Kwac and Cho, 2003a; Roy *et al.*, 2011).

Experimental Approaches

A 2D IR spectrometer requires an ultrafast IR light source, delay lines, a sample chamber, and a detection system. Various implementations exist for each stage of the 2D IR spectrometer, each of which has its own advantages and limitations.

Ultrafast IR Light Source

Ultrafast IR light sources that can create femtosecond pulses at μJ pulse energies and kHz repetition rates form the technical foundation for 2D IR spectroscopy. Because 2D IR signals are generated via a third-order nonlinear optical process, high pulse energies are required. The combination of solid state Ti:sapphire regenerative amplifier lasers and optical parametric amplification (OPA) provides a reliable approach to deliver femtosecond tunable mid-IR pulses for 2D IR spectroscopy. Here we give only a brief introduction on how this setup works; for detailed knowledge about OPA, please refer to an excellent review article (Cerullo and De, 2003). The light source setup typically begins with a commercial oscillator-seeded regenerative amplifier which outputs 800 nm pulses with <100 fs pulse durations and mJ pulse energies at 1 kHz. The 800 nm pulse is sent into a multi-stage OPA system, which converts a single 800 nm photon into a pair of photons at near IR wavelength. The OPA conversion is coherent and therefore requires energy and momentum conservation, as summarized in Eq. (1a,b):

$$\omega_{800\text{ nm}} = \omega_{\text{Sig}} + \omega_{\text{Idl}} \quad (1a)$$

$$k_{800\text{ nm}} = k_{\text{Sig}} + k_{\text{Idl}} \quad (1b)$$

Here, ω is the frequency of the pulses and k is the corresponding wavevector. Sig represent the emitted photon that has higher frequency, which is called signal, and Idl represents the one with lower frequency, which is called idler. The condition for conservation of momentum (Eq. 1b) is called phase matching and is achieved using birefringent non-linear optical crystals such as β -barium borate (BBO). To tune the frequency of the signal and idler, the birefringent crystal is rotated to the appropriate phase-matching tilt angle. To generate mid-IR pulses, the signal and idler pulses need to be overlapped spatially and temporally on a second non-linear crystal to undergo a difference frequency generation (DFG) process, as summarized in Eq. (2a,b):

$$\omega_{\text{sig}} - \omega_{\text{idl}} = \omega_{\text{IR}} \quad (2a)$$

$$k_{\text{sig}} - k_{\text{idl}} = k_{\text{IR}} \quad (2b)$$

The mid-IR pulse is generated at a frequency that is the difference between frequencies of signal and idler pulses. Birefringent crystals such as AgGaS_2 are widely used in DFG processes. To tune the frequency of mid-IR pulse, the tilt angle of the crystals for both OPA and DFG need to be adjusted in order to have the optimal signal and idler frequencies for the DFG process.

Although the combination of regenerative amplification and OPA is the most popular method to generate intense mid-IR pulses for 2D IR spectroscopy, the majority of these laser systems operate at 1 kHz. To improve sensitivity and broaden applications to samples with weak vibrational transitions, it is necessary to maintain pulse energies while increasing the repetition rate. The Krummel group has recently developed a home-built optical parametric chirped-pulse amplification (OPCPA) setup to implement 2D IR spectroscopy at 100 kHz (Tracy *et al.*, 2016; Luther *et al.*, 2016). They showed that a 100 kHz system is able to collect a 2D IR spectrum with high signal-to-noise in less than 1 s. This development can open new opportunities for resolving fast chemical and biological process that occur at the sub-second time scale.

Delay-Lines and Sample Chamber

After mid-IR is generated, the 2D IR pulse sequence needs to be generated and the IR pulses overlapped spatially and temporally at the sample. Here we discuss the two most popular approaches and their advantages and limitations.

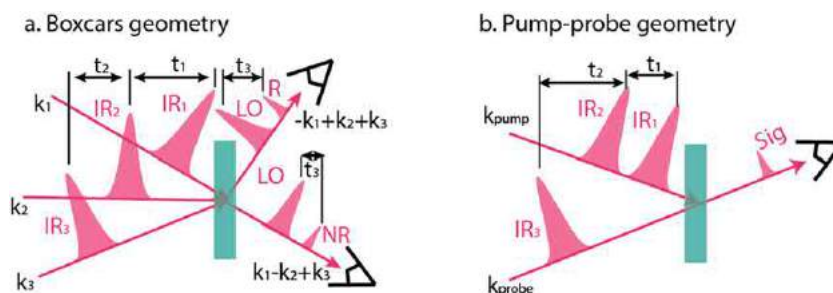


Fig. 2 Geometry of sample chamber. (a) Boxcars geometry. All pulses approach the sample at different angles, and the rephrasing (R) and nonrephasing (NR) signals emit at different phase matching angles. Signals are heterodyned by LO. (b) Pump probe geometry. IR₁ and IR₂ together form the pump beam and IR₃ is the probe beam. The phase matching directions of both rephrasing and nonrephasing signals are the same as IR₃. Therefore, IR₃ also acts as LO to heterodyne the signal, a process known as self-heterodyning.

Fig. 2(a) shows the traditional “boxcars” geometry in which three IR beams are aligned so that they approach the sample from three different angles (Khalil *et al.*, 2003b; Rock *et al.*, 2013; Asplund *et al.*, 2000). The phase-matching conditions dictate that the 2D IR signal is emitted at a fourth, unique angle. The benefit of this setup is that rephrasing and nonrephasing 2D IR spectra can be measured separately because they have different phase-matching directions. In addition, the local oscillator can be tuned independently from the other pulses to optimize the signal-to-noise ratio. The limitation of the “boxcars” geometry is that it requires long acquisition times since both t_1 and t_3 need to be scanned using mechanical stages. Thus, it usually takes several hours to complete a spectrum, which makes the setup vulnerable to phase drift and limits it from studying processes such as protein folding that take place on the minute time scale. In addition, if purely absorptive 2D IR spectra are desired, rephrasing and nonrephrasing spectra must be collected separately and their phase drift must be corrected manually before adding them together (Khalil *et al.*, 2003b).

One way to overcome the above-mentioned limitations is to use a pump-probe geometry for acquiring 2D IR spectra (Shim and Zanni, 2009; Shim *et al.*, 2009; Hamm *et al.*, 2000; Cervetto *et al.*, 2004; Rock *et al.*, 2013). In the pump-probe geometry, IR₁ and IR₂ are arranged collinearly, whereas IR₃ is sent to the sample at a different angle. Because IR₁ and IR₂ are collinear, phase matching dictates that the signal is emitted in the same direction as IR₃. This arrangement not only simplifies the instrumental setup, but also allows self-heterodyne detection; because IR₃ and the 2D IR signal propagate in the same direction, residual IR₃ can serve as the local oscillator to heterodyne the signal. Thus, no additional local oscillator is needed and the phase between signal and local oscillator does not drift. Self-heterodyned signals are often detected by sending IR₃ and signal into a monochromator and projecting them onto a detector. In this way, the time domain signal is experimentally Fourier transformed along t_3 axis into ω_3 . Thus, in the pump-probe geometry only t_1 needs to be scanned and the data acquisition time is shorter than “boxcars” geometry.

There are many ways of realizing pump-probe geometry, but the most popular and convenient method is using a mid-IR pulse shaper to construct a double mid-IR pulse pair that serves as IR₁ and IR₂ (Shim and Zanni, 2009; Shim *et al.*, 2009). This shaper-based pump-probe geometry is preferred because other experimental tricks, such as phase cycling and rotating frame, can be implemented easily and because it can scan the t_1 time delay on a shot-to-shot basis, which not only reduces noise from long term fluctuations but also enables rapid data acquisition on the time scale of seconds. Thus, it is possible to use pulse-shaper-based 2D IR setups to study fast protein folding kinetics. Another popular implementation of the pump-probe geometry is the “hole-burning” experiment, which uses a narrow band pump pulse and broad band probe pulse. Instead of taking data in the time domain, “hole-burning” experiments scan the center frequency of the narrow band pump and then plots a series of “hole-burning” pump-probe spectra at different pump frequencies to obtain a 2D IR spectrum. This method is easy to maintain and operate, but it lacks temporal resolution and the spectral lineshapes are often distorted (Hamm *et al.*, 2000; Cervetto *et al.*, 2004).

2D IR Signal Detection Systems

The detection of 2D IR signals is the final part of the 2D IR spectrometer. While the fast development of Si-based photon detectors has benefited detection in the visible regime, mid-IR detectors are both rare and expensive. The most common mid-IR detectors are HgCdTe (MCT) detectors. HgCdTe is a low band gap semiconductor material that can be excited by mid-IR radiation. To reduce charge carrier noise due to thermal fluctuations, MCT detectors often require cryogenic cooling. Both single element and array MCT detectors are used for 2D IR spectroscopy.

Single element MCT detectors are used in the traditional boxcars geometry, in which both t_1 and t_3 are scanned in the time domain. To eliminate fluctuations from the local oscillator, balanced heterodyne detection is often implemented (Mukherjee *et al.*, 2005; Fulmer *et al.*, 2004). The benefit of using a single element MCT detector is that it is economical and its fast read-out time enables shot-to-shot averaging and lock-in-detection, which improve signal to noise. However, as mentioned in the previous section, the data acquisition is slow.

Alternatively, 64- or 128-element MCT array detectors, in combination with a spectrograph, can be used for signal detection in pump-probe geometries, as discussed in the Section Delay-Lines and Sample Chamber. MCT array detectors maintain most of the

advantages of the single element detector, while the pump-probe geometry significantly reduces data acquisition times since only one time delay needs to be scanned. The only caveat is that MCT arrays are relatively expensive. Most recently, two-dimensional focal plane array (FPA) IR detectors have become available for scientific research. Since an FPA has more pixels than a traditional array detector, it allows a reference beam to be monitored simultaneously with data collection, which improves signal-to-noise. Furthermore, this detector makes new capabilities such as 2D IR imaging possible (Serrano *et al.*, 2015; Ostrander *et al.*, 2016).

Besides directly detecting mid-IR signals, an alternative approach is to upconvert IR photons into the visible regime, where sensitive visible detectors can detect the signal. Kubarych and co-workers pioneered this approach by mixing a chirped narrowband 800 nm pulse with the emitted IR signal on a non-linear crystal to upconvert the emitted IR signal into a visible signal via sum-frequency generation (Nee *et al.*, 2008, 2007; Anna and Kubarych, 2010). Because the upconversion process is non-resonant, the temporal profile of the original 2D IR signal is well-preserved. The upconversion process enables the use of visible light detectors that are typically cheaper and more sensitive to be used for 2D IR spectroscopy. So far, both charge-coupled device (CCD) and complementary metal-oxide-semiconductor (CMOS) cameras have been implemented in 2D IR spectrometers (Rock *et al.*, 2013).

Applications

The experimental methods discussed above represent the most popular approaches to obtain 2D IR spectra. Each has its own advantages and limitations, but all have allowed 2D IR spectroscopy to provide unprecedented insights about molecular structure and dynamics. Below, we discuss a few examples in which 2D IR spectroscopy has been applied to problems across chemistry, materials, and biology.

Material Sciences

Solid state materials

2D IR spectroscopy has been implemented to gain unique insights into the structures and dynamics of materials, which has implications on designing and controlling materials at molecular level. 2D IR spectroscopy can better resolve molecular inhomogeneities than FTIR (Xiong *et al.*, 2009; Oudenhoven *et al.*, 2015; Barbour *et al.*, 2006), which is very useful in studying heterogeneous solid state materials. Furthermore, the cross peaks of 2D IR spectra are often used to follow dynamic processes in materials, such as ion-exchange and charge (Kwon and Park, 2015; Fulfer and Kuroda, 2016; Xiong *et al.*, 2009). Lastly, by measuring 2D IR spectra at different t_2 , it is feasible to probe the dynamics of molecular ligands or solvent molecules near the material surfaces (Yan *et al.*, 2015, 2016; Nishida *et al.*, 2014; Cui *et al.*, 2016; Huber *et al.*, 2015; Huber and Massari, 2014; Ren *et al.*, 2015, 2014; Dutta *et al.*, 2015).

Charge dynamics of dye-sensitized solar cells

2D IR spectroscopy reveals that charge transfer dynamics in dye-sensitized solar cells (DSSCs) are conformation dependent (Xiong *et al.*, 2009; Oudenhoven *et al.*, 2015). In DSSCs, multiple timescale charge transfer dynamics have been observed by transient absorption spectroscopy. It is speculated that dye molecules can form different conformations on the surface of nanoparticles, each of which has its own charge transfer pathway which therefore lead to the multiple time scale dynamics. However, it is difficult to resolve conformations using FTIR spectroscopy, in which spectral peaks are congested. It is also difficult to distinguish conformations in traditional visible-pump IR-probe experiments because the vibrational peaks are on top of a broad free-electron absorption background, which makes resolving vibrational features difficult (Anderson and Lian, 2005). Using 2D IR spectroscopy, Xiong, Zanni and co-workers studied the surface conformation and charge dynamics of $\text{Re}(\text{CO})_3(\text{Bi-pyridine})\text{COOHCl}$ sensitized TiO_2 nanoparticles. Three vibrational peaks of the $A'(1)$ mode of the CO stretch of the dye molecules are clearly resolved (Fig. 3(a)). In solution phase, the $A'(1)$ mode only results in one vibrational peak. Thus, the three distinct peaks reflect that three different conformations are formed at the TiO_2 surface. The reason that 2D IR can differentiate vibrational peaks better than FTIR is that 2D IR signal is proportional to transition dipole moment, μ , to the fourth order (μ^4) whereas FTIR depends on μ^2 .

To understand how different conformations contribute to the charge transfer dynamics, a UV pulse is inserted between IR_2 and IR_3 during waiting time t_2 . This allows 2D IR to correlate conformations before and after charge transfer has occurred (Bredenbeck *et al.*, 2004). When a new cross peak appears in the 2D IR spectrum, it reveals that charge transfer has occurred and the corresponding conformation is in an electronic excited state. At the same time, diagonal peaks of all three conformations are bleached, demonstrating that the ground state populations of all three conformations are decreased. Thus, all conformations are excited by the UV pulse and involved in charge transfer. However, the conformation at $\omega_{\text{pump}} = 2040 \text{ cm}^{-1}$ does not have the excited state cross peak that the lower frequency conformations do (Fig. 3(b)). This difference indicates that at the time that the dye molecules are probed, the lower frequency conformations are still in electronic excited states, but the conformation at $\omega_{\text{pump}} = 2040 \text{ cm}^{-1}$ is in neither a ground nor excited electronic state. Therefore, it is highly likely that the electron has transferred from this conformation to the TiO_2 (Xiong *et al.*, 2009). Further investigations using 2D IR spectroscopy show that these three conformations are three different aggregates of dye at the surface of TiO_2 (Oudenhoven *et al.*, 2015). Thus, 2D IR has a unique capability to resolve multiple molecular conformations in materials and further resolve charge transfer dynamics with conformational specificity when combined with UV or visible excitation pulses.

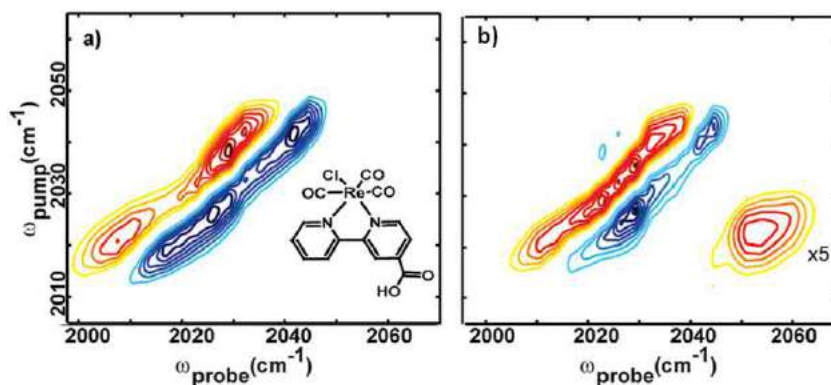


Fig. 3 2D IR spectra of dye-sensitized solar cells. (a) A static 2DIR spectrum shows three vibrational peaks, indicating three different conformations when dyes attach to the surface of TiO_2 . (b) A transient 2D IR spectrum after charge transfer is initiated by UV pulse. Since all three diagonal peaks appear, all three conformations are involved in the charge transfer. However, a large cross peak only appears at $\omega_{\text{pump}} = 2020 \text{ cm}^{-1}$, whereas no cross peak appears at higher frequency. This result indicates that different conformations have different charge transfer dynamics. Details refer to main text. Adapted from Xiong, W., Laaser, J.E., Paoprasert, P., *et al.*, 2009. Transient 2D IR spectroscopy of charge injection in dye-sensitized nanocrystalline thin films. *Journal of the American Chemical Society* 131, 18040–18041.

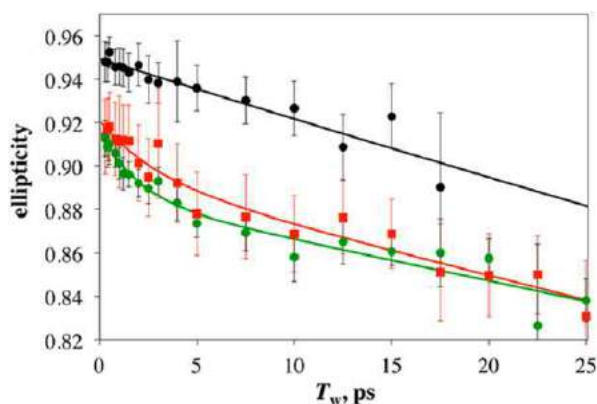


Fig. 4 Spectral diffusion of Si-H on the surface of nanoscale pores, infiltrated with pentane (black circles), isopropanol (green circles), and isopropanol replaced with pentane (red squares). Overlaid solid lines are multiexponential fits to the data. After pentane replaces isopropanol as the solvent in nanoscale pore, spectral diffusion dynamics remain the same as isopropanol-infiltrated pores. This suggests that isopropanol strongly binds to the silica surface and it cannot be displaced by pentane. Adapted from Huber, C.J., Massari, A.M., 2014. Characterizing solvent dynamics in nanoscopic silica sol-gel glass pores by 2D-IR spectroscopy of an intrinsic vibrational probe. *Journal of Physical Chemistry C* 118, 25567–25578.

Nanoscale materials

Nanoscale materials are attractive to scientific and technological developments due to their large surface area and tunable mechanical, optical and electromagnetic properties. In particular, nanoscale porous materials have been developed for chemical and energy storage purposes. Molecules inside these materials are spatially confined and experience molecule-pore surface interactions which can change the structure and dynamics of confined molecular ensembles (Nishida *et al.*, 2014; Huber *et al.*, 2015; Huber and Massari, 2014). Understanding molecule-pore surface interactions and how they affect confined molecular ensembles is critical to optimizing the capacity and selectivity of nanoscale porous materials. To understand these effects, different vibrational probes have been developed for 2D IR spectroscopy to investigate the structure and dynamics of molecules confined in nanoscale pores.

Si-H vibrational modes on silica surfaces have been used to probe local solvent dynamics in porous silica nanoparticles. It is shown that spectral diffusion measurements of the Si-H modes are sensitive to solvent dynamics up to 5 nm away (Huber *et al.*, 2015). Using the same vibrational probes, solvent-pore surface interactions are studied for nanoscale silica sol-gel glass pores. Huber and Massari observe different spectral diffusion dynamics when the pores are infiltrated with pentane versus isopropanol. More interestingly, when isopropanol in the nanopores is replaced by pentane, the spectral diffusion dynamics remain similar to the isopropanol-infiltrated nanopores (Fig. 4). This result suggests that although pentane could replace the majority of solvents molecules inside the pores, isopropanol strongly bound to silica surface and the binding process is irreversible (Huber and Massari, 2014).

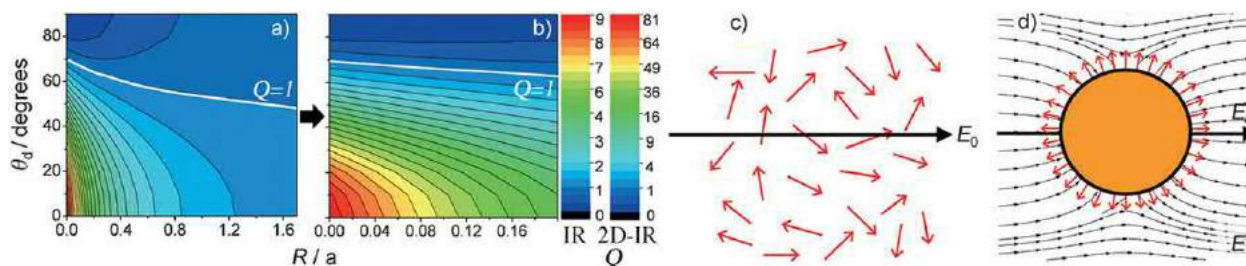


Fig. 5 Signal enhancement of 2D IR spectroscopy on metal nanoparticles. Plots of the IR (a) and 2D-IR (b) signal enhancement factors as a function of dipole tilt angle, θ_d , and distance from the surface, R , for an orientationally-averaged ensemble of dipoles located around a spherical metal nanoparticle of radius a . Illustration of relative angles between electric field and transition dipole moments in isotropic (c) and nanoparticle (d) situations. Adapted from Donaldson, P.M., Hamm, P., 2013. Gold nanoparticle capping layers: Structure, dynamics, and surface enhancement measured using 2D-IR spectroscopy. *Angewandte Chemie – International Edition* 52 (2), 634–638.

Another nanoporous material that has been studied by 2D IR is metal-organic-frameworks (MOFs). $[\text{FeFe}](\text{dcbdt})(\text{CO})_6$ is attached to a UiO-66 MOF to probe the dynamics of inner pores (Nishida *et al.*, 2014). Nishida, Fayer, and coworkers observed a slow 670 ps spectral diffusion constant in empty MOF pores, which is attributed to a large-scale motion of the MOFs. When filled with DMF, the spectral diffusion becomes even slower. This result indicates that MOF structures become more rigid and the large scale motions are further slowed by hosting DMF. More interestingly, no large homogeneous broadening is observed. This is in sharp contrasts with the spectral lineshape of $[\text{FeFe}](\text{dcbdt})(\text{CO})_6$ in bulk DMF solutions, which is significantly homogeneously broadened. Since in solution the broad homogeneous lineshape can be attributed to solvent motions that occur faster than or on the same time scale as the time resolution of 2D IR, the lack of homogeneous broadening in the 2D IR spectra of DMF-infiltrated MOFs indicates that the motion of DMF molecules is largely constrained. These pioneering works demonstrated how 2D IR spectroscopy could provide insights into interactions between molecules and surface of materials in nanoscale, confined spaces.

Nanoscale materials are not only interesting scientific systems to be investigated by 2D IR spectroscopy, they can also advance 2D IR spectroscopic techniques (Donaldson and Hamm, 2013; Selig *et al.*, 2015). Donaldson and Hamm showed theoretically and experimentally that the 2D IR signal can be enhanced for molecules bound at the surface of metal nanoparticles (Donaldson and Hamm, 2013). This can be understood in a simplified scenario: when molecules bind to a metal nanoparticle, certain vibrational transition dipoles may orient perpendicular to the surface. These transition dipoles are also parallel to the near-field electric fields, which orient parallel to the surface normal. Therefore, the vibrational modes of surface-bound molecules are aligned uniformly with the electric fields (Fig. 5(d)). As a result, the dipole interaction with the electric field is strengthened relative to the solution phase, where the relative angles between transition dipoles and electric fields are isotropically distributed (Fig. 5(c)). In this case, the 2D IR signal is enhanced 81 times. The more general scenario is described in Fig. 5(a) and (b), in which the enhancement factor was calculated as a function of distance to the metal surface, R , and tilt angle to surface normal, θ_d . The enhancement is strongest when R is small and θ_d is 0. As R becomes large and θ_d approaches to 90, the enhancement factor decrease to zero. Experimentally, 2D IR signal enhancement is measured using amide-PEG coated gold nanoparticles, where the distance to gold surface and particle sizes are varied. It is found that enhancement occurs up to 1 nm from the surface and the enhancement is larger when large (radius = 10 nm) nanoparticles are used.

Molecular electrolytes

Ion structure and dynamics of electrolytes

Molecular properties of electrolytes directly influence the performance of batteries, particularly in lithium ion batteries which are the most commonly used energy storage devices. A wide range of ion-molecule and ion-ion interactions exist in the electrolyte which can influence ion diffusions, pairing and solvation (Kwon and Park, 2015; Cui *et al.*, 2016; Fulfer and Kuroda, 2016). When complexes are formed between ions or between ions and molecules, the vibrational frequencies of the ions and molecules can shift due to differing electrostatic environments. As a result, the dissociated and associated configurations have two distinct vibrational peaks in FTIR. Similarly, 2D IR spectra contain diagonal peaks of different configurations and cross peaks that can reveal the exchange dynamics between these configurations. Thus, 2D IR spectroscopy is very useful to resolve dynamics of ion-ion or ion-molecule complex formations in electrolytes. The structure and dynamics of the electrolyte between LiPF_6 and organic carbonates have been studied using 2D IR spectroscopy (Fulfer and Kuroda, 2016). Together with DFT calculations, Fulfer and Kuroda find evidence to support that Li^+ and organic carbonates form two types of ion pairs, contact and solvent-separated ion pairs, which has implications in rational design of electrolyte composition for better Li batteries. In another work, a model Li^+ electrolyte system, Li^+ and CH_3SCN in acetonitrile, is investigated by Kwon and Park using 2D IR spectroscopy, where the SCN vibrational modes of CH_3SCN are probed (Kwon and Park, 2015). Two distinct diagonal peak pairs are resolved in the 2D IR spectrum at 1 ps (Fig. 6). The high frequency peak (C) corresponded to the $\text{Li}^+ \cdot \text{CH}_3\text{SCN}$ complex, and the peak at low frequency (F) is assigned to the free CH_3SCN . Please note the second plot convention is used in these 2D IR spectra, as discussed in Section Brief Theory. As t_2 increases, cross peaks start to appear. Due to interference with the overtone signal of the diagonal peaks, part of the cross peaks (X2) does not appear clearly. Nevertheless, peaks that are well-separated from diagonal peaks (X1 and X3) can be identified.

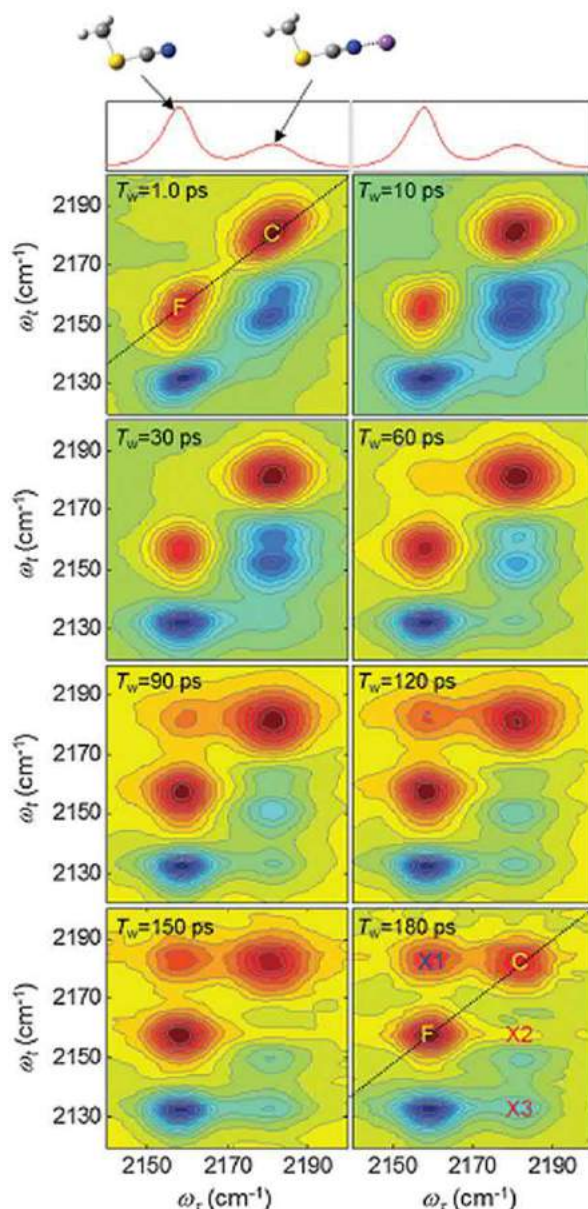


Fig. 6 Time-dependent 2D IR spectral series of Li^+ and CH_3SCN . The free CH_3SCN (F) and Li^+ - CH_3SCN complex (C) are resolved along 2D IR diagonal peaks. As time increases, the free and complex forms exchange conformation and therefore cross peak (X) appears. Adapted from Kwon, Y., Park, S., 2015. Complexation dynamics of CH_3SCN and Li^+ in acetonitrile studied by two-dimensional infrared spectroscopy. *Physical Chemistry Chemical Physics* 17, 24193–24200.

X1 corresponds to transformation from F to C, which is the association process, and X3 corresponds to the dissociation reactions. By following the cross peak growth dynamics and fitting them using a kinetic model, the dissociation time constant is determined to be 160 ± 10 ps and the association time constant is 360 ± 20 ps. These quantitative results provide some fundamental parameters for understanding dissociation and association dynamics of electrolytes for Li^+ batteries.

Another special electrolyte material is an ionic liquid (Dutta *et al.*, 2015; Brinzer *et al.*, 2015; Ren *et al.*, 2015, 2014). Ionic liquids are molten salts at room temperature that are composed of organic cations and anions. Besides their use as electrolytes, ionic liquids also have applications in carbon capture and tribology. Multiple molecular forces exist in ionic liquids, which leads them to have unique but complicated molecular properties. For instance, it is observed that a decrease in hydrogen bonding can lead to a large increase in viscosity, which is counterintuitive. 2D IR spectroscopy has been used to understand the molecular origins of an ionic liquid's viscosity (Ren *et al.*, 2014). In this work, two ionic liquids, 1-butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide ($[\text{C}_4\text{C}_1\text{im}][\text{NTf}_2]$) and 1-butyl-2,3-dimethylimidazolium bis(trifluoromethylsulfonyl)imide ($[\text{C}_4\text{C}_1\text{C}_1^2\text{im}][\text{NTf}_2]$), are studied. These two ionic liquids only differ by one H atom that is replaced by CH_3 in $[\text{C}_4\text{C}_1\text{im}][\text{NTf}_2]$, which breaks the hydrogen bonds between the cation and anions. Ren, Garrett-Roe and co-workers measure spectral diffusion of

both compounds and find that the time constants of spectral diffusion dynamics, i.e. the ion complex breaking and reforming dynamics, are 47 ± 15 ps for the methylated ionic liquid ($[\text{C}_4\text{C}_1\text{C}_1^2\text{im}][\text{NTf}_2]$) and 26 ± 3 ps for ($[\text{C}_4\text{C}_1\text{im}][\text{NTf}_2]$). The authors further suggest that this difference in the ion-complex reorganization rate is responsible for the difference in viscosity between the two liquids. In a Stokes-Einstein picture, the fundamental limit to activate translational diffusion and viscosity is the time to break up the local ion-complex; the faster ion complex reorganization is, the lower viscosity will be. Besides viscosity, other properties of ionic liquids, such as solute-solvent interactions and interactions between CO_2 and ionic liquids, have also been investigated (Brinzer *et al.*, 2015; Ren *et al.*, 2015; Dutta *et al.*, 2015).

Organometallic chemical catalysts

Organometallic chemical catalysts have been developed and used for many reactions, such as CO_2 reduction (Kumar *et al.*, 2012), H_2O splitting (Weston *et al.*, 2011), and CH activation (Crabtree, 2004). During catalytic reaction cycles, organometallic catalysts undergo transitions from one intermediate state to another on a fast time scale that cannot be followed by techniques such as NMR spectroscopy. 2D IR spectroscopy has been used to explore the transitions between different intermediates with fs or ps temporal resolution (Jones *et al.*, 2011; Nee *et al.*, 2008; Anna and Kubarych, 2010; Cahoon *et al.*, 2008).

Dynamics of homogeneous catalysts

One beautiful example is a study of the famous Berry's pseudorotation of $\text{Fe}(\text{CO})_5$ by Cahoon, Harris and co-workers (Cahoon *et al.*, 2008). Berry's pseudorotation predicts that $\text{Fe}(\text{CO})_5$, which has a D_{3h} geometry, can undergo transformations through C_{4v} geometry and return back to the original D_{3h} geometry. During this process, however, the two axial CO groups exchange with two equatorial CO groups (Fig. 7(B)). $\text{Fe}(\text{CO})_5$ has two IR active modes. The doubly degenerate e' modes that involve three equatorial CO groups absorb at 1999 cm^{-1} and the a'' mode, which is composed of vibrations of axial CO groups, absorbs at 2022 cm^{-1} . It is observed that as t_w increases, cross peaks between the e' and a'' modes appear in the 2D IR spectra (Fig. 7(A–C)). The cross peak dynamics reflect Berry's pseudorotation: as axial and equatorial CO groups exchange their positions, the vibrational energy they carry would also exchange from a'' to e' and vice versa. The authors model the vibrational energy transfer process and simulate corresponding 2D IR spectra (Fig. 7(D–F)). They found the pseudorotation occurs within 8.0 ± 0.6 ps, with an activation barrier of 2.13 kJ/mol (Fig. 7(G)). Besides this example, there are many other interesting studies on the isomerization and vibrational dynamics of organometallic catalysts in solution phase using 2D IR spectroscopy (Anna *et al.*, 2009; Nee *et al.*, 2008; Anna and Kubarych, 2010; Jones *et al.*, 2011). These works provide valuable insights on the dynamics of transition states in organometallic catalysts.

Surface attached catalysts

Another exciting application is to use reflection mode 2D IR spectroscopy to study surface-attached catalysts (Rosenfeld *et al.*, 2011, 2013; Yan *et al.*, 2015, 2016; Nishida *et al.*, 2016; Wang *et al.*, 2015; Li *et al.*, 2016a,b). Organometallic catalysts have been attached to metals and semiconductors to perform electrochemistry. Yet, there is a lack of fundamental knowledge on their structure and dynamics, which could be affected by the surface attachment. The Fayer group studied a set of fac-Re organometallic catalyst attached by click chemistry onto gold and silica surfaces through various lengths of linkers (Rosenfeld *et al.*, 2011, 2013, 2012; Yan *et al.*, 2015, 2016; Nishida *et al.*, 2016). By carefully controlling the surface density of these Re compounds, it is shown that the spectral diffusion

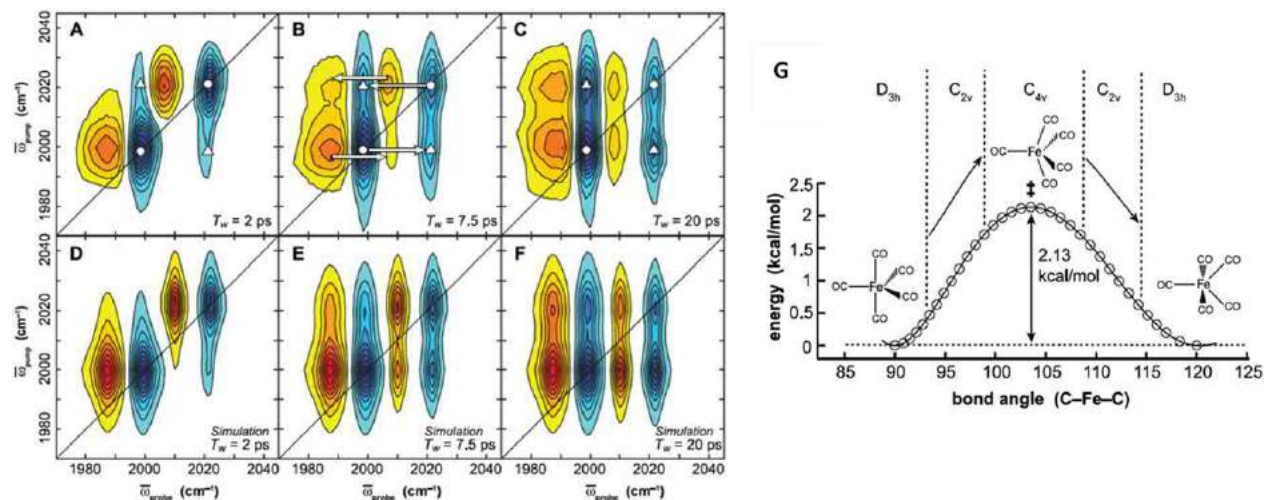


Fig. 7 2D IR measurement of Berry's pseudo rotation on $\text{Fe}(\text{CO})_5$. (A–C) Experimental 2D IR spectra at different time delays show cross peak growth as waiting time increases. (D–F) Simulated 2D IR spectra based on a vibrational energy transfer model. (G) Depiction of the mechanism and energetics of a pseudorotation plotted as a function of the bond angle (C–Fe–C) between one axial and one equatorial CO ligand that exchange during the pseudorotation. Adapted from Cahoon, J.F., Sawyer, K.R., Schlegel, J.P., Harris, C.B., 2008. Determining transition-state geometries in liquids using 2D-IR. *Science* 319, 1820–1823.

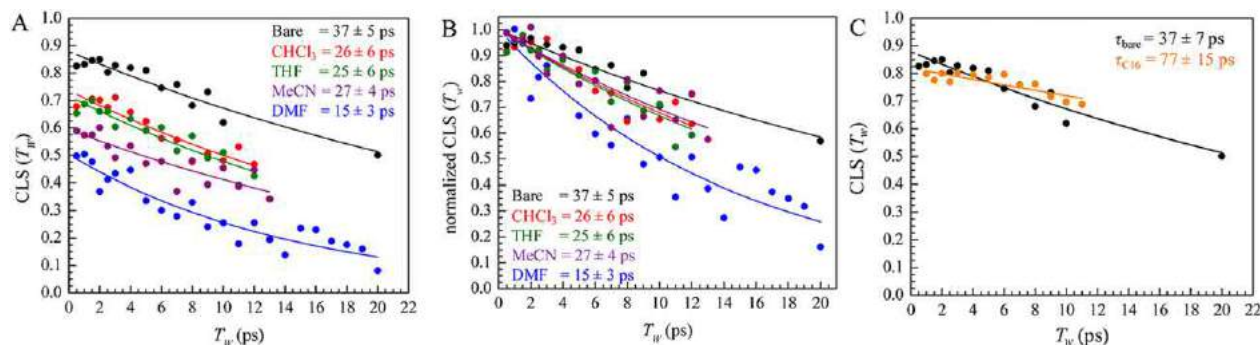


Fig. 8 Spectral diffusion measurements of surface-immobilized catalysts with different solvent environments. (A) Spectral diffusion and (B) normalized spectral diffusion dynamics in organic polar solvents. (C) Spectral diffusion in non-polar solvents. When surface-immobilized catalysts are immersed in solvents that can dissolve the catalysts, they have faster dynamics than under solvent-free environments. Adapted from Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2013. Structural dynamics at monolayer – liquid interfaces probed by 2D IR spectroscopy. *Journal of Physical Chemistry C* 117, 1409–1420.

measurement is primarily affected by structural dynamics such as constrained-rotation motions and surrounding solvent dynamics (Rosenfeld *et al.*, 2012, 2013). The spectral diffusion time constants are especially sensitive to the surrounding solvents. When polar solvents that can solvate the Re compounds are used, the corresponding spectral diffusion is faster than on a “dry” catalyst surface (Rosenfeld *et al.*, 2013). However, when solvents that do not solvate the Re compounds are used, the spectral diffusion dynamics are significantly slower – even slower than the “dry” surface (Fig. 8). The difference in spectral diffusion demonstrates that the local solvent structures differ greatly depending on the type of solvent used. The polar organic solvents could swell the surface and penetrate into the catalytic layer, whereas the other solvents might cause a compaction of the surface. Besides being highly sensitive to the solvent, it is also shown that the spectral dynamics are sensitive to the length of the linker, the surface coverage, and the substrate (Yan *et al.*, 2014, 2015). In general, the longer the linker, the faster the dynamics are. This agrees with the intuition that longer linkers are more flexible. Furthermore, it is found that the dynamics on gold surfaces is much slower than on SiO_2 , because the thiol anchoring groups used on the gold surface are mobile and allow 2D “recrystallization” into a more compact monolayer, which slows the structural dynamics of the catalysts. A recent combination of 2D IR spectroscopy and MD simulations explores the molecular origin of the spectral diffusion of these catalysts. It is found many types of linker dynamics contribute to spectral diffusion, including dihedral flips in the alkyl chain and triazole ring flips, which drive subsequent chain motions (Yan *et al.*, 2016). These novel developments and applications to surface attached organometallic catalysts allow in-depth knowledge to be gained about the structure and dynamics of the immobilized catalysts and the surrounding local solvents.

Protein Structure and Folding Dynamics

The function of a protein is determined by its structure and all biological processes involve changes in protein conformation. A variety of techniques exist to study protein structure, including X-ray crystallography, NMR spectroscopy, and electron microscopy. These techniques are best applied to static structures, however, and rarely allow structural dynamics to be observed directly. In contrast, 2D IR spectroscopy can be used to observe protein conformational dynamics that occur over a wide range of timescales, from picoseconds to hours (Ganim *et al.*, 2008; Kim and Hochstrasser, 2009; Hamm *et al.*, 2008; Woutersen and Hamm, 2002; Zhuang *et al.*, 2009).

Most infrared studies of proteins focus on the backbone amide I mode, which is generated primarily by the $\text{C}=\text{O}$ stretching motion with a smaller contribution from N-H stretching. Within a protein, the amide I modes of individual amino acids vibrate in unison to create delocalized normal mode vibrations that extend across multiple amino acids. This coupling, which depends strongly on the secondary structure of the protein, causes a change in the frequency of the amide I mode. For example, the random couplings between residues within disordered peptides cause the amide I mode to appear as a broad peak centered around 1650 cm^{-1} , while the regular coupling patterns within β -sheets structures causes the amide I mode to narrow and shift to lower frequencies, which can range from 1620 cm^{-1} for highly ordered, extended β -sheets to 1635 cm^{-1} for smaller β -sheets. Thus, changes in the frequency of the amide I mode indicate changes in protein secondary structure (Fig. 9). A few 2D IR studies of proteins have examined additional vibrational modes, including the backbone amide II mode (Maekawa *et al.*, 2009), which comprises N-H bending and C-N stretching, and sidechain modes (Buchanan *et al.*, 2014).

While the amide I mode can be used to determine the overall structural content of a protein, a 2D IR spectrum of a wild-type protein generally does not contain enough information to determine a more detailed protein structure. Yet, 2D IR spectroscopy is capable of residue-specific structural resolution when applied to labeled proteins. Isotope labels can be incorporated into the protein during synthesis or expression (Middleton *et al.*, 2010; Moran *et al.*, 2012) and do not perturb the structure or the dynamics. Commonly, either $^{13}\text{C}=^{16}\text{O}$ or $^{13}\text{C}=^{18}\text{O}$ labeling of the backbone carbonyl is used. The heavier nuclei cause the amide I frequency of the labeled residues to shift by 40 or 60 cm^{-1} , respectively, so that the labeled residues are spectrally resolved from the remaining unlabeled residues (Fig. 9(b)). As a result, the secondary structure of the labeled residue can be determined by analyzing the frequency, linewidth, and crosspeaks of the labeled mode.

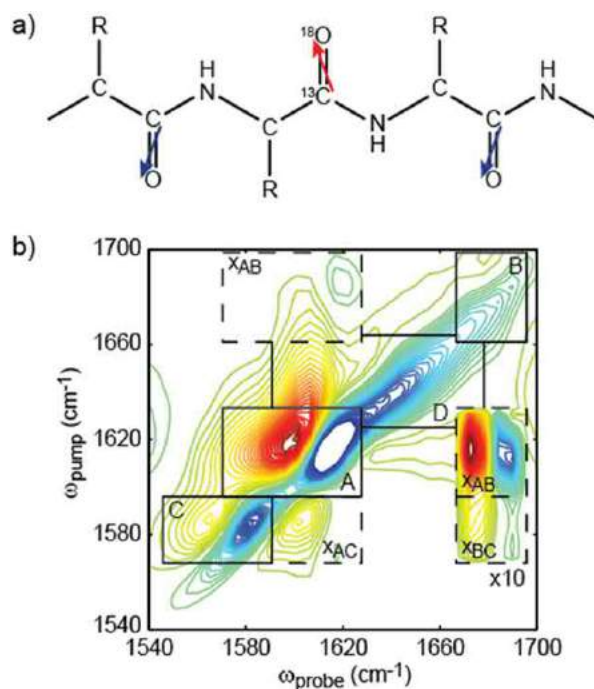


Fig. 9 Overview of protein 2D IR spectroscopy. (a) Peptide sequence with amide I modes marked with arrows. The dipole lies about 20° off the $\text{C}=\text{O}$ bond. One carbonyl is isotope labeled with $^{13}\text{C}^{18}\text{O}$, which shifts the frequency of that mode by 60 cm^{-1} . (b) Example spectrum for an isotope-labeled peptide with diagonal peaks, enclosed in solid boxes, characteristic of β -sheet (A), β -turn (B), and disordered (D) structures. The isotope-labeled mode (C) appears well separated from the other amide I modes. Crosspeaks between diagonal modes, enclosed in dashed boxes, appear when the corresponding regions of secondary structure are coupled. Spectrum reproduced from Buchanan, L.E., Dunkelberger, E.B., Zanni, M.T., 2012. Examining amyloid structure and kinetics with 1D and 2D infrared spectroscopy and isotope labeling. In: Fabian, H., Naumann, D. (Eds.), *Protein Folding and Misfolding: Shining Light by Infrared Spectroscopy*, vol. 1. Springer, pp. 217–237.

In the follow sections, we will highlight a few examples of how 2D IR spectroscopy has been applied to amyloid fibrils and membrane proteins.

Amyloid proteins

The misfolding and aggregation of proteins into amyloid fibrils is associated with more than 20 human diseases including type II diabetes, Alzheimer's disease, prion diseases, and cataracts. Yet, there are also cases of functional amyloids that are required for necessary for biological functions such as the formation of melanin (Chiti and Dobson, 2006) and recent studies indicate that the ability to form amyloids may be common to all proteins (Goldschmidt *et al.*, 2010). Thus, an understanding of amyloid protein structure and formation is important not only for therapeutic applications but also for the broader field of protein folding.

Amyloid fibers are characterized by extended, intermolecular β -sheets in which the individual β -strands are oriented perpendicular to the fiber axis. Atomic-level structures of amyloid proteins are difficult to determine. While a few X-ray crystallography and NMR spectroscopy studies have been completed on amyloid fragments and polypeptides, the most common techniques for studying amyloids and their formation are circular dichroism spectroscopy, fluorescence spectroscopy, and transmission electron microscopy, which provide limited structural information beyond identifying the presence of fibers. Additionally, a mechanistic understanding of amyloid formation requires temporal resolution so that intermediate states may be identified. This is especially true in the case of amyloid diseases, where growing evidence suggests that prefibrillar intermediates rather than the fibrils themselves may be the pathogenic species.

2D IR spectroscopy has been used to study numerous amyloid proteins, including human amylin (type II diabetes) (Shim *et al.*, 2009; Buchanan *et al.*, 2013), amyloid beta (Alzheimer's disease) (Kim *et al.*, 2008, 2009), polyglutamine (Huntington's disease) (Buchanan *et al.*, 2014), and crystallins (cataracts) (Moran *et al.*, 2012). A variety of labeling schemes are used in these studies ranging from uniform labeling to residue-specific labeling.

Structure of polyglutamine fibrils

Several neurodegenerative diseases, including Huntington's disease, are associated with the formation of amyloid fibers by proteins containing mutationally-expanded polyglutamine tracts. A wide variety of structures has been proposed for the polyglutamine aggregates, but the two most widely supported structures are the β -arc and β -turn models (Fig. 10, top; Kar *et al.*, 2013; Kajava *et al.*, 2016). Each model has the polyglutamine sequence forming two antiparallel β -strands separated by a short turn, but differ in the backbone hydrogen-bonding structure. In the β -arc model, the two β -strands within each monomer contribute to the formation of

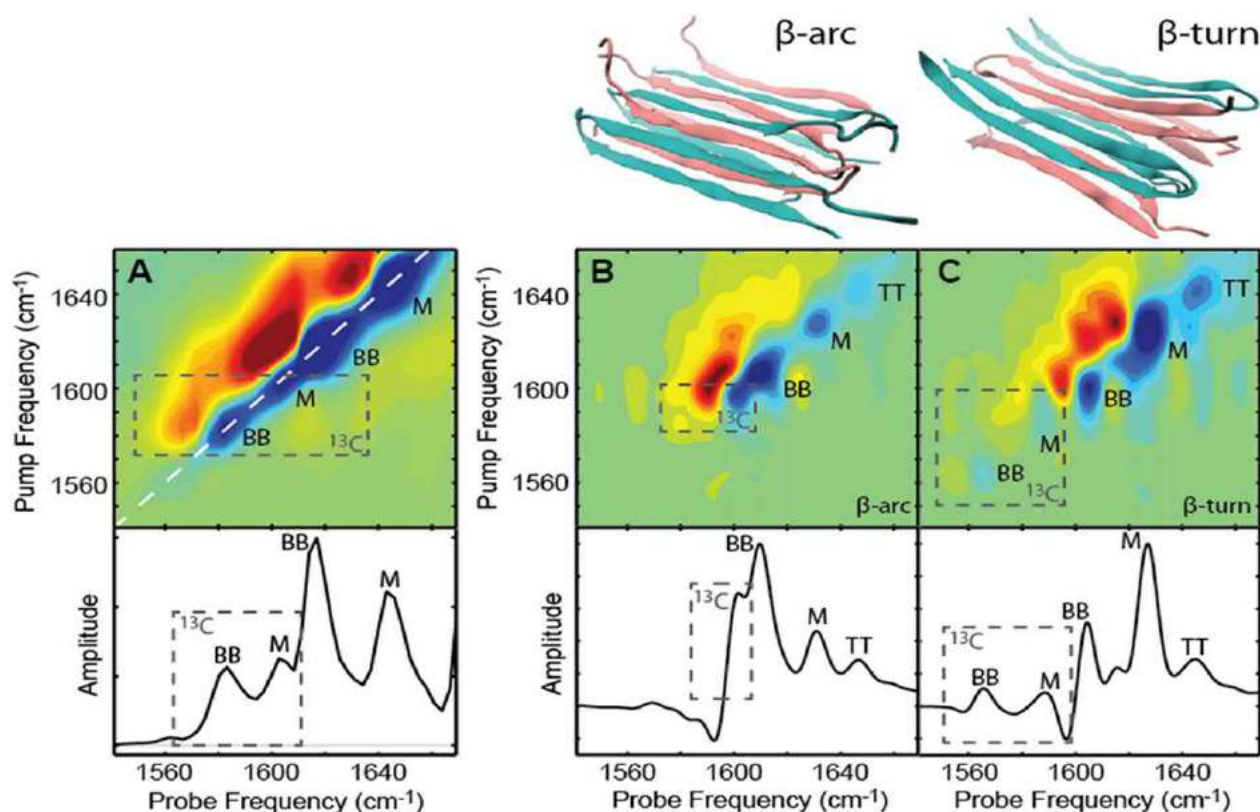


Fig. 10 Experimental and simulated 2D IR spectra of isotope-diluted Q24 fibrils. (A) Experimental 2D IR spectrum and diagonal intensity slice of fibrils formed from 10% ^{13}C -labeled Q24 mixed with natural abundance ^{12}C Q24. Simulated 2D IR spectra and slices for the (B) β -arc and (C) β -turn models of isotope-diluted Q24 fibrils of the same composition. The peaks are labeled BB for backbone modes, M for mixed backbone-sidechain modes, and TT for disordered modes at the turns and termini. Schematic visualizations of the fibril models are shown above their respective simulated spectra, with monomeric units shown in alternating colors for clarity. Adapted from Moran, S.D., Woys, A.M., Buchanan, L.E., *et al.*, 2012. Two-dimensional IR spectroscopy and segmental ^{13}C labeling reveals the domain structure of human γ D-crystallin amyloid fibrils. *Proceedings of the National Academy of Sciences* 109 (9), 3329–3334.

two separate β -sheets. In the β -turn model, the two β -strands contribute to the same β -sheet. In both structures, the side chains are intercalated to stabilize the stacked β -sheets.

While 2D IR spectra of Q24, a polyglutamine peptide with 24 glutamine residues, agree with the prediction that polyglutamine forms antiparallel β -sheets, infrared spectroscopy alone cannot differentiate between the two models described above because the backbone structures of the β -sheets are nearly identical between models. However, if it were possible to obtain the spectrum of an individual monomer within the fibril, the structure could be distinguished by determining whether the two β -strands were coupling to each other within a β -sheet (β -turn) or if they were isolated in separate β -sheets with minimal coupling between strands (β -arc). Buchanan *et al.* used mixtures of uniformly labeled Q24 peptides to vibrationally isolate monomers within the fibrils (Buchanan *et al.*, 2014). Q24 peptides were expressed in *E. coli* so they could be either uniformly ^{13}C -labeled or left at natural abundance ^{12}C . Samples were prepared of 10% ^{13}C -labeled Q24 mixed with 90% natural abundance Q24. At that ratio, it is statistically unlikely that any two ^{13}C -labeled peptides will be stacked consecutively within a fibril. Thus, the frequency of the ^{13}C amide I mode will reflect only couplings between strands of the same monomer. The experimental spectra (Fig. 10(A)) clearly identify the β -turn model as the structure for this peptide and show strong agreement with spectra computed from molecular dynamics simulations of these models (Fig. 10(B) and (C)).

Kinetics of amyloid formation by human amylin

Human amylin is a 37 amino acid polypeptide implicated in type II diabetes. Based on solid-state NMR measurements and molecular dynamics simulations, Tycko and co-workers proposed a model for amylin in which the monomers form two β -strands connected by a short disordered region (Luca *et al.*, 2007). The strands stack into two separate parallel β -sheets. This structure was confirmed by 2D IR measurements in which a series of fibers were formed in which a single amino acid had been labeled with ^{13}C = ^{18}O via solid-phase peptide synthesis (Wang *et al.*, 2011). When amylin aggregates into in-register parallel β -sheets, the isotope labeled amino acids align within the sheets; this creates a column of coupled oscillators that form their own set of normal modes within the β -sheets, red-shifting the frequency of the isotope-labeled mode just as coupling within a β -sheet shifts the unlabeled amide I. Thus, the frequency of an isotope-labeled mode is an excellent indicator of whether the labeled residue adopts a β -sheet structure.

With the development of mid-IR pulse shaping, 2D IR spectra can be collected rapidly and continuously over the course of amyloid formation. Shim *et al.* used 2D IR spectroscopy to follow the aggregation of amylin (Shim *et al.*, 2009). They studied six separate samples of amylin, each having a different amino acid labeled with $^{13}\text{C}^{18}\text{O}$. As the disordered monomers assemble into parallel β -sheets, the uncoupled isotope-labeled feature at 1595 cm^{-1} disappears and a coupled isotope-labeled feature grows in between 1570 and 1585 cm^{-1} . By comparing the rate at which the coupled peak developed for each labeled residue, they determined the order in which the residues are incorporated into the fiber β -sheets. More recently, Buchanan and coworkers performed an analogous study that focused on labeling residues within the disordered turn of amylin (Buchanan *et al.*, 2013). They discovered an early β -sheet intermediate on the fibril formation pathway. The intermediate contains a transient β -sheet that is broken to form the disordered turn in the final fiber structure. The discovery of the intermediate represents a significant advance in the understanding of amyloid formation, as it both provides a rationale for the initial lag phase observed in kinetics measurements before fibril β -sheets are observed and presents a target for future inhibitor studies and drug design.

Membrane proteins

Membrane proteins are involved in many basic cellular activities, such as ion transport and energy transduction (Engel and Gaub, 2008). Despite their prevalence in biology, our understanding of membrane proteins is limited by their hydrophobicity. 2D IR spectroscopy is capable of determining the structure of membrane proteins either solubilized by detergent or embedded in protein bilayer. Woys and coworkers used site-specific $^{13}\text{C}^{18}\text{O}$ labeling to measure the 2D IR lineshapes for 15 of the 18 residues of ovispirin, an antibiotic polypeptide that binds to the surfaces of membranes as an α -helix (Woys *et al.*, 2010). A regular oscillation was observed in inhomogeneous linewidth with a 15 cm^{-1} difference between the largest and smallest linewidths. The period of the oscillation matches the periodicity of an α -helix. As the inhomogeneous linewidth measures the structural disorder of the backbone and surrounding electrostatic environment, broader lineshapes indicate residues that lie closer to the membrane surface while narrower lineshapes indicate residues that lie within the hydrophobic membrane interior. Comparison with simulations allowed the orientation and depth of ovispirin within the membrane to be determined.

Beyond structural measurements, the application of 2D IR spectroscopy to ion channels, a special class of membrane proteins, has provided significant insight into their function.

Ion permeation through the K^+ ion channel

Potassium ion channels are responsible for setting the resting membrane potential of many cells and shaping the action potential of excitable cells such as neurons. Ion permeation through the channel is highly selective and determined by the selectivity filter, a narrow pore whose structure is highly conserved across all K^+ channels. The pore is lined by backbone carbonyls that coordinate the K^+ ions through the carbonyl oxygens. The series of carbonyls create four K^+ binding sites through which the ions must move sequentially. Kratochvil and coworkers used 2D IR spectroscopy to determine the mechanism for K^+ permeation through the KcsA ion channel (Kratochvil *et al.*, 2009). Previous experiments suggest that “knock-on” permeation, in which the pore is simultaneously occupied by two K^+ ions separated by water molecules, is the most likely mechanism for the transport of K^+ ions through the pore. However, these studies can not rule out alternative models such as “hard-knock” permeation, in which a pair of K^+ ions occupy adjacent sites in the pore with no intervening water molecules. The inherent picosecond time resolution of 2D IR spectroscopy allows it to provide a nearly instantaneous snapshot of the ion binding configurations in the pore. Using protein semisynthesis, two adjacent residues at the center of the pore were labeled with $^{13}\text{C}^{18}\text{O}$; these residues participate in the three of the four binding sites and their frequencies depend strongly on the occupancy of these sites. In combination with molecular dynamics simulations, the 2D IR data provides new experimental evidence that water and ions alternate through the pore during conduction and reveals a new conformation of the protein backbone that had only been observed in MD simulations.

Role of water in the influenza M2 channel

The M2 channel of the influenza virus transports protons across the viral envelope to acidify the interior, a process that is vital to the replication of the virus. Water is an integral part of the M2 proton channel as it is required for assisting proton conduction. The transmembrane domain of the M2 channel (M2TM) consists of a bundle of four identical α -helices that rearrange at acidic pH to open the channel and allow proton transport (Manor *et al.*, 2009). X-ray crystallography revealed a water cluster near the center of M2TM which is stabilized by the carbonyls of residues Gly34 and His37. Hochstrasser and coworkers used 2D IR spectroscopy to further explore the nature of the water within the channel (Ghosh *et al.*, 2011). Spectral diffusion measurements, in which the evolution of the 2D IR line shape is tracked as a function of the delay between pump and probe pulses, can be used to identify mobile water molecules. The rapid exchange of hydrogen bonds causes the vibrational frequencies of nearby amide groups to vary rapidly in time, leading to increasingly homogeneous peaks as the waiting time between pulses is increased. They specifically probed the water dynamics at Gly24 by incorporating $^{13}\text{C}^{18}\text{O}$ isotope label at the backbone carbonyl. At pH 8, the water cluster near Gly24 is found to be immobilized on the 10 ps timescale as no appreciable spectral diffusion was observed. In contrast, at pH 5.2 when the channel is open, spectral diffusion is observed on the timescale of 1.3 ps, indicating the presence of highly mobile water molecules (Fig. 11). The flow of water in the open state of the channel thus facilitates the diffusion of the proton through the channel. A subsequent study found that channel-blocking drugs increase the mobility of water molecules within the channel, even at non-acidic pH (Ghosh *et al.*, 2014). Thus, drug binding leads to an increase in the entropy of the channel water, increasing the thermodynamic favorability of the drug-binding process as a whole.

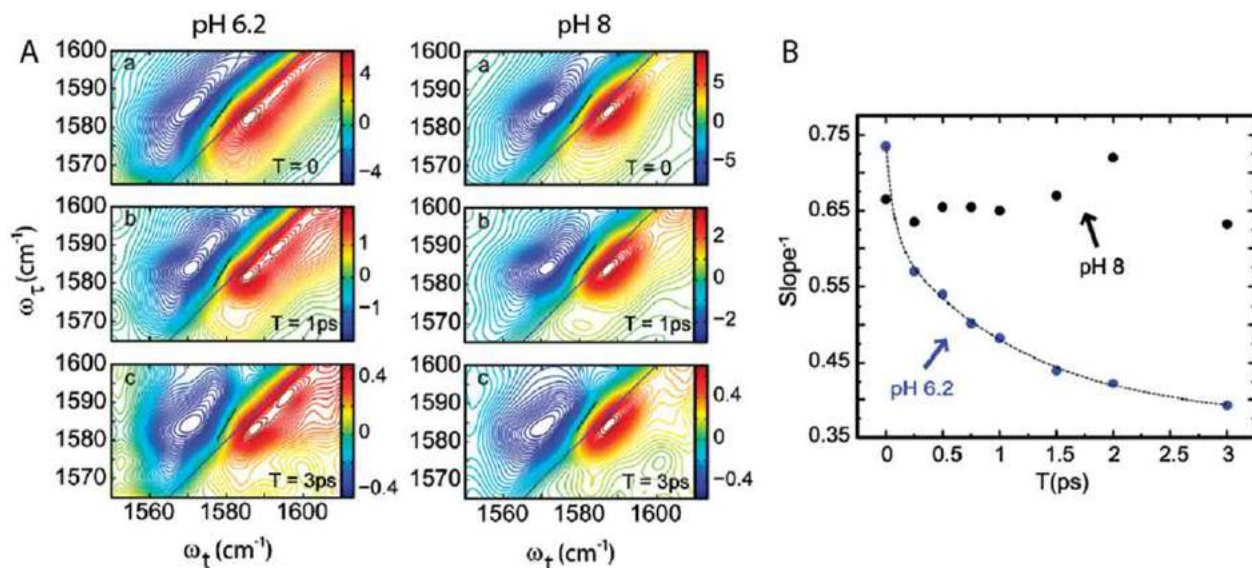


Fig. 11 Spectral diffusion measurements of the M2 proton channel. (A) 2D IR spectra of the isotope-labeled region of M2TM when Gly34 is labeled with $^{13}\text{C}^{18}\text{O}$. Spectra are collected at pH 6.2 (left) and pH 8.0 (right) at waiting times between 0 and 3 ps. Nodal slopes are indicated with black lines. (B) Plot of inverse nodal slope vs waiting time for M2TM. Adapted from Ghosh, A., Qiu, J., Degrad, W.F., Hochstrasser, R.M., 2011. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. *Proceedings of the National Academy of Sciences* 108 (15), 6115–6120.

Dynamics of Water

Although water is critical for many chemical and biological processes and therefore has been studied extensively using both experimental and theoretical approaches, much about water remains unclear. While repulsive forces typically dictate the structure of liquids, extensive hydrogen bonding causes water to exhibit many anomalous properties. Each water molecule can accept and donate two hydrogen bonds which rapidly interchange on the timescale of tens of femtoseconds to picoseconds (Roberts *et al.*, 2009). This rapidly evolving hydrogen bond network governs many aqueous processes including ion solvation and proton transport.

2D IR spectroscopy has provided a wealth of insight into the hydrogen bond switching dynamics of water (Loparo *et al.*, 2006). 2D IR anisotropy measurements of the OH stretching vibration of HOD were used to track the reorientation of water molecules as they undergo hydrogen bond exchange (Roberts *et al.*, 2009; Ramasesha *et al.*, 2011). The results suggest that the water hydrogen bond network rearranges through concerted motions that large angle molecular reorientations. Other 2D IR studies have explored the slowing of hydrogen bond dynamics as liquid water is cooled from ambient conditions to the metastable supercooled regime (Perakis and Hamm, 2011; Kraemer *et al.*, 2008). Beyond studies of pure water, 2D IR spectroscopy has been used to observe the exchange of hydrogen bonds between water molecules and ions (Moilanen *et al.*, 2009). These studies show that while water dynamics are slowed in salt solutions compared to pure water, they are only slowed by a factor of approximately 2. The influx of new insights into water structure and dynamics from 2D IR spectroscopy, as well as other ultrafast nonlinear optical techniques, has raised many questions about the suitability of existing water models and led theorists to develop new models for water to better explain experimental observations (Schmidt *et al.*, 2007; Ni and Skinner, 2014).

A recent 2D IR study by the Tokmakoff group focused on the proton-transport mechanism in water (Thamer *et al.*, 2015). Proton transport governs most biological redox processes as well as acid-base chemistry but little was known about how the water hydrogen-bonding network accommodates excess protons. The primary point of debate involved the dominant structure of the proton-water complex and how various species interconvert during proton transport. By exciting O-H stretching vibrations and detecting the response over a broad spectral window (1500–4000 cm⁻¹), they observed cross peaks at 3200 and 1760 cm⁻¹ which are characteristic of the water stretching and bending vibrations of the Zundel complex $\text{H}(\text{H}_2\text{O})_2^+$, a structure in which the proton is shared equally between two water molecules (Fig. 12(A)). The lifetime of the Zundel complex was determined to be at least 480 fs (Fig. 12(B)); this is long for the complex to be merely a transition state, which suggests it must play a key role in aqueous proton transfer.

Perspective

2D IR spectroscopy continues to rapidly progress with new implementations and applications. Several new directions have already emerged and could lead to major breakthroughs in the next 10 years.

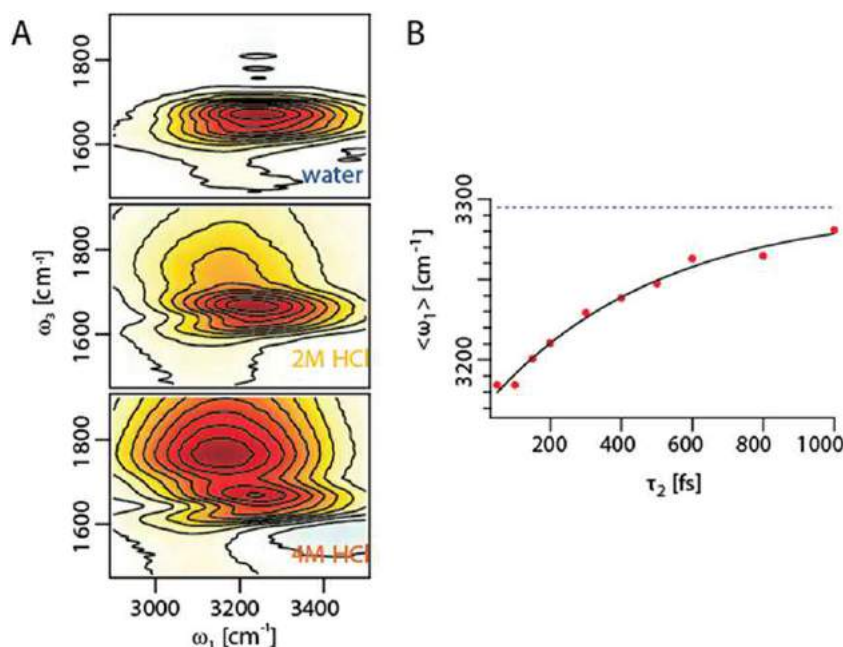


Fig. 12 2D IR crosspeaks reveal existence of Zundel complex in water. (A) Background subtracted 2D IR spectra for water, 2M HCl, and 4M HCl. A previously unobserved cross peak corresponding to the Zundel bend transition appears at $\omega_3 = 1760$ cm⁻¹ in the acid spectra and increases linearly in intensity with acid concentration. (B) As waiting time increases, the Zundel cross peak blue-shifts along the excitation axis, indicating that either proton transport or vibrational energy transfer is occurring between bulk water and the Zundel complex on the time scale of 480 fs. Adapted from Thamer, M., De Marco, L., Ramasesha, K., *et al.*, 2015. Ultrafast 2D IR spectroscopy of the excess proton in liquid water. *Science* 350 (6256), 78–82.

Surface- and Interface-Sensitive 2D Vibrational Spectroscopy

By combining 2D IR spectroscopy with surface- and interface-sensitive techniques, intrinsic surface- and interface-sensitive 2D vibrational spectroscopic techniques have been developed. These techniques have found applications in the study of catalytic surfaces (Xiong *et al.*, 2011; Wang *et al.*, 2015; Li *et al.*, 2016a,b), protein-functionalized surfaces (Laaser and Zanni, 2013; Ghosh *et al.*, 2015; Laaser *et al.*, 2014; Ho *et al.*, 2015), and air/water interfaces (Hsieh *et al.*, 2014; Piatkowski *et al.*, 2014; Livingstone *et al.*, 2016; Bredenbeck *et al.*, 2009, 2008; Zhang *et al.*, 2011a,b; Singh *et al.*, 2012, 2016; Inoue *et al.*, 2015; Nihonyanagi *et al.*, 2013) which are difficult to be study using bulk techniques. There are two main versions of surface- and interface-sensitive 2D vibrational spectroscopy. The Zanni, Bonn, Tahara, and Xiong groups implemented 2D vibrational sum frequency generation (2D SFG) spectroscopy, which combines the pulse sequences of 2D IR and SFG. The detected 2D SFG signal is a fourth-order nonlinear optical response; thus, it is intrinsically sensitive to surfaces and interfaces. An alternative approach is 2D attenuated total reflection (ATR) spectroscopy, developed by the Hamm group (Kraack *et al.*, 2014, 2015, 2016; Lotti *et al.*, 2016; Kraack and Hamm, 2016b). This technique applies the 2D IR pulse sequence at the surface of an ATR crystal, taking advantage of evanescent waves at the crystal surface to achieve interface- and surface-sensitivity. With improved signal-to-noise, these surface- and interface-sensitive 2D vibrational techniques are expected to unravel molecular conformations and dynamics at complex interfaces. An excellent review article has become available on this specific topic just recently (Kraack and Hamm, 2016a).

Broadband Light Sources

The Tokmakoff, Khalil and Peterson groups have worked to develop broadband IR laser pulses for 2D IR applications (Petersen and Tokmakoff, 2010; Calabrese *et al.*, 2012; Cheng *et al.*, 2012; Balasubramanian *et al.*, 2016). These broadband light sources are based on filamentation in gaseous media. When a 800 nm pulse and its harmonics are focused into the gaseous media, pulses that span from <400 to <5000 cm⁻¹ have been realized (Calabrese *et al.*, 2012). It is feasible to compress the pulses close to the transform limit using a deformable mirror pulse shaper (Balasubramanian *et al.*, 2016). The broadband IR pulses have sufficiently high pulse energies to be used as IR₃ in the 2D IR pulse sequence to obtain partial broadband 2D IR spectra. This approach has been used to study couplings between stretching and bending motions of water in Zundel complex form (Thamer *et al.*, 2015) and electronic vibrational coupling (Gaynor *et al.*, 2016). In the future, there is great interest in developing broadband IR pulses with μ J pulse energies to enable full broadband 2D IR spectroscopy.

2D IR Microscopy

As with many spectroscopic techniques, developing imaging capabilities to spatially visualize samples with 2D IR spectral signals is an emerging field. This development will allow spatial heterogeneity to be resolved in 2D IR spectra and also enrich microscope

imaging with molecular information that is not available from traditional absorption- or fluorescence-based optical microscopy. To the best of our knowledge, there are only two pioneering 2D IR microscopy reports (Baiz *et al.*, 2014; Ostrander *et al.*, 2016). Both used a pulse shaper based approach to collect 2D IR spectra. FPA detectors reduced data acquisition times and allowed 3.48 μm spatial resolution to be achieved (Ostrander *et al.*, 2016). Although still in its infancy, these proof-of-principle works demonstrate the feasibility of collecting microscope images with 2D IR spectral information and this field could blossom in the next 10 years.

2D Electronic-Vibrational Spectroscopy and 2D Vibrational-Electronic Spectroscopy

Another notable development related to 2D IR spectroscopy is 2D Electronic-Vibrational (EV) spectroscopy and 2D Vibrational-Electronic (VE) spectroscopy, which probe the coupling between electronic and vibrational excitations of molecules. The Fleming group has pioneered 2D EV spectroscopy, in which electronic states of molecules are excited and the subsequent vibrational motions are probed. They showed that this technique is useful in understanding the interplay between electronic states and vibrational manifolds responsible for nonradioactive electronic relaxation pathways (Oliver *et al.*, 2014). 2D EV spectroscopy has also revealed electronic energy transfer between segments in photosynthetic complexes (Lewis and Fleming, 2016; Lewis *et al.*, 2016). Theoretical frameworks were developed to quantify correlated electronic and vibrational dynamics and to correlate electronic state populations with structural information (Lewis *et al.*, 2015a,b). 2D VE spectroscopy, developed by the Khalil group, is complementary to 2D EV spectroscopy. 2D VE spectroscopy excites vibrational modes of molecules and probes electronic states. They used this technique to study organometallic charge transfer compounds and were able to distinguish coherent and incoherent vibrational energy transfer among coupled vibrational modes and charge transfer transitions (Courtney *et al.*, 2015a,b). These pioneering studies demonstrate that 2D VE spectroscopy can reveal how molecular vibrations modulate energy and charge transfer in molecular systems. Together, 2D EV and 2D VE spectroscopy extend the ability of 2D spectroscopic techniques from probing dynamics and couplings between the same type of molecular motions to probing interactions between molecular motions with different energy and time scales.

Acknowledgements

Lauren E. Buchanan and Wei Xiong thank all the researchers who contributed to the development of 2D IR spectroscopy in the past 18 years. Lauren E. Buchanan and Wei Xiong thank Prof. Martin T. Zanni for being a fantastic mentor and for introducing them to 2D IR spectroscopy. Wei Xiong acknowledges his group members and the financial support from UC San Diego. Lauren E. Buchanan acknowledges her group members and financial support from Vanderbilt University.

See also: Tutorial on Multidimensional Coherent Spectroscopy

References

- Anderson, N.A., Lian, T., 2005. Ultrafast electron transfer at the molecule–semiconductor nanoparticle interface. *Annual Review of Physical Chemistry* 56, 491–519.
- Anna, J.M., Kubarych, K.J., 2010. Watching solvent friction impede ultrafast barrier crossings: A direct test of Kramers theory. *Journal of Chemical Physics* 133, 174506.
- Anna, J.M., Ross, M.R., Kubarych, K.J., 2009. Dissecting enthalpic and entropic barriers to ultrafast equilibrium isomerization of a flexible molecule using 2DIR chemical exchange spectroscopy. *Journal of Physical Chemistry A* 113, 6544–6547.
- Asbury, J.B., Steinell, T., Stromberg, C., *et al.*, 2004. Water dynamics: Vibrational echo correlation spectroscopy and comparison to molecular dynamics simulations. *Journal of Physical Chemistry A* 108, 1107–1119.
- Asplund, M.C., Zanni, M.T., Hochstrasser, R.M., 2000. Two-dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. *Proceedings of the National Academy of Sciences* 97 (15), 8219–8224.
- Auer, B., Kumar, R., Schmidt, J.R., Skinner, J.L., 2007. Hydrogen bonding and Raman, IR, and 2D-IR spectroscopy of dilute HOD in liquid D₂O. *Proceedings of the National Academy of Sciences* 104 (36), 14215–14220.
- Baiz, C.R., Schach, D., Tokmakoff, A., 2014. Ultrafast 2D IR microscopy. *Optics Express* 22 (15), 18724.
- Bakulin, A.A., Liang, C., Jansen, T.L.C., *et al.*, 2009. Hydrophobic solvation: A 2D IR spectroscopic inquest. *Accounts of Chemical Research* 42 (9), 1229–1238.
- Balasubramanian, M., Courtney, T.L., Gaynor, J.D., Khalil, M., 2016. Compression of tunable broadband mid-IR pulses with a deformable mirror pulse shaper. *Journal of the Optical Society of America B* 33 (10), 2033–2037.
- Barbour, L.W., Hegadorn, M., Asbury, J.B., 2006. Microscopic inhomogeneity and ultrafast orientational motion in an organic photovoltaic bulk heterojunction thin film studied with 2D IR vibrational spectroscopy. *Journal of Physical Chemistry B* 110, 24281–24286.
- Bredenbeck, J., Ghosh, A., Nienhuys, H., Bonn, M., 2009. Interface-specific ultrafast two-dimensional vibrational spectroscopy. *Accounts of Chemical Research* 42 (9), 1332–1342.
- Bredenbeck, J., Ghosh, A., Smits, M., Bonn, M., 2008. Ultrafast two dimensional-infrared spectroscopy of a molecular monolayer. *Journal of the American Chemical Society* 130, 2152–2153.
- Bredenbeck, J., Helbing, J., Hamm, P., 2004. Labeling vibrations by light: Ultrafast transient 2D-IR spectroscopy tracks vibrational modes during photoinduced charge transfer. *Journal of the American Chemical Society* 126, 990–991. Available at: <http://dx.doi.org/10.1063/1.4932983>.
- Brinzer, T., Berquist, E.J., Ren, Z., *et al.*, 2015. Ultrafast vibrational spectroscopy (2D-IR) of CO₂ in ionic liquids: Carbon capture from carbon dioxide's point of view. *Journal of Chemical Physics* 142, 212425.

- Buchanan, L.E., Carr, J.K., Fluiitt, A.M., *et al.*, 2014. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. *Proceedings of the National Academy of Sciences* 111 (16), 5796–5801.
- Buchanan, L.E., Dunkelberger, E.B., Tran, H.Q., *et al.*, 2013. Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. *Proceedings of the National Academy of Sciences* 110 (48), 19285–19290.
- Buchanan, L.E., Dunkelberger, E.B., Zanni, M.T., 2012. Examining amyloid structure and kinetics with 1D and 2D infrared spectroscopy and isotope labeling. In: Fabian, H., Naumann, D. (Eds.), *Protein Folding and Misfolding: Shining Light by Infrared Spectroscopy*, vol. 1. Springer, pp. 217–237.
- Cahoon, J.F., Sawyer, K.R., Schlegel, J.P., Harris, C.B., 2008. Determining transition-state geometries in liquids using 2D-IR. *Science* 319, 1820–1823.
- Calabrese, C., Stingel, A.M., Shen, L., Petersen, P.B., 2012. Ultrafast continuum mid-infrared spectroscopy: Probing the entire vibrational spectrum in a single laser shot with femtosecond time resolution. *Optics Letters* 37 (12), 2265–2267.
- Cerullo, G., De, S.S., 2003. Ultrafast optical parametric amplifiers. *Review of Scientific Instruments* 74 (1), 1–18.
- Cervetto, V., Helbing, J., Bredenbeck, J., Hamm, P., 2004. Double-resonance versus pulsed Fourier transform two-dimensional infrared spectroscopy: An experimental and theoretical comparison. *Journal of Chemical Physics* 121 (12), 5935–5942.
- Cheng, M., Reynolds, A., Widgren, H., Khalil, M., 2012. Generation of tunable octave-spanning mid-infrared pulses by filamentation in gas media. *Optics Letters* 37 (11), 1787–1789.
- Chiti, F., Dobson, C.M., 2006. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry* 75, 333–366.
- Cho, M., 2009. *Two-Dimensional Optical Spectroscopy*. CRC Press.
- Courtney, T.L., Fox, Z.W., Estergreen, L., Khalil, M., 2015. Measuring coherently coupled intramolecular vibrational and charge-transfer dynamics with two-dimensional vibrational-electronic spectroscopy. *Journal of Physical Chemistry Letters* 6, 1286–1292.
- Courtney, T.L., Fox, Z.W., Slenkamp, K.M., Khalil, M., 2015b. Two-dimensional vibrational-electronic spectroscopy T. *Journal of Chemical Physics* 143, 154201. Available at: <http://dx.doi.org/10.1063/1.4932983>.
- Crabtree, R.H., 2004. Organometallic alkane CH activation. *Journal of Organometallic Chemistry* 689, 4083–4091.
- Cui, Y., Fulfer, K.D., Ma, J., *et al.*, 2016. Solvation dynamics of an ionic probe in chloride-based deep eutectic solvents. *Physical Chemistry Chemical Physics* 18, 31471–31479.
- Demirdöven, N., Khalil, M., Tokmakoff, A., 2002. Correlated vibrational dynamics revealed by two-dimensional infrared spectroscopy. *Physical Review Letters* 89 (23), 237401.
- Donaldson, P.M., Hamm, P., 2013. Gold nanoparticle capping layers: Structure, dynamics, and surface enhancement measured using 2D-IR spectroscopy. *Angewandte Chemie – International Edition* 52 (2), 634–638.
- Dutta, S., Ren, Z., Brinzer, T., Garrett-Roe, S., 2015. Two-dimensional ultrafast vibrational spectroscopy of azides in ionic liquids reveals solute-specific solvation. *Physical Chemistry Chemical Physics* 17, 26575–26579.
- Eaves, J.D., Loparo, J.J., Fecko, C.J., *et al.*, 2005. Hydrogen bonds in liquid water are broken only fleetingly. *Proceedings of the National Academy of Sciences* 102 (37), 13019–13022.
- Engel, A., Gaub, H.E., 2008. Structure and mechanics of membrane proteins. *Annual Review of Biochemistry* 77, 127–148.
- Fayer, M.D., 2009. Dynamics of liquids, molecules, and proteins measured with ultrafast 2D IR vibrational echo chemical exchange spectroscopy. *Annual Review of Physical Chemistry* 60, 21–38.
- Fayer, M.D., 2013. *Ultrafast Infrared Vibrational Spectroscopy*. CRC Press.
- Fulfer, K.D., Kuroda, D.G., 2016. Solvation structure and dynamics of the lithium ion in organic carbonate-based electrolytes: A time-dependent infrared spectroscopy study. *Journal of Physical Chemistry C* 120, 24011–24022.
- Fulmer, E.C., Mukherjee, P., Krummel, A.T., Zanni, M.T., 2004. A pulse sequence for directly measuring the anharmonicities of coupled vibrations: Two-quantum two-dimensional infrared spectroscopy. *Journal of Chemical Physics* 120 (7), 8067–8078.
- Ganim, Z., Chung, H.O.I.S., Smith, A.W., *et al.*, 2008. Amide I two-dimensional infrared spectroscopy of proteins. *Accounts of Chemical Research* 41 (3), 432–441.
- Gaynor, J.D., Courtney, T.L., Balasubramanian, M., Khalil, M., 2016. Fourier transform two-dimensional electronic-vibrational spectroscopy using an octave-spanning mid-IR probe. *Optics Letters* 41 (12), 2895–2898.
- Ghosh, A., Ho, J., Serrano, A.L., *et al.*, 2015. Two-dimensional sum-frequency generation (2D SFG) spectroscopy: Summary of principles and its application to amyloid fiber monolayers. *Faraday Discussions* 177, 493–505.
- Ghosh, A., Qiu, J., Degrado, W.F., Hochstrasser, R.M., 2011. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. *Proceedings of the National Academy of Sciences* 108 (15), 6115–6120.
- Ghosh, A., Wang, J., Moroz, Y.S., *et al.*, 2014. 2D IR spectroscopy reveals the role of water in the binding of channel-blocking drugs to the influenza M2 channel. *Journal of Chemical Physics* 140, 235105.
- Goldschmidt, L., Teng, P.K., Riek, R., Eisenberg, D., 2010. Identifying the amyloids, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences* 107 (8), 3487–3492.
- Golonzka, O., Khalil, M., Demirdöven, N., Tokmakoff, A., 2015. Vibrational anharmonicities revealed by coherent two-dimensional infrared spectroscopy. *Physical Review Letters* 115 (10), 2154–2157.
- Hamm, P., 2006. Three-dimensional-IR spectroscopy: Beyond the two-point frequency fluctuation correlation function. *Journal of Chemical Physics* 124, 124506.
- Hamm, P., Helbing, J., Bredenbeck, J., 2008. Two-dimensional infrared spectroscopy of photoswitchable peptides. *Annual Review of Physical Chemistry* 59, 291–317.
- Hamm, P., Lim, M., DeGrado, W.F., Hochstrasser, R.M., 2000. Pump/probe self heterodyned 2D spectroscopy of vibrational transitions of a small globular peptide. *Journal of Chemical Physics* 112 (4), 1907–1916.
- Hamm, P., Lim, M., Hochstrasser, R.M., 1998. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *Journal of Physical Chemistry B* 102, 6123–6138.
- Hamm, P., Zanni, M.T., 2011. *Concepts and Methods of 2D Infrared Spectroscopy*. New York, NY: Cambridge University Press.
- Hasegawa, T., Tanimura, Y., 2008. Nonequilibrium molecular dynamics simulations with a backward-forward trajectories sampling for multidimensional infrared spectroscopy of molecular vibrational modes. *Journal of Chemical Physics* 128, 64511.
- Hochstrasser, R.M., 2001. Two-dimensional IR-spectroscopy: Polarization anisotropy effects. *Chemical Physics* 266, 273–284.
- Ho, J., Sko, D.R., Ghosh, A., Zanni, M.T., 2015. Structural characterization of single-stranded DNA monolayers using two-dimensional sum frequency generation spectroscopy. *Journal of Physical Chemistry B* 119, 105868–105896.
- Hsieh, C., Okuno, M., Hunger, J., *et al.*, 2014. Aqueous heterogeneity at the air/water interface revealed by 2D-HD-SFG spectroscopy. *Angewandte Chemie International Edition* 53, 8146–8149.
- Huber, C.J., Egger, S.M., Spector, I.C., *et al.*, 2015. 2D-IR spectroscopy of porous silica nanoparticles: Measuring the distance sensitivity of spectral diffusion. *Journal of Physical Chemistry C* 119, 25135–25144.
- Huber, C.J., Massari, A.M., 2014. Characterizing solvent dynamics in nanoscopic silica sol – gel glass pores by 2D-IR spectroscopy of an intrinsic vibrational probe. *Journal of Physical Chemistry C* 118, 25567–25578.
- Hunt, N.T., 2009. 2D-IR spectroscopy: Ultrafast insights into biomolecule structure and function. *Chemical Society Reviews* 38 (7), 1837–1848.
- Hybl, J.D., Yu, A., Farrow, D.A., Jonas, D.M., 2002. Polar solvation dynamics in the femtosecond evolution of two-dimensional fourier transform spectra. *Journal of Physical Chemistry A* 106, 7651–7654.

- Inoue, K., Nihonyanagi, S., Singh, P.C., *et al.*, 2015. 2D heterodyne-detected sum frequency generation study on the ultrafast vibrational dynamics of H₂O and HOD water at charged interfaces. *Journal of Chemical Physics* 142, 212431.
- Ishizaki, A., Tanimura, Y., 2006. Modeling vibrational dephasing and energy relaxation of intramolecular anharmonic modes for multidimensional infrared spectroscopies. *Journal of Chemical Physics* 125, 84501.
- Ishizaki, A., Tanimura, Y., 2007. Dynamics of a multimode system coupled to multiple heat baths probed by two-dimensional infrared spectroscopy. *Journal of Physical Chemistry A* 111, 9269–9276.
- Jones, B.H., Huber, C.J., Massari, A.M., 2011. Solvation dynamics of Vaska's complex by 2D-IR spectroscopy. *Journal of Physical Chemistry C* 115, 24813–24822.
- Kajava, A.V., Baxa, U., Steven, A.C., 2016. B-arcades: Recurring motifs in naturally occurring and disease-related amyloid fibrils. *The FASEB Journal* 24 (5), 1311–1319.
- Kar, K., Hoop, C.L., Drombosky, K.W., *et al.*, 2013. β -Hairpin-mediated nucleation of polyglutamine amyloid formation. *Journal of Molecular Biology* 425 (7), 1183–1197.
- Khalil, M., Demirdo, N., Tokmakoff, A., 2003a. Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra. *Physical Review Letters* 90 (4), 47401.
- Khalil, M., Demirdoven, N., Tokmakoff, A., 2003b. Coherent 2D IR spectroscopy: Molecular structure and dynamics in solution. *Journal of Physical Chemistry A* 107, 5258–5279.
- Kim, Y.S., Hochstrasser, R.M., 2005. Chemical exchange 2D IR of hydrogen-bond making and breaking. *Proceedings of the National Academy of Sciences* 102 (32), 11185–11190.
- Kim, Y.S., Hochstrasser, R.M., 2009. Applications of 2D IR spectroscopy to peptides, proteins, and hydrogen-bond dynamics. *Journal of Physical Chemistry B* 113 (24), 8231–8251.
- Kim, Y.S., Liu, L., Axelsen, P.H., Hochstrasser, R.M., 2008. Two-dimensional infrared spectra of isotopically diluted amyloid fibrils from A β 40. *Proceedings of the National Academy of Sciences of the United States of America* 105, 7720–7725.
- Kim, Y.S., Liu, L., Axelsen, P.H., Hochstrasser, R.M., 2009. 2D IR provides evidence for mobile water molecules in beta-amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America* 106 (42), 17751–17756.
- Kraack, J.P., Hamm, P., 2016a. Surface-sensitive and surface-specific ultrafast two-dimensional vibrational spectroscopy. *Chemical Reviews* 117 (16), 10623–10664. doi:10.1021/acs.chemrev.6b00437.
- Kraack, J.P., Hamm, P., 2016b. Vibrational ladder-climbing in surface-enhanced, ultrafast infrared spectroscopy. *Physical Chemistry Chemical Physics* 18, 16088–16093.
- Kraack, J.P., Kaech, A., Hamm, P., 2016. Surface enhancement in ultrafast 2D ATR IR spectroscopy at the metal–liquid interface. *Journal of Physical Chemistry C* 120, 3350–3359.
- Kraack, J.P., Lotti, D., Hamm, P., 2014. Ultrafast, multidimensional attenuated total reflectance spectroscopy of adsorbates at metal surfaces. *Journal of Physical Chemistry Letters* 5, 2325–2329.
- Kraack, J.P., Lotti, D., Hamm, P., 2015. 2D attenuated total reflectance infrared spectroscopy reveals ultrafast vibrational dynamics of organic monolayers at metal–liquid interfaces. *Journal of Chemical Physics* 142, 212413.
- Kraemer, D., Cowan, M.L., Paarmann, A., *et al.*, 2008. Temperature dependence of the two-dimensional infrared spectrum of liquid H₂O. *Proceedings of the National Academy of Sciences* 105 (2), 437–442.
- Kratochvil, H.T., Carr, J.K., Matulef, K., *et al.*, 2009. Instantaneous ion configurations in the K⁺ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* 353, 1040–1044.
- Kumar, B., Llorente, M., Froehlich, J., *et al.*, 2012. Photochemical and photoelectrochemical reduction of CO₂. *Annual Review of Physical Chemistry* 63, 541–569.
- Kurochkin, D.V., Naraharisetty, S.R.G., Rubtsov, I.V., 2007. A relaxation-assisted 2D IR spectroscopy method. *Proceedings of the National Academy of Sciences* 104 (36), 14209–14214.
- Kwac, K., Cho, M., 2003a. Molecular dynamics simulation study of N-methylacetamide in water. II. Two-dimensional infrared pump–probe spectra. *Journal of Chemical Physics* 119 (4), 2256–2263.
- Kwac, K., Cho, M., 2003b. Two-color pump–probe spectroscopies of two- and three-level systems: 2-Dimensional line shapes and solvation dynamics. *Journal of Physical Chemistry A* 107, 5903–5912.
- Kwac, K., Lee, H., Cho, M., 2004. Non-gaussian statistics of amide I mode frequency fluctuation of N-methylacetamide in methanol solution: Linear and nonlinear vibrational spectra. *Journal of Chemical Physics* 120 (3), 1477–1490.
- Kwon, Y., Park, S., 2015. Complexation dynamics of CH₃SCN and Li⁺ in acetonitrile studied by two-dimensional infrared spectroscopy. *Physical Chemistry Chemical Physics* 17, 24193–24200.
- Laaser, J.E., Sko, D.R., Ho, J., *et al.*, 2014. Two-dimensional sum-frequency generation reveals structure and dynamics of a surface-bound peptide. *Journal of the American Chemical Society* 136, 956–962.
- Laaser, J.E., Zanni, M.T., 2013. Extracting structural information from the polarization dependence of one- and two-dimensional sum frequency generation spectra. *Journal of Physical Chemistry A* 117, 5875–5890.
- Lewis, N.H.C., Dong, H., Oliver, T.A.A., Fleming, G.R., 2015a. A method for the direct measurement of electronic site populations in a molecular aggregate using two-dimensional electronic-vibrational spectroscopy. *Journal of Chemical Physics* 143, 124203.
- Lewis, N.H.C., Dong, H., Oliver, T.A.A., Fleming, G.R., 2015b. Measuring correlated electronic and vibrational spectral dynamics using line shapes in two-dimensional electronic-vibrational spectroscopy. *Journal of Chemical Physics* 142, 174202.
- Lewis, N.H.C., Fleming, G.R., 2016. Two-dimensional electronic-vibrational spectroscopy of chlorophyll a and b. *Journal of Physical Chemistry Letters* 7, 831–837.
- Lewis, N.H.C., Gruenke, N.L., Oliver, T.A.A., *et al.*, 2016. Observation of electronic excitation transfer through light. *Journal of Physical Chemistry Letters* 7, 4197–4206.
- Lin, Z., Rubtsov, I.V., 2012. Constant-speed vibrational signaling along polyethyleneglycol chain up to 60-Å distance. *Proceedings of the National Academy of Sciences* 109 (5), 1413–1418.
- Livingstone, R.A., Zhang, Z., Piatkowski, L., *et al.*, 2016. Water in contact with a cationic lipid exhibits bulklike vibrational dynamics. *Journal of Physical Chemistry B* 120, 10069–10078.
- Li, Y., Wang, J., Clark, M.L., *et al.*, 2016a. Characterizing interstate vibrational coherent dynamics of surface adsorbed catalysts by fourth-order 3D SFG spectroscopy. *Chemical Physics Letters* 650, 1–6.
- Li, Z., Wang, J., Li, Y., Xiong, W., 2016b. Solving the “magic angle” challenge in determining molecular orientation heterogeneity at interfaces. *Journal of Physical Chemistry C* 120 (36), 20239–20246.
- Loparo, J.J., Roberts, S.T., Tokmakoff, A., *et al.*, 2006. Multidimensional infrared spectroscopy of water. II. Hydrogen bond switching dynamics multidimensional infrared spectroscopy of water. *The Journal of Chemical Physics* 125, 194522.
- Lotti, D., Hamm, P., Kraack, J.P., 2016. Surface-sensitive spectro-electrochemistry using ultrafast 2D ATR IR spectroscopy. *Journal of Physical Chemistry C* 120, 2883–2892.
- Luca, S., Yau, W., Leapman, R., Tycko, R., 2007. Peptide conformation and supramolecular organization in amylin fibrils: Constraints from solid-state NMR. *Biochemistry* 46, 13505–13522.
- Luther, B.M., Tracy, K.M., Gerrity, M., *et al.*, 2016. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. *Optics Express* 24 (4), 10095–10100.
- Maekawa, H., De Poli, M., Toniolo, C., Ge, N.-H., 2009. Couplings between peptide linkages across a 3(10)-helical hydrogen bond revealed by two-dimensional infrared spectroscopy. *Journal of the American Chemical Society* 131 (6), 2042–2043.
- Manor, J., Mukherjee, P., Lin, Y., *et al.*, 2009. Article gating mechanism of the influenza A M2 channel revealed by 1D and 2D IR spectroscopies. *Structure/Folding and Design* 17 (2), 247–254. Available at: <http://dx.doi.org/10.1016/j.str.2008.12.015>.

- Middleton, C.T., Woys, A.M., Mukherjee, S.S., Zanni, M.T., 2010. Residue-specific structural kinetics of proteins through the union of isotope labeling, mid-IR pulse shaping, and coherent 2D IR spectroscopy. *Methods* 52 (1), 12–22.
- Moilanen, D.E., Wong, D., Rosenfeld, D.E., *et al.*, 2009. Ion–water hydrogen-bond switching observed with 2D IR vibrational echo chemical exchange spectroscopy. *Proceedings of the National Academy of Sciences* 106 (2), 375–380.
- Moran, S.D., Woys, A.M., Buchanan, L.E., *et al.*, 2012. Two-dimensional IR spectroscopy and segmental ^{13}C labeling reveals the domain structure of human γ D-crystallin amyloid fibrils. *Proceedings of the National Academy of Sciences* 109 (9), 3329–3334.
- Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. *Annual Review of Physical Chemistry* 51, 691–729.
- Mukherjee, P., Kass, I., Arkin, I.T., Zanni, M.T., 2005. Picosecond dynamics of a membrane protein revealed by 2D IR. *Proceedings of the National Academy of Sciences* 103 (10), 3528–3533.
- Nee, M.J., Baiz, C.R., Anna, J.M., *et al.*, 2008. Multilevel vibrational coherence transfer and wavepacket dynamics probed with multidimensional IR spectroscopy. *Journal of Chemical Physics* 129, 84503.
- Nee, M.J., McCanne, R., Kubarych, K.J., 2007. Two-dimensional infrared spectroscopy detected by chirped pulse upconversion. *Optics Letters* 32 (6), 713–715.
- Nihonyanagi, S., Mondal, J.A., Yamaguchi, S., Tahara, T., 2013. Structure and dynamics of interfacial water studied by vibrational sum-frequency generation. *Annual Review of Physical Chemistry* 64, 579–603.
- Nishida, J., Tamimi, A., Fei, H., *et al.*, 2014. Structural dynamics inside a functionalized metal–organic framework probed by ultrafast 2D IR spectroscopy. *Proceedings of the National Academy of Sciences* 111 (52), 18442–18447.
- Nishida, J., Yan, C., Fayer, M.D., 2014. Dynamics of molecular monolayers with different chain lengths in air and solvents probed by ultrafast 2D IR spectroscopy. *Journal of Physical Chemistry C* 118, 523–532.
- Nishida, J., Yan, C., Fayer, M.D., 2016. Orientational dynamics of a functionalized alkyl planar monolayer probed by polarization-selective angle-resolved infrared pump–probe spectroscopy. *Journal of the American Chemical Society* 138, 14057–14065.
- Ni, Y., Skinner, J.L., 2014. Ultrafast pump–probe and 2DIR anisotropy and temperature-dependent dynamics of liquid water within the E3B model. *Journal of Chemical Physics* 141, 24509.
- Oliver, T.A.A., Lewis, N.H.C., Graham, R., 2014. Correlating the motion of electrons and nuclei with two-dimensional electronic-vibrational spectroscopy. *Proceedings of the National Academy of Sciences* 111 (28), 10061–10066.
- Ostrander, J.S., Serrano, A.L., Ghosh, A., Zanni, M.T., 2016. Spatially resolved two-dimensional infrared spectroscopy via wide-field microscopy. *ACS Photonics* 3, 1315–1323.
- Oudenhoven, T.A., Joo, Y., Laaser, J.E., *et al.*, 2015. Dye aggregation identified by vibrational coupling using 2D IR spectroscopy. *Journal of Chemical Physics* 142, 212449.
- Paarmann, A., Hayashi, T., Mukamel, S., Miller, R.J.D., 2009. Nonlinear response of vibrational excitons: Simulating the two-dimensional infrared spectrum of liquid water. *Journal of Chemical Physics* 130, 204110.
- Park, K., Cho, M., 1998. Time- and frequency-resolved coherent two-dimensional IR spectroscopy: Its complementary relationship with the coherent two-dimensional Raman scattering spectroscopy. *Journal of Chemical Physics* 109 (24), 10559–10569.
- Perakis, F., Hamm, P., 2011. Two-dimensional infrared spectroscopy of supercooled water. *Journal of Physical Chemistry B* 115, 5289–5293.
- Petersen, P.B., Tokmakoff, A., 2010. Source for ultrafast continuum infrared and terahertz radiation. *Optics Letters* 35 (12), 1962–1964.
- Piatkowski, L., Zhang, Z., Backus, E.H.G., *et al.*, 2014. Extreme surface propensity of halide ions in water. *Nature Communications* 5, 4083. Available at: <http://dx.doi.org/10.1038/ncomms5083>.
- Ramasesha, K., Roberts, S.T., Nicodemus, R.A., *et al.*, 2011. Ultrafast 2D IR anisotropy of water reveals reorientation during hydrogen-bond switching. *Journal of Chemical Physics* 135, 54509.
- Ren, Z., Brinzer, T., Dutta, S., Garrett-Roe, S., 2015. Thiocyanate as a local probe of ultrafast structure and dynamics in imidazolium-based ionic liquids: Water-induced heterogeneity and cation-induced ion pairing. *Journal of Physical Chemistry B* 119, 4699–4712.
- Ren, Z., Ivanova, A.S., Couchot-Vore, D., Garrett-Roe, S., 2014. Ultrafast structure and dynamics in ionic liquids: 2D-IR spectroscopy probes the molecular origin of viscosity. *Journal of Physical Chemistry Letters* 5, 1541–1546.
- Roberts, S.T., Ramasesha, K., Tokmakoff, A., 2009. Structural rearrangements in water viewed through two-dimensional infrared spectroscopy. *Accounts of Chemical Research* 42 (9), 1239–1249.
- Rock, W., Li, Y., Pagano, P., Cheatum, C.M., 2013. 2D IR spectroscopy using four-wave mixing, pulse shaping, and IR upconversion: A quantitative comparison. *Journal of Physical Chemistry A* 117, 6073–6083.
- Rosenfeld, D.E., Gengeliczki, Z., Smith, B.J., *et al.*, 2011. Structural dynamics of a catalytic monolayer probed by ultrafast 2D IR vibrational echoes. *Science* 334 (6056), 634–639.
- Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2012. Dynamics of functionalized surface molecular monolayers studied with ultrafast infrared vibrational spectroscopy. *Journal of Physical Chemistry C* 116, 23428–23440.
- Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2013. Structural dynamics at monolayer–liquid interfaces probed by 2D IR spectroscopy. *Journal of Physical Chemistry C* 117, 1409–1420.
- Roy, S., Pshenichnikov, M.S., Jansen, T.L.C., 2011. Analysis of 2D CS spectra for systems with non-gaussian dynamics. *Journal of Physical Chemistry B* 115, 5431–5440.
- Rubtsova, N.I., Qasim, L.N., Kurnosov, A.A., *et al.*, 2015. Ballistic energy transport in oligomers. *Accounts of Chemical Research* 48, 2547–2555.
- Rubtsova, N.I., Rubtsov, I.V., 2013. Ballistic energy transport via perfluoroalkane linkers. *Chemical Physics* 422, 16–21.
- Rubtsova, N.I., Rubtsov, I.V., 2015. Vibrational energy transport in molecules studied by two-dimensional infrared spectroscopy. *Annual Review of Physical Chemistry* 66, 717–738.
- Sanda, F., Mukamel, S., 2006. Stochastic simulation of chemical exchange in two dimensional infrared spectroscopy. *Journal of Chemical Physics* 125, 14507.
- Schmidt, J.R., Roberts, S.T., Loparo, J.J., *et al.*, 2007. Are water simulation models consistent with steady-state and ultrafast vibrational spectroscopy experiments? *Chemical Physics* 341, 143–157.
- Selig, O., Siffels, R., Rezus, Y.L.A., 2015. Ultrasensitive ultrafast vibrational spectroscopy employing the near field of gold nanoantennas. *Physical Review Letters* 114, 233004.
- Serrano, A.L., Ghosh, A., Ostrander, J.S., Zanni, M.T., 2015. Wide-field FTIR microscopy using mid-IR pulse shaping. *Optics Express* 23 (14), 17815.
- Shim, S.-H., Gupta, R., Ling, Y.L., *et al.*, 2009. Two-dimensional IR spectroscopy and isotope labeling defines the pathway of amyloid formation with residue-specific resolution. *Proceedings of the National Academy of Sciences* 106 (16), 6614–6619. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672516&tool=pmcentrez&rendertype=abstract>.
- Shim, S., Zanni, M.T., 2009. How to turn your pump–probe instrument into a multidimensional spectrometer: 2D IR and Vis spectroscopies via pulse shaping. *Physical Chemistry Chemical Physics* 11, 748–761.
- Singh, P.C., Inoue, K., Nihonyanagi, S., *et al.*, 2016. Femtosecond hydrogen bond dynamics of bulk-like and bound water at positively and negatively charged lipid interfaces revealed by 2D HD-VSFG spectroscopy. *Angewandte Chemie International Edition* 55, 10621–10625.
- Singh, P.C., Nihonyanagi, S., Yamaguchi, S., Tahara, T., 2012. Ultrafast vibrational dynamics of water at a charged interface revealed by two-dimensional heterodyne-detected vibrational sum frequency generation. *Journal of Chemical Physics* 137, 94706.
- Sueur, A.L., Le, Horness, R.E., Thielges, M.C., 2015. Applications of two-dimensional infrared spectroscopy. *Analyst* 140, 4336–4349.
- Thamer, M., De Marco, L., Ramasesha, K., *et al.*, 2015. Ultrafast 2D IR spectroscopy of the excess proton in liquid water. *Science* 350 (6256), 78–82.

- Tracy, K.M., Barich, M.V., Carver, C.L., *et al.*, 2016. High-throughput two-dimensional infrared (2D IR) spectroscopy achieved by interfacing microfluidic technology with a high repetition rate 2D IR spectrometer. *Journal of Physical Chemistry Letters* 7, 4865–4870.
- Wang, J., Clark, M.L., Li, Y., *et al.*, 2015. Short-range catalyst – surface interactions revealed by heterodyne two-dimensional sum frequency generation spectroscopy. *Journal of Physical Chemistry Letters* 6, 4204–4209.
- Wang, L., Middleton, C.T., Singh, S., *et al.*, 2011. 2DIR spectroscopy of human amylin fibrils reflects stable β -sheet structure. *Journal of the American Chemical Society* 133 (40), 16062–16071.
- Weston, M., Britton, A.J., Shea, J.N.O., 2011. Charge transfer dynamics of model charge transfer centers of a multicenter water splitting dye complex on rutile TiO₂ (110). *Journal of Chemical Physics* 134, 54705.
- Woutersen, S., Hamm, P., 2002. Nonlinear two-dimensional vibrational spectroscopy of peptides. *Journal of Physics: Condensed Matter* 14, R1035–R1062.
- Woys, A.M., Lin, Y.S., Reddy, A.S., *et al.*, 2010. 2D IR line shapes probe ovispirin peptide conformation and depth in lipid bilayers. *Journal of the American Chemical Society* 132 (8), 2832–2838.
- Wright, J.C., 2011. Multiresonant coherent multidimensional spectroscopy. *Annual Review of Physical Chemistry* 62, 209–230.
- Xiong, W., Laaser, J.E., Mehlenbacher, R.D., Zanni, M.T., 2011. Adding a dimension to the infrared spectra of interfaces using heterodyne detected 2D sum-frequency generation (HD 2D SFG) spectroscopy. *Proceedings of the National Academy of Sciences* 108 (52), 20902–20907.
- Xiong, W., Laaser, J.E., Paoprasert, P., *et al.*, 2009. Transient 2D IR spectroscopy of charge injection in dye-sensitized nanocrystalline thin films. *Journal of the American Chemical Society* 131, 18040–18041.
- Yan, C., Yuan, R., Nishida, J., Fayer, M.D., 2015. Structural influences on the fast Dynamics of alkylsiloxane monolayers on SiO₂ surfaces measured with 2D IR spectroscopy. *Journal of Physical Chemistry C* 119, 16811–16823.
- Yan, C., Yuan, R., Pfalzgraff, W.C., *et al.*, 2016. Unraveling the dynamics and structure of functionalized self-assembled monolayers on gold using 2D IR spectroscopy and MD simulations. *Proceedings of the National Academy of Sciences* 113 (18), 4929–4934.
- Zhang, Z., Piatkowski, L., Bakker, H.J., Bonn, M., 2011a. Communication: Interfacial water structure revealed by ultrafast two-dimensional surface vibrational spectroscopy. *Journal of Chemical Physics* 135, 21101.
- Zhang, Z., Piatkowski, L., Bakker, H.J., Bonn, M., 2011b. Ultrafast vibrational energy transfer at the water/air interface revealed by two-dimensional surface vibrational spectroscopy. *Nature Chemistry* 3 (11), 888–893. Available at: <http://dx.doi.org/10.1038/nchem.1158>.
- Zheng, J., Kwak, K., Asbury, J., *et al.*, 2005. Ultrafast dynamics of solute–solvent complexation observed at thermal equilibrium in real time. *Science* 309, 1338–1343.
- Zheng, J., Kwak, K., Fayer, M.D., 2007. Ultrafast 2D IR vibrational echo spectroscopy. *Accounts of Chemical Research* 40 (1), 75–83.
- Zhuang, W., Hayashi, T., Mukamel, S., 2009. Coherent multidimensional vibrational spectroscopy of biomolecules: Concepts, simulations, and challenges. *Angewandte Chemie International Edition* 48, 3750–3781.

Two-Dimensional Electronic Spectroscopy

Yin Song, University of Michigan, Ann Arbor, MI, United States

Xiaoqin Li, The University of Texas at Austin, Austin, TX, United States

Jennifer P Ogilvie, University of Michigan, Ann Arbor, MI, United States

© 2018 Elsevier Inc. All rights reserved.

Introduction

At present, the state of the art of optical spectroscopy of condensed phase systems aims to perform multidimensional spectral characterization with ultrafast time resolution. Since the first experimental demonstrations, multidimensional visible and infrared (IR) spectroscopy have been applied to address many fundamental questions in condensed phase dynamics. Several excellent textbooks (Hamm and Zanni, 2011; Cho, 2009) and reviews (Jonas, 2003; Ogilvie and Kubarych, 2009; Cho, 2008) discuss the principles and applications of multidimensional spectroscopy. Here we focus on studies in the visible and near-IR. In this frequency regime, 2D spectroscopy is commonly referred to as “2D electronic spectroscopy (2DES),” to be distinguished from its infrared (2DIR) and ultraviolet (2DUV) counterparts. We discuss experimental considerations and common implementations of 2DES, and summarize recent applications to a broad range of condensed phase systems.

2DES is an extension of one-dimensional (1D) pump–probe spectroscopy that overcomes some of its limitations by resolving the third order spectroscopic signal as a function of excitation frequency. This additional frequency dimension can separate homogeneous and inhomogeneous broadening, reveal electronic coupling and expose elusive states that can be hidden in systems with rich excited-state structure and dynamics. In pump–probe spectroscopy, the electric field of the pump pulse interacts (via the dipole operator) with the molecules twice to promote the molecules from the ground state to the excited state. After a variable waiting time (T), the electric field of a probe pulse interacts with the system of interest to generate the third-order polarization, which radiates the signal field. The probe pulse also acts as a reference pulse, referred to as a ‘local oscillator’ to interfere with the signal field, enabling heterodyne-detection (here the term heterodyne-detection refers to the detection of the interference between the signal field with a “local oscillator” field with the same spectral content. This process enables signal amplification and detection of the complex signal field (spectrally-resolved phase and amplitude of the signal. The pump–probe signal ($S(T, \omega)$) is often presented as a function of waiting time and detection frequency/wavelength to reveal the time evolution of the excited-state populations. Since the pump pulse can only be either a spectrally narrow (temporally long), or a spectrally broad (temporally short) pulse, there is a trade-off between time and frequency resolution with respect to the excitation axis. Thus, pump–probe spectroscopy is limited in its ability to uncover dynamic correlations between direct photoexcited states and probing states in complex systems with spectrally-overlapping transitions and dynamics occurring on an ultrafast time scale.

Fourier transform 2D spectroscopy differs from pump–probe spectroscopy in that it utilizes two ultrashort pulses with a variable time delay (coherence time, τ) to excite the sample (see Fig. 1(a)). The first field-matter interaction creates a coherence (a superposition of the ground and excited states) while the second interaction promotes the molecules to a population or another coherence. After a waiting time (T), the third-order polarization is generated in the sample via the third field-matter interaction and the radiated signal is heterodyne-detected with a fourth ‘local oscillator’ pulse. For each waiting time, the coherence time is scanned from $-\tau$ to $+\tau$ to obtain a 2D map ($S(\tau, T, \omega_3)$), where ω_3 is the detection frequency. Fourier transformation of the signal along the coherence time axis gives the signal in the frequency domain $S(\omega_1, T, \omega_3)$, (where ω_1 is the excitation axis). This 2D dynamic map directly reveals the correlation of the photoexcited and probing states as a function of waiting time T . As the excitation information is recorded by using the time-domain technique, the excitation frequency resolution is determined by the length of the coherence time scan rather than the excitation bandwidth. In other words, 2DES circumvents the time-frequency resolution trade-off in pump–probe spectroscopy.

Challenges

A key challenge in Fourier-transform 2DES is creating and delivering the appropriate pulse sequence with variable, accurate and phase-stable relative time delays. Time delays must be recorded with high accuracy if they are to yield accurate Fourier-transformed frequencies. Phase stability, which to first order is equivalent to a timing jitter between the pulses, should be as high as possible to enable clean separation of the complex signal components with a high signal-to-noise ratio. Typically one requires an interferometric precision of $\sim \lambda/100$, corresponding to timing errors of ~ 0.017 fs at 500 nm. As a consequence, the challenges of implementing 2DES increase for shorter wavelengths, where small mechanical vibrations or air current fluctuations lead to optical pathlength variations. Another challenge is that the 2D signal is relatively weak and must be distinguished from the incident pulse as well as from pump–probe signals, free-induction decays and transient-grating signals. Isolating the 2D signal from the background and noise (e.g. the scattering) can be achieved by using phase-matching and/or phase-cycling. In addition, to maximize the information available from a 2D measurement, both “rephasing” and “nonrephasing” complex signal fields must be recorded. To

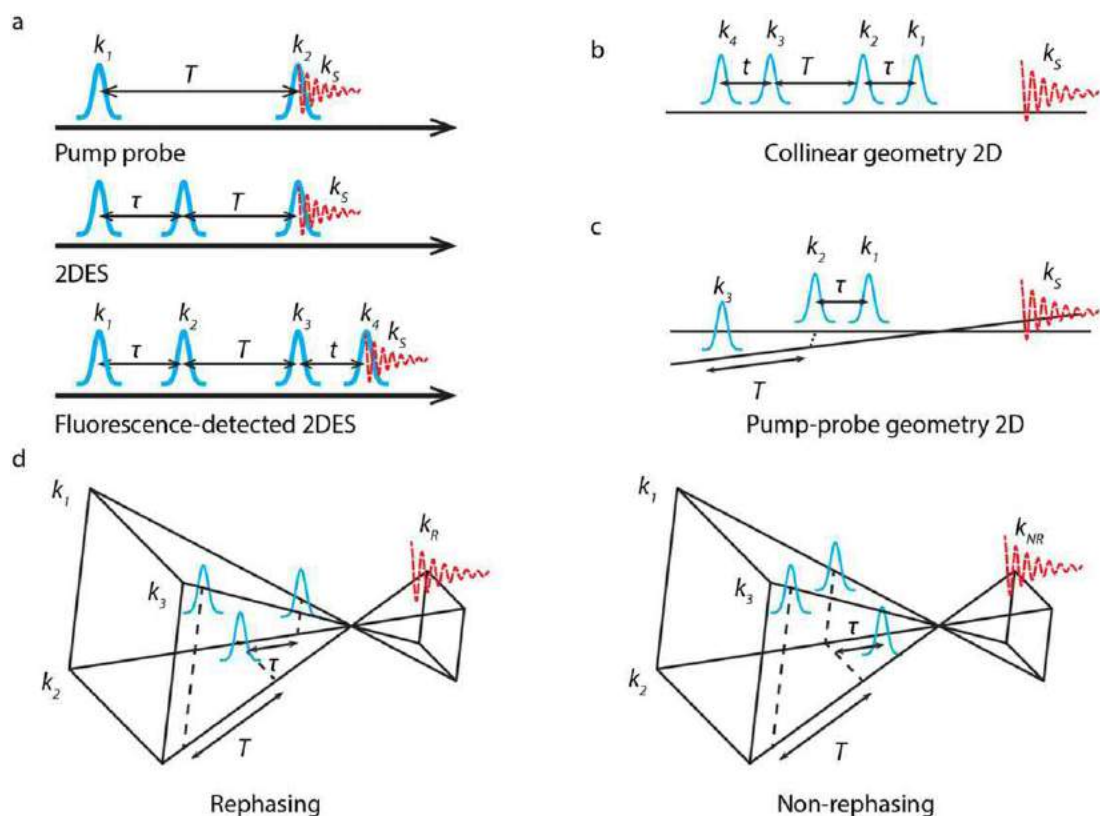


Fig. 1 (a) Pulse sequence in pump-probe, 2DES and fluorescence-detected 2DES; (b)–(d) cartoon illustrations of collinear geometry 2DES (fluorescence-detected), pump-probe geometry 2DES, BOXCARS rephasing and non-rephasing 2DES.

resolve the complex signal fields, the majority of 2DES setups employ spectral interferometry, in which interference between the signal and a reference pulse is measured in the frequency domain. A number of different experimental implementations of 2D spectroscopy have met the many challenges of 2D spectroscopy in different ways (recently reviewed by Fuller and Ogilvie (2015)) as discussed below.

Experimental Implementations

The first multidimensional optical spectroscopy measurements were pioneered independently by Joffe and Jonas. In 1996 the Joffe group, who developed the spectral interferometry method, used 2D spectroscopy to record the second-order response of a nonlinear crystal. The first third-order 2DES setup was reported by the Jonas group in 1998. In their implementation, two excitation pulses, the probe pulse and the signal were arranged in the BOXCARS geometry and the signal was emitted in a background-free direction. The coherence time was scanned using a translation stage that was calibrated by spectral interferometer, with the time zero determined by finding the symmetry-point of a pump-probe signal. The complex signal was recorded via spectral interferometry. A tracer pulse with pre-determined phase (by frequency-resolved optical gating (FROG)) was sent along the signal path and interferometry between the tracer and reference pulses was used to determine the signal phase. The phase stability of the interferometer-based setup was $\sim \lambda/27$ at 800 nm for 20 min (i.e. 0.1 fs rms drift of time).

Following these pioneering works, a number of approaches to 2DES were developed, aiming to improve phase stability, signal-to-noise ratio and ease of implementation. Three geometries, shown in Fig. 1(b)–(d), are now commonly used to perform 2DES: the collinear geometry, the pump-probe geometry and the BOXCARS geometry. Below we describe commonly-used implementations of 2DES and how they overcome the various implementation challenges. Fig. 2 shows the experimental setup for several typical 2DES implementations. The pros and cons of various approaches to 2DES are given in Table 1.

The BOXCARS Geometry

The BOXCARS geometry was the first to be implemented by the 2D community and remains the most common geometry. Its key advantage is that the signal emits in a background-free direction (spatially separated from the three pulses that generate the 2D signal). In 2DES, two different phase-matched signals are typically recorded and combined to obtain the “absorptive” 2DES

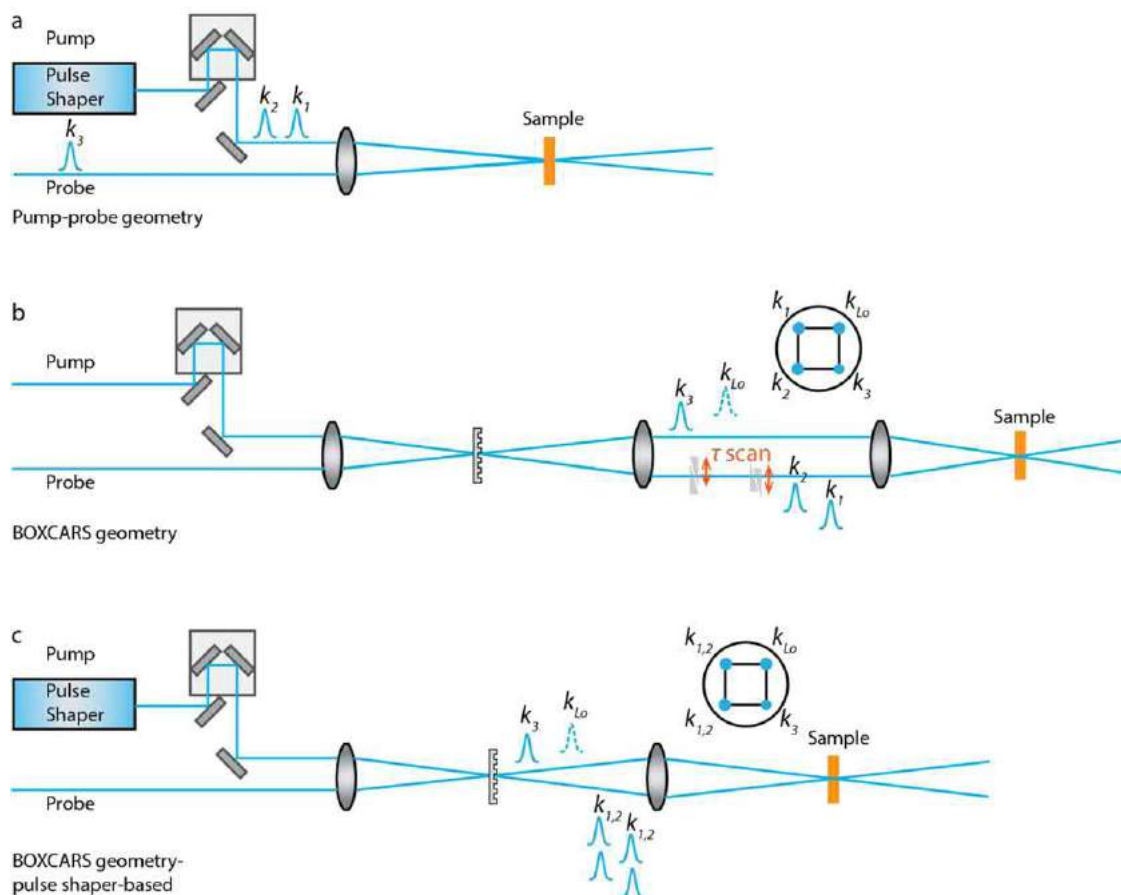


Fig. 2 Experimental implementations of several different 2DES setups. (a) Pulse-shaper-based pump-probe geometry 2DES; (b) BOXCARS geometry 2DES with refractive delays (τ), employing a diffractive optic (D.O.) beam splitter; and (c) pulse-shaper based BOXCARS geometry 2DES, employing a D.O. beam splitter.

spectrum, free from broadening refractive contributions. These are the “rephasing” and “nonrephasing” signals, emitted in the $\mathbf{k}_2 - \mathbf{k}_1 + \mathbf{k}_3$ and $\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$ directions, respectively. A related signal, with the phase-matching direction $\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$, yields information about double-quantum coherences. For mechanical delay lines, the rephasing and nonrephasing 2D signals are collected by scanning coherence time τ over positive and negative delays, respectively. As all four laser beams typically experience different optical paths in the BOXCARS geometry, the phase of the complex 2D signal is undetermined. The method of “phasing” the 2D spectrum will be discussed in more detail below.

Several methods have been used to generate the pulse pairs for 2DES in the BOXCARS geometry. The first 2DES setup used conventional beam splitters. Beam splitters pose three main problems for broadband pulses: (1) the beams may not be perfectly phase-matched at the focus (sample position); (2) the signal propagation direction may be slightly different from that of the local oscillator, reducing signal and resulting in spectral distortions; and (3) pulse-front tilt reduces the time resolution of the experiment. By constructing robust interferometers with carefully mounted optics and balancing the glass thickness in the pulse pairs, the phase differences and fluctuations can be minimized. To avoid the problem of the pulse-front tilt and imperfect phase matching, diffractive optic (D.O.) beam splitters were pioneered by the Miller and Nelson groups to generate pulse pairs, initially for transient-grating spectroscopy. Depending on the type of the light source and the applications, some groups have used a single 2D D.O. or 2 perpendicularly-mounted 1D D.O. to split a single beam into four beams at the corners of a square; other groups use two light sources as pump and probe beam, crossing them at a 1D D.O. to split each beam into two. The use of a single laser source gives better phase-matching while the use of two light sources with a 1D D.O. enables two-color 2D measurements. The use of D.O.s provides easy alignment of the signal and local oscillator for heterodyne detection. The D.O. based setup has the limitation of the spectral bandwidth since the diffracted beams are overlapped when the bandwidth reaches an octave of the wavelength.

Another method to generate the pulse pairs is to use a spatial filter with four holes at corners of a square. This method was employed first by the Nelson group and later by the Piefier group to generate four pulses from one original beam. Nelson and coworkers demonstrated that there is considerable power loss of the beam by using a spatial filter and that scattering from the edges of the holes can increase background signals.

Table 1 Summary of various implementations of 2DES, pros and cons and bandwidth limitations.

<i>Method</i>	<i>Pros</i>	<i>Cons</i>	<i>Bandwidth limitations</i>
Standard interferometer	Broadband and heterodyne detection	Motorized delays can cause time zero uncertainties and timing errors Requires either active stabilization to achieve linearly spaced time delays or high-precision time-delay measurements Rephasing and nonrephasing spectra are not acquired simultaneously Requires phasing to obtain absorptive spectra	Limited only by optical components (beam splitters, mirrors)
Diffractive optics	High passive phase stability, ease of alignment and heterodyne detection	Requires phasing to obtain absorptive spectra Rephasing and nonrephasing spectra are not acquired simultaneously	Limited by overlapping octaves and optical components (grating efficiency, mirrors, etc.)
Pulse shaping	Phase cycling to remove scatter and separate signals Automatic absorptive spectra in pump-probe geometry No uncertainty in interpulse pump delay Phase locking allows for spectra to be collected in fewer time steps.	Polarization control may be difficult, depending on pulse shaper and geometry May require phasing to obtain absorptive spectra, depending on geometry Pulse shapers can be lossy	Likely limited by pulse shaper Often limited by coatings for spatial-light-modulator shapers
Fluorescence-detected two-dimensional Fourier transform via phase modulation	Insensitive to mechanical/phase instabilities 1/f noise suppressed by lock-in detection No need for phasing data Single element detector, facilitating use with high repetition rate laser sources.	Requires the construction of reference signals Requires an additional motorized time delay to be scanned	Likely limited by acousto-optic modulators
Single shot	Reduced acquisition time, insensitive to laser noise	Requires separate recording of rephasing and nonrephasing signals Requires phasing to obtain absorptive spectra Laser spatial mode quality requirements are more stringent Requires a 2D camera to record the spatial dimension	Limited only by optical components (beam splitters, mirrors)
All reflective approaches	Broadband	Requires phasing to obtain absorptive spectra, depending on geometry May have reduced phase stability, depending on exact approach used Rephasing and nonrephasing spectra are not acquired simultaneously Motorized delays can cause time zero uncertainties and timing errors	Limited only by optical components (beam splitters, mirrors)
Birefringent interferometer (TWINS)	Collection in a partially rotating frame allows for spectra to be collected in fewer time steps.	Cannot implement phase cycling, or modulation of single pulses in a pulse pair	Limited only by optical components (beam splitters, mirrors, birefringent crystals)

(Continued)

Table 1 Continued

<i>Method</i>	<i>Pros</i>	<i>Cons</i>	<i>Bandwidth limitations</i>
	Compensates group delay dispersion introduced by changing wedge thickness over a time scan Permits pump-probe geometry measurements without need for a pulse-shaper	Motorized delays can cause time zero uncertainties and timing errors (reduced in the case of refractive delay lines)	

Source: Fuller, F.D., Ogilvie, J.P., 2015. Annual Reviews of Physical Chemistry 66, 667–690.

The use of D.O.s or spatial filters to generate the four beams for a BOXCARS 2DES experiment promises high passive phase stability since the phases of any beams originating from a single beam are correlated. Thus when common optics are used to direct them to the sample, relative phase errors between pulse pairs can be minimized, producing a passively phase stable signal. To precisely scan the time delay and have the cancellation of the phase fluctuations in the signal, two designs of optical paths have been developed. One design uses a geometry similar to D.O.-based transient grating experiments where all four beams hit common reflective mirrors between the D.O. and the sample. In order to precisely scan the coherence time while maintaining passive-phase stability, refractive delays may be used, either by adjusting the material insertion using a pair of wedges mounted on a translation stage, or by rotating cover slides mounted on a rotational stage. The use of the refractive delays introduces additional chirp during the coherence scan and may pose some problems for broadband laser pulses and longer time scans. For a 10 fs Fourier-transform limited visible pulse centered at 600 nm, the temporal width can be distorted by $\sim 6\%$ when scanning up to 300 fs. The refractive delay works better for the coherence scan where optical dephasing < 100 fs. As an alternative to refractive delays, other designs using all reflective optics may be used. This approach was pioneered by the Miller group, who showed that a carefully designed optical path in which appropriate pulse pairs hit common optics can minimize phase drifts in the detected signal. They also improved the phase stability with a design that decouples the coherence time scan from the waiting time delays.

As an alternative to the passive phase-stabilization approach, the Cundiff group was the first implement an actively phase-stabilized 2DES experiment. In their approach, the four beams for the BOXCARS geometry are generated by two interferometers, one between the two excitation pulses and another between the probe and reference pulse. A third interferometer is used to actively control the relative phase of the output from the other two interferometers. Changes in the relative time delays of the four pulses are measured by monitoring the interference of a continuous HeNe laser co-propagating through the interferometers. The HeNe measurements provide feedback to actively phase-stabilize the setup by adjusting mirrors mounted on piezoelectric transducers. In this approach, the time delays are scanned by conventional delay stages and actively stabilized at each coherence and population time step, for which the 2D signal may be averaged over the desired number of laser shots. This setup is particularly powerful for enabling the measurement of double quantum coherence signals that require phase-stabilization over the population time delay T , and for experiments that require the scanning of long coherence time delays.

Pulse-shapers

Pulse-shapers can be used to replace one or two delay lines in 2DES experiments. Removing the mechanical delay lines can improve the phase stability and provide the opportunity to control the relative pulse phases, enabling “phase-cycling” for scatter reduction and the isolation of particular signals of interest. The Nelson group has developed a spatial-temporal pulse-shaper in which the four laser pulses, generated either by a D.O. or by a spatial light modulator (SLM), are incident at the different locations of a 2D SLM pulse-shaper, enabling independent control of each pulse delay and phase. They have shown that space-time coupling effects in the setup can be largely corrected. The Ogilvie group has combined the use of an acousto-optic (Dazzler) pulse-shaper and D.O. for 2DES. In their implementation, the pulse-shaper is placed before the D.O. to generate a pair of collinear pump pulses with controlled relative phase. Thus instead of having four laser pulses (Fig. 2), there are six pulses incident on the sample. Two pairs of the pulses are time overlapped and generate two transient grating signals at different waiting times in the direction of the local oscillator. Rephasing and nonrephasing signals are emitted simultaneously in the phase-matching direction and can be separated via phase-cycling. Simultaneous collection of both rephasing and nonrephasing signals facilitates excellent “phasing” of the data. In general, the use of pulse-shapers offers the advantages of ease-of-use and phase-cycling that can isolate signals of interest and suppress background scatter that differs from the desired signal in its phase dependence on the incident pulses. Pulse-shapers that can operate at high update rates also enable detection of the signal in the rotating frame with coarse time steps on a shot-to-shot basis, speeding up the experiments. The disadvantages of pulse-shaper-based 2DES designs stem from the spectral bandwidth (typically ~ 100 – 200 nm are achievable), maximum time delays (typically picoseconds) and overall throughput (ranging from ~ 10 – 75%), each of which is pulse-shaper and implementation-dependent.

“Phasing” the 2DES signal

In principle, if all the pulses generated from a single light source experience the same optical elements and the reference beam follows the signal path, the absolute phase from the different pulses cancels in the signal. However, in the BOXCARS geometry, the

absorptive and dispersive part of the 2D signal are mixed even with the use of a D.O., requiring that a polynomial phase be added to separate the real and imaginary part of the complex signal. The phase distortion of the 2D signal may be due to the imperfect optical surfaces, the zero time offset between pump pulses and unbalanced material dispersion in the beam paths. A few methods have been proposed to phase the 2D spectrum.

The Jonas group first proposed to determine the absolute signal phase by utilizing a tracer pulse to follow the optical path of the signal. Spectral interferometry could then be used to determine the difference between the relative phases of the tracer and the reference, and the reference and signal. With measurement of the spectral phase of the tracer as determined by frequency resolved optical gating (FROG), the absorptive and refractive parts of the 2D spectrum could be separated. Alternatively, the Cundiff and Hamm groups have employed spatial interference for phasing the signal.

The projection-slice theorem is the most commonly used approach for phasing 2D spectra. It has been shown by the Jonas group that the integration of the 2D signal over the excitation axis is equivalent to the pump-probe signal. Thus, the signal phase can be determined by a fitting procedure that finds the optimum variable spectral phase for which the integrated 2D signal matches the spectrally-resolved pump-probe signal for the same waiting time delay. Two potential issues may occur when applying this theorem. First, the pump-probe signal obtained in the BOXCARS geometry is weaker than the 2D signal and can be noisy, in particular for scattering samples. Secondly, the pump-probe signal is often obtained by blocking one pump beam and one probe beam in the BOXCARS geometry. Since the geometrical configuration of the pump-probe and 2DES measurements are different this may result in a difference of the signal amplitude and potentially spectral differences between the integrated 2DES and pump-probe data.

To resolve the issue of the weak pump-probe signal, the Hauer group has employed two D.O.s in the 2DES setup in such a way that both the rephasing and nonrephasing signals can be collected simultaneously. Through balanced detection, an automatically-phased heterodyne-detected transient grating signal (TG) can also be obtained in this configuration. The phased heterodyne-detected TG is much stronger than the pump-probe signal, and can be used in its place to phase the 2D spectrum using the projection-slice theorem.

2DES in the pump-probe geometry

2DES in the pump-probe geometry was first proposed by the Jonas group in 1999 and implemented first in the infrared by Zanni and Tokmakoff and later in the visible by Ogilvie. The pulse sequence is showed in Fig. 1(c), where a collinear pair of pump pulses excite the sample, which is then probed by a third noncollinear probe pulse. Since the wavevectors of the pump pulses are identical, the rephasing and non-rephasing signals are emitted in the same direction, along with two pump-probe signals and the probe beam. An advantage of the pump-probe geometry is that the relevant pulse pairs hit common optics and phase fluctuations cancel in the signal, providing excellent passive phase stability. In addition there is no unknown relative phase between the local oscillator (the transmitted probe) and the signal. This obviates the need for a phasing procedure as the absorptive part of the 2D spectrum is automatically obtained. Compared to the BOXCARS geometry, the phase-matching of broadband pulses is easier to achieve in the pump-probe geometry. As the probe beam follows the signal path and enters the detector, an attenuator may need to be placed before the detector in order to avoid detector saturation. As a result, the signal is also attenuated and the weak signal becomes more susceptible to noise, including scatter. The inability to control the relative amplitude of signal and local oscillator limits the ability to optimize the signal-to-noise ratio. Recently, the Jonas group showed that this problem can be solved with a modified pump-probe geometry in which the sample is placed in a Sagnac-interferometer. This setup enables the attenuation of the local oscillator without affecting the intensity of the probe beam. Thus a stronger 2D signal can be generated without saturating the detector and the signal-to-noise ratio can be optimized.

The coherence time scan in the pump-probe geometry is not as straightforward as the BOXCARS geometry since two excitation pulses are overlapped in space. Several methods have been used to implement the coherence scan. Tokmakoff has used conventional delay stages, while the Zanni and Ogilvie groups have utilized an AOM-based pulse-shapers, either in a 4f geometry or using a transmissive shaper (Dazzler) to generate the pump pulse pairs and scan the coherence time. As discussed previously, the use of pulse-shapers makes it possible to collect the signal in a rotating frame scheme on an (almost) shot-to-shot basis which greatly decreases the experimental time. Phase cycling can be used to reduce scatter and isolate signals of interest such as rephasing and nonrephasing components. The insertion of a pulse-shaper into the pump arm of a frequency-resolved pump-probe setup enables its straightforward conversion into a 2DES experiment.

The Cerullo group developed a method called Translating-Wedge-Based Identical Pulses eNcoding System (TWINS) to generate the time-delayed pulse pairs for 2DES in the pump-probe geometry. The method employs a 45° polarized beam incident upon several birefringent elements, including wedges mounted on a translation stage, to generate two orthogonally-polarized pulses with a variable time delay. After the birefringent materials, a 45° polarizer projects the time-delayed collinear pulse pair to the same polarization. Although refractive delays are used, this setup is capable of broadband operation, with ~ 10 fs pulses in either UV, visible or mid-IR regimes by a careful selection of the birefringent materials. In this setup, the coherence delay is calibrated using spectral interferometry between two pump pulses in order to achieve an accurate excitation frequency axis. A phasing procedure owing to the uncertainty of the time zero offset is required to obtain the absorptive 2D spectrum.

The Fully Collinear Geometry

The collinear geometry 2D setup was pioneered by the Warren group in 2003, using an acousto-optic pulse-shaper to generate a sequence of three pulses for the experiment. In this geometry the 2D signal is spatially overlapped with the incident pulse sequence, as well as the linear and other higher order signals. In this case a combination of fluorescence detection to enable

spectral filtering of the incident pulses, and phase-cycling can be used to extract the desired signal. The collinear geometry avoids possible spectral distortions that can arise from imperfect phase-matching in other 2DES geometries. This is particularly useful for studies of samples such as molecular clusters or single molecules, the size of which is smaller than the optical wavelength.

The Marcus group has also pioneered fluorescence-detected 2DES in the collinear geometry. Instead of using a pulse-shaper, they employ a phase-modulation approach, using two Mach-Zender interferometers to generate two independently phase-modulated collinear pulse pairs. Each beam is tagged with a unique frequency via an acousto-optic modulator (AOM), such that each pulse pair is modulated by the difference frequency of the AOMs. A third delay line scans the waiting time between the excitation and probing pulse pairs. The nonlinear signal of interest can be selected via lock-in detection using an appropriately constructed reference frequency. This reference frequency can be generated by combining the pulse-pair interference signals from each Mach-Zender interferometer, detected in separate monochromators. The double phase modulation can significantly suppress noise and remove the effect of time jitter caused by timing errors in the mechanical delay lines.

We note that the use of fluorescence signal detection requires that the final pulse leave the system in an excited state population. Compared to the standard 2DES three pulse sequence, an additional field-matter interaction is required. This can be realized by adding a fourth pulse, as is done by the Marcus group, or by having one of the three pulses interact with the sample twice. In the Warren implementation, the second pulse plays this role, yielding a 2D spectrum at zero waiting time. The additional field-matter interaction means that there are differences between the quantum pathways accessed in fluorescence-detected and standard 2DES experiments, making them complementary approaches. As an alternative to fluorescence detection, the fully collinear phase-modulation approach can be combined with other measurements that detect excited state populations. For example, both the Marcus and Cundiff groups have used this approach in combination with photocurrent signal detection. Aeschlimann and coworkers have used photoemission electron microscopy to read-out the excited state population, achieving nanoscopic spatial resolution.

Noise Suppression

2DES signals are often contaminated by unwanted spurious signals as well as scattering. In the BOXCARS geometry, a linear response contribution from the local oscillator, as well as a third-order signal will be detected along with the desired 2D signal. These unwanted signals can be removed via Fourier transformation along the detection frequency axis. However, scattering from excitation and probe pulses can also emit along the signal detection direction, and can significantly degrade the signal quality. An early solution was to use a shutter/chopper in the probe beam path to remove the scattering from the excitation pulses. Recently, the Zigmantas group implemented a double modulation scheme to suppress the noise and signals from slowly accumulating long-lived species.

Pulse-shapers also offer easy noise suppression methods. If scattering or other unwanted signals have different phase dependence than the 2D signal, phase cycling can be used to effectively isolate the desired 2D response. This method is particularly useful in the pump-probe and fully collinear geometries where the signal is overlapped with large background optical responses. The particular experimental implementation and desired signal dictate the appropriate phase-cycling schemes.

The Cerullo group has employed a noise suppression approach wherein a loudspeaker is attached to a probe mirror mount. The vibration induced by the speaker modulates the 2D signals on a shot-to-shot basis. The scattering can be suppressed by carefully tuning the amplitude and phase of the vibration. A similar approach that uses a wobbling Brewster window or photoelastic modulator has been used by the Hamm group in 2D infrared experiments.

Single-Shot Measurements

The Engel group developed a single-shot 2DES method called gradient-rapid-assisted-pulse-echo spectroscopy (GRAPES). In analogy to approaches used in single-shot pulse characterization methods, they use spatial encoding of a time delay. This is achieved by focusing the two excitations pulses using a cylindrical lens. As the two excitation beams cross at an angle near the focus, the wavefront tilt produces a spatial encoding of the coherence time delay for a fixed waiting time. The third-order signal is then imaged onto a two-dimensional CCD camera in which the vertical axis resolves the signal at different coherence time delays. In this approach rephasing and nonrephasing signals are collected independently and the focusing geometry defines the range of coherence times that can be scanned, which is typically ~ 300 fs. The transient grating signal can be extracted from either the rephasing or nonrephasing spectrum at $\tau=0$. Since the transient grating signal is equivalent to 2D spectrum integrated over the excitation axis, the rephasing and nonrephasing spectrum can be phased independently by using the projection slice theorem. The single-shot approach has the advantage of being robust against laser fluctuations that can degrade the signal-to-noise ratio in 2DES experiments. The rapid acquisition time also enables higher order multidimensional spectroscopies as recently demonstrated by the Harel group.

Applications of 2D Electronic Spectroscopy

Early 2D Studies

The development of 2D spectroscopy was inspired by Ernst, who suggested in 1976 that the technique of multidimensional NMR could be extended to other frequency regimes. With the development of ultrafast lasers, this idea was realized by Joffe in 1996. In the first implementation of 2D spectroscopy, Joffe and coworkers characterized the second-order response of a nonlinear crystal (potassium dihydrogen phosphate) (KDP). The signal $\chi^{(2)}$ was presented as a function of both the difference and sum frequencies

between the two excitation frequencies. It was suggested that such a 2D map could provide relevant information about the coupling between excited states.

The first third-order Fourier transform 2D spectroscopy was implemented by the Jonas group in 1998. A laser dye–IR144 was used to demonstrate the feasibility and capability of the experiment. The solvation dynamics of IR144 was investigated later by the same group. They found that the lineshape of 2D absorptive spectra evolves from a symmetrical diagonally elongated shape at 0 fs to a round square-like shape at 100 fs. The lineshape evolution reflects the fast loss of the correlation between excitation and detection frequency, induced by vibrational relaxation and solvation.

Since these early studies, 2DES is being used in an increasing number of applications. These applications exploit the strengths of the method that are particularly suited to the system of interest. Below we detail several examples of 2DES applications to a wide range of systems.

Photosynthesis

In photosynthetic systems, arrays of multi-pigment antenna complexes harvest sunlight and transfer the excitation energy to reaction centers where the energy is converted to a stable charge-separated state that fuels the slower biochemical steps of photosynthesis. The pigment-protein antennae and reaction center complexes make ideal targets for 2DES studies for a number of reasons. Designed to absorb light energy over a wide range of wavelengths, photosynthetic systems use a relatively small number of types of pigments, held in place by the surrounding protein. The protein environment tunes their electronic resonances and fixes their relative separations and orientations to define the energy landscape that guides excitation energy transfer and charge separation pathways. As a result, the linear absorption spectra of photosynthetic complexes are highly inhomogeneously broadened, and the electronic coupling between pigments varies from weak, where excitations are localized on individual pigments, to strong, where excitations are shared among pigments and an “excitonic” picture is appropriate. 2DES provides unique information about the electronic structure of photosynthetic systems that is buried in the linear absorption spectrum. 2DES as a function of waiting time details the photosynthetic energy transfer and charge separation pathways, enabling tests of transfer mechanisms. Below we give a brief overview of a subset of the work in this growing field and refer the reader to reviews ([Ogilvie and Kubarych, 2009](#); [Ginsberg *et al.*, 2009](#); [Lewis and Ogilvie, 2012](#); [Mukamel *et al.*, 2009](#)) for a more comprehensive discussion.

In 2005 the first 2DES study of a photosynthetic complex was made by Fleming and coworkers, who examined the Fenna–Matthews–Olson (FMO) complex from green sulfur bacteria. Its simple and well-characterized structure has made the FMO complex an important model system to investigate photosynthetic energy transfer. The FMO complex is a trimer of identical subunits, consisting of 7 bacteriochlorophylls (BChl) (although an 8th BChl is present in some preparations) that serves as a bridge structure to transfer energy from the light-harvesting antennae to the reaction center. Within each subunit, the intermolecular electronic coupling in FMO leads to seven excitonic states delocalized over a subset of BChl molecules as depicted in [Fig. 3](#). Fleming and coworkers mapped the energy transfer pathways in FMO by examining the waiting-time dependence of the cross peak amplitudes that indicate population transfer between the different excitonic states. Although consistent with overall downhill energy flow, their results suggested that there are two dominant pathways in FMO. Recently, Zigmantas and coworkers

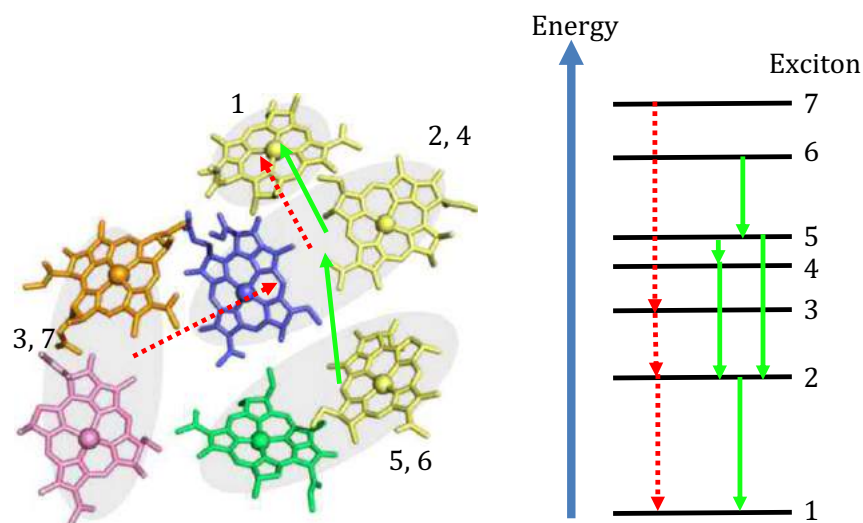


Fig. 3 The seven bacteriochlorophyll pigments of FMO (left), and the excitonic level structure (right). Gray shading indicates the delocalized nature of the excitons. Energy transport pathways identified by Fleming and coworkers ([Brixner *et al.*, 2005](#)) are depicted in red (dashed) and green (solid). Adapted from Ogilvie, J.P., Kubarych, K.J., 2009. *Advances in Atomic, Molecular, and Optical Physics* 57, 249–321.

revisited this system by using 2DES with full polarization-controlled beams. Controlling the polarizations of excitation and probing beams make it possible to suppress diagonal (or cross) peaks and enhance visibility of the cross (or diagonal) ones. This technique was first utilized by Hochstrasser in the infrared and has been employed by several groups for 2DES studies. Through global analysis of the polarization-controlled 2D data Zigmantas and coworkers were able to resolve the 8th exciton state and further unravel the complex energy-transfer pathways.

Another example in which 2DES was used to map energy-transfer pathways was a study by the Scholes group of the light-harvesting II (LH2) complex from purple bacteria. LH2 contains two types of chromophores: carotenoids (Car) and BChl. The electronic structure of Car is often described by a three-level model – a ground state, a dark low-lying state (S_1 , $2A_g^-$) and a bright high-lying state (S_2 , $1B_u^+$). Two energy-transfer pathways from Car to BChl have been suggested: a direct energy transfer from S_2 state to the BChl Q_x state and an indirect energy-transfer pathway via S_1 state to the BChl Q_y . The theoretically estimated overall energy-transfer efficiency from the two pathways was in poor agreement with experimental measurements. It had been proposed that a dark/weakly allowed state (X) might exist between S_2 and S_1 which was missing in the energy-transfer calculation and unresolved in 1D spectra. The Scholes group showed that this state could be clearly resolved by 2DES. To study the role of the dark state in energy transfer, they performed a global-target analysis, finding that the 2D evolution associated spectra (EAS) exhibited multiple cross peaks corresponding to the X, S_1 , S_2 and Q_x states. Their analysis revealed a clear map of energy transfer via the dark state, resolving the discrepancy between theoretical calculations and experiments.

The Ogilvie group was the first to study photosynthetic reaction centers with 2DES. They examined the photosystem II reaction center (PSII RC), which is at the heart of oxygenic photosynthesis, taking excitation energy from neighboring antennae complexes and forming a charge separated state on ultrafast timescales. The PSII RC contains six chlorophylls (Chl) and two pheophytins (Pheo) arranged in two symmetric branches, with charge separation occurring on one side. The spectral overlap of all of the RC pigments, and the similar timescales of energy transfer and charge separation in this system have made it particularly challenging to understand both the electronic structure and charge separation mechanism of the PSII RC. To extract the heterogeneous kinetics of the system they fit the 2D dataset at each frequency-frequency point to a function composed of four exponential-decay terms. From this they constructed 2D decay-associated spectra to expose connections between the initially excited states and subsequent intermediates throughout the processes of rapid energy transfer to the charge separated state. In collaboration with Abramavicius and Mukamel, the Ogilvie group has used 2DES to test excitonic and charge separation models of the PSII RC (see Fig. 4). More recently they have performed 2DES experiments on the PSII “core” complex, revealing the spectral signatures of energy equilibration and transfer between neighboring antennae and the RC.

Coherence

Excitation with broadband pulses can readily excite superpositions of electronic, vibrational and mixed electronic-vibrational (vibronic) states, leading to coherent oscillations observable in time-resolved spectroscopies. Of particular interest to the photosynthesis and artificial light-harvesting communities have been observations of coherent processes on timescales concurrent with energy transfer and charge separation, leading to considerable speculation about the possible functional importance of the observations. The first report of coherent dynamics in a photosynthetic system was made by Vos, Martin and coworkers in the early 1990s in their pioneering pump-probe studies of the bacterial reaction center. Their work stimulated considerable experimental and theoretical effort. The interest in coherence in photosynthetic function was revived by the Fleming group in 2007 when they reported long-lived coherent amplitude oscillations in 2DES studies of the FMO complex at 77 K. These oscillations appeared at cross-peaks in the 2DES data as a function of waiting time. They originally attributed the observations to electronic coherence between exciton states, and proposed that the long-lived nature of the coherence could have important functional implications for efficient and coherent energy transfer. Similar coherent dynamics have now been observed by 2DES in a number of other light-harvesting antennae and reaction centers over a wide range of temperatures.

Considerable effort by the 2DES community has been devoted to understanding the physical origin of coherent dynamics in 2D spectra. Experiments by the Kauffman and Ogilvie groups on simple laser dyes showed that purely vibrational coherence can generate coherent oscillations in 2DES spectra. When the electronic structure of a system is well understood, the ways in which coherence between different states can be generated by a 2D pulse sequence can be readily enumerated and mapped to the appropriate (ω_1 , ω_3) position on a 2D spectrum. Turner and coworkers performed this analysis for simple systems as illustrated in Fig. 5. They find that purely electronic coherence produces oscillations in the cross peaks of rephasing spectra and the diagonal peaks of nonrephasing spectra, while purely vibrational coherence produces a distinctly different pattern of oscillating signals. Egorova has also performed a similar study, as have Butkus and coworkers. The latter have Fourier transformed the 2D spectra (or rephasing or nonrephasing spectra) along the waiting time T , extracting Fourier-transformed maps (often called coherence amplitude maps or beating maps). Such maps distinguish electronic and vibrational coherence and separate vibrational coherence from ground- and excited- electronic states at certain peak positions. Inspired by this work, Seibt and Pullerits Fourier transformed the complex 2D signal and found that the coherence amplitude maps further separates into positive and negative frequencies, which can help distinguish signals from different pathways. In addition to these data analysis methods, the Ogilvie group developed an experimental approach termed ‘two-color rapid acquisition coherence spectroscopy’ (T-RACS), to separate vibrational coherence from ground- and excited-electronic states, demonstrating that this could be done for some vibrational modes of chlorophyll in solution. This experiment is a variant of 2D spectroscopy and has the advantage to greatly reduce the experimental time.

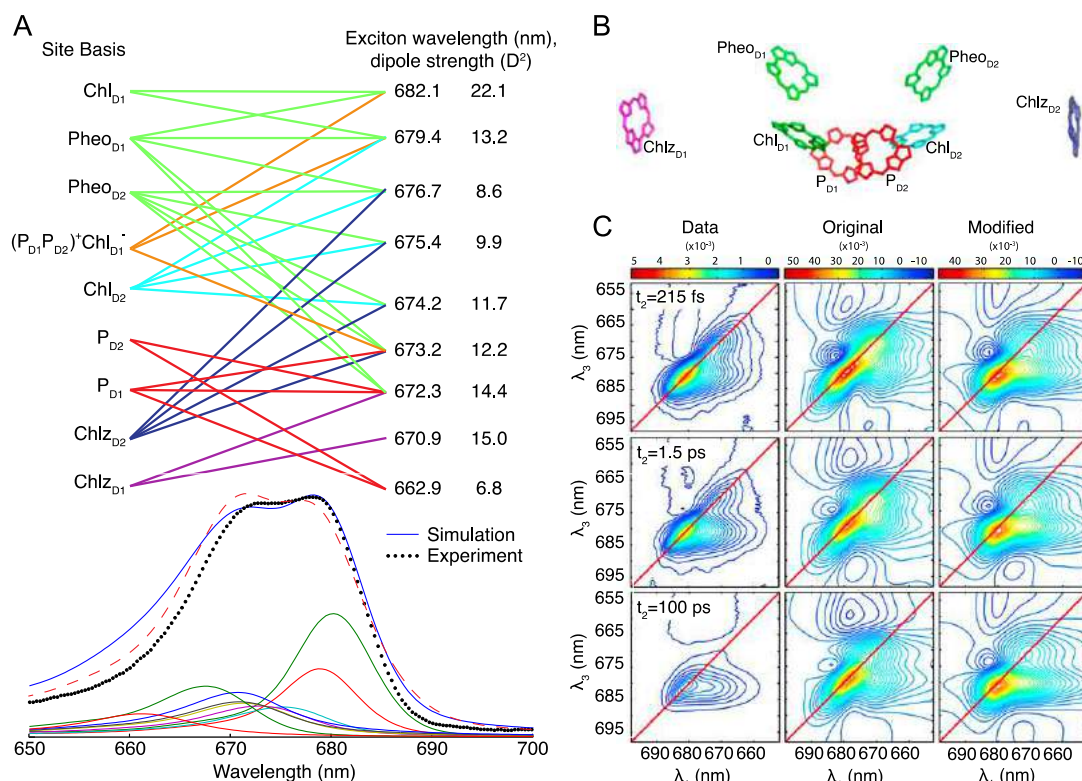


Fig. 4 (A) Excitonic model of the PSII RC. Lines denote any pigments with greater than 10% probability of participating in the connecting exciton state. Also given are the dipole strengths. Bottom, the experimental (black dotted) and simulated 77 K absorption spectrum of the Qy band for the modified (blue solid) and original Novoderezhkin model (red dashed). Also shown are the underlying excitonic contributions calculated for 5000 realizations of disorder for the modified model. (B) Structural arrangement of the PSII RC chlorophyll (Chl) and Pheophytin (Pheo) pigments (carotenoids present in the structure are not shown), where the D1 and D2 subscripts reflect the location of the pigments in the D1 and D2 proteins of the PSII RC. The most strongly coupled pair of Chls are denoted P_{D1} and P_{D2}. (C) Two-dimensional electronic spectra of the PSII RC at 77 K for waiting times $T=215$ fs, 1.5 ps and 100 ps. Experimental data (left) and simulations based on the Novoderezhkin and modified excitonic models are shown in the middle and right. Adapted from Lewis, K.L.M., Fuller, F.D., Myers, J.A., *et al.*, 2013. *Journal of Physical Chemistry A* 117 (1), 34–41.

Although the physical origin of coherence can be readily established in simple systems, photosynthetic complexes have poorly understood electronic structure with higher-lying excited states, electronic and vibrational coupling, protein-pigment interactions and dark states. Nevertheless there has been considerable progress made in understanding the 2DES observations of coherence in photosynthetic systems. Jonas and coworkers considered a vibronic dimer and realized that the nonadiabatic electronic and vibrational mixing leads to new pathways for vibrational coherence in the ground electronic states. Within this model they were able to describe the main 2DES signatures of coherence in FMO. In agreement with work by Moran on a different photosynthetic system, they noted that resonance between electronic energy gaps and pigment vibrational modes could have important implications for energy transfer. The Ogilvie and van Grondelle groups have reported coherent dynamics in the PSII RC and have noted the existence of electronic-vibrational resonances in this system, proposing their possible importance for charge separation. In general there is a growing consensus that the coherent dynamics that have been observed to date in photosynthetic system by 2DES are largely vibrational and/or vibronic in origin. It remains an active area of research to understand the functional relevance of electronic-vibrational resonances for energy transfer and charge separation.

Excitonic Electronic Structure and Dynamics in J-Aggregates

Molecular aggregates are clusters of molecules that self-assemble into superstructures with separations typically on the order of their component size. Similar structures are found in nature, for example in the chlorosome antenna of green sulfur bacteria. J-aggregates from cyanine dyes, have been developed as artificial light-harvesting systems aimed at mimicking the efficient energy transfer properties of photosynthetic antennae. These aggregates can be prepared in aqueous solution and form a double-layered tubular strand of ~ 11 nm outer diameter, total tube wall thickness of ~ 4 nm, with a lamellar spacing of the individual chromophore layers of ~ 2.2 nm. Within either chromophore layer, the dye molecules can be considered to be arranged in a brickwork structure. The different molecular electronic coupling strength within layers and between layers leads to a complex electronic structure and rich excitonic dynamics. The linear absorption spectrum of this system exhibits four main peaks, but it is unclear whether or not some

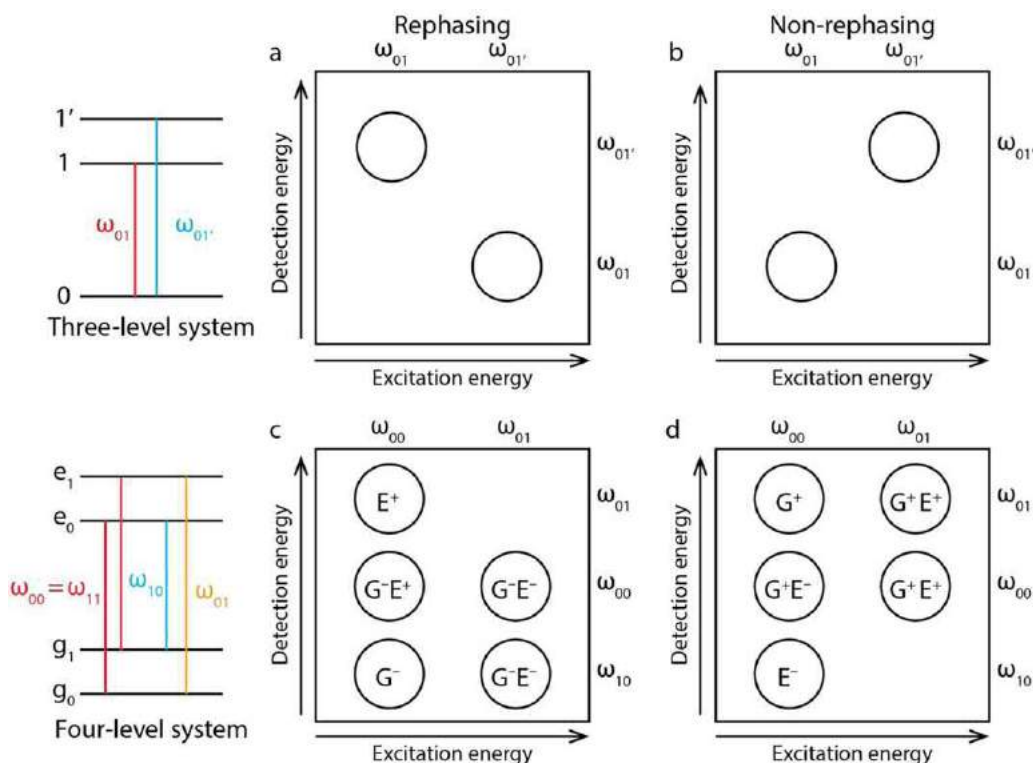


Fig. 5 Coherence amplitude maps for a three-level system (composed of a ground and two excited electronic states) (a, b) and a four-level system (composed of ground and excited electronic states each coupled to one vibrational mode with a single excited vibrational level shown on each electronic state) (c, d). Both rephasing (a, c) and non-rephasing maps (b, d) are displayed. The maps are generated by using double-sided Feynmann diagrams under the assumption that all transitions are allowed. In Fig. (c) and (d), G and E denote vibrational coherence from ground- and excited-electronic states, respectively. $+/-$ denote the signal appears at the positive/negative frequency (f_2).

peaks are from non-aggregated monomers. Kauffman, Hauer and coworkers have performed a series of 2DES studies to investigate the electronic structure and energy transfer pathways within these aggregates. The cross peaks in the 2D map provide clear evidence of the delocalized nature of the absorption bands. By using this model system, the authors also demonstrated the capability to perform global analysis on the 2D dataset to reveal the detailed excitonic relaxation pathways in this system.

Charge Transfer in Organic Photovoltaic Materials

Organic photovoltaics (OPVs) use a blend of an electron donor such as a conjugated polymer, and an electron acceptor such as fullerene to convert sunlight into electricity. The electron donor often serves as the main light-harvester for the system. The primary processes in OPVs are the light absorption by electron donors, the transfer of energy to the donor/acceptor interface, and charge separation at the interface. Mapping the pathways of energy transfer and charge separation and elucidating their mechanism in OPVs is not only of fundamental importance, but also aids in the optimization of devices. Scholes and coworkers investigated charge transfer dynamics in a blend of poly-(3-hexylthiophene) (P3HT) and [6,6]-phenyl-C61 butyric acid methyl ester (PCBM) by using 2D spectroscopy. This blend has very rapid electron transfer. This process is seen as a rise of the P3HT hole signal, indicated by a photoinduced absorption (PIA) that is clearly resolved as a cross-peak in 2DES. One intriguing finding in this study is that vibrational coherence corresponding to C=C stretching mode has been transferred from the excited state to the P3HT hole. It is likely that the coherence here is a spectator mode and does not affect the charge-transfer dynamics. However, the existence of coherence on the P3HT hole implies that charge transfer precedes vibrational relaxation, exposing the significance of hot electron transfer.

Solid State Systems

2DES has been applied to study solids (almost exclusively semiconductors) in the past 10 years. Semiconductors lie at the foundation of the electronic and photonic industry. They are a class of materials with an energy band gap of 0.5 eV to a few eV. There are two different types of semiconductors, indirect gap and direct gap materials. Si and Ge are the most widely applied in indirect gap materials while GaAs, GaP, and InAs are the well-known direct gap materials. Direct gap materials are efficient light absorbers and emitters, thus widely used in light emitting devices such light emitting diodes (LEDs) and lasers. Most LEDs and lasers are built on quantum confined structures such as quantum wells and quantum dots in which the electrons are confined to a

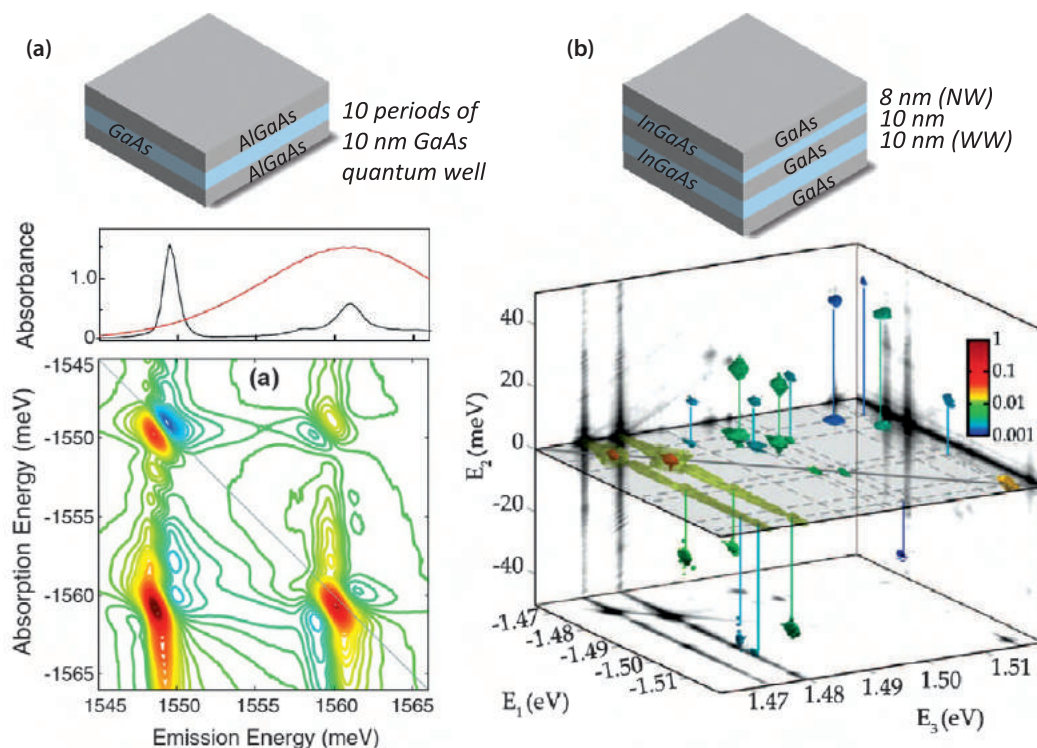


Fig. 6 Illustration of a semiconductor single quantum well (a), a double quantum well (b) and measured 2D spectra from them. (a) The dispersive line shape in the real part of the nonlinear signal originates from many-body interaction induced energy shifts of the excitons. (b) A total of 33 peaks are identified in a 3D spectrum from a double quantum well. Weak diagonal peaks are from dark excitons. Coupling between bright and dark excitons enhances the signal from dark excitons.

two-dimensional plane or localized in all three dimensions. Understanding electronic coupling and quantum coherence are critical for building both conventional optoelectronic devices and future quantum information processing devices.

2DES was first applied to study many body interactions between excitons in quantum wells by Cundiff and coworkers. Excitons in semiconductors are known as Wannier excitons with larger Bohr radius than the Frenkel excitons found in molecular and organic materials. A number of mechanisms such as the local field, biexciton formation, excitation induced dephasing, and excitation induced shift are known to contribute to many body interactions. Using conventional 1D spectroscopy, however, it is difficult to distinguish between these different mechanisms. As one of the earliest applications of 2DES to solids, Cundiff and coworkers explored the phase resolved nonlinear signal in 2DES, which provides valuable information for determining the dominant interaction mechanism. In the example shown in Fig. 6(a), heavy-hole and light-hole exciton resonances in a GaAs/AlGaAs quantum well are simultaneously excited and identified as the diagonal peaks while the coupling between them is manifested as off-diagonal peaks. Most interestingly, the dispersive lineshape of these peaks suggests that the leading many-body effect is excitation induced energy shift.

As a natural extension, 3DES has been used to probe semiconductor asymmetric double quantum wells. Double-well potentials are encountered in many important problems and systems such as diffusion of defects in solids, structural relaxation in glasses, and proton tunneling between DNA bases. A semiconductor double quantum well hosts a rich variety of exciton resonances with tunable energies, thus serving as a useful model system for investigating coupling and dynamics encountered in double well potentials in general. In addition to bright excitons, Davis and coworkers have observed two types of dark excitons: parity-forbidden excitons and spatial indirect excitons in asymmetric InGaAs/GaAs double quantum wells. The observation of these dark excitons with small oscillator strengths was made by possible via off-diagonal coupling peaks between the bright and dark excitons. Such coupling effectively amplifies the signal from dark excitons. In the spectrum shown in Fig. 6(b), a total of 33 peaks were identified, providing unprecedented and comprehensive information on excitonic resonances, coupling among them, and associated quantum dynamics in double quantum wells.

In semiconductor quantum dots, excitons are confined in all three dimensions, leading to bound many-body states. For example, trions (excitons bound with an additional charge) and biexcitons (a bound state of two excitons) have all been clearly identified in 2DES performed on an ensemble of quantum dots. The biexcitons are clearly identified as an off-diagonal peak corresponding to absorption at the exciton state and emission at an energy shifted toward lower energy by an amount corresponding to the biexciton binding energy. The high sensitivity of 2DES allows nonlinear signals from just a few quantum dots to be measured. The excitons confined in each quantum dot are slightly shifted in energy due to different confinement potentials, leading to multiple isolated diagonal peaks. Quantum dots that are coupled to each other can be identified by off-diagonal peaks.

Spatial imaging of individual quantum dots further determines the separation between the coupled quantum dots. As expected, the coupling strength reduces as the separation between quantum dots increases.

References

- Brixner, T., Stenger, J., Vaswani, H.M., *et al.*, 2005. *Nature* 434 (7033), 625–628.
- Cho, M., 2009. *Two-Dimensional Optical Spectroscopy*. New York: CRC Press, p. 385.
- Cho, M.H., 2008. *Chemical Reviews* 108 (4), 1331–1418.
- Fuller, F.D., Ogilvie, J.P., 2015. *Annual Reviews of Physical Chemistry* 66, 667–690.
- Ginsberg, N.S., Cheng, Y.C., Fleming, G.R., 2009. *Accounts of Chemical Research* 42 (9), 1352–1363.
- Hamm, P., Zanni, M.T., 2011. *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge: Cambridge University Press, p. 286.
- Jonas, D.M., 2003. *Annual Review Of Physical Chemistry* 54, 425–463.
- Lewis, K.L.M., Ogilvie, J.P., 2012. *Journal of Physical Chemistry Letters* 3, 503–510.
- Mukamel, S., Abramavicius, D., Yang, L.J., *et al.*, 2009. *Accounts of Chemical Research* 42 (4), 553–562.
- Ogilvie, J.P., Kubarych, K.J., 2009. *Advances in Atomic, Molecular, and Optical Physics* 57, 249–321.

Multidimensional Terahertz Spectroscopy

Michael Woerner, Klaus Reimann, and Thomas Elsaesser, Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

© 2018 Elsevier Inc. All rights reserved.

Introduction

Over the last two decades, nonlinear terahertz (THz) spectroscopy (Luo *et al.*, 2004; Gaal *et al.*, 2006) has developed as a new field of condensed matter research, due to both substantial progress in the generation of THz pulses with high electric field amplitudes and new experimental concepts such as, e.g., THz pump–midinfrared (MIR) probe experiments (Gaal *et al.*, 2007; Günter *et al.*, 2009; Huber *et al.*, 2001, 2005; Kampfrath *et al.*, 2011; Pashkin *et al.*, 2011; Hwang *et al.*, 2015). THz pulses have a subpico- to picosecond duration and cover a frequency range from 0.3 to 30 THz or 10–1000 cm^{−1}, corresponding to photon energies from 1.25 to 125 meV and far-infrared wavelengths between 1 mm and 10 μm. Elementary excitations of condensed matter in this energy range include nuclear motions, i.e., translations, intra- and intermolecular vibrations as well as phonons in solids, and low-frequency electronic excitations such as excited-state absorption in large molecules or intraband, interband, defect-related and intra-excitonic transitions in solids (Tonouchi, 2007). Nonequilibrium dynamics and correlations of low-frequency excitations are not fully understood and call for time-resolved nonlinear spectroscopies in that spectral range. A second class of phenomena are processes driven by strong nonresonant electric fields such as carrier transport in the quantum kinetic regime or field-induced shifts of electronic states and transitions between them.

A preeminent feature of THz spectroscopy consists in the phase-resolved detection of electric field transients, allowing for direct time-domain measurements of THz fields in amplitude and absolute phase (Wu and Zhang, 1995). Both the spectral (Somma *et al.*, 2015) and the dynamic range of the field measurement has been improved significantly, now allowing for sophisticated experimental techniques involving phase-locked sequences of THz pulses (Leitenstorfer *et al.*, 1999; Huber *et al.*, 2000; Brodschelm *et al.*, 2001; Kübler *et al.*, 2004, 2005; Sell *et al.*, 2008; Junginger *et al.*, 2010; Schubert *et al.*, 2011). Nonlinear two-dimensional (2D) THz spectroscopy is a prominent method of this type, allowing for a measurement of couplings between electronic or vibrational excitations, for mapping structural fluctuations via time-dependent lineshapes, and for following chemical exchange processes in time (Mukamel, 2000; Jonas, 2003; Hamm *et al.*, 1998; Asplund *et al.*, 2000).

Multi-dimensional experiments in nuclear magnetic resonance (NMR) are typically performed in the nonperturbative regime using more than three radio-frequency pulses (Ernst *et al.*, 1997), while 2D optical spectroscopy has mainly been performed in the third-order or $\chi^{(3)}$ limit by applying pump–probe or three-pulse photon-echo techniques in noncollinear four-wave-mixing geometries (Hamm and Zanni, 2011). In a three-pulse photon-echo experiment, the pulses interact with the sample in a noncollinear geometry with wavevectors $\vec{k}_A$, $\vec{k}_B$, and $\vec{k}_C$. The first two interactions with the sample are separated by the coherence time τ . The first pulse propagating along $\vec{k}_A$ generates a coherent polarization of the sample, which is transformed into a transient population by the interaction with the second pulse propagating along $\vec{k}_B$. After the waiting period T , during which the population evolves freely, the nonlinear signal is generated by interaction with the electric field of the last pulse propagating along $\vec{k}_C$. Third order, i.e., $\chi^{(3)}$ nonlinear signals are emitted into the directions $-\vec{k}_A + \vec{k}_B + \vec{k}_C$ and $\vec{k}_A - \vec{k}_B + \vec{k}_C$, which are different from those of the three pulses interacting with the sample. In noncollinear 2D spectroscopy, a phase-resolved detection of the emitted third-order signal can be accomplished, e.g., by heterodyning it with a fourth synchronized pulse, the so called local oscillator (LO) (Hamm and Zanni, 2011).

In the THz frequency range, noncollinear n-wave-mixing geometries are inherently difficult if not impossible to realize due to the pronounced diffraction of beams at the large wavelengths involved. As an alternative approach, the fully phase-resolved collinear 2D spectroscopy has been developed for the THz range and been applied to a variety of condensed-phase systems (Kuehn *et al.*, 2009, 2011a,b; Junginger *et al.*, 2012; Woerner *et al.*, 2013; Bowlan *et al.*, 2014a,b; Folpini *et al.*, 2015; Somma *et al.*, 2014, 2016a,b). In this article, we describe the basic concepts and the most recent development in multidimensional THz spectroscopy. The article is organized as follows. In Section “Concepts of Collinear Multidimensional THz Spectroscopy”, we introduce the concept of frequency vectors which is fully equivalent to that of wavevectors in noncollinear 2D spectroscopy. Special diagrams exploiting chains of frequency vectors in 2D frequency space allow for disentangling quantum pathways in Liouville space, even in experiments where different orders of nonlinearities such as $\chi^{(3)}$, $\chi^{(5)}$, etc. occur simultaneously. In Section “Two-dimensional THz Spectroscopy with Three Phase-locked THz Pulses” we introduce an experimental setup for 2D THz spectroscopy with three independent THz pulses, followed by a discussion of two prototype experiments on intersubband coherences in semiconductor quantum wells (Section “THz-driven AC Stark Effect of Intersubband Transitions in Semiconductor Quantum Wells”) and two-photon interband coherences in the semiconductor InSb (Section “Three-pulse 2D THz Spectroscopy on InSb”). A brief summary and outlook are given in Section “Conclusions”.

Concepts of Collinear Multidimensional THz Spectroscopy

Multidimensional THz spectroscopy is typically implemented in a geometry where a sequence of THz pulses propagating in collinear direction interacts with the sample. The resulting nonlinear polarization (or current) emits a THz electric field which is

detected in the same direction. A pulse sequence applied in three-pulse experiments is schematically shown in Fig. 1(a) with pulses A (red) and B (green) separated by the coherence time τ and pulses B and C (blue) by the waiting or population time T . After the last interaction, the nonlinear polarization or current in the sample emits the electric field $E_{NL}(\tau, T, t)$ (black curve).

In the following, the coherence time τ is measured relative to the maximum of pulse B, whereas the population time T and the real time t are measured relative to the maximum of pulse C. As a result, the field $E_A(\tau, T, t)$ depends on τ , T and the real time t , the electric field $E_B(T, t)$ depends on T and t , and $E_C(t)$ on t only. A sequence with pulse A preceding pulse B preceding pulse C implies $\tau < 0$ and $T < 0$. Thus, the center of pulse A occurs at the real time $t_A = T + \tau$, the center of pulse B at $t_B = T$ and the center of pulse C at $t_C = 0$. Fig. 1(b) shows a 2D plot of the timing scheme as a function of both real time t and coherence time τ with the colored lines representing the centers of the THz pulses. A 3D Fourier transform along τ , T , and t provides the frequency dependent fields $\tilde{E}_A(v_\tau, v_T, v_t)$, $\tilde{E}_B(v_T, v_t)$, and $\tilde{E}_C(v_t)$. As a result, the orientations of frequency vectors which are perpendicular to the phase fronts of the pulses in a space with three time dimensions, are linearly independent of each other (Kuehn *et al.*, 2011a). For the pulse sequence shown in Fig. 1(b) the corresponding frequency vectors are:

$$\vec{v}_A = \begin{pmatrix} v_\tau \\ v_T \\ v_t \end{pmatrix} = \begin{pmatrix} -v_0 \\ v_0 \\ v_0 \end{pmatrix} \quad \vec{v}_B = \begin{pmatrix} 0 \\ v_0 \\ v_0 \end{pmatrix} \quad \vec{v}_C = \begin{pmatrix} 0 \\ 0 \\ v_0 \end{pmatrix} \quad (1)$$

The orientation of the frequency vectors of pulses A, B, and C are shown as red, green, and blue arrows in Fig. 1(b), respectively. A nonlinear signal due to a certain interaction sequence in Liouville space can be identified by the orientation of its phase fronts, in reference to the phase fronts of the three pulses. For third-order, i.e., $\chi^{(3)}$ signals, two different types of interaction sequences are distinguished, the so-called rephasing response, in which the phase differences acquired during the coherence period τ are reversed by interaction with the third pulse, and the nonrephasing response in which phase differences from the different interactions add up (Khalil *et al.*, 2003). For the timing scheme shown in Fig. 1(b), the phase fronts of the rephasing and nonrephasing nonlinear signals are shown as magenta and cyan lines, respectively. The phase fronts of the nonrephasing signal and of pulse 1 are parallel while the rephasing signal exhibits phase fronts perpendicular to those of pulse 1.

A standard approach for describing Liouville space pathways connected with a third-order nonlinear response are double-sided Feynman diagrams (Mukamel, 2000; Hamm and Zanni, 2011; Yee and Gustafson, 1978; Boyd and Mukamel, 1984; Mukamel, 1995; Yang *et al.*, 2008). The time evolution of the density matrix describing a rephasing contribution (left) and its nonrephasing counterpart (right) are shown in Fig. 1(c). Double-sided Feynman diagrams clearly show the time evolution of the $\langle \text{ket} \rangle$ and $\langle \text{bra} \rangle$

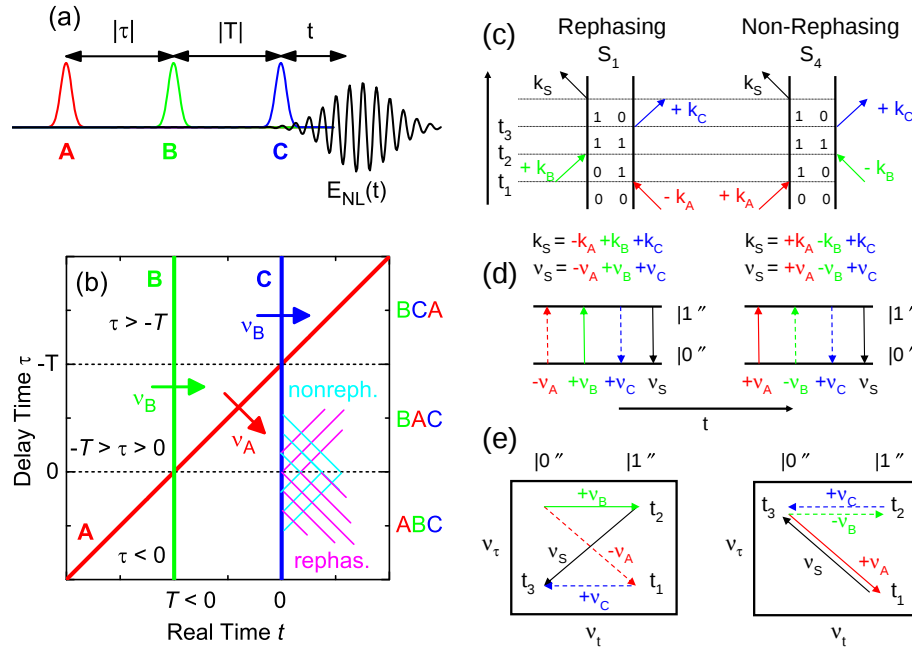


Fig. 1 (a) Timing scheme of a nonlinear, two-dimensional (2D) experiment using three THz pulses, pulse A (red), pulse B (green), and pulse C (blue). After the last pulse in the timing sequence the nonlinear current in the sample emits the electric field $E_{NL}(t)$ (black curve). (b) 2D plot of the timing scheme as a function of both real time t and delay time τ . Shown are the centers of pulses. For $\tau \leq 0$ (pulse sequence ABC) pulses A and B are separated by the coherence time τ , pulses B and C by the waiting time T , and the detection time t starts with pulse C. Arrows: orientation of the corresponding frequency vectors. The phase fronts of the rephasing and nonrephasing nonlinear signals are shown as magenta and cyan lines, respectively. (c) Standard double-sided Feynman diagrams describing the Liouville pathway of a third-order nonlinear response for a rephasing (left) and a nonrephasing contribution (right). (d) Liouville pathway diagrams in the style of Pollard *et al.* (1992). (e) Alternative diagrams exploiting frequency vectors used in the THz community (Kuehn *et al.*, 2009a,b).

Table 1 Coherence, waiting, and detection times and frequency vectors for different pulse sequences.

pulse sequence	Coherence time	Waiting time	Detection time	Rephasing freq. vector	Nonrephasing freq. vector
ABC	$t_B - t_A = -\tau$	$t_C - t_B = -T$	t	$\begin{pmatrix} v_0 \\ 0 \\ v_0 \end{pmatrix}$	$\begin{pmatrix} -v_0 \\ 0 \\ v_0 \end{pmatrix}$
BAC	$t_A - t_B = \tau$	$t_C - t_A = -T - \tau$	t	$\begin{pmatrix} -v_0 \\ 0 \\ v_0 \end{pmatrix}$	$\begin{pmatrix} v_0 \\ 0 \\ v_0 \end{pmatrix}$
BCA	$t_C - t_B = -T$	$t_A - t_C = T + \tau$	$t - T - \tau$	$\begin{pmatrix} -v_0 \\ 0 \\ v_0 \end{pmatrix}$	$\begin{pmatrix} -v_0 \\ 2v_0 \\ v_0 \end{pmatrix}$

side of the density matrix in Liouville space. Energetic aspects and phase matching conditions in wavevector or frequency vector space are not fully obvious in this picture. In panel (d), we show the same Liouville pathways in the style of Pollard *et al.* (1992). Here, the $|l\rangle$ evolution is represented by a chain of solid arrows whereas the $\langle b|$ evolution by dashed arrows. Transition energies are clearly visible in this type of diagram, frequently used in recent 2D THz experiments (Somma *et al.*, 2016a,b). In the 2D THz community (Kuehn *et al.*, 2009, 2011a) a third, alternative type of diagrams has been developed [cf. Fig. 1(e)], which exploits the frequency vector components v_r and v_t of Eq. (1) to visualize the Liouville pathway in 2D frequency space. The advantage of the latter diagrams is that the frequency vector chain on the $|l\rangle$ side of the density matrix (solid vectors) and that on the $\langle b|$ side (dashed vectors) determine both the orientation of the phase fronts of the driving and nonlinearly emitted fields. For instance, the diagrams shown in Fig. 1(e) predict correctly the orientation of the phase fronts of the rephasing $-\vec{v}_A + \vec{v}_B + \vec{v}_C$ (magenta) and nonrephasing $\vec{v}_A - \vec{v}_B + \vec{v}_C$ (cyan) contributions for pulse sequence ABC as shown in panel (b). For the different pulse sequences ABC, BAC, and BCA both the frequency vectors and the coherence, waiting, and detection times change their roles. Details are summarized in Table 1.

In the collinear THz 2D experiments performed so far (Somma *et al.*, 2016a,b), the pulse separation $t_C - t_B = -T$ has been fixed with the implication that the waiting time period T and correspondingly the v_r components of the frequency vectors are not accessible. Consequently, one cannot separate rephasing from nonrephasing signal contributions for pulse sequence BCA (last row in Table 1 and $\tau \geq -T$ in Fig. 1(b)). For the pulse sequence BAC ($-T \geq \tau \geq 0$) the interpretation of the experiment is complicated by the fact that the waiting time $-T - \tau$ is simultaneously varied with the coherence time τ . Thus, the pulse sequence ABC ($\tau \leq 0$) is typically used to analyze nonlinear signals in the $\chi^{(3)}$ limit. As we shall see below, the simple $\chi^{(3)}$ classification scheme according to Table 1 breaks down for higher-order nonlinearities $\chi^{(n)}$ with $n > 3$. In this regime, all pulse sequences ABC, BAC, and BCA are useful for the data analysis.

Two-Dimensional THz Spectroscopy With Three Phase-Locked THz Pulses

The experimental implementation of fully phase-resolved collinear 2D THz spectroscopy is discussed in this section (Kuehn *et al.*, 2009, 2011a,b; Junginger *et al.*, 2012; Woerner *et al.*, 2013; Bowlan *et al.*, 2014a,b; Folpini *et al.*, 2015; Somma *et al.*, 2014). The experiments of Somma *et al.* (2016a,b) were performed with three phase-locked THz pulses in the setup schematically shown in Fig. 2. A mode-locked Ti:sapphire oscillator generates linearly polarized 12 fs pulses centered at 800 nm at an 85 MHz repetition rate. A small part of the oscillator output is separated by a beam splitter to serve as a probe in free-space electrooptic (EO) sampling (Wu and Zhang, 1997). The major fraction of the oscillator pulses feeds two multi-pass amplifiers. Each amplifier delivers 25 fs pulses with an energy of up to 1 mJ at a repetition rate of 1 kHz. Pulses of the first amplifier are split into two parts of identical energy and mutual delay τ while the output pulse of the second amplifier is delayed by an additional time interval T . The three 800-nm beams are sent onto three separate GaSe crystals to generate the sub-picosecond phase-locked THz pulses A, B, and C. Phase-matched optical rectification within the pulse spectrum (Reimann *et al.*, 2003) provides almost octave-spanning pulses with a center frequency at 20 THz. Behind the GaSe crystals, the 800 nm beams are blocked by silicon wafers. Using a 90° off-axis parabolic mirror the three THz pulses are tightly focused on the sample. A couple of off-axis parabolic mirrors is used to image the THz electric fields transmitted through the sample onto a 10- μ m-thick (110)-oriented ZnTe crystal to be detected by free-space electrooptic sampling (Wu and Zhang, 1997). By varying the real time t between the three THz pulses and the probe pulse from the Ti:sapphire oscillator, the electric field of the THz pulses is measured in amplitude and phase. In Fig. 3(b) we show the driving fields used in the 2D THz experiments of Somma *et al.* (2016a,b).

Mechanical choppers are placed in each beam path in order to retrieve the nonlinear response of the sample. The three choppers are synchronized to the 1 kHz repetition rate of the laser system and allow for measuring seven transients, one when all three pulses ABC interact with the sample, three for the pairs AB, BC, or CA, and three for the single pulses A, B, or C. From the

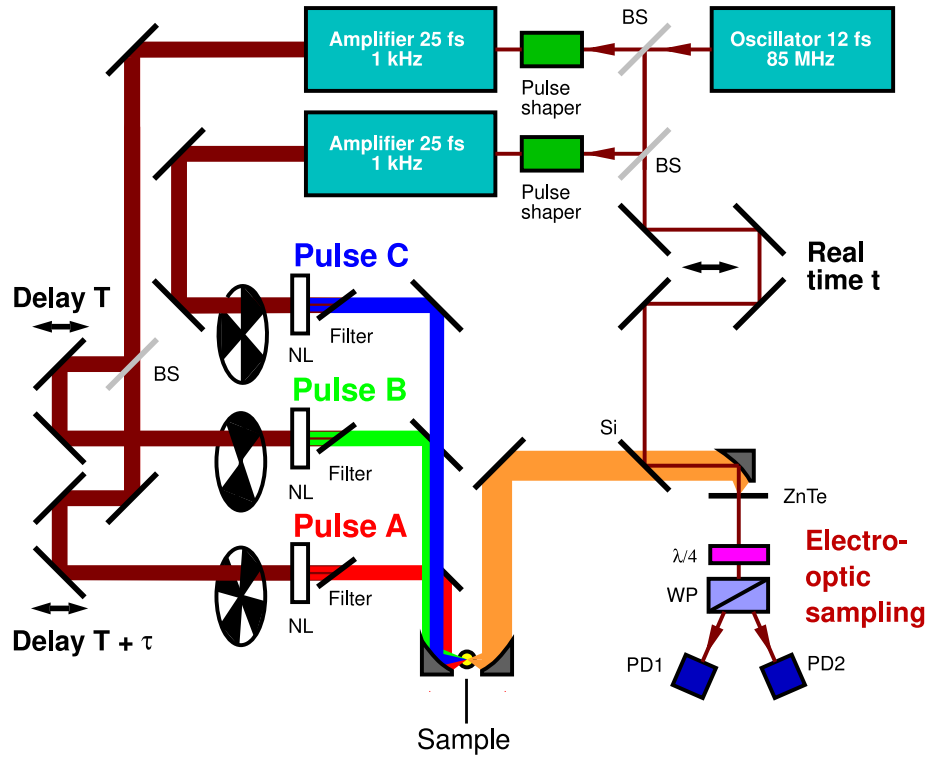


Fig. 2 Schematic view of the experimental setup for 2D THz spectroscopy used in [Somma *et al.* \(2016a,b\)](#). Intense THz electric fields are generated via optical rectification of near-infrared pulses in nonlinear crystals (NL) and detected by electrooptic (EO) sampling in a thin ZnTe crystal. Two acousto-optic pulse shapers, implemented in both amplifiers, serve to optimize the THz generation by tailoring the spectral components of the pulses from the oscillator.

measured transients, the nonlinear field emitted by the sample is given by

$$\begin{aligned}
 E_{NL}(\tau, T, t) &= E_{ABC}(\tau, T, t) \\
 &- E_{AB}(\tau, T, t) - E_{BC}(T, t) - E_{CA}(\tau, T, t) \\
 &+ E_A(\tau, T, t) + E_B(T, t) + E_C(t).
 \end{aligned} \quad (2)$$

Most of the fully phase-resolved 2D collinear THz experiments reported so far were performed with just two THz pulses ([Kuehn *et al.*, 2009, 2011a,b](#); [Junginger *et al.*, 2012](#); [Woerner *et al.*, 2013](#); [Bowlan *et al.*, 2014a,b](#); [Folpini *et al.*, 2015](#); [Somma *et al.*, 2014](#)). A special version with two pulses are experiments with an intense THz transient $E_{THz}(\tau, t)$ and a weaker electric field transient in the midinfrared spectral range $E_{MIR}(t)$ ([Gaal *et al.*, 2006](#); [Folpini *et al.*, 2015](#)). In the two-pulse case the nonlinear signal has a simpler expression:

$$E_{NL}(\tau, t) = E_{THz+MIR}(\tau, t) - E_{THz}(\tau, t) - E_{MIR}(t). \quad (3)$$

THz-Driven AC Stark Effect of Intersubband Transitions in Semiconductor Quantum Wells

So far, nearly all studies based on 2D spectroscopy have concentrated on resonant interactions between light and the system under study, i.e., a sequence of pulses interacts with single- and/or multi-photon resonances. The application of distinctly off-resonant electric field transients for controlling resonant optical excitations has remained scarce ([Wagner *et al.*, 2010](#)). Off-resonant coherent control requires a high amplitude of the off-resonant field with frequency (ν) to make the interaction energy with the system comparable to the transition energy of the optical resonance.

A basic concept for describing off-resonant coherent control is the so called dressed-state picture developed by [Autler and Townes \(1955\)](#) and [Bonch-Bruевич and Khodovoi \(1967\)](#). In this picture, one considers not just the electronic states but combined states of electron and light, e.g., the state consisting of the electron in level 1 and one photon with energy $h\nu$, $|n=1, 1 h\nu\rangle$ (see [Fig. 1\(a\)](#) of [Folpini *et al.* \(2015\)](#)). Because of the interaction with the electric field, the states $|n=1, 1 h\nu\rangle$ and $|n=2, 0 h\nu\rangle$ repel each other, leading to a blueshift of the transition frequency for $\nu < \nu_0$ and to a redshift for $\nu > \nu_0$ (ν_0 is the unperturbed transition frequency between states $n=1$ and $n=2$). We show in the following how an off-resonant perturbation of a midinfrared dipole transition by a strong THz field is mapped by 2D spectroscopy.

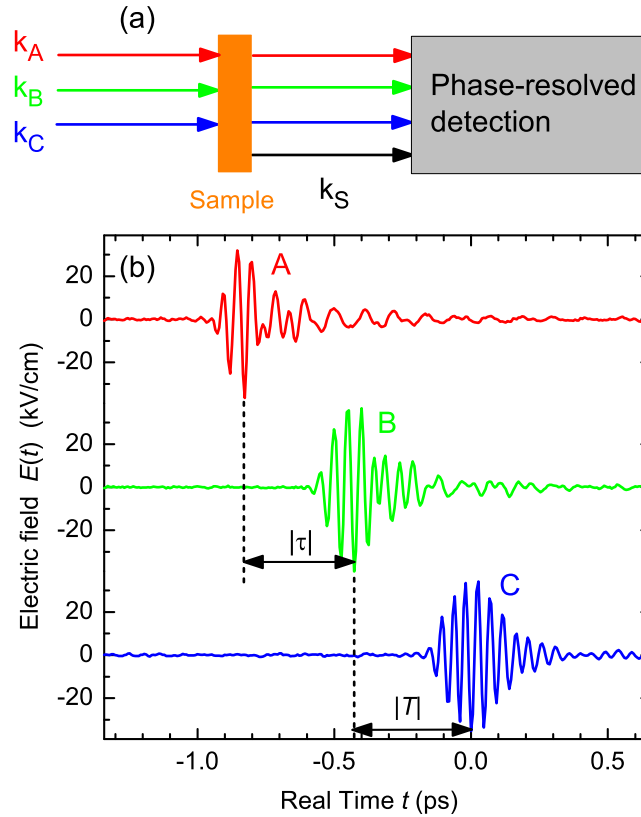


Fig. 3 (a) Schematic of the collinear interaction geometry in 2D terahertz (THz) spectroscopy. The electric field transmitted through the sample is measured in amplitude and phase by electrooptic sampling. (b) Sequence of three phase locked THz pulses: A and B are separated by the coherence time $\tau = -0.4$ ps and B and C are separated by the waiting time $T = -0.43$ ps. The electric field is plotted as a function of the real time t .

The model system under study is the optical transition between different confined subbands in a semiconductor quantum well (QW), a fundamental elementary excitation of free electrons confined in two dimensions (Helm, 2000). Such intersubband (IS) excitations represent sensitive probes of the complex many-body and scattering behavior of electrons and have been studied by both experiment and theory (Luo *et al.*, 2004; Delteil *et al.*, 2012; Wagner *et al.*, 2011; Dietze *et al.*, 2012; Shih *et al.*, 2005). In the present experiments, a sample grown by molecular beam epitaxy on an insulating GaAs substrate was investigated. It contains 20 GaAs QWs of $L = 11$ nm width, separated by 20 nm wide $\text{Al}_{0.35}\text{Ga}_{0.65}\text{As}$ barriers. The barriers are δ -doped to achieve an electron concentration of $5 \times 10^{11} \text{ cm}^{-2}$ per QW. As the thickness of the QW stack is small compared to both THz and MIR wavelengths, all QWs experience a similar electric field (Stroucken *et al.*, 1996). The $1 \rightarrow 2$ IS absorption is centered at a frequency of $\nu_0 = 21$ THz.

In the experiment, the nonresonant THz field manipulates the linear IS polarization, an effect manifested in a well-defined phase shift of the coherent IS emission. Fig. 4 shows the time-domain two-color 2D signal from the semiconductor quantum well sample interacting with a THz pulse centered at 1.2 THz and a mid-infrared pulse centered at 21 THz. In panel (a) the sum of the electric fields $E_{\text{THz}}(\tau, t) + E_{\text{MIR}}(t)$ transmitted through the sample is plotted as a function of the real time t and the delay or coherence time τ . Panel (b) displays the nonlinear signal $E_{\text{NL}}(t, \tau)$ in the mid-infrared according to Eq. (3). The dashed lines mark the centers of the two pulses in time and $\tau = 0$. A prototype electric field transient $E_{\text{THz}}(\tau = 0, t) + E_{\text{MIR}}(t)$ in the pulse overlap is shown in panel (c). The corresponding nonlinearly emitted field $E_{\text{NL}}(\tau = 0, t)$ is shown in panel (d).

The nonlinear signal represents, in principle, a response up to arbitrary order in the incident fields. For the present, comparably small THz field amplitude, however, the nonlinear signal is well accounted for by considering the third-order response. The signal occurs predominantly at positive values of τ and mainly after the MIR pulse at long t . This time structure gives direct evidence of a perturbed free induction decay (PFID) (Chachisvilis *et al.*, 1995). To separate the different nonlinear contributions, we derive the 2D spectrum $E_{\text{NL}}(\nu_i, \nu_\tau)$ as a function of ν_i and ν_τ by a two-dimensional Fourier transform (Kuehn *et al.*, 2009, 2011a). As shown in Fig. 5(b), only a THz-pump–MIR-probe signal E_{NL} at the frequency position $(\nu_i = \nu_0, \nu_\tau = 0)$ is present while MIR-pump–THz-probe and four-wave-mixing signals are absent. For a higher signal-to-noise ratio, $E_{\text{NL}}(\nu_i, \nu_\tau)$ is filtered to keep only the THz-pump–MIR-probe component and then back transformed to the time domain as shown in Fig. 4(d). Using the latter nonlinear transient one can calculate the measured spectrally resolved THz-pump–MIR-probe signal as a function of ν_i and pump–probe delay τ , which is plotted in Fig. 5(c). The dispersive lineshape shows clearly a THz-field-induced blue shift of the intersubband transition. Panel (d) displays the spectrally resolved pump–probe signal calculated with help of perturbation theory which will be discussed next.

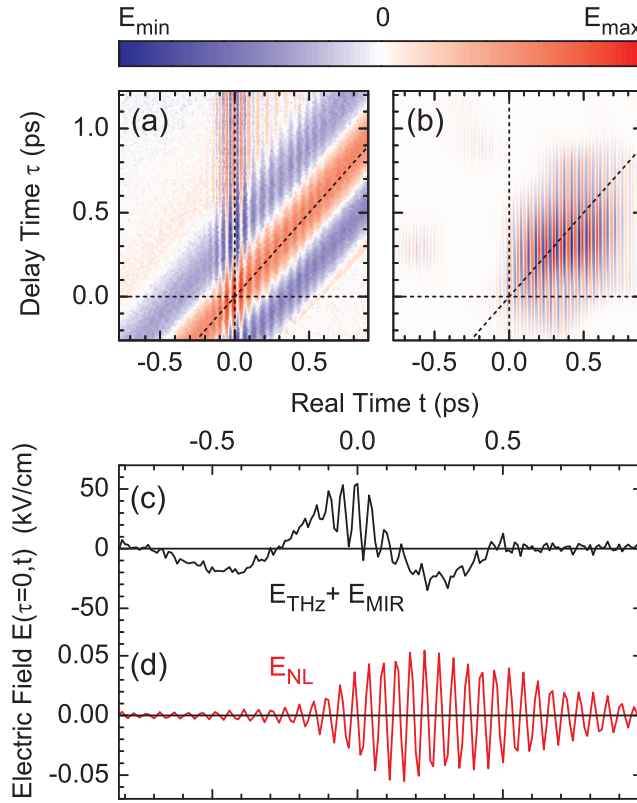


Fig. 4 Two-color 2D signal from a semiconductor quantum well sample interacting with a THz pulse centered at 1.2 THz and a mid-infrared pulse centered at 21 THz. (a) The sum of the electric fields $E_{\text{THz}}(\tau, t) + E_{\text{MIR}}(t)$ transmitted through the sample is plotted as a function of the real time t and the delay τ . (b) Nonlinear signal $E_{\text{NL}}(t, \tau)$ in the mid-infrared. The dashed lines show the centers of the two pulses and $\tau=0$. (c) Electric field transient $E_{\text{THz}}(\tau=0, t) + E_{\text{MIR}}(t)$. (d) The corresponding nonlinearly emitted field $E_{\text{NL}}(\tau=0, t)$ shows a perturbed free induction decay of the intersubband polarization as a function of real time t .

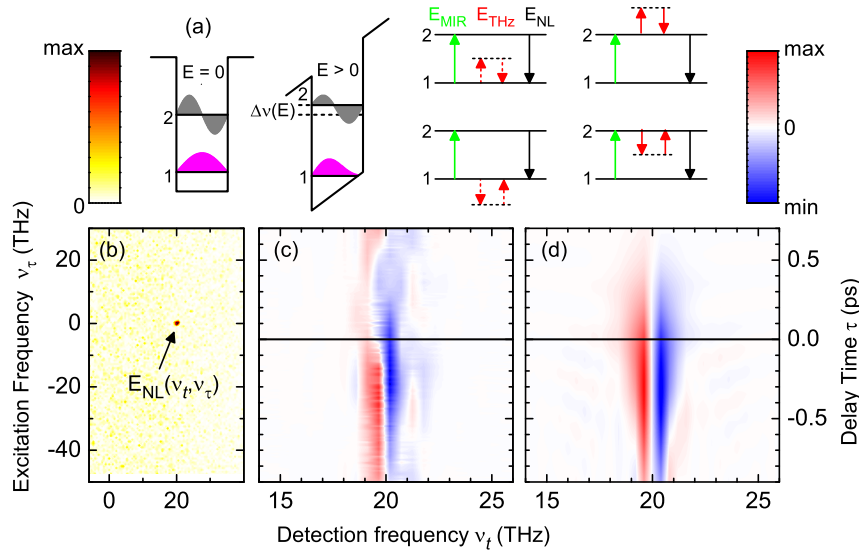


Fig. 5 (a) Schematic of the potential and wave functions of the quantum well without (left) and with (middle) applied THz field. For clarity, the electric field and the level shifts shown are much larger than in the actual experiment. (b) Right: lowest order (i.e. $\chi^{(3)}$) Liouville pathway diagrams in the style of Pollard *et al.* (1992) describing the AC Stark effect on intersubband transitions. (b) A 2D Fourier transform of the nonlinear signal shown in Fig. 4(b) gives the two-dimensional spectral density $E(v_t, v_t)$ as a function of the detection frequency v_t and the excitation frequency v_t . (c) Measured spectrally resolved THz-pump-MIR-probe signal as a function of v_t and pump-probe delay τ . (d) Calculated spectrally resolved pump-probe signal corresponding to the experiment shown in (c).

The electronic levels are determined by the confinement potential and wave functions of the quantum well without the applied THz field (lhs of Fig. 5(a)). In contrast to the dressed state picture, i.e., a modified QW potential (middle of Fig. 5(a)), perturbation theory treats the interaction of the intersubband transition with both the THz and the MIR field on equal footing. The rhs of Fig. 5(a) shows the relevant lowest order, i.e., $\chi^{(3)}$ Liouville pathway diagrams in the style of Pollard *et al.* (1992). The short arrows describe interactions with the nonresonant THz field while the levels and the long arrows stand for IS excitations. The off-resonant interaction with the THz pulse allows for sequences in which first a THz photon is emitted and later reabsorbed by the system. Such ‘acausal’ interaction sequences are important when stepping through ‘virtual’ states. All the diagrams shown are non-rephasing as the THz field cannot revert the phase of the IS coherence. It is important to note that such nonrephasing diagrams do not survive the rotating wave approximation which is the basis of standard analysis in 2D spectroscopy (Hamm and Zanni, 2011). Moreover, the nonrephasing diagrams shown on the rhs of Fig. 5(a) do not have any rephasing counterparts. As a consequence, it is impossible to construct a purely absorptive 2D spectrum as the sum of rephasing and nonrephasing diagrams (Hamm and Zanni, 2011), whenever off-resonant interaction sequences in Liouville space dominate.

Three-Pulse 2D THz Spectroscopy on InSb

Off-resonant interaction sequences also dominate in our second example, in which ‘dark’ two-photon coherences in the semiconductor InSb are investigated. InSb is the III-V semiconductor with the smallest band gap, which has a value at room temperature of $E_g = 0.17$ eV corresponding to a frequency of $\nu_{\text{THz}} = E_g/h = 41$ THz. The well known electronic band structure of InSb (Kane, 1957) is shown in Fig. 5(a) of Somma *et al.* (2016a). The THz pulses of Fig. 3(b) with a center frequency of 21 THz are neither resonant to the bandgap nor to any other resonance in the unexcited crystal. The THz peak field strength $E \approx 50$ kV/cm and the extraordinarily large interband transition dipole $d_{cv} = e_0 \cdot (4 \text{ nm})$ at the Γ point (e_0 : elementary charge) (Kane, 1957, 1959) result in a Rabi frequency $\Omega_{\text{Rabi}} = E d_{cv}/\hbar = 3 \times 10^{13} \text{ s}^{-1}$, i.e., comparable to the THz carrier frequency. As a result, the 2D experiments presented in the following are in the strongly nonperturbative regime of light-matter interaction, which is characterized by the simultaneous occurrence of nonlinear polarizations of different orders $\chi^{(n)}$, similar to Rabi oscillations on intersubband transitions in quantum wells (Kuehn *et al.*, 2009; Folpini *et al.*, 2015). In particular, the nonperturbative regime allows for multiple interactions of a single THz transient with the InSb crystal. In Somma *et al.* (2016a,b) we reported the ultrafast dynamics of two-phonon coherences and signals related to multiple two-photon excitations of electron-hole pairs in the III-V semiconductor. In the following, we concentrate on the dynamics of ‘dark’ two-photon coherences in InSb, studying a 70 μm thick (100)-oriented InSb single crystal with a low n -type doping ($n \leq 10^{16} \text{ cm}^{-3}$). All measurements were performed at ambient temperature (300 K).

Results of the 2D THz experiment performed with the three pulses A, B, and C of Fig. 3(b) are summarized in Fig. 6. In panel (a), the sum of electric field transients $E_A(\tau, T, t) + E_B(T, t) + E_C(t)$ transmitted through the InSb sample is shown in a contour plot as

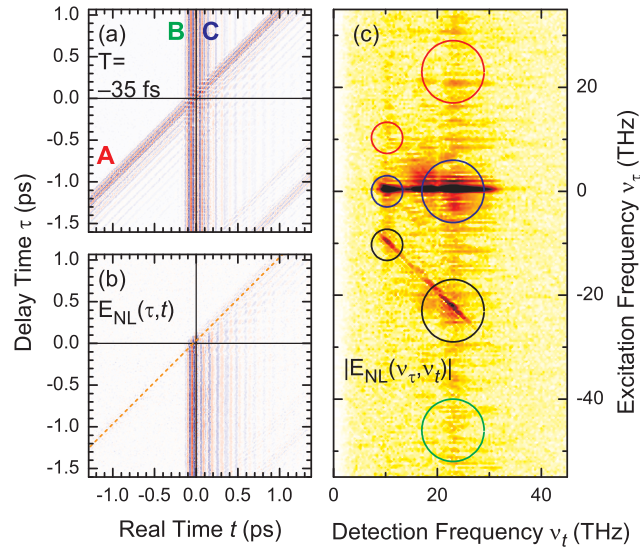


Fig. 6 Two-dimensional spectroscopy on InSb using the three THz pulses shown in Fig. 3(b). (a) Contour plot of the sum of electric field transients $E_A(\tau, T, t) + E_B(T, t) + E_C(t)$ transmitted through the InSb sample as a function of the coherence time τ and real time t for the waiting time $T = -35$ fs. (b) Nonlinear signal $E_{\text{NL}}(\tau, T, t)$ according to Eq. (2). The orange dashed line indicates the center of pulse A. (c) Contour plot of the amplitude $|E_{\text{NL}}(\nu_\tau, \nu_t)|$ which is the 2D Fourier transform of $E_{\text{NL}}(\tau, T, t)$. The colored circles indicate the position of relevant signals in the 2D frequency space. The linear amplitude scales of the 2D scans range over (a) ± 76.5 kV/cm and (b) ± 30 kV/cm.

a function of the coherence time τ and real time t for the waiting time $T = -35$ fs. In the following all contour plots are normalized to their respective maximum signals. The corresponding nonlinear signal $E_{\text{NL}}(\tau, T, t)$ derived with the help of Eq. (2) is shown in panel (b). In accordance with the law of causality a nonvanishing $E_{\text{NL}}(\tau, T, t)$ is exclusively observed starting with or after the last pulse in the timing sequence. For better orientation, the center of pulse A is indicated by the orange dashed line, which intersects the horizontal $\tau = 0$ line (black) at $t = -T$. In Fig. 6(c) the contour plot of the amplitude $|E_{\text{NL}}(\nu_\tau, \nu_t)|$ is displayed, obtained by a 2D Fourier transform of $E_{\text{NL}}(\tau, T = -35 \text{ fs}, t)$.

The circles of different size and color indicate the positions of relevant signals in the 2D frequency space. Concerning the detection frequency ν_t , strong nonlinear signals occur in the spectral range of the driving pulses $15 \text{ THz} < \nu_t < 25 \text{ THz}$ (large circles) and at the two-phonon resonance $\nu_t = 10 \text{ THz}$ (small circles). The latter signals have been discussed in detail in Somma *et al.* (2016a). As a function of the excitation frequency ν_τ we observe significant nonlinear signals for $\nu_\tau = 0$ (blue circles), $\nu_\tau = -\nu_t$ (black circles), $\nu_\tau = +\nu_t$ (red circles), and $\nu_\tau = -2\nu_t$ (green circle). The discussion in the following focuses on the signal in the green circle. It should be emphasized that the fully phase-resolved detection of electric field transients allows for a highly accurate measurement of the waiting time T . This knowledge becomes important in situations in which coherences evolve along the waiting period, the phase of which is determined by the exact value of T . For instance, recent 2D experiments on dipole-dipole interactions in a rubidium gas were detected via the time evolution of the two-atom coherence evolving along the waiting period T (Dai *et al.*, 2012; Gao *et al.*, 2016). The ‘dark’ two-photon coherence excited via two-photon excitation of InSb is measured in a similar scheme.

Fig. 7(a) and (b) shows the time domain nonlinear signal derived from the peak at $(\nu_\tau = -44 \text{ THz}, \nu_t = +22 \text{ THz})$ (green circle in Fig. 6(c)) by Gaussian frequency filtering and Fourier back-transform. The corresponding results for the red, blue, and black circles at frequency vectors $(\nu_\tau, \nu_t = 22 \text{ THz})$ have been presented and discussed in Fig. 7 of Somma *et al.* (2016b), whereas data for $(\nu_\tau, \nu_t = 10 \text{ THz})$ have been discussed in detail in Somma *et al.* (2016a).

The orange dashed lines in Fig. 7(a) and (b) indicate the center of pulse A (waiting time $T = -35$ fs) with wavefronts parallel to the $t = \tau$ diagonal. In contrast, the nonlinear signal shows tilted phase fronts perpendicular to the frequency vector $(\nu_\tau = -44 \text{ THz}, \nu_t = +22 \text{ THz})$ (black dashed arrow in Fig. 7(b)). This signal is different from the $\chi^{(3)}$ signals typically observed in 2D spectroscopy (cf. Fig. 1(b) and Table 1). In the set of frequency vectors (Eq. (1)), $\vec{\nu}_A$ is the only one with a nonvanishing $\nu_\tau = -22 \text{ THz}$ component. Thus, generation of a nonlinear signal at $\vec{\nu}_S = (\nu_\tau = -44 \text{ THz}, \nu_t = +22 \text{ THz})$ (dashed vector in Fig. 7(b) and (d)) requires at least two interactions of pulse A with the InSb sample. This fact excludes a $\chi^{(3)}$ scenario, because each pulse has interacted at least once with the sample, i.e., there is a minimum number of four interactions. On the other hand, the ν_t component of $\vec{\nu}_S$ is consistent with an odd total number of interactions only. Thus, the lowest order Liouville pathway corresponds at least to a $\chi^{(5)}$ susceptibility. The fact that the THz pulses are off-resonant to any dipole-allowed transition from the ground state of the sample and the observation that the electric field emitted by the nonlinear polarization is extremely short as a function of t (Fig. 7(a) and (b)), point to a scenario in which the respective last pulse in the interaction sequence interacts with a very broad band excited absorption of the crystal.

A $\chi^{(5)}$ scenario compatible with all experimental observations is represented by the Liouville space pathway shown in Fig. 7(d) and (e). Two interactions with pulse A create a two-photon interband coherence in the crystal which evolves freely along the coherence time $t_A - t_B = -\tau$. Pulse B which overlaps in time with pulse A ($\tau = 0$), completes the two-photon absorption event by two more interactions (Pidgeon *et al.*, 1979; Miller *et al.*, 1979). Finally, a single interaction with pulse C probes the broadband free carrier absorption in the conduction band of InSb. The time evolution of the two-photon interband coherence is plotted as a

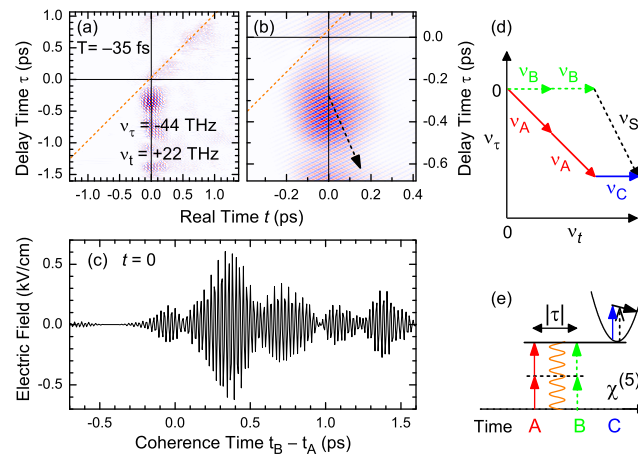


Fig. 7 (a), (b) Nonlinear signal for $T = -35$ fs at the frequency vector $(\nu_\tau = -44 \text{ THz}, \nu_t = 22 \text{ THz})$ (green circle in Fig. 6(c)) as a function of the coherence time $t_A - t_B = -\tau$ and real time t . Panel (b) shows an enlarged view for small τ and t values. The tilted phase fronts are perpendicular to the direction of the frequency vector (black dashed arrow). (c) The cut along the ($t = 0$) line of panel (a) shows directly the transient of the ‘dark’ two-photon interband coherence as a function of the coherence time $-\tau$. (d) Corresponding Liouville pathway diagram exploiting frequency vectors in 2D frequency space (cf. Fig. 1(e)). (e) Corresponding Liouville pathway diagram in the style of Pollard *et al.* (1992).

function of the coherence time $-\tau$ in Fig. 7(c). The fifth-order nonlinear signal shows quite a large amplitude (500 V/cm) and decays on a surprisingly long time scale. So far, there is no theoretical analysis of the microscopic dephasing mechanisms of two-photon coherences in semiconductors. One may speculate that reduced carrier-carrier scattering rates result in a surprisingly long decoherence time of the two-photon interband polarization. Four-wave-mixing experiments on semiconductor nanostructures (Kim *et al.*, 1992) have shown that electron-electron scattering is inhibited at the Fermi edge of the electron distribution, connected with the so-called Fermi edge singularity.

Conclusions

Multidimensional THz spectroscopy, a novel field of ultrafast science, has made rapid progress in recent years. The overview presented here combines an account of technological developments with selected prototype results for solids and nanostructures. Three-pulse 2D spectroscopy has been implemented in a collinear propagation geometry of three THz pulses and with fully phase-resolved detection of electric fields by electrooptic sampling. A key concept for analyzing 2D THz spectra is the introduction of frequency vectors in the multi-dimensional frequency space, which allow for identifying the dominating quantum pathways in Liouville space. Typical 2D experiments in the THz frequency range involve both off-resonant and multiple (i.e., beyond $\chi^{(3)}$) interaction sequences entering easily the nonperturbative regime of light-matter interaction. Two prototype experiments on intersubband coherences in quantum wells and the ‘dark’ two-photon coherence in the narrow gap semiconductor InSb illustrate the potential of multidimensional THz methods.

The generation of complex multicolor pulse sequences represents an important topic of current and future work. Another key issue is the generation of THz pulses of higher energy to study comparably weak dipole transitions such as intermolecular vibrations, which play a key role in the structural dynamics of liquids and other disordered systems. Such developments will broaden the range of nonlinear interactions to be studied and establish 2D THz techniques as a versatile tool, approaching the variability of multidimensional NMR methods.

See also: Coherent Terahertz Sources. Terahertz Detectors

References

- Asplund, M.C., Zanni, M.T., Hochstrasser, R.M., 2000. Two-dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. *Proc. Natl. Acad. Sci. USA* 97, 8219–8224. doi:10.1073/pnas.140227997.
- Autler, S.H., Townes, C.H., 1955. Stark effect in rapidly varying fields. *Phys. Rev.* 100, 703–722. doi:10.1103/PhysRev.100.703.
- Bonch-Bruевич, A.M., Khodovoi, V.A., 1967. Current methods for the study of the Stark effect in atoms. *Usp. Fiz. Nauk* 93, 71–110. doi:10.1070/PU1968v010n05ABEH005850.
- Bowlan, P., Martinez-Moreno, E., Reimann, K., Elsaesser, T., Woerner, M., 2014a. Ultrafast terahertz response of multi-layer graphene in the nonperturbative regime. *Phys. Rev. B* 89, 041408(R). doi:10.1103/PhysRevB.89.041408.
- Bowlan, P., Martinez-Moreno, E., Reimann, K., Elsaesser, T., Woerner, M., 2014b. Terahertz radiative coupling and damping in multilayer graphene. *New J. Phys.* 16, 013027. doi:10.1088/1367-2630/16/1/013027.
- Boyd, R.W., Mukamel, S., 1984. Origin of spectral holes in pump-probe studies of homogeneously broadened lines. *Phys. Rev. A* 29, 1973–1983. doi:10.1103/PhysRevA.29.1973.
- Brodschelm, A., Tauser, F., Huber, R., Sohn, J.Y., Leitenstorfer, A., 2001. Amplitude and phase resolved detection of tunable femtosecond pulses with frequency components beyond 100 THz. In: Elsaesser, T., Mukamel, S., Murnane, M.M., Scherer, N.F. (Eds.), *Ultrafast Phenomena XII*. Berlin: Springer, pp. 215–217.
- Chachisvilis, M., Fidler, H., Sundström, V., 1995. Electronic coherence in pseudo two-colour pump-probe spectroscopy. *Chem. Phys. Lett.* 234, 141–150. doi:10.1016/0009-2614(95)00041-2.
- Dai, X., Richter, M., Li, H., *et al.*, 2012. Two-dimensional double-quantum spectra reveal collective resonances in an atomic vapor. *Phys. Rev. Lett.* 108, 193201. doi:10.1103/PhysRevLett.108.193201.
- Delteil, A., Vasanelli, A., Todorov, Y., *et al.*, 2012. Charge-induced coherence between intersubband plasmons in a quantum structure. *Phys. Rev. Lett.* 109, 246808. doi:10.1103/PhysRevLett.109.246808.
- Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. *Opt. Express* 20, 23053. doi:10.1364/OE.20.023053.
- Ernst, R.R., Bodenhausen, G., Wokaun, A., 1997. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford: Clarendon Press.
- Folpini, G., Morrill, D., Somma, C., *et al.*, 2015. Nonresonant coherent control-intersubband excitations manipulated by a nonresonant terahertz pulse. *Phys. Rev. B* 92, 085306. doi:10.1103/PhysRevB.92.085306.
- Gaal, P., Kuehn, W., Reimann, K., *et al.*, 2007. Internal motions of a quasiparticle governing its ultrafast nonlinear response. *Nature* 450, 1210–1213. doi:10.1038/nature06399.
- Gaal, P., Reimann, K., Woerner, M., *et al.*, 2006. Nonlinear terahertz response of *n*-type GaAs. *Phys. Rev. Lett.* 96, 187402. doi:10.1103/PhysRevLett.96.187402.
- Gao, F., Cundiff, S.T., Li, H., 2016. Probing dipole-dipole interaction in a rubidium gas via double-quantum 2D spectroscopy. *Opt. Lett.* 41, 2954–2957. doi:10.1364/OL.41.002954.
- Günter, G., Anappara, A.A., Hees, J., *et al.*, 2009. Sub-cycle switch-on of ultrastrong light-matter interaction. *Nature* 458, 178–181. doi:10.1038/nature07838.
- Hamm, P., Lim, M., Hochstrasser, R.M., 1998. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *J. Phys. Chem. B* 102, 6123–6138. doi:10.1021/jp9813286.
- Hamm, P., Zanni, M., 2011. *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge: Cambridge University Press.
- Helm, M., 2000. The basic physics of intersubband transitions. In: Liu, H.C., Capasso, F. (Eds.), *Intersubband Transitions in Quantum Wells: Physics and Device Applications*. Academic Press, pp. 1–100.
- Huber, R., Brodschelm, A., Tauser, F., Leitenstorfer, A., 2000. Generation and field-resolved detection of femtosecond electromagnetic pulses tunable up to 41 THz. *Appl. Phys. Lett.* 76, 3191–3193. doi:10.1063/1.126625.

- Huber, R., Kübler, C., Tübel, S., *et al.*, 2005. Femtosecond formation of coupled phonon-plasmon modes in InP: ultrabroadband THz experiment and quantum kinetic theory. *Phys. Rev. Lett.* 94, 027401. doi:10.1103/PhysRevLett.94.027401.
- Huber, R., Tauser, F., Brodschelm, A., *et al.*, 2001. How many-particle interactions develop after ultrafast excitation of an electron-hole plasma. *Nature* 414, 286–289. doi:10.1038/35104522.
- Hwang, H.Y., Fleischer, S., Brandt, N.C., *et al.*, 2015. A review of non-linear terahertz spectroscopy with ultrashort tabletop-laser pulses. *J. Mod. Opt.* 62, 1447–1479. doi:10.1080/09500340.2014.918200.
- Jonas, D.M., 2003. Two-dimensional femtosecond spectroscopy. *Annu. Rev. Phys. Chem.* 54, 425–463. doi:10.1146/annurev.physchem.54.011002.103907.
- Junginger, F., Mayer, B., Schmidt, C., *et al.*, 2012. Nonperturbative interband response of a bulk InSb semiconductor driven off resonantly by terahertz electromagnetic few-cycle pulses. *Phys. Rev. Lett.* 109, 147403. doi:10.1103/PhysRevLett.109.147403.
- Junginger, F., Sell, A., Schubert, O., *et al.*, 2010. Single-cycle multiterahertz transients with peak fields above 10 MV/cm. *Opt. Lett.* 35, 2645–2647. doi:10.1364/OL.35.002645.
- Kampfrath, T., Sell, A., Klatt, G., *et al.*, 2011. Coherent terahertz control of antiferromagnetic spin waves. *Nature Photon.* 5, 31–34. doi:10.1038/nphoton.2010.259.
- Kane, E.O., 1957. Band structure of indium antimonide. *J. Phys. Chem. Solids* 1, 249–261. doi:10.1016/0022-3697(57)90013-6.
- Kane, E.O., 1959. Zener tunneling in semiconductors. *J. Phys. Chem. Solids* 12, 181–188. doi:10.1016/0022-3697(60)90035-4.
- Khalil, M., Demirdöven, N., Tokmakoff, A., 2003. Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. *J. Phys. Chem. A* 107, 5258–5279. doi:10.1021/jp0219247.
- Kim, D.-S., Shah, J., Cunningham, J.E., *et al.*, 1992. Carrier-carrier scattering in a degenerate electron system: strong inhibition of scattering near the Fermi edge. *Phys. Rev. Lett.* 68, 2838–2841. doi:10.1103/PhysRevLett.68.2838.
- Kübler, C., Huber, R., Leitenstorfer, A., 2005. Ultrabroadband terahertz pulses: generation and field-resolved detection. *Semicond. Sci. Technol.* 20, S128–S133. doi:10.1088/0268-1242/20/7/002.
- Kübler, C., Huber, R., Tübel, S., Leitenstorfer, A., 2004. Ultrabroadband detection of multi-terahertz field transients with GaSe electro-optic sensors: approaching the near infrared. *Appl. Phys. Lett.* 85, 3360–3362. doi:10.1063/1.1808232.
- Kuehn, W., Reimann, K., Woerner, M., Elsaesser, T., 2009. Phase-resolved two-dimensional spectroscopy based on collinear *n*-wave mixing in the ultrafast time domain. *J. Chem. Phys.* 130, 164503. doi:10.1063/1.3120766.
- Kuehn, W., Reimann, K., Woerner, M., Elsaesser, T., Hey, R., 2011a. Two-dimensional terahertz correlation spectra of electronic excitations in semiconductor quantum wells. *J. Phys. Chem. B* 115, 5448–5455. doi:10.1021/jp1099046.
- Kuehn, W., Reimann, K., Woerner, M., *et al.*, 2011b. Strong correlation of electronic and lattice excitations in GaAs/AlGaAs semiconductor quantum wells revealed by two-dimensional terahertz spectroscopy. *Phys. Rev. Lett.* 107, 067401. doi:10.1103/PhysRevLett.107.067401.
- Leitenstorfer, A., Hunsche, S., Shah, J., Nuss, M.C., Knox, W.H., 1999. Detectors and sources for ultrabroadband electro-optic sampling: experiment and theory. *Appl. Phys. Lett.* 74, 1516–1518. doi:10.1063/1.123601.
- Luo, C.W., Reimann, K., Woerner, M., *et al.*, 2004. Phase-resolved nonlinear response of a two-dimensional electron gas under femtosecond intersubband excitation. *Phys. Rev. Lett.* 92, 047402. doi:10.1103/PhysRevLett.92.047402.
- Miller, A., Johnston, A., Dempsey, J., *et al.*, 1979. Two-photon absorption in InSb and Hg_{1-x}Cd_xTe. *J. Phys. C* 12, 4839–4849. doi:10.1088/0022-3719/12/22/025.
- Mukamel, S., 1995. *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.
- Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. *Annu. Rev. Phys. Chem.* 51, 691–729. doi:10.1146/annurev.physchem.51.1.691.
- Pashkin, A., Kübler, C., Ehrke, H., *et al.*, 2011. Ultrafast insulator-metal phase transition in VO₂ studied by multiterahertz spectroscopy. *Phys. Rev. B* 83, 195120. doi:10.1103/PhysRevB.83.195120.
- Pidgeon, C.R., Wherrett, B.S., Johnston, A.M., Dempsey, J., Miller, A., 1979. Two-photon absorption in zinc-blende semiconductors. *Phys. Rev. Lett.* 42, 1785–1788. doi:10.1103/PhysRevLett.42.1785.
- Pollard, W.T., Dexheimer, S.L., Wang, Q., *et al.*, 1992. Theory of dynamic absorption spectroscopy of nonstationary states. 4. Application to 12-fs resonant impulsive Raman spectroscopy of bacteriorhodopsin. *J. Phys. Chem.* 96, 6147–6158. doi:10.1021/j100194a013.
- Reimann, K., Smith, R.P., Weiner, A.M., Elsaesser, T., Woerner, M., 2003. Direct field-resolved detection of terahertz transients with amplitudes of megavolts per centimeter. *Opt. Lett.* 28, 471–473. doi:10.1364/OL.28.000471.
- Schubert, O., Riek, C., Junginger, F., *et al.*, 2011. Ultrashort pulse characterization with a terahertz streak camera. *Opt. Lett.* 36, 4458–4460. doi:10.1364/OL.36.004458.
- Sell, A., Leitenstorfer, A., Huber, R., 2008. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. *Opt. Lett.* 33, 2767–2769. doi:10.1364/OL.33.002767.
- Shih, T., Reimann, K., Woerner, M., *et al.*, 2005. Nonlinear response of radiatively coupled intersubband transitions of quasi-two-dimensional electrons. *Phys. Rev. B* 72, 195338. doi:10.1103/PhysRevB.72.195338.
- Somma, C., Folpini, G., Gupta, J., *et al.*, 2015. Ultra-broadband terahertz pulses generated in the organic crystal DSTMS. *Opt. Lett.* 40, 3404–3407. doi:10.1364/OL.40.003404.
- Somma, C., Folpini, G., Reimann, K., Woerner, M., Elsaesser, T., 2016a. Two-phonon quantum coherences in indium antimonide studied by nonlinear two-dimensional terahertz spectroscopy. *Phys. Rev. Lett.* 116, 177401. doi:10.1103/PhysRevLett.116.177401.
- Somma, C., Folpini, G., Reimann, K., Woerner, M., Elsaesser, T., 2016b. Phase-resolved two-dimensional terahertz spectroscopy including off-resonant interactions beyond the $\chi^{(3)}$ limit. *J. Chem. Phys.* 144, 184202. doi:10.1063/1.4948639.
- Somma, C., Reimann, K., Flytzanis, C., Elsaesser, T., Woerner, M., 2014. High-field terahertz bulk photovoltaic effect in lithium niobate. *Phys. Rev. Lett.* 112, 146602. doi:10.1103/PhysRevLett.112.146602.
- Stroucken, T., Knorr, A., Thomas, P., Koch, S.W., 1996. Coherent dynamics of radiatively coupled quantum-well excitons. *Phys. Rev. B* 53, 2026–2033. doi:10.1103/PhysRevB.53.2026.
- Tonouchi, M., 2007. Cutting-edge terahertz technology. *Nature Photon.* 1, 97–105. doi:10.1038/nphoton.2007.3.
- Wagner, M., Helm, M., Sherwin, M.S., Stehr, D., 2011. Coherent control of a THz intersubband polarization in a voltage controlled single quantum well. *Appl. Phys. Lett.* 99, 131109. doi:10.1063/1.3644988.
- Wagner, M., Schneider, H., Stehr, D., *et al.*, 2010. Observation of the intraexciton Autler-Townes effect in GaAs/AlGaAs semiconductor quantum wells. *Phys. Rev. Lett.* 105, 167401. doi:10.1103/PhysRevLett.105.167401.
- Woerner, M., Kuehn, W., Bowlan, P., Reimann, K., Elsaesser, T., 2013. Ultrafast two-dimensional terahertz spectroscopy of elementary excitations in solids. *New J. Phys.* 15, 025039. doi:10.1088/1367-2630/15/2/025039.
- Wu, Q., Zhang, X.-C., 1995. Free-space electro-optic sampling of terahertz beams. *Appl. Phys. Lett.* 67, 3523–3525. doi:10.1063/1.114909.
- Wu, Q., Zhang, X.-C., 1997. Free-space electro-optics sampling of mid-infrared pulses. *Appl. Phys. Lett.* 71, 1285–1286. doi:10.1063/1.119873.
- Yang, L., Zhang, T., Bristow, A.D., Cundiff, S.T., Mukamel, S., 2008. Isolating excitonic Raman coherence in semiconductors using two-dimensional correlation spectroscopy. *J. Chem. Phys.* 129, 234711. doi:10.1063/1.3037217.
- Yee, T.K., Gustafson, T.K., 1978. Diagrammatic analysis of the density operator for nonlinear optical calculations: pulsed and cw responses. *Phys. Rev. A* 18, 1597–1617. doi:10.1103/PhysRevA.18.1597.

Second Harmonic Generation Spectroscopy of Hidden Phases

Liuyan Zhao, University of Michigan, Ann Arbor, MI, United States

Darius Torchinsky, Temple University, Philadelphia, PA, United States

John Harter, University of California, Santa Barbara, CA, United States

Alberto de la Torre and David Hsieh, California Institute of Technology, Pasadena, CA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

A cornerstone in condensed matter research is the emergence of the concept of symmetry and symmetry breaking upon an order formation (Landau, 1936, 1937), an idea that an ordered state can only have a true subset of symmetries of its unordered counterpart. Being able to experimentally capture the evolution of symmetries across a phase transition of an order is essential for understanding its microscopic mechanism and its macroscopic mechanical, electronic and magnetic properties as well (Birss, 1964; Nye, 1985). Traditionally, X-ray (Warren, 1969), neutron (Squires, 2012) and electron (Zou *et al.*, 2011) diffraction techniques are widely used in determining the symmetries of lattice, magnetic and charge orders, while resonant X-ray diffraction has recently been exploited to probe symmetries of more exotic orders such as orbital (Murakami *et al.*, 1998) and multipolar orders (Kuramoto *et al.*, 2009; Santini *et al.*, 2009). However, the accuracy of symmetry space group assignment through the above diffraction techniques critically depends on the ability of performing a unique fit to the diffraction pattern. This is not always feasible because of technical challenges including the presence of elements with a large adsorption or weak cross-section, the limited crystal size smaller than the probe beam spot, the coexistence of multiple orders with different symmetries, or the small domain size of multiple degenerate states of an order.

SHG (Franken *et al.*, 1961), the lowest-order degenerate nonlinear optical generation, is an alternative to diffraction techniques for resolving the symmetries of a material through the structure of its second order nonlinear optical susceptibility tensor (Boyd, 2003; Shen, 1984). Of particular interest is its extreme sensitivity to inversion symmetry because the leading order electric dipole (ED) contribution to SHG is only allowed wherever inversion symmetry is broken such as at interfaces of heterostructures of centrosymmetric crystals, i.e., the optical susceptibility tensor for this ED SHG process is zero if inversion symmetry is present. To experimentally determine the structure of the SHG susceptibility tensor for a crystal, a SHG rotation anisotropy (RA) measurement is typically performed, in which the intensity of the SHG light generated from the crystal is recorded as the crystal rotates about its surface normal.

The application of SHG in determining symmetry point group dates back to 1980s for studying lattice reconstructions on semiconductor surfaces (Heinz *et al.*, 1985; Tom *et al.*, 1983; Yamada and Kimura, 1993) and metal-electrolyte interfaces (Shannon *et al.*, 1987a,b; Shen, 1989), extends later on for probing magnetic ordering on metal surfaces (Dähn *et al.*, 1996; Gridnev *et al.*, 2001; Kirilyuk and Rasing, 2005; Nývlt *et al.*, 2008; Pan *et al.*, 1989; Reif *et al.*, 1991), in bulk semiconductors (Lafrentz *et al.*, 2012) and multiferroics (Fiebig *et al.*, 1994, 2000, 2005; Kumar *et al.*, 2008), and recently to characterizing two-dimensional atomic crystals (Clark *et al.*, 2014; Janisch *et al.*, 2014; Kim and Lim, 2015; Kumar *et al.*, 2013; Li *et al.*, 2013; Malard *et al.*, 2013; Seyler *et al.*, 2015; Wu *et al.*, 2014; Yin *et al.*, 2014) and topological materials (Hamh *et al.*, 2016; Hsieh *et al.*, 2011; McIver *et al.*, 2012; Wu *et al.*, 2016). However, it is only after overcoming a number of technical challenges very recently that this technique has been introduced as an effective probe for hidden phases in correlated electron systems. The most outstanding one is the need for sensitivity to the entire SHG susceptibility tensor, which requires performing the SHG RA measurements in the oblique incidence reflection geometry. This, under the traditional rotating-sample-based scheme, is achievable at the room temperature despite a number of alignment difficulties, but was almost impossible at cryogenic temperatures where electronic phase transitions usually happen. A novel SHG RA design based on a rotation of the light scattering-plane instead of the sample was first reported in 2014 (Harter *et al.*, 2015; Torchinsky *et al.*, 2014), and has been successfully applied in a few systems to determine their lattice (Harter *et al.*, 2016; Hogan *et al.*, 2016; Torchinsky *et al.*, 2015) and electronic symmetries (Zhao *et al.*, 2016a,b; Harter *et al.*, 2017). Moreover, by keeping the sample stationary during the RA measurements, it allows to perform experiments not only at cryogenic temperatures, but also under static magnetic or strain fields.

This review is organized as follows. In Section “Second Harmonic Generation in Crystals”, we introduce the theoretical background of SHG and its relationship to the lattice and electronic symmetries of a crystal. In Section “Generations of Rotational Anisotropy Experimental Designs”, we describe the advances in the SHG RA designs with an emphasis on their capabilities and limitations. In Section “Hidden Phases in Correlated Electron Systems”, we discuss hidden phases in correlated electron systems and the challenges for their experimental detections. In Section “Hidden Phases Revealed by SHG-based Techniques”, we take examples of SHG-based techniques revealing hidden phases. Finally, in Section “Conclusions and Prospects”, we discuss the prospects of generalizing SHG to non-degenerate second order nonlinear optical generations in detecting novel electronic orders.

Second Harmonic Generation in Crystals

Electromagnetic waves of the incident light propagating through a medium can induce electric dipole ($\vec{P}$), magnetic dipole ($\vec{M}$), electric quadrupole ($\vec{Q}$), or even higher rank multipolar densities, all of which together compose the source term ($\vec{S}$; note that

this is not the Poynting vector) for the radiated light (Boyd, 2003; Fiebig *et al.*, 2005).

$$\vec{S} = \mu_0 \frac{\partial^2 \vec{P}}{\partial t^2} + \mu_0 \left(\nabla \times \frac{\partial \vec{M}}{\partial t} \right) - \mu_0 \left(\nabla \frac{\partial^2 \vec{Q}}{\partial t^2} \right) + \dots \quad (1)$$

The electric dipole term ($\propto \vec{P}$) is the leading order contribution to $\vec{S}$, and is typically $\sim \lambda/a$ times stronger than the second order magnetic dipole ($\propto \vec{M}$) and electric quadrupole ($\propto \vec{Q}$) terms, where λ and a are the wavelength of the light and the lattice constant of the crystal respectively. This is precisely the reason why linear optical processes can be well described by the electric dipole approximation.

Generally, the induced dipolar/multipolar densities can be expressed as linear, second and higher order nonlinear expansions of electric $\vec{E}(\omega)$ and magnetic $\vec{H}(\omega)$ fields of the incident light.

$$\vec{P}(\omega, 2\omega, \dots) \propto \chi^{pe} \vec{E}(\omega) + \chi^{pm} \vec{H}(\omega) + \chi^{pee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{pem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{pmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O} \left[\left(\vec{E}, \vec{H} \right)^3 \right] \quad (2)$$

$$\vec{M}(\omega, 2\omega, \dots) \propto \chi^{me} \vec{E}(\omega) + \chi^{mm} \vec{H}(\omega) + \chi^{mee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{mem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{mmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O} \left[\left(\vec{E}, \vec{H} \right)^3 \right] \quad (3)$$

$$\vec{Q}(\omega, 2\omega, \dots) \propto \chi^{qe} \vec{E}(\omega) + \chi^{qm} \vec{H}(\omega) + \chi^{qee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{qem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{qmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O} \left[\left(\vec{E}, \vec{H} \right)^3 \right] \quad (4)$$

The coefficients (χ) in Eqs. (2–4), i.e. optical susceptibility tensors, describe the optical processes happening in the crystal upon the driving fields, and reflects the physical properties of the crystal. For example, $\chi_{ijk}^{pee}(2\omega)$ characterizes the electric dipole SHG response in the crystal via two consecutive interactions with the incident electric field, where the first superscript denotes the electric dipole origin (p) of the induced source, the second and third superscripts denote the electric (e) nature of the driving field, and the subscripts denote the polarization components.

Neumann's principle dictates that property tensors, such as χ above, should remain invariant under transformations that respect the symmetries of the crystal (Birss, 1964). That is

$$\chi_{ijk\dots} = R_{ii'} R_{jj'} R_{kk'} \dots \chi_{i'j'k'\dots} \quad (5)$$

where $R_{ii'}$, $R_{jj'}$, $R_{kk'}$ are matrices for the symmetry transformations, and $\chi_{i'j'k'\dots}$ and $\chi_{ijk\dots}$ are the property tensors before and after the transformation respectively. This enforces a set of relationships among the tensor elements, and therefore significantly reduces the number of independent non-zero tensor elements. As a result, the symmetry properties of the crystal are encoded in the structure of the property tensors, and we can in principle determine the lattice and electronic symmetries of the crystal via measuring the property tensors. Linear responses with the lowest rank property tensors have limited symmetry sensitivity due to their degenerate structure for the symmetry point groups within the same crystal system, while nonlinear responses with higher rank tensors gain greater symmetry resolutions through different tensor forms for different point groups. The second order nonlinear optical responses, including SHG, sum frequency generation (SFG) and difference frequency generation (DFG), stand out in particular because of their extreme sensitivity to inversion symmetry, i.e., their leading order electric dipole contribution (χ_{ijk}^{pee}) is only allowed wherever inversion symmetry is broken such as surfaces and interfaces of centrosymmetric crystals. However, here we draw your attention to that the higher order multipolar contributions, such as magnetic dipole (χ_{ijk}^{mee}) and electric quadrupole (χ_{ijk}^{qee}), are present even under inversion symmetry, despite their much lower radiation efficiency. In this review, we will focus on SHG, the degenerate second order nonlinear optical response, and its recent applications in revealing hidden electronic phases in correlated condensed matter systems.

Quantum mechanical description of optical SHG susceptibility is expressed via equations, for example, in the electric dipole approximation (Boyd, 2003; Zhao *et al.*, 2016a),

$$\chi_{ijk}^{pee}(2\omega) \propto \sum_{\vec{g}} \sum_{s,m,f} \left[\frac{\langle \vec{g}, \vec{k} | P_i | f, \vec{k} \rangle \langle f, \vec{k} | P_j | m, \vec{k} \rangle \langle m, \vec{k} | P_k | g, \vec{k} \rangle}{(\hbar\omega_{fg} - 2\hbar\omega - i\gamma_{fg}) (\hbar\omega_{mg} - \hbar\omega - i\gamma_{mg})} + \dots \right] f_g(\vec{k}) \quad (6)$$

which involves a two-photon adsorption process driven by two consecutive electric dipole transitions at an energy of $\hbar\omega$ from the initial state $|g, \vec{k}\rangle$ to the intermediate state $|m, \vec{k}\rangle$ and from the intermediate state $|m, \vec{k}\rangle$ to the final state $|f, \vec{k}\rangle$ respectively, followed by a single-photon emission process driven by one electric dipole transition at an energy of $2\hbar\omega$ from $|f, \vec{k}\rangle$ back to $|g, \vec{k}\rangle$. The energy difference between the initial state and the intermediate (final) state is given by $\hbar\omega_{mg}$ ($\hbar\omega_{fg}$), and the damping rate for the transition between the initial and the intermediate (final) state is represented by γ_{mg} (γ_{fg}) with $\gamma_{m(f)g} = \frac{1}{2}(\tau_{m(f)}^{-1} + \tau_g^{-1})$ where τ_g and $\tau_{m(f)}$ are the lifetime for the initial and the intermediate (final) states respectively. $f_g(\vec{k})$ is the Fermi distribution function for the initial state $|g, \vec{k}\rangle$. Similar as linear optics, the dominant contributions to $\chi_{ijk}^{pee}(2\omega)$ are from the resonant optical transitions where $2\hbar\omega$ ($\hbar\omega$) matches the energy difference between the initial and final (intermediate) states, $\hbar\omega_{fg}$ ($\hbar\omega_{mg}$), and therefore $\chi_{ijk}^{pee}(2\omega)$ responds most when the resonant transition happens at energies with most spectra-weight transfer across a phase transition. This statement is generally true not only for $\chi_{ijk}^{pee}(2\omega)$, but also for any single optical process in Eqs. (2–4). More than linear optics, SHG has a better ability of probing the symmetry properties of the relevant states. Furthermore, SHG can be sensitive to inversion symmetry breaking phases even if the photon energy is off resonance with the most responsive electronic

states because of the stronger radiation efficiency of the activated leading order electric dipole contribution ($\chi_{ijk}^{pee}(2\omega)$) from the inversion symmetry breaking order.

The current understanding of SHG as a symmetry probe of electronic states is at a qualitative level where we only discuss if the SHG susceptibility tensor elements are zero or not so as to infer what symmetry operations are present in the system. Compared to linear optics, we lack comprehensive understanding on the amplitudes of SHG susceptibility tensor elements which in fact contain key information of the nature and the strength of the order parameters. Thereby, theoretical inputs in SHG, or more generally nonlinear optical generations, from novel electronic states will be greatly appreciated. In the following parts of this review, we will mainly discuss the experimental aspects of SHG from hidden electronic phases.

Generations of Rotational Anisotropy Experimental Designs

In addition to its sensitivity to inversion symmetry, SHG is capable of detecting rotational symmetries, mirror symmetries and time reversal symmetry by performing a RA measurement to access as many SHG susceptibility tensor elements as possible.

A SHG RA measurement is to record the SHG intensity ($I^{2\omega}$) as a function of the azimuthal angle ((φ)) about the crystal surface normal shown in Fig. 1. The incident light at an optical frequency of ω and the reflected light at a frequency of 2ω form a scattering plane. The angle between the scattering plane and the crystal axis a is defined as the azimuthal angle (φ). The polarizations of the incident and reflected light can be independently selected to be either within ($P_{in/out}$) or normal ($S_{in/out}$) to the scattering plane. In total, it gives four independent polarization geometries ($P/S_{in} - P/S_{out}$) that together access all the SHG susceptibility tensor elements and each select different combinations of the tensor elements.

Traditionally, the full RA measurement is obtained via rotating the crystal about its surface normal (Fig. 1), and has been successfully applied in studies of molecule self-assembly on crystal surfaces, semiconductors and transition metal oxide interfaces. However, this scheme has a number of alignment challenges including beam walking on the sample and precession of the reflected beam during the RA acquirement, and therefore it not only requires large crystal sizes, but also causes aberrations in the RA patterns. More importantly, the rotation of the crystal prohibits the measurements taken under various experimental conditions such as at cryogenic temperature, in magnetic fields, under high pressure and so on. In strongly correlated electron systems, temperature, magnetic field and pressure are precisely the experimental parameters that we tune to realize various electronic phases of matter.

To overcome the limitations on experimental conditions posed by the rotation of the crystal, a different design was introduced to keep the sample stationary during the RA measurements. As shown in Fig. 2(a), instead of oblique incidence in Fig. 1, it takes the normal incidence geometry. In this design, the polarizations of the incoming and the reflected/transmitted light can be selected to be either parallel or orthogonal to each other, and the RA data is taken via rotating the polarizations relative to the crystal axis a . As we see in Fig. 2(a), both polarizations are parallel to the crystal surface (xy plane in the lab coordinate), missing the component in the direction of the surface normal (z direction). According to the relationships in Eqs. (2–4), such as $P_i = \chi_{ijk}^{pee} E_j E_k$, this scheme lacks the accessibility to the SHG susceptibility tensor elements whose subscripts includes z , e.g., χ_{ijk}^{pee} elements with $i=z$ or $j=z$ or $k=z$. For the high symmetry systems where the inaccessible SHG susceptibility tensor elements are zero, this scheme works successfully, and we will see one such example on $\text{Cd}_2\text{Re}_2\text{O}_7$ (Petersen *et al.*, 2006) in Section “A Parity-broken Nematic Order in $\text{Cd}_2\text{Re}_2\text{O}_7$ ”. However, it faces limitations in the systems with lower symmetry where the inaccessible tensor elements are necessary for differentiating between symmetry point groups.

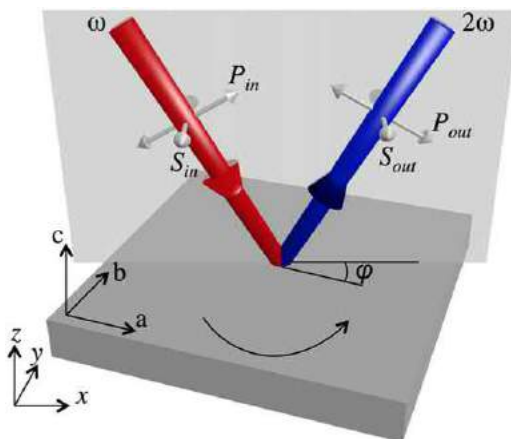


Fig. 1 Schematic of the rotating-sample-based SHG RA setup. The electric field polarizations of the obliquely incident fundamental beam (in) and outgoing SHG beam (out) can be independently selected to be either parallel (P) or perpendicular (S) to the light scattering plane (shaded vertical plane). SHG RA data are acquired by measuring the SHG intensity ($I^{2\omega}$) as a function of the angle (φ) between the fixed light scattering plane and the ac plane of the rotating sample. xyz is the lab coordinate and abc is the sample coordinate.

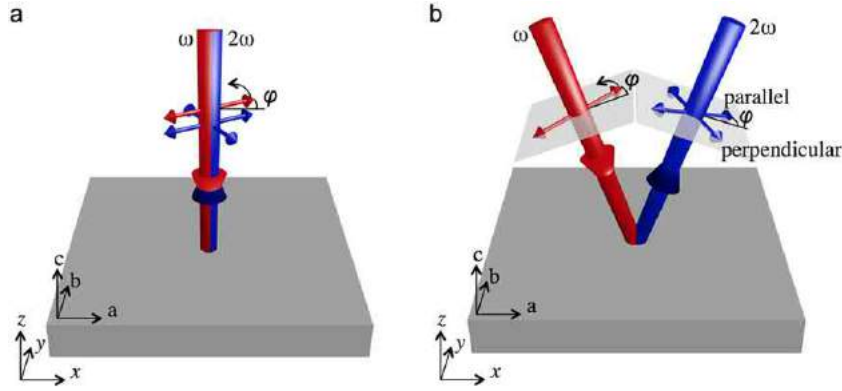


Fig. 2 Schematic of the rotating-polarization-based SHG RA setups. (a) In the normal incidence geometry, the polarizations of the incoming (ω) and outgoing (2ω) beams can be selected to be either parallel or perpendicular to each other. The SHG RA data is to record the SHG intensity $I^{2\omega}$ as a function of the angle (φ) between the a -axis and the polarization of the incoming light. (b) A derivative of (a) to the oblique incidence geometry. The polarizations of the incoming (ω) and outgoing (2ω) beams can be selected to be parallel, same angle (φ) between the polarization and the intersect of ac -plane and the wave-plane (shaded gray planes), or perpendicular, angle (φ) for incoming and ($\varphi + 90^\circ$) for outgoing.

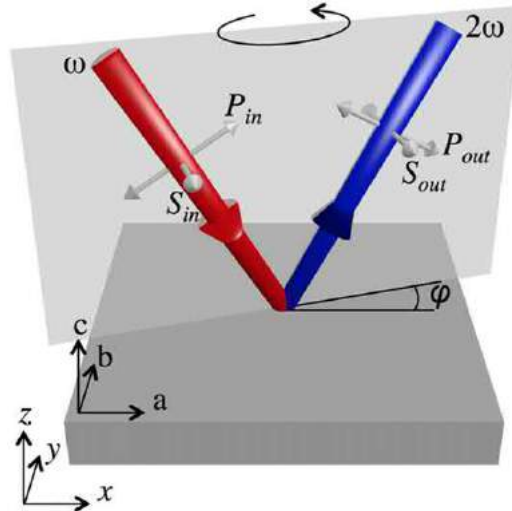


Fig. 3 Schematic of the rotating-scattering-plane-based SHG RA setups. This is equivalent to the one in Fig. 1 but with a rotating scattering-plane instead of a rotating sample.

A derivative of this scheme was therefore developed, as shown in Fig. 2(b). It keeps the idea of rotating the polarizations of both incident and reflected light, but deliberately chooses the oblique incidence to have polarization component along z direction. However, it still faces a couple of limitations. First, different from the last two setups in Figs. 1 and 2(a), the measurement process from this design itself breaks rotational symmetries, and consequently the RA data would seemingly obscure the rotational symmetries of the crystal. Second, there are only two independent polarization channels, i.e., parallel- or cross-polarized between incoming and reflected/transmitted beams, in this scheme, as compared to four in the Fig. 1 scheme. This limits the ability of accessing specific individual SHG susceptibility tensor elements, and therefore sometimes can overlook signatures of emergent phases. We will discuss this later in a concrete example in Section “An Odd-Parity Hidden Order in Sr_2IrO_4 ”.

Recently, a rotating-scattering-plane-based RA setup design has been proposed and constructed (Torchinsky *et al.*, 2014). In the design shown in Fig. 3, instead of rotating crystals or polarizations, it rotates the scattering-plane while maintaining the polarizations locked with respect to the scattering plane. By doing so, it is in fact equivalent to the rotating-sample-based scheme in Fig. 1, but unifies the advantages of all three designs above. (1) By keeping the sample stationary in the current geometry, various sample environment conditions such as cooling the crystal to a cryogenic temperature, and directing a magnetic or strain field along a particular crystallographic direction are compatible with the SHG RA measurements, which empowers this technique to study the vast amount of emergent electronic phases in correlated materials. (2) The oblique incidence enables the accessibility to all susceptibility tensor elements, which is necessary to the sensitivity of differentiating symmetry point groups. In addition, the reflection geometry is especially suitable for the bulk single crystals, the common form for correlated electron materials, because

their thickness typically exceeds the penetration depth of light at the inter-band resonance frequencies. (3) The full rotation of 360° of the light scattering plane during the RA acquirement preserves all rotational symmetries, and therefore the RA patterns directly visualize the rotational symmetries in the crystal about its surface normal axis. Furthermore, the total four polarization geometries relative to the scattering plane maximize the number of independent SHG RA measurements of the susceptibility tensor for one crystallographic facet. (4) The design of $4f$ optical system, described in details in the following paragraph, minimizes misalignments including beam walking on the sample and precession of reflected beam. As a result, not only SHG RA experiments on small size single crystals (<1 mm) or spatially inhomogeneous samples can be easily performed, but also aberrations in RA SHG patterns are significantly reduced.

In practice, the rotation of the light scattering plane is realized by rotating a customized fused silica phase mask (PM) which diffracts the normal incident beam into $+1$ and -1 orders at an angle α relative to the optical axis. So far, two generations of RA designs have been developed. In the first generation shown in Fig. 4, the collimated incoming beam of a pulsed laser at a given photon energy first passes through a polarizer (P) to determine the incoming power and a half-wave plate (WP) to choose the incoming polarization with respect to the scattering plane (S_{in} for perpendicular and P_{in} for parallel), then gets focused by a lens (L1) onto the PM to produce $+/-1$ orders forming the light scattering plane. The two diffracted orders get collimated simultaneously and parallel to each other after a second lens (L2). One order is then blocked by a beam block (B) while the other one is filtered through a long pass filter (LPF) to eliminate any parasitic higher harmonics. The collimated incident beam is finally focused on the sample by a Cassegrain reflective objective (RO). The fundamental and higher harmonic beams reflected from the sample exit the RO from its diametrically opposite side, and get picked up by a D-cut silver coated pick-off mirror (DM). The SHG beam with a defined polarization of S_{out} or P_{out} is then selected through a polarization analyzer (A) and a set of spectra-filters including two short pass filters (SPF) and one interference filter (IF), and eventually directed into a photo-multiplier tube (PMT). In this design, the set of PM, L2, RO and sample form a $4f$ system that overcomes the alignment challenges in the rotating-sample-based RA design. During the RA data acquisitions, PM together with WP steps at discrete angles to illuminate the investigated area on the sample at a series of azimuthal angle (φ), and the detection arm consisting of DM, A, SPF/IF and PMT mounted on a big rotation stage rotates in sync with PM and WP to record the induced SHG intensity $I_{P/S_{in}-P/S_{out}}^{2\omega}(\varphi)$ at a given polarization geometry $P/S_{in}-P/S_{out}$ and an angle (φ). The stepped fashion of this design requires a long time (in order of 10^3 s) to complete the full 360° sweep of (φ) for acquiring RA data, which is disadvantageous especially when the laser characteristics (power, point, pulse duration, etc.) have long-term fluctuations.

In order to overcome the drawback of long acquisition time, a second generation of the rotating-scattering-plane-based RA setup was developed, aiming at much shorter rotation cycles (Harter *et al.*, 2015). Shown in Fig. 5, a circularly polarized beam, produced from a linearly polarized beam using a quarter-wave plate (QWP), first passes through a linear polarizer (LP1) to select a polarization, either S_{in} or P_{in} , and then gets focused onto a fused silica binary PM to diffract the beam from the optical axis. The $+1$ order of the diffracted beam is then brought collimated and parallel to the optical axis via a lens (L1), further transmits through two tilted dichroic mirrors (DM1 and DM2 that transmit the fundamental beam but reflects the higher harmonic parasitic beams), and finally gets focused on the sample via an achromatic lens (L2). The reflected fundamental and higher harmonic beams return on the diametrically opposite side of L2, upon which it gets re-collimated. The SHG signal gets picked out by two tilted dichroic mirrors (DM2 and DM3) and one set of spectra filters (SF) with a chosen polarization of S_{out} or P_{out} via a linear polarizer (LP2), and draws a circle with varying intensity $I_{S/P_{in}-S/P_{out}}^{2\omega}(\varphi)$ on a cooled electron-multiplying CCD camera as the scattering plane rotates. In this design, LP1, PM and LP2 are coupled to a common axle to eliminate any possible mechanical phase slip, and therefore can rotate co-axially at a high speed (4 Hz or even more), which is greater than the first generation by a factor of 10^3 .

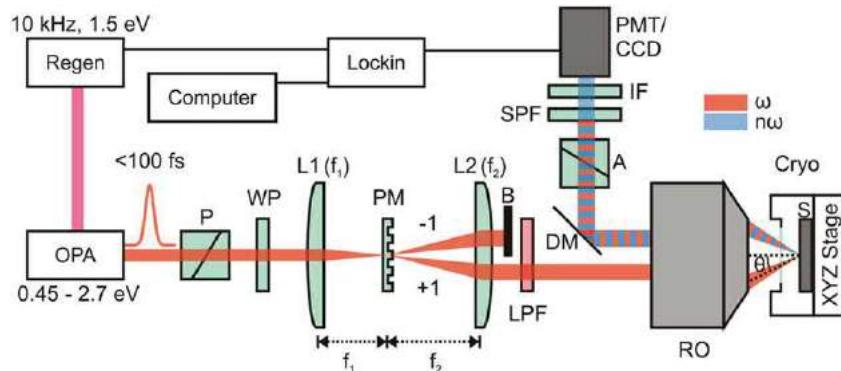


Fig. 4 Layout of the first generation rotating-scattering-plane-based SHG RA setup. A <100 fs pulsed laser beam from an OPA, seeded by a Ti:sapphire regenerative amplifier, passes through a polarizer (P) and waveplate (WP), and is focused by the first lens (L1) on a phase mask (PM). PM, the second lens (L2), a reflective objective (RO) and a sample (S) form a $4f$ optical system. A first order ($+1$) diffracted beam gets focused on S through this $4f$ system while the other one (-1) is blocked. The reflected beam is collimated through RO, picked off by a D-cut mirror (DM), filtered through an analyzer (A), short-pass filters (SPF) and interference filters (IF), and finally detected by a photomultiplier tube (PMT) using lock-in detection or a CCD camera. Adapted from Torchinsky, D.H., *et al.*, 2014. A low temperature nonlinear optical rotational anisotropy spectrometer for the determination of crystallographic and electronic symmetries. Review of Scientific Instruments 85(8), 083102.

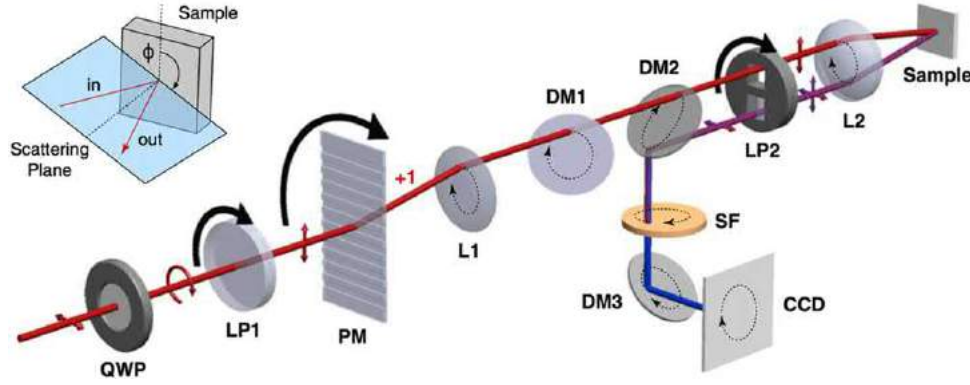


Fig. 5 Layout of the second generation rotating-scattering-plane-based SHG RA setup. Pulsed laser light passes through a series of optics including a quarter-wave plate (QWP), linear polarizers (LP1, 2), a phase mask (PM), collimating lens (L1), dichroic mirrors (DM1-3), achromatic objective lens (L2), spectral filter (SF), and CCD camera (CCD). LP2 is mounted on the detection side of a rotating wheel while the incoming side has a through-hole. The dashed arrows show the path traced by the beam on L1, L2, DM1-3, SF, and CCD during the RA acquisition. Top left inset shows the simplified geometry of a reflection-based SHG RA measurement. Adapted from Harter, J.W., *et al.*, 2015. High-speed measurement of rotational anisotropy nonlinear optical harmonic generation using position-sensitive detection. *Optics Letters* 40(20), 4671–4674.

Using the rotating-scattering-plane-based RA technique, it has been shown successful in detecting hidden symmetry breaking orders in a pyrochlore $\text{Cd}_2\text{Re}_2\text{O}_7$ (Harter *et al.*, 2017), a perovskite iridate Sr_2IrO_4 (Zhao *et al.*, 2016b) and a cuprate $\text{YBa}_2\text{Cu}_3\text{O}_y$ (Zhao *et al.*, 2016a) as discussed in Sections “A Parity-broken Nematic Order in $\text{Cd}_2\text{Re}_2\text{O}_7$, An Odd-Parity Hidden Order in Sr_2IrO_4 , and A Magnetic Quadrupole Order in the Pseudogap of $\text{YBa}_2\text{Cu}_3\text{O}_y$ ”.

Hidden Phases in Correlated Electron Systems

In condensed matter systems, a hidden phase can either be a state of matter that cannot be reached in the thermodynamic equilibrated conditions (Huai and Nasu, 2002; Ichikawa *et al.*, 2011; Koshihara and Adachi, 2006; Nasu, 2004; Tokura, 2006; Zhang *et al.*, 2015), or be one in equilibrium that however cannot be easily accessed by conventional detection techniques (Fechner *et al.*, 2016; Fu, 2015; Kuramoto *et al.*, 2009; Santini *et al.*, 2009; Witczak-Krempa *et al.*, 2013). In this review, we take examples from the latter category and discuss their microscopic nature revealed using optical SHG based techniques.

In a correlated electron system, the interplay among electronic degrees of freedom often leads to a macroscopic realization of the system's ground state that involves spontaneous symmetry breaking and the emergence of an order parameter. Often, the microscopic nature of the ordered state can be investigated through direct external probes. For example, the electric dipole arrangement in ferroelectrics can be studied by applying an external electric field. However, there are occasions where the ordering element is not of such conventional scalar or vector form as charge density or electric/magnetic dipole, but instead is of the tensor form. Such ordering parameters of the tensor character couple with the usual scalar or vector probe fields so weakly that they can be invisible to these conventional probes. As the most well known example, the hidden order phase in the heavy fermion compound URu_2Si_2 has been theoretically proposed to be a multipolar (tensor) order (Mydosh and Oppeneer, 2011, 2014; Shah *et al.*, 2000), but so far both the microscopic nature of the order parameter and the macroscopic symmetries of its long range ordered state remain mysterious with few direct experimental results.

In classical electrodynamics, the concept of electric and magnetic multipolar expansions is introduced to account for the spatial distribution of charge or magnetization density (Jackson, 1999). Electric multipolar expansions include terms of total charge Q (or monopole term), dipole moment p_i , quadrupole moment q_{ij} , and even higher multipolar moments where

$$Q = \int d\vec{r}' \rho(\vec{r}') \quad (7)$$

$$p_i = \int d\vec{r}' r'_i \rho(\vec{r}') \quad (8)$$

$$q_{ij} = \int d\vec{r}' (3r'_i r'_j - r'^2 \delta_{ij}) \rho(\vec{r}') \quad (9)$$

Similarly, magnetic multipolar expansions consist the first term of magnetic moment m_i with

$$m_i = \int d\vec{r}' \mu_i(\vec{r}') \quad (10)$$

the second term μ_{ij} with

$$\mu_{ij} = \int d\vec{r}' r'_i \mu_j(\vec{r}') \quad (11)$$

that can be decomposed into three irreducible groups:

- (1) the pseudo-scalar from the trace of the tensor that contains a single component,

$$a = \frac{1}{3} \mu_{ii} = \frac{1}{3} \int d\vec{r}' r'_i \mu_i(\vec{r}') \quad (12)$$

- (2) the toroidal moment pseudo-vector that is the anti-symmetric part of the tensor and contains three independent components,

$$T_i = \frac{1}{2} \epsilon_{ijk} \mu_{jk} = \frac{1}{2} \int d\vec{r}' \epsilon_{ijk} r'_j \mu_k(\vec{r}') \quad (13)$$

- (3) the quadrupole moment tensor that is the traceless symmetric part of $\vec{\mu}$ and contains five independent components,

$$\tilde{q}_{ij} = \frac{1}{2} \left(\mu_{ij} + \mu_{ji} - \frac{2}{3} \delta_{ij} \mu_{kk} \right) = \frac{1}{2} \int d\vec{r}' \left[r'_i \mu_j(\vec{r}') + r'_j \mu_i(\vec{r}') - \frac{2}{3} \delta_{ij} r' \cdot \vec{\mu}(\vec{r}') \right] \quad (14)$$

where the toroidal moment t_i and the quadrupole moment q_{ij} couple to the curl and the gradient of the magnetic field respectively, while the pseudo-scalar a is coupled to the divergent of the magnetic field and therefore represents a monopole moment; and even higher order terms.

In condensed matter systems, the electric and magnetic multipolar order parameters can be defined in a similar way as above (Fechner *et al.*, 2016), given the integration volume in Eqs. (7–14) is now replaced by the unit cell of a material. Taking the second term of magnetic multipolar expansions as an example (Fig. 6), the monopole, toroidal and quadrupole terms can be realized by the arrangement of localized magnetic moment within the unit cell. Each of the patterns in Fig. 6 has its time reversal symmetry related counterpart, and therefore its long range ordered state naturally has at least two degenerate ground states. Different patterns have different sets of symmetries, and consequently their long range ordered states have different point groups. For example, $\tilde{q}_{z^2}$ pattern breaks all three mirror planes m_{xy} , m_{xz} and m_{yz} , but obeys the symmetry operations of each mirror operation followed by a time reversal symmetry operation m'_{xy} , m'_{xz} and m'_{yz} with the prime standing for the time reversal symmetry operation. In contrast, $\tilde{q}_{xz}$ pattern only breaks m_{xz} while remaining symmetric under m'_{xz} , m_{xy} and m_{yz} operations. Therefore, assuming these two patterns ferroically order in a long range in a $4/mmm$ crystal lattice, the resulting magnetic point group will be $m'm'm'$ for the $\tilde{q}_{z^2}$ type quadrupole order, and $m'mm$ for the $\tilde{q}_{xz}$ type one.

In the end of this section, let us take the example of ferroic $\tilde{q}_{xz}$ - type magnetic quadrupole order to briefly summarize the challenges of direct experimental detection of tensor (i.e., multipolar) orders. First, the summation of all the magnetic moment within a unit cell is zero, and therefore there is no net magnetization per unit cell that can couple directly with the external

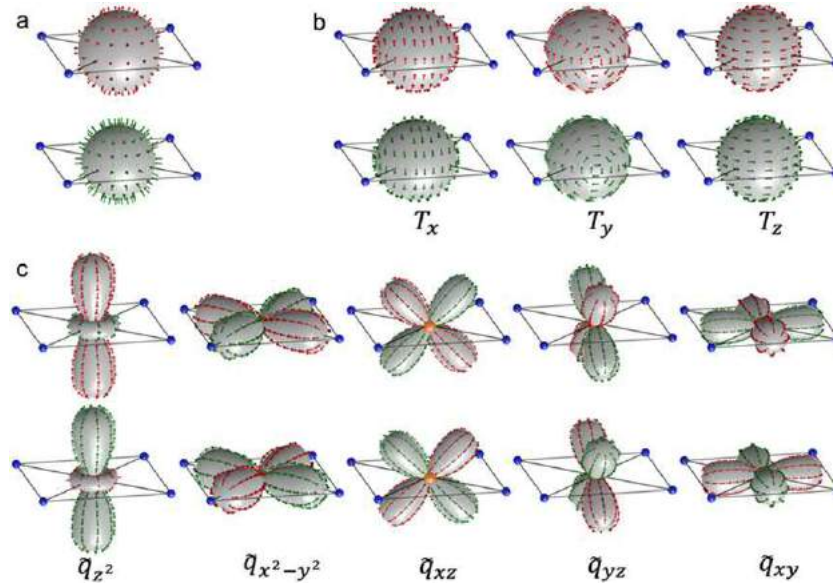


Fig. 6 Patterns of local magnetic dipole moments $\vec{\mu}$ within a unit cell of a square plaque that generates magnetic quadrupoles, created based on Fechner, M., *et al.*, 2016. Quasistatic magnetoelectric multipoles as order parameter for pseudogap phase in cuprate superconductors. Physical Review B 93(17), 174419. (a) Magnetic monopole component; (b) three toroidal components, T_x , T_y and T_z ; (c) five magnetic quadrupole components, $\tilde{q}_{z^2}$, $\tilde{q}_{x^2-y^2}$, $\tilde{q}_{xz}$, $\tilde{q}_{yz}$, and $\tilde{q}_{xy}$. Patterns in the 2nd and 4th rows are the time-reversed counterparts of that in the 1st and 3rd rows.

magnetic field. Second, the order parameter here is described by a rank-2 tensor, so the conjugate field to couple with needs to be a tensor field. In principle, the multiple copies of electric and magnetic fields in the nonlinear optics can act as the tensor field that couples with the tensor order parameter. Third, the time reversal symmetry related two ground states lead to two domains in the ordered state. Therefore, the symmetry properties of ordered state can be easily concealed by the domain averaged measurements. Forth, the ferroic order preserves the translational symmetries of the underlying crystal lattice, which is the $q=0$ problem known in diffraction measurements. As a result, the diffraction pattern from the ferroic tensor order overlays on top of the nuclear diffraction pattern, and requires extremely high resolution to resolve.

Hidden Phases Revealed by SHG-based Techniques

So far, SHG has been successfully exploited to reveal the otherwise hidden phases in several correlated electron systems because of its high sensitivity to the symmetry point groups, tensor field from multiple copies of electromagnetic fields, and spatial resolution of optical diffraction limit $\sim 1\mu\text{m}$.

A Ferrotoroidal Order in LiCoPO_4

LiCoPO_4 has the olivine crystal structure and is well characterized by the mmm symmetry point group in the paramagnetic state (Rabe, 2007; Spaldin *et al.*, 2008; Van Aken *et al.*, 2007; Zimmermann *et al.*, 2014). It contains planes of CoO_6 octahedra that are buckled by sharing of corners with PO_4 tetrahedra (gray) above (red Co^{2+}) and below (green Co^{2+}) the plane as shown in Fig. 7 (a). Looking into the ab -plane, two pairs of Co^{2+} cations within one unit cell are separated into two sets with slightly different radius from their center (Fig. 7(c) left column without magnetic moments). At 21.8 K, the localized magnetic moments at Co^{2+} form a collinear antiferromagnetic order with alternating sign from row to row in the plane, and are oriented along a direction that is 4.6° away from the axis of the row, i.e., b -axis. Considering the clockwise ($+4.6^\circ$) and anti-clockwise (-4.6°) tilt of the Co^{2+} magnetic moment and their time reversal symmetry related counterparts, there are in total four distinct but energetically equivalent magnetic ground states, as sketched in Fig. 7(b). In practice, each of the states can be split into two parts – one part has the b -axis components of the magnetic moments, and the other has the rest of the magnetic moments that are aligned along a -axis. The first

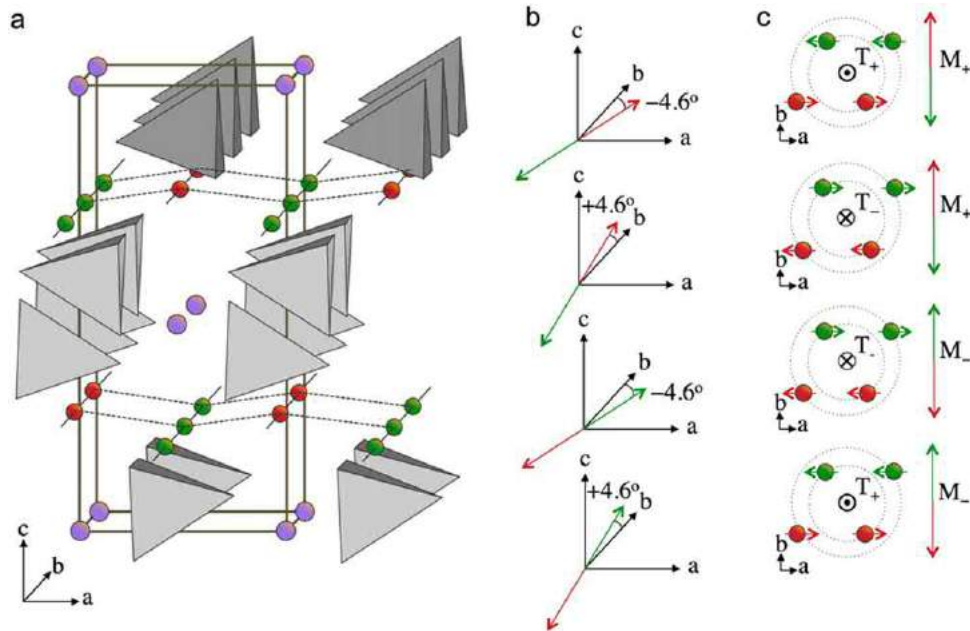


Fig. 7 Illustration of ferrotoroidicity in LiCoPO_4 . (a) Crystal structure of LiCoPO_4 , where cobalt ions (green and red spheres) form planar layers. These cobalt layers are buckled (red for the upward-buckled rows and green for the downward-buckled rows) by the presence of phosphorus-centered oxygen tetrahedra (gray) above and below them. And lithium ions (purple spheres) are located between the tetrahedra. (b) The magnetic moments of cobalt are collinearly aligned, but tilted at an angle of 4.6° with respect to the rows. There are in total four arrangements for the magnetic moments, two directions of the tilt relative to the rows and their time-reversed counterparts. (c) Decompositions of the magnetic moments into the a -axis component that gives rise to the non-zero toroidization out of (T^+) or into (T^-) the ab -plane (left column), and b -axis component that forms the antiferromagnetic order of the non-centrosymmetric mmm' symmetry with two time-reversal related structures (M^- and M^+) (right column). Corresponding to the four arrangements in (b), there are four combinations between the toroidal moment and the magnetic moment, namely, T^+M^+ , T^-M^+ , T^-M^- and T^+M^- . Adapted from Rabe, K.M., 2007. Solid-state physics: Response with a twist. *Nature* 449 (7163), 674–675.

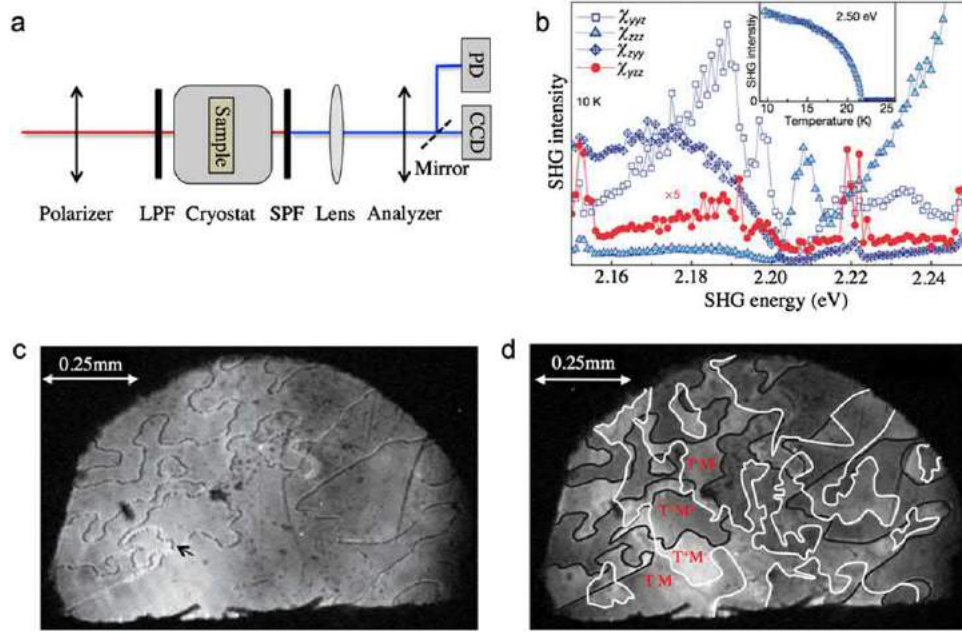


Fig. 8 Observation of ferrotoroidic domains by optical SHG. (a) Schematic of the transmission-based SHG setup. Incoming light passes through Polarizer and long pass filter (LPF) before normal incidence onto the sample in the cryostat. The transmitted light is filtered by the short pass filter (SPF) and the selected SHG light gets detected by either CCD camera (imaging) or photodiode (integrated intensity); (b) Energy dependent contributions from independent non-zero tensor components χ_{ijk} are plotted. The inset shows the temperature dependent SHG intensity from χ_{zzz} with an order parameter like onset at 21.8 K; (c) Antiferromagnetic domains that are obtained using light from χ_{yyz} are imaged. The dark lines (with dark arrow pointing one example) are the domain boundaries between the two time-reversal related antiferromagnetic domains. (d) Coexistence of antiferromagnetic and ferrotoroidic domains that are obtained using the interfering light from χ_{yyz} and χ_{zyy} is shown. The dark and white lines indicate the antiferromagnetic and ferromagnetic domain walls respectively. Examples of four areas are labeled with the order parameter combinations T^+M^+ , T^+M^- , T^-M^+ , and T^-M^- . Adapted from Van Aken, B.B., *et al.*, 2007. Observation of ferrotoroidic domains. *Nature* 449 (7163), 702–705 and Spaldin, N.A., M. Fiebig, Mostovoy, M., 2008. The toroidal moment in condensed-matter physics and its relation to the magnetoelectric effect. *Journal of Physics: Condensed Matter* 20(43), 434203 (All authors contributed equally to this work).

part forms an antiferromagnetic order that is characterized by the mmm' symmetry point group, and has two degenerate ground states related by the time reversal symmetry operation (M_+ and M_- in the right column in Fig. 7(c)). The second part gives rise to a ferrotoroidal order $T_i = \sum_{a=1}^N \epsilon_{ijk} r_{a,j} \mu_{a,k}$ that is characterized by $2'$ symmetry point group with a -axis as the 2-fold rotation axis, and also has two time reversal symmetry related ground states that originate from the clockwise and anti-clockwise tilt of the magnetic moments (T_+ and T_- in the right column in Fig. 6(c)). Four combinations of the antiferromagnetic order and the ferrotoroidic order, i.e., M_+T_+ , M_+T_- , M_-T_+ and M_-T_- , fully reproduce the four magnetic structures in Fig. 7(b).

Van Aken and colleagues established (Van Aken *et al.*, 2007), for the first time, the presence of ferrotoroidic order by detecting domains of opposite toroidizations through the coherent interference between the SHG from the ferrotoroidic order and the antiferromagnetic order in LiCoPO_4 . By measuring SHG response of a polished (100)-oriented LiCoPO_4 film with a thickness $122\mu\text{m}$ using a transmission SHG setup with a normal incidence geometry shown in Fig. 8(a), independent nonzero electric dipole induced SHG susceptibility tensor components χ_{yyz} , χ_{zzz} , χ_{zyy} and χ_{yzz} shows up at the Néel temperature $T_N = 21.8\text{K}$ (Fig. 8(b)), which is only compatible with the low symmetry point group $2'$ for the proposed ferrotoroidic order. On the other hand, χ_{yyz} is also present in the electric dipole induced SHG susceptibility tensor under mmm' symmetry point group for the antiferromagnetic order. SHG image on the LiCoPO_4 crystal with light from χ_{yyz} component shows dark lines on a constant-intensity background in Fig. 8(c), where the dark lines are the domain boundaries between two 180° -rotated antiferromagnetic domains. The image of the interference between light from χ_{yyz} and χ_{zyy} components displays a patchwork with bright and dark areas in Fig. 8(d), which indicates two SHG contributions from antiferromagnetic and ferrotoroidic order are present here and interfere constructively (bright) and destructively (dark). In Fig. 8(d), the black lines highlight the antiferromagnetic domain boundaries while the white lines outline the ferrotoroidic domains. Examples of four antiferromagnetic and ferrotoroidic order combinations are also marked in Fig. 8(d).

Following the work by Van Aken *et al.*, Anne Zimmermann and colleagues further confirmed the ferrotoroidic nature of the ordered state in LiCoPO_4 below the Néel temperature by observing hysteretic poling of ferrotoroidic domains in the conjugate toroidal field $\vec{S} \propto \vec{E} \times \vec{B}$ (Zimmermann *et al.*, 2014). SHG images in Fig. 9(a) show that magnetic or electric field applied alone during cooling results in a multi-domain state while the product of magnetic and electric field promotes a ferrotoroidic single-domain state. A striking pronounced hysteresis is observed in the integrated SHG intensity in Fig. 9(b) when cycling the magnetic field within a $\pm 2\text{T}$ interval at a fixed electric field of 11 kV cm^{-1} . Furthermore, the SHG intensity v.s. $\mu_0 H \cdot E$ spectra from separate runs where the sign and amplitude of the fixed fields varied all collapse onto one another in Fig. 9(c), confirming that solely the

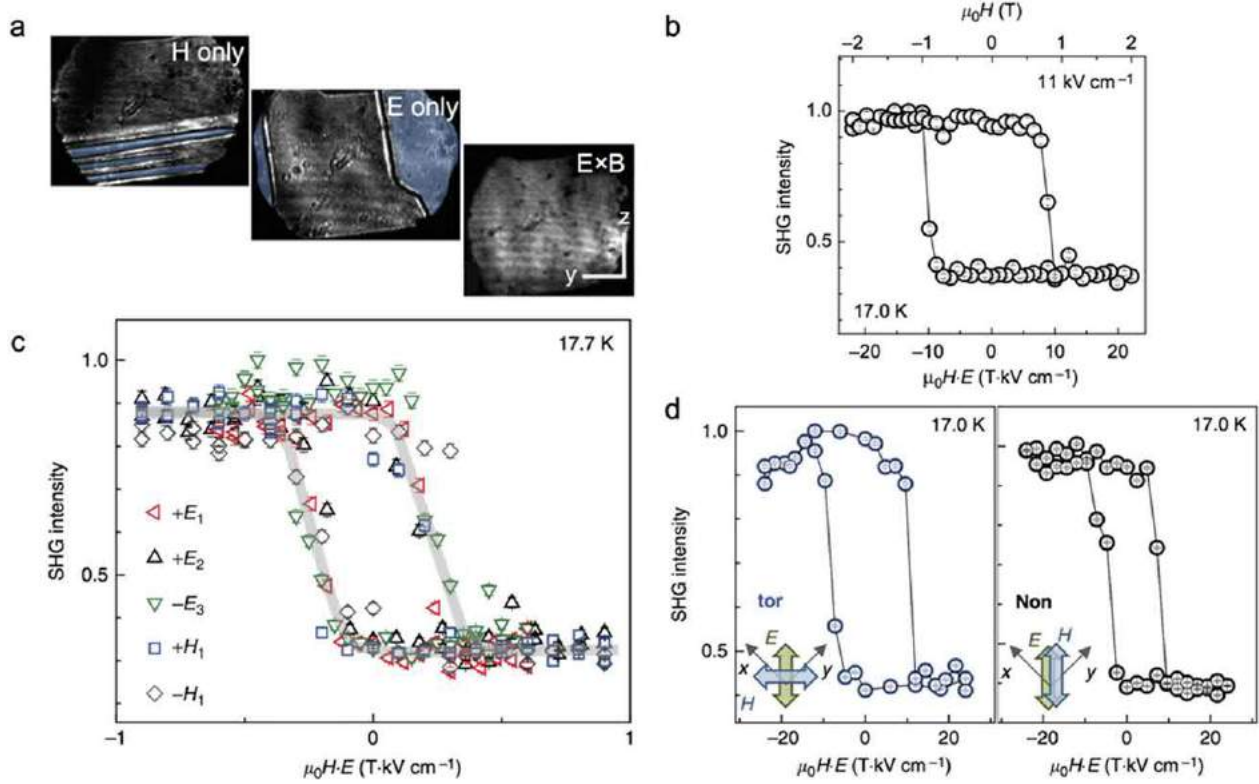


Fig. 9 Ferroic nature of magnetic toroidal order. (a) Multi-domain states (first two images) and single domain state (third image) emerge after cooling the LiCoPO_4 sample in a magnetic or electric field alone and in a toroidal field of orthogonal magnetic (H_x) and electric (E_y) fields respectively; (b) SHG intensity at 17.0 K as a function of a static magnetic field $\mu_0 H_x$ ranging from -2 T to 2 T in the presence of a static electric field $E_y = 11\text{ kVcm}^{-1}$ shows a hysteresis loop; (c) Hysteresis loops of SHG intensity collapse into one (gray guide-of-eye line) for various combinations in signs and amplitudes of magnetic and electric fields ($+E_1 = +1.2\text{ kVcm}^{-1}$, $+E_2 = +1.8\text{ kVcm}^{-1}$, $-E_3 = +1.0\text{ kVcm}^{-1}$, $+H_1 = +0.5\text{ T}$, $-H_1 = -0.5\text{ T}$) with the same $E_y \cdot H_x$ product; (d) Hysteresis loops in purely toroidal field and non-toroidal field with the same $E \cdot H$ product show different coercive fields. The insets show the corresponding experimental field configurations. Adapted from Zimmermann, A.S., D. Meier, D., Fiebig, M., 2014. Ferroic nature of magnetic toroidal order. Nature communications 5, 4796.

product of the magnetic and electric field determines the shape of hysteresis loop. Finally, the toroidal and non-toroidal poling are compared through measuring the coercive field under two field configurations as shown in Fig. 9(d). The experimental value of $\frac{|EH|_{\text{tor}}}{|EH|_{\text{non}}} = 1.21 \pm 0.14$ significantly differs from the calculated value of 3 assuming equal efficiency between toroidal and non-toroidal poling, showing that toroidal field is much more effective in promoting the ferrotoroidic single-domain state.

A Parity-Broken Nematic Order in $\text{Cd}_2\text{Re}_2\text{O}_7$

The correlated metallic pyrochlore $\text{Cd}_2\text{Re}_2\text{O}_7$ undergoes a continuous phase transition from an inversion symmetric cubic structure (space group $Fd\bar{3}m$, Fig. 10(a)) to a parity-broken tetragonal structure (space group $I\bar{4}m2$, Fig. 10(b)) at $T_s = 200\text{ K}$, with the lattice distortion of the E_u symmetry (Castellan et al., 2002; Kendziora et al., 2005; Petersen et al., 2006; Weller et al., 2004; Yamaura and Hiroi, 2002). Recent theory predicts that spin-orbit coupled metal can host a variety of inversion asymmetric ordered phases resulting from Pomeranchuk type instabilities in the spin channel (Fu, 2015). These phases preserve the translational symmetries but break the point group symmetries of underlying crystal lattice, and therefore regarded as new examples of electronic liquid crystals (i.e., electronic nematicity). Moreover, the spin-orbit coupling leads to spin-split Fermi surfaces with characteristic spin textures, and generally induces structural distortions from the electronic parity-breaking orders. In particular, $\text{Cd}_2\text{Re}_2\text{O}_7$ is one of the proposed pyrochlore oxide candidates hosting a primary multipolar ordered phase of either E_u or T_{2u} type order parameters that further induces the observed secondary structural phase transition.

Prior to 2006, the low temperature structural order parameters was not settled between the distinct but nearly degenerated crystallographic structures $I\bar{4}m2$ (Fig. 10(b), $\chi = \{\chi_{xyz} = \chi_{xzy} = \chi_{yxz} = \chi_{yzx} = \chi_{zyx}\}$) and $I4_122$ (Fig. 10(c), $\chi = \{\chi_{xyz} = \chi_{xzy} = -\chi_{yxz} = -\chi_{yzx}\}$) (Castellan et al., 2002; Kendziora et al., 2005; Weller et al., 2004; Yamaura and Hiroi, 2002). Jesse C. Petersen and colleagues resolved this mystery by using SHG ellipsometry, and identified the structural symmetry below $T_s = 200\text{ K}$ to be $I\bar{4}m2$ (Petersen et al., 2006). The SHG response of the cubic (111) oriented $\text{Cd}_2\text{Re}_2\text{O}_7$ were measured on a reflection SHG setup of the normal incidence geometry, with independent control on the incident and reflected polarizations, a_{ω} and $a_{2\omega}$, by a half-wave plate ($\lambda/2$) and a crystal polarizer (P) respectively (Fig. 11(a)). In order to minimize the effects from linear birefringence,

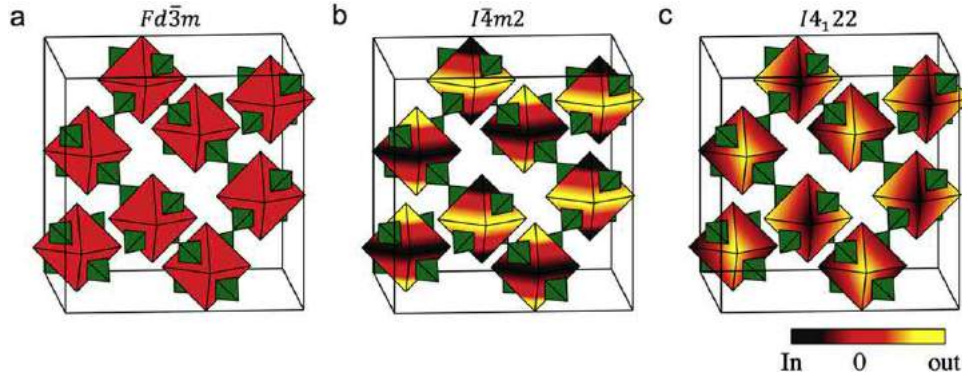


Fig. 10 The structural phase transition in $\text{Cd}_2\text{Re}_2\text{O}_7$. (a) Cubic pyrochlore unit cell with $Fd\bar{3}m$ symmetry. The vertices of green tetrahedral indicate Re sites and red octahedra indicate O_1 sites. Cd and O_2 sublattice are not shown here. (b-c) Pseudocubic unit cells of the distorted lattices with $I\bar{4}m2$ and $I4_122$ symmetry respectively. The neighboring octahedra have displacements with opposite sign, thus breaking inversion symmetry. Adapted from Petersen, J.C., *et al.*, 2006. Nonlinear optical signatures of the tensor order in $\text{Cd}_2\text{Re}_2\text{O}_7$. *Nature Physics* 2(9), 605–608.

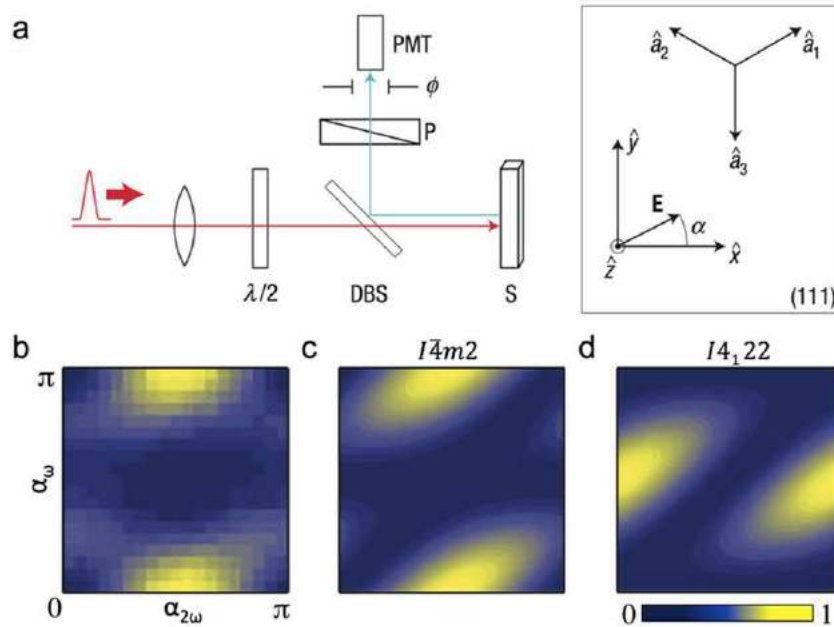


Fig. 11 Identification of crystal structure below the transition temperature $T_s=200\text{K}$. (a) (Left) Schematic of the SHG ellipsometry setup, where femtosecond laser pulses were focused through a dichroic beamsplitter (DBS) to a cubic (111) surface of $\text{Cd}_2\text{Re}_2\text{O}_7$ and the reflected SHG was detected by a photomultiplier tube (PMT) after passing through an iris (Φ). Input and output polarizations are selected with a half-wave plate ($\lambda/2$) and a crystal polarizer (P), respectively. (Right) Cubic coordinate ($\hat{a}_1, \hat{a}_2, \hat{a}_3$) for the high-temperature phase, and optical coordinate ($\hat{x}, \hat{y}, \hat{z}$), with $\hat{z}$ -axis along the optical axis. (b) Normalized SHG intensity map as a function of input and output polarizations taken at 4 K. (c–d) Simulations of SHG intensity map under $I\bar{4}m2$ and $I4_122$ symmetry respectively. Adapted from Petersen, J.C., *et al.*, 2006. Nonlinear optical signatures of the tensor order in $\text{Cd}_2\text{Re}_2\text{O}_7$. *Nature Physics* 2(9), 605–608.

crystal twinning, experimental misalignment etc. on the SHG signal variations, the authors obtained 2D data sets of SHG intensity in by varying a_ω and $a_{2\omega}$. Comparing the experimental map (Fig. 11(b)) with the simulated maps under $I\bar{4}m2$ (Fig. 11(c)) and $I4_122$ (Fig. 11(d)) symmetries, it clearly identifies $I\bar{4}m2$ as the proper crystallographic structure for $\text{Cd}_2\text{Re}_2\text{O}_7$ at $T_s=200\text{K}$.

John Harter and colleagues recently revisited $\text{Cd}_2\text{Re}_2\text{O}_7$ to identify the electronic origin of the primary order parameter (Harter *et al.*, 2017).

An Odd-Parity Hidden Order in Sr_2IrO_4

The single layer perovskite iridate Sr_2IrO_4 realizes a novel spin orbit coupled $J_{\text{eff}}=1/2$ mott insulating ground state, and transits into a canted antiferromagnetic ordered state that lowers system symmetry from tetragonal ($4/mmm$ point group, Fig. 12(a)) to orthorhombic ($mmm1'$ point group, Fig. 12(b)) (Kim *et al.*, 2008, 2009). Despite the striking similarities in crystallographic,

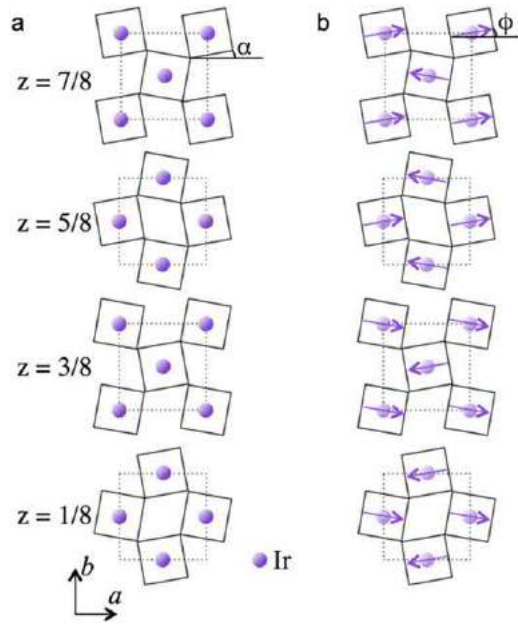


Fig. 12 Crystal structure and antiferromagnetic order in Sr_2IrO_4 . (a) Crystal structure of Sr_2IrO_4 , consisting four layers of IrO_6 octahedra (square in ab -plane) per unit cell. There are two sets of IrO_6 octahedra sublattices, clockwise and anti-clockwise rotated for $\alpha \sim 12^\circ$ about c -axis. (b) Canted antiferromagnetic ordering of $J_{\text{eff}} = 1/2$ moments (purple arrows) within the ab -plane and their stacking pattern along c -axis. The magnetic moments perfectly lock with the IrO_6 octahedra edges ($\phi/\alpha = 1$).

magnetic structures and electronic ground state between Sr_2IrO_4 and the cuprate parent compound La_2CuO_4 , recent experimental realizations of pseudogap and d -wave gap behaviors in the doped Sr_2IrO_4 systems further bridge the analogy between Sr_2IrO_4 and cuprates. More recently, the observation of an odd-parity multipolar order in Sr_2IrO_4 and the Rh-doped counterpart is one more surprising reproduction of cuprate phenology in this iridate system.

It was a mystery why the magnetic moment at the Ir^{4+} locks perfectly with the IrO_6 octahedron orientation ($\phi/a = 1$ in Fig. 12 (b)), despite there are two opposite canting directions. Darius H Torchinsky and colleagues resolved a fine staggered tetragonal distortion of IrO_6 octahedra that enhances the magnetoelastic coupling to lock the magnetic moment with octahedra rotation perfectly (Torchinsky *et al.*, 2015), using first version of rotating-scattering-plane-based RA integer harmonic generation setup (Fig. 4 in Section “Generations of Rotational Anisotropy Experimental Designs”). The third harmonic generation (THG) RA patterns (Fig. 13(a)) were observed to rotate away from the crystal axes a and b for an angle of $\sim 11.5^\circ$, and the two in $P_{\text{in}} - S_{\text{out}}$ and $S_{\text{in}} - P_{\text{out}}$ geometries show alternating big-small lobes. Both observations prove broken mirror planes m_{ac} and m_{bc} in the data, therefore in the crystal as well. The largest subgroup of previously assigned structural symmetry point group $4/mmm$ is the centrosymmetric $4/m$, and the THG RA patterns are well fit by the bulk electric dipole induced THG under $4/m$. Microscopically, the loss of the two mirror planes originates from a staggered tetragonal distortion of the crystal lattice where the two sets of IrO_6 octahedral sublattices have unequal tetragonal splitting (Δ_1 and Δ_2 shown in Fig. 13(b)). In particular, theoretical calculations found that the magnetoelastic locking is remarkably insensitive (i.e., $\phi/a = 1$) to both the tetragonal distortion size and spin orbit coupling strength when the signs of the tetragonal distortion is staggered between sublattices ($\Delta_1 = -\Delta_2$), shown in Fig. 13(c). In contrast, calculations on the uniform tetragonal distortion model ($\Delta_1 = \Delta_2$) shows a sharp decrease of the locking when the tetragonal distortion size increases at a given spin orbit coupling strength (Fig. 13(d)). The comparison between the two models suggests that the magnitude of the locking is more strongly influenced by the spatially averaged value of the tetragonal distortion rather than its local value on an individual IrO_6 octahedron, allowing the existence of large local tetragonal distortion to be reconciled with the observation of perfect magnetoelastic locking.

Cooling Sr_2IrO_4 down to the cryogenic temperature, Liuyan Zhao and colleagues revealed an odd-parity order that breaks the global inversion symmetry (Zhao *et al.*, 2016b), using the SHG RA setup in Fig. 4. At room temperature, the SHG RA pattern, in the $P_{\text{in}} - S_{\text{out}}$ polarization geometry in Fig. 14(a), results from the bulk electric quadrupole induced SHG from the centrosymmetric crystal structure with $4/m$ symmetry. In contrast, the low temperature SHG RA pattern, $P_{\text{in}} - S_{\text{out}}$ in Fig. 14(b), clearly breaks rotational symmetries and inversion symmetry, which is fitted by an interference between the structural contribution of bulk electric quadrupole induced SHG under $4/m$ and the emergent order contribution of bulk electric dipole induced SHG under $2'/m$ (or $m1'$). Here, it is worth noting that this new order is most obvious in the $P_{\text{in}} - S_{\text{out}}$ geometry while barely notable in the $S_{\text{in}} - P_{\text{out}}$ channel (Fig. 14(c) and (d)). Take this hidden order as an example, if we measure with the setup in Fig. 3(b) that only has cross and parallel polarized channels, the $P_{\text{in}} - S_{\text{out}}$ and $S_{\text{in}} - P_{\text{out}}$ signals above will be combined in the cross polarized channel. As a result, the small signal from the new order will be a tiny change on top of a large background from the structural signal, and therefore can be easily concealed, just as in the case of $S_{\text{in}} - P_{\text{out}}$ above.

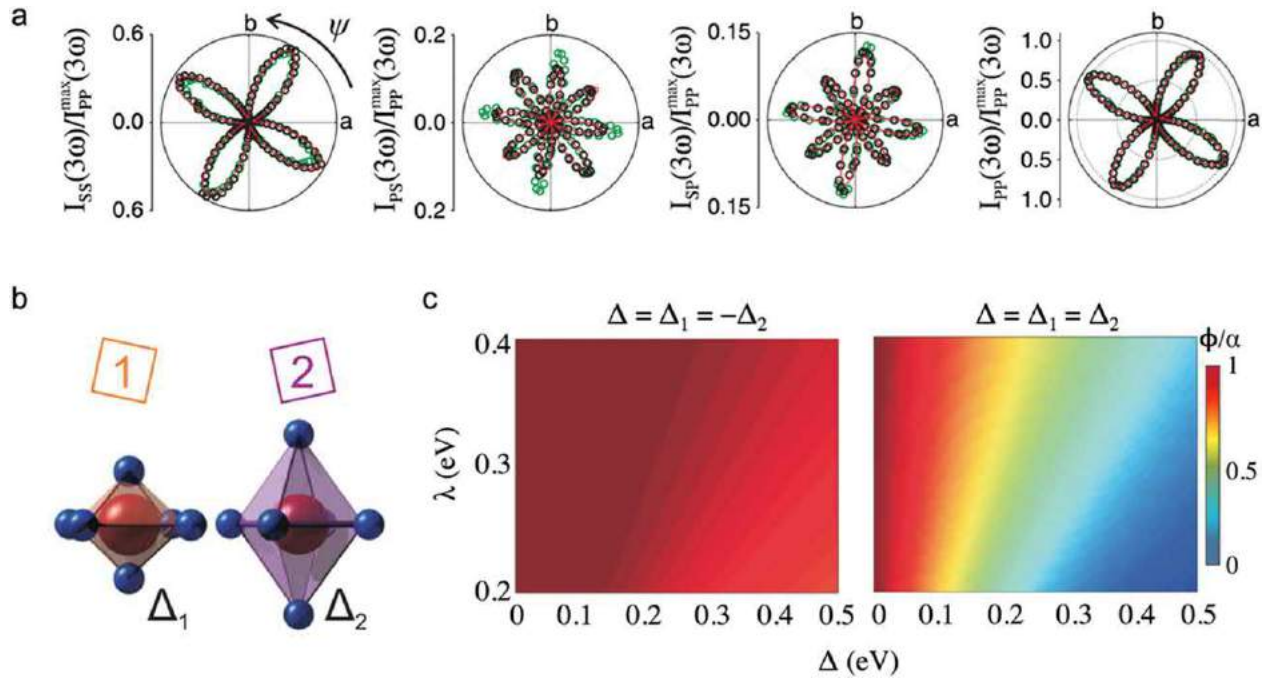


Fig. 13 Staggered tetragonal distortions in Sr_2IrO_4 revealed by THG. (a) THG RA patterns in four polarization geometries, $S_{in} - S_{out}(SS)$, $P_{in} - S_{out}(PS)$, $S_{in} - P_{out}(SP)$ and $P_{in} - P_{out}(PP)$, with THG intensity scaled to the PP maxima. Black circles are taken at 295K while green are at 180K, and red lines are best fits to the 295K data using bulk electric dipole induced THG RA patterns calculated under centrosymmetric $4/m$ point group. (b) An unequal tetragonal distortion (Δ_1 and Δ_2) on the two sublattices as required by the $4/m$ point group. (c-d) The calculated ratio of ϕ/a as a function of both λ and Δ for the case of staggered ($\Delta = \Delta_1 = -\Delta_2$) and uniform ($\Delta = \Delta_1 = \Delta_2$) distortion respectively. Adapted from Torchinsky, D.H., *et al.* 2015. Structural distortion-induced magnetoelastic locking in Sr_2IrO_4 revealed through nonlinear optical harmonic generation. *Physical Review Letters* 114(9), 096404.

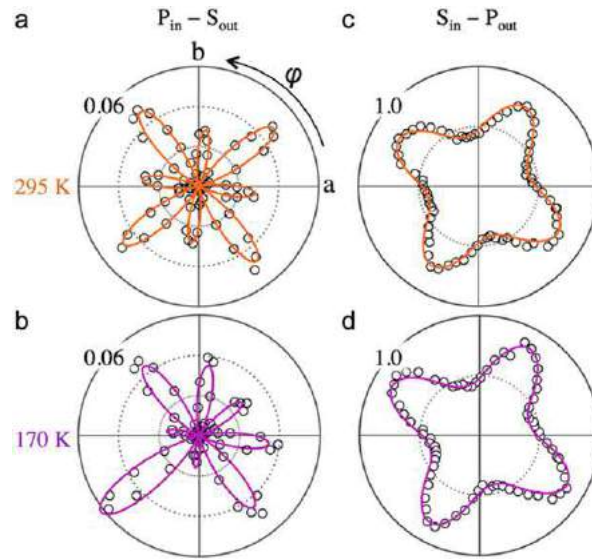


Fig. 14 Symmetry of the hidden order in Sr_2IrO_4 . (a-b) SHG RA data collected in the $P_{in} - S_{out}$ polarization geometry at $T=295\text{K}$ and $T=170\text{K}$, respectively. (c-d) Analogous data taken in the $S_{in} - P_{out}$ polarization geometry. The orange lines are best fits to the data at 295K using electric quadrupole induced SHG under the centrosymmetric $4/m$ point group, and the purple lines are best fits to the data at 170K using a coherent interference between electric quadrupole (centrosymmetric $4/m$ point group) and electric dipole (non-centrosymmetric $2'/m$ or $m1'$ point group) induced SHG. Adapted from Zhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. *Nature Physics* 12(1), 32–36.

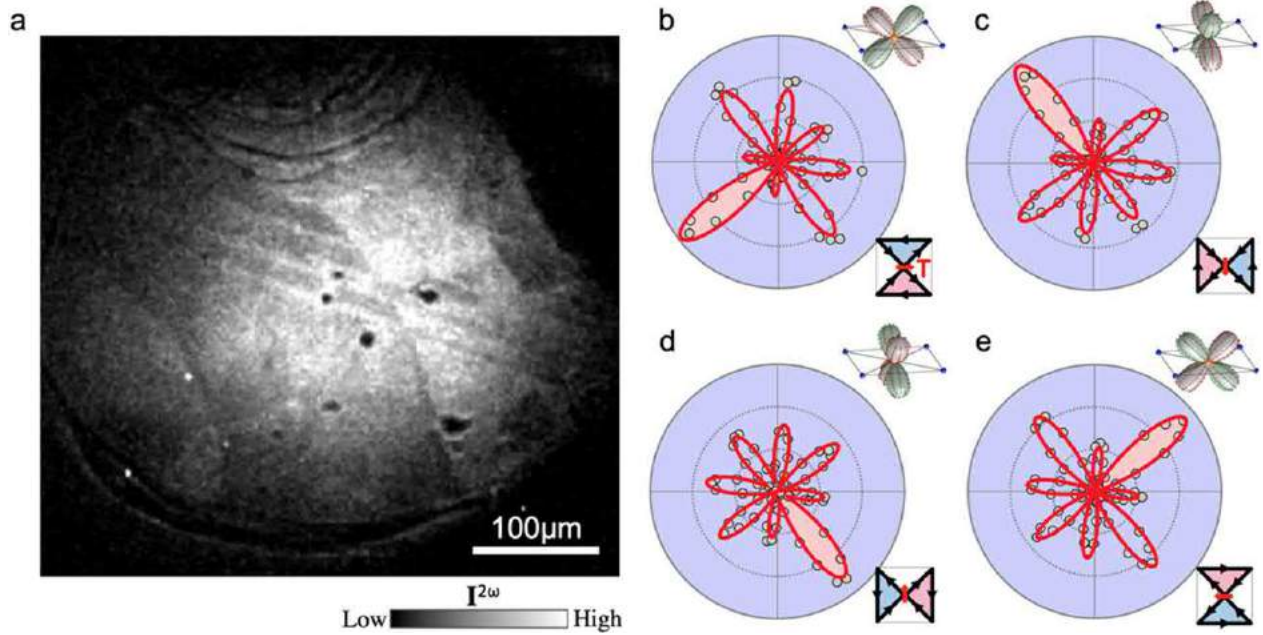


Fig. 15 Degenerate ground states of the hidden order in Sr_2IrO_4 . (a) Wide-field reflection SHG image of Sr_2IrO_4 (001) plane measured under $P_{in} - S_{out}$ polarization geometry at $T = 175\text{K}$. The brighter and darker patchwork in the image arise from the domains of the hidden order. (b–e) Four different types of SHG RA patterns found by performing local measurements within all the domains in (a). The largest lobes are shaded in pink to highlight the orientation of the patterns that are rotated every 90° . Schematics of the four degenerate magnetic quadrupole configurations are in the upper right insets while that of magneto-electric loop-current order are shown in the lower right insets. Adapted from Zhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. *Nature Physics* 12(1), 32–36.

Wide field SHG image at 175K in Fig. 15(a) shows bright and dark patches that are separated by domain boundaries straight over a length scale of tens of micrometers. Local RA SHG measurements within each of the patches across the crystal reveal a total of four distinct of SHG RA patterns in Fig. 15(b–e), which are exactly 0° , 90° , 180° and 270° -rotated copies of that shown in Fig. 14(b). These results suggest four degenerate ground states of the noncentrosymmetric ordered state, and are consistent with the microscopic model of magneto-electric loop-current order and magnetic quadrupole order. On one hand, the magneto-electric loop-current order, consisting a pair of counter-circulating current loops in each IrO_2 plaquet (Kung *et al.*, 2014; Orenstein, 2011; Varma, 1997; Weber *et al.*, 2009; Yakovenko, 2015), satisfies $2'/m$ point group and can be described by a toroidal pseudovector order parameter defined as $\vec{T} = \sum \vec{r}_i \times \vec{\mu}_i$, where $\vec{r}_i$ is the location of the orbital magnetic moment $\vec{\mu}_i$ inside the plaquette. Four degenerate configurations exist in a tetragonal lattice because the two intra-unit-cell current loops can lie along either of the two diagonals in the square plaquet and can have either of two time-reversed configurations, which correspond to four 90° -rotated directions of the pseudovector T as shown in the lower right inset of Fig. 15(b–e). On the other hand, because of the tetragonal distortion of IrO_6 octahedron, $\tilde{q}_{xy}$ and $\tilde{q}_{yz}$ magnetic quadrupoles remain degenerate but split away from $\tilde{q}_{xz}$, $\tilde{q}_{yz}$ and their time-reversed counterparts form the four degenerate configurations shown in the upper right inset of Fig. 15(b–e), and each of their ferroic orders satisfies $2'/m$ point group. SHG RA data so far cannot differentiate between these two models, and certainly cannot rule out any other microscopic models that obey the same set of symmetries.

The temperature and doping evolution of the odd-parity hidden order in $\text{Sr}_2\text{Ir}_{1-x}\text{Rh}_x\text{O}_4$ is summarized in the phase diagram in Fig. 16. The hidden phase transition observed in the parent Sr_2IrO_4 persist in the hole doped $\text{Sr}_2\text{Ir}_{1-x}\text{Rh}_x\text{O}_4$, even though the transition temperature T_Q is suppressed with increasing x . Remarkably, the splitting between T_Q and the Néel temperature T_N grows monotonically from approximately 2 to 75K between $x=0$ and $x \sim 0.11$, indicating that the Néel and hidden order are not trivially tied, but are independent and distinct electronic phase. But the relationship between this hidden phase and the pseudogap phenomena remains uncertain in Rh-doped Sr_2IrO_4 system because of lacking of the pseudogap onset temperature here.

A Magnetic Quadrupole Order in the Pseudogap of $\text{YBa}_2\text{Cu}_3\text{O}_\gamma$

An enigmatic pseudogap region in the phase diagram of cuprates, characterized by a partial suppression of low energy electronic excitations, has been subjected to the debate of its nature for almost 30 years. Taking $\text{YBa}_2\text{Cu}_3\text{O}_\gamma$ as an example, its parent $\text{YBa}_2\text{Cu}_3\text{O}_6$ has a tetragonal crystal structure (point group $4/mmm$, Fig. 17(a)) while its detwinned hole doped counterparts $6 < \gamma \leq 7$ have an orthorhombic crystal structure because of the O occupation in the CuO chains along b -axis (point group mmm , Fig. 17(b)) (Shaked, 1994). Recently, polarized neutron diffraction (Fauqué *et al.*, 2006; Mangin-Thro *et al.*, 2015; Mook *et al.*, 2008), Nernst effect (Daou *et al.*, 2010), THz polarimetry (Lubashevsky *et al.*, 2014) and ultrasound (Shekhter *et al.*, 2013)

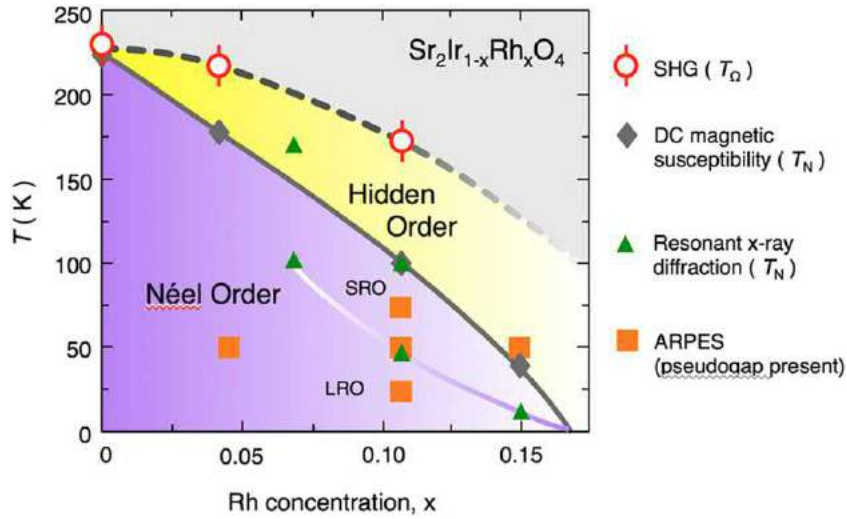


Fig. 16 Temperature versus doping phase diagram of $\text{Sr}_2\text{Ir}_{1-x}\text{Rh}_x\text{O}_4$. Onset temperatures of the hidden order (red circle), long range (LRO, green triangle) and short range (SRO, green triangle) (Clancy *et al.*, 2014) antiferromagnetic order (gray diamond) (Cao *et al.*, 2014; Qi *et al.*, 2012) are marked at multiple Rh doping levels. Points where a pseudogap (Cao *et al.*, 2014) (orange square) present are also marked, even though a pseudogap onset temperature is not known yet. Adapted from Zhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. *Nature Physics* 12(1), 32–36.

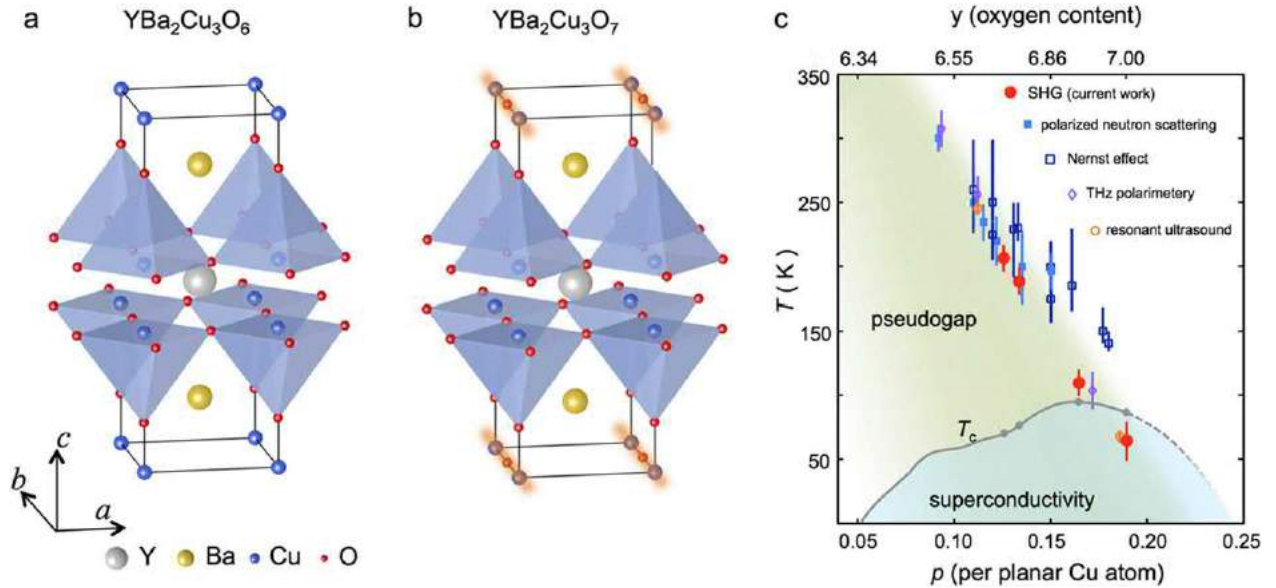


Fig. 17 Structure and phase diagram of $\text{YBa}_2\text{Cu}_3\text{O}_y$. (a–b) Tetragonal $4/mmm$ crystal structure for $\text{YBa}_2\text{Cu}_3\text{O}_6$ and orthorhombic mmm crystal structure for $\text{YBa}_2\text{Cu}_3\text{O}_7$, where the extra one oxygen per unit cell fully occupies the Cu-O chain along b-axis (orange shaded chains) in $\text{YBa}_2\text{Cu}_3\text{O}_7$. (c) Temperature versus oxygen doping (or hole doping) phase diagram of $\text{YBa}_2\text{Cu}_3\text{O}_y$ showing broken symmetry phases at the boundary of pseudogap region detected by SHG (red circles), polarized neutron scattering (blue squares) (Fauqué *et al.*, 2006; Mangin-Thro *et al.*, 2015; Mook *et al.*, 2008), Nernst effect (open navy squares) (Daou *et al.*, 2010), THz polarimetry (purple diamonds) (Lubashevsky *et al.*, 2014) and resonant ultrasound (orange circles) (Shekhter *et al.*, 2013) measurements. Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $\text{YBa}_2\text{Cu}_3\text{O}_y$ *Nature Physics* 13(3), 250–254.

measurements on $\text{YBa}_2\text{Cu}_3\text{O}_y$ suggest that the pseudogap onset temperature T^* coincides with a phase transition that breaks time-reversal, four-fold rotation, and mirror symmetries respectively (Fig. 17(c)). However, the microscopic nature of the pseudogap and its fate in the overdoped region remains poorly understood.

Zhao and colleagues approaches the nature of pseudogap by tracking its full symmetry evolution over a wide range of doping levels from underdoped ($y=6.67, 6.75$) to optimal doped ($y=6.92$) to overdoped ($y=7.0$) (Zhao *et al.*, 2016a), using the RA setup described in Fig. 4. Above T^* , linear RA patterns from $\text{YBa}_2\text{Cu}_3\text{O}_y$ measured in $S_{in} - S_{out}$ geometry (Fig. 18(a–d) insets) exhibits not

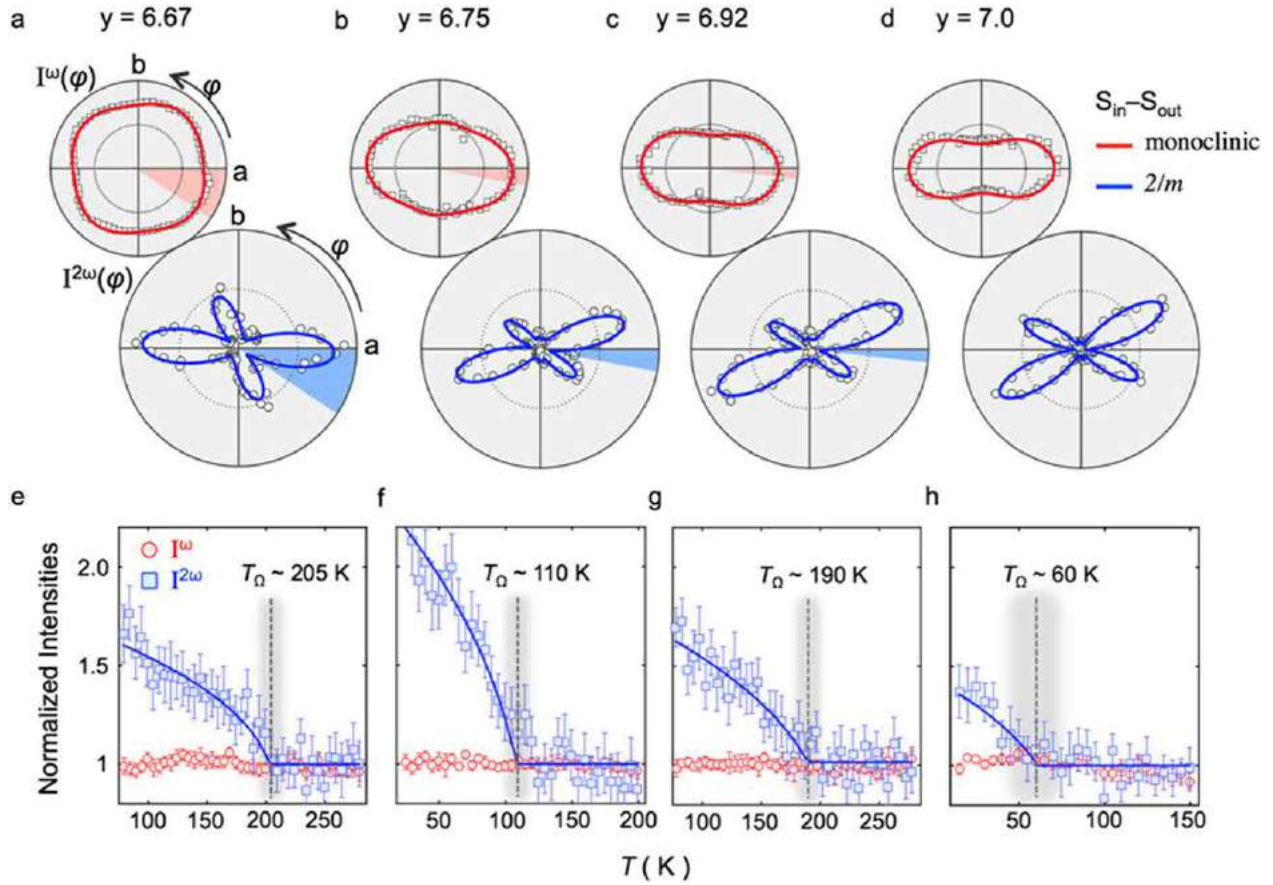


Fig. 18 Doping dependence of crystal symmetry and multipolar order in $\text{YBa}_2\text{Cu}_3\text{O}_y$. (a–d) Polar plots of linear (upper left insets) and SHG RA patterns measured at $T = 295\text{K}$ in $S_{in} - S_{out}$ polarization geometry from $\text{YBa}_2\text{Cu}_3\text{O}_y$ with $y = 6.67, 6.75, 6.92$ and 7.0 respectively. The former are fit to electric dipole induced linear response of monoclinic crystal system (red lines) and the latter are fit to electric quadrupole induced SH response of the centrosymmetry $2/m$ point group (blue line). The angular deviations of the maxima of linear RA patterns and lobe bisector of SHG RA patterns away from the a -axis are shaded with red and blue respectively. (e–h) Temperature dependence of linear and SH response of $\text{YBa}_2\text{Cu}_3\text{O}_y$ with $y = 6.67, 6.75, 6.92$ and 7.0 scaled to their room temperature values. Blue curves overlaid on SH data sets are guides to the eye. The inversion symmetry breaking transition temperature T_Ω determined from SH data are marked with dashed gray lines. The width of the shaded gray intervals represents the uncertain of T_Ω . Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $\text{YBa}_2\text{Cu}_3\text{O}_y$ Nature Physics 13(3), 250–254.

only more pronounced anisotropy with increasing hole doping as expected from the filling of CuO chains, but also a rotation of the intensity maxima and minima away from the a and b -axes indicating an absence of m_{ac} and m_{bc} symmetries consistent with a monoclinic distortion. SHG RA data (Fig. 18(a–d)) further narrows down the symmetry point group to be the centrosymmetric $2/m$ out of the monoclinic class, and the source of SHG response therefore to be bulk electric quadrupole. The degree of monoclinicity, qualitatively tracked by via the angular deviations of the RA patterns away from a -axis, decreases monotonically over the range between $y = 6.67$ and $y = 7$, suggesting that this structural refinement from mmm to $2/m$ originates from the vacancy induced monoclinic distortions in the oxygen sublattice.

For all four doping levels shown in Fig. 18(e–h), the linear responses have no observable temperature dependence while, in sharp contrast, SHG responses exhibit a significant order parameter like upturn below a doping dependent critical temperature T_Ω . This dichotomy is naturally reconciled if the bulk inversion symmetry is broken below T_Ω , and turns on a stronger electric dipole induced SHG radiation on top of the pre-existing electric quadrupole response. More importantly, the onset of this inversion symmetry broken order T_Ω coincides with the pseudogap onset T^* in the underdoped and optimal doped region, and the boundary of this order extends into the superconducting dome in the overdoped region as shown in Fig. 17(c). The absence of any measurable anomalies in the SHG intensity across either the superconducting or the charge order temperature indicates that this non-centrosymmetric phase is independent of and coexists with both of them.

In fact, the enhanced SHG signal in $S_{in} - S_{out}$ geometry below T_Ω rules out the possibility of two-fold rotation symmetry C_2 present in the inversion symmetry broken phase, because the ED induced SHG inferred above is strictly forbidden by symmetry in $S_{in} - S_{out}$ geometry for any point that contains C_2 . As a result, the phase below T_Ω breaks both inversion symmetry and C_2 , and has the symmetry of either of the two magnetic point group $2'/m$ and $m1'$, or their subgroups. Therefore, the SHG RA pattern below T_Ω

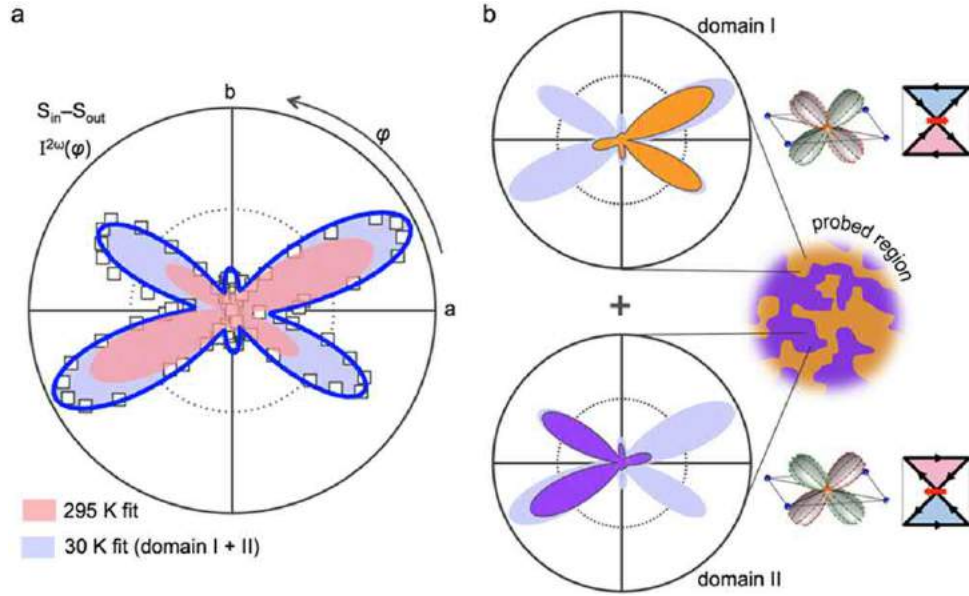


Fig. 19 Point group symmetry in the pseudogap region of $\text{YBa}_2\text{Cu}_3\text{O}_y$. (a) Polar plot of SH RA pattern measured at $T=30\text{K}$ in $S_{in} - S_{out}$ polarization geometry from $\text{YBa}_2\text{Cu}_3\text{O}_{6.92}$ (squares). The blue curve is a best fit to the average of two 180° rotated domains with $2'/m$ point group symmetry. The shaded pink area is the fit to the SH RA pattern at $T=295\text{K}$ that is overlaid for comparison. (b) Decomposition of the fit to the $R=30\text{K}$ data (shaded blue area) into its individual domain contributions (shaded orange and purple areas), with the schematics of two degenerate magnetic quadrupole or magneto-electric loop current configurations shown on the right side correspondingly. A schematic of the domain structure inside our beam spot is shown to the right. Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $\text{YBa}_2\text{Cu}_3\text{O}_y$ Nature Physics 13(3), 250–254.

($S_{in} - S_{out}$ geometry, Fig. 19(a)) seemingly preserves two-fold rotation symmetry because of spatial averaging over domains of two degenerate ground states whose order parameters are related by 180° -rotation about c -axis (Fig. 19(b)). This implies a characteristic domains size that is much smaller than the laser spot size. Moreover, the low symmetry ($2'/m$ or $m1'$) underlying the pseudogap cannot be explained by stripe or nematic type orders alone, but instead suggest the presence of an odd-parity magnetic order parameter, which is consistent with theoretical proposals involving a ferroic ordering of local Cu-site magnetic-quadrupoles (inset 1 of Fig. 19(b)) (Fechner *et al.*, 2016; Lovesey *et al.*, 2015), current loops circulating within Cu-O octahedra (inset 2 of Fig. 19(b)) (Simon and Varma, 2002, 2003; Varma, 1997), O-site moments (Moskvin, 2012) or magneto-electric multipoles (Fechner *et al.*, 2016).

Although RA measurements above cannot identify the specific microscopic origin out of a number of candidates above, the results suggest a boundary of an odd-parity magnetic phase for the pseudogap region that extends inside the superconducting dome slightly beyond optimal doping. It is worth noting that this phase is similar to the one discussed in the pseudogap region of Sr_2IrO_4 , a cuprate analogue, in part iii above, which indicates a robust connection between pseudogap and this odd-parity hidden order.

Conclusions and Prospects

SHG RA experiments in correlated electron systems show that nonlinear optics is not only a powerful complement to existing diffraction techniques to refine subtle structural distortions, but also a unique probe for detecting multipolar orders hidden to most existing techniques. While the elementary symmetry point group analysis can already tell a lot about the symmetry properties of the multipolar order, such as the odd-parity magnetic order in both Sr_2IrO_4 and $\text{YBa}_2\text{Cu}_3\text{O}_y$, a more sophisticated group theory analysis together with Landau theory can advance the understanding on the orders and their couplings to a significant depth, such as the primary odd-parity nematic order and the secondary structural order in $\text{Cd}_2\text{Re}_2\text{O}_7$.

The relative SHG amplitude of the hidden phase contributions to the crystallographic structure background exceeds the linear counterpart by orders of magnitude. In the case of multipolar order, the vector field of linear optics fails to couple with the tensor order parameter effectively, so that the nonlinear optical techniques become especially powerful. For example, in $\text{YBa}_2\text{Cu}_3\text{O}_y$, linear optical response can only determine the monoclinic crystal class for the room temperature structure, and shows no observable changes upon the emergence of the odd-parity magnetic order and pseudogap. On the other hand, SHG refines the structure to the centrosymmetric $2/m$ point group, and further resolves the broken inversion and two-fold rotation symmetries in the pseudogap region. Such advantages of SHG not only comes from its great sensitivity to symmetries, in particular, inversion symmetry, but also benefits from the tensor fields formed by more than one copy of magneto-electric fields.

So far, we have only focused on SHG, and would like to briefly discuss some possibilities of other second order nonlinear optics including SFG and DFG in this paragraph. First, the degenerate nature of SHG enforces the permutation symmetry of the last two indices of electric dipole induced SHG susceptibility tensor, i.e., $\chi_{ijk} = \chi_{ikj}$, which can forbid SHG response from some non-centrosymmetric system (Boyd, 2003). For example, $\chi = \{\chi_{xyz} = -\chi_{xzy} = \chi_{yxz} = -\chi_{yzx} = \chi_{zxy} = -\chi_{zyx}\}$ for cubic point group 432, and it will turn into zero after the SHG permutation symmetry $\{\chi_{xyz} = \chi_{xzy} = \chi_{yzx} = \chi_{yxz} = \chi_{zyx} = \chi_{zxy}\}$. In order to probe systems with symmetry such as 432, it is essential to apply SFG or DFG, rather than SHG. Second, as we have discussed in Sections “Generations of Rotational Anisotropy Experimental Designs and An Odd-Parity Hidden Order in Sr_2IrO_4 ”, the more control we have on selecting tensor elements, the higher chance we have to capture hidden phases. However, in SHG measurements, there is no independent control over the two copies of incoming electromagnetic fields, so that it is in general hard to select individual tensor elements. Alternatively, SFG and DFG overcome this challenge well thanks to their two independent incoming fields. Third, in correlated electron systems, the impact of emergent orders on the electronic states usually happens to the ones within a couple of hundreds meV around the Fermi level whose resonant energy is much smaller than optical photon energy and hardly accessible by optical SHG or SFG. On the other hand, DFG, with the photon energy difference between the two incoming fields matching that of the changing states, can be especially sensitive to such phase transitions.

In this review, we have discussed only a few examples that demonstrate the possibilities of SHG RA measurements probing hidden multipolar orders in correlated electron system. We believe that the great potential of nonlinear optics and second order nonlinear optics in particular are far from being well explored, and foresee its power in revealing unconventional symmetry broken phases in correlated materials. Besides static measurements that we discussed through the review, the ultrashort laser pulses meant for strong magnetoelectric fields to enhance nonlinear responses are naturally compatible with time-resolved pump-and-nonlinear-probe measurements, to explore the dynamics of the hidden orders, disentangle their couplings with coexisting orders, or even probe the transient photo-induced symmetry breaking phases.

References

- Birss, R.R., 1964. *Symmetry and Magnetism*, 863. Amsterdam: North-Holland.
- Boyd, R.W., 2003. *Nonlinear optics*. New York, NY: Academic Press.
- Cao, Y., Wang, Q., Waugh, J.A., *et al.*, 2016. Hallmarks of the Mott-Metal crossover in the hole doped $J=1/2$ Mott insulator Sr_2IrO_4 . *Nature communications* 7, 11367.
- Castellan, J.P., Gaulin, B.D., Van Duijn, J., *et al.*, 2002. Structural ordering and symmetry breaking in $\text{Cd}_2\text{Re}_2\text{O}_7$. *Physical Review B* 66 (13), 134528.
- Clancy, J.P., Lupascu, A., Gretarsson, H., *et al.*, 2014. Dilute magnetism and spin-orbital percolation effects in $\text{Sr}_2\text{Ir}_{1-x}\text{Rh}_x\text{O}_4$. *Physical Review B* 89 (5), 054409.
- Clark, D.J., Senthikumar, V., Le, C.T., *et al.*, 2014. Strong optical nonlinearity of CVD-grown MoSe_2 monolayer as probed by wavelength-dependent second-harmonic generation. *Physical Review B* 90 (12), 121409.
- Dähn, A., Hübner, W., Bennemann, K.H., 1996. Symmetry analysis of the nonlinear optical response: Second harmonic generation at surfaces of antiferromagnets. *Physical Review Letters* 77 (18), 3929.
- Daou, R., Chang, J., LeBoeuf, D., *et al.*, 2010. Broken rotational symmetry in the pseudogap phase of a high- T_c superconductor. *Nature* 463 (7280), 519–522.
- Fauqué, B., Sidis, Y., Hinkov, V., *et al.*, 2006. Magnetic order in the pseudogap phase of high- T_c superconductors. *Physical Review Letters* 96 (19), 197001.
- Fechner, M., Fierz, M. J. A., Thole, F., *et al.*, 2016. Quasistatic magnetoelectric multipoles as order parameter for pseudogap phase in cuprate superconductors. *Physical Review B* 93 (17), 174419.
- Fiebig, M., Fröhlich, D., Krichevskov, B.B., *et al.*, 1994. Second harmonic generation and magnetic-dipole-electric-dipole interference in antiferromagnetic Cr_2O_3 . *Physical Review Letters* 73 (15), 2127.
- Fiebig, M., Fröhlich, D., Kohn, K., *et al.*, 2000. Determination of the magnetic symmetry of hexagonal manganites by second harmonic generation. *Physical Review Letters* 84 (24), 5620.
- Fiebig, M., Pavlov, V.V., Pisarev, R.V., 2005. Second-harmonic generation as a tool for studying electronic and magnetic structures of crystals: Review. *JOSA B* 22 (1), 96–118.
- Franken, P.A., Hill, A.E., Peters, C.W., *et al.*, 1961. Generation of optical harmonics. *Physical Review Letters* 7 (4), 118.
- Fu, L., 2015. Parity-breaking phases of spin-orbit-coupled metals with gyrotropic, ferroelectric, and multipolar orders. *Physical Review Letters* 115 (2), 026401.
- Gridnev, V.N., Pavlov, V.V., Pisarev, R.V., *et al.*, 2001. Second harmonic generation in anisotropic magnetic films. *Physical Review B* 63 (18), 184407.
- Hamh, S.Y., Park, S.H., Jerng, S.K., *et al.*, 2016. Surface and interface states of Bi_2Se_3 thin films investigated by optical second-harmonic generation and terahertz emission. *Applied Physics Letters* 108 (5), 051609.
- Harter, J.W., Niu, L., Woss, A.J., *et al.*, 2015. High-speed measurement of rotational anisotropy nonlinear optical harmonic generation using position-sensitive detection. *Optics Letters* 40 (20), 4671–4674.
- Harter, J.W., Chu, H., Jiang, S., *et al.*, 2016. Nonlinear and time-resolved optical study of the 112-type iron-based superconductor parent $\text{Ca}_{1-x}\text{La}_x\text{FeAs}_2$ across its structural phase transition. *Physical Review B* 93 (10), 104506.
- Harter, J.W., Zhao, Z.Y., Yan, J.Q., Mandrus, D., Hsieh, D., 2017. A parity-breaking electronic nematic phase transition in the spin-orbit coupled correlated metal $\text{Cd}_2\text{Re}_2\text{O}_7$. *Science* 356 (6335), 295.
- Heinz, T.F., Loy, M.M.T., Thompson, W.A., 1985. Study of Si (111) surfaces by optical second-harmonic generation: reconstruction and surface phase transformation. *Physical Review Letters* 54 (1), 63.
- Hogan, T., Bjaalie, L., Zhao, L., *et al.*, 2016. Structural investigation of the bilayer iridate $\text{Sr}_3\text{Ir}_2\text{O}_7$. *Physical Review B* 93 (13), 134110.
- Hsieh, D., McIver, J.W., Torchinsky, D.H., *et al.*, 2011. Nonlinear optical probe of tunable surface electrons on a topological insulator. *Physical Review Letters* 106 (5), 057401.
- Huai, P., Nasu, K., 2002. Difference between photoinduced phase and thermally excited phase. *Journal of the Physical Society of Japan* 71 (4), 1182–1188.
- Ichikawa, H., Nozawa, S., Sato, T., *et al.*, 2011. Transient photoinduced ‘hidden’ phase in a manganite. *Nature Materials* 10 (2), 101–105.
- Jackson, J.D., 1999. *Classical Electrodynamics*. New York, NY: Wiley.
- Janisch, C., Wang, Y., Ma, D., *et al.*, 2014. Extraordinary second harmonic generation in tungsten disulfide monolayers. *Scientific Reports* 4, 5530.
- Kendziora, C.A., Sergienko, I.A., Jin, R., *et al.*, 2005. Goldstone-mode phonon dynamics in the pyrochlore $\text{Cd}_2\text{Re}_2\text{O}_7$. *Physical Review Letters* 95 (12), 125503.
- Kim, B.J., Jin, H., Moon, S.J., *et al.*, 2008. Novel $J=1/2$ Mott state induced by relativistic spin-orbit coupling in Sr_2IrO_4 . *Physical Review Letters* 101 (7), 076402.
- Kim, B.J., Ohsumi, H., Komesu, T., *et al.*, 2009. Phase-sensitive observation of a spin-orbital Mott state in Sr_2IrO_4 . *Science* 323 (5919), 1329–1332.
- Kim, D.H., Lim, D., 2015. Optical second-harmonic generation in few-layer MoSe_2 . *Journal of the Korean Physical Society* 66 (5), 816–820.

- Kirilyuk, A., Rasing, T., 2005. Magnetization-induced-second-harmonic generation from surfaces and interfaces. *JOSA B* 22 (1), 148–167.
- Koshihara, S.-y., Adachi, S.-i., 2006. Photo-induced phase transition in an electron–lattice correlated system – future role of a time-resolved X-ray measurement for materials science. *Journal of the Physical Society of Japan* 75 (1), 011005.
- Kumar, A., Rai, R.C., Podraza, N.J., *et al.*, 2008. Linear and nonlinear optical properties of BiFeO₃. *Applied Physics Letters* 92 (12), 121915.
- Kumar, N., Najmaei, S., Cui, Q., *et al.*, 2013. Second harmonic microscopy of monolayer MoS₂. *Physical Review B* 87 (16), 161403.
- Kung, Y.F., Chen, C.C., Moritz, B., *et al.*, 2014. Numerical exploration of spontaneous broken symmetries in multiorbital Hubbard models. *Physical Review B* 90 (22), 224507.
- Kuramoto, Y., Kusunose, H., Kiss, A., 2009. Multipole orders and fluctuations in strongly correlated electron systems. *Journal of the Physical Society of Japan* 78 (7), 072001.
- Lafrentz, M., Brunne, D., Kaminski, B., *et al.*, 2012. Optical third harmonic generation in the magnetic semiconductor EuSe. *Physical Review B* 85 (3), 035206.
- Landau, L., 1936. The theory of phase transitions. *Nature* 138, 840–841.
- Landau, L.D., 1937. On the theory of phase transitions. I. *Zhurnal Eksperimentalnoi Teoreticheskoi Fiziki* 11, 19.
- Li, Y., Rao, Y., Mak, K.F., *et al.*, 2013. Probing symmetry properties of few-layer MoS₂ and h-BN by optical second-harmonic generation. *Nano Letters* 13 (7), 3329–3333.
- Lovesey, S.W., Khalyavin, D.D., Staub, U., 2015. Ferro-type order of magneto-electric quadrupoles as an order-parameter for the pseudo-gap phase of a cuprate superconductor. *Journal of Physics: Condensed Matter* 27 (29), 292201.
- Lubashevsky, Y., Pan, L., Kirzhner, T., *et al.*, 2014. Optical birefringence and dichroism of cuprate superconductors in the THz regime. *Physical Review Letters* 112 (14), 147001.
- Malard, L.M., Alencar, T.V., Barboza, A.P.M., *et al.*, 2013. Observation of intense second harmonic generation from MoS₂ atomic crystals. *Physical Review B* 87 (20), 201401.
- Mangin-Thro, L., Sidis, Y., Wildes, A., Bourges, P., 2015. Intra-unit-cell magnetic correlations near optimal doping in YBa₂Cu₃O_{6.85}. *Nature Communications* 6, 7705.
- McIver, J.W., Hsieh, D., Drapcho, S.G., *et al.*, 2012. Theoretical and experimental study of second harmonic generation from the surface of the topological insulator Bi₂Se₃. *Physical Review B* 86 (3), 035327.
- Mook, H.A., Sidis, Y., Fauque, B., *et al.*, 2008. Observation of magnetic order in a superconducting YBa₂Cu₃O_{6.6} single crystal using polarized neutron scattering. *Physical Review B* 78 (2), 020506.
- Moskvin, A.S., 2012. Pseudogap phase in cuprates: Oxygen orbital moments instead of circulating currents. *JETP Letters* 96 (6), 385–390.
- Murakami, Y., Hill, J.P., Gibbs, D., *et al.*, 1998. Resonant X-ray scattering from orbital ordering in LaMnO₃. *Physical Review Letters* 81 (3), 582.
- Mydosh, J.A., Oppeneer, P.M., 2011. Colloquium: Hidden order, superconductivity, and magnetism: The unsolved case of URu₂Si₂. *Reviews of Modern Physics* 83 (4), 1301.
- Mydosh, J.A., Oppeneer, P.M., 2014. Hidden order behaviour in URu₂Si₂ (A critical review of the status of hidden order in 2014). *Philosophical Magazine* 94 (32–33), 3642–3662.
- Nasu, K., 2004. Photoinduced phase transitions. Singapore: World Scientific.
- Nye, J.F., 1985. Physical properties of crystals: Their representation by tensors and matrices. Oxford: Oxford university press.
- Nývlt, M., Bisio, F., Kirschner, J., 2008. Second harmonic generation study of the antiferromagnetic NiO (001) surface. *Physical Review B* 77 (1), 014435.
- Orenstein, J., 2011. Optical nonreciprocity in magnetic structures related to high-T_c superconductors. *Physical Review Letters* 107 (6), 067002.
- Pan, R.-P., Wei, H.D., Shen, Y.R., 1989. Optical second-harmonic generation from magnetized surfaces. *Physical Review B* 39 (2), 1229.
- Petersen, J.C., Caswell, M.D., Dodge, J.S., *et al.*, 2006. Nonlinear optical signatures of the tensor order in CdRe₂O₇. *Nature Physics* 2 (9), 605–608.
- Qi, T.F., Korneta, O.B., Li, L., *et al.*, 2012. Spin-orbit tuned metal-insulator transitions in single-crystal Sr₂Ir_{1-x}Rh_xO₄ (0 ≤ x ≤ 1). *Physical Review B* 86 (12), 125105.
- Rabe, K.M., 2007. Solid-state physics: Response with a twist. *Nature* 449 (7163), 674–675.
- Reif, J., Zink, J.C., Schneider, C.M., *et al.*, 1991. Effects of surface magnetism on optical second harmonic generation. *Physical Review Letters* 67 (20), 2878.
- Santini, P., Carretta, S., Amoretti, G., *et al.*, 2009. Multipolar interactions in f-electron systems: The paradigm of actinide dioxides. *Reviews of Modern Physics* 81 (2), 807.
- Seyler, K.L., Schaibley, J.R., Gong, P., *et al.*, 2015. Electrical control of second-harmonic generation in a WSe₂ monolayer transistor. *Nature Nanotechnology* 10 (5), 407–411.
- Shah, N., Chandra, P., Coleman, P., Mydosh, J.A., 2000. Hidden order in URu₂Si₂. *Physical Review B* 61 (1), 564.
- Shaked, H., 1994. Crystal structures of the high-T_c superconducting copper-oxides. Elsevier Science Publishers BV.
- Shannon, V.L., Koos, D.A., Richmond, G.L., 1987a. Changes in the second harmonic rotational anisotropy for a silver (111) electrode as a function of bias potential. *Journal of Physical Chemistry* 91 (22), 5548–5551.
- Shannon, V.L., Koos, D.A., Richmond, G.L., 1987b. Second harmonic generation for in situ analysis of electrode surface structure. *Applied Optics* 26 (17), 3579–3583.
- Shekhter, A., Ramshaw, B.J., Liang, R., *et al.*, 2013. Bounding the pseudogap with a line of phase transitions in YBa₂Cu₃O₆ + [dgr]. *Nature* 498 (7452), 75–77.
- Shen, Y.-R., 1984. Principles of nonlinear optics.
- Shen, Y.R., 1989. Optical second harmonic generation at interfaces. *Annual Review of Physical Chemistry* 40 (1), 327–350.
- Simon, M.E., Varma, C.M., 2002. Detection and implications of a time-reversal breaking state in underdoped cuprates. *Physical Review Letters* 89 (24), 247003.
- Simon, M.E., Varma, C.M., 2003. Symmetry considerations for the detection of second-harmonic generation in cuprates in the pseudogap phase. *Physical Review B* 67 (5), 054511.
- Spaldin, N.A., Fiebig, M., Mostovoy, M., 2008. The toroidal moment in condensed-matter physics and its relation to the magnetoelectric effect. *Journal of Physics: Condensed Matter* 20 (43), 434203. (All authors contributed equally to this work).
- Squires, G.L., 2012. Introduction to the theory of thermal neutron scattering. Cambridge: Cambridge University Press.
- Tokura, Y., 2006. Photoinduced phase transition: A tool for generating a hidden state of matter. *Journal of the Physical Society of Japan* 75 (1), 011001.
- Tom, H.W.K., Heinz, T.F., Shen, Y.R., 1983. Second-harmonic reflection from silicon surfaces and its relation to structural symmetry. *Physical Review Letters* 51 (21), 1700.
- Torchinsky, D.H., Chu, H., Qi, T., *et al.*, 2014. A low temperature nonlinear optical rotational anisotropy spectrometer for the determination of crystallographic and electronic symmetries. *Review of Scientific Instruments* 85 (8), 083102.
- Torchinsky, D.H., Chu, H., Zhao, L., *et al.*, 2015. Structural distortion-induced magnetoelastic locking in Sr₂IrO₄ revealed through nonlinear optical harmonic generation. *Physical Review Letters* 114 (9), 096404.
- Van Aken, B.B., Rivera, J.-P., Schmid, H., Fiebig, M., 2007. Observation of ferrotoroidic domains. *Nature* 449 (7163), 702–705.
- Varma, C.M., 1997. Non-Fermi-liquid states and pairing instability of a general model of copper oxide metals. *Physical Review B* 55 (21), 14554.
- Warren, B.E., 1969. X-ray Diffraction. Chelmsford, MA: Courier Corporation.
- Weber, C., Lauchli, A., Mila, F., Giamarchi, T., 2009. Orbital currents in extended Hubbard models of high-T_c cuprate superconductors. *Physical Review Letters* 102 (1), 017005.
- Weller, M.T., Hughes, R.W., Rooke, J., Knee, C.S., Reading, J., 2004. The pyrochlore family – a potential panacea for the frustrated perovskite chemist. *Dalton Transactions* 19, 3032–3041.
- Witczak-Krempa, W., Chen, G., Kim, Y. B., Balents, L., 2014. Correlated quantum phenomena in the strong spin-orbit regime. *Annual Review of Condensed Matter Physics* 5 (1), 57.
- Wu, L., Patankar, S., Morimoto, T., *et al.*, 2016. Giant anisotropic nonlinear optical response in transition metal monophosphide Weyl semimetals. *Nature Physics* 13, 350–355. (Advance online publication).
- Wu, W., Wang, L., Li, Y., *et al.*, 2014. Piezoelectricity of single-atomic-layer MoS₂ for energy conversion and piezotronics. *Nature* 514 (7523), 470–474.
- Yakovenko, V.M., 2015. Tilted loop currents in cuprate superconductors. *Physica B: Condensed Matter* 460, 159–164.

- Yamada, C., Kimura, T., 1993. Anisotropy in second-harmonic generation from reconstructed surfaces of GaAs. *Physical Review Letters* 70 (15), 2344.
- Yamaura, J.-I., Hiroi, Z., 2002. Low temperature symmetry of pyrochlore oxide $\text{Cd}_2\text{Re}_2\text{O}_7$. *Journal of the Physical Society of Japan* 71 (11), 2598–2600.
- Yin, X., Ye, Z., Chenet, D.A., *et al.*, 2014. Edge nonlinear optics on a MoS_2 atomic monolayer. *Science* 344 (6183), 488–490.
- Zhang, J., Tan, X., Liu, M., *et al.*, 2015. Cooperative photoinduced metastable phase control in strained manganite films. *Nature Physics* 15, 956.
- Zhao, L., Belvin, C.A., Liang, R., *et al.*, 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $\text{YBa}_2\text{Cu}_3\text{O}_y$. *Nature Physics* 13 (3), 250–250.
- Zhao, L., Torchinsky, D., Chu, H., *et al.*, 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. *Nature Physics* 12 (1), 32–36.
- Zimmermann, A.S., Meier, D., Fiebig, M., 2014. Ferroic nature of magnetic toroidal order. *Nature Communications* 5, 4796, 2014.
- Zou, X., Hovmöller, S., Oleynikov, P., 2011. *Electron Crystallography: Electron Microscopy and Electron Diffraction*, 16. Oxford: Oxford University Press.

Saturated Absorption Spectroscopy for Diode Laser Locking

Bachana Lomsadze, University of Michigan, Ann Arbor, MI, United States

© 2018 Elsevier Inc. All rights reserved.

Introduction

Since their development about three decades ago, external-cavity diode lasers (ECDLs) have become very attractive coherent light sources for many applications. They are typically used for atomic and molecular spectroscopy, laser cooling and trapping, optical telecommunications, etc. The features that make ECDLs attractive are: their narrow linewidth, broadly tunable wavelength, low cost, compactness, and relative ease of construction.

Using external-cavity feedback configurations, one is able to achieve laser linewidths of hundreds of kHz. However, temperature fluctuations, vibrations, electronic noise, change in atmospheric pressure, etc., will broaden the laser linewidth and shift the center frequency. For some applications, for example, cooling and trapping, a drift of the center frequency of a few MHz for even a short period of time can be problematic. For these applications laser frequency locking is required.

Over the years, multiple laser frequency locking techniques have been developed. One of the most commonly used methods is the Pound–Drever–Hall (PDH) method which locks the frequency of a laser to a stable Fabry–Perot cavity reference. Alternatively, if the laser frequency must be locked to some atomic resonance then a setup using saturated absorption spectroscopy can be employed. This method has many different variations itself.

In this article the technique of locking a laser to a Rb hyperfine transition using saturated absorption spectroscopy with a room temperature Rb atomic vapor cell as a reference will be discussed. But, the method described here can be used with any reference cell.

This article is divided into four sections in which (1) the external-cavity feedback configurations are reviewed, (2) discusses the principles of saturation absorption spectroscopy, (3) shows the experimental arrangement used for laser locking, and (4) concludes the article.

External-Cavity Configurations

Two ECDL configurations are widely used today: the Littrow and Littman–Metcalf configurations. Simplified schematic diagrams of these configurations are shown in Fig. 1. They consist of a diode laser (injection current diode) with a highly reflective ($> 90\%$) rear facet and antireflection coated front facet, a collimating lens, and a dispersive element for wavelength tuning (typically a diffraction grating). In both configurations the 1st-order diffracted beam is used for laser wavelength tuning.

For the Littrow configuration (Fig. 1(a)), the 1st-order diffracted beam from the grating is retro-reflected back into the laser diode for feedback and the laser wavelength tuning is accomplished by rotating the grating. For some applications, for example, when broad wavelength tuning is required, this configuration is not preferred because a change in the grating orientation causes a change in the laser output beam (0th order) pointing. However, this problem can be resolved with a pointing stabilization feedback arrangement.

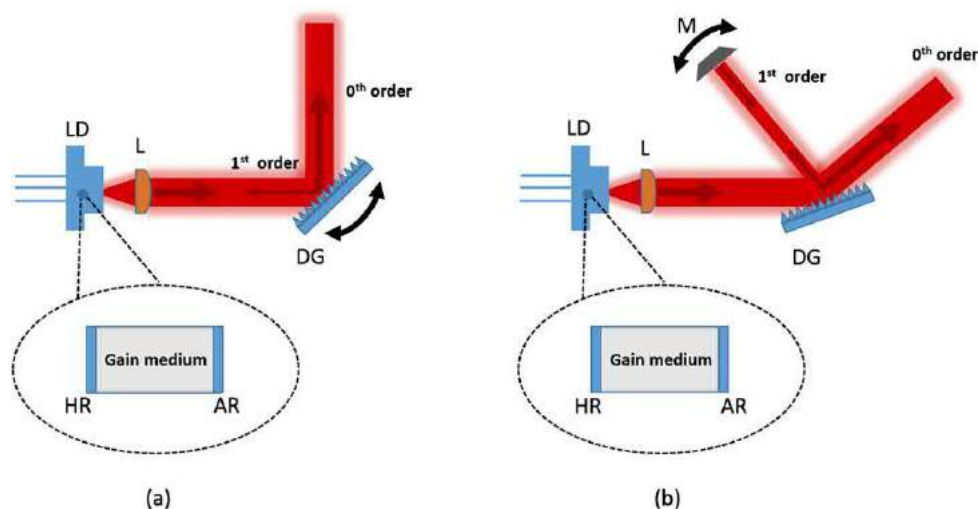


Fig. 1 Schematic diagram of Littrow (a) and Littman–Metcalf (b) external-cavity configurations. AR, antireflective coated facet; DG, diffraction grating; HR, high reflective facet; L, collimating Lens; LD, laser diode; M, mirror.

For the Littman–Metcalf configuration (Fig. 1(b)), the grating position is fixed and the wavelength is tuned by rotating a mirror. This configuration usually generates an output beam with a narrower linewidth because the 1st order beam refracts twice from the grating.

Principles of Saturated Absorption Spectroscopy

Fig. 2 shows the hyperfine energy level diagrams of Rubidium atoms' D₂ lines for both isotopes (⁸⁷Rb and ⁸⁵Rb).

The natural linewidth of the hyperfine lines (full-width at half maximum) is about 6 MHz. However, the measured transmission spectrum of a room temperature Rb vapor cell shows broadening of these lines up to 500 MHz (see Fig. 3) which is due to the velocity distribution of atoms in the reference cell and the Doppler Effect. This data was obtained by measuring the trans-

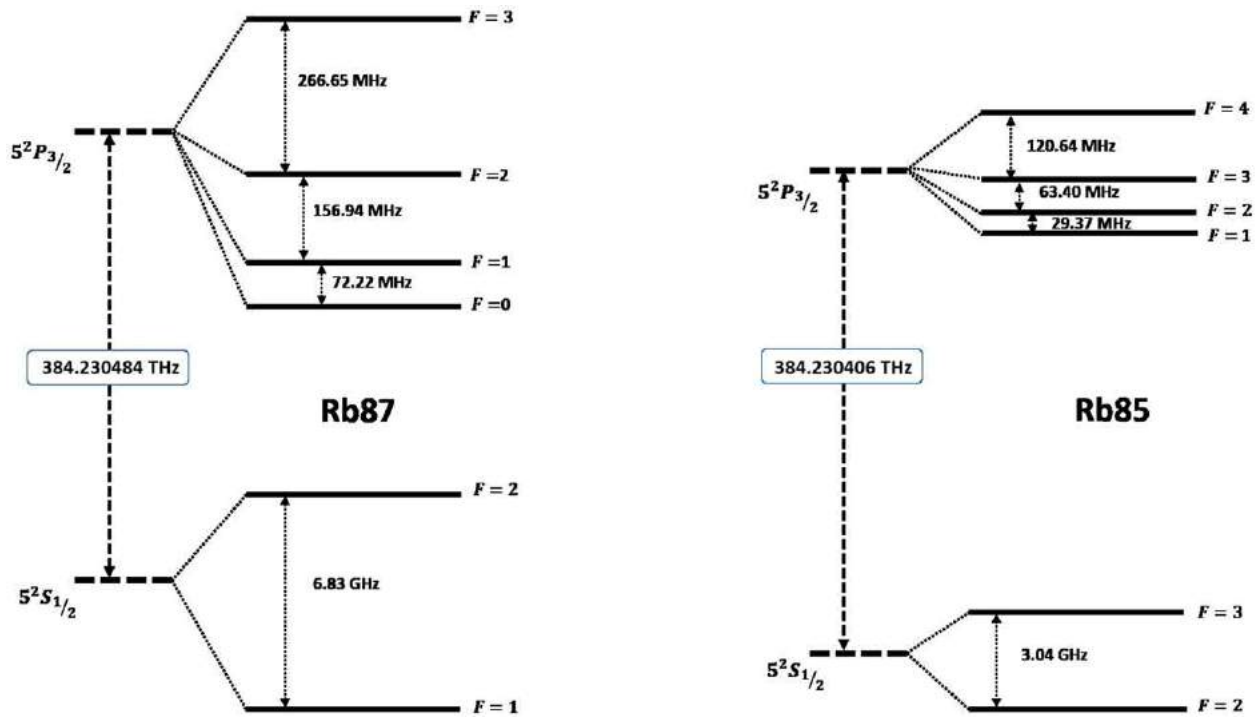


Fig. 2 Hyperfine energy level diagram for ⁸⁷Rb (left) and ⁸⁵Rb (right) D₂ lines (not to scale).

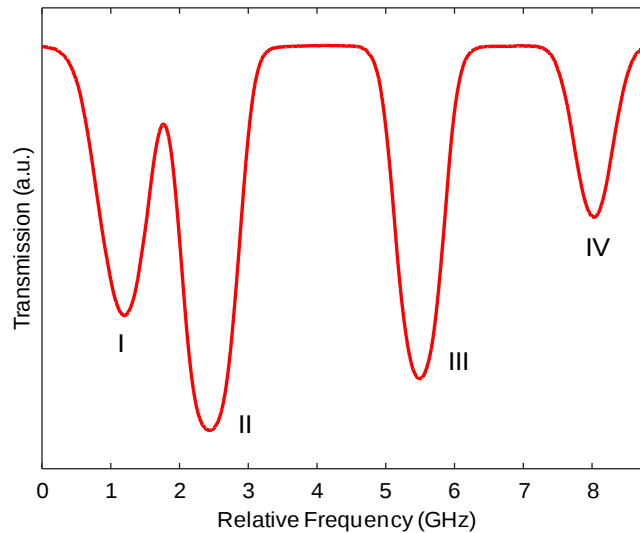


Fig. 3 Transmission profile of a room temperature Rb cell near 780.24 nm.

mission of a beam of laser light from an ECDL through the vapor cell while scanning its center frequency over a 9 GHz range centered near 780.24 nm (384.23 THz).

I and IV dips correspond to ^{87}Rb transitions from $5S_{1/2}$ $F=2$ and $5S_{1/2}$ $F=1$ to $5P_{3/2}$ states, respectively. II and III dips correspond to ^{85}Rb transitions from $5S_{1/2}$ $F=3$ and $5S_{1/2}$ $F=2$ to $5P_{3/2}$ states, respectively. Obviously, none of these broad absorption dips can be used to lock a laser onto a hyperfine resonance line.

One way to eliminate Doppler broadening is to cool the atoms down to hundreds of microkelvin temperatures using a complex optical cooling and trapping experimental setup.

Doppler Free Spectroscopy

Doppler broadening can also be eliminated without having to cool the atoms by using the simple arrangement shown in Fig. 4. Both strong (pump) and weak (probe) beams are generated from the same laser source (hence they have the same wavelength) and counter-propagate in the Rb cell.

The measured transmission spectrum of the probe beam after overlapping with the pump beam in the reference cell is shown in Fig. 5(I). The structure shown represent all dipole allowed transitions between hyperfine energy levels of both Rb isotopes. The spectrum, showing only the $5S_{1/2}$ ($F=2$) to $5P_{3/2}$ (all allowed hyperfine states) transitions for ^{87}Rb is plotted in Fig. 5(II). The elimination of Doppler broadening by counter propagating beams can be explained using a simple 2 level system.

Fig. 6(a) shows a two level system and the number density function of ground state Rb atoms as a function of their velocities (Maxwell-Boltzmann distribution) along the probe beam propagation axis at room temperature. Positive and negative velocities correspond to atoms moving in the same and opposite directions as the probe beam, respectively. When the laser frequency is detuned from the resonance frequency ($\omega_L < \omega$ or $\omega_L > \omega$), then both pump and probe beams interact with the same magnitude but the opposite sign velocity group atoms (due to the Doppler Effect). This interaction brings atoms from the ground state ($|g\rangle$) into the excited state ($|e\rangle$) and hence modifies the density function (Fig. 6(b) and (c)). Clearly the effect from the pump beam is stronger because of its high intensity. This effect is called the "hole burning." When the laser frequency is close to the atomic

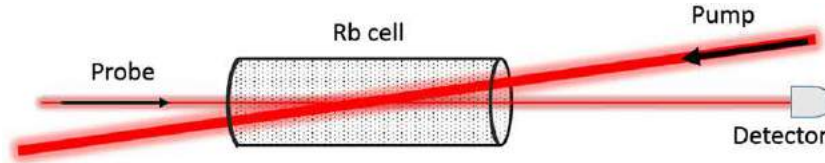


Fig. 4 Schematic diagram of counter propagating pump and probe beams overlapped inside the Rb cell.

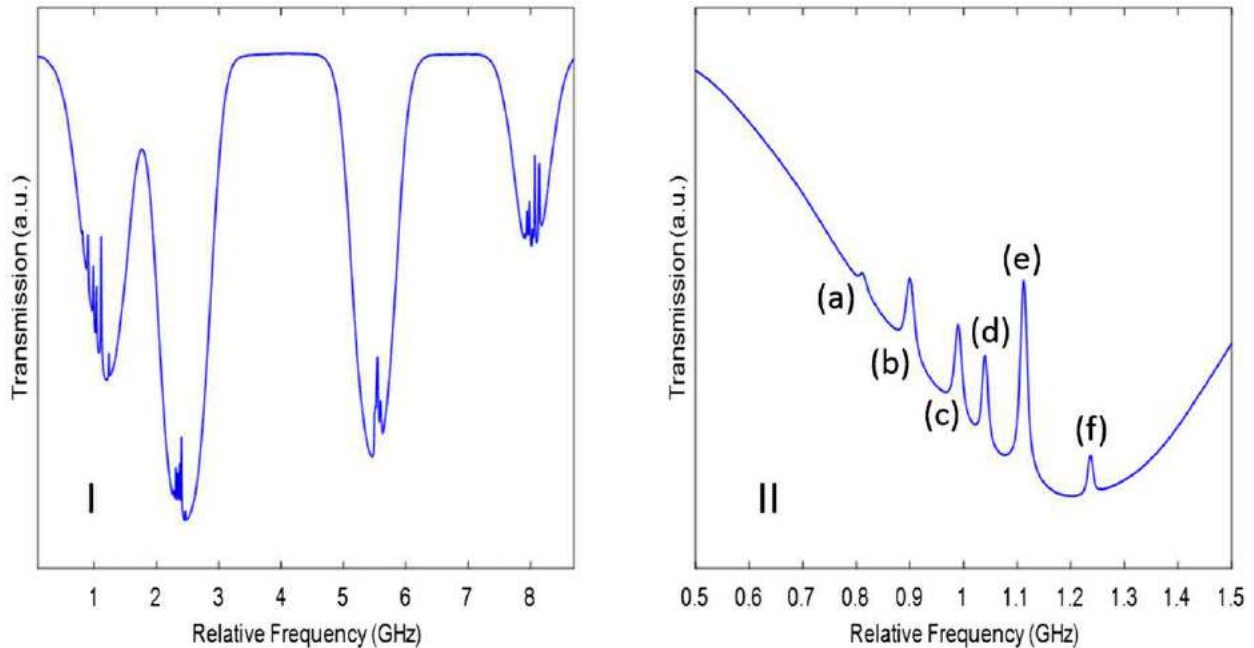


Fig. 5 (I) Transmission profile of the probe beam overlapped with the pump beam in a room temperature Rb reference cell. (II) Hyperfine structure of the $5S_{1/2}$ to $5P_{3/2}$ transitions for ^{87}Rb isotope (a) $F=2$ to $F=1$, (c) $F=2$ to $F=2$, (f) $F=2$ to $F=3$, (b) "crossover" resonance between (a) and (c), (d) "crossover" resonance between (a) and (f), (e) "crossover" resonance between (c) and (f).

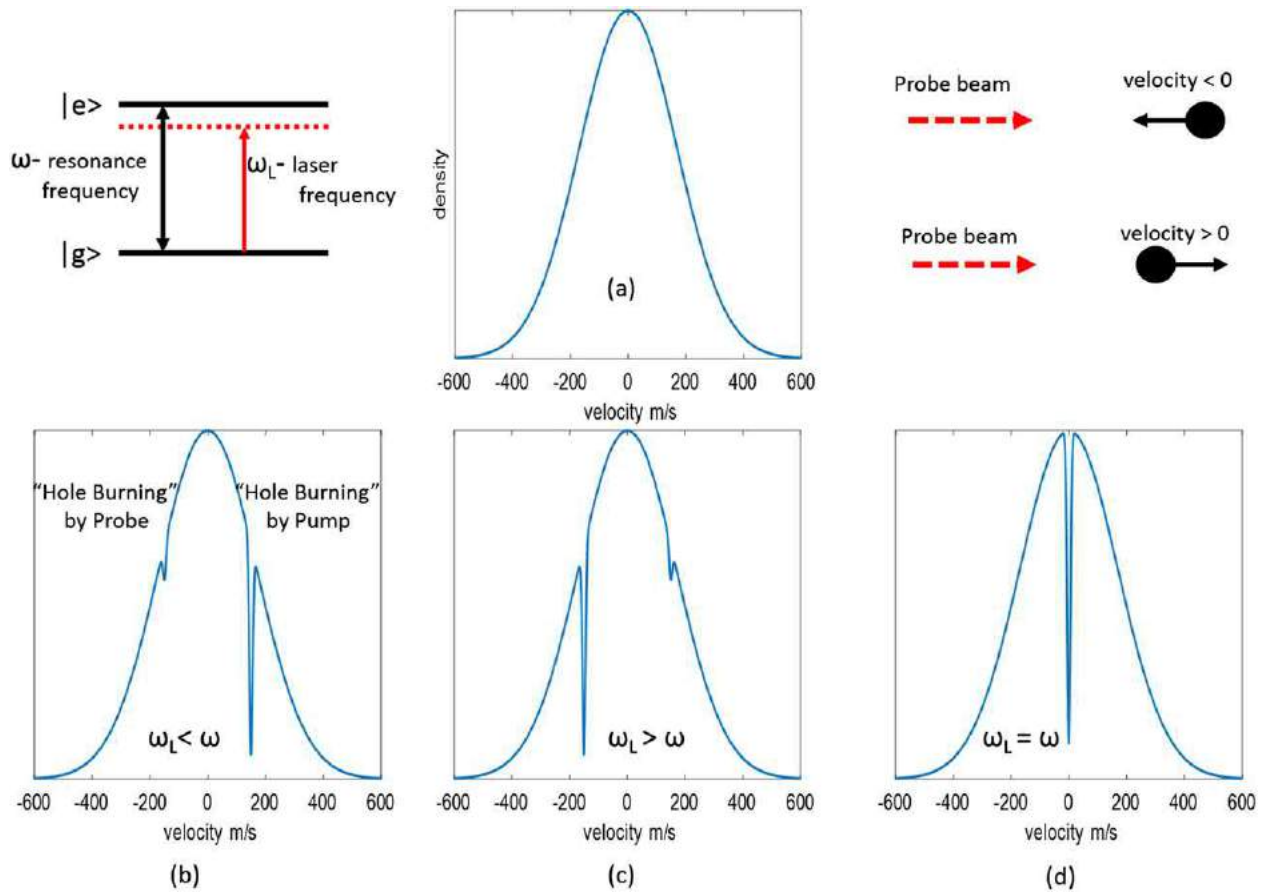


Fig. 6 (a) Two level system and the number density function of ground state Rb atoms as a function of their velocity. (b) “Hole burning” when the laser frequency is less than the resonance frequency. (c) Laser frequency is higher than the resonance frequency. (d) Laser frequency is equal to the resonance frequency.

resonance frequency, then both pump and probe beams interact with the zero-velocity group of atoms (**Fig. 6(d)**). In this case the strong pump beam depletes the ground state population and will even saturate the absorption if intense enough. Consequently, the probe beam experiences modified (decreased) absorption. The hyperfine lines shown in (a), (c), and (f) on **Fig. 5** represent decreased absorption of the probe beam in the presence of the strong pump beam.

Crossover Resonances

Fig. 5 also shows additional peaks (b), (d) and (e) that do not directly correspond to any atomic resonances. These peaks are referred to as “crossover” resonances. They emerge when a system has multiple excited states (e.g., Rb atoms – see **Fig. 2**) that are within the Doppler width and share the same ground state. For simplicity I will consider a 3 level “V” system that has one ground state ($|g\rangle$) and two excited states ($|e1\rangle$ and $|e2\rangle$) such as that shown in **Fig. 7**. When the laser frequency is in between these two resonances, then there are two different velocity groups of atoms that the pump beam interacts with: one that is redshifted (dip 2 which corresponds to $|g\rangle \rightarrow |e1\rangle$ transition) and one that is blue shifted (dip 4 that corresponds to $|g\rangle \rightarrow |e2\rangle$ transition). At the same time the probe beam interacts with exactly the opposite sign velocity groups of atoms. In **Fig. 7** they correspond to dip 1 ($|g\rangle \rightarrow |e2\rangle$ transition) and dip 3 ($|g\rangle \rightarrow |e1\rangle$ transition). When the laser frequency is exactly halfway between these two resonances, the absorption dips from pump and probe beams overlap (1 with 2 and 3 with 4). In the presence of the strong pump field this causes reduction of the probe beam absorption. The peaks (b), (d), and (e) in **Fig. 5** correspond to all possible crossover resonances (please see the caption for details).

Experimental Setup

Schematic diagram of an experimental setup that is typically used for laser locking is shown in **Fig. 8**.

A beam from an ECDL is divided using a polarizing beam splitter. Most of the laser power is sent to an experiment (e.g., cooling and trapping) whereas only a small portion (less than 1 mW) is used for generating a correction signal for laser

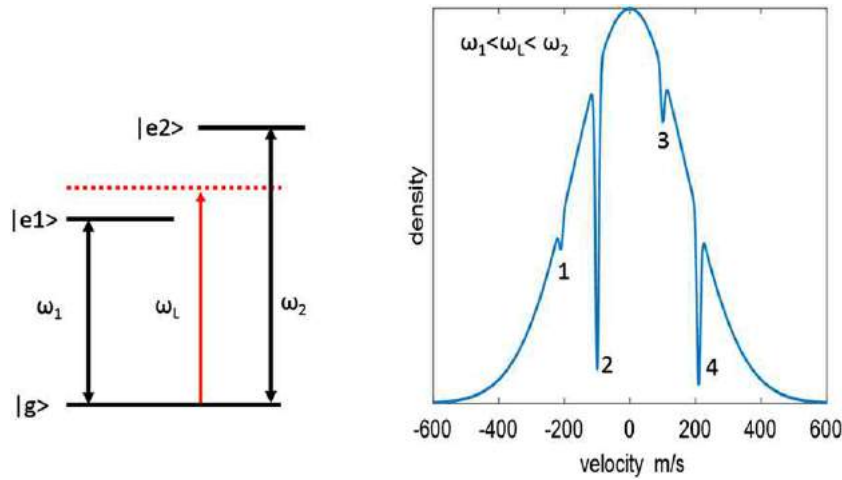


Fig. 7 Three level “V” system. “Hole burning” when the laser frequency is in between two resonance frequencies.

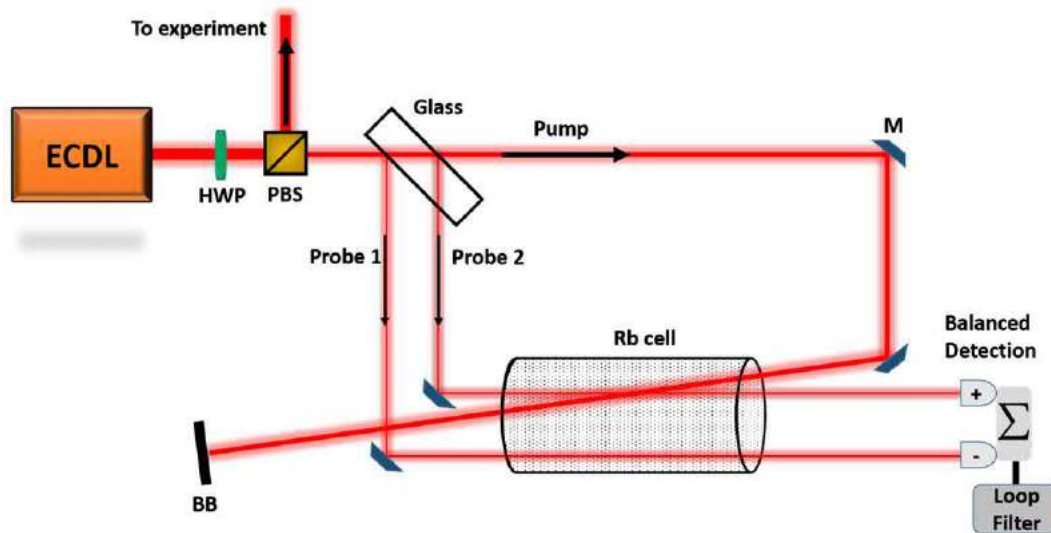


Fig. 8 Experimental setup for saturated absorption spectroscopy. BB, beam block; ECDL, external cavity diode laser; HWP, half wave plate; M, mirror; PBS, polarizing beam splitter.

locking. Using a thick glass this beam is split again into 3 different beams: 2 probe (weak) beams reflected from the front and rear surfaces of the glass and one (strong) transmitted pump beam. The pump beam is overlapped with one of the probe beams in the reference cell and both probe beams are monitored using a balanced photodetector. The signal from the balanced detector is sent to a loop filter that generates a correction signal for the laser.

Balanced detection has the following advantages:

1. As shown in Fig. 6(a) the measured transmission profile of the probe beam overlapped by the pump beam has the Doppler broadened background. In Fig. 8 that corresponds to the transmission of the probe 2 beam. The transmission profile of the probe 1 beam is shown in Fig. 3 and is pure Doppler broadened. Balanced detection takes the difference between these spectrums and the resulting spectrum is Doppler background free (Fig. 9).
2. Balanced detection helps to eliminate any intensity fluctuations that the ECDL might have. This is especially important if, for example, the laser must be locked on the side of one of the hyperfine lines. Without the balanced detection, any unwanted intensity fluctuation would cause the transmission spectrum and hence the correction signal to fluctuate as well. This would result in broadening of the laser frequency.

If the laser frequency must be locked to the peak of an absorption line then current or phase modulation techniques can be used. Using these methods one obtains a spectrum that is the derivative of the saturated absorption spectrum in which zero crossings correspond to the peaks of the absorption lines.

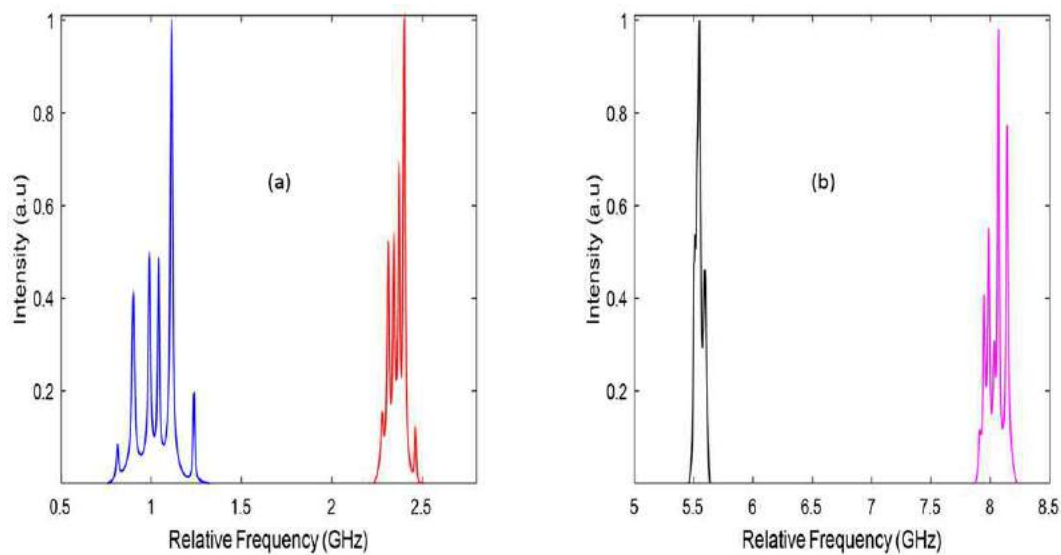


Fig. 9 Doppler background free saturated absorption spectra of Rb atoms. (a) 0–2.7 GHz range. (b) 5–8.5 GHz range.

Conclusions

In this article the use of saturated absorption spectroscopy for laser locking was reviewed. I have discussed the basic principles of saturation absorption spectroscopy and reviewed how room temperature reference cells can provide Doppler free absorption lines. I have explained the appearance of crossover resonances in multilevel systems and reviewed an experimental technique that is commonly used for generating a correction signal for laser locking.

Acknowledgment

I would like to thank Brad Smith for obtaining the saturated absorption spectra.

Further Reading

- Arnold, A.S., Wilson, J.S., Boshier, M.G., 1998. A simple extended-cavity diode laser. *Review of Scientific Instruments* 69 (3). doi:10.1063/1.1148756.
- Black, E.D., 2001. An introduction to Pound–Drever–Hall laser frequency stabilization. *American Journal of Physics* 69 (1). doi:10.1119/1.1286663.
- MacAdam, K.B., Steinbach, A., Wieman, C., 1992. A narrow-band tunable diode laser system with grating feedback and a saturated absorption spectrometer for Cs and Rb. *American Journal of Physics* 60 (12). doi:10.1119/1.16955.

Attosecond Spectroscopy

Agnieszka Jaroń-Becker and Andreas Becker, University of Colorado, Boulder, CO, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The quest for studying dynamics in matter on ultrashort timescales has driven the development of a variety of technologies in ultrafast science. The most prominent among these are ultrashort electron pulses, femtosecond laser pulses, X-ray free electron lasers and, currently the shortest off all these probes, attosecond light pulses. The atomic unit of time is $24.2 \text{ as} = 24.2 \times 10^{-18} \text{ s}$, corresponding to the period of an electron circling in the first Bohr orbit divided by 2π . Attosecond pulse technology therefore allows for resolution of dynamical quantum processes on the timescales of electron motion in atoms and other forms of matter, as compared to the observation of atomic motion using femtosecond laser pulses.

The goal of this brief review is to first present the concepts and physical understanding behind the generation of attosecond optical pulses via the currently most common source, the highly nonlinear optical high harmonic generation process driven by strong laser pulses. In the second part of the review a number of spectroscopic techniques and applications for the temporal resolution on the attosecond timescale are discussed. For a more detailed presentation of the topics we refer to a number of recent books and review articles (Kapteyn *et al.*, 2005; Corkum and Krausz, 2007; Krausz and Ivanov, 2009; Chang, 2011; Popmintchev *et al.*, 2011; Gallmann *et al.*, 2012; Dahlström *et al.*, 2012; Vrakking, 2014; Leone *et al.*, 2014; Peng *et al.*, 2015; Pazourek *et al.*, 2015; Ramasesha *et al.*, 2016; Calegari *et al.*, 2006; Wu *et al.*, 2016; Leone and Neumark, 2016) as well as the original work cited therein.

In this review we use, if not stated differently, Hartree atomic units, which is an unit set particularly useful in atomic, molecular and optical physics. In this unit set the values for the electron mass m_e , the elementary charge e , Planck's constant $\hbar$; and the Coulomb's constant $k_e = 1/4\pi\epsilon_0$ are unity, i.e., $m_e = e = \hbar = k_e = 1$.

Generation of Attosecond Optical Pulses

The progress of ultrashort laser pulse technology since the invention of mode-locking is summarized in Fig. 1. After a continuous decrease of the pulse duration until the mid 1980s there was only a marginal progress for more than a decade. In the late 1980s, femtosecond laser pulses with a duration of a few cycles and high peak intensities became available due to techniques for pulse compression and chirped pulse amplification. Such pulses approached a single optical cycle with an optical period around 2.7 fs at a central wavelength of 800 nm.

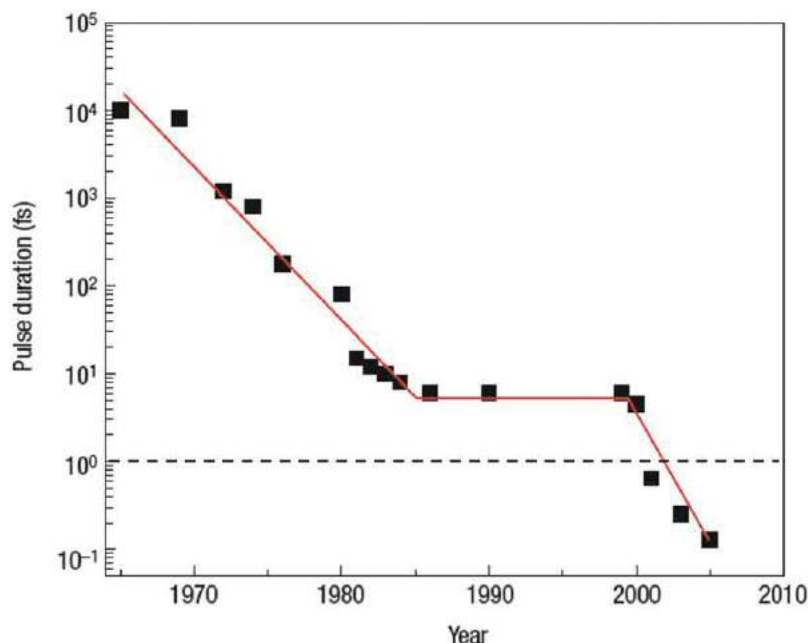


Fig. 1 Development of the laser pulse duration (in femtoseconds): The minimum duration first continuously decreased until the mid of the 1980s. Following marginal progress for more than a decade, the femtosecond barrier (dashed line) was broken in 2001. Adapted with permission from Mcmillan Publishers Ltd., Corkum, P.B., Krausz, F., 2007. Attosecond science. *Nat. Phys.* 3, 381, copyright (2007).

The basic idea towards engineering sub-femtosecond pulses, proposed in the early 1990s and first realized in 2001, is similar to the principle of mode-locking used for femtosecond lasers. If a range of phase-locked frequencies is combined via Fourier synthesis, the temporal profile is a sequence of pulses with a duration inverse proportional to the number of frequencies involved in the synthesis. Consequently, the generation of attosecond optical pulses relies on the availability of a comb of coherent light at equidistant frequencies in the ultraviolet, extreme ultraviolet or even the soft X-ray spectral regime. Before reviewing the physics underlying high-order harmonic generation as the most common source of attosecond pulse generation we briefly recap basic concepts and mathematical expressions used for the description of ultrashort laser pulses via their electric fields.

Electric Laser Field in Time and Frequency Domain

High-order harmonic and attosecond pulse generation is primarily focused on linearly polarized radiation. Since the wavelength of the involved radiation is, in general, much larger than the size of the target, typically an atom, the electric dipole approximation can be used and a linearly polarized pulse can be expressed by its time-dependent electric field:

$$\mathbf{E}(t) = \varepsilon_0(t) \cos(\omega_0 t + \phi_{CE}) \hat{\mathbf{e}} \quad (1)$$

where $\varepsilon_0(t)$ is the time-dependent envelope function, ω_0 the central angular frequency, related to the central wavelength $\lambda_0 = 2\pi c/\omega_0$, $(\phi)_{CE}$ is the carrier-(to-)envelope phase of the electric field and $\hat{\mathbf{e}}$ represents the polarization direction of the field.

For Gaussian pulses the envelope function is given by

$$\varepsilon_0(t) = \varepsilon_0 \exp \left[-2 \ln 2 \left(\frac{t^2}{\Delta t^2} \right) \right] \quad (2)$$

where ε_0 is the peak amplitude. The intensity profile of the pulse is then

$$I(t) = I_0 \exp \left[-4 \ln 2 \left(\frac{t^2}{\Delta t^2} \right) \right] \quad (3)$$

with I_0 is the peak intensity and Δt is the full-width-half-maximum (FWHM) pulse duration. Via Fourier transform the electric field of a laser pulse can be also described in the frequency domain. For a Gaussian pulse in the time domain the spectral amplitude is also a Gaussian function:

$$\tilde{\varepsilon}(\omega) = \tilde{\varepsilon}_0 \exp \left[-2 \ln 2 \left(\frac{(\omega - \omega_0)^2}{\Delta \omega^2} \right) \right] \quad (4)$$

with full-width half-maximum bandwidth $\Delta \omega$.

High-Order Harmonic Generation

The currently most commonly used method for the generation of coherent light in form of attosecond pulses is the physical process of high-order harmonic generation. Harmonic generation is a nonlinear frequency up conversion effect occurring in the interaction of laser fields with matter. Observation of the second harmonic of laser light in 1961, immediately after the invention of the laser, marked the emergence of the field of perturbative nonlinear optics. In the perturbative interaction regime, the intensities of generated higher harmonics decrease rapidly due to the weak perturbation of matter by the laser field.

The generation of high-order harmonics was discovered in 1987 via the observation of a radiation spectrum characterized by a large number of odd harmonics of the fundamental driving laser frequency with intensities that did not decrease significantly, as it was expected in perturbative nonlinear optics. In the respective experiments, atoms interacted with femtosecond laser pulses having peak intensities of the order of $10^{13} - 10^{14} \text{ W cm}^{-2}$. At these intensities the electric field strength of the laser pulse becomes comparable to those of the fields due to the Coulomb interaction between charged particles in atoms and molecules. Consequently, the electric field cannot be considered as a perturbation anymore. In such intense laser pulses, matter is ionized even if the photon energy of the field is only a small fraction of the ionization potential. This strong-field regime of laser-matter interaction is restricted by an upper limit at intensities leading to the onset of relativistic effects and those due to the influence of the magnetic field of the laser.

Microscopic description

The power spectrum of the radiation emitted by an accelerating charge is proportional to the magnitude of the acceleration. Assuming that the electric field of a strong laser pulse interacts with a single electron in an atom, a time-dependent dipole is created by the electron wavepacket and the parent ion given by

$$\mathbf{d}(t) = \langle \Psi(\mathbf{r}; t) | \mathbf{r} | \Psi(\mathbf{r}; t) \rangle \quad (5)$$

where $\Psi(\mathbf{r}; t)$ represents the electron wave function, resulting from the solution of the nonrelativistic Schrödinger equation describing the interaction between the atom (in the single-active-electron approximation) with the electric field of an intense laser

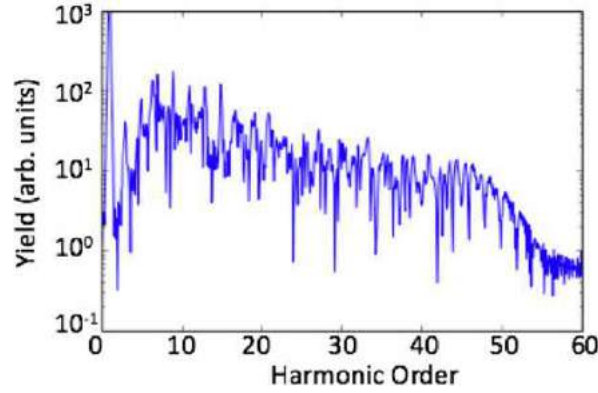


Fig. 2 Exemplary result of numerical simulations of the single-atom high harmonic spectrum for atomic hydrogen interacting with a laser pulse with central wavelength $\lambda_0=800$ nm and peak intensity $I_0=3 \times 10^{14}$ W cm $^{-2}$. From Miller, M.R., 2016. PhD thesis, JILA, University of Colorado, Boulder.

pulse, and $\mathbf{r}$ is the space vector. The dipole acceleration can be either determined via the second time derivative of the dipole

$$\mathbf{a}(t) = \frac{d^2}{dt^2} \langle \Psi(\mathbf{r}; t) | \mathbf{r} | \Psi(\mathbf{r}; t) \rangle \quad (6)$$

or using the Ehrenfest theorem as

$$\mathbf{a}(t) = \langle \Psi(\mathbf{r}; t) | -\nabla V(\mathbf{r}) + \mathbf{E}(t) | \Psi(\mathbf{r}; t) \rangle \quad (7)$$

where $V(\mathbf{r})$ is the Coulomb potential between the propagating electron and the parent ion. The high-order harmonic spectrum is then obtained as the Fourier transform of the dipole acceleration.

An example of a high-order harmonic spectrum is shown in **Fig. 2** for a hydrogen atom driven by a linearly polarized electric field at a central wavelength of 800 nm and a peak intensity of 3×10^{14} W cm $^{-2}$. The structure of the radiation spectrum possesses the following characteristic features of high-order harmonic generation: strong emission at the photon energy of the driving field, followed by a quick drop off due the exponential decay for perturbative low-order harmonics, subsequent plateau of harmonic generation at constant strength and rapid decrease of radiation efficiency beyond a cut-off energy.

Three-step model of HHG

An intuitive physical picture of HHG emerges from an analysis known as the three-step model, depicted graphically in **Fig. 3(a)**. The strong electric field suppresses the Coulomb barrier binding the electron to the atomic core, which permits the tunneling of an electron wave packet. The wave packet then propagates in the oscillating electric field and, depending on the time of release, it can be steered back to the parent ion. With some probability, the wave packet can recombine with the parent ion to the ground state. The energy absorbed by the electron during the propagation is emitted in form of a high energy photon. In a quantum mechanical model description of HHG this picture is reflected via a product of three probability amplitudes, accounting for ionization, propagation and recombination of the electron wave packet.

Due to the quasiperiodic repetition of this process every half cycle in a multicycle laser field, the resulting dipole emission spectrum is discrete, consisting of odd harmonics of the central driver laser frequency. In an ultrashort few-cycle driver pulse, the discrete structure becomes less and less pronounced and eventually disappears in particular in the cut-off region of the spectrum. Predictions of this quantum mechanical model are in qualitative agreement, in particular concerning the form and extension of the plateau, with experimental observation and the results of ab-initio numerical calculations, based on **Eq. (6)**.

Some basic features of the HHG process and spectrum can be understood from a simplified classical analysis, based on the Newton's equation in one dimension, for propagation of a point-like electron in an oscillating linearly polarized electric field:

$$\frac{d^2x}{dt^2} = -\varepsilon(t)x \quad (8)$$

with initial conditions $\dot{x} = 0$ and $x=0$. Predictions of the corresponding classical trajectory analysis, emerging during a center cycle of the field used to generate the HHG spectrum shown in **Fig. 2**, are shown in **Fig. 3(b)**. Each color corresponds to a different ionization time and represents trajectories that return to the parent ion and, hence, contribute to HHG; while gray lines indicate ionization events that do not lead to return of the electron. The classical analysis predicts a cutoff in the HHG spectrum at a maximum energy of

$$E_{\max} = I_p + 3.17U_p \approx I\lambda_0^2 \quad (9)$$

where I_p is the ionization potential of the atom, $U_p = I_0/4\omega_0^2$ is the ponderomotive potential. This prediction is in good agreement with experimental observations of HHG spectra in multicycle driver laser pulses, whereas for few-cycle pulses the cutoff can be shifted to slightly higher HHG harmonics. The excursion distance for the classical trajectory corresponding to the cut-off harmonic

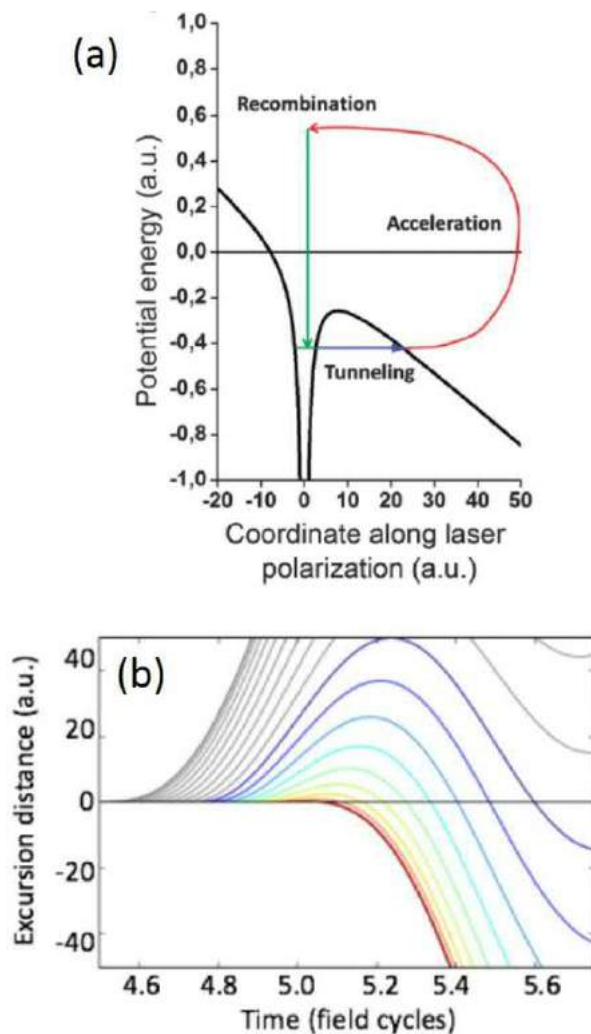


Fig. 3 (a) Conceptual picture of the three-step model of high harmonic generation (Reproduced from Vrakking, M.J.J., 2014. Attosecond imaging. *Phys. Chem. Chem. Phys.* 16, 2775 with permission of the PCCP Owner Societies.); (b) Classical trajectories of a point-like electron propagating in a strong laser field. Laser parameters are the same as for the HHG spectrum in Fig. 2. Reproduced from Miller, M.R., 2016. PhD thesis, JILA, University of Colorado, Boulder.

energy is given by the quiver radius $\alpha_0 = U_p/\omega_0^2$. Propagation of the electron via two different trajectories, following excursion distances either shorter or longer than the quiver radius, contribute to the harmonic generation process for each of the other photon energies.

The scaling of the harmonic cut-off energy with the intensity and the square of the driver wavelength (Eq. (9)) indicates that in principle the spectral bandwidth of the harmonic emission, which is important for the attosecond pulse generation (see Section Formation of Attosecond Pulses), can be extended by increasing the intensity or the wavelength of the driver laser. However, in practice HHG in a specific charge state of the single atom gets terminated by the depletion of the electron population in the ground state, which limits the applied laser intensity. Furthermore, the efficiency of the single-atom harmonic emission decreases rapidly with increase of the wavelength due to the diffusion of the electron wave packet along the longer and longer excursion distances in the continuum.

Macroscopic phase-matching aspects

In order to achieve efficient high harmonic emission and, hence, attosecond pulse generation the high harmonic signal from the many atoms in the generating medium must be phase matched. This means that the phase velocity of the driving laser field and the harmonic light fields must be matched in order to guarantee coherent addition of the harmonic emission from the individual atoms. Understanding of phase matching in the non-perturbative interaction regime is different to the concepts known from perturbative nonlinear optics. Besides factors arising due to the geometry of the HHG experiment and intrinsic atomic phases, the phase mismatch depends on the relation of neutral and ionic species in the focal region and also the pressure of the gaseous

medium. Accurate analysis of these effects recently enabled the extension of high harmonic generation up to the keV X-ray regime with ultraviolet and mid-infrared driving lasers.

Controlling polarization of high-order harmonics

High harmonic generation driven by an strong electric field is very sensitive to the ellipticity of the pump pulse, and the intensity of the harmonic signal decreases rapidly as the polarization moves away from linear. This can be easily understood within the classical picture, since in an elliptically polarized field the electron wave packet acquires a transverse component of the velocity, prohibiting its return to the core of the parent ion. Consequently, for research in HHG is mainly focused on the generation of linearly polarized high-order harmonics. The polarization of the harmonics can be controlled using two laser pulses. Setups using either collinear bichromatic beams or noncollinear beams in both cases with counter-rotating polarization have been demonstrated.

Formation of Attosecond Pulses

The temporal structure of high harmonic emission yields the basic for the generation of attosecond pulses, either in the form of a train of pulses or as isolated pulses. The generation of a train does not necessitate any particular requirement for harmonic driver pulse. For spectroscopic applications it can be however desirable to isolate a single pulse from the train by controlling the underlying harmonic emission process.

Attosecond pulse trains

A phase-locked periodic spectrum of equidistant frequencies results in a periodic intensity profile in the time domain, i.e. a comb or train of pulses. With a multi-cycle laser pulse a high harmonic spectrum with pronounced peaks at odd multiples of the fundamental frequency is generated. In the plateau regime of the spectrum the emission peaks are of similar strength and have a smooth relative phase relation. Consequently, the Fourier synthesis of this part of the spectrum leads to a train of pulses with ultrashort duration, which depending on the spectral bandwidth can be shorter than a femtosecond.

Alternatively, the generation of a train of attosecond pulses via HHG can be understood in the time domain directly: In a multi-cycle laser pulse the emission and recombination of the electron wave packet is repeated periodically in every half cycle of the electric field. Hence, a burst of harmonic radiation is emitted over a fraction of the femtosecond driving field cycle, forming each time an attosecond laser pulse. These pulses are separated by the half period of the driving field, and a train of attosecond pulses is generated. Since the harmonic generation process strongly depends on the electric field at the instant of emission of the electron wave packet and during its propagation, in a driving pulse the spectral content and the temporal structure of each attosecond pulse is however different. Furthermore, the electron wave packet returns from alternate side of the atom during subsequent half cycles resulting in attosecond pulses with alternating up- and down-chirp in frequency.

Isolated attosecond pulses

While attosecond pulse trains are used for spectroscopic purposes, a greater flexibility is gained through the generation of isolated attosecond pulses. The extraction or isolation of a single attosecond pulse from a train is too fast for usual electronic devices. However, understanding of the high-order harmonic process gives rise to techniques to select harmonic emission from one recombination event. An overview of the most common techniques is discussed below.

Spectral selection

According to the physical picture of high harmonic generation the maximum harmonic energies close to the cutoff of the spectrum are emitted during the most strongest electric field cycle near the peak of the driving pulse. The generation of an isolated attosecond pulse can be therefore realized by selecting photon energies that are emitted through just one recombination event. This can be achieved by filtering spectral components at lower-energy harmonics, that arise from multiple recombination events produced during other cycles than the central field cycle.

Such a spectral selection scheme has been implemented using a few-cycle laser pulse, in which the electric field strength and hence the extension of the generated harmonic spectrum varies greatly from cycle to cycle. Due to the sensitivity of the high harmonic process to the actual field strength, a control of the carrier-envelope phase is essential as well. The spectral selection method has been also demonstrated in laser pulses, in which the atom is fully ionized during the rising part of the pulse and, hence, the harmonic process is terminated at a certain field cycle.

Temporal gating

Another set of methods for isolated attosecond pulse generation relies on the restriction of the HHG process to just one single half cycle of the electric field. Currently, the most effective technique (polarization gating, Fig. 4) makes use of the strong sensitivity of the harmonic process on the ellipticity of the pulse. Since the efficiency of the harmonic emission process decreases quickly by varying the polarization from linear, an isolated attosecond pulse can be produced with a driver laser pulse that is linearly polarized during a single half cycle of the field, while it is elliptically polarized during the remainder of the pulse.

Such a driving pulse with variation of its polarization can be produced by superimposing two counter-rotating circularly polarized laser pulse. By time delaying two such few-cycle laser pulses, the resulting pulse is linearly polarized over the central half

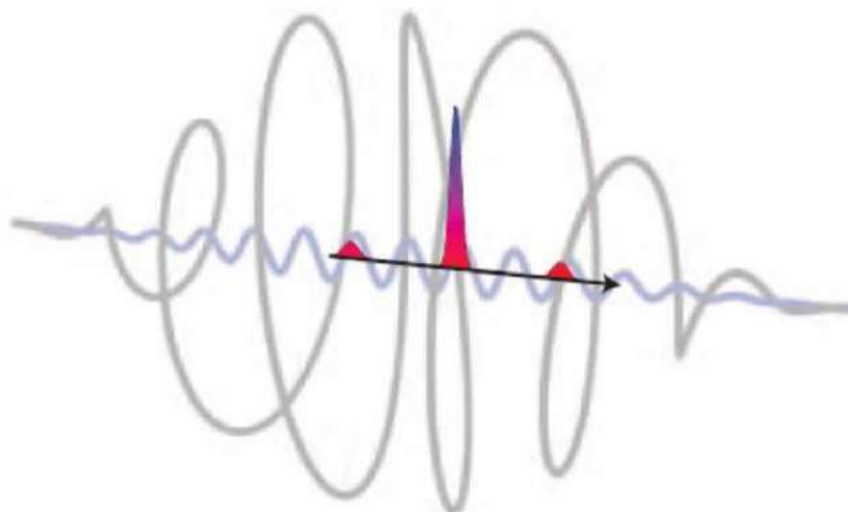


Fig. 4 Principle of polarization gating. A femtosecond driver laser pulse with time-dependent polarization is used. The pulse is linearly polarized during a single half-cycle of the pulse, leading to the restriction of efficient attosecond pulse generation to an isolated event. Reprinted with permission from Mcmillan Publishers Ltd., Chini, M., *et al.*, 2014. *Nat. Photonics* 8, 178, copyright (2014).

field cycle only, creating the desired temporal gate for efficient harmonic emission and isolated attosecond pulse generation. This concept was further extended by combining a driver pulse with time-dependent ellipticity and a linearly polarized second harmonic pulse. In this technique, known as double optical gating, an asymmetric electric field with time-varying polarization is created. As a consequence, the high harmonic and attosecond pulse generation occurs only once per optical cycle, instead of once per half cycle, providing more flexibility for the application in the polarization gating technique.

Attosecond light house

A further route to the generation of isolated attosecond pulses is provided by a technique, dubbed as the attosecond light house effect (**Fig. 5**). Usually, in a HHG process the attosecond pulses are emitted in a spatially confined beam, having a divergence much lower than the driving laser beam. By introducing a time-dependent rotation of the wavefront, the propagation direction of each harmonic emission event varies depending on the wavefront direction at the moment of the HHG process. While a train of attosecond pulses is generated in this technique, the pulses in the train are angularly separated. If the rotation of the wavefront is larger than the divergence of the attosecond pulses, an isolated attosecond pulse can be spatially selected in the far field. The method has been demonstrated using plasma mirrors as well as gas targets.

Phase matching gating

Each of the previous methods achieves isolation of a single attosecond pulse via control of the HHG process on the microscopic (single-atom) level. Harmonic emission and attosecond pulse production however also depend on the phase matching of the radiation generated from the many atoms in the medium. An essential part of the phase-matching conditions is the degree of ionization of the medium, which increases during the interaction with a strong laser pulse. Consequently, on the leading edge of the pulse the laser phase velocity in the medium may be lower than the speed of light, c , due to the dominance of neutral atoms in the medium. In contrast, due to the higher fraction of ions in the medium in the trailing edge of the pulse, the laser phase velocity may be larger than c . As a result, the phase matching conditions are satisfied only during a temporal window within the laser pulse (**Fig. 6**) and efficient harmonic emission and attosecond pulse generation is restricted to the recombination events falling in this phase matching window. For midinfrared driver lasers the phase matching window is found to be much shorter than for Ti: sapphire lasers. This is, in particular, due to the fact that at longer wavelengths optimal phase conditions are reached at higher gas pressures and ionization densities. As practical consequence, the temporal phase matching window can be narrowed to a single half cycle leading to robust isolated attosecond pulse generation.

Spectroscopy Techniques and Applications

Measurements on the attosecond timescale are still in its infancy. One of the open challenges is the realization of a conventional pump-probe spectroscopy set-up using two isolated attosecond laser pulses, in which time resolution is achieved by the delay and durations of the two pulses. So far, the intensity of attosecond laser pulses has however been too low to achieve significant attosecond pump-probe signals. But, using the knowledge how to manipulate and control electrons with laser pulses, a number of techniques have been developed in which attosecond measurements are performed either without the applications of attosecond

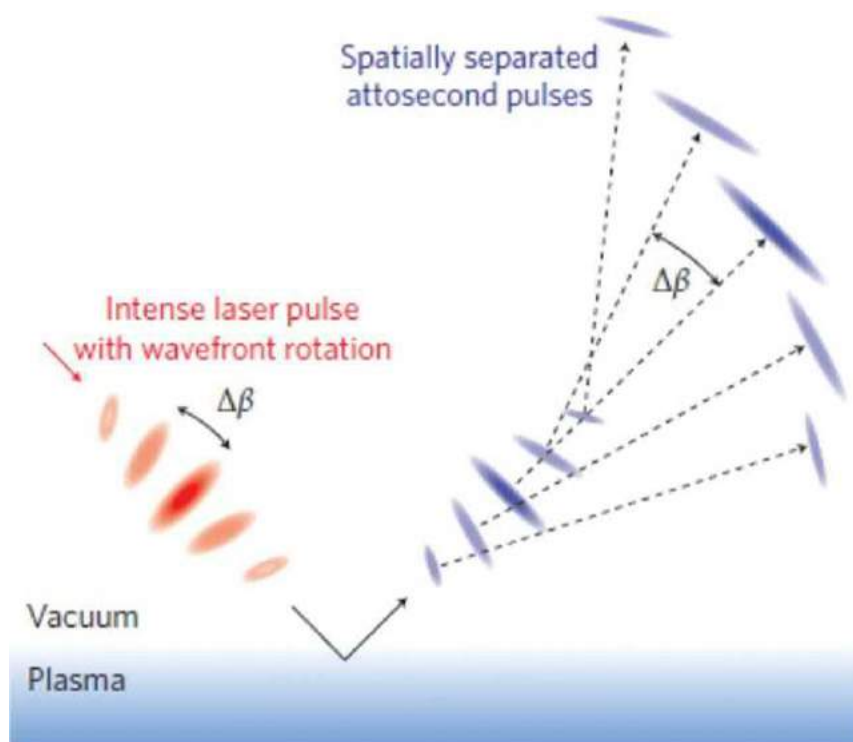


Fig. 5 Attosecond light house principle. Using a time-dependent rotation of the laser wavefront, the attosecond pulses in the train are emitted spatially in different directions and an isolated pulse can be selected. Reproduced with permission from Mcmillan Publishers Ltd., Wheeler, J.A., *et al.*, 2012. Nat. Photonics 6, 829, copyright (2012).

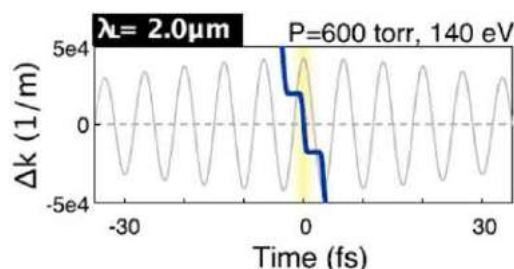


Fig. 6 Calculated phase mismatch Δk for a many-cycle long mid-infrared driver pulse. The phase matching window ($\Delta k \approx 0$, highlighted in yellow) is limited to one half-cycle of the driver pulse, restricting efficient build-up of harmonic emission and attosecond pulse generation to an isolated event. From Chen, M.C., *et al.*, 2014. Proc. Natl. Acad. Sci. USA 111, E2361.

pulses at all or by combining attosecond and femtosecond laser pulses. Applications of attosecond measurements using these spectroscopic techniques are made to reveal electron dynamics in atoms, molecules and solid state systems such as metals and semiconductors.

Attosecond Spectroscopy With Femtosecond Pulses

Recollision spectroscopy

The basic physical process of ionization and return of the electron wave packet in strong laser fields itself (see Section Three-step model of HHG) enables observations with attosecond time resolution. In the spirit of pump-probe spectroscopy, the emission of the electron wave packet may also initiate (pump) a process in the parent ion, e.g., the evolution in excited states or vibration/dissociation in case of a molecular system, while the return of the electron acts as a probe. In this scenario the time between emission and return of the wave packet can be considered as the pump-probe delay. Within a field cycle, the return of the electron wave packet with different energies occur at different time delays (**Fig. 7**), enabling simultaneous snapshots of the system within one physical process. The variation of the time delays is in the attosecond range for laser wavelengths in the near-infrared and mid-infrared regime. Further control of the delay, within certain limits, can be achieved via tuning of the wavelength and hence the duration of the field cycle of the driving laser field.

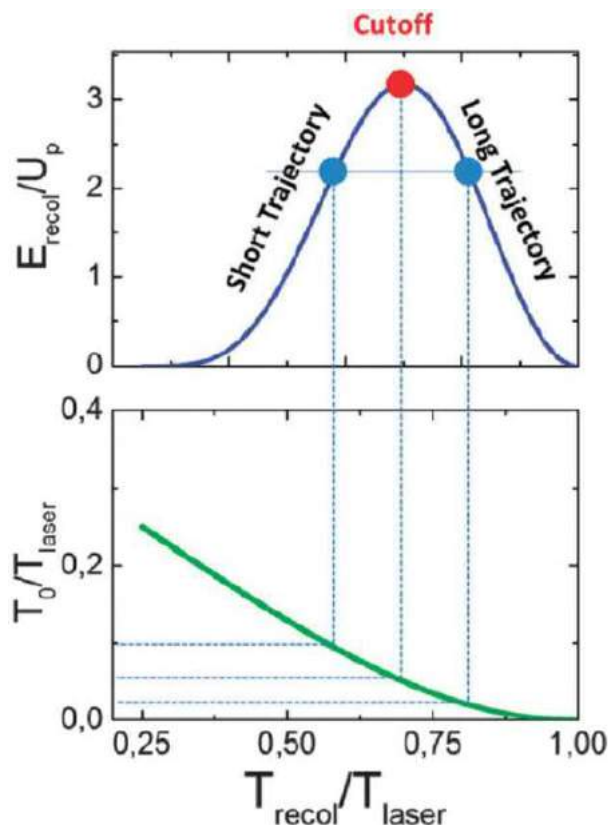


Fig. 7 Relation between the electron return energy and the return time (upper panel) and the times of ionization and return of the electron (lower panel), according to the classical rescattering model (Eq. (8)). Reproduced from Vrakking, M.J.J., 2014. Attosecond imaging. *Phys. Chem. Chem. Phys.* 16, 2775 with permission of the PCCP Owner Societies.

On its return the electron wave packet can recombine with, diffract on or further ionize the parent ion. The resulting harmonic or photoelectron spectra contain information about changes in the structure of the ion as well as the electronic distribution inside the ion with an attosecond time resolution. For the analysis and interpretation of the spectra it is often assumed that the amplitude for the full process can be factorized in one for the propagation of the electron wave packet and one for its (field-free) scattering or recombination.

Attoclock: Angular streaking

Another way to realize attosecond time resolution in the interaction of a femtosecond laser pulse with matter relies on the application of an ultrashort two-cycle nearly circularly polarized laser pulse. Here, the resolution is achieved because of the rotation of the electric field vector in the polarization plane, which serves as the hand of a clock. If a target atom or molecule gets ionized the final emission angle of the electron wave packet is related to the instant of ionization. Thus, the observed photoelectron angular momentum distribution contains time information about the ionization process. Peak positions in such distributions can be measured with an angular resolution of about 1° , which corresponds to a temporal resolution of about 7.4 as at a center wavelength of 800 nm. Further calibration of this attoclock concept also relies on the accurate analysis of the propagation of the electron wave packet in the combined fields of the circularly polarized field and the Coulomb field between the electron and the parent ion.

Such angular-resolved photoelectron measurements with ultrashort nearly circularly polarized femtosecond laser pulses have been applied to assess the time for tunneling of an electron wave packet through the strong field suppressed Coulomb barrier, the time delay in the successive steps of electron emission in sequential strong-field double ionization and the emission time of an electron wave packet from a molecule. Besides the challenges in the accurate analysis of the results, it remains technically difficult to generate circularly polarized pulses in the few-cycle regime. A residual amount of ellipticity in the polarization results in small oscillations of the electric field magnitude on the sub-cycle timescale and may distort the resulting momentum distributions.

Photoelectron and Ion Spectroscopy With Attosecond Pulses

Although pump-probe spectroscopy with attosecond pulses in both steps could not be realized up to date, in experiments employing different two-color schemes with a femtosecond (near-infrared) pulse and isolated attosecond (XUV) pulse or attosecond pulse trains high time resolution in measurements were achieved. As a common principle the attosecond time resolution is

achieved by controlling the application of the attosecond pulse over the optical field cycle of the near-infrared pulse. Since the near-infrared pulse is usually a replica of the pulse used to generate the attosecond pulses, perfect synchronization and control of the time delay between the pulses can be assured. In general, the use even of a moderately strong femtosecond pulse however complicates the analysis and interpretation of the results, as the impact of the near-infrared field on the observed dynamics needs to be taken into account.

Attosecond streak camera

One set of attosecond measurements makes use of the streak camera principle by applying a linearly polarized isolated attosecond pulse along with a moderately strong femtosecond near-infrared laser pulse in interaction with an atom. A photoelectron, which is instantaneously liberated into the continuum by the attosecond pulse, experiences a momentum shift due to the presence of the near-infrared laser pulse (Fig. 8). The magnitude of the shift depends on the vector potential at the time of ionization and the subsequent propagation of the electron wave packet in the combined fields of the Coulombic interaction between the electron and the ion as well as with the near-infrared streaking pulse. By recording the momentum (or energy) shift as a function of the time delay between the two pulses a streaking trace is obtained.

The attosecond streak camera concept was initially used to characterize isolated attosecond pulses, but soon after it was applied to obtain dynamic information about processes in atoms and solids, in which usually the streaking traces for two events at different photoelectron energies are compared with each other. Proof-of-principle experiment of this kind was a measurement of the lifetime of an inner-shell vacancy in an atom, which was produced by excitement of a core electron by the attosecond XUV pulse. Sampling of the emission of the Auger electron, following the filling of the hole from higher energy level, by the femtosecond near-infrared field provides access to time resolution of this inner-shell process. Later, observation of a relative shift between streaking traces for photoelectrons emitted from different levels in atoms and solids was interpreted as due to the difference in the time delays acquired during the emission and propagation of the respective electron wave packets.

Reconstruction of attosecond beating by interference of two-photon transitions

Using a train of attosecond pulses, instead of an isolated attosecond pulse, in conjunction with a long weak near-infrared pulse leads to a technique referred to as reconstruction of attosecond beating by interference of two-photon transitions (RABBITT). Assuming that the spectral bandwidth of the attosecond pulses in the train spans a number of harmonics, ionization of the target in the combined fields leads to a series of main bands (corresponding to the photoelectron energies upon absorption of one of harmonic photons) and side bands (corresponding to the additional absorption or emission of a near-infrared photon) in the photoelectron energy spectrum (Fig. 9). Since there are two pathways leading to the same side band, there appears an interference pattern in the spectrum as a function of the time the attosecond pulse is applied over the optical half cycle of the near-infrared field.

The time resolution in the RABBITT method is related to the modulation of the probability of the side band peaks, that depends on the phase difference between consecutive harmonics (or the time of application of the attosecond pulses) and an

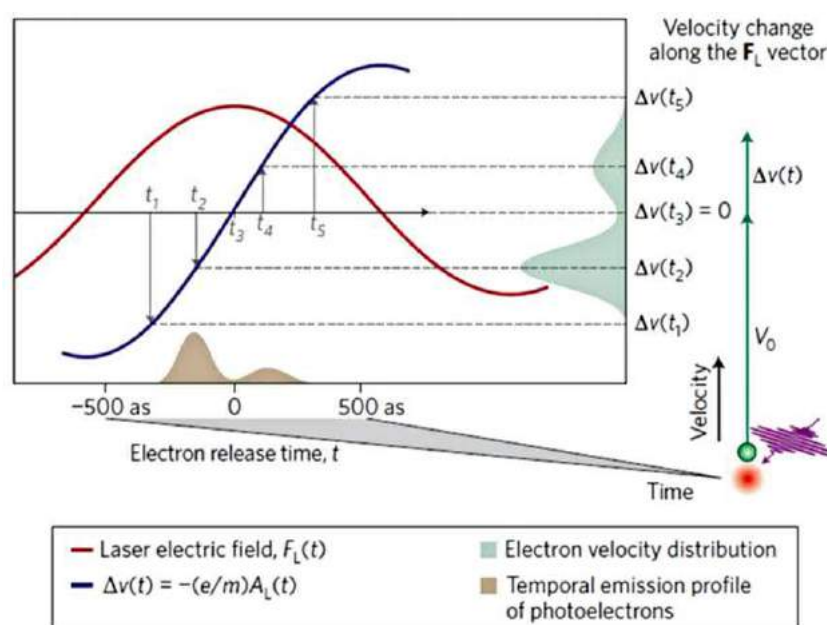


Fig. 8 Concept of attosecond streak camera: Photoelectrons emitted by an attosecond pulse are probed by the presence of a femtosecond (streaking) pulse. The final photoelectron momentum is related to the vector potential of the streaking pulse at the instant of release of the electron wave packet. Reprinted with permission from Mcmillan Publishers Ltd., Krausz, F., Stockmann, M.I., 2014. Nat. Photonics 8, 205, copyright (2014).

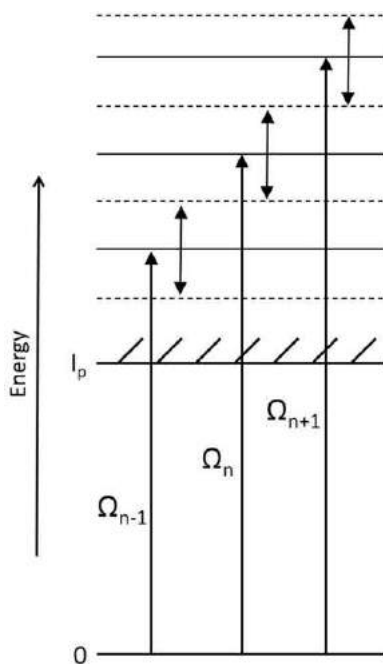


Fig. 9 Illustration of energy levels with main bands (solid lines) and side bands (dashed lines) in the continuum of the spectrum, as well as transitions involved in RABBITT principle.

intrinsic phase related to the target. In combination with measurement of the amplitudes of the harmonics the phase information can be used to reconstruct the temporal profile of the attosecond pulses and the electron wave packets initiated by these pulses. In the ideal (limiting) case of a CW monochromatic field and a periodic repetition of identical attosecond pulses in the train, the RABBITT scheme provides equivalent information as the attosecond streak camera technique, discussed in Section Spectral selection. Since it is usually less demanding experimentally to generate attosecond pulse trains and the probe infrared field can be weaker than in the case of streaking, the RABBITT is often considered more practical. Accordingly, applications of the RABBITT technique include delay measurements in atomic and condensed matter systems.

Similar kind of photoelectron spectroscopy has been performed using stronger femtosecond laser pulses as well, which induce higher-order near-infrared photon transitions in the continuum (or, equivalently larger photoelectron momentum shifts). Consequently, interferences between the main bands in the angular resolved photoelectron energy spectrum appear and information about photoelectron emission and recollision can be retrieved.

Ion spectroscopy

As an alternative to observation of the photoelectron, the population of ionic states and migration and localization of charges in the course of molecular dissociation or fragmentation can be detected to observe attosecond electron dynamics. In this set-up the attosecond pulse is used to ionize and/or excite the molecule and induces an electron wave packet dynamics between different electronic states on the attosecond timescale. The electric field of the femtosecond laser pulse, applied at a variable time delay, probes this dynamics either via ionization of the electron or by driving it inside the dissociating molecular system. The population in the ion channels as well as the localization of the electron charges on the fragments depends on the time delay between the two pulses and contains time information about the attosecond electron dynamics inside the molecule. Analysis of the attosecond dynamics can be complicated by correlation between the photoelectron and the residual molecular ion as well as nonadiabatic response of the electron wave packet to the driving oscillating electric field.

Absorption Spectroscopy With Attosecond Pulses

In another class of attosecond measurements time-resolved spectral information is obtained by applying a femtosecond near-infrared laser pulse either before or after the attosecond pulse. In these pump-probe techniques, which are complementary to those discussed in Section Photoelectron and Ion Spectroscopy With Attosecond Pulses, the time information is contained in the radiation transmitted through the gas medium, liquid or solid.

Attosecond transient absorption spectroscopy

In transient absorption spectroscopy the absorption by a medium following the interaction with a pump and a probe pulses, that are time-delayed with respect to each other, is measured. In the application to attosecond measurements a weak attosecond XUV

pulse is used together with an intense femtosecond near-infrared pulse. The technique can be applied for either sequence of the pulses, where the near-infrared pulse either precedes (conventional pulse sequence) or follows (unconventional pulse sequence) the XUV pulse. Although the time-integrated absorption spectrum is observed, fast transient dynamics may be imprinted onto the spectrum. The response of the medium depends on the interaction at the single-atom level, which is proportional to the imaginary part of the product of Fourier transforms of the one-electron dipole moment (Eq. (5)) and the two-color electric field, and the collective response from all atoms via the macroscopic polarization. Temporal resolution in such a transient absorption experiment is achieved by recording the spectrum as a function of the delay between the two pulses.

In the conventional sequence the attosecond pulse probes the dynamics initiated by the near-infrared pump pulse. With this set-up electron wave packet dynamics in the valence shells of atomic systems coherently excited by an intense femtosecond pulse has been measured. In another application the time dependence of excitations in the conduction band of silicon dioxide have been observed. Using the opposite pulse sequence, in which the attosecond pulse comes first as a pump, spectral features corresponding to light-induced states in the continuum, in particular the lifetimes and lineshapes related to autoionizing states, have been analyzed.

Attosecond four-wave mixing spectroscopy

Very recently, earlier proposals to extend the concepts of 2D Fourier transform and four-wave mixing spectroscopy by using attosecond pulses have been implemented in the experiment. To this end, the spectrum of the attosecond pulses were comprised to just one harmonic, which is resonant with a transition to an excited state in the atomic target. Further coupling of states in the atom by a near-infrared pulse, leads to the population of a state via the absorption of a combination of one XUV and two near-infrared photons. The decay of the state back to the ground state induces the emission of a (fourth) photon at a specific energy. Detection of the emitted radiation as a function of the time delay between the attosecond XUV pulse train and the femtosecond near-infrared pulse allows to retrieve time-resolved information about the dynamics in the coupled states.

References

- Calegari, F., Sansone, G., Stagira, S., Vozzi, C., Nisoli, M., 2006. Advances in attosecond science. *J. Phys. B: At., Mol. Opt. Phys.* 49, 062001.
- Chang, Z., 2011. *Fundamentals of Attosecond Optics*. CRC Press, Taylor & Francis Group.
- Corkum, P.B., Krausz, F., 2007. Attosecond science. *Nat. Phys.* 3, 381–387.
- Dahlström, J.M., L'Huillier, A., Maquet, A., 2012. Introduction to attosecond delays in photoionization. *J. Phys. B: At. Mol. Opt. Phys.* 45, 183001.
- Gallmann, L., Cirelli, C., Keller, U., 2012. Attosecond science: recent highlights and future trends. *Annu. Rev. Phys. Chem.* 63, 447–469.
- Kapteyn, H.C., Murnane, M.M., Christov, I.P., 2005. Extreme nonlinear optics: coherent X rays from lasers. *Phys. Today* 59 (3), 39–44.
- Krausz, F., Ivanov, M., 2009. Attosecond physics. *Rev. Mod. Phys.* 81, 163–234.
- Leone, S.R., McCurdy, C.W., Burgdörfer, J., *et al.*, 2014. What will it take to observe processes in 'real time'? *Nat. Photonics* 8, 162–166.
- Leone, S.R., Neumark, D.M., 2016. Attosecond science in atomic, molecular, and condensed matter physics. *Faraday Discuss.* 194, 15–39.
- Pazourek, R., Nagele, S., Burgdörfer, J., 2015. Attosecond chronoscopy of photoemission. *Rev. Mod. Phys.* 87, 765–802.
- Peng, L.-Y., Jiang, W.-C., Geng, J.-W., Xiong, W.-H., Gong, Q., 2015. Tracing and controlling electronic dynamics in atoms and molecules by attosecond pulses. *Phys. Rep.* 575, 1–71.
- Popmintchev, T., Chen, M.-C., Arpin, P., Murnane, M.M., Kapteyn, H.C., 2011. The attosecond nonlinear optics of bright coherent X-ray generation. *Nat. Photonics* 4, 822–832.
- Ramasesha, K., Leone, S.R., Neumark, D.M., 2016. Real-time probing of electron dynamics using attosecond time-resolved spectroscopy. *Annu. Phys. Rev. Chem.* 67, 41–63.
- Vrakking, M.J.J., 2014. Attosecond imaging. *Phys. Chem. Chem. Phys.* 16, 2775–2789.
- Wu, M., Chen, S., Camp, S., Schafer, K.J., Gaarde, M.B., 2016. Theory of strong-field attosecond transient absorption spectroscopy. *J. Phys. B: At. Mol. Opt. Phys.* 49, 062003.

Introduction

Chemistry is concerned with the induction and observation of changes in matter, where the changes are to be understood at the molecular level. Spectroscopy is the principal experimental tool for connecting the macroscopic world of matter with the microscopic world of the molecule, and is, therefore, of central importance in chemistry. Since its invention, the laser has greatly expanded the capabilities of the spectroscopist. In linear spectroscopy, the monochromaticity, coherence, high intensity, and high degree of polarization of laser radiation are ideally suited to high-resolution spectroscopic investigations of even the weakest transitions. The same properties allowed, for the first time, the investigation of the nonlinear optical response of a medium to intense radiation. Shortly after the foundations of nonlinear optics were laid, it became apparent that these nonlinear optical signals could be exploited in molecular spectroscopy, and since then a considerable number of nonlinear optical spectroscopies have been developed. This short article is not a comprehensive review of all these methods. Rather, it is a discussion of some of the key areas in the development of the subject, and indicates some current directions in this increasingly diverse area.

Nonlinear Optics for Spectroscopy

The foundations of nonlinear optics are described in detail elsewhere in the encyclopedia, and in some of the texts listed in the Further Reading section at the end of this article. The starting point is usually the nonlinearity in the polarization, P_i , induced in the sample when the applied electric field, E , is large:

$$P_i = \epsilon_0 \left[\chi_{ij}^{(1)} E_j + \frac{1}{2} \chi_{ijk}^{(2)} E_j E_k + \frac{1}{4} \chi_{ijkl}^{(3)} E_j E_k E_l + \dots \right] \quad (1)$$

where ϵ_0 is the vacuum permittivity, $\chi^{(n)}$ is the n th order nonlinear susceptibility, the indices represent directions in space, and the implied summation over repeated indices convention is used. The signal field, resulting from the nonlinear polarization, is calculated by substituting it as the source polarization in Maxwell's equations and converting the resulting field to the observable, which is the optical intensity.

The nonlinear polarization itself arises from the anharmonic motion of electrons under the influence of the oscillating electric field of the radiation. Thus, there is a microscopic analog of Eq. (1) for the induced molecular dipole moment, μ_i :

$$\mu_i = \left[\epsilon_0 \alpha_{ij} d_j + \epsilon_0^{-2} \beta_{ijk} d_j d_k + \epsilon_0^{-3} \gamma_{ijkl} d_j d_k d_l + \dots \right] \quad (2)$$

in which α is the polarizability, β the first molecular hyperpolarizability, γ the second, etc. The power series is expressed in terms of the displacement field $\mathbf{d}$ rather than $\mathbf{E}$ to account for local field effects. This somewhat complicates the relationship between the molecular hyperpolarizabilities, for example, γ_{ijkl} and the corresponding macroscopic susceptibility, $\chi_{ijkl}^{(3)}$, but it is nevertheless generally true that a molecule exhibiting a high value of γ_{ijkl} will also yield a large third-order nonlinear polarization. This relationship between the macroscopic and microscopic parameters is the basis for an area of chemistry which has influenced nonlinear optics (rather than the inverse). The synthesis of molecules which have giant molecular hyperpolarizabilities has been an active area, because of their potential applications in lasers and electro-optics technology. Such molecules commonly exhibit a number of properties, including an extended linear π electron conjugation and a large dipole moment, which changes between the ground and excited electronic states. These properties are in accord with the predictions of theory and of quantum chemical calculations of molecular hyperpolarizabilities.

Nonlinear optical spectroscopy in the frequency domain is carried out by measuring the nonlinear signal intensity as a function of the frequency (or frequencies) of the incident radiation. Spectroscopic information is accessible because the molecular hyperpolarizability, and therefore the nonlinear susceptibility, exhibits resonances: the signal is enhanced when one or more of the incident or generated frequencies are resonant with the frequency of a molecular transition. The rich array of nonlinear optical spectroscopies arises in part from the fact that with more input fields there are more accessible resonances than is the case with linear spectroscopy. As an example we consider the third-order susceptibility, $\chi_{ijkl}^{(3)} E_j E_k E_l$, in the practically important case of two incident fields at the same frequency, ω_1 , and a third at ω_2 . The difference between the two frequencies is close to the frequency of a Raman active vibrational mode, Ω_{rg} (Fig. 1). The resulting susceptibility can be calculated to have the form:

$$\chi_{ijkl}^{(3)}(-2\omega_1 + \omega_2; \omega_1, \omega_1, -\omega_2) = \frac{N \Delta \rho_{rg} \Omega_{rg} (\alpha_{ij}^R \alpha_{kl}^R + \alpha_{ik}^R \alpha_{jl}^R)}{12 \hbar [\Omega_{rg}^2 - (\omega_1 - \omega_2)^2 + \Gamma^2 - 2i(\omega_1 - \omega_2)\Gamma]} \quad (3)$$

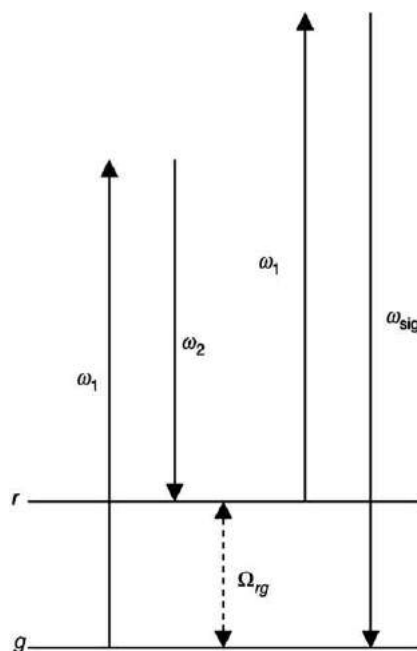


Fig. 1 An illustration of the resonance enhancement of the CARS signal, $\omega_{\text{sig}} = 2\omega_1 - \omega_2$ when $\omega_1 - \omega_2 = \Omega_{rg}$.

in which the α^R are elements of the Raman susceptibility tensor, Γ is the homogeneous linewidth of the Raman transition, $\Delta\rho_{rg}$ a population difference, and N the number density. A diagram illustrating this process is shown in **Fig. 1**, where the signal field is at the anti-Stokes Raman frequency ($2\omega_1 - \omega_2$). The spectroscopic method which employs this scheme is called coherent anti-Stokes Raman spectroscopy (CARS) and is one of the most widely applied nonlinear optical spectroscopies (see below). Clearly, from **Eq. (3)** we can see that the signal will be enhanced when the difference frequency is resonant with the Raman transition frequency.

With essentially the same experimental geometry there will also be nonlinear signals generated at both the Stokes frequency ($2\omega_2 - \omega_1$) and at the lower of the incident frequencies, ω_2 . These signals have distinct phase matching conditions (see below), so they can easily be discriminated from one another, both spatially and energetically. Additional resonance enhancements are possible if either of the individual frequencies is resonant with an electronic transition of the molecule, in which case information on Raman active modes in the excited state is also accessible.

It is worthwhile noting here that there is an equivalent representation of the n th order nonlinear susceptibility tensor $\chi^{(n)}(\omega_{\text{sig}} : \omega_1, \dots, \omega_n)$ as a time domain response function, $R^{(n)}(\tau_1, \dots, \tau_n)$. While it is possible to freely transform between them, the frequency domain representation is the more commonly used. However, the response function approach is increasingly applied in time domain nonlinear optical spectroscopy when optical pulses shorter than the homogeneous relaxation time are used. In that case, the time ordering of the incident fields, as well as their frequencies, is of importance in defining an experiment.

An important property of many nonlinear optical spectroscopies is the directional nature of the signal, illustrated in **Fig. 2**. The directional nonlinear signal arises from the coherent oscillation of induced dipoles in the sample. Constructive interference can lead to a large enhancement of the signal strength. For this to occur the individual induced dipole moments must oscillate in phase – the signal must be phase matched. This requires the incident and the generated frequencies to travel in the sample with the same phase velocity, $k_\omega/\omega = c/n_\omega$ where n_ω is the index of refraction and k_ω is the wave propagation constant at frequency ω . For the simplest case of second-harmonic generation (SHG), in which the incident field oscillates at ω and the generated field at 2ω , phase matching requires $2k_\omega = k_{2\omega}$. The phase-matching condition for the most efficient generation of the second harmonic is when the phase mismatch, $\Delta k = 2k_\omega - k_{2\omega} = 0$, but this is not generally fulfilled due to the dispersion of the medium, $n_\omega < n_{2\omega}$. In the case of SHG, a coherence length L can be defined as the distance traveled before the two waves are 180° out of phase, $L = |\pi/\Delta k|$. For cases in which the input frequencies also have different directions, as is often the case when laser beams at two frequencies are combined (e.g., CARS), the phase-matching condition must be expressed as a vectorial relationship, hence, for CARS, $\Delta \mathbf{k} = \mathbf{k}_{\text{AS}} - 2\mathbf{k}_1 + \mathbf{k}_2 \approx 0$, where $\mathbf{k}_{\text{AS}}$ is the wavevector of the signal at the anti-Stokes frequency (**Fig. 2**). Thus for known input wavevectors one can easily calculate the expected direction of the output signal, $\mathbf{k}_s$. This is illustrated for a number of important cases in **Fig. 2**.

Nonlinear Optical Spectroscopy in Chemistry

As already noted, there is a vast array of nonlinear optical spectroscopies, so it is clear some means of classification will be required. For coarse graining the order of the nonlinear process is very helpful, and that is the scheme we will follow here.

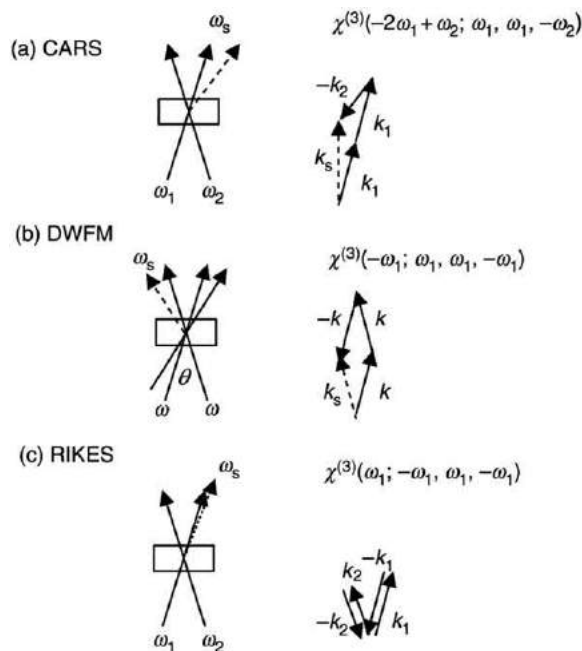


Fig. 2 Experimental geometries and corresponding phase matching diagrams for (a) CARS (b) DFWM (c) RIKES. Note that only the RIKES geometry is fully phase matched.

Second-Order Spectroscopies

Inspection of Eq. (1) shows that in gases, isotropic liquids, and solids where the symmetry point group contains a center of inversion, $\chi^{(2)}$ is necessarily zero. This is required to satisfy the symmetry requirement that polarization must change sign when the direction of the field is inverted, yet for a quadratic, or any even order dependence on the field strength, it must remain positive. Thus second-order nonlinearities might not appear very promising for spectroscopy. However, there are two cases in which second-order nonlinear optical phenomena are of very great significance in molecular spectroscopy, harmonic generation and the spectroscopy of interfaces.

Harmonic conversion

Almost every laser spectroscopist will have made use of second-harmonic or sum frequency generation for frequency conversion of laser radiation. Insertion of an oscillating field of frequency ω into Eq. (1) yields a second-order polarization oscillating of 2ω . If two different frequencies are input, the second-order polarization oscillates at their sum and difference frequencies. In either case, the second-order polarization acts as a source for the second-harmonic (or sum, or difference frequency) emission, provided $\chi^{(2)}$ is nonzero. The latter can be arranged by selecting a noncentrosymmetric medium for the interaction, the growth of such media being an important area of materials science. Optically robust and transparent materials with large values of $\chi^{(2)}$ are available for the generation of wavelengths shorter than 200 nm to longer than 5 μm . Since such media are birefringent by design, a judicious choice of angle and orientation of the crystal with respect to the input beams allows a degree of control over the refractive indices experienced by each beam. Under the correct phase matching conditions $n_\omega \approx n_{2\omega}$, and very long interaction lengths result, so the efficiency of signal generation is high.

Higher-order harmonic generation in gases is an area of growing importance for spectroscopy in the deep UV and X-ray region. The generation of such short wavelengths depends on the ability of amplified ultrafast solid state lasers to generate extremely high instantaneous intensities. The mechanism is somewhat different to the one outlined above. The intense pulse is injected into a capillary containing an appropriate gas. The high electric field of the laser causes ionization of atoms. The electrons generated begin to oscillate in the applied laser field. The driven recombination of the electrons with the atoms results in the generation of the high harmonic emission. Although the mechanism differs from the SHG case, questions of phase matching are still important. By containing the gas in a corrugated waveguide phase matching is achieved, considerably enhancing the intensity of the high harmonic. Photon energies of hundreds of electronvolts are possible using this technique. The generation of such high energies is not yet routine, but a number of potential applications are already apparent. A powerful coherent source of X-ray and vacuum UV pulses will certainly aid surface analysis techniques such as UV photo-emission and X-ray photoelectron spectroscopy. Much excitement is currently being generated by the possibility of using intense ultrashort X-ray pulses to record structural dynamics on an ultrafast timescale.

Surface second-order spectroscopy

At an interface inversion symmetry is absent by definition, so the second-order nonlinear susceptibility is finite. If the two bulk phase media are themselves isotropic, then even a weak second-order signal necessarily arises from the interface. This surface-specific

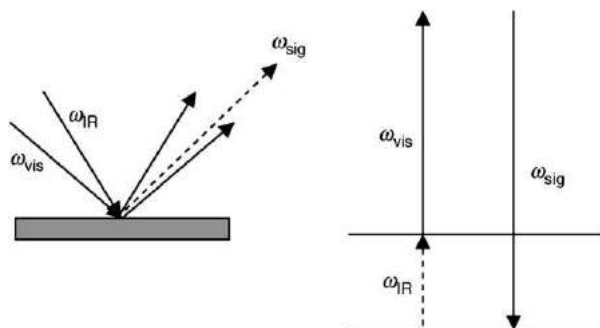


Fig. 3 The experimental geometry for SFG, and an illustration of the resonance enhancement of $\omega_{\text{sig}} = \omega_{\text{IR}} + \omega_{\text{vis}}$ at a Raman and IR allowed vibrational transition.

all-optical signal is unique, because it can be used to probe the interface between two condensed phases. This represents a great advantage over every other form of surface spectroscopy. In linear optical spectroscopy, the signal due to species at the interface are usually swamped by contributions from the bulk phase. Other surface-specific signals do exist, but they rely on the scattering of heavy particles (electrons, atoms) and so can only be applied to the solid vacuum interface. For this reason the techniques of surface SHG and SFG are widely applied in interface spectroscopy.

The most widely used method is sum frequency generation (SFG) between temporally overlapped tuneable infrared and fixed frequency visible lasers, to yield a sum frequency signal in the visible region of the spectrum. The principle of the method is shown in Fig. 3. The surface nonlinear susceptibility exhibits resonances at vibrational frequencies, which are detected as enhancements in the visible SFG intensity. Although the signal is weak, it is directional and background free, so relatively easily measured by photon counting techniques. Thus, SFG is used to measure the vibrational spectra of essentially any optically accessible interface. There are, however, some limitations on the method. The surface must exhibit a degree of order – if the distribution of adsorbate orientation is isotropic the signal again becomes zero by symmetry. Second, a significant enhancement at the vibrational frequency requires the transition to be both IR and Raman allowed, as suggested by the energy level diagram (Fig. 3).

The SHG signal can also be measured as a function of the frequency of the incident laser, to recover the electronic spectrum of the interface. This method has been used, particularly in the study of semiconductor surfaces, but generally the electronic spectra of adsorbates contain less information than their vibrational spectra. However, by measuring the SHG intensity as a function of time, information on adsorbate kinetics is obtained, provided some assumptions connecting the surface susceptibility to the molecular hyperpolarizability are made. Finally, using similar assumptions, it is possible to extract the orientational distribution of the adsorbate, by measuring the SHG intensity as a function of polarization of the input and output beams. For these reasons, SHG has been widely applied to analyze the structure and dynamics of interfaces.

Third-Order Spectroscopies

The third-order coherent Raman spectroscopies were introduced above. One great advantage of these methods over conventional Raman is that the signal is generated in a coherent beam, according to the appropriate phase matching relationship (Fig. 2). Thus, the coherent Raman signal can easily be distinguished from background radiation by spatial filtering. This has led to CARS finding widespread application in measuring spectra in (experimentally) hostile environments. CARS spectroscopy has been widely applied to record the vibrational spectra of flames. Such measurements would obviously be very challenging for linear Raman or IR, due to the strong emission from the flame itself. The directional CARS signal in contrast can be spatially filtered, minimizing this problem. CARS has been used both to identify unstable species formed in flames, and to probe the temperature of the flame (e.g., from measured population differences Eq. (3)). A second advantage of the technique is that the signal is only generated when the two input beams overlap in space. Thus, small volumes of the sample can be probed. By moving the overlap position around in the sample, spatially resolved information is recovered. Thus, it is possible to map the population of a particular transient species in a flame.

CARS is probably the most widely used of the coherent Raman methods, but it does have some disadvantages, particularly in solution phase studies. In that case the resonant $\chi^{(3)}$ signal (Eq. (3)) is accompanied by a nonresonant third-order background. The interference between these two components may result in unusual and difficult to interpret lineshapes. In this case, some other coherent Raman methods are more useful. The phase matching scheme for Raman-induced Kerr effect spectroscopy (RIKES) is shown in Fig. 2. The RIKES signal is always phase matched, which leads to a long interaction length. However, the signal is at ω_2 and in the direction of ω_2 , which would appear to be a severe disadvantage in terms of detection. Fortunately, if the input polarizations are correctly chosen the signal can be isolated by its polarization. In an important geometry, the signal (ω_2) is isolated by transmission through a polarizer oriented at 45° to a linearly polarized pump (ω_1). The pump is overlapped in the sample with the probe (ω_2) linearly polarized at -45° . Thus the probe is blocked by the polarizer, but the signal is transmitted. This geometry may be viewed as pump-induced polarization of the isotropic medium to render it birefringent, thus inducing ellipticity in the transmitted probe, such that the signal leaks through the analyzing polarizer (hence the alternative name optical Kerr effect).

In this geometry, it is possible to greatly enhance signal-to-noise ratios by exploiting interference between the probe beam and the signal. Placing a quarterwave plate in the probe beam with its fast axis aligned with the probe polarization, and reorienting it slightly ($<1^\circ$) yields a slightly elliptically polarized probe before the sample. A fraction of the probe beam, the local oscillator (LO), then also leaks through the analyzing polarizer, temporally and spatially overlapped with the signal. Thus, the signal and LO fields are seen by the detector, which measures the intensity as:

$$I(t) = \frac{nc}{8\pi} |\mathbf{E}_{\text{LO}}(t) + \mathbf{E}_s(t)|^2 = I_{\text{LO}}(t) + I_s(t) + \frac{nc}{4\pi} \text{Re}[\mathbf{E}_s^*(t) \cdot \mathbf{E}_{\text{LO}}(t)] \quad (4)$$

where the final term may be very much larger than the original signal, and is linear in $\chi^{(3)}$. This term is usually isolated from the strong I_{LO} by lock-in detection. This method is called optical heterodyne detection (OHD), and generally leads to excellent signal to noise. It can be employed with other coherent signals by artificially adding the LO to the signal, provided great care is taken to ensure a fixed phase relationship between LO and signal. In the RIKES experiment, however, the phase relationship is automatic. The arrangement described yields an out-of-phase LO, and measures the real part of $\chi^{(3)}$, the birefringence. Alternatively, the quarterwave plate is omitted, and the analyzing polarizer is slightly reoriented, to introduce an in-phase LO, which measures the imaginary part of $\chi^{(3)}$, the dichroism. This is particularly useful for absorbing media. The OHD-RIKES method has been applied to measure the spectroscopy of the condensed phase, and has found particularly widespread application in transient studies (below).

Degenerate four wave mixing (DFWM) spectroscopy is a simple and widely used third-order spectroscopic method. As the name implies, only a single frequency is required. The laser beam is split into three, and recombined in the sample, in the geometry shown in Fig. 2. The technique is also known as laser-induced grating scattering. The first two beams can be thought of as interfering in the sample to write a spatial grating, with fringe spacing dependent on the angle between them. The signal is then scattered from the third beam in the direction expected for diffraction from that grating. The DFWM experiment has been used to measure electronic spectra in hostile environments, by exploiting resonances with electronic transitions. It has also been popular in the determination of the numerical value of $\chi^{(3)}$, partly because it is an economical technique, requiring only a single laser, but also because different polarization combinations make it possible to access different elements of $\chi^{(3)}$. The technique has also been used in time resolved experiments, where the decay of the grating is monitored by diffraction intensity from a time delayed third pulse.

Two-photon or, more generally, multiphoton excitation has applications in both fundamental spectroscopy and analytical chemistry. Two relevant level schemes are shown in Fig. 4. Some property associated with the final state permits detection of the multiphoton absorption, for example, fluorescence in (a) and photocurrent in (b).

Excitation of two-photon transitions, as in Fig. 4(a), is useful in spectroscopy because the selection rules are different to those for the corresponding one-photon transition. For example the change in angular momentum quantum number, ΔL , in a two-photon transition is 0, ± 2 , so, for example, an atomic S to D transition can be observed. High spectroscopic resolution may be attained using Doppler free two-photon absorption spectroscopy. In this method, the excitation beams are arranged to be counter-propagating, so that the Doppler broadening is cancelled out in transitions where the two excitation photons arise from beams with opposing wavevectors. In this case, the spectroscopic linewidth is governed only by the homogeneous dephasing time.

The level scheme in Fig. 4(b) is also widely used in spectroscopy, but in this case the spectrum of the intermediate state is obtained by monitoring the photocurrent as a function of ω_1 . The general technique is known as resonance enhanced multiphoton ionization (REMPI) and yields high-quality spectra of intermediate states which are not detectable by standard methods, such as fluorescence. The sensitivity of the method is high, and it is the basis of a number of analytical applications, often in combination with mass spectrometry.

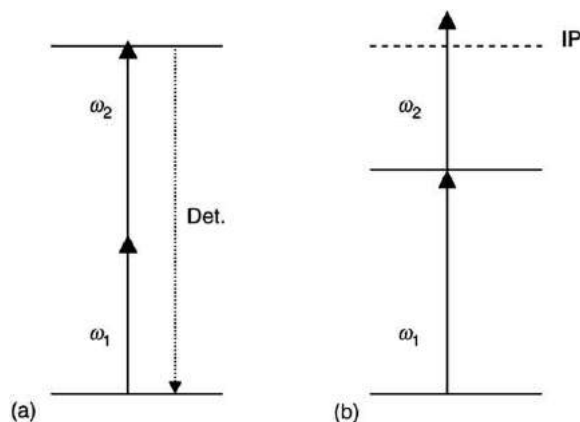


Fig. 4 Two cases of resonant two photon absorption. In (a) the excited state is two-photon resonant, and the process is detected by the emission of a photon. In (b) the intermediate state is resonant, and the final energy is above the ionization potential (IP) so that photocurrent or mass detection can be used.

Ultrafast Time Resolved Spectroscopy

The frequency and linewidth of a Raman transition may be extracted from the CARS measurement, typically by combining two narrow bandwidth pulsed lasers, and tuning one through the resonance while measuring the nonlinear signal intensity. The time resolved analogue requires two pulses, typically of a few picoseconds duration (and therefore a few wavenumbers bandwidth) at ω_1 and ω_2 to be incident on the sample. This pair coherently excites the Raman mode. A third pulse at ω_1 is incident on the sample a time t later, and stimulates the CARS signal at $2\omega_1 - \omega_2$ in the phase-matched direction. The decay rate of the vibrational coherence is measured from the CARS intensity as a function of the delay time. Thus, the frequency of the mode is measured in the frequency domain, but the linewidth is measured in the time domain. If very short pulses are used (such that the pulsewidth is shorter than the inverse frequency of the Raman active mode) the Raman transition is said to be impulsively excited, and the CARS signal scattered by the time delayed pulse reveals an oscillatory response at the frequency of the Raman active mode, superimposed on its decay. Thus, in this case, spectroscopic information is measured exclusively in the time domain. In the case of nonlinear Raman spectroscopy, similar information is available from the frequency and the time domain measurements, and the choice between them is essentially one of experimental convenience. For example, time domain CARS, RIKES, and DFWM spectroscopy have turned out to be particularly powerful routes to extracting low-frequency vibrational and orientational modes of liquids and solutions, thus providing detailed insights into molecular interactions and reaction dynamics in the condensed phase.

Other time domain experiments contain information that is not accessible in the frequency domain. This is particularly true of photon echo methods. The name suggests a close analogy with nuclear magnetic resonance (NMR) spectroscopy, and the (optical) Bloch vector approach may be used to describe both measurements, although the transition frequencies and time-scales involved differ by many orders of magnitude. In the photon echo experiment, two or three ultrafast pulses with carefully controlled interpulse delay times are resonant with an electronic transition of the solute. In the two-pulse echo, the echo signal is emitted in the phase match direction at twice the interpulse delay, and its intensity as a function of time yields the homogeneous dephasing time associated with the transition. In the three-pulse experiment the pulses are separated by two time delays. By measuring the intensity of the stimulated echo as a function of both delay times it is possible to separately determine the dephasing time and the population relaxation time associated with the resonant transition. Such information is not accessible from linear spectroscopy, and can be extracted only with difficulty in the frequency domain. The understanding of photon echo spectroscopy has expanded well beyond the simple description given here, and it now provides unprecedented insights into optical dynamics in solution, and thus informs greatly our understanding of chemistry in the condensed phase. The methods have recently been extended to the infra red, to study vibrational transitions.

Higher-Order and Multidimensional Spectroscopies

The characteristic feature of this family of spectroscopies is the excitation of multiple resonances, which may or may not require measurements at $\chi^{(n)}$ with $n > 3$. Such experiments require multiple frequencies, and may yield weak signals, so they only became experimentally viable upon the availability of stable and reliable solid state lasers and optical parametric generators. Measurements are made in either the time or the frequency domain, but in either case benefit from heterodyne detection.

One of the earliest examples was two-dimensional Raman spectroscopy, where multiple Raman active modes are successively excited by temporally delayed pulse pairs, to yield a fifth-order nonlinear signal. The signal intensity measured as a function of both delay times (corresponding to the two dimensions) allows separation of homogeneous and inhomogeneous contributions to the line shape. This prodigiously difficult $\chi^{(5)}$ experiment has been completed in a few cases, but is plagued by interference from third-order signals.

More widely applicable are multidimensional spectroscopies using infrared pulses or combinations of them with visible pulses. The level scheme for one such experiment is shown in Fig. 5 (which is one of many possibilities). From the scheme, one can see that the nonlinear signal in the visible depends on two resonances, so both can be detected. This can be regarded as a multiply resonant nondegenerate four-wave mixing (FWM) experiment. In addition, if the two resonant transitions are coupled, optical excitation of one affects the other. Thus, by measuring the signal as a function of both frequencies, the couplings between transitions are observed. These appear as cross peaks when the intensity is plotted in the two frequency dimensions, very much as with 2D NMR. This technique is already providing novel information on molecular structure and structural dynamics in liquids, solutions, and proteins.

Spatially Resolved Spectroscopy

A recent innovation is nonlinear optical microscopy. The nonlinear dependence of signal strength on intensity means that nonlinear processes are localized at the focal point of a lens. When focusing is strong, such as in a microscope objective, spatial localization of the nonlinear signal can be dramatic. This is the basis of the two-photon fluorescence microscopy method, where a high repetition rate source of low-energy ultrafast pulses is focused by a microscope objective into a sample labeled with a fluorescent molecule, which has absorption at half the wavelength of the incident photons (Fig. 4). The fluorescence is necessarily localized at the focal point because of its dependence on the square of the incident intensity. By measuring intensity while scanning the position of the focal point in space, a 3D image of the distribution of the fluorophore is constructed. This technique turns out to have a number of advantages over one photon fluorescence microscopy, most notably in terms of ease of implementation, minimization of sample damage, and depth resolution. The technique is widely employed in cell biology.

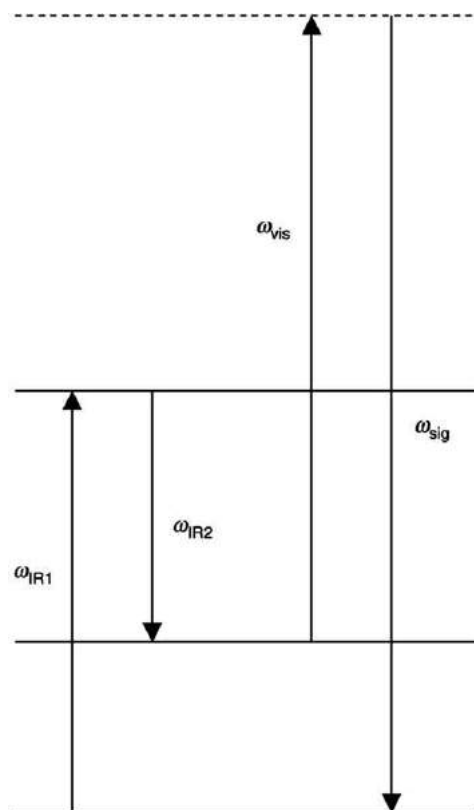


Fig. 5 Illustration of multiple resonance enhancements in a FWM geometry, from which 2D spectra may be generated.

Stimulated by the success of two-photon microscopy, further nonlinear microscopies have been developed, all relying on the spatial localization of the signal. CARS microscopy has been demonstrated to yield 3D images of the distribution of vibrations in living cells. It would be difficult to recover such data by linear optical microscopy. SHG has been applied in microscopy. In this case, by virtue of the symmetry selection rule referred to above, a 3D image of orientational order is recovered. Both these and other nonlinear signals provide important new information on complex heterogeneous samples, most especially living cells.

See also: Transient Holographic Grating Techniques in Chemical Dynamics

Further Reading

- Andrews, D.L., Allcock, P., 2002. *Optical Harmonics in Molecular Systems*. Weinheim, Germany: Wiley-VCH.
- Andrews, D.L., Demidov, A.A. (Eds.), 2002. *An Introduction to Laser Spectroscopy*. New York: Kluwer Academic.
- Butcher, P.N., Cotter, D., 1990. *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.
- Cheng, J.X., Xie, X.S., 2004. Coherent anti-Stokes Raman scattering microscopy: Instrumentation, theory, and applications. *Journal of Physical Chemistry B* 108, 827.
- de Boei, W.P., Pshenichnikov, M.S., Wiersma, D.A., 1998. Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49, 99.
- Eisenthal, K.B., 1992. Equilibrium and dynamic processes at interfaces by 2nd harmonic and sum frequency generation. *Annual Review of Physical Chemistry* 43, 627.
- Fleming, G.R., 1986. *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Oxford University Press.
- Fleming, G.R., Cho, M., 1996. Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47, 109.
- Fourkas, J.T., 2002. Higher-order optical correlation spectroscopy in liquids. *Annual Review of Physical Chemistry* 53, 17.
- Hall, G., Whitaker, B.J., 1994. Laser-induced grating spectroscopy. *Journal of Chemistry Society Faraday Transactions* 90, 1.
- Heinz, T.F., 1991. Second order nonlinear optical effects at surfaces and interfaces. In: Ponath, H.E., Stegeman, G.I. (Eds.), *Nonlinear Surface Electromagnetic Phenomena*, pp. 353.
- Hesselink, W.H., Wiersma, D.A., 1983. Theory and experimental aspects of photon echoes in molecular solids. In: Hochstrasser, R.M., Agranovich, V.M. (Eds.), *Spectroscopy and Excitation Dynamics of Condensed Molecular Systems*. Amsterdam: North Holland (Chapter 6).
- Levenson, M.D., Kano, S., 1988. *Introduction to Nonlinear Laser Spectroscopy*. San Diego, CA: Academic Press.
- Meech, S.R., 1993. Kinetic application of surface nonlinear optical signals. In: Lin, S.H., Fujimura, Y., Villaeys, A. (Eds.), *Advances in Multiphoton Processes and Spectroscopy*, vol. 8. Singapore: World Scientific, pp. 281.
- Mukamel, S., 1995. *Principles of Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.
- Rector, K.D., Fayer, M.D., 1998. Vibrational echoes: a new approach to condensed-matter vibrational spectroscopy. *International Review of Physical Chemistry* 17, 261.
- Richmond, G.L., 2001. Structure and bonding of molecules at aqueous surfaces. *Annual Review of Physical Chemistry* 52, 357.

- Shen, Y.R., 1984. *The Principles of Nonlinear Optics*. New York: Wiley.
- Smith, N.A., Meech, S.R., 2002. Optically heterodyne detected optical Kerr effect – Applications in condensed phase dynamics. *International Review of Physical Chemistry* 21, 75.
- Tolles, W.M., Nibler, J.W., McDonald, J.R., Harvey, A.B., 1977. Review of theory and application of coherent anti-Stokes Raman spectroscopy (CARS). *Applied Spectroscopy* 31, 253.
- Wright, J.C., 2002. Coherent multidimensional vibrational spectroscopy. *International Review of Physical Chemistry* 21, 185.
- Zipfel, W.R., Williams, R.M., Webb, W.W., 2003. Nonlinear magic multiphoton microscopy in the biosciences. *Nature Biotechnology* 21, 1368.
- Zyss, J. (Ed.), 1994. *Molecular Nonlinear Optics*. San Diego: Academic Press.

Alternative Plasmonic Materials

Gururaj V Naik, Rice University, Houston, TX, United States

© 2018 Elsevier Ltd. All rights reserved.

The contents of this chapter are adapted from P. R. West *et al.*, *Laser & Photon. Rev.* **4**(6) 795–808 (2010), and G. V. Naik *et al.*, *Adv. Mater.* **25**(24) 3264–3294 (2013).

Introduction

Modern science has enabled techniques to precisely control the flow of heat, light, charge, spin and other quantities at nanometer scale. Each of these new capabilities have given rise to new technologies that have or on their way to revolutionize our society. For example, controlling the flow of charge at nanoscale has led to nanoelectronics – a technology that brought us into the information age. An impact of such large magnitude was possible only because scientists from many different disciplines worked together to solve major technological problems allowing us to build billions of transistors on a millimeter sized chip. Electronics in 1980s used only a few elements such as silicon, oxygen, boron, phosphorous, and aluminum to build the devices. However, within two decades, the number of elements expanded to cover more than half the periodic table. Many ground-breaking material innovations led to the use of new materials which enabled electronic devices to be simultaneously smaller and faster by nearly 3 orders. Such explosion of material innovation transformed electronics from a niche technology to a revolutionizing technology. Similar to nanoelectronics, other technologies – especially nanophotonics can make an impact of the same or even greater scale if only such big scale material innovation happens. Nanophotonics had an explosion of revolutionary concepts in the past decade with the advent of metamaterials and transformation optics. However, practical implementation of these concepts and their commercialization require material innovation. Thus, the promise of nanophotonics as the next generation technology relies upon material innovation.

In order to better appreciate the need for materials innovation in nanophotonics, it is useful to understand how nanophotonic devices operate and what they are made of. Nanophotonics enables confining light to nanoscale – a feat that is possible either by high refractive index or resonant materials. While materials with very high refractive index at optical frequencies are not available in nature, coupling light to electronic resonances in materials is a common approach. Free electron plasma as in metals can support electronic or plasmon-polaritonic resonance over a broad spectrum: infrared, visible and ultra-violet. Thus, metals are referred to as plasmonic materials and form one of the basic building blocks of nanophotonic devices. Since metals have high optical losses, nanophotonic devices often pick the metals with lowest ohmic losses – the noble metals. Thus, the material library of nanophotonics has been limited for a long time to only noble metals and some common dielectrics and semiconductors. Drawing analogy to nanoelectronics, the situation of nanophotonics resembled that of electronics in 1980s. But in the recent past, the quest for new materials in the nanophotonics library has intensified which has led to a new research direction – searching alternatives to metals. Here, in this article, the recent breakthroughs in materials innovation for nanophotonics – especially in searching alternatives to metals – will be discussed. In the next few sections, the need for alternative plasmonic materials, what they are, how to find them, and what applications do they benefit will be discussed. Given that the plasmonic material research is currently rapidly growing, this article has limited itself to summarizing the major breakthroughs in this emerging field of research.

A Case for Alternative Plasmonic Materials

Metals such as gold and silver are commonly used in plasmonic and optical metamaterial devices because of their small ohmic losses or high DC conductivity. However, at optical frequencies another loss mechanism namely interband transitions plays an important role in these metals. Loss arising from interband transitions occurs when a valence electron in a metal absorbs a photon to jump to the Fermi surface or an electron near the Fermi-surface absorbs a photon to jump to the next unoccupied conduction band. This is the loss mechanism that is responsible for the color of copper and gold. [Fig. 1](#) shows the imaginary part of permittivity of gold in the optical range adopted from [Johnson and Christy \(1972\)](#). The loss depicted by the imaginary part of permittivity can be split into two parts: interband and intraband losses. The intraband losses (or Drude losses) in gold are high in the near-infrared (NIR) and are lower for shorter wavelengths. On the other hand, interband losses in gold are high for the shorter wavelengths in the visible range. These additional losses at optical frequencies caused by interband transitions make metals such as gold unsuitable for many plasmonic and metamaterial devices.

For applications like transformation optics (TO) that require that the imaginary part of a metal's dielectric function be small, even if interband transitions are absent in the operating spectral range, intraband transitions and scattering losses are invariably present and often result in large overall losses. The high loss in conventional plasmonic materials is but one of the major disadvantages in using metals like gold and silver in plasmonics, optical metamaterials and TO. One issue is that the magnitude of the real part of the permittivity is very large in conventional metals. This is a problem in designing many TO devices because these devices often require meta-molecules with a nearly balanced polarization response ([Cai and Shalaev, 2009](#)). In other words, the polarization response from the metallic components should be of the same order as that from the dielectric components within

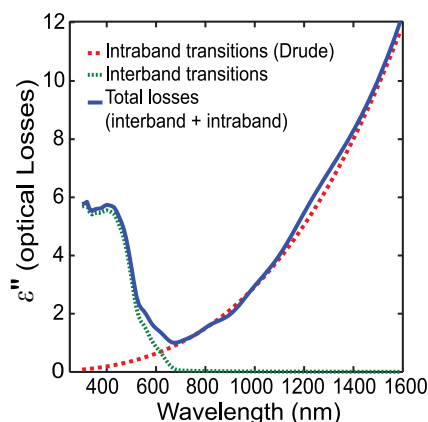


Fig. 1 The imaginary part of permittivity of gold in the optical range (solid line), data from Johnson, P.B., Christy, R.W., 1972. Phys. Rev. B 6, 4370. The individual contributions from free electron losses (intraband transitions) and interband transition losses are shown in dotted lines.

each meta-molecule. When the real parts of permittivity of the metal and dielectric are on the same order, the geometric fill fractions of the metal and dielectric can be readily tuned to match the design requirements. On the other hand, if the magnitude of the real part of permittivity of the metal is a few orders larger than that of the dielectric, the metal fill fraction in the meta-molecule will be a few orders smaller than that of the dielectric. This constraint would necessitate very tiny metal inclusions in the meta-molecule, which poses a number of problems especially in terms of successful nanofabrication. Thus, having smaller magnitudes of real permittivity for plasmonic materials would be advantageous in many applications. Since, the origin of the large magnitude of real permittivity in noble metals can be traced to their very large carrier concentrations, reducing the carrier concentration in metals would help applications such as TO metamaterials.

Aside from the issues of loss and not adjustable dielectric permittivity described above, metals also pose nanofabrication challenges, especially when grown as thin films. Metal thin films exhibit quite different morphologies when compared to bulk metal, which can lead to the degradation of optical properties. First of all, ultra-thin metal films deposited by common techniques such as evaporation or sputtering often grow as semi-continuous or discontinuous films. Metal films exhibit a percolation threshold in the film thickness when grown on commonly used substrates such as glass, quartz, sapphire and silicon. Overcoming this limit would require extra efforts such as using a wetting layer (Chen *et al.*, 2010) or a lattice-matched substrate (Pashley, 1959). However, these approaches have limitations in terms of design integration and scalability. Generally, thin metal films exhibit a structure composed of many small grains. In contrast, thick films are typically composed of large grains, and their optical properties resemble those of the bulk material. When the films are thin, the grainy structure causes additional grain-boundary scattering for free electrons and increases the losses in the metal. Another loss mechanism arising from the microstructure of thin metal films is related to surface roughness. Nanoscale patterning invariably results in rough surfaces and edges, which cause additional scattering and optical losses.

Losses in thin metal films can increase nearly threefold due to grain-boundary scattering. In order to avoid such additional losses in conventional plasmonic materials, single-crystal growth of noble metal films has been attempted (Wang *et al.*, 2015). The improvement in losses is evident, but it is not substantial due to the limitations that arise from the nanopatterning of the metal films.

Another important issue to consider in the context of realistic devices and integration is the chemical stability of the materials. The degradation of metals on exposure to air/oxygen or humidity would pose additional problems in fabrication and integration of devices. Among conventional plasmonic materials, silver and copper are well known to degrade in air, but gold is very stable in air. While copper forms a native oxide layer in air, silver is sensitive to sulfidation and tarnishes to form a layer of silver sulfide. The tarnishing of these metals has a direct consequence on their optical properties, and the optical losses increase, which in turn results in larger values of the imaginary part of the dielectric function.

Another important technological challenge associated with noble metals is that they are not compatible with standard silicon manufacturing processes. This precludes plasmonic and metamaterial devices from leveraging on standard nanofabrication technologies. This also diminishes the possibility of integrating plasmonic and metamaterial components with nanoelectronic components. The compatibility issue with noble metals arises from the fact that these metals can diffuse into silicon to form deep traps, which severely affects the performance of nanoelectronics devices. Hence the integration of noble metals into silicon manufacturing processes is a difficult challenge. Recently, copper has been incorporated into silicon processes, but additional, special processing steps are needed to create diffusion barriers between the silicon and the copper. Gold and silver still remain outside the realm of feasibility for silicon manufacturing processes.

Major drawback of metals is that their optical properties cannot be tuned or adjusted easily. For example, the carrier concentration of metals cannot be changed much with the application of moderate electric fields, optical fields, or temperature, etc. Hence, in applications where switching or modulation of the optical properties is essential, metals cannot be used as tunable elements.

With all the shortcomings of conventional plasmonic materials, researchers have been motivated to search for better alternatives. Many alternatives to metals have been proposed that overcome one or more of the drawbacks mentioned above. The significance of a particular alternative depends on the end application, but general criteria for the choice of an alternative plasmonic material can be outlined from the issues raised in the preceding discussions.

Optical Properties of Materials

Since the rest of the article focuses on optical constants of materials and their influence on device performance, a closer look at the optical response of materials is helpful. Since plasmonic materials support a free electron cloud/plasma, their free-electron response can be described by the Drude model (Dressel *et al.*, 2002) as shown in Eq. (1).

$$\varepsilon_{\text{Drude}}(\omega) = \varepsilon'_{\text{Drude}} + i\varepsilon''_{\text{Drude}} = \varepsilon_b - \frac{\omega_p^2}{(\omega^2 + \gamma^2)} + i \frac{\omega_p^2 \gamma}{(\omega^2 + \gamma^2)\omega} \quad (1)$$

Here ε_b is the polarization response from the core electrons (background permittivity), ω_p is the plasma frequency and γ is the Drude relaxation rate. Relaxation rate is responsible for scattering/ohmic losses and scales directly with the imaginary part of the dielectric function. And, the square of the plasma frequency, ω_p^2 – proportional to carrier concentration (n) – scales directly with the real and imaginary part of the permittivity. In optics, it has been noticed that the Drude scattering rate γ is a phenomenological parameter that is dependent the microstructure of the plasmonic structure. For example, the internal grain size of the metal film can increase the carrier scattering rate. When the grain size is large, as in a bulk metal, the relaxation rate is given by γ_0 . When the grain size is small, as in the case of thin metal films, Eq. (2) describes the enhanced relaxation rate that arises from additional grain-boundary scattering (Kreibig and Vollmer, 1995).

$$\gamma = \gamma_0 + A \frac{v_F}{d} \quad (2)$$

where A is a dimensionless empirical constant, v_F is the Fermi-velocity of the electrons in the metal and d is the average grain size. Smaller grains result in a larger γ and, hence, higher losses.

Surface roughness is yet another factor which might lead to a larger effective γ . All such effects can be empirically described by a factor called *loss factor* as described in Eq. (3). Experimental findings suggest that a loss factor of 3 to 5 is common in nano-patterned gold and silver films.

$$\gamma = (\text{Loss factor}) \times \gamma_0 \quad (3)$$

While Drude model effectively captures the response from free carriers (or intraband transitions in quantum picture), the response from interband transitions of electrons are harder to model. Interband transitions form a significant loss mechanism in materials at optical frequencies, and occur when electrons jump from one energy band to a higher empty energy band by absorbing incident photons. In metals, when a bound electron absorbs an incident photon, the electron can shift from a lower energy band to the Fermi surface or from near the Fermi surface to the next higher empty energy band. Both of these processes – described as interband transitions – result in high loss at optical frequencies.

In optics, often the response of interband transitions is captured by a series of Lorentz oscillators. Lorentz oscillator is the basic model describing the light-matter interaction in a two-level system, and its form is as given by Eq. (4) (Dressel *et al.*, 2002):

$$\varepsilon_{lk} = \frac{f_{lk}\omega_{p,lk}^2}{\omega_{lk}^2 - \omega^2 - i\omega\Gamma_{lk}} \quad (4)$$

Here, f_{lk} corresponds to the strength of the oscillator at energy levels l and k , ω_{lk} is the resonant frequency corresponding to the difference between the energies of levels l and k , Γ_{lk} is the damping in the oscillator accounting for non-zero line-width of the peak, and $\omega_{p,lk}$ is similar to the plasma frequency given by Eq. (1) with the difference that n here refers to the density of states l or k . When there are many of such interacting energy levels, the effective permittivity can be expressed as a summation over all allowed Lorentzian terms. This is a popular approach referred to as Drude-Lorentz model (see Eq. (5)) to reasonably approximate the dielectric function of metals.

$$\varepsilon(\omega) = \varepsilon_{\text{Drude}}(\omega) + \sum_{lk} \varepsilon_{lk}(\omega) \quad (5)$$

In general, solids with a periodic lattice have electronic energy levels which exist as bands instead of discrete levels, requiring a Joint-Density-of-States (JDOS) description and integration over all the allowed transitions at any given photon energy (a more detailed discussion of this formulation is found in reference Dressel *et al.* (2002)).

Ideal Plasmonic Material

A material with a purely real and negative permittivity ($\varepsilon''=0$ and $\varepsilon'<0$) would be an ideal candidate to replace metals in most of the plasmonic and metamaterial devices. Such a material produces a metallic response to light while exhibiting zero losses. However, it is impossible to have zero losses and negative permittivity simultaneously for all frequencies in any dispersive material

due to the causality condition. All is not lost, though – there can be a frequency interval in which the permittivity is purely real and negative. This is possible without violating causality only when there are large losses present at lower frequencies. A theoretical solution satisfying this requirement has been proposed by [Khurgin and Sun \(2010\)](#). The fact that none of the known metals have this specific band structure explains why natural metals are always with associated losses. In other words, there are no naturally occurring materials that are simultaneously lossless and have negative real permittivity in any part of spectrum. Reference [Khurgin and Sun \(2010\)](#) suggests a way to engineer sodium metal for zero losses in the NIR spectrum: stretching sodium lattice by nearly two times. However, it is not clear if such extreme engineering is practical. Thus, a zero loss-plasmonic material is elusive, and a practical alternative is to minimize the losses to an extent it is tolerable.

The dielectric function described by the Drude model (Eq. (1)) suggests that reducing γ can directly scale down the losses. Many possibilities have been explored in reducing the carrier-damping losses in metals including cooling metals to cryogenic temperatures ([Bouillard et al., 2012](#)) and using monolithic metal crystals ([Wang et al., 2015](#)). However, the reduction in losses was either not adequate or its practical implementation was a challenge (as in cooling).

Another possibility for reducing loss in metals is to reduce ω_p , which can reduce the magnitudes of both of the real permittivity and the imaginary permittivity. However, ω_p should not be reduced too much to retain a metallic response in the desired wavelength range. This lower limit on ω_p can be deduced from the Drude-model description of $\epsilon'(\omega)$. If we require metallic behavior ($\epsilon'(\omega) < 0$) for frequencies less than the cross-over frequency ($\omega < \omega_c$), then $\omega_p > \sqrt{\epsilon_b(\omega_c^2 + \gamma^2)}$. This also sets the lower limit on the imaginary part of permittivity: $\epsilon''(\omega) > \frac{\epsilon_b \gamma}{\omega_c}$ for $\omega < \omega_c$. Both of these limits are useful in assessing and engineering different materials and their optical properties to produce alternative plasmonic materials.

There are two possibilities in producing alternative plasmonic materials based on the Drude description. One of them is to dope semiconductors heavily and create enough free carriers that the material's optical properties become metallic in the desired wavelength range. The other option is to remove excess free carriers from metals so as to reduce the carrier concentration to the desired value. Both of these techniques have their own benefits from the perspective of material design and are discussed in the following sections. In addition to small losses, an ideal plasmonic material would possess other useful properties such as adjustable real permittivity, thermal and chemical stability, easy to fabricate nanostructures and integrate them into devices, and earth abundance in availability. Certain applications may demand additional properties such as CMOS compatibility and bio-compatibility. In the following sections, promising material candidates are discussed after a brief review of conventional plasmonic materials.

Conventional Plasmonic Materials

Metals are candidates for plasmonic applications because of their high conductivity. Among metals, silver and gold are the two most often used for plasmonic applications due to their relatively low loss in the visible and NIR ranges. In fact, almost all of the significant experimental work on plasmonics has used either silver or gold as the plasmonic material. While metals other than silver and gold have been used in plasmonics, their use is quite limited, as their losses are higher than those of silver and gold. [Blaber et al. \(2010\)](#) review plasmonic properties of many metals and intermetallics, and present a table listing the highest figure-of-merit of metals in the periodic table. The figure-of-merit defined for localized surface plasmon applications (Q_{LSP}) is defined as the negative ratio of the real to the imaginary parts of the dielectric function. It is clear from reference [Blaber et al. \(2010\)](#) that silver and gold perform the best followed by copper, aluminum and alkali metals. Given the high chemical reactivity of alkali metals, silver, gold, copper and aluminum have been the conventional plasmonic materials.

Among all metals, silver has the lowest optical loss (see Drude damping rate in [Table 1](#)) in the visible and NIR ranges. However, in terms of fabrication, silver degrades relatively quickly and the thickness threshold for uniform continuous films is around 12–23 nm, making it difficult to integrate silver into nanophotonic devices. Additionally, silver losses are strongly dependent on the surface roughness. Gold is the next-best material in terms of loss in the visible and NIR ranges. Compared with silver, gold is chemically stable and can form a continuous film even at thicknesses around 1.5–7 nm, however losses are at least three times higher than in silver. Silver and gold films can be fabricated by various physical vapor deposition (PVD) techniques such as electron-beam/thermal evaporation and sputtering. Nanoparticles and metal-coated nanoparticles can be synthesized by colloidal and electrochemical techniques.

Copper has the second-best conductivity among metals (next to silver), and is expected to exhibit promising plasmonic properties. Indeed, ϵ'' of Cu is comparable to that of gold from 600 to 750 nm. Considering the cost of silver and gold, copper would be a good alternative to silver and gold if its chemical stability were better. Copper easily oxidizes and forms Cu_2O and

Table 1 Drude model parameters for metals. ω_{int} is the frequency of onset for interband transitions. Drude parameters tabulated are not valid beyond this frequency

	ϵ_b	ω_p (eV)	γ (eV)	ω_{int} (eV)
Silver (41,62,63)	3.7	9.2	0.02	3.9
Gold (41,63)	6.9	8.9	0.07	2.3
Copper (41,62,63)	6.7	8.7	0.07	2.1
Aluminum (64,65)	0.7	12.7	0.13	1.41

CuO which degrades its optical properties as well. However, copper is CMOS compatible and if oxidation is prevented by protective layers, copper can be a good plasmonic material.

Aluminum has not been an attractive plasmonic material due to the existence of an interband transition around 800 nm (1.5 eV), resulting in large ε'' values in the visible wavelength range (see [Table 1](#)). However, in the UV range, ε' is negative even below 200 nm where ε'' is still relatively low. Thus, aluminum is a better plasmonic material than either gold or silver in the blue and UV range. Similar to copper, aluminum is easily oxidized. Aluminum very rapidly forms a thin self-limiting layer of aluminum oxide (Al_2O_3) under atmospheric conditions, making device fabrication challenging. Despite these challenges, aluminum is an earth-abundant, inexpensive and CMOS compatible metal attracting significant attention in the recent past ([Knight et al., 2014](#)).

Conventional plasmonic materials – gold, silver, aluminum, copper and other – offer a broad set of optical and material properties useful for many nanophotonic applications. However, the emerging nanophotonic devices such as TO metamaterials need even broader set of properties. As outlined in the previous section, plasmonic materials with adjustable real permittivity, smaller imaginary permittivity, chemically stable, and easy to fabricate and integrate into devices are necessary. Thus, finding alternatives to conventional plasmonic materials is necessary.

Finding Alternatives

Designing materials with a given set of properties is not easy, and often researchers employ a variety of strategies to tackle this problem. For example, in the section on ideal plasmonic materials, we noticed how smaller carrier concentration in plasmonic materials can lead to lower imaginary permittivity and smaller magnitude of real permittivity. Two strategies to achieve the goal of lowering or adjusting the carrier concentration would be: doping semiconductors to create just enough free carriers, and removing excess free carriers from conventional plasmonic materials by alloying and compounding. The following subsections elaborate on these two approaches to design plasmonic materials. In addition, a new class of materials – 2D or ultrathin materials – will be discussed as candidates for alternative plasmonic materials.

Semiconductors to Metals

Increasing the carrier concentration in semiconductors enough to cause them to behave like metals is accomplished through doping. Achieving metal-like optical properties ($\varepsilon' < 0$) in the spectrum of interest puts a lower limit on the carrier concentration that must be achieved by doping. The optical response of free carriers as described by the Drude model ([Eq. \(1\)](#)) can be used to estimate this minimum carrier concentration. To obtain metal-like properties ($\varepsilon' < 0$) for $\omega < \omega_C$, the lower limit on the plasma frequency (ω_p) and hence the carrier concentration (n) is given by [Eq. \(4\)](#)

$$\omega_p^2 > \varepsilon_b(\omega_C^2 + \gamma^2), \quad n > \frac{\varepsilon_0 m^*}{e^2} \varepsilon_b(\omega_C^2 + \gamma^2), \quad (4)$$

where ε_0 is the vacuum permittivity, e is the electron charge, and m^* is the effective mass of the carrier. [Table 2](#) shows the estimates of carrier concentration that would be required for many common semiconductors in order to obtain $\varepsilon' = -1$ at the technologically important telecommunication wavelength ($\lambda = 1.55 \mu\text{m}$). In this scenario, ω_C (cross-over frequency) is slightly higher than the telecommunication frequency. For many of the common semiconductors such as silicon, the minimum carrier concentration to obtain metal-like optical properties in the NIR is about 10^{21} cm^{-3} . Smaller ε_b and m^* values would slightly reduce the minimum carrier concentration, but a high carrier density is inevitable for achieving metal-like properties in the NIR. Such high

Table 2 Comparison of different heavily-doped semiconductors as potential alternative plasmonic materials. The table evaluates the carrier concentration required to reduce the real permittivity of semiconductors to $\varepsilon' = -1$ at telecommunication wavelength ($\lambda = 1.55 \mu\text{m}$). The electronic parameters of the semiconductors reported in the literature are used in Drude model to evaluate the optical properties in the NIR.

Material	Background permittivity (ε_b)	Carrier mobility when heavily doped ($\text{cm}^2/\text{V-s}$)	Effective mass (m^*)	Relaxation rate (eV)	Carrier concentration required to achieve $\text{Re}\{\varepsilon\} = -1$ at $\lambda = 1.55 \mu\text{m}$ ($\times 10^{20} \text{ cm}^{-3}$)	$\text{Im}\{\varepsilon\}$ or losses at $\lambda = 1.55 \mu\text{m}$
n-Si	11.70	80	0.270	0.0536	16.0	0.8508
p-Si	11.70	60	0.390	0.0495	23.1	0.7853
n-SiGe	15.10	50	0.24	0.0965	18.2	1.9414
n-GaAs	10.91	1000 ^a	0.068	0.017	3.76	0.2534
p-GaAs	10.91	60	0.44	0.0438	24.4	0.6528
n-InP	9.55	700	0.078	0.0212	3.82	0.2796
n-GaN	5.04	50	0.24	0.0965	6.83	0.7283
p-GaN	5.24	5	1.4	0.1654	42.3	1.290
Al: ZnO	3.80	47.6	0.38	0.064	8.52	0.384
Ga: ZnO	3.80	30.96	0.38	0.0984	8.59	0.5904
ITO	3.80	36	0.38	0.0846	8.56	0.5077

^aMobility corresponding to the free electron concentration of $1 \times 10^{19} \text{ cm}^{-3}$

carrier densities require ultra-high doping densities, which poses major limitations. The challenges in ultra-high doping will be discussed in the latter part of this section.

Another issue to be considered when choosing a semiconductor for creating metal-like behavior is the mobility of the carriers. The Drude relaxation rate (γ) can be related to the mobility (μ) of the charge carrier at a given optical frequency by Eq. (5).

$$\gamma = e/\mu m^* \quad (5)$$

Reducing the damping loss requires that the product of mobility and effective mass must be as large as possible. Mobility degrades significantly with higher doping levels due to increased impurity scattering. Hence, it is important to consider appropriate mobility numbers when assessing various semiconductors. Table 2 shows the approximate values of γ evaluated using Eq. (5) where low-frequency m^* and high-doping mobility (about 10^{19} cm^{-3}) values are used. The damping losses in semiconductors are comparable to those of bulk gold and silver as reported by Johnson and Christy (1972).

Another consideration in choosing semiconductors is their optical bandgap. The optical bandgap corresponds to the onset of interband transitions, which cause additional optical losses. Hence, the optical bandgap needs to be larger than the frequency spectrum of interest. At telecommunication wavelength of $1.55 \mu\text{m}$, the corresponding photon energy is 0.8 eV . Clearly there are many semiconductors with bandgaps greater than 0.8 eV that can be useful for applications in the NIR and longer wavelengths only if they can be doped heavily. From Table 2, we note that the only major bottleneck in turning semiconductors into low-loss plasmonic elements at optical frequencies is accomplishing the required ultra-high doping. A deeper understanding of the doping mechanism in semiconductors can provide insights for accomplishing this ultra-high doping, and hence we turn our attention to this process next.

Doping is a process of incorporating foreign atoms or impurities into the lattice of a semiconductor to controllably change the properties of the semiconductor. When certain atomic species are incorporated into the lattice sites of the semiconductor, the free-carrier concentration in the material can be changed proportionally. For our purposes, we will only consider doping that acts to increase the free-carrier concentration.

When dopant atoms replace (substitute) the semiconductor atoms in the lattice, this form of doping is called substitutional doping, which effectively contributes to an increase in the charge-carrier density in the material. When the dopant atom occupies an interstitial site, it is called interstitial doping. Interstitial doping is an ineffective doping method because it does not contribute any free carriers and, therefore, does not produce any electrical doping. There is another mechanism called doping compensation that can also result in ineffective doping. In this mechanism, dopants such as silicon in GaAs can behave as both an n-type and p-type dopant and self-compensate any net doping effect.

Doping is limited on the higher side by the solid solubility of the dopant in the semiconductor. Introducing dopants more than the solid-solubility limit may result in phase separation of dopants or compounds of dopants. This phase separation of excess dopants leads to ineffective doping. Often, ultra-high doping results in more crystal defects, which could act as traps for free carriers. These trap states can counter the effect of doping and thus reduce the carrier concentration. In general, the doping mechanism is complicated and, most often, not all dopant atoms succeed in contributing free charge carriers due to a number of reasons, including those mentioned above.

Doping efficiency is an important quantity and can be defined as the fraction of dopants that contribute to the charge carrier density. Doping efficiency decreases sharply for high doping concentrations in many semiconductors, making ultra-high doping a tough challenge. Although there are many different material engineering approaches that can provide elegant solution to this problem (e.g., delta doping (Weir *et al.*, 1994)), a straightforward approach to the problem of achieving ultra-high doping would be to choose a dopant that has very high solid-solubility limit in the selected semiconductor material. This approach has been utilized in demonstrating many semiconductor-based plasmonic materials. Heavily doped silicon, III-V systems based on GaAs, and transparent conducting oxides (TCOs) such as indium-tin-oxide (ITO), and aluminum zinc oxide (AZO) are widely explored as alternative plasmonic materials.

III-V semiconductors

Compounds of III-V semiconductors, such as those based on GaAs and InP, are semiconductors with an optical bandgap in the NIR. This is useful to know because a bandgap comparable to that of silicon results in a value of ϵ_b that is comparable to that of silicon for these materials. In addition, the electron mobility is very high in these materials due to a small effective mass. This relaxes the carrier concentration requirement for observing metallic properties in the NIR (see Table 1). In these materials, a carrier concentration in excess of 10^{20} cm^{-3} is required to observe plasmonic properties at the telecommunication wavelength. However, such high doping is very challenging in these materials owing to lower solid solubilities of the dopants and poor doping efficiency (Walukiewicz, 2001). Doping higher than 10^{19} cm^{-3} is known to produce effects such as doping compensation. For instance, silicon is a common n-type dopant in GaAs, but at high doping levels, silicon can behave not only as an n-type substitutional dopant, but also as a p-type dopant. Thus, silicon can compensate itself at high doping densities, resulting in low doping efficiency. N-type doping beyond 10^{19} cm^{-3} is clearly difficult in GaAs films. Many other studies report similar trends, and it was pointed out that the carrier concentration in GaAs shows a saturating trend for donor densities higher than $5 \times 10^{18} \text{ cm}^{-3}$. On the other hand, p-doping with Be or C can be used to achieve carrier concentrations in excess of 10^{20} cm^{-3} (Yamada *et al.*, 1989). However, holes have a higher effective mass and poor carrier mobility, which raises the minimum bar on carrier concentration to turn p-GaAs plasmonic at the telecommunication frequency (see Table 1). Other III-V semiconductor families such as InGaAs, AlGaAs, GaInSb, InP, and GaInAsP have properties similar to GaAs and doping them high is challenging. However, they all have high electron mobility and make low loss plasmonic materials in the infrared.

Transparent conducting oxides

Oxide semiconductors such as zinc oxide, cadmium oxide and indium oxide can be highly doped to make them conducting films (Minami, 2005). Since these semiconductors have a large bandgap, they are transparent in the visible range. Hence, these materials are known as transparent conducting oxides (TCOs). Because TCOs can be doped very heavily, TCOs exhibit high DC conductivity. It is exactly this property that gives them metal-like optical properties in the NIR range. Like any other semiconductor, the optical properties of TCOs can be tuned by changing the carrier concentration/doping. They can be grown into thin films and many different nanostructures, polycrystalline and crystalline structures, patterned by standard fabrication procedures and integrated with many other standard technologies. Thus, TCOs form an obvious choice as alternative plasmonic materials in the NIR. Among the many TCOs, previous studies have shown that heavily-doped ZnO and ITO are good candidates for NIR applications (Naik and Boltasseva, 2010). Hence, in the subsequent discussion on TCOs, we focus on doped ZnO and ITO.

Thin films of TCO materials may be grown by many different techniques such as sputtering, laser ablation, evaporation, solution processing and chemical vapor deposition. Since TCOs are non-stoichiometric oxides, their properties depend on the deposition technique used. Deposition schemes such as laser ablation and sputtering are more suitable in cases where the stoichiometry needs to be controlled to produce desired film properties. In an example, Naik *et al.* deposit aluminum-doped ZnO (AZO), gallium-doped ZnO (GZO) and ITO thin films using the pulsed laser deposition (PLD) technique. Fig. 2 shows the carrier concentration in these TCO films as a function of doping concentration. The carrier concentration was extracted by fitting the Drude model to the measured data. Optical characterization was carried out using a spectroscopic ellipsometer (J.A. Woollam Co.). The films were deposited under an oxygen partial pressure of 1 mTorr. Their results indicate that AZO shows lower losses with higher doping because of improved crystallinity in the highly-doped films. However, the highest carrier concentration achieved is smaller in AZO than in ITO or GZO. The losses are higher with GZO and ITO, but much higher carrier concentrations are possible in these films. When the oxygen partial pressure was decreased below 1 mTorr, the resultant AZO films exhibited higher carrier concentrations and performed the best compared to the other two materials. Fig. 3 shows the optical properties of TCO films extracted from ellipsometry measurements. The retrieval of the dielectric functions of TCOs was based on Drude-Lorentz model. Drude model was added to account for the free carriers and the Lorentz oscillator was added in UV to account for the interband transitions at the band edge. Table 3 shows the parameters of Drude-Lorentz model for AZO, GZO and ITO retrieved from ellipsometry measurements. Clearly, AZO is the TCO with the lowest loss, followed by ITO and then GZO.

Other than semiconductors and TCOs, there are non-stoichiometric compounds such as In_2O_3 , TiO_2 , IrO_2 and metallic VO_2 that are intrinsically doped. All of these materials can support large carrier concentrations ($> 10^{19} \text{ cm}^{-3}$) to provide Drude metal-like behavior, though their losses are typically higher. However, there are niche applications in which these materials can be very useful as plasmonic components. For example, perovskite oxides show magnetoresistance properties useful for data storage applications (Rao, 1998). Vanadium dioxide shows a metal-semiconductor transition at temperatures smaller than 100°C , which can be useful in switching applications (Dicken *et al.*, 2009).

Metal Alloys

Metallic alloys, intermetallics and metallic compounds are potential candidates for alternative plasmonic materials owing to their large free electron densities. Because of the strong plasmonic performance of noble metals, one approach to improve these materials is by shifting their interband transitions to another (unimportant) part of the spectrum. This can be achieved by

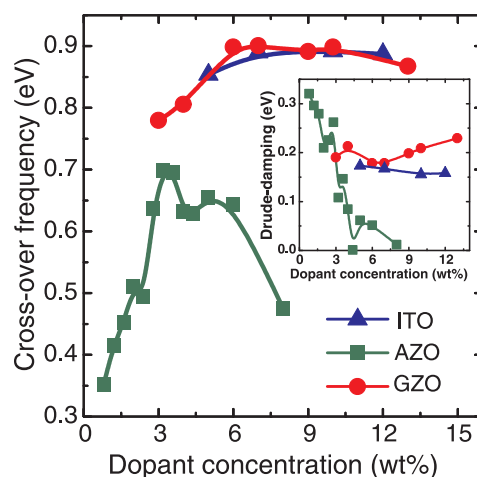


Fig. 2 Cross-over frequency (frequency at which $\text{Re}(\epsilon)=0$) or screened plasma frequency) as a function of dopant concentration in three TCOs: indium-tin-oxide (ITO), Ga:ZnO (GZO) and Al:ZnO (AZO) (adapted from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.) The inset shows the Drude damping coefficient as a function of dopant concentration. All the films were deposited using pulsed laser deposition and the optical parameters are extracted using a spectroscopic ellipsometer.

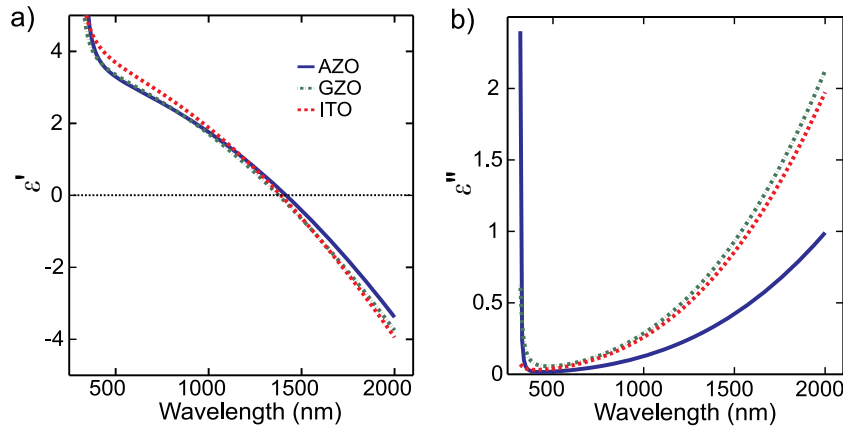


Fig. 3 Optical constants of TCO thin films deposited using pulsed laser deposition. The optical constants are measured using spectroscopic ellipsometry. (a) Real (b) Imaginary parts of dielectric function of TCO films: Al:ZnO (2 wt%), Ga:ZnO (4 wt%) and ITO (10 wt%). The deposition conditions were optimized to produce lowest losses and highest plasma frequency. Figure reproduced with permission from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. *Opt. Mater. Express* 1, 1090.

Table 3 Drude-Lorentz parameters of five alternative plasmonic materials retrieved from ellipsometry measurements. The dielectric function of the materials may be approximated in the wavelength range of 350–2000 nm by the equation: $\varepsilon_m(\omega) = \varepsilon_b - \frac{\omega_p^2}{\omega(\omega + i\gamma)}$ + $\frac{f_1 \omega_1^2}{\omega_1^2 - \omega^2 - i\omega\Gamma_1}$, where the values of the parameters are as listed in the table

	AZO (2 wt%)	GZO (4 wt%)	ITO (10 wt%)	TiN	ZrN
ε_b	3.5402	3.2257	3.528	4.855	3.4656
ω_p (eV)	1.7473	1.9895	1.78	7.06	8.018
γ (eV)	0.04486	0.1229	0.155	0.291	0.5192
f_1	0.5095	0.3859	0.3884	2.125	2.4509
ω_1 (eV)	4.2942	4.050	4.210	3.862	5.48
Γ_1 (eV)	0.1017	0.0924	0.0919	1.45	1.7369

alloying/reacting two or more elements to create unique band structures that can be fine-tuned by adjusting the proportion of each alloyed material/reactant.

Alloys of noble metals have received significant attention given that noble metals are the conventional plasmonic materials. Studies on alloys of conventional plasmonic materials such as Ag–Cu, Au–Ag, Ag–Al, Au–Cd, Ag–In, and Ag–Mg show that the optical properties generally become worse than the best of the metals being alloyed. However, alloying helps in shifting the interband transitions where a simple mixing rule has been observed to be a good approximation. Similar studies and observations have been made on various intermetallics as summarized in reference [Blaber et al. \(2010\)](#). It has been concluded that intermetallics with more atoms in a unit cell usually suffer from higher optical losses. Thus, gold and silver continue to dominate as metallic plasmonic materials.

Metals to “Less-Metals”

Another approach to decrease the excessive optical losses in metals is to reduce the carrier concentration in the material. A smaller carrier concentration may be achieved in intermetallic or metallic compounds, which we will term “dilute metals” or “less-metals” in this work. In general, introducing non-metallic elements into a metal lattice reduces the carrier concentration. This also alters the electronic band structure significantly and might result in undesirable consequences, such as large interband absorption losses and higher Drude-damping losses. However, there is a lot of room for optimization, and this direction of research has yet to receive much attention from researchers. There are many advantages of plasmonic intermetallic or metallic compounds such as tunability and ease of fabrication and integration, which could out-weigh the disadvantages of these materials. Thus, it is worth pursuing a search for materials that are dilute metals.

A survey of the literature shows that the optical properties of many different classes of dilute-metals have been studied previously. Among them are metal silicides and germanides, ceramics such as oxides, carbides, borides and nitrides, which were all found to exhibit negative real permittivities in different parts of the optical spectrum. None of these materials possess all the desirable properties, and in particular none of them exhibit low optical losses. However, there has been a growing interest in materials that are important from the view point of technological applications. Specifically, silicides, and metal nitrides which are already used in silicon CMOS technology have received significant attention.

The dielectric functions of these silicides show metallic behavior in the mid-infrared (MIR), NIR and visible ranges. The losses are quite high in these materials due to interband transitions, especially in the NIR where their real permittivities are small in magnitude. Comparing to noble metals, silicides exhibit very high losses. While they are useful at longer wavelengths ($\lambda > 3 \mu\text{m}$) where there are no interband transitions, their real permittivity values are large in magnitude, similar to noble metals. Nevertheless, the manufacturing advantages of silicides make them promising materials for infrared plasmonics. Recently, Soref *et al.* proposed using PdSi_2 as plasmonic material for MIR plasmonics (Soref *et al.*, 2008) and Cleary *et al.* demonstrated an application of silicides for an SPP-based infrared biosensor where PdSi_2 was used as a plasmonic material (Cleary *et al.*, 2008). Blaber *et al.* (2010) provided a brief review of silicides for plasmonic applications which shows that the calculated quality factors for LSPR applications can be high for some silicides such as TiSi_2 .

Metal nitrides such as titanium nitride, zirconium nitride, tantalum nitride and hafnium nitride exhibit metallic properties in the visible and longer wavelengths (Naik *et al.*, 2011). These are non-stoichiometric, interstitial compounds with large free-carrier concentrations. These materials are also refractory, stable and hard and allow for tuning of their optical properties by varying their composition. Technologically, they are important because many of them are currently used in silicon CMOS technology as barrier layers in copper-Damascene processes and as gate metals to n-type and p-type transistors. Thus, metal nitrides offer fabrication and integration advantages which could be useful in integrating plasmonics with nanoelectronics.

Refractory metal nitrides can be deposited by many techniques such as chemical vapor deposition (CVD), atomic layer deposition (ALD), physical vapor deposition (PVD) such as reactive sputtering pulsed laser deposition and ion-assisted reactive evaporation, and wet-chemical techniques. Since metal nitrides have nearly three orders of magnitude smaller surface energy, they can be grown into ultra-thin ($< 10 \text{ nm}$) and epitaxial or single crystal films. Many of these nitrides are grown epitaxially on various substrates such as c-sapphire, MgO and Si. Epitaxial growth leading to ultra-smooth, ultra-thin films is very important for plasmonic applications, and noble metals usually fail in this aspect.

Optical properties of metal nitrides are receiving significant attention only recently. Naik *et al.* recently studied the optical properties of titanium nitride, tantalum nitride, hafnium nitride and zirconium nitride (Naik *et al.*, 2011). Fig. 4 shows the dielectric functions of thin films of these metal nitrides deposited by DC reactive sputtering. The dielectric functions were retrieved from the spectroscopic ellipsometer measurements on a thin film sample. The Drude-Lorentz model parameters of these metal nitrides extracted from ellipsometry measurements are listed in Table 3. Note that the parameters listed for TiN in the table are for films with improved optical properties. Small improvements in the deposition conditions such as chamber seasoning allowed better quality TiN films. For all nitride films excepting titanium nitride, the deposition was carried out in nitrogen-deficient ambient, which resulted in metal-rich films. When the composition of sputtering ambient was varied from 100% nitrogen to 20% nitrogen and 80% argon, all the metal nitrides except titanium nitride exhibited optical properties that changed from dielectric behavior to metallic behavior. Titanium nitride shows a lower carrier concentration under nitrogen-rich deposition conditions. However, the metallic property in TiN is not lost because the change in the carrier concentration is small. It is also important to note that most of these nitrides have larger carrier relaxation rates than those of the noble metals, which increases the optical losses in these nitrides. Additionally, interstitial nitrides exhibit interband transition losses in the visible and ultraviolet ranges. Thus, their optical properties are undesirable in these ranges unless significant growth-parameter optimization is employed. The deposition conditions can change the Drude relaxation rate, the plasma frequency and the strength of the interband transitions. While nitrogen-rich films showed a monotonic decrease in the plasma frequency with increasing nitrogen content, the dependence was not monotonic for the other parameters. Thus, careful optimization of the deposition conditions is necessary to achieve a low-loss TiN film.

Surface plasmon-polaritons are shown to be supported by titanium nitride films. Steinmüller-Nethl *et al.* demonstrated SPP excitation on an approximately 30-nm-thick TiN film deposited by sputtering (Steinmüller-Nethl *et al.*, 1994). In that study, SPP

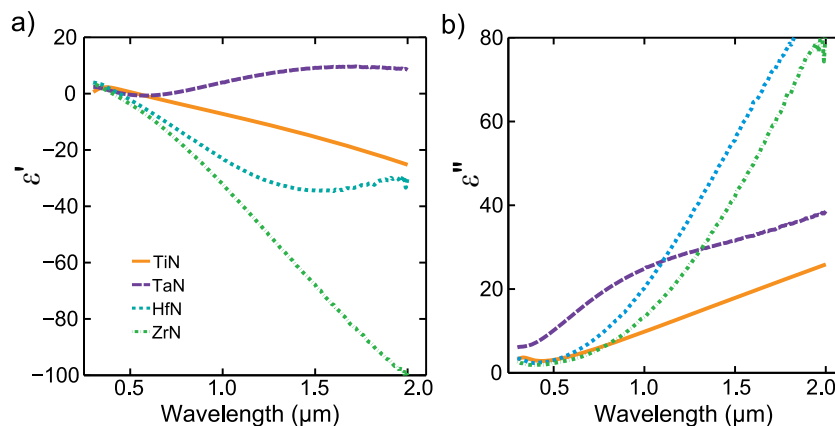


Fig. 4 Optical constants of metal nitride thin films deposited using DC reactive sputtering. The (a) real and (b) imaginary parts of dielectric functions extracted from spectroscopic ellipsometry measurements. Figure reproduced with permission from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.

excitation was demonstrated at wavelengths of 700, 750, 800 and 850 nm. Hibbins *et al.* extended this work to many wavelengths in the range from 500 to 900 nm (Hibbins *et al.*, 1998). Recently, Chen *et al.* demonstrated SPP excitation on TiN films in the visible wavelengths using the Kretschmann geometry (Chen *et al.*, 2011). Naik *et al.* also reported SPP coupling on epitaxial TiN thin films deposited by DC reactive sputtering on c-sapphire substrates (Naik *et al.*, 2012b). The SPP modes were coupled onto the TiN film using a dielectric grating coupler formed by patterning a thin layer of polymer (electron-beam resist, ZEP-520A) on top of the TiN film. The quality factor of surface plasmon polaritons on TiN thin films were predicted to be comparable to that on gold films. Given that TiN is a much harder material than gold, stable at temperatures in excess of 1200 °C, and possess low surface energy allowing it to form epitaxial ultra-thin films, TiN has systematically started to replace gold in plasmonic devices.

Though 'less-metals' such as metal nitrides exhibit plasmonic behavior in the optical range, their optical loss is no smaller than that of silver. Nevertheless, materials such as TiN have optical properties similar to gold, the second best noble metal in terms of losses. Given the other advantages of metal nitrides of noble metals, metal nitrides hold great promise as alternative plasmonic materials in the visible and NIR ranges.

Two-Dimensional Plasmonic Materials

Recently, plasmonics with 2D materials, or flatland plasmonics has gained much attention (Maier, 2012). Two-dimensional materials such as graphene have many advantages over bulk three dimensional (3D) materials from both scientific and technological perspectives. The optical properties of 2D materials are quite different from those of bulk, 3D materials, which results in significantly different plasmon dispersion relationships. Two-dimensional materials are technologically important because their properties are useful in many other applications, including electronics. The properties of 2D materials can be dynamically tuned by electrical, chemical, electrochemical and other means. For example, the optical properties of graphene can be tuned by electrical field-effect methods (Novoselov *et al.*, 2012). In addition, 2D materials can be processed via conventional, planar fabrication techniques. Currently, the synthesis of large-area single, crystalline 2D materials is a challenge that limits the set of suitable applications for these materials. However, there is now significant effort invested in overcoming this bottleneck.

There are reports of the observation of plasmons in 2D electron gas (2DEG) systems such as semiconductor inversion layers, semiconductor heterostructures, and graphene. Except in the case of graphene, plasmons were observed in other cases only at low temperatures because of the high losses (low carrier mobility or high carrier scattering rates) occurring at room temperature. All of the observations of plasmons were made in the MIR or longer wavelength ranges. Plasmons in the visible or NIR ranges were not observed in any of these 2D materials because of insufficient carrier densities.

Applications

The search for alternative plasmonic materials has already led to the discovery of many new materials for plasmonics. Yet, the exploration continues because there is not one material that works for all applications. Depending on the dominant physical phenomenon underlying the application, the performance metrics differ and so do the required optical properties of the plasmonic material. Hence, different plasmonic materials work the best for different applications. Based on the applications, plasmonic devices can be broadly classified into integrated plasmonics, sub-diffraction imaging, chemical sensing, high-temperature plasmonics, and metamaterials & transformation optics. Plasmonic materials for each of these applications are reviewed in the following.

Integrated Plasmonics

Integrated plasmonics is an analogue of integrated electronics aimed at information processing applications. Plasmonics can confine light to nanoscale and offer huge bandwidth unlike electronics, making plasmonics a promising technology to substitute/complement nanoelectronics. Some of the on-chip plasmonic devices are interconnects or waveguides, resonators, de-mux/multiplexers, and modulators or switches. Except for switches, all other components are passive devices and rely upon propagation of SPPs.

SPP propagation at metal/dielectric interface is characterized by two important metrics: mode confinement and propagation length. A small mode size and long propagation length are desirable, though these characteristics trade-off with each other. In general, for SPP waveguiding applications, the real and imaginary parts of the metal permittivity influence the trade-off between confinement and propagation loss significantly. A larger negative real permittivity gives rise to smaller field penetration into the metal and larger mode size, and a smaller imaginary permittivity leads to lower losses. Thus, the figure of merit would take the form $[\text{Re}(\epsilon_m)]^2/\text{Im}(\epsilon_m)$, where ϵ_m is the permittivity of the plasmonic material. This quantity is plotted for different alternative plasmonic materials in Fig. 5. Noble metals outperform all other materials because of their large negative real permittivity. The large negative real permittivity is a consequence of large plasma frequency or large carrier concentration. Since the figure-of-merit, $[\text{Re}(\epsilon_m)]^2/\text{Im}(\epsilon_m)$ exhibits stronger dependence on real permittivity than imaginary permittivity, having a larger plasma frequency results in a more significant improvement in performance than reducing the losses. Because all the alternative plasmonic materials we considered have lower plasma frequencies than noble metals, their performance metrics as integrated plasmonic elements are worse than those offered by noble metals.

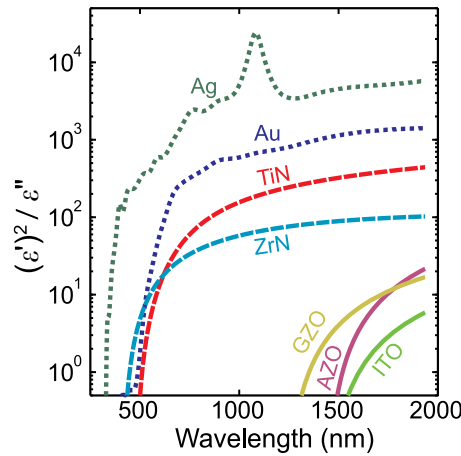


Fig. 5 The performance of many 2D plasmonic waveguides, non-spherical LSPR structures, resonant metamaterial devices such as negative-index-metamaterials depend on the quantity $\frac{Re(\epsilon_m)^2}{Im(\epsilon_m)}$ (ϵ_m is the permittivity of plasmonic material). This quantity is plotted in the figure for various plasmonic materials.

However, alternative plasmonic materials offer advantages like CMOS compatibility which could be a bigger factor than a slightly lower figure-of-merit. Aluminum, a CMOS compatible metal has been used in plasmonic systems for SPP waveguiding (MacDonald *et al.*, 2008), color filters and hot carrier photo-detectors (Zheng *et al.*, 2014). Copper waveguides are also shown as CMOS-compatible solutions to SPP waveguiding (Fedyanin *et al.*, 2016). Titanium nitride, another CMOS compatible material is also explored for integrated plasmonic applications. Kinsey *et al.* demonstrated TiN interconnects whose performance was comparable to SPP waveguides made from gold (Kinsey *et al.*, 2014). Though alternative plasmonic materials have not been able to perform as well as or better than silver in integrated plasmonic applications, they offer other advantages which makes them promising.

Optical modulator being the key element in an information processing application, plasmonic modulators offer superior performance in terms of modulation contrast. Semiconductor based plasmonic materials such as ITO and 2D material – graphene has been shown to work well as electro-optic materials. Often, these switchable materials together with other plasmonic metals have been incorporated into modulators. While silicon photonic modulators have extinction ratio in the order of 1 dB/mm, plasmonic modulators with extinction ratio greater than 3 dB/ μ m have been realized (Liu *et al.*, 2015). When a plasmonic modulator is incorporated into silicon photonics, high insertion loss has been observed due to the mismatch in plasmonic and photonic modes. However, in a completely plasmonic technology, insertion loss would not be a problem. Thus, alternative plasmonic materials are promising for nanoscale light switches.

Chemical Sensing

Chemical sensing applications rely upon strong field enhancement near plasmonic resonators. Metallic nanostructures exhibiting localized surface plasmon resonance (LSPR) concentrate light to nanoscale enhancing the local electric field by many times. Using a dipole approximation in the electrostatic limit, the near-field enhancement, absorption cross-section and extinction cross-section of spherical plasmonic nanoparticles can be calculated for conventional and alternative plasmonic materials. Fig. 6 plots the maximum field enhancement ($|E_{max}/E_0|$) on the surface of spherical nanoparticles of gold, silver, titanium nitride, zirconium nitride and TCOs of diameter $\lambda/10$. The refractive index of the host medium surrounding the nanoparticles is assumed to be 1.33. Clearly, silver outperforms all other materials with resonances in the blue part of visible spectrum. Other materials show comparable performance in different parts of the NIR and visible ranges. Notably, TCOs could enable LSPR applications in the NIR without the need for complex geometries such as core-shell structures. Titanium nitride and zirconium nitride have similar performance and exhibit LSPR modes in the visible part of the spectrum. In particular, titanium nitride could be of significant interest for biological applications because its LSPR mode lies in the biological transparency window. Also, TiN may be useful in plasmonic heating applications because it is refractory.

When the particle geometry is more complex than a sphere, the LSPR wavelength and its strength change. In general, small spherical particles exhibit a resonance strength that is proportional to $Re\{-\epsilon_m\}/Im\{\epsilon_m\}$, where ϵ_m is the permittivity of the plasmonic material. When the geometry changes from a sphere to cigar-like elongated structures, for example, the resonance quality varies as $[Re\{\epsilon_m\}]^2/Im\{\epsilon_m\}$. Alternative plasmonic materials have smaller magnitudes of real permittivity due to their smaller plasma frequencies, and hence they perform poorly when compared to noble metals for such complicated geometries.

High Temperature Plasmonics

Application such as heat-assisted magnetic recording (HAMR), photothermal cancer treatment, and thermophotovoltaics (TPV) require plasmonic devices operating at elevated temperatures. Refractory metals are more preferred than low-melting noble metals

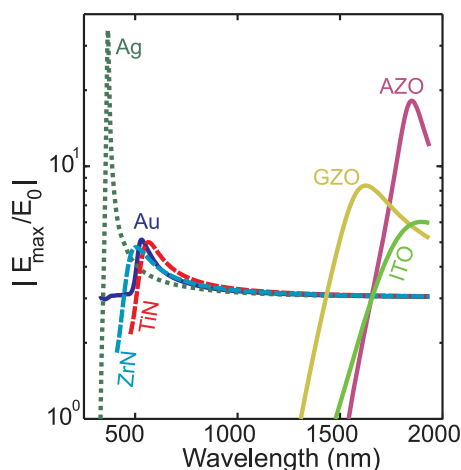


Fig. 6 Calculated maximum field enhancement on the surface of a spherical nanoparticle of the plasmonic material embedded in a dielectric host of refractive index 1.33. The calculations use quasistatic approximation. Reproduced from Naik, G.V., Shalaev, V.M., Boltasseva, A., 2013. *Adv. Mater.* 25 (24), 3264–3294. doi: 10.1002/adma.201205076.

in such applications. Though noble metals have high melting points ($\sim 1000^\circ\text{C}$), their nanostructures with high ratio of surface area to volume deform at relatively low temperatures. One of the primary factors contributing to this deformation is high surface energy of these metals. On the other hand, refractory metals overcome such limitations and hence are promising for high temperature plasmonic applications.

Since high temperature plasmonic devices also rely upon LSPR phenomenon, the same trend as observed in Fig. 6 holds. Given the problems of noble metals at elevated temperatures, alternative materials such as TiN and Zn are very promising. Li *et al.* demonstrated the contrasting high-temperature stability between gold and TiN nanostructures (Li *et al.*, 2014). While 800°C annealing destroyed the shape and optical properties of gold nanostructures, similar TiN nanostructures survived the annealing cycle and preserved their optical properties intact.

Recently, Guler *et al.* demonstrated LSPR in lithographically fabricated TiN nanodisks and showed better photothermal response than in similar gold nanostructures (Guler *et al.*, 2013). As an additional advantage, TiN nanostructures showed resonances in the biological transparency window of 800–1200 nm wavelength. Given the bio-compatibility of TiN, it is a promising material for photothermal cancer therapy. Ishii *et al.* demonstrated better solar steam generation in water suspended TiN nanoparticles than in similar suspensions of carbon-black or gold nanoparticles (Ishii *et al.*, 2016).

Recently, high temperature optical constants of gold and TiN are measured at high temperatures, and it is clear that TiN is stable at temperatures greater than 1200°C and undergoes smaller degradation in its optical losses than gold (Briggs *et al.*, 2017). While gold doubles its Drude damping rate at 600°C , TiN doubles at 1200°C . Clearly, titanium nitride is a promising plasmonic material for high temperature applications and other refractory metal nitrides are yet to be explored.

Metamaterials and Transformation Optics

Metamaterials may be classified into two categories: LSPR-based and non-LSPR metamaterials. LSPR-based metamaterials such as negative index metamaterials rely upon LSPR phenomenon and hence the best performing plasmonic materials are noble metals followed by metal nitrides, further followed by TCOs (Tassin *et al.*, 2012). Non-LSPR metamaterials rely upon effective medium behavior of composite materials and examples are cloaks, hyperbolic metamaterials, and epsilon-near-zero (ENZ) materials. As a general note, most of the metamaterial implementations of transformation optics fall into the category of non-LSPR metamaterials. Such metamaterials benefit a lot from alternative plasmonic materials because of their smaller $\text{Im}\{\epsilon_m\}$.

Hoffman *et al.* showed negative refraction at an operating wavelength of $8\ \mu\text{m}$ in a so-called hyperbolic metamaterial consisting of planar, alternating layers (a superlattice) of heavily doped ($1\text{--}4 \times 10^{18}\ \text{cm}^{-3}$) $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ and undoped $\text{Al}_{0.48}\text{In}_{0.52}\text{As}$ deposited by MBE (Hoffman *et al.*, 2007). This metamaterial device was shown to have a performance figure-of-merit of about 20, the highest reported so far. Yet another demonstration is the epsilon-near-zero (ENZ) properties of the InAsSb material when doped heavily. Adams *et al.* show that heavy doping ($1\text{--}2 \times 10^{19}\ \text{cm}^{-3}$) of InAsSb causes the real permittivity to cross zero and turn metal-like in the MIR range (Adams *et al.*, 2011). At the zero cross-over of real permittivity (occurring at about $8\ \mu\text{m}$ wavelength), the material behaves as an ENZ material and exhibits special properties such as photon funneling.

In the near-IR, Kim *et al.* showed ENZ property of TCO substrates pin the resonance of gold nanorods always to ENZ wavelength irrespective of their geometry (Kim *et al.*, 2016). Naik *et al.* demonstrated highest performing hyperbolic metamaterial in the near-infrared using Al-doped ZnO and undoped ZnO alternating layers (Naik *et al.*, 2012a). The figure-of-merit of this hyperbolic metamaterial was nearly 10 while similar noble-metal based structures have 3 orders of magnitude smaller

figures-of-merit. The implication of this demonstration is that TCO-based metamaterials are best suited for applications such as hyperlens – a sub-diffraction imaging metamaterial device.

In the visible spectral range, TiN based metamaterials have been demonstrated. In an example of a metamaterial for quantum optical applications, Naik *et al.* developed an epitaxial metal/dielectric superlattice based on TiN and AlN, and showed that the structure exhibits hyperbolic dispersion (Naik *et al.*, 2014). Layer thickness down to 5 nm – not possible with noble metals – was possible in TiN based metamaterial only because the surface energy of TiN is nearly 3 orders of magnitude smaller than that of noble metals. Such a metamaterial showed a huge enhancement in local photonic density of states, and a fluorophore probe near such metamaterial experienced about 80x enhancement in its radiative decay rate. Such huge enhancement in photon emission rate is useful in many light emitting devices including a single photon source for quantum optical applications. Shalaginov *et al.* extended this work to demonstrate fast single photon emission from NV centers of nano-diamonds (Shalaginov *et al.*, 2013).

Overall, alternative plasmonic materials are very promising for non-LSPR metamaterials owing to their smaller magnitude of real permittivity and small imaginary permittivity.

Summary and Outlook

With the rapid development of nanophotonics, it is clear that there will not be a single plasmonic material that is suitable for all applications at all frequencies. Rather, a variety of material combinations must be fine-tuned and optimized for individual situations or applications. While several approaches and materials have been discussed, many more have been and are being proposed. The problem of finding better plasmonic materials remains open-ended. An improved plasmonic material holds the key in transforming nanophotonics from a laboratory science to a ubiquitous technology for the coming generations.

References

- Adams, D.C., Inampudi, S., Ribaud, T., *et al.*, 2011. Phys. Rev. Lett. 107, 1090.
 Blaber, M.G., Arnold, M.D., Ford, M.J., 2010. J. Phys. Condens. Matter 22, 143201.
 Bouillard, J.S.G., Dickson, W., O'Connor, D.P., Wurtz, G.A., Zayats, A., 2012. Nano Lett. 12, 1561.
 Briggs, J.A., Naik, G.V., Zhao, Y., *et al.*, 2017. Appl. Phys. Lett. 1.(in press).
 Cai, W., Shalaev, V., 2009. Optical Metamaterials: Fundamentals and Applications. Springer Verlag.
 Chen, N.C., Lien, W.C., Liu, C.R., *et al.*, 2011. J. Appl. Phys. 109, 43104.
 Chen, W., Thoreson, M.D., Ishii, S., Kildishev, A.V., Shalaev, V.M., 2010. Opt. Express 18, 5124.
 Cleary, J., Peale, R., Shelton, D., *et al.*, 2008. in (Cambridge Univ Press).
 Dicken, M.J., Aydin, K., Pryce, I.M., *et al.*, 2009. Opt. Express 17, 18330.
 Dressel, M., Grüner, G., Bertsch, G.F., 2002. Am. J. Phys. 70, 1269.
 Fedyanin, D.Y., Yakubovsky, D.I., Kirtaev, R.V., Volkov, V.S., 2016. Nano Lett. 16, 362.
 Guler, U., Ndukaife, J.C., Naik, G.V., *et al.*, 2013. Cleo 2013 QTu1A 2.
 Hibbins, A.P., Sambles, J.R., Lawrence, C.R., 1998. J. Mod. Opt. 45, 2051.
 Hoffman, A.J., Alekseyev, L., Howard, S.S., *et al.*, 2007. Nat. Mater. 6, 946.
 Ishii, S., Sugavaneshwar, R.P., Nagao, T., 2016. J. Phys. Chem. C 120, 2343.
 Johnson, P.B., Christy, R.W., 1972. Phys. Rev. B 6, 4370.
 Khurgin, J.B., Sun, G., 2010. Appl. Phys. Lett. 96, 181102.
 Kim, J., Dutta, A., Naik, G.V., *et al.*, 2016. Optica 3, 4.
 Kinsey, N., Ferrera, M., Naik, G.V., *et al.*, 2014. Opt. Express 22, 12238.
 Knight, M.W., King, N.S., Liu, L., *et al.*, 2014. ACS Nano 8, 834.
 Kreibig, U., Vollmer, M., 1995. Springer Ser. Mater. Sci. 535.
 Liu, K., Ye, C.R., Khan, S., Sorger, V.J., 2015. Laser Photonics Rev. 9, 172.
 Li, W., Guler, U., Kinsey, N., *et al.*, 2014. Adv. Mater. 26, 7959.
 MacDonald, K.F., Sármson, Z.L., Stockman, M.I., Zheludev, N.I., 2008. Nat. Photonics 3, 55.
 Maier, S.A., 2012. Nat. Phys. 8, 581.
 Minami, T., 2005. Semicond. Sci. Technol. 20, S35.
 Naik, G.V., Boltasseva, A., 2010. Phys. Status Solidi – Rapid Res. Lett. 4, 295.
 Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.
 Naik, G.V., Liu, J., Kildishev, A.V., Shalaev, V.M., 2012a. Proc. Natl. Acad. Sci. USA 109, 8834.
 Naik, G.V., Saha, B., Liu, J., *et al.*, 2014. Proc. Natl. Acad. Sci. USA 111, 7546.
 Naik, G.V., Schroeder, J.L., Ni, X., *et al.*, 2012b. Opt. Mater. Express 2, 478.
 Novoselov, K.S., Fal, V.I., Colombo, L., *et al.*, 2012. Nature 490, 192.
 Pashley, D.W., 1959. Philos. Mag. 4, 316.
 Rao, C.N.R., 1998. Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci. 356, 23.
 Shalaginov, M.Y., Ishii, S., Liu, J., *et al.*, 2013. Appl. Phys. Lett. 102, 2011.
 Soref, R., Peale, R.E., Buchwald, W., 2008. Opt. Express 16, 6507.
 Steinmüller-Nethl, D., Kovacs, R., Gornik, E., Röthhammer, P., 1994. Thin Solid Films 237, 277.
 Tassin, P., Koschny, T., Kafesaki, M., Soukoulis, C.M., 2012. Nat. Photonics 6, 259.
 Walukiewicz, W., 2001. Phys. B. Condens. Matter 302–303, 123.
 Wang, C.-Y., Chen, H.-Y., Sun, L., *et al.*, 2015. Nat. Commun. 6, 7734.
 Weir, B.E., Feldman, L.C., Monroe, D., *et al.*, 1994. Appl. Phys. Lett. 65, 737.
 Yamada, T., Tokumitsu, E., Saito, K., *et al.*, 1989. J. Cryst. Growth 95, 145.
 Zheng, B.Y., Wang, Y., Nordlander, P., Halas, N.J., 2014. Adv. Mater. 6318.

Raman Lasers

Marco Santagiustina, University of Padova, Padova, Italy

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Since their first appearance, Raman lasers have represented a simple and reliable approach to obtain laser sources at wavelengths not available through stimulated emission. Several experimental setups were developed (single pass, Fabry–Pérot cavities, ring cavities, intra- and extra cavity pumping etc.) each one with its own advantages, and disadvantages. As well, there were several materials that could be exploited, with a particular emphasis on solid-state lasers, which were mainly based on special crystals and optical fibers.

The main advantages of solid-state devices, with respect to those exploiting gases or liquids, are their reliability, long life, high power handling and high beam quality emission. In particular, crystals are known to be suitable materials for high power Raman lasers, due to their very good thermal properties, high damage threshold, high Raman gain. All these properties had finally made crystalline media Raman lasers a class of very efficient devices. Another class of solid-state devices was represented by fiber Raman lasers, realized in silica glass and showing two unique features, not possessed by Raman lasers based on crystalline media. The first was a broadband gain, due to the homogeneous broadening effect introduced by the amorphous glass. Such property enabled the achievement of wavelength tuning, a property lacking in crystal based devices. The second, unique feature was the seamless integration, through fusion splices, of the amplifying section (a fiber) with all the other devices required for the cavity, that are the mirrors (by means of fiber Bragg gratings) and the output ports (by means of fiber isolators, fiber directional couplers etc.). This design greatly simplified the assembly and the stability of the laser and enabled the realization of cascaded Raman lasers, in nested cavities, that maintained a high degree of compactness.

In the most recent years, the field of Raman lasers has taken advantage of the ongoing developments in photonics, extending such progresses into the specific content of Raman amplification and lasing. In particular, one should mention the extremely large advancements in the realization of photonic waveguides and circuits, which is leading to realize high quality on-chip devices that so far required bulk optics. As well, one should mention the new possibilities demonstrated by a new class of optical fibers based on photonic bandgap guidance. In the following sections we first briefly recall the fundamentals of Raman lasers and then we introduce the most relevant, recent advancements in Raman lasers.

Basic Principles

Raman lasers are based on the stimulated Raman scattering, an inelastic scattering process in which photons of a wave at one frequency, defined as the pump, are converted into photons at a smaller frequency, the Stokes wave. If the energy of the generated photon is larger than the pump photon energy, then the output wave is defined as anti-Stokes. In these processes part of the energy of the optical waves is exchanged with some internal form of energy of the Raman medium, typically molecular vibrations and rotations or lattice waves. So, Raman scattering can occur in several different media like gases, atomic vapors, liquids, crystals and glasses. In Fig. 1, the energy diagram of the stimulated generation of the Stokes and anti-Stokes waves is depicted.

The anti-Stokes interaction is less probable, because it requires the existence of a higher energy level ($v+1$) much more populated than that of smaller energy (v); however, at thermal equilibrium, the populations are proportional to the Boltzmann distribution, $\exp(-E_v/K_B T)$ where E_v is the energy of level v , K_B is the Boltzmann constant and T the temperature. Therefore, the higher energy levels are less populated and generally Stokes waves (at wavelengths longer than the pump) are generated in a

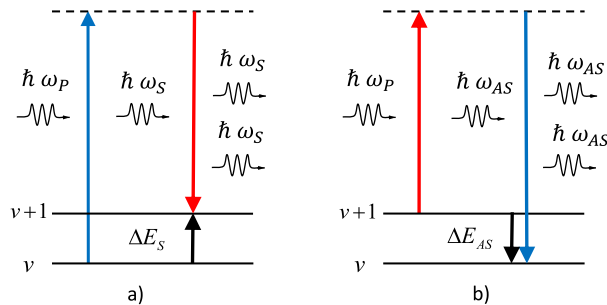


Fig. 1 Energy diagram of the stimulated Raman scattering; solid horizontal lines, v and $v+1$, indicate two medium energy levels, separated by an energy difference of $\Delta E_{S,AS}$; the dashed horizontal lines represent the unstable energy level reached by the medium during the process; $\hbar\omega_P$, $\hbar\omega_S$ and $\hbar\omega_{AS}$ are respectively the pump, Stokes and anti-Stokes photon energies. (a) refers to the Stokes interaction while (b) refers to the anti-Stokes interaction.

Raman laser. Though the pump to Stokes frequency shift is dictated by the medium energy levels, the lasing wavelength can be set to an arbitrary value, provided that a pump wave at a suitable wavelength is available. The efficiency of the processes depicted in Fig. 1 depends on several medium parameters and is summarized by the Raman gain coefficient at resonance g_R , measured in m W^{-1} . Energy levels are also characterized by a certain linewidth (with the typical Lorentzian distribution) that depends on which vibrational or rotational mode is excited, on the temperature, and on the coupling between the excited state and other states.

By imposing the conservation of the number of photons, that characterizes the interaction depicted in Fig. 1(a), a set of two nonlinear equations can be found for the evolution along the propagation direction (z) of the Stokes and pump wave intensities (I_S , I_P , measured in W m^{-2}):

$$\frac{dI_S}{dz} = (g_R I_P - \alpha_S) I_S, \quad \frac{dI_P}{dz} = \pm \left(\frac{\omega_P}{\omega_S} g_R I_S + \alpha_P \right) I_P$$

where α_S and α_P are the medium attenuation coefficients (measured in m^{-1}) respectively at the Stokes and pump frequencies, the $- (+)$ sign in the second equation applies to the case when the pump is co-propagating (counter-propagating) with respect to the Stokes wave. The equations are a valuable model to predict Raman laser properties and must be solved with boundary conditions that take into account the presence of a cavity. They are valid for the continuous and quasi-continuous wave cases; in other regimes (e.g. pulsed), propagation equations like the modified Schrödinger equations must be used instead.

Advancements in Crystalline Raman Lasers

The availability of new Raman crystalline materials, with large thermal damage thresholds and of suitable continuous wave pumping diode lasers have enabled to broaden the set of wavelengths covered by this type of laser sources and to achieve more easily the continuous wave emission.

The problem of thermal handling has been efficiently tackled by the growth of high quality Tungstate and Vanadate crystals (see Table 1). The heat generated during Raman scattering is an intrinsic problem of Raman lasers; in fact, it must be recalled that the energy difference between input and output photons is released to the medium in the form of vibrational or rotational energy so it amounts to

$$P_{\text{heat}} = P_S \left(\frac{\omega_P}{\omega_S} - 1 \right)$$

where $P_S = \hbar \omega_S I_S$ is the Stokes wave power. In the near infrared region and with crystalline media the ratio between heat dissipated power and Stokes power (P_{heat}/P_S) can be as high as 10%. For this reason, the crystals used in Raman lasers should present a high thermal conductivity. Tungstates and Vanadates, as shown in Table 1, have a much higher thermal conductivity of Barium nitrate that was previously identified as the most performing crystalline media. In addition, Tungstates show a negative thermo-optic coefficient (refractive index decreases with temperature) and so thermal lensing, that can decrease the damage threshold, is avoided. The improved heating handling properties highly compensate the gain that is relatively smaller than that of Barium nitrate.

Crystalline Raman lasers pumped by an external pump can provide high output powers (up to about 2 W) with efficiencies that span from 5% to 17%. Pulse durations are in the range between 1 and 50 ns, with very few examples of sub-nanosecond operation, and repetition rates are in the range of 10–50 kHz.

A recent, major innovation is the possibility of achieving pump and Raman lasing in the same crystal (see Fig. 2), to finally realize the so called intra-cavity self-Raman lasers.

Table 1 Raman frequency shift, linewidth, gain coefficient, thermal conductivity and thermo-optic coefficient for various crystalline Raman media.

	Raman shift [cm^{-1}]	Raman linewidth [cm^{-1}]	Raman gain coefficient @1064 nm [cm GW^{-1}]	Thermal conductivity @25°C [$\text{W m}^{-1} \text{K}^{-1}$]	Thermo-optic coefficient [10^{-6}K^{-1}]
Barium nitrate $\text{Ba}(\text{NO}_3)_2$	1047	0.4	11	1.17	–20
Calcium tungstate CaWO_4	908	4.8	3	16	–7.1, ^a –10.2 ^a
Potassium gadolinium tungstate $\text{KGd}(\text{WO}_4)_2$	768	6.4	4.4	2.5–3.4	–0.8, ^a –5.5 ^a
Yttrium vanadate YVO_4	892	2.6	>4.5	5.2	3
Diamond C	1332	1.5	12	2000	9.6

^aValue depending on the crystal orientation.

Source: All data except for diamond from Piper, J.A., Pask, H.M., 2007. Crystalline raman lasers. IEEE Journal Of Selected Topics In Quantum Electronics 13 (3), 692–704; data for diamond are from Lubeigt W., Bonner, G.M., Hastie, J.E., et al. Continuous-wave diamond Raman laser. Optics Letters 35 (17), 2994–2996.

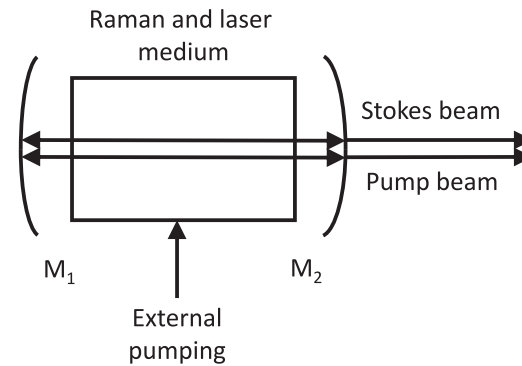


Fig. 2 Experimental setup of an intra-cavity, self-Raman oscillator. Mirrors M_1 and M_2 are highly reflective at the Stokes and at the pump wavelengths.

However, it must be remarked that, though simplifying the setup by eliminating the need for a separate Raman crystal, the thermal loading of the crystal in intra-cavity self-Raman lasers is extremely high, because of the coexistence of both amplification phenomena.

Another recent, remarkable advancement in the experimental setup of crystalline Raman lasers is the increased efficiency of intra-cavity frequency-doubling that has lead to high power output beams (1.8 W, with 9% efficiency) in the visible yellow-orange region.

The most significant recent development in crystalline Raman lasers is certainly represented by continuous lasing operation, also enabled by the exploitation of intra-cavity self-Raman lasing. Continuous wave laser output powers of about 700–800 mW have been achieved, with efficiencies spanning from 4% to 10%. Output power is still limited due to the high thermal loading that clearly affects the continuous wave regime.

From the thermal viewpoint, diamonds are the most promising crystalline media for realizing solid state Raman lasers because diamond thermal conductivity is the highest among solids (see [Table 1](#)). The quality of diamonds grown by means of chemical vapor deposition are by now comparable to the natural ones, with extremely broad transparency windows ($0.225\text{--}2.6\text{ }\mu\text{m}$, $>6.2\text{ }\mu\text{m}$). Therefore, diamond Raman lasers have reached, in a few years, a high degree of maturity. Though intra-cavity diamond Raman lasers may still suffer the thermal loading of lasing material, the external cavity setup enabled high conversion efficiencies (from 50% to 70%) and a very high output power: from about 10 W for continuous wave, to about 100 W in quasi-continuous wave and to a few kW in a pulsed regime (typically with kHz repetition rates and 20 ns pulses). The large diamond thermal conductivity also allows for the implementation of multiple pumping, with further enhancement of the output power (up to 6.7 kW was demonstrated). But the applications of diamond Raman laser can be extended also to the regime of very short pulses. In fact, though phonon dephasing time is in the order of 7 ps, short pulses (down to 25 fs) have been achieved exploiting the strong (soliton-like) compression effects due to the interplay between the cavity dispersion and the strong nonlinear chirp due to the Kerr nonlinearity. Finally, both monolithic and on-chip diamond Raman lasers can be realized and feature low thresholds (less than 100 mW) and Raman cascading up to the 4th order in the continuous wave regime.

On-Chip Raman Lasers

The evolution of the manufacturing technology of photonic waveguides and circuits has lead, in the recent past, to a process of miniaturization of several devices, including Raman lasers. The route for on-chip devices is particularly advanced with semi-conductors, a field in which the high footprint technologies developed for electronics (epitaxy, e-beam lithography etc.) have been transferred to the photonic domain.

The most striking result in the field of Raman lasers is the demonstration and realization of silicon Raman lasers. Though silicon represents the by far most important material in electronics, its exploitation in photonics is hampered by the fact that the amplification through stimulated emission of radiation is hard to achieve, because it is an indirect bandgap semiconductor. To circumvent the lack of lasing action, Raman amplification has been proposed and successfully demonstrated.

Silicon presents a Raman shift of about 520 cm^{-1} with a linewidth of 4.6 cm^{-1} and a very large Raman gain coefficient (estimated about 76 cm GW^{-1} at $1.55\text{ }\mu\text{m}$ pumping). The physical process of Raman scattering in semiconductors is slightly different from that occurring in insulators, because it is mediated by electrons. In fact, as depicted in [Fig. 3](#), the input pump photon first excites an electron-hole pair; the electron interacts with the medium and generates the lattice phonon.

The radiative recombination of electron-hole pairs finally results into the emission of the scattered Stokes photon. In silicon (and semiconductor media in general) other nonlinear phenomena can occur that can hamper Raman amplification. A particularly limiting effect is the two-photon absorption, which can cause pump depletion. Such phenomenon highly depends on the effective recombination lifetime of the free carriers, which can be reduced by a proper design of the waveguide. Several approaches can be used that include the interaction, recombination and diffusion with interface states at the boundary between the silicon and the buried oxide layer in SOI (silicon-on-insulator) waveguides or the realization of reverse-bias p-n junctions.

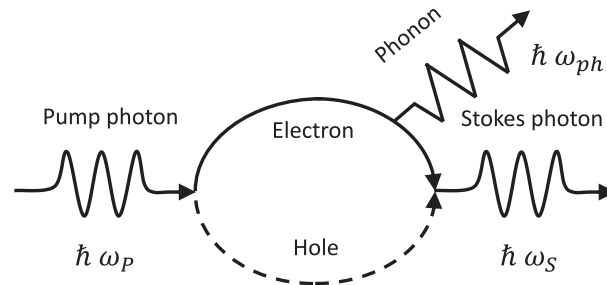


Fig. 3 Feynman diagram describing the Raman scattering process in semiconductors. A pump photon first generates an electron–hole pair. The electron loses energy by interaction with the lattice and a phonon is generated. The electron–hole recombination generates a Stokes (less energetic) photon. Based on Jalali, B., Raghunathan, V., Dimitropoulos, D., Boyraz, O. 2006. Raman-based silicon photonics. *IEEE Journal of Selected Topics in Quantum Electronics* 12 (3), 412–421. doi:10.1109/JSTQE.2006.872708.

Several types of silicon Raman lasers have been demonstrated by means of the previously mentioned approaches and with different type of cavities (that is ring, Fabry-Peròt, photonic crystals) getting continuous wave emission with a threshold as low as 20 mW, output power of about 50 mW and conversion efficiency of 12%. Slightly inferior performance (26 mW threshold, 10 mW output) can be achieved without any electrical bias. Remarkably, the Raman cascade lasing to the second order of continuous wave Stokes was also achieved with 5 mW of output power.

On chip Raman lasers has been manufactured also with other materials and techniques. Examples are the ultralow threshold (74 μ W), high efficient (45%) monolithic silica micro-cavities and III-V Raman injection lasers that exploited triply resonant Raman scattering to enhance the low gain coefficient of such materials that is rather low (0.008 cm GW^{-1}) in comparison to other crystalline insulators or to silicon.

Advancements in Fiber Raman Lasers

Fiber Raman lasers based on step index single mode silica fibers proved to be one of the most reliable and effective types of Raman sources. Silica glass fibers present negligible losses over a rather large window ($<0.5 \text{ dB/km}$ between 1.2 and 1.7 μm). Moreover, the seamless integration of all the devices needed for the lasing make this type of source one of the most appealing in several practical applications. In fact, active lasing materials (in particular rare earths like neodymium, erbium, ytterbium, thulium, holmium etc.) can be easily embedded in the fiber glass matrix to realize an intra-cavity self-Raman pumping scheme, very similarly to what mentioned above for crystalline media. Highly reflective, narrow band mirrors can be permanently inscribed in the same fibers in the form of Bragg gratings that offer reliable and stable performance and, in principle, a certain degree of tuning, by means of the fiber temperature and fiber strain control. Other fundamental devices like power splitter, circulators and isolators are as well realized in optical fibers. Finally, all parts can be assembled by means of fusion splices, that present low losses (including low return losses) and that enable high power handling.

From the thermal handling viewpoint optical fibers are also very convenient. In silica fibers, the small frequency shift (about 490 cm^{-1}) reduces the total amount of heat power to 6.8% of the Stokes power. Moreover, the long length and the very small radius of the optical fibers by far compensate the value of the thermal conductivity (about $1.4 \text{ K m}^{-1} \text{ W}^{-1}$) that is smaller than many crystalline media. The use of double-clad fibers enables cladding-pumped setups in which high total pump power can be launched and propagates in the internal cladding, while the Stokes signal is generated in the core. The outcome of all the previously mentioned combined properties and design characteristics is that extremely high power loading can be realized in fiber Raman lasers.

Continuous wave Raman fiber lasers have, in fact reached, the record kW regime (about 1.5 kW) at 1120 nm; in that case intra-cavity pumping was realized by doping a 26 m long fiber with ytterbium ions, made active by means of external semiconductor laser diodes emitting at 915 and 976 nm. Fiber Bragg gratings provided the cavity for both Stokes and pump waves and the overall optical efficiency was also very good (75%). Similar examples of this technology have been demonstrated with emission at other wavelengths, enabled by means of different co-doping, including some examples in the 2 μm range (thulium and holmium doped intra-cavity fiber Raman lasers). One of the main issues to achieve high spectral purity is that of suppressing high-order Stokes lasing. This is achieved by keeping the fiber short enough (typically 20–30 m) or by inserting intra-cavity filters that in fibers can be realized by designing fibers with specific cut-off wavelengths, so keeping, once again, the valuable seamless integration of the whole device.

Conversely, when cascaded Raman generation is a target this can be also achieved with very high power output and very large overall efficiency. A fifth order converter (1117–1480 nm) reported about 300 W output power at 1480 nm with a total conversion efficiency of about 64%.

Fiber Raman lasers, as well as the previously mentioned Raman laser types, also leveraged onto the most recent progresses in material science and technology to improve their performance and broadens the spectrum of emitted frequencies and therefore

applications. Though silica glass is still the most preferred medium for handling high power beams, other glasses have been tested. Fluoride glasses (Raman shift 572 cm^{-1}) enabled lasing beyond $2\text{ }\mu\text{m}$ by pumping at 1981 nm through an intra-cavity thulium doped fiber with an output power in the order of a few Watts. The chalcogenide glasses seem more promising for extending Raman lasing beyond $3\text{ }\mu\text{m}$. In fact, losses of chalcogenide materials rapidly decay at wavelengths larger than $3\text{ }\mu\text{m}$ (for example As_2S_3 fiber has $>1\text{ dB/m}$ at $3\text{ }\mu\text{m}$ but 0.1 dB/m at $3.345\text{ }\mu\text{m}$). The Raman gain coefficient is much larger than that of silica (about 780 times) and that of fluoride glasses, so partly compensating the increase in linear losses. Thermal and mechanical properties of these glasses are, however, poorer with respect to silica glass and therefore it can be expected that their usage will be limited to achieve emission in spectral regions where silica fiber losses would be too high.

Another fundamental improvement in the field of optical fibers, namely the realization of photonic crystal fibers has also opened new perspectives for Raman fiber lasers. In photonic crystal fibers the guiding effect is no longer based on the total internal reflection, but on the photonic bandgap principle. The cladding is therefore substituted by a periodic structure, inhibiting light transmissions (bandgap) at the wavelengths that are to be guided so that they can only propagate in the photonic crystal defect, which is the fiber core. Among the various unusual properties of photonic crystal fibers, the one that is particularly attractive for Raman lasing is the property of guiding light in low refractive index materials, including vacuum or gases. Gases were among the first Raman media to be used in realizing single pass and cavity Raman lasers, because they presented rather large Raman gain coefficients (from tens to a few hundred times that of silica glass), large transparency spectra and, moreover, Raman frequency shifts are the largest among Raman media. However, the vessels used to contain the gas were bulky and the nonlinearity was severely limited by diffraction. In fact, no light confinement was actually possible and therefore the Raman interaction can take place only in the region where the pump beam was focused, whose effective volume is diffraction limited. Hollow core photonic crystal fibers fully eliminated such problem and high intensity beams (fiber mode effective areas are in the order of tens of μm^2) can be maintained over much longer distances (from ten to a few hundred meters) before fiber loss becomes too large, so enabling Raman generation with a very high conversion efficiency. Experimental demonstrations were essentially made with hydrogen, that was the most widely studied in bulk vessels. A conversion efficiency of about 50% (line shifted by about 600 cm^{-1}) with 2.25 W of input power was achieved in a 30 m long hydrogen filled hollow core photonic crystal fiber. Conversion efficiency seems essentially limited by fiber loss, as demonstrated by the fact that 99.99% of output power was contained in the Stokes wave. Fully exploiting seamless fiber integration, experiments including fiber Bragg gratings mirrors (cavity) reduced the threshold to sub-Watt levels (600 mW). The very high efficiency derives also from the suppression of higher order Stokes waves, due to the large loss of the fiber at longer wavelengths.

Conversely, the availability of hollow core fiber (the so called Kagomé photonic crystal fibers) in which fiber loss coefficient is small (e.g. $<2\text{ dB m}^{-1}$) over a large band (a few hundreds of nm at wavelengths in the visible and infrared spectral ranges) enabled the realization of a Raman cascade and even the coherent excitation of multiple rotational or vibrational Raman lines. In the latter case, a completely novel feature for Raman lasers, which is the emission of a multi-octave frequency comb can be achieved. In fact, leveraging on the coherence of the Raman lines, all excited by the same pump wave, Raman combs have a potential to deliver sub-femtosecond pulses, as the spectrum can span over almost 1000 THz with a variable line spacing (from a few tens to more than one hundred THz).

Applications

Raman lasers continue to be applied in a large variety of fields. The availability of new wavelengths has broadened the range of applications in remote sensing for safety, like in atmospheric science but also in new areas like food quality evaluation. Biomedical applications are, by far, the area in which the recent progresses have had the largest impact. The availability of unconventional wavelengths (like the yellow-orange emission of crystalline lasers), the compactness and high-power delivery (like for fiber Raman lasers pumped by diode lasers) have made these devices a fundamental tool in diagnostics and therapies, in several medical areas like dermatology, ophthalmology, surgery and dentistry. Raman lasers remain fundamental tools for spectroscopic applications, now entailing also coherent combs and sub-femtosecond pulses, which also opened a completely new line of application in metrology. In optical communications, fiber Raman lasers are the most exploited solution for pumping Raman amplification along the fiber link, because of the reliability, efficiency and seamless integration with the rest of the network.

See also: Raman Spectroscopy

Further Reading

- Abdolvand, A., Walser, A.M., Ziemenczuk, M., Nguyen, T., Russell, P. St. J., 2012. Generation of a phase-locked Raman frequency comb in gas-filled hollow-core photonic crystal fiber. *Optics Letters* 37, 4362–4364. doi:10.1364/OL.37.004362.
- Bernier, M., Fortin, V., El-Amraoui, M., Messaddeq, Y., Vallée, R., 2014. $3.77\text{ }\mu\text{m}$ fiber laser based on cascaded Raman gain in a chalcogenide glass fiber. *Opt. Lett.* 39, 2052–2055. doi:10.1364/OL.39.002052.
- Bernier, M., Fortin, V., Caron, N., *et al.*, 2013. Mid-infrared chalcogenide glass Raman fiber laser. *Optics Letters* 38, 127–129. doi:10.1364/OL.38.000127.
- Boyras, O., Jalali, B., 2004. Demonstration of a silicon Raman laser. *Optics Express* 12, 5269–5273. doi:10.1364/OPEX.12.005269.

- Cerný, P., Jelínková, H., Zverev, P.G., Basiev, T.T., 2004. Solid state lasers with Raman frequency conversion. *Progress in Quantum Electronics* 28, 113–143. doi:10.1016/j.pquantelec.2003.09.003.
- Chen, Y.F., 2004. Efficient 1521-nm Nd: GdVO₄ Raman laser. *Optics Letters* 29, 2632–2634. doi:10.1364/OL.29.002632.
- Chen, Y.F., 2004. Compact efficient all-solid-state eye-safe laser with self-frequency Raman conversion in a Nd: YVO₄ crystal. *Optics Letters* 29, 2172–2174. doi:10.1364/OL.29.002172.
- Couny, F., Benabid, F., Light, P.S., 2007. Subwatt threshold cw Raman fiber-gas laser based on H₂-filled hollow-core photonic crystal fiber. *Physical Review Letters* 99, 143903. doi:10.1103/PhysRevLett.99.143903.
- Feng, Y., Taylor, L.R., Bonaccini Calia, D., 2009. 150 W highly-efficient Raman fiber laser. *Optics Express* 17, 23678–23683. doi:10.1364/OE.17.023678.
- Fève, J.-P.M., Shortoff, K.E., Bohn, M.J., Brasseur, J.K., 2011. High average power diamond Raman laser. *Optics Express* 19, 913–922. doi:10.1364/OE.19.000913.
- Georgiev, D., Gapontsev, V.P., Dronov, A.G., et al., 2005. Watts-level frequency doubling of a narrow line linearly polarized Raman fiber laser to 589 nm. *Optics Express* 13, 6772–6776. doi:10.1364/OPEX.13.006772.
- Grabitchikov, A.S., Linsinetskii, V.A., Orlovich, V.A., et al., 2004. Multimode pumped continuous-wave solid-state Raman laser. *Optics Letters* 29, 2524–2526. doi:10.1364/OL.29.002524.
- Granados, E., Spence, D.J., Mildren, R.P., 2011. Deep ultraviolet diamond Raman laser. *Optics Express* 19, 10857–10863. doi:10.1364/OE.19.010857.
- Hausmann, B.J.M., Bulu, I., Venkataraman, V., Deotare, P., Lončar, M., 2014. Diamond nonlinear photonics. *Nature Photonics* 8, 369–374. doi:10.1038/nphoton.2014.72.
- Jackson, S.D., Anzueto-Sánchez, G., 2006. Chalcogenide glass Raman fiber laser. *Applied Physics Letters* 88, 221106. doi:10.1063/1.2208369.
- Jalali, B., Raghunathan, V., Dimitropoulos, D., Boyraz, O., 2006. Raman-based silicon photonics. *IEEE Journal of Selected Topics in Quantum Electronics* 12 (3), 412–421. doi:10.1109/JSTQE.2006.872708.
- Jiang, H., Zhang, L., Feng, Y., 2015. Silica-based fiber Raman laser at > 2.4 μm. *Optics Letters* 40, 3249–3252. doi:10.1364/OL.40.003249.
- Kippenberg, T.J., Spillane, S.M., Armani, D.K., Vahala, K.J., 2004. Ultralow-threshold microcavity Raman laser on a microelectronic chip. *Opt. Lett.* 29, 1224–1226. doi:10.1364/OL.29.001224.
- Kitzler, O., McKay, A., Mildren, R.P., 2012. Continuous-wave wavelength conversion for high-power applications using an external cavity diamond Raman laser. *Optics Letters* 37, 2790–2792. doi:10.1364/OL.37.002790.
- Kivistö, S., Hakulinen, T., Guina, M., Okhotnikov, O.G., 2007. Tunable Raman soliton source using mode-locked Tm-Ho fiber laser. *IEEE Photonics Technology Letters* 19, 934–936. doi:10.1109/LPT.2007.898877.
- Latawiec, P., Venkataraman, V., Burek, M.J., et al., 2015. On-chip diamond Raman laser. *Optica* 2, 924–928. doi:10.1364/OPTICA.2.000924.
- Lee, A.J., Spence, D.J., Piper, J.A., Pask, H.M., 2010. A wavelength-versatile, continuous-wave, self-Raman solid-state laser operating in the visible. *Optics Express* 18, 20013–20018. doi:10.1364/OE.18.020013.
- Li, B.-B., Xiao, Y.-F., Yan, M.-Y., Clements, W.R., Gong, Q., 2013. Low-threshold Raman laser from an on-chip, high-Q, polymer-coated microcavity. *Optics Letters* 38, 1802–1804. doi:10.1364/OL.38.001802.
- Lin, J., Spence, D.J., 2016. 25.5 fs dissipative soliton diamond Raman laser. *Optics Letters* 41, 1861–1864. doi:10.1364/OL.41.001861.
- Lubeigt, W., Bonner, G.M., Hastie, J.E., et al., 2010. Continuous-wave diamond Raman laser. *Optics Letters* 35, 2994–2996. doi:10.1364/OL.35.002994.
- McKay, A., Spence, D.J., Coutts, D.W., Mildren, R.P., 2017. Diamond-based concept for combining beams at very high average powers. *Laser & Photonics Reviews* 11, 1600130. doi:10.1002/lpor.201600130.
- Mildren, R.P., Convery, M., Pask, H.M., Piper, J.A., McKay, T., 2004. Efficient, all-solid-state, Raman laser in the yellow, orange and red. *Optics Express* 12, 785–790. doi:10.1364/OPEX.12.000785.
- Mildren, R.P., Pask, H.M., Ogilvy, H., Piper, J.A., 2005. Discretely tunable, all-solid-state laser in the green, yellow, and red. *Optics Letters* 30, 1500–1502. doi:10.1364/OL.30.001500.
- Mildren, R.P., Sabella, A., 2009. Highly efficient diamond Raman laser. *Optics Letters* 34, 2811–2813. doi:10.1364/OL.34.002811.
- Parrotta, D.C., Kemp, A.J., Dawson, M.D., Hastie, J.E., 2013. Multiwatt, continuous-wave, tunable diamond Raman laser with intracavity frequency-doubling to the visible region. *IEEE Journal on Selected Topics in Quantum Electronics* 19, 6471175. doi:10.1109/JSTQE.2013.2249046.
- Pask, H.M., Dekker, P., Mildren, R.P., Spence, D.J., Piper, J.A., 2008. Wavelength-versatile visible and UV sources based on crystalline Raman lasers. *Progress in Quantum Electronics* 32, 121–158. doi:10.1016/j.pquantelec.2008.09.001.
- Piper, J.A., Pask, H.M., 2007. Crystalline Raman Lasers. *IEEE Journal of Selected Topics in Quantum Electronics* 13 (3), 692–704. doi:10.1109/JSTQE.2007.897175.
- Qin, G., Liao, M., Suzuki, T., Mori, A., Ohishi, Y., 2008. Widely tunable ring-cavity tellurite fiber Raman laser. *Optics Letters* 33, 2014–2016. doi:10.1364/OL.33.002014.
- Reilly, S., Savitski, V.G., Liu, H., et al., 2015. Monolithic diamond Raman laser. *Opt. Lett.* 40, 930–933. doi:10.1364/OL.40.000930.
- Rong, H., Jones, R., Liu, A., et al., 2005. A continuous-wave Raman silicon laser. *Nature* 433, 725–728. doi:10.1038/nature03346.
- Rong, H., Kuo, Y.-H., Xu, S., et al., 2006. Monolithic integrated Raman silicon laser. *Optics Express* 14, 6705–6712. doi:10.1364/OE.14.006705.
- Rong, H., Xu, S., Cohen, O., et al., 2008. A cascaded silicon Raman laser. *Nature Photonics* 2, 170–174. doi:10.1038/nphoton.2008.4.
- Rong, H., Xu, S., Kuo, Y.-H., et al., 2007. Low-threshold continuous-wave Raman silicon laser. *Nature Photonics* 1, 232–237. doi:10.1038/nphoton.2007.29.
- Russell, P. St. J., Hölzer, P., Chang, W., Abdolvand, A., Travers, J.C., 2014. Hollow-core photonic crystal fibres for gas-based nonlinear optics. *Nature Photonics* 8, 278–286. doi:10.1038/nphoton.2013.312.
- Sabella, A., Piper, J.A., Mildren, R.P., 2010. 1240 nm diamond Raman laser operating near the quantum limit. *Optics Letters* 35, 3874–3876. doi:10.1364/OL.35.003874.
- Sabella, A., Piper, J.A., Mildren, R.P., 2014. Diamond Raman laser with continuously tunable output from 3.38 to 3.80 μm. *Optics Letters* 39, 4037–4040. doi:10.1364/OL.39.004037.
- Supradeepa, V.R., Feng, Y., Nicholson, J.W., 2017. Raman fiber lasers. *Journal of Optics* 19 (2), 023001. doi:10.1088/2040-8986/19/2/023001.
- Takahashi, Y., Inui, Y., Chihara, M., et al., 2013. A micrometre-scale Raman silicon laser with a microwatt threshold. *Nature* 498, 470–474. doi:10.1038/nature12237.
- Taylor, L.R., Feng, Y., Calia, D.B., 2010. 50W CW visible laser source at 589 nm obtained via frequency doubling of three coherently combined narrow-band Raman fibre amplifiers. *Optics Express* 18, 8540–8555. doi:10.1364/OE.18.008540.
- Troccoli, M., Belyanin, A., Capasso, F., et al., 2005. Raman injection laser. *Nature* 433, 845–848. doi:10.1038/nature03330.

GaN Lasers

Harumasa Yoshida, Mie University, Tsu, Mie, Japan

© 2018 Elsevier Ltd. All rights reserved.

Introduction

There are three main types of wide bandgap semiconductors, which are group III-nitrides as GaN, II-chalcogenides as ZnSe and II-oxides as ZnO. Diamond and silicon carbide are also important wide bandgap semiconductors. Among those materials, the III-nitrides are direct bandgap semiconductor and one of the most promising candidates for the semiconductor light sources such as laser diodes and light emitting diodes (LEDs). The III-nitride-based semiconductors can provide a wide range of bandgap energies from 0.7 to 6.2 eV, compounding with InN, GaN, and AlN. In principle, it is possible to tune the emission spectral range of the GaN system down to infrared or toward deep ultraviolet region. At present, InGaN- and AlGaN-based LEDs cover the spectral range from red (> 600 nm) to deep ultraviolet (210 nm) wavelengths. In contrast, the laser diodes require a more complex structure, thicker layers and higher crystalline quality than the LEDs to satisfy all the requirements of suitable optical and electrical confinements as well as high emission efficiency for lasing. As a result of the tight material requirements, the laser diodes are confined to a much narrower spectral range from green (530 nm) to near ultraviolet (320 nm) wavelengths.

The InGaN-based laser diodes are already in markets. In contrast, the AlGaN-based laser diodes set into the markets through a stage of research and development. Based on the 405-nm InGaN laser diodes, Blu-ray disk systems have been in widespread use for the high-density optical data storage application, instead of the conventional DVD storage system based on the 650-nm InGaP laser diodes. The blue InGaN laser diodes would be useful for metallic material processing and high brightness illuminations with scintillators because of the high power characteristic. For laser display application, the green InGaN laser diodes become alternative light source to bulky second-harmonic-generation YAG solid-state lasers. The ultraviolet AlGaN laser diodes could offer an alternative to the costly and bulky lasers such as excimer, HeCd and nitrogen gas lasers, and as third and fourth harmonic-generation YAG lasers. These conventional ultraviolet gas and solid-state lasers are widely used in the scientific and industrial fields. The merits and advantages of flexible wavelength, compactness, low-power consumption, and rapid startup of the laser diodes offer much convenience of the laser applied equipment, and provide a chance to create novel applications. For example, the ultraviolet laser diodes can provide excellent performance for the long range sensing applications such as light detection and ranging (LiDAR), and 3-dimensional mapping because the maximum permissible exposure for human eyes is much higher in 300-nm-band than that in 900-nm-band of the conventional GaAs-based laser diodes under the condition of pulsed mode operation.

In this section, after the brief overview on group III-nitrides, typical fabrication procedures and characteristics of both the AlGaN laser diodes for ultraviolet region and the InGaN laser diodes for visible region are described. Impacts on device characteristics caused by unique nature of the nitrides are also mentioned.

Group III-Nitrides

Group III-nitrides are compounds composed of group III-elements of boron (B), aluminum (Al), gallium (Ga) and indium (In), and group V-element of nitrogen (N). AlN, GaN, and InN are the binary nitrides suited to optical and electronic devices. Ternary and quaternary alloys such as AlGaN, InGaN, and AlInGaN are also synthesized for the complicated device fabrication. The GaN and the related nitrides crystallize in wurtzite or zincblende structure (Morkoç, 2008; Takahashi *et al.*, 2007; Nakamura *et al.*, 1997). The wurtzite structure is much stable at room temperature and pressure, whereas the zincblende structure is quasi-stable due to the difference of the degree of ionicity in the ionic bonding components. The crystalline structure of wurtzite is hexagonal, and the [0001] axis perpendicular to the hexagon structured plane is called as *c*-axis, as schematically illustrated in Fig. 1. The $\langle \bar{1}100 \rangle$ axes perpendicular to the lateral plane of hexagonal column and the $\langle 11\bar{2}0 \rangle$ axes parallel to that plane are called as *m*-axes and *a*-axes, respectively. The III-nitrides exhibit remarkable polarization effects due to the asymmetry in atomic configuration along the [0001] direction. The strain-induced piezoelectric and spontaneous polarization fields have substantial impacts on the device characteristics. In contrast, the *m*-direction and *a*-direction are free from polarization due to symmetric structures along the directions. The *c*-plane is described as polar plane, and the *m*-plane and *a*-plane are described as nonpolar planes after the polarization properties. Other polarization-controlled planes can be obtained by slicing the crystal at some angles. Those are called as semi-polar planes. The typical one is $\{20\bar{2}1\}$ planes, as shown in Fig. 2. Table 1 lists typical crystal lattice parameters for AlN, GaN, and InN. The wurtzite structures of AlN, GaN, and InN exhibit direct bandgap properties, whereas zincblende ones show indirect bandgap behavior. The typical values of bandgaps are also listed on the Table 1. Bandgaps of 6.2 (AlN), 3.4 (GaN), and 0.7 eV (InN) correspond to wavelengths of 200, 360, and 1800 nm, respectively. By controlling the composition of ternary compound AlGaN and InGaN, the optical emission is expected to be tuned extensively from deep ultraviolet to infrared spectral regions in principle.

The differences of lattice constants between AlN and GaN, GaN and InN, and AlN and InN are 2.5, 11, and 14%, respectively. It could be hard to grow quality crystals of AlInN alloy due to the large differences in optimal growth temperatures as well as lattice

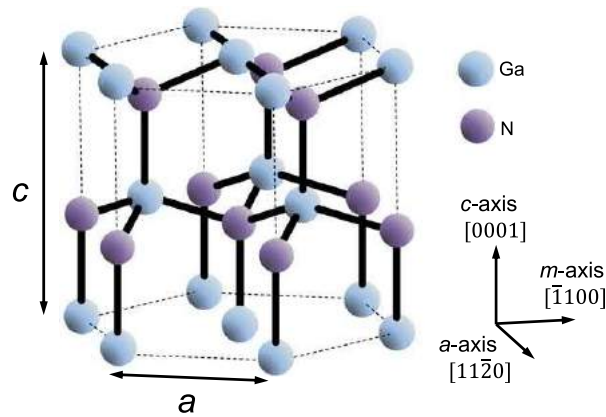


Fig. 1 Schematic illustration of wurtzite GaN crystalline structure. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

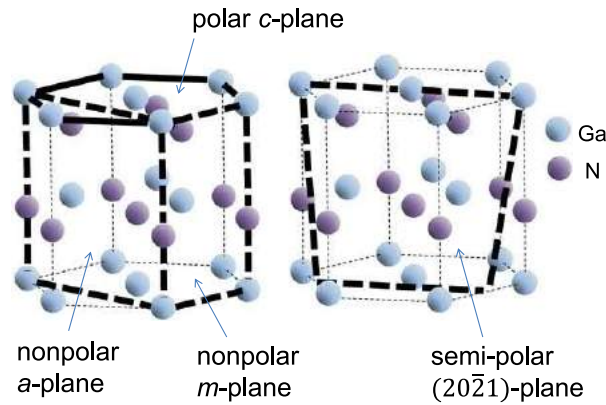


Fig. 2 Polar, nonpolar and semi-polar planes of GaN. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

Table 1 Lattice constants and bandgaps (wurtzite)

	a (Å)	c (Å)	Bandgap (eV)
AlN	3.111	4.980	6.2
GaN	3.189	5.186	3.4
InN	3.538	5.703	0.7

Source: Morkoç, H. (Ed.), 2008. Handbook of Nitride Semiconductors and Devices, vol. 1, Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA and Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

constants between AlN and InN. The calculated lattice parameters a and c of AlGaIn and InGaIn alloys can be approximately predicted by applying Vegard's law from several reports:

$$a_{\text{AlGaIn}} = 3.1986 - 0.0891x \text{ Å and } c_{\text{AlGaIn}} = 5.2262 - 0.2323x \text{ Å} \quad (1)$$

$$a_{\text{InGaIn}} = 3.1986 + 0.3862y \text{ Å and } c_{\text{InGaIn}} = 5.2262 + 0.574y \text{ Å} \quad (2)$$

where x and y are the AlN and InN molar fractions in the alloys, respectively. The bandgaps of these alloys can also be approximated, respectively, as:

$$E_{\text{AlGaIn}} = 6.1x + 3.4(1 - x) - b_{\text{AlGaIn}} \cdot x(1 - x) \text{ eV} \quad (3)$$

$$E_{\text{InGaIn}} = 0.7y + 3.4(1 - y) - b_{\text{InGaIn}} \cdot y(1 - y) \text{ eV} \quad (4)$$

Among a lot of reported parameters, the typical bowing parameters will be $b_{\text{AlGaIn}} = 1.0$ eV and $b_{\text{InGaIn}} = 1.43$ eV.

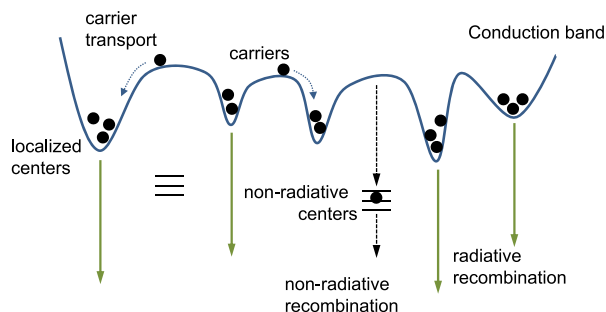


Fig. 3 Schematic representation on potential fluctuation. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. *Wide Bandgap Semiconductors*. Berlin Heidelberg: Springer-Verlag and Nakamura, S., Pearton, S., Fasol, G., 1997/2000. *The Blue Laser Diode*. Berlin Heidelberg: Springer-Verlag.

For the growth of nitride semiconductors, sapphire substrates are widely used because of the thermal and chemical stabilities. Although the large differences in lattice constants and thermal expansion coefficients between the sapphire ($a=4.763 \text{ \AA}$, $\Delta a/a=7.0 \times 10^{-6} \text{ K}^{-1}$) and GaN ($a=3.189 \text{ \AA}$, $\Delta a/a=5.59 \times 10^{-6} \text{ K}^{-1}$), a low-temperature AlN or GaN buffer layer enables the growth of quality GaN film on the sapphire substrates. Commercially available bulk GaN substrates are more suitable for the growth of high quality GaN film because of the lattice matching. However, it is difficult to grow high quality AlGaIn film with high AlN molar fraction due to the lack of lattice matched substrates such as bulk AlN substrates.

There is a strong localization effect especially in the InGaIn alloy. The localization effect fortunately leads to an excellent emission efficiency of the InGaIn-based optical devices such as LEDs. The fluctuations of well width and alloy composition result in the potential fluctuation in active layers. The degree of potential fluctuation is caused by alloy disorder and correlates to the energy gap difference between the two binaries such as InN and GaN. InN and GaN provide an energy gap difference of 2.7 eV which is much larger than a typical value of a few hundred milli-electron volt in order for the II–VI and III–V semiconductor alloys. And the InGaIn alloy is easy to generate a compositional modulation due to a large difference of 11% in the lattice constants between InN and GaN. The localized centers act as quantum dots and reduce the trap probability of excitons in nonradiative centers, as schematically illustrated in Fig. 3. However, the excess potential fluctuation reduces the gain for lasing of the laser diodes and relates to the difficulty in green InGaIn laser diodes, as mentioned later.

AlGaIn-Based Laser Diodes

The AlGaIn laser diodes, in which the active layer was composed of the In-free AlGaIn alloys, were reported for the first time in 2008 (Yoshida *et al.*, 2008). The AlGaIn laser diodes were grown on the foreign substrates such as sapphire and silicon carbide because of lack of lattice matched substrates. The lattice mismatch results in the dislocations and the cracking of layers as well as the wafer bending. Epitaxial lateral overgrowth (ELO) method is one of solutions to relax the strain between the AlGaIn layers and the lattice mismatched substrates, and to reduce the dislocations of the AlGaIn layers grown on the substrates.

The quality AlGaIn crystal films can be grown by the ELO method on the sapphire substrates. Fig. 4 shows a schematic illustration of typical layer structure. The material growth for crack-free and low-dislocation-density AlGaIn layers is carried out by metalorganic vaporphase epitaxy (MOVPE). Trimethylgallium (TMG), trimethylaluminum (TMA), and ammonia (NH_3) are used as the source materials. A *c*-plane (0001) sapphire is used as the substrate. After the deposition of a buffer layer with a thickness of several tens of nanometers at a relatively low growth temperature of 500°C on the sapphire substrate, a $2.5\text{-}\mu\text{m}$ -thick GaN layer is grown at 1050°C . Following the deposition of SiO_2 striped masks along the *m*-axis on the GaN layer, inclined-facet GaN seed crystals can be selectively grown at 900°C through the masks by facet-controlled epitaxy. Both the width and the spacing of stripe masks are $2\text{--}3 \text{ }\mu\text{m}$. Then, a subsequent AlGaIn layer is laterally overgrown at 1150°C on the inclined facets of the GaN seeds. A coalescence of each overgrown AlGaIn layer from opposite facets of the GaN seeds is achieved. There is no crack over the entire area of the 2 in. diameter wafer. On the contrary, a number of cracks are easily generated in the wafer grown by the conventional growth sequence without the inclined-facet GaN seed crystals. The cathode-luminescence observation reveals the low dislocation density of the AlGaIn film with an average dark spot density in the order of 10^8 cm^{-2} . The dislocation density of AlGaIn film grown by the ELO method is almost two orders of magnitude lower compared with that of the film directly grown on the sapphire substrate with the low-temperature buffer layer.

The AlGaIn laser diodes are fabricated on the quality AlGaIn film. The adequate AlN molar fractions of the AlGaIn film depend on the emission wavelengths of laser diodes to satisfy the sufficient optical and electrical confinements in the active region. However, the layer configuration and the compositional allocation are constrained and compromised not to generate the cracks in film due to the lattice mismatched crystal growth despite the ELO method. As a typical example, the laser diodes with emission wavelengths of $320\text{--}340 \text{ nm}$ are based on the AlGaIn film with an AlN molar fraction of 0.3. The layer structure on the $\text{Al}_{0.3}\text{Ga}_{0.7}\text{N}$ film is composed of a $2.8\text{-}\mu\text{m}$ -thick $n\text{-Al}_{0.3}\text{Ga}_{0.7}\text{N}$ contacting layer, a 600-nm -thick $n\text{-Al}_{0.3}\text{Ga}_{0.7}\text{N}$ cladding layer, a 90-nm -thick AlGaIn guiding layer, AlGaIn multiple quantum wells (MQWs), a 120-nm -thick AlGaIn guiding layer, a 20-nm -thick $p\text{-AlGaIn}$

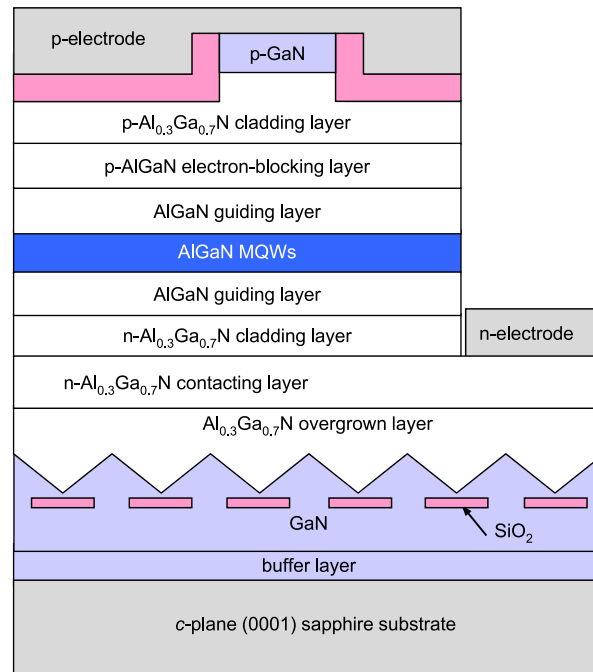


Fig. 4 Layer structure of AlGaN laser diodes. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaN multiple-quantum-well laser diode. *Nature Photonics* 2, 551–554. Available at: <http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html>.

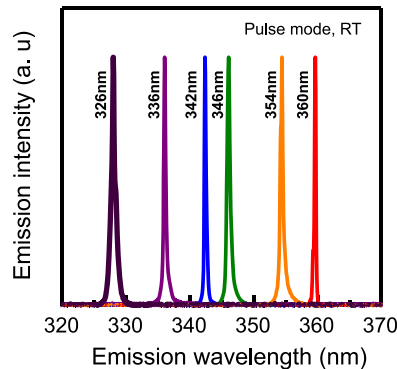


Fig. 5 A series of spectra for AlGaN laser diodes. Reproduced from Yoshida, H., Kuwabara, M., Yamashita, Y., *et al.*, 2009. AlGaN-based laser diodes for the short-wavelength ultraviolet region. *New Journal of Physics* 11, 125013. Available at: <http://iopscience.iop.org/article/10.1088/1367-2630/11/12/125013/meta>.

electron-blocking layer, a 500-nm-thick p-Al_{0.3}Ga_{0.7}N cladding layer and a 25-nm-thick p-GaN contacting layer. The devices are processed as ridge waveguide lasers in accordance with the following procedure. After forming laser stripes and exposing the n-AlGaN contacting layer by conventional dry etching, an n-electrode is deposited on the exposed n-AlGaN contacting layer. A p-electrode is deposited on the p-GaN contacting layer. Facets of the laser cavities are also formed by dry etching. It is difficult to employ a highly reflective coating exactly for the cavity facets due to the obstruction of remaining unetched terraces in front of the facets.

Fig. 5 shows a series of lasing spectra for the AlGaN laser diodes with the AlN molar fraction being from 0 to 6% in the AlGaN wells. These spectra are measured under a pulsed-current mode. The emission wavelengths depend on both the MQW compositions and the well widths. An emission wavelength of 336 nm is nearly identical to 337 nm of the conventional nitrogen gas lasers. At present, the lasing wavelength has reached up to 326 nm corresponding to 325 nm of the HeCd gas lasers.

Fig. 6 shows a typical light output–current (L – I) characteristic of the AlGaN laser diodes with an emission wavelength of 342 nm. The device exhibits clear nonlinear behavior in the L – I characteristic. The threshold current of 390 mA corresponds to a threshold current density of 8.7 kA cm^{-2} for the 5- μm -ridge stripe and the 900- μm -long cavity. The peak output power from one

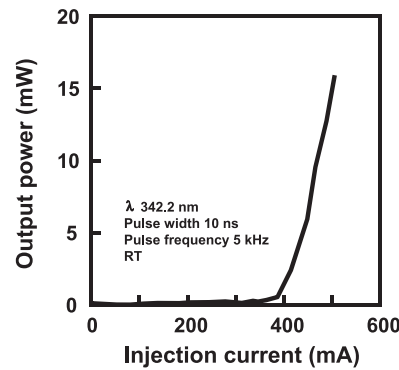


Fig. 6 Light output-current characteristic of AlGaIn laser diode on sapphire substrate. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaIn multiple-quantum-well laser diode. *Nature Photonics* 2, 551–554. Available at: <http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html>.

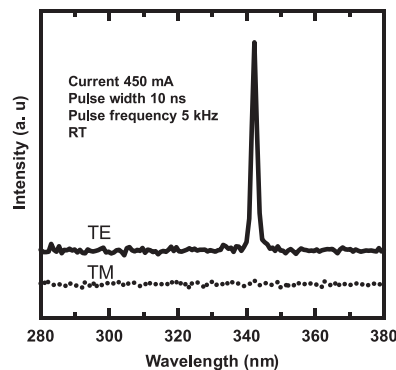


Fig. 7 Polarized emission spectra of AlGaIn laser diode. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaIn multiple-quantum-well laser diode. *Nature Photonics* 2, 551–554. Available at: <http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html>.

side of the facets approaches as high as 15 mW. The differential external quantum efficiency (DEQE) for the output from both facets is estimated to be 8.2%.

The typical transverse electric (TE) and transverse magnetic (TM) emission spectra from the AlGaIn laser diode are shown in **Fig. 7**. Above the threshold, a very strong polarized emission can be observed. The emission is TE polarized due to the dominance of TE optical gain in the AlGaIn MQWs with relatively low AlN molar fraction as well as the high reflectivity for the TE mode as a result of the cavity facets. Due to the large negative crystal field splitting in AlN compared with the positive value in GaN, an unusual valence band structure of AlN gives rise to unique optical polarization properties in the AlGaIn alloys. The optical emission is dominantly TE polarized ($E \perp c$) for GaN, whereas the emission is predominantly TM polarized ($E \parallel c$) for AlN because a crystal-field split-off-hole state lies quite lower than heavy and light hole-states. The introduction of Al significantly changes the polarized optical gain property in the MQW. For the AlGaIn MQWs with rather high AlN molar fraction, the TM emission is expected to be the dominant emission due to the polarization property of AlGaIn alloys.

The peak output power of AlGaIn laser diodes on the sapphire substrates are around several tens of milli-watts, and higher powers are expected to satisfy various application requirements. A broad-area striped structure is one of approaches for the higher output powers. The laser diodes fabricated on sapphire unavoidably accept lateral conduction due to the insulating property of sapphire. The laser diodes on sapphire substrates are unsuitable for the broad-area stripes because the lateral conduction leads to an un-uniform current injection into the wide emitter stripes. Moreover, the mirror facets of AlGaIn laser diodes on the sapphire substrates are formed typically by dry etching due to the difficulty in cleaving the sapphire. Far-field patterns (FFPs) of the laser diodes are deteriorated by optical interference between the direct beam and the reflected beam from the lying down terrace in front of the etched mirror facet. The commercially available bulk GaN substrates are a potential candidate for both the vertical conducting structure and the cleaved mirror facets because of the excellent conductivity and cleavability of GaN.

The broad-area striped AlGaIn laser diodes can be fabricated by the similar manner as mentioned above on the GaN substrates as an alternative to the sapphire substrates (Aoki *et al.*, 2015; Taketomi *et al.*, 2016). **Fig. 8** shows a typical L – I characteristic of the 50- μ m-ridge stripe AlGaIn laser diodes with an emission wavelength of 338 nm under the pulsed operation. The peak output power exhibits a clear nonlinear behavior around a threshold current of 7 A corresponding to a current density of 38.9 kA cm^{-2} ,

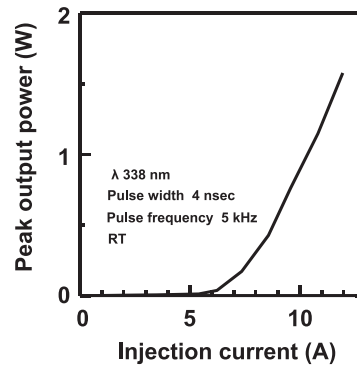


Fig. 8 Light output-current characteristic of broad-area AlGaIn laser diode on GaN substrate. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaIn multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: <http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta>.

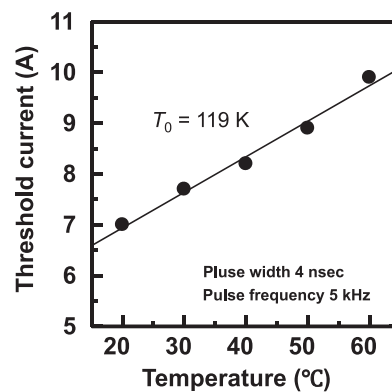


Fig. 9 Temperature dependence on threshold currents of AlGaIn laser diode. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaIn multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: <http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta>.

and linearly increases above the threshold. A peak power of over 1 W is observed at operating currents of over 10 A. The slope efficiency and DEQE are estimated to be 0.31 W A^{-1} and 8.5%, respectively. The threshold current density and the slope efficiency are inferior to those of InGaIn laser diodes (see next section of InGaIn-based laser diodes). It could result from both the weak optical confinement due to the smaller difference in refractive index between the guiding layers and the cladding layers, and the electron overflow from the active region into the p-AlGaIn cladding layer due to the low hole concentration as well as being free from the In effect.

Fig. 9 shows the temperature dependence on the threshold currents under the pulsed operation in a temperature range from 20 to 60°C. The threshold currents are increased from 7.0 to 9.9 A with increasing the temperature. The characteristic temperature T_0 of threshold current can be estimated to be 119K from the fitted line. This value is comparable to that of the AlGaIn laser diodes on sapphire substrates. **Fig. 10** shows the temperature dependence on emission wavelengths at a peak power of approximately 500 mW. The wavelengths move from 338.4 to 339.8 nm with a temperature rise of 40°C. The temperature coefficient of emission wavelengths is also estimated to be 0.033 nm K^{-1} . The small temperature coefficient is comparable to that of the infrared GaAs-based distributed feedback laser diodes with Bragg grating inside. The excellent spectral stability is favorable characteristic for various applications.

Fig. 11(a) shows the horizontal and vertical FFPs of the laser diode with GaN/AlGaIn MQWs on the GaN substrate. The ridge width and cavity length are 1.5 and 300 μm , respectively. For the comparison, the FFPs of similar device on sapphire substrate are also shown in **Fig. 11(b)**. The two dimensional images of FFPs are also shown on the graphs. Both the horizontal and vertical FFPs of the device on GaN substrate exhibit unimodal profiles, respectively. In contrast, the distorted FFPs can be seen for the device on sapphire substrate. The deteriorated FFP of the laser diode on sapphire substrate are considered to be caused by the interference between the direct beam from emitting area and the reflected beam from the residual terrace lying down in front of the etched mirror facet. The cleaved mirror facets provide the excellent FFP for flexible optical beam control owing to the cleavability of GaN substrates.

There are difficulties in the crystal growth of the AlGaIn layer with higher AlN molar fraction due to the lack of lattice matched substrates, and in the conductivity control of Mg doped p-type AlGaIn due to the low hole concentration originated from the high

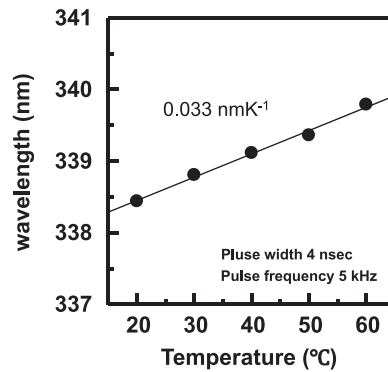


Fig. 10 Temperature dependence on emission wavelengths of AlGaIn laser diode. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaIn multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: <http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta>.

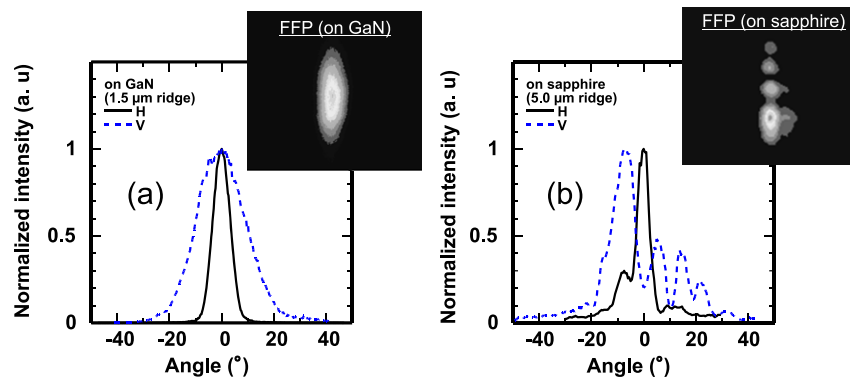


Fig. 11 FFPs of AlGaIn laser diodes on (a) GaN and (b) sapphire substrates. Reproduced from Aoki, Y., Kuwabara, M., Yamashita, Y., *et al.*, 2015. A 350-nm-band GaN/AlGaIn multiple-quantum-well laser diode on bulk GaN. Applied Physics Letters 107, 151103-1-151103-4. Available at: <http://aip.scitation.org/doi/abs/10.1063/1.4933257>.

binding energy of Mg. The Mg acceptor activation energy of 510 meV in AlN is much higher than 150 meV in GaN. These difficulties limit continuous wave (CW) operations and extending the wavelengths toward deep ultraviolet.

InGaIn-Based Laser Diodes

The first InGaIn laser diodes, in which the active layer was composed of the InGaIn alloys, were demonstrated by Nakamura *et al.* in 1996 (Nakamura *et al.*, 1996, 1997/2000). The laser diodes were fabricated on the sapphire substrates because the bulk GaN substrates were unavailable commercially in the early stages. Nowadays, the bulk GaN substrates are commonly used for the InGaIn laser diodes. According to the report on the first InGaIn laser diodes, the laser diodes were grown by MOVPE on the *c*-plane (0001) sapphire substrate. Fig. 12 shows a schematic illustration of the layer structure. The device is composed of a low temperature GaN buffer layer on the sapphire, a n-GaN contacting layer, a n-InGaIn buffer layer, a n-AlGaIn cladding layer, a n-GaN guiding layer, InGaIn MQWs, a p-AlGaIn layer, a p-GaN guiding layer, a p-AlGaIn cladding layer, and a p-GaN contacting layer. The mirror cavity facets were formed by dry etching due to the difficulty in cleaving the epilayers grown on the sapphire. In order to reduce the threshold current, high reflection facet coating was employed. Evaporated Ni/Au and Ti/Al were used for the p-GaN and n-GaN contact layers, respectively. Reportedly, the threshold current was 1.7 A, corresponding to a current density of 4 kA cm^{-2} , under the pulsed operation of a pulse width of 2 μs and a period of 2 ms. A voltage of 34 V was applied at the threshold. The DEQE is estimated to be 13% from a pulsed output power of 215 mW per facet at an injection current of 2.3 A at room temperature. The laser exhibited a dominantly strong emission at a wavelength of 417 nm with a full width of half maximum (FWHM) of 1.6 nm. Because of the etched facets, it was difficult to obtain a unimodal FFP.

It was not easy to operate the InGaIn laser diodes in CW mode. Under the direct current operation, the devices were easily failed and burned in a moment at room temperature due to heat generation caused by the high operating voltages. The room-temperature CW operation of InGaIn laser diodes with a long lifetime of 40 min was achieved in 1997 by optimizing the growth conditions, the ohmic contacts and the doping profiles. Further long life CW laser diodes were demonstrated using the GaN

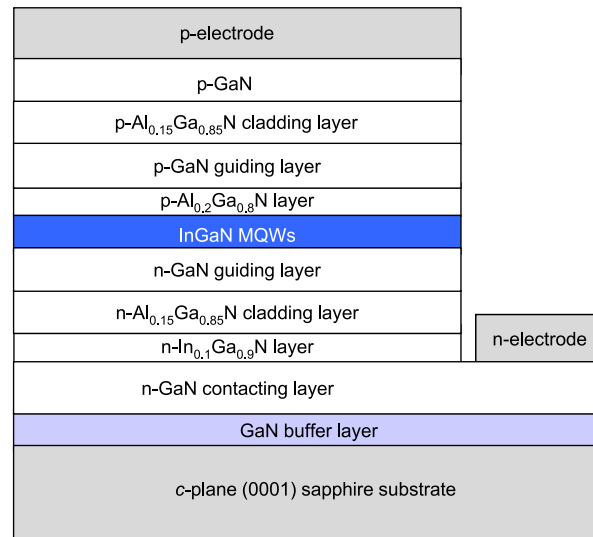


Fig. 12 Layer structure of InGaN laser diodes. Reproduced from Nakamura, S., Pearton, S., Fasol, G., 1997/2000. *The Blue Laser Diode*, Berlin Heidelberg: Springer-Verlag.

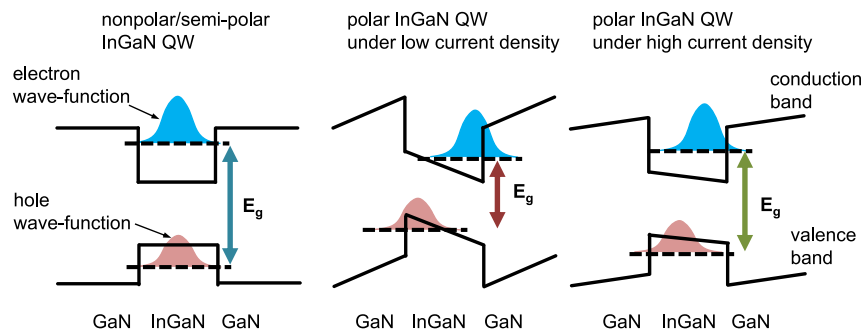


Fig. 13 Schematic band models of nonpolar and polar InGaN QWs. Reproduced from Kafafi, Z.H., Martin- Palma, R.J., Nogueira, A.F., *et al.*, 2015. The role of photonucs in energy. *Journal of Photon Energy* 5 (1), 050997. Available at: <http://photonicsforenergy.spiedigitallibrary.org/article.aspx?articleid=2463108>.

substrates as alternative to the sapphire substrates. Owing to an excellent GaN thermal conductivity of $1.3 \text{ W cm}^{-1} \text{ K}^{-1}$ higher than $0.5 \text{ W cm}^{-1} \text{ K}^{-1}$ of the sapphire, as well as the low dislocation density of the GaN substrates, the laser diodes were operated for 800 h. The 405-nm violet InGaN laser diodes led to Blu-ray disk technology in early 2000s.

For the violet InGaN laser diodes, the InN molar fractions of InGaN well layer are around 0.15. With increasing the InN molar fraction in the InGaN well layer of MQWs, the emission can shift theoretically to longer wavelengths. The blue laser diodes with an emission wavelength of 450 nm were reported in 2000. An estimated lifetime of 200 h for the blue laser diodes was seriously inferior to that of 2000 h for the violet laser diodes. This relates to poor crystalline quality of the InGaN MQW structure caused by dissociation of the InGaN active layer with a high InN molar fraction of 0.3. In recent years, a high output power of over 4 W has been marked for the blue laser diodes by the optimization of growth condition and the improvement in thermal management. The high output-power blue laser diodes are applied to laser displays, high brightness lamps, and material processing.

The red AlInGaP laser diodes are already available. Green laser diodes are the last key semiconductor-light-source for compact full-color-display applications consisting with three primary colors of red, blue, and green. Ideally, wurzite InGaN alloys grown in the nonpolar (m and a directions) and the semi-polar (for instance, $\{20\bar{2}1\}$) directions are the most potential candidate because of less blue-shift of the emission and higher internal quantum efficiency (IQE) in QWs compared to the polar c -plane QWs. The QWs on the nonpolar and semi-polar planes are free from and/or less quantum confined Stark effect (QCSE). The internal electric field is induced along the c -axis of the QWs by spontaneous and piezoelectric polarizations due to the asymmetry of the crystal as well as the strain resulting from the lattice mismatch between the layers. The internal electric field of QWs can be screened with increasing the injection current. The counter effect for the QCSE by the screening results in the increase of bandgap, or the blue-shift of emission wavelength. Moreover, the internal electric field spatially separates an electron wave-function and a hole wave-function in the QWs. The separate wave-functions reduce the radiative recombination probability or IQE. These phenomena are observed not only in the InGaN system but also in the AlGaIn system. **Fig. 13** schematically illustrates the phenomena. The

blue-shift behavior as well as the less IQE inhibits the shift to longer wavelengths of the InGaN laser diodes. The less blue-shift and the sufficient wave-function overlap are the important properties of nonpolar and semi-polar QWs.

There are other issues on In compositional fluctuation with increasing In contents in the InGaN active layer caused by the immiscibility of InN with GaN. Although the moderate In fluctuation is much effective for the improved emission of blue InGaN LEDs because of the localized carriers, as already shown in Fig. 3, the excessive fluctuation generate much defects in the QWs. The defects act as non-radiative centers and result in less IQE of the green InGaN laser diodes. The In fluctuation also leads to broad emission spectrum, and reduces the gain for lasing of the laser diodes.

Several groups demonstrated the CW operations of green laser diodes taking advantages of the nonpolar and semi-polar InGaN layers (Okamoto *et al.*, 2009; Takagi *et al.*, 2012; Miyoshi *et al.*, 2009). However, it is difficult to grow or to fabricate the large size nonpolar and semi-polar GaN substrates. The poor availability of non- or semi-polar substrates constrains the developments of green laser diodes. Contrarily, the polar *c*-GaN substrates with a diameter of 2 in. are commercially and widely available. The diameter reaches up to 6 in. at present. By optimizing the growth conditions of InGaN even with high InN molar fraction, the watt-class 537-nm green laser diodes have been realized on the *c*-GaN substrates. Based on the technological progress of the blue and green InGaN laser diodes as well as the red AlInGaP laser diodes, the standard wavelengths of three primary colors have been determined to be 467, 532, and 630 nm for the super high vision displays (ITU-R BT.2020-1).

References

- Aoki, Y., Kuwabara, M., Yamashita, Y., *et al.*, 2015. A 350-nm-band GaN/AlGaIn multiple-quantum-well laser diode on bulk GaN. *Applied Physics Letters* 107, 151103-1–151103-4.
- Miyoshi, T., Masui, S., Okada, T., *et al.*, 2009. 510–515nm InGaN-based green laser diodes on *c*-plane GaN substrate. *Applied Physics Express* 2, 062201-1–062201-3.
- Morkoç, H. (Ed.), 2008. *Handbook of Nitride Semiconductors and Devices*, vol. 1. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.
- Nakamura, S., Pearson, S., Fasol, G., 1997/2000. *The Blue Laser Diode*. Berlin Heidelberg: Springer-Verlag.
- Nakamura, S., Senoh, M., Nagahama, S., *et al.*, 1996. InGaN-Based Multi-Quantum-Well-Structure Laser Diodes. *Japanese Journal Applied Physics* 35, L74–L76.
- Okamoto, K., Kashiwagi, J., Tanaka, T., Kubota, M., 2009. Nonpolar *m*-plane InGaN multiple quantum well laser diodes with a lasing wavelength of 499.8 nm. *Applied Physics Letters* 94, 071105-1–071105-3.
- Takagi, S., Enya, Y., Kyono, T., *et al.*, 2012. High-power (over 100mW) green laser diodes on semipolar {20 $\bar{2}$ 1} GaN substrates operating at wavelengths beyond 530 nm. *Applied Physics Express* 5, 082102-1–082102-3.
- Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. *Wide Bandgap Semiconductors*. Berlin Heidelberg: Springer-Verlag.
- Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338nm AlGaIn multiple-quantum-well laser diode on a GaN substrate. *Japanese Journal Applied Physics* 55, 05FJ05.
- Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaIn multiple-quantum-well laser diode. *Nature Photonics* 2, 551–554.

Optically Pumped Semiconductor Lasers

Jerome V Moloney and Alexandre Laurain, University of Arizona, Tucson, AZ, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Optically pumped semiconductor lasers (OPSLs), also referred to as vertical external cavity surface emitting lasers (VECSELs), or semiconductor disk laser (SDL) exhibit tremendous performance in terms of power, spatial and temporal coherence, noise characteristic, short pulse generation, and offer unique features such as tunability, power scalability, or intracavity frequency conversion (second harmonic generation (SHG) or difference frequency generation (DFG)). They combine the advantage of a semiconductor laser technology: compact and highly efficient gain medium that can be engineered to emit at a targeted wavelength, covering the visible up to the mid-infrared (IR) range, with the advantage of diode-pumped solid-state lasers: power scalability with circular low divergence beam and high-Q optical cavity. Unlike most semiconductor lasers, the laser light is amplified perpendicular to the surface of the semiconductor rather than the edge of the semiconductor chip. In this sense, they can be viewed as close cousins of the vertical cavity surface emitting laser (VCSEL), a widely employed commercial source that is generally electrically pumped and exhibits a typical low power of a few mill-watts in single mode operation. Fundamentally different, however, is the physical layout of the OPSL cavity depicted in Fig. 1.

Compared to a VCSEL, the optical cavity is extended by using an external mirror instead of a top distributed Bragg reflector (DBR), leaving an air gap between the OPSL chip and the external output coupler. The external optical cavity can be independently designed to provide a single transverse mode operation, by matching the cavity fundamental mode size with the pumped area. The external cavity also facilitate the insertion of various optical elements like an etalon and birefringent filter for high power single frequency operation, a saturable absorber for passive modelocking of ultrashort pulses, or a non-linear crystal that can exploit the very high intra-cavity circulating field to efficiently convert the fundamental wavelength to visible and UV frequencies via harmonic generation down to terahertz frequencies via DFG. Thanks to the very broad absorption spectrum of semiconductor media and extremely thin active region, OPSL chips can be excited with a spatially and spectrally incoherent pump beam, permitting the use of commercially available high power fiber-coupled diode laser modules. The chip itself consists of a high reflectivity distributed Bragg mirror (DBR) consisting of multiple repeats of two semiconductor materials, followed by an extremely short gain region (typically $\sim 2 \mu\text{m}$) consisting of pump absorbing barriers and multiple quantum wells (QWs) or quantum dots (QDs) placed strategically according to the standing wave of the optical field. For high power operation, the chip is typically bonded to a chemical vapor deposition (CVD) diamond heat spreader which is placed in close contact with a water cooled copper heat sink. Thermal management is critically important as most of the pump induced heat is generated in an extremely small volume (disk of diameter $500 \mu\text{m}$ and thickness $2 \mu\text{m}$).

The wavelength versatility and tunability of these compact sources result from the many possible semiconductor gain material combinations that can be grown to provide direct lasing in the UV (InGa_N, AlGa_N, Ga_N), visible GaInP/AlGaIn/GaAs, near-IR (InGaAs, AlGaAs, GaAs), and mid-IR (InGaSb, AlGaSb, GaSb). Due to the relatively recent development and immaturity of some material systems, especially in the visible and UV, it is sometimes more advantageous to generate these wavelengths via SHG using a more mature material system like InGaAs/GaAs. This usually offers even greater power at the expense of added cavity complexity. It is thus not surprising that many companies such as Coherent (Santa Clara) commercialize high power semiconductor lasers with a technology based on InGaAs QWs, and offer visible and ultraviolet wavelengths via efficient intra-cavity SHG in OPSLs.

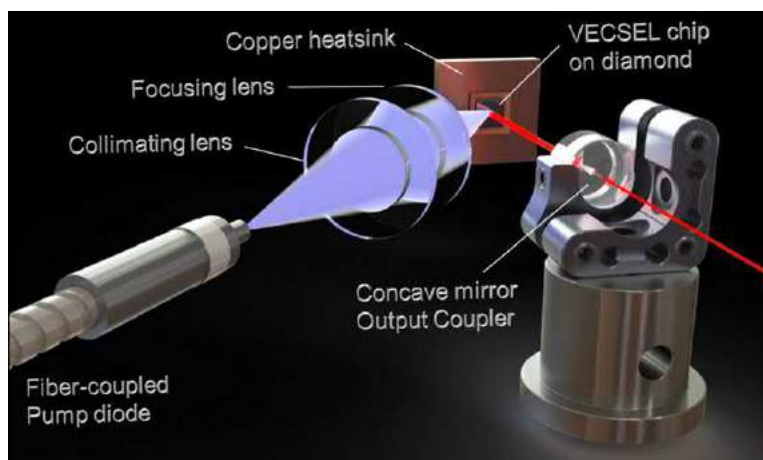


Fig. 1 Representation of an optically pumped semiconductor laser (OPSL) setup in a linear cavity configuration.

Another remarkable feature of OPSLs is their ability to generate ultrashort pulses at repetition rates unreachable with solid state or fiber laser technology due to Q-switching instabilities. The optical cavity can be arranged to integrate a semiconductor saturable absorber mirror (SESAM) to passively modelock the modes with repetition rates ranging from tens of megahertz up to 100 GHz. This is a major advantage for the development of compact femtosecond mode-locked source with output peak powers exceeding kilowatts and average output powers exceeding 5 W.

Background

The history of the development of OPSLs is relatively recent, dating back to the mid-1990s. The semiconductor active region of the OPSL chip typically consists of 8–12 QWs arranged to lie close to the anti-nodes of the signal standing-wave within the OPSL, as illustrated in Fig. 2. Due to the micro-cavity or micro-resonator formed by the DBR and the semiconductor/air interface, the field intensity within the structure can be significantly increased at the nominal wavelength. This resonance can be exploited to enhance the relatively low gain resulting from the fact that light is extracted to normal rather than along the QWs plane. A typical resonant periodic gain (RPG) design provides a gain of about 5 to 10%, depending on the number of QWs used and the pumping rate. The resulting device has marked advantages relative to other solid state lasers and semiconductor edge emitters in particular. Specifically, power scaling is simply achieved by increasing the diameter of the pump and mode waist on the chip. For single transverse mode operation, the fundamental mode is matched to the pump areas which serve as a soft aperture to discriminate the high order transverse modes. Indeed, since the transverse modes are not guided like in an edge emitter, they naturally diffract in the free space external cavity and the higher order modes spatially extend beyond than the fundamental mode and are absorbed more in the unpumped region of the OPSL chip. TEM₀₀ operation with large diameter provide a very low divergence beam (<1 degree), often close to the diffraction limit ($M^2 < 1.1$) with a circular symmetry. These large beam diameters also strongly reduce the risk of a light induced surface damage or facet damage, the plague of diode edge emitters. In contrast to high power solid disk lasers, more than 80% of the pump power is absorbed in a single pass, simplifying the pumping scheme.

Most modern OPSLs are now grown via MBE or MOCVD as bottom emitters, namely, the RPG QW stack is grown first on a lattice matched substrate. This growth mode greatly facilitates heat extraction, as the poorly thermal conductive thick substrate (GaAs for InGaAs/AlGaAs/GaAs) can be entirely removed by chemical etching after bonding to a CVD diamond or other heat spreader. Bottom emitters require growth of an etch stop layer (typically AlGaAs or InGaP) just before a cap and the QW stack. The cap layer is needed to confine carriers within barrier regions so that they are efficiently captured in the QWs. The large refractive index of the semiconductor material relative to air ensures strong confinement of the signal field within the resulting semiconductor micro-cavity. In some instances, it is desirable to AR coat the surface of the chip to facilitate mode-locked pulse generation or internal signal conversion via SHG or DFG. The linear cavity depicted in Fig. 1 represents the simplest and optimal arrangement for raw power generation although a V-cavity arrangement provides a better control of the transverse modes on the OPSL chips and can include a SESAM for mode-locking.

Overview

OPSL Design Strategies

While other solid state disk laser sources are limited to wavelengths associated with specific transitions in dopant materials, the semiconductor gain medium supports wide tunability by simply changing the relative composition of the semiconductor

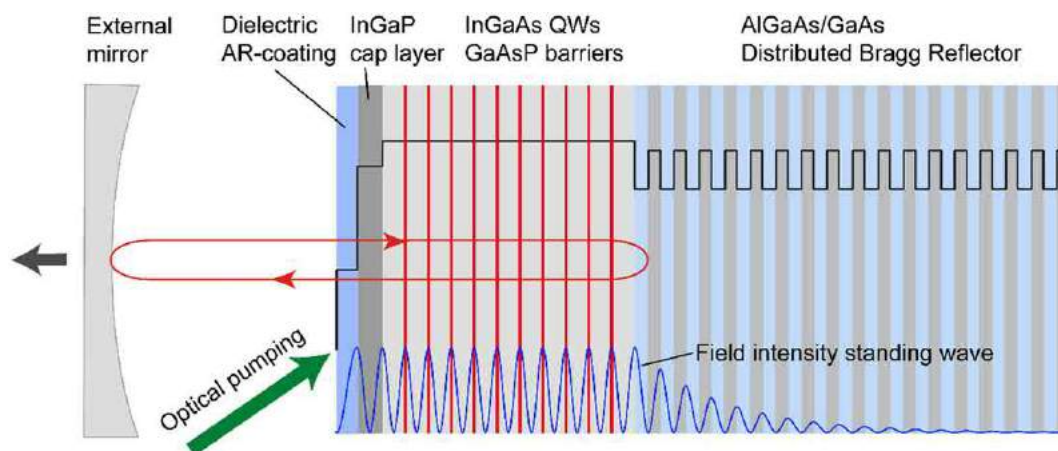


Fig. 2 Schematic representation of a typical GaAs-based vertical external cavity surface emitting laser (VECSEL) structure with a resonant periodic gain (quantum wells (QWs) placed on the field antinodes).

materials. For example, InGaAs QWs can be tuned to lase between 920 and 1200 nm by changing the composition of the Indium in the well. Increasing Indium increases strain and requires that strain compensating layers to be grown between individual QWs in an RPG stack, for example. The tuning range can be extended further toward 1.5 μm by slightly doping with atomic nitrogen. The nitrogen with concentration of a few percent is not incorporated into atom lattice sites but acts as an independent atom inclusion that shifts the valence upwards and reduces the bandgap in InGaAs. The 1.5 μm window can also be accessed via InGaAsP QWs grown on InP. DBRs are difficult to grow for this material system and the combination of InP/InGaAlAs has a low index contrast and consequently requires very thick structures. It is possible to wafer fuse the active region to a GaAs/AlAs Bragg mirror on a GaAs wafer, but the fabrication of such hybrid structures is difficult.

Performance of OPSLs for different targeted applications such as power scaling, high power single frequency operation or mode-locked short pulse generations is strongly dependent on many factors including semiconductor epitaxy design and thermal management. Typical diagnostics of growth accuracy prior to chip processing involve measuring surface and/or edge photoluminescence (PL). The two PL diagnostics differ in that the surface PL, which can be done on a full semiconductor wafer, is convolved with the DBR reflectivity, whereas the edge PL gives a direct measure of the individual QW quality and growth accuracy.

However, the wavelength of the PL at room temperature is usually shorter than the wavelength of the gain peak in a running OPSL. In designing the OPSL chip for targeted performance, one must know accurately the temperature increase under varying pump power as the gain peak itself shifts strongly with temperature increase. Therefore the QWs need to be designed with PL at shorter wavelength anticipating the location of the peak gain at the final laser operating temperature (typically around 400K). One must also take into account the spectral shift of the micro-cavity resonance, as it will also shift with temperature but at a slower rate than the QWs gain. Fig. 3 shows a typical evolution of the unsaturated gain spectrum and micro-cavity resonance under pumping and it's correlation with the output power characteristic. At low pump power the structure is close to room temperature and the QW gain peak is significantly detuned (~ 30 nm) from the nominal wavelength of 1000 nm, whereas the resonance is only slightly detuned (~ 7 nm). As the pump power increase and thus the temperature, the gain spectrum gets higher, broader and shifts at a rate of about 0.3 nm/K whereas the resonance shifts at a rate of about 0.07 nm/K. The gain peak and the resonance will be spectrally aligned when the internal temperature reaches $\sim 400\text{K}$ giving the maximum output power. If the pump power is further increased, the spectral shift and the gain amplitude decrease with temperature lead to a thermal rollover of the output power.

Obviously, the temperature increase for a given pump power depends on the thermal impedance of the device and is the main limiting factor of high power OPSLs. The thermal impedance depends critically on the bonding quality of the OPSL chip to the CVD diamond heat spreader, the thickness and thermal conductivity of the semiconductor membrane and the pump beam size. Increasing the pump beam size reduces the thermal impedance but also increases the lasing threshold. As long as the thermal impedance scales with the beam area, a linear power scaling is possible but, for very large pump beam (~ 1 mm diameter), the lateral thermal heat flow becomes a limiting factor and the impedance no longer scales with the beam area and limits the achievable power output. This limitation together with the possible gain depletion from lateral lasing with large pump beam has set the limit of raw power from a single device to just above 100 W to date, which is remarkable coming from a semiconductor laser.

Current sophisticated software tools (SimuLase) utilize the microscopic physics of the semiconductor active structure to aid in targeting a specific wavelength and power performance. Such software tools reduce dramatically the usual required multi-parameter ad hoc phenomenological treatment needed in less sophisticated approaches. In addition to the optical and modal

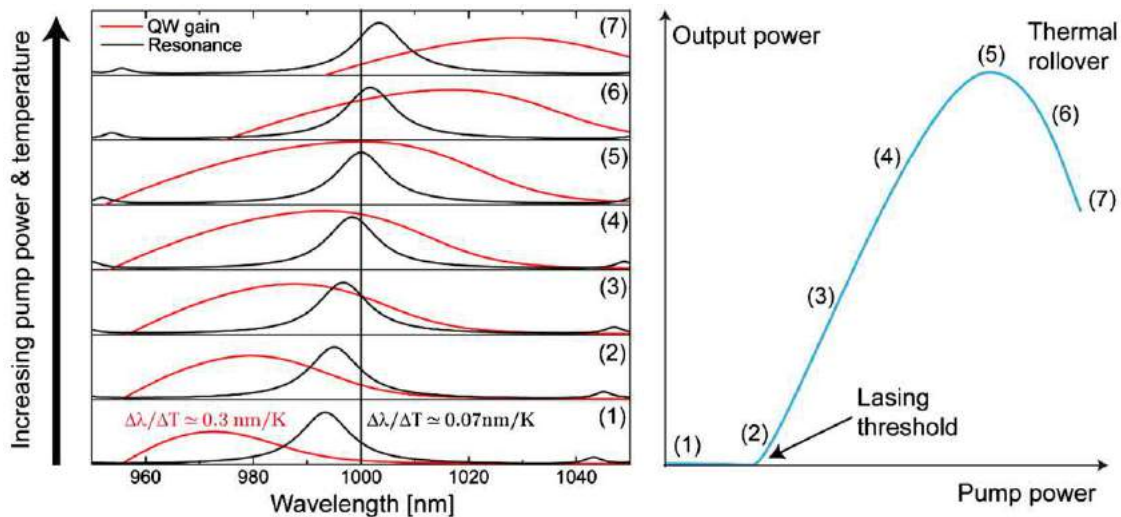


Fig. 3 Evolution of the intrinsic quantum wells (QWs) gain and micro-cavity resonance with increasing pump power for a nominal emission wavelength of 1000 nm at maximum power, and correlation with the output power characteristic.

gain, the OPSL performance also depends strongly on both radiative and nonradiative carrier losses (spontaneous emission and Auger recombination). Auger recombination tends to become increasingly important at high temperature and for longer wavelengths (i.e., lower bandgap).

OPSL Gain Structure

OPSL gain chips are typically designed as RPG structures. Here individual QWs are placed at the anti-nodes of the standing wave laser signal field that is confined within the semiconductor micro-cavity. This gain structure has a high reflectivity distributed Bragg Mirror (DBR) at one end such that the structure acts as a "gain mirror." The DBR is grown of similar materials to that of the active structure, for example, AlGaAs/AlAs alternating layers are employed in structures with InGaAs QWs. This simple optical component can be placed in different ways relative to other elements (mirrors, SESAMs, etc.) to assemble different types of cavities.

The design of the optical gain structure is crucial when optimizing performance for direct continuous wave (CW), intra-cavity SHG/DFG or ultrashort pulse generation. A global optimization scheme of the structure before wafer growth must take into account the "cold" and "hot" lasing properties of the device. When pumped either optically or electrically, the chip heats up and the optical properties can be strongly modified. The goal is to minimize the influence of heating in the active structure making an efficient means of heat extraction critical. The growth mode can strongly impact the thermal management scheme (Fig. 4).

For example, many of the original OPSL devices were grown as top emitters – the DBR stack is first grown on the substrate followed by the multiple QW stack and carrier confinement cap layers. The rather high thermal impedance of semiconductor materials limits this kind of devices to relatively low power (< 10 W). For higher power they generally require that optically transparent heat spreaders (single crystal diamond or silicon carbide) are capillary or wafer bonded to the top of the structure. Ring-shaped copper heatsinks are then bonded to the heat spreader to allow for accessible space for the pump and laser light to reach the OPSL chip. The main disadvantage of this technique is that the heatspreader is inside the laser cavity, requiring very high optical quality and transparent materials. This added element might also induce parasitic reflection, additional losses and birefringence or require a specific antireflection coating, etc., which can alter the performance of the device. It is however, very useful and sometime the only alternative with substrate materials which cannot be easily removed from the structure like GaSb for mid-IR wavelength.

Bottom emitters are grown in reverse order (flip-chip design) – the cap layer is grown first followed by the QW stack and finally the DBR. In some instances it is necessary to first grow an etch stop layer on the substrate. The bottom emitter can be directly bonded with Ti/Au/In or Cr/Au/In to a lower optical quality CVD diamond heat spreader. The great advantage of the bottom emitter is that the entire substrate can be removed with chemical etchants leaving behind an extremely thin OPSL chip. Not surprisingly bottom emitters exhibit dramatically better performance due to very efficient cooling.

OPSL Loss Mechanisms

Individual optical components inserted in OPSL cavities and output coupled transmission are generic optical loss mechanisms that need to be balanced against the relatively low gain of an OPSL chip. Due to the high finesse of the cavity, scattering losses on the surface of the chip and to a lesser extent on the surface of the dielectric mirror(s) cannot always be neglected. Typical scattering losses of 0.25 to 0.75% are expected, which can become increasingly important with low modal gain structures (long wavelengths, anti-resonant design, etc.), and need to be accounted for the design of the OPSL chip and cavity.

For passive mode-locking operation, saturable and non-saturable losses from the absorber (SESAM, Graphene-SAM, etc.) contribute further to the overall cavity loss when trying to achieve lasing threshold and will impact the dynamics of the pulsed lasing regime.

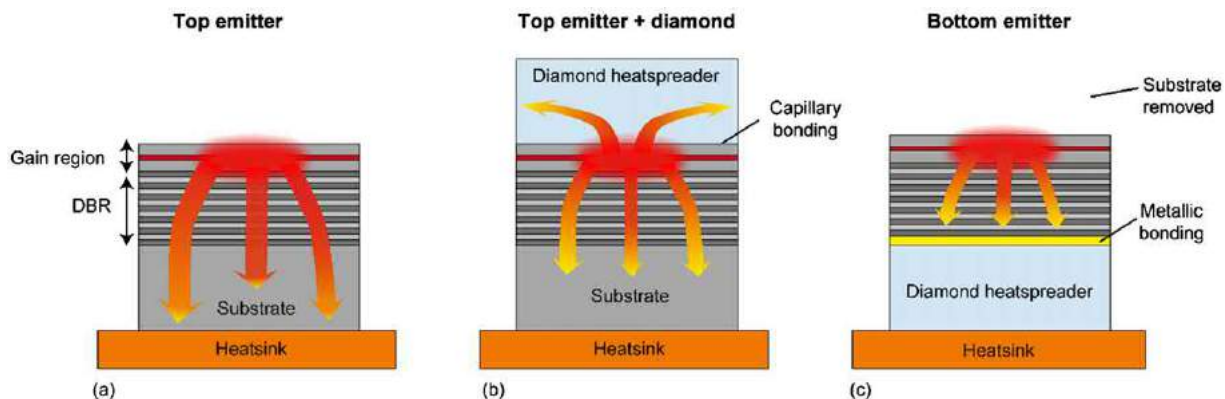


Fig. 4 Thermal management strategies for high power vertical external cavity surface emitting laser (VECSEL). (a) Top emitter: high thermal impedance from the substrate limits to low power regimes; (b) top emitter with intracavity diamond heatspreader: low thermal impedance but increased optical losses; and (c) bottom emitter bonded to a diamond: low thermal impedance and low losses.

On the other hand, intrinsic carrier losses in semiconductor active media directly affect the material gain and efficiency of the device and must be considered or even better microscopically calculated for a given material system, in order to fully optimize a structure. The carrier losses are usually dominated by radiative (spontaneous emission) and non-radiative (Auger losses) recombinations. Defect assisted recombinations which are assumed proportional to the carrier density are negligible for high quality grown structures. Auger losses, generally considered proportional to the carrier density cubed (often not the case!), become dominant at longer wavelengths but are also players at high operating temperatures at shorter wavelengths. The latter can suppress the gain to such an extent that the OPSL does not lase at all or, more typically, suddenly shuts off under strong pumping due to the added heat. Recently, novel Type-II W gain structures have been grown in an effort to suppress Auger losses. These designs displace the electron confining layer from the hole confining layers so as to minimize Auger losses. This is done at the expense of reducing the gain but has been successfully demonstrated for both electrical and optically pumped OPSLs.

OPSL Pumping Schemes

As mentioned previously, the best performing OPSLs are typically pumped by an external multimode fiber-coupled diode laser. The quality of the pump laser is not important nor is the pumping wavelength if the OPSL is designed for barrier pumping. In this scenario, the QWs are grown between pump absorbing barrier layers. For example, InGaAs QW can be surrounded by layers such as GaAs, AlGaAs or GaAsP to provide broad absorption features that are guaranteed to absorb over a broad spectral bandwidth. For GaAs-based materials, one can benefit from commercial pump modules emitting at 808 nm available at relatively low cost, which were initially developed for the pumping of solid state lasers. The disadvantage of barrier absorption is that the carrier confinement in the QWs imposes an energy difference between the pump and the lasing photons. This energy or wavelength difference between the shorter wavelength pump (808 nm) and the emitting wavelength (> 1000 nm) can be quite large leading to a large Stoke shift or quantum defect (Fig. 5(a)). This energy difference is wasted in the form of heat when the excited carriers relax into the wells via phonon interactions with the crystal lattice. This is one of the major sources of pump induced heat in the structure and also limits the achievable optical-to-optical efficiency of the device. Another source of heat comes from the incomplete absorption of the pump in the barriers, typically between 10 and 20% of the incident pump power, which will be absorbed in the DBR or in the bonding materials such as Ti or Cr at the back end of the DBR if it is transparent at the pump wavelength.

A second option is to directly pump the QW in the RPG stack (Fig. 5(b)). This strongly reduces the quantum defect, but it is at the expense of the absorption. The reduced interaction length (< 100 nm) between absorbing layers and the pump requires a recycling of the unabsorbed pump to increase the overall efficiency and to avoid unnecessary heating. This form of in-well pumping is analogous to a multi-pass pumping scheme employed with solid state thin disk lasers. While the QW absorption per pass is considerably stronger than the thin disk absorbing medium, it still requires numerous passes for efficient absorption (> 4) which is impractical to implement. The requirements on the pump laser wavelength and linewidth are also more stringent in this case. If the pump wavelength is too close to the gain peak, the absorption will decrease with the carrier density in the QW as the gain bandwidth gets broader and becomes more transparent at the pump wavelength. Nevertheless, in-well pump OPSL has been demonstrated with different material system with rather impressive power performance but much less than barrier pumping. It is also a good solution for OPSL emitting in the visible (red), where barrier pumping becomes impractical due to the lack of available pump sources.

Although not strictly characterized as OPSLs, it is important to mention that a structure can also be pumped electrically. Even though they are technologically more challenging, electrically pumped VECSELs (EP-VECSEL) have been demonstrated in the near

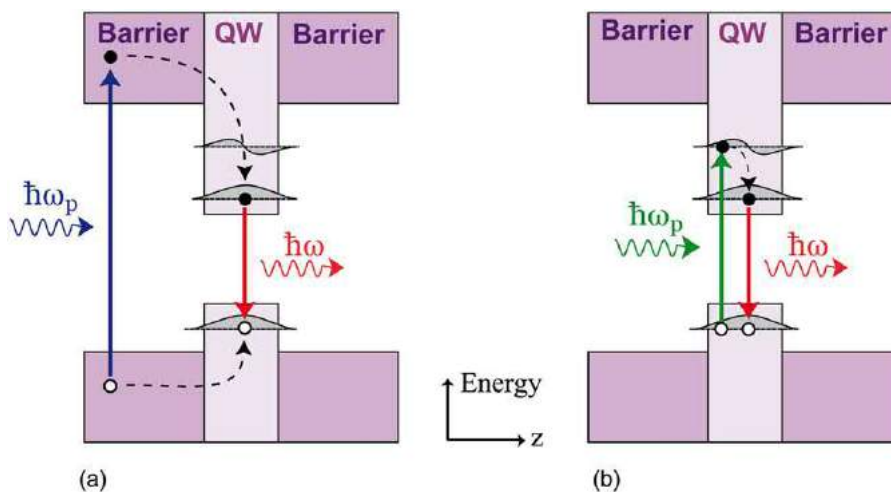


Fig. 5 Optical pumping scheme of OPSLs. (a) Barrier pumping: strong absorption, large quantum defect and (b) in-well pumping: weak absorption, low quantum defect.

IR and have exhibited impressive output powers in a TEM₀₀ beam. This technology was used to generate both CW operation and mode-locked picosecond pulse trains. Just like the OPSL, thermal management plays a crucial role in the performance of these lasers – electrically pumped systems suffer additionally from Joule heating whereas optically pumped systems produce heat via the quantum defect and incomplete pump absorption in the active region. EP-VECSELS have a gain structure similar to that of a VCSEL, where the active area is sandwiched between a circular electrode on the bottom (mesa) and a ring electrode on the top to permit light extraction. In order to compensate the increased optical losses from free carrier absorption in the required doped layers, the structure usually contain a partial DBR reflector on the top of the active region to boost the micro cavity resonance and increase the modal gain. Unfortunately, this technique cannot be power scaled easily, as the contact geometry does not allow a uniform injection of carriers in the active region with a large area. The weak lateral carrier diffusion in the center of the active area is a major limitation for the output power in a single transverse mode. So far, the powers achievable with such devices appear to be limited to the order of 4 W. The compactness of this technology is, however, very attractive for numerous applications and have been successfully commercialized by NECSELS, primarily for visible displays and laser lighting markets.

OPSL Cavity Configurations

The simplest cavity configuration is the linear cavity depicted in Fig. 1. Here, the OPSL chip with backing DBR is mounted on a CVD diamond heat spreader and the latter soldered to a copper heat sink. An air gap on the order of a few millimeters to tens of centimeter exists between the active OPSL chip and an external output curved mirror with transmission varying between typically 0.5 and 10% for the higher gain chips. The external pump beam is imaged on the OPSL chip at a finite angle, typically between 20 and 45 degrees. This simple cavity setup is best for initially testing OPSL chip performance and for scaling of power beyond tens of Watts. This external free space cavity is a stable multimode optical resonator as long as the cavity length is shorter than the radius of curvature of the concave mirror ($L < R_c$). This cold cavity supports a very large number of transverse modes and the length of the cavity and curved mirror will dictate the fundamental and higher modes waists. By finely adjusting the fundamental mode waist to the pumped area, it is possible to spatially filter all the higher order mode with the help of the gain/absorption profile. However, to stabilize large TEM₀₀ beams the cavity length required might be too long and the thermal lensing in the semiconductor can destabilize the output beam at high pump levels. In this case, it is preferable to use a more sophisticated cavity configuration, like a V-cavity where a large beam waist can be more easily obtained.

In a V-shaped cavity, the OPSL chip can be placed either at the vertex of the V (Fig. 6(a)) or as an end mirror (Fig. 6(b)), depending on the application targeted.

In a mode-locking setup, the SESAM absorber is typically placed at the end of the V-cavity and the other end mirror acts as an output coupler (Fig. 6(a)).

For SHG, the non-linear crystal might be placed between a flat and a curved mirror to ensure a double pass in the crystal before SHG light extraction through the curved mirror (Fig. 6(b)). There are many other variants of cavity geometries, a few are depicted in Fig. 6, including a Z-cavity, a ring cavity which offers further advantages over the V-cavity for modelocking. In some cases, it may be desirable to pump multiple OPSL chips for further power scaling within a single cavity, like in the W-shaped cavity of Fig. 6(f). It has been demonstrated that power scales close to linearly with number of OPSL chips. There are much more possible geometry which is not shown here, that might serve best a particular application.

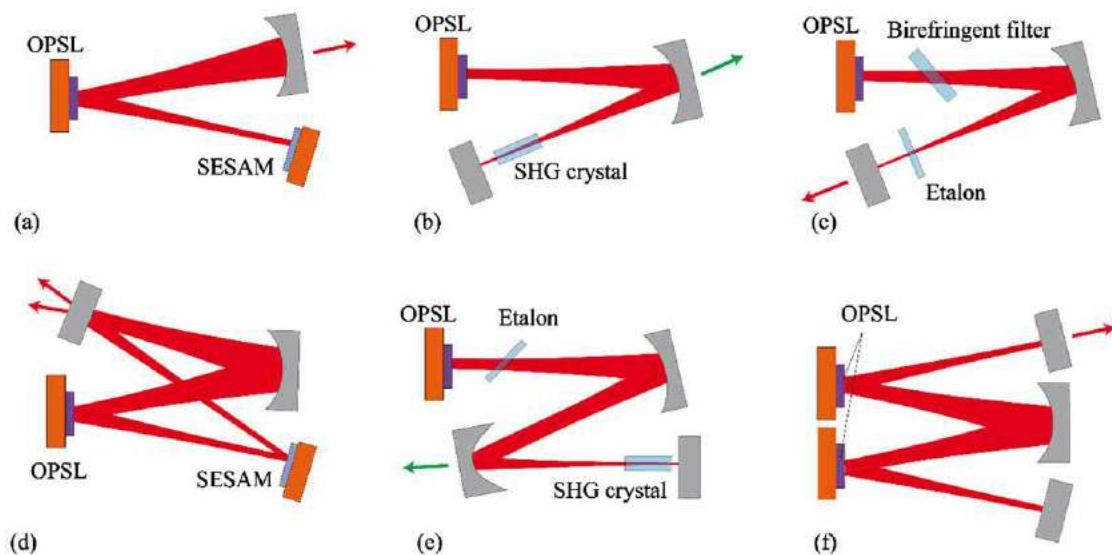


Fig. 6 Examples of VECSEL cavity geometries. (a) V-cavity for passive modelocking; (b) V-cavity for second harmonic generation; (c) V-cavity for single frequency operation; (d) folded ring cavity for colliding pulse modelocking; (e) Z-cavity for SHG; and (f) W-cavity for multi-chip power scaling.

High Power, Wavelength Tunable, High Brightness CW OPSLs

CW powers exceeding 100 W have already been demonstrated from a single OPSL chip. No attempt was made to enforce a single TEM₀₀ mode at these power levels and thermal management will offer the greatest challenge. These deliver the final circular beam with $M^2 < 10$ which is much less than what was demonstrated by comparable power diode bars. TEM₀₀ mode operation has been demonstrated up to powers of about 40 W and commercial OPSL intracavity SHG OPSL sources are available at powers up to 10 W. The broad gain bandwidth of the semiconductor QW allows for tuning over 30–40 nm albeit at lower multi-Watt powers.

The high brightness circular beam delivered by the OPSL chip offers many advantages over established high power broad area and diode bar emitters. Firstly, significantly improved brightness relative to these sources means that it can be imaged onto a smaller target spot or into a fiber – high power broad area edge emitters deliver a highly aberrated, highly divergent beam with an elliptical beam shape due to the rectangular shape and small vertical dimension of the emitting diode. Moreover, edge emitters are limited by the high power density at the emitted aperture leading to potential facet damage. The OPSL chip can maintain a much lower power density (power per unit area) as the emitter surface is much larger and power scaling is achieved by increasing the area of the pump and signal spot on the chip while maintaining a relatively constant power density.

High Power Narrow Linewidth CW OPSL Sources

Commercial single frequency edge emitters typically consist of a lower power single frequency DFB seed integrated with a flared planar amplifier (MOPA). These have been shown to perform up to 10 W but are extremely sensitive to any stray feedback or residual reflectivity at the end of the flare amplifier. The latter is present even with the best anti-reflection coatings. Moreover, they suffer from the same beam aberrations as their broad area high power counterparts.

Single mode OPSL sources with linewidths in the tens of kilohertz regime have been demonstrated with powers up to 23 W at a wavelength of around 1 μm . At this power level, single longitudinal mode operation needs to be enforced by inserting additional optical elements such as birefringent filters and etalon into a V-cavity. At lower power, it has been shown that single frequency operation is achievable with a very high spectral purity without any need for intra-cavity filter. This is accomplished thanks to the nearly ideal homogenous gain broadening of the OPSL combined to the high finesse cavity which strongly enforces the effect of the spectral gain curvature, and lead to high temporal coherence.

Fundamentally, the intensity and frequency noise of a single frequency laser is limited by the amount of spontaneous emission contained in the lasing mode. Because of the random nature of its phase, the spontaneous emission creates a fluctuating electrical field in time, which converts to a phase and frequency fluctuation of the output. In the high-Q cavity of a single frequency OPSL, the spontaneous emission is strongly filtered by the cold-cavity mode which gives rise to the emitted laser mode. For this reason, the fundamental limits in term of intensity and frequency noise are extremely low for an OPSL (sub-hertz linewidth). On the contrary, in most laser diodes the lasing mode propagates in an optical waveguide where the spontaneous emission is directly coupled and confined with the lasing mode and can be largely amplified along the cavity (amplified spontaneous emission (ASE)), which increases the fundamental limits by several orders of magnitude.

The current limitations in term of linewidth are however, not from the spontaneous emission but from technical contributions, such as mechanical vibrations and intensity noise from the pump source. These contributions have however, a spectral signature in a relatively low frequency range (< 500 kHz), giving the possibility of a noise stabilization for ultralow noise operation. It also means that OPSL can reach shot-noise limited operation over a very wide frequency range in the RF-microwave domain, in spite of poor quality multimode pumping, since the high-Q cavity will filter out the noise of the pump above the cold-cavity cutoff frequency. These OPSLs properties provide a frequency noise and linewidth much lower than other laser technologies, surpassing conventional diode-laser technologies by orders of magnitude.

Quantum Dot CW OPSL Sources

OPSLs incorporating QDs as a gain medium have also been demonstrated with CW output power exceeding 8 W. In most cases the QDs are grown using the Stranski–Krastanov growth method, in place of the QWs at the anti-nodes of the standing wave in the micro-cavity. These have the advantage of being wavelength adjustable by controlling the average dot size and the inevitable inhomogeneous size distribution of the dots provides a broad but lower gain relative to the QWs. InP QDs facilitate lasing in the visible around 600–700 nm whereas InAs QD VECSELs operate at longer wavelengths around 1 μm . Since QDs inherently exhibit strong inhomogeneous gain broadening, they may not be the medium of choice for single frequency operation, since the weaker mode coupling in the gain will likely lead to multimode operation. However, the increased gain bandwidth and expected spectral hole burning could provide an excellent platform for dual or multiple frequencies operation, to generate terahertz via DFG, for example. The increased gain bandwidth should also be able to support short pulse durations in a mode-locked state. Mode-locking of QD OPSLs has been reported with pulses of picosecond duration in the visible using InP and near-IR with gain layers based on self-assembled InAs QDs. Even though QDs have proven to be a suitable candidate for the gain medium of a CW or modelocked OPSL gain medium, it is not clear at this point if QDs are superior to QWs. The intrinsic low saturation fluence of QDs, which is an advantage for a SESAM, could be a major drawback for a modelocked gain medium where a strong saturation fluence is usually desired.

Low Noise Multi-Gigahertz Repetition Rate Femtosecond Pulsed OPSLs

OPSLs are emerging as new sources of low noise, multi-gigahertz repetition rate femtosecond duration mode-locked sources. The compact V-cavity, Z-cavity, or ring cavity configuration facilitates replacement of one of the cavity HR mirrors by a SESAM. The SESAM typically consists of a single QW whose absorption spectrum is matched to the central wavelength of the desired pulse. It is grown on a DBR to provide for high reflectivity when the QW absorption is saturated at high intra-cavity field intensities. SESAMs need to be designed to have as fast a carrier recombination rate as possible to ensure short pulses. This can be achieved by low temperature growth of the QW creating defects to cause rapid defect recombination or by growing the single quantum well in close vicinity to the semiconductor surface (few nanometer) in order to facilitate fast recombination with surface states.

The low gain and compact cavity of a mode-locked VECSEL creates challenges in regards to the implementation of dispersion management to minimize group delay dispersion for femtosecond pulse generation. Unlike solid state mode-locked lasers such as Ti:Sapphire, one cannot afford to add many internal dispersion compensating optical components due to their intrinsic loss even when AR-coated. Typically, dispersion control is implemented at the semiconductor epitaxial design phase and further complemented with AR-coatings on the VECSEL and SESAM chips. The modelocking of OPSLs has been demonstrated at repetition rates ranging from 85 MHz to 100 GHz, and at operating wavelength ranging from 665 to 1960 nm. To date, the best performance in terms of power and pulse duration has been demonstrated between 960 and 1080 nm, where the semiconductor media benefit from the maturity and advantages of GaAs-based materials. To date, the record peak power obtained in a modelocked state exceeds 6.3 kW with 410 fs pulses at a repetition rate of 390 MHz. The shortest pulse duration was demonstrated at around 1030 nm with pulses of 107 fs, externally compressed to 96 fs.

A compact variant of a mode-locked OPSL is the MIXSEL which involves integrating the saturable absorber into the same epitaxial structure as the gain element. The saturable absorber elements can be either a QW or a single quantum dot layer. This approach is ideally suited very high repetition rate as the cavity can be extremely compact. It appears however, that this approach is less flexible and leaves less room for growth error, which has limited the performance to lower power and longer pulses (sub-300 fs) than the external SESAM approach.

Intra-Cavity SHG and DFG OPSLs

OPSLs offer maximum flexibility in implementing various modes of intra-cavity SHG and DFG generation. The potential of achieving high kW-level internal circulating fields within the cavity guarantee high power signal generation at wavelengths spanning UV to terahertz. Frequency conversion via intra-cavity SHG has been extensively employed to generate a host of wavelengths in the visible and UV from an IR fundamental wavelength. Short UV wavelengths around 200 nm typically requires two SHG stages starting with IR wavelength originating from an InGaAs QW OPSLs frequency doubled in the OPSL cavity, the visible second harmonic output can be further converted to UV in an external resonator containing a BaB₂O₄ crystal, for example. This technique has been employed to generate stabilized single frequency output with a couple hundreds of mill watts in the UV.

OPSL are also promising for the development of higher-power terahertz sources via DFG. By designing the semiconductor structure to provide a flat broadband gain, one can exploit the unique carrier dynamics of QWs or QDs gain medium, to favor the emission of multiple wavelengths simultaneously. The spectral spacing of these multiple wavelengths can be enforced by a thin etalon in the cavity to allow DFG in the terahertz regime. The broad gain bandwidth and the etalon thickness selection and tilt can provide a large and fine tunability of the terahertz emission, from 100 GHz to several terahertz. The stability and gain competition dynamics of multiple wavelength OPSLs have been investigated and have shown that with high intracavity powers in high-Q cavities, a relatively stable simultaneous two-color operation is feasible, probably due to strong kinetic hole burning effects and/or spatial hole burning. Moreover, the circular nearly diffraction-limited beam profile is ideal for driving nonlinear processes with high efficiency. However, one has yet to investigate the influence of a nonlinear crystal, such as a periodically poled lithium niobate (PPLN) crystal, on the dynamics of an OPSL operating in a multi-color regime. A DFG in an external resonant cavity would be an alternative to prevent against instabilities caused by the frequency conversion in the crystal.

Applications

Current commercial applications of OPSLs are primarily based on intra-cavity SHG based on the InGaAs/AlGaAs/GaAs material system. These lasers emit in the near-IR and can generate impressive power in the visible and UV wavelength range. Coherent's commercial OPSL based products include OBIS Lasers for life sciences applications, from the UV (375 nm) to the near IR (785 nm); OBIS LG Lasers offer unparalleled OPSL laser performance in the smallest form factor for plug-and-play capability; Sapphire Lasers: CW Lasers available at multiple wavelengths from 458 to 594 nm and ultra-narrow linewidth versions; Verdi G Series family of OPSL lasers with powers up to 20 W at 532 nm and Tracer a portable and lightweight green forensic laser system. The commercial EP-VECSEL implemented by NEXSEL provides a host of single and multiple wavelengths SHG generated sources covering the RGB spectrum mainly for commercial lighting applications. In recent years, M Squared has been commercializing modelocked OPSLs under the name of Dragonfly to replace the more conventional titanium sapphire laser. The Dragonfly series is a simple turnkey laser system providing source of picosecond mode-locked pulses with excellent beam quality near the 1 μ m region. Innoptics is another emerging company developing a compact commercial VECSEL module capable to generate a low noise continuously tunable single frequency output in the near IR around 1 μ m (OPSilent series) and in the mid-infrared around

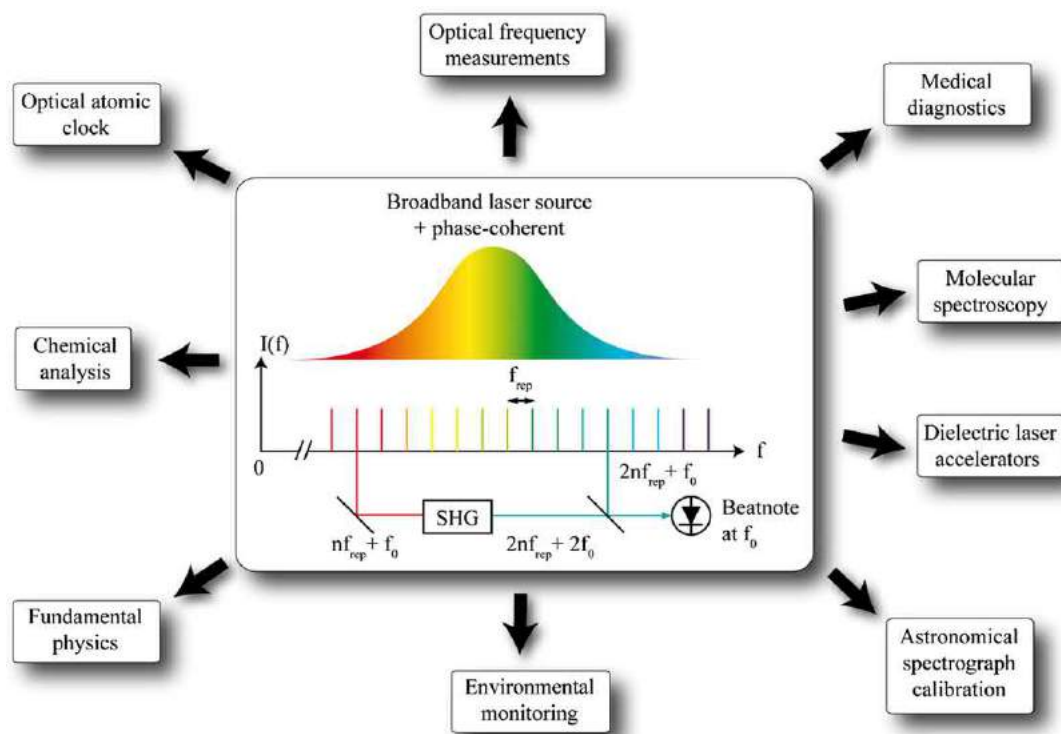


Fig. 7 Principle of a self-referenced frequency comb and potential applications.

2.3 μm (OPScan series). These wavelengths are interesting for gas analysis as they benefit from an atmospheric transparency window and from the presence of numerous gas signatures (methane, carbon monoxide, etc.).

However, OPSLs remain a nascent technology with a plethora of potential applications. Laboratory demonstrations in the CW single frequency regime have shown tremendous performance in terms of coherence, tunability and power, which are extremely interesting for applications such as remote guidestar (at a wavelength of 1141 or 1178 nm doubled to 589 nm), high resolution spectroscopy, remote sensing, radar/lidar, optical metrology (frequency reference, gyroscope), or atom trapping to cite a few.

The high finesse external cavity of OPSLs give access to an intense optical field and prolonged interaction with an intracavity medium, like a nonlinear crystal for efficient SHG conversion or DFG for coherent terahertz emission, or like a dilute gas for intracavity laser-absorption spectroscopy (ICLAS) to improve the sensitivity of in-situ trace-gas analyzers.

When modelocked, OPSLs can provide a high peak power at high repetition rate, which is advantageous for nonlinear applications such as multiphoton imaging, tomography, surgery or material processing. Ultrashort pulse generation from an OPSL could also find applications in high-resolution time-domain terahertz spectroscopy with asynchronous optical sampling, ultrafast communication systems or to generate a low noise self-referenced gigahertz frequency combs. The realization of OPSLs-based frequency comb would open up a lot more applications that are summarized in Fig. 7.

Finally, OPSL cavities can be tailored to generate more exotic transverse mode and wavefront such as a vortex mode carrying an orbital angular momentum, that find applications in optical handling of microscopic particles, atoms manipulation, sub-diffraction limit microscopy, material processing, and quantum telecommunication.

Disclaimer

This material is based upon work supported under Air Force Contract No. FA9550-14-1-0062. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force.

Further Reading

- Hader, J., Scheller, M., Laurain, A., *et al.*, 2017. Ultrafast non-equilibrium carrier dynamics in semiconductor laser mode-locking. *Semiconductor Science and Technology* 32, 013002.
- Moloney, J.V., Hader, J., Koch, S.W., 2007. Quantum design of semiconductor active materials: Laser and amplifier applications. *Laser & Photonics Reviews* 1, 24–43.
- Okhotnikov, O.G., 2010. *Semiconductor Disk Lasers*. New York, NY: John Wiley & Sons.

Relevant Website

www.nlcstr.com

Nonlinear Control Strategies Inc. (SimuLase™).

Optical Parametric Amplifiers

Giulio Cerullo, Sandro De Silvestri, and Cristian Manzoni, Institute of Photonics and Nanotechnology, Milano, Italy

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Thanks to spectacular technological progress over the last three decades, nowadays sources of ultrashort (femtosecond and picosecond) laser pulses have become reliable, rugged and power scalable, and are used in many fields, ranging from fundamental studies in physics, chemistry and biology to industrial and biomedical applications. Such sources are typically based on solid state gain media, such as Ti:sapphire or Nd:Yb-doped crystals, which typically emit at specific wavelengths (800 nm for Ti:sapphire and $\approx 1 \mu\text{m}$ for Yb and Nd) with very limited tunability.

On the other hand, many applications require tunability of the ultrashort pulses, over a broad range from the ultraviolet (UV) to the mid-infrared (mid-IR). As there are no classical laser active media, based on population inversion, capable of providing gain over such a broad frequency range, frequency tunability must be achieved by a nonlinear optical effect, exploiting the very high peak powers produced by ultrashort pulse laser sources. In most cases, frequency tunability is achieved by the second order nonlinear optical effect known as optical parametric amplification (OPA) (Giordmaine and Miller, 1965; Baumgartner and Byer, 1979). The principle of OPA is quite simple (Fig. 1(a)): in a suitable nonlinear crystal, energy is transferred from an high frequency and high intensity beam (the pump beam, at frequency ω_3) to a lower frequency, lower intensity beam (the signal beam, at frequency ω_1) which is thus amplified; in addition a third beam (the idler beam, at frequency ω_2) is generated, to fulfill energy conservation. The OPA process can have a simple corpuscular interpretation (see Fig. 1(b)): a photon at frequency ω_3 is absorbed by a virtual level of the material and a photon at frequency ω_1 stimulates the emission of two photons at frequencies ω_1 and ω_2 . In this interaction, energy conservation:

$$\hbar\omega_3 = \hbar\omega_1 + \hbar\omega_2 \quad (1)$$

is fulfilled. The signal frequency to be amplified can vary in principle from 0 to ω_3 , and correspondingly the idler varies from ω_3 to 0 (as a matter of fact, the lowest frequency is limited by absorption of the nonlinear crystal). The so-called degeneracy condition is achieved when $\omega_1 = \omega_2 = \omega_3/2$. As will be discussed in the next section, the OPA process is efficient when, in addition to energy conservation, the phase matching condition

$$\Delta k = k_3 - k_1 - k_2 = 0 \quad (2)$$

is satisfied, where k_i are the wave vectors of the interacting beams.

In summary, the OPA process transfers energy from a high-power, fixed frequency pump beam to a low-power, variable frequency signal beam, generating an idler beam to satisfy energy conservation. The OPA can thus be seen as a “photon cutter,” cutting a high energy photon $\hbar\omega_3$ into the sum of two lower energy photons $\hbar\omega_1$ and $\hbar\omega_2$. The OPA process thus provides an optical amplifier with continuously variable center frequency (determined by the phase-matching condition) and represents an easy way of tuning over a broad range the carrier frequency of an otherwise fixed wavelength laser system. On the other hand, if suitably designed, an OPA can simultaneously fulfill the phase matching condition for a very broad range of signal frequencies. The OPA thus acts as a broadband amplifier, efficiently transferring energy from a narrowband pump pulse to a broadband signal (idler) pulse; it can therefore be used to dramatically shorten, by more than an order of magnitude, the duration of the pump pulse. The concept of broadband OPA is very flexible and can be applied to produce few-optical-cycle pulses over a very wide range of frequencies, provided that a broadband yet weak seed is available and the proper phase-matching conditions are identified. This article is a self-consistent review of OPAs and is organized as follows: Section Theory of Optical Parametric Amplification introduces the theory of OPA; Section Phase Matching discusses the concept of phase matching; Section Architecture of an OPA presents the general architecture of an OPA and separately analyzes the different components; Section Ultra-Broadband OPAs describes ultra-broadband OPAs; finally, Section Optical Parametric Chirped pulse Amplification introduces the concept of the optical parametric chirped pulse amplifier (OPCPA).

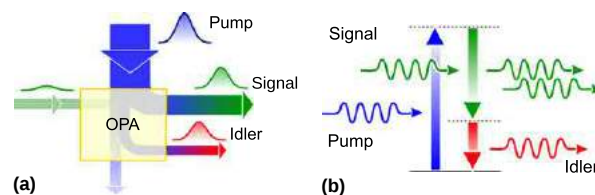


Fig. 1 (a) General scheme of the optical parametric amplifier (OPA) process and (b) corpuscular interpretation of the OPA.

Theory of OPA

Let us consider an optical field, consisting of the superposition of three monochromatic waves, at frequencies ω_1 , ω_2 , and ω_3 :

$$E(z, t) = \frac{1}{2} \{A_1(z) \exp[i(\omega_1 t - k_1 z)] + A_2(z) \exp[i(\omega_2 t - k_2 z)] + A_3(z) \exp[i(\omega_3 t - k_3 z)] + c.c.\} \quad (3)$$

satisfying the condition $\omega_1 + \omega_2 = \omega_3$, impinging on a non-centrosymmetric medium generating a second order nonlinear polarization:

$$P_{NL}(z, t) = \epsilon_0 \chi^{(2)} E^2(z, t) \quad (4)$$

Such a situation is known as “nonlinear second-order parametric interaction” and corresponds to an exchange of energy between the three fields mediated by the second order nonlinearity. The nonlinear polarization contains three components at frequencies ω_1 , ω_2 and ω_3 , given by

$$P_{1NL}(z, t) = \frac{\epsilon_0 \chi^{(2)}}{2} A_2^* A_3 \exp\{i[\omega_1 t - (k_3 - k_2)z]\} + c.c. \quad (5a)$$

$$P_{2NL}(z, t) = \frac{\epsilon_0 \chi^{(2)}}{2} A_1^* A_3 \exp\{i[\omega_2 t - (k_3 - k_1)z]\} + c.c. \quad (5b)$$

$$P_{3NL}(z, t) = \frac{\epsilon_0 \chi^{(2)}}{2} A_1 A_2 \exp\{i[\omega_3 t - (k_1 + k_2)z]\} + c.c. \quad (5c)$$

Obviously there are other terms on P_{NL} at different frequencies, such as, for example, $2\omega_1$, $2\omega_2$, $\omega_1 - \omega_2$... Here we consider only the terms at ω_1 , ω_2 and ω_3 because we assume that only the interaction between these three fields is efficient, due to the phase-matching condition.

The nonlinear polarization at frequency ω_j can thus be written as:

$$P_{NLj}(z, t) = \frac{1}{2} p_{NLj}(z) \exp\{i[\omega_j t - k_{pj}z]\} + c.c. \quad (6)$$

where we emphasize that the wave vector of the polarization k_{pj} is in general different from that of the corresponding field. The propagation equation for each field in the nonlinear medium can be written, in the monochromatic limit, as:

$$\frac{\partial A_j}{\partial z} = -i \frac{\mu_0 \omega_j c}{2n_j} p_{NLj} \exp[-i\Delta k_j z] \quad (7)$$

where $\Delta k_j = k_{pj} - k_j$ is the so-called “wave-vector mismatch” between the nonlinear polarization and the field. By inserting Eq. (5) into Eq. (7), we can derive the following three equations for the fields at ω_1 , ω_2 and ω_3 (Boyd, 2003):

$$\frac{\partial A_1}{\partial z} = -i\kappa_1 A_2^* A_3 \exp[-i\Delta k z] \quad (8a)$$

$$\frac{\partial A_2}{\partial z} = -i\kappa_2 A_1^* A_3 \exp[-i\Delta k z] \quad (8b)$$

$$\frac{\partial A_3}{\partial z} = -i\kappa_3 A_1 A_2 \exp[i\Delta k z] \quad (8c)$$

where we have defined the nonlinear coupling coefficients: $\kappa_i = \frac{\omega_i \chi^{(2)}}{4cn_i}$ and the “wave-vector mismatch” as: $\Delta k = k_3 - k_1 - k_2$. Note that the first two equations are formally identical, indicating that the fields at ω_1 and ω_2 play the same role. According to the boundary initial conditions $A_i(0)$ ($i=1, 2, 3$), two main nonlinear processes can arise: sum frequency generation (SFG) and difference frequency generation (DFG). In SFG the input fields are $A_1(0)$ and $A_2(0)$ at ω_1 and ω_2 ; their interaction produces the field A_3 at frequency $\omega_3 = \omega_1 + \omega_2$. Second harmonic generation (SHG) is just a particular case of SFG in which $\omega_1 = \omega_2$. In DFG, the input fields are $A_3(0)$ at ω_3 and $A_1(0)$ at ω_1 ; the field A_3 loses energy in favor of A_1 and of the newly generated field A_2 at $\omega_2 = \omega_3 - \omega_1$. The OPA is a mechanism similar to DFG, except for the strength of the interacting fields: DFG arises when the fields at ω_3 and ω_1 have comparable intensities, while for an OPA the field at ω_1 is much weaker. In the OPA process, therefore, an intense beam at ω_3 (the pump) transfers energy to the beam at ω_1 (the signal), thereby amplifying it, and generates light at ω_2 (the idler). Note that the DFG/OPA processes also occur starting from the fields $A_3(0)$ at ω_3 and $A_2(0)$ at ω_2 ; in this case the newly generated field is at ω_1 . This means that the signal and idler play an interchangeable role in the OPA process.

Eq. (8) can be solved by making the assumption that the pump field is not depleted during the interaction ($A_3 \approx \text{const.}$), and that there is no initial idler beam ($A_{20} = 0$). After proper manipulation, the signal evolution along the nonlinear medium can be written as:

$$\frac{d^2 A_1}{dz^2} = -i\Delta k \frac{dA_1}{dz} + \Gamma^2 A_1 \quad (9)$$

where $\Gamma^2 = \frac{2d_{\text{eff}}^2 \omega_1 \omega_2}{c_0^3 \epsilon_0 n_1 n_2 n_3} I_3$ and $I_3 = \frac{1}{2} n_3 c \epsilon_0 |A_3|^2$ is the pump beam intensity. Γ is known as the small signal gain of the OPA and $d_{\text{eff}} = \frac{\chi^{(2)}}{2}$ is the element of the $\chi^{(2)}$ tensor that keeps trace of the polarization of the beams and their propagation directions within the nonlinear crystal lattice.

In the hypothesis of an initial signal field amplitude $A_1(0)=A_{10}$ (the “seed beam”) and no initial idler ($A_2(0)=0$), the solutions of Eq. (9) are:

$$I_1(z) = I_{10} \left\{ 1 + \left[\frac{\Gamma}{g} \sinh(gz) \right]^2 \right\} \quad (10a)$$

$$I_2(z) = I_{10} \frac{\omega_2}{\omega_1} \left[\frac{\Gamma}{g} \sinh(gz) \right]^2 \quad (10b)$$

where $g = \sqrt{\Gamma^2 - \left(\frac{\Delta k}{2}\right)^2}$. For the case of large gain ($gz \gg 1$) Eq. (10) further simplify to:

$$I_1(z) = \frac{I_{10}}{4} \left(\frac{\Gamma}{g} \right)^2 \exp(2gz) \quad (11a)$$

$$I_2(z) = \frac{\omega_2}{\omega_1} I_1(z) \quad (11b)$$

giving an exponential growth of both signal and idler intensities with crystal length, characteristic of an optical amplifier. It should also be noted, that, in the large gain limit, Eq. (11b) can be rewritten as:

$$\frac{I_2(z)}{\hbar\omega_2} = \frac{I_1(z)}{\hbar\omega_1} \quad (12)$$

which means that, in the large gain limit, signal and idler intensities are related by energy conservation, since for each annihilated pump photon a signal and an idler photon are simultaneously generated.

The parametric gain for the signal beam, in the large gain limit, can be written as:

$$G = \frac{I_1(z)}{I_{s0}} = \frac{1}{4} \left(\frac{\Gamma}{g} \right)^2 \exp(2gz) \quad (13)$$

which, for perfect phase matching ($\Delta k=0$) becomes:

$$G = \frac{1}{4} \exp(2\Gamma z) \quad (14)$$

The exponential growth of the gain with propagation length is qualitatively different from the quadratic growth that occurs in other second-order processes like SFG and SHG and shows that the OPA behaves like a real amplifier. With respect to a classical optical amplifier based on population inversion in an atomic or molecular transition, however, an OPA has three important differences:

1. it does not have any energy storage capability, i.e., the gain is present only during the temporal overlap of the interacting pulses;
2. the gain center frequency is not fixed, but can be continuously adjusted by varying the phase-matching condition;
3. as we will see in Section Phase Matching Bandwidth of an OPA, the gain bandwidth is not limited by the linewidth of an electronic/vibrational transition, but rather by the possibility of satisfying the phase-matching condition over a broad range of frequencies.

Let us now consider the factors influencing the parametric gain G :

1. $G \propto \exp(g)$ exponentially depends on the parameter g , which is maximum when $\Delta k=0$ (phase-matching condition). G rapidly decreases for nonzero values of Δk , suggesting that phase-matching is a key condition to be fulfilled in order to get significant amplification from the OPA process.
2. $G \propto \exp(d_{\text{eff}})$ depends exponentially on the second order nonlinear optical coefficient of the crystal d_{eff} ; one should therefore select the crystal with the largest nonlinear response. There are however other considerations leading to the choice of the crystal, such as phase matching range, dispersive properties, and optical damage threshold.
3. $G \propto \exp(\sqrt{I_3})$ scales as the exponential of the square root of the pump intensity. This indicates the suitability of ultrashort pulses for OPAs, due to their high peak powers. One should try to use the highest possible pump intensity before the onset of other nonlinear optical phenomena such as self-focusing, self-phase modulation and beam breakup. In order to be able to use high pump intensities, it is however important to have a spatially clean beam profile, without hot spots.
4. $G \propto \exp(L)$ scales as the exponential of the crystal length L , as in an optical amplifier. With ultrashort light pulses, however, the optimum crystal length has to be chosen considering the durations and group velocities of the interacting pulses.
5. $G \propto \exp(\sqrt{\omega_1\omega_2})$ scales as the exponential of the square root of the product of signal and idler frequencies. This seems to indicate an advantage to use higher pump frequencies. However this advantage is often offset by the larger difference in group velocities of the interacting pulses.

Phase Matching

Interaction Types in an OPA

Eq. (8) describes the nonlinear interaction between pump, signal and idler waves, assuming the same propagation direction (collinear interaction). In general, the interaction involves fields with different polarizations and different propagation directions; a complete description of the most general case goes beyond the purpose of this article. To take polarization of the fields into account, it is often sufficient to reduce the nonlinear susceptibility tensor to the scalar coefficient d_{eff} , which accounts for the average interaction strength among the fields. For the sake of clarity, we will limit ourselves here to the simpler, and more commonly adopted, uniaxial birefringent crystals; the reader is referred to (Dmitriev *et al.*, 1999) for a complete treatment of the case of biaxial crystals. In uniaxial crystals, the best reference frame to describe the electric field of an electromagnetic wave is defined by the optical axis of the medium and the wave-vector $\mathbf{k}$ of the propagating beam, which form a plane σ and two normal polarization directions, called ordinary (o) and extraordinary (e). According to the polarization configurations of the three interacting fields, we group them in the following interaction types:

Type 0: the three interacting waves have equal polarization states, either ordinary (ooo) or extraordinary (eee).

Type I: the low-frequency fields (ω_1, ω_2) propagate with the same polarization, either ordinary or extraordinary. The resulting configurations are therefore ooe or eeo, respectively.

Type II: the low-frequency fields are cross-polarized, and the possible configurations are eoe, oee, eeo, and oeo.

From the polarization configuration of the fields it is possible to evaluate the two important parameters governing the nonlinear interaction: the nonlinear effective coefficient d_{eff} , which accounts for the second-order nonlinear response of the medium, and the frequency-dependent phase-mismatch Δk , which determines the efficiency of the energy flow between the interacting fields.

Phase Matching Methods

Eq. (10) shows that the nonlinear parametric interaction is optimized when $\Delta k=0$, at perfect phase matching. In a collinear configuration, recalling that $k=\omega n/c_0$, the phase-matching condition is equivalent to

$$n_3\omega_3 - n_2\omega_2 - n_1\omega_1 = 0 \quad (15)$$

which links the refractive indexes n_i seen by the interacting waves. In general, it can be shown that it is not possible to fulfill this condition in an isotropic medium, due to material dispersion. To achieve phase matching, two approaches can be followed:

1. exploit birefringence of the non-centrosymmetric nonlinear crystal by propagating the three interacting beams with different polarizations, leading to Type-I and Type-II interaction schemes (see Fig. 2(a));
2. apply a periodic modulation to the sign of the nonlinear coefficient d_{eff} , leading to the so-called quasi-phase-matching (QPM) regime (Hum and Fejer, 2007) (see Fig. 2(c)); in this case the fields can propagate with parallel polarization (Type 0). Phase matching is not locally satisfied, but the modulation of the sign of d_{eff} leads to an average macroscopic net exchange of energy between the fields.

To study phase matching by birefringence, for simplicity we will again limit ourselves to uniaxial crystals. These crystals have two frequency-dependent refractive indexes: the ordinary index n_o and the extraordinary index n_e ; if $n_o < n_e$, the crystal is called positive uniaxial crystal, whereas if $n_o > n_e$ it is called negative uniaxial crystal. The ordinary polarization propagates with index of refraction n_o , while the extraordinary polarization experiences an index of refraction which varies from n_o to n_e as a function of the

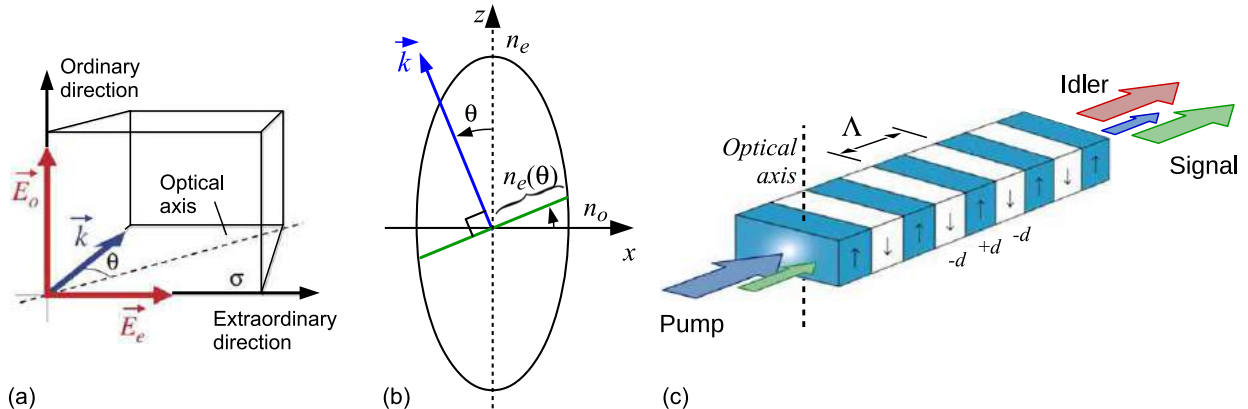


Fig. 2 (a) Ordinary and extraordinary directions in uniaxial birefringent crystals. (b) Section of the index ellipsoid and graphical meaning of ordinary and extraordinary refractive indices, for a positive uniaxial crystal. The optical axis is in the z direction. (c) Sketch of a periodically-poled crystal and the propagating fields of an optical parametric amplifier (OPA).

angle θ between $\mathbf{k}$ and the optical axis, according to the equation:

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2(\theta)}{n_o^2} + \frac{\sin^2(\theta)}{n_e^2} \quad (16)$$

By changing the propagation direction inside the crystal, $n_e(\theta)$ can thus be tuned from n_o to n_e (see Fig. 2(b)): in this way, there often exists one value of θ giving the right n_e which fulfills the phase-matching condition of Eq. (15). This angle is indicated with θ_m and is called phase-matching angle. If only one field has extraordinary polarization, θ_m can be exactly calculated from Eqs. (15) and (16), while the calculation is more involved if two fields have extraordinary polarization.

The QPM approach to phase matching is typically used with those nonlinear crystals whose $\chi^{(2)}$ tensor is such that the highest nonlinear coefficient manifests when the three interacting beams have the same polarizations, as for stoichiometric lithium tantalate (SLT) or lithium niobate (LiNbO_3 , LN). In these crystals the strongest element of the nonlinear tensor is d_{33} , which contributes to the nonlinear process when the three interacting beams share the same extraordinary polarization (Type 0 configuration). In this case birefringence phase-matching can never be achieved, as discussed above. In QPM the nonlinear properties of the medium are spatially modulated by periodically changing the sign of the $\chi^{(2)}$ coefficient, reversing the alignment of the ferroelectric domains in the birefringent crystal upon application of a suitable voltage. When the modulation of the nonlinear coefficient is periodic with grating period Λ , the phase-matching condition now includes the grating wave-vector $K_g = 2\pi/\Lambda$ and reads:

$$\Delta k = k_3 - k_1 - k_2 - K_g = 0 \quad (17)$$

Finding the phase-matching condition, for the case of QPM, corresponds to calculating the poling period Λ_m .

Phase Matching Bandwidth of an OPA

Let us now discuss which conditions determine the gain bandwidth of an OPA. Ideally one would like to have a broadband amplifier, i.e., an amplifier which, for a fixed pump frequency $\bar{\omega}_3$, provides a more or less constant gain over an as broad as possible range of signal frequencies. To this end, one needs to keep the phase mismatch Δk as small as possible over a large bandwidth. Practically, however, the phase-matching condition can be satisfied only for a given set of frequencies $(\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3)$, such that

$$\Delta k = k(\bar{\omega}_3) - k(\bar{\omega}_1) - k(\bar{\omega}_2) = 0 \quad (18)$$

If the pump frequency is fixed at $\bar{\omega}_3$ and the signal frequency changes to $\bar{\omega}_1 + \Delta\omega$, then by energy conservation the idler frequency changes to $\bar{\omega}_2 - \Delta\omega$. The ensuing wave vector mismatch becomes

$$\Delta k = k(\bar{\omega}_3) - [k(\bar{\omega}_1) + \Delta k_1] - [k(\bar{\omega}_2) + \Delta k_2] = -\Delta k_1 - \Delta k_2 \quad (19)$$

which can be approximated to the first order as

$$\Delta k = -\frac{\partial k_1}{\partial \omega_1} \Delta\omega + \frac{\partial k_2}{\partial \omega_2} \Delta\omega = \delta_{12} \Delta\omega \quad (20)$$

where $\delta_{12} = \frac{1}{v_{g2}} - \frac{1}{v_{g1}}$ is the group velocity mismatch (GVM) between signal and idler pulses. The full width at half maximum parametric gain bandwidth for a crystal of length L can then be calculated from Eq. (13), within the large gain and low pump-depletion approximations, as:

$$\Delta\omega = \frac{4\log(2)^{1/2}}{\pi} \left(\frac{\Gamma}{L}\right)^{1/2} \frac{1}{|\delta_{12}|} \quad (21)$$

Eq. (21) shows that the gain bandwidth is inversely proportional to the GVM between signal and idler and has only a square root dependence on small-signal gain Γ and crystal length L . For the case when $v_{g1} = v_{g2}$ (group-velocity matched OPA), Eq. (21) loses validity and Eq. (20) must be expanded to the second order in $\Delta\omega$, giving

$$\Delta\omega = \frac{4\log(2)^{1/4}}{\pi} \left(\frac{\Gamma}{L}\right)^{1/4} \frac{1}{\left|\frac{\partial^2 k_1}{\partial \omega_1^2} + \frac{\partial^2 k_2}{\partial \omega_2^2}\right|^{1/2}} \quad (22)$$

where $\frac{\partial^2 k}{\partial \omega^2}$ is the group velocity dispersion (GVD) of the pulses. In this case the gain bandwidth is thus inversely proportional to the square root of the sum of the GVDs of signal and idler pulses. The conditions for group-velocity matched OPAs will be discussed in a later section.

Architecture of an OPA

General OPA Scheme

The conceptual scheme of an OPA pumped by ultrashort pulses is summarized in Fig. 3. It fundamentally consists of three subsystems (Cerullo and De Silvestri, 2003):

1. A seed pulse generation stage, which exploits a nonlinear optical process to generate, starting from the pump pulse, from its second harmonic (SH) or from its third harmonic (TH), a weak pulse at the signal frequency that initiates the OPA process.

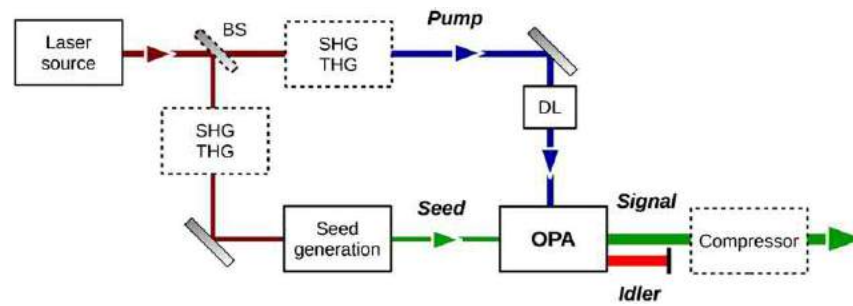


Fig. 3 General scheme of an optical parametric amplifier (OPA). BS: beam splitter; DL: delay line; SHG, THG: second- and third-harmonic generation stages.

2. Parametric amplification in one or more gain stages, pumped either by the fundamental wavelength (FW) of the driving laser or by its SH/TH.
3. An optional pulse compression stage by a dispersive delay line, to obtain the minimum pulse duration compatible with its bandwidth, the so-called transform-limited (TL) pulse duration.

Analyzing the operation of these three blocks provides the key for understanding the properties of the OPA and in particular the design required to achieve specific working parameters.

Seed Generation

In standard OPA systems, the two most widely adopted schemes for producing the seed pulse are optical parametric generation (OPG) and white-light continuum (WLC) generation. OPG, also known as parametric superfluorescence (PSF), is parametric amplification of the vacuum or quantum noise, which can also be thought of as two-photon spontaneous emission from a virtual level excited by the pump field (Harris *et al.*, 1967). In practice it is simply achieved by pumping a suitable second order nonlinear crystal, which is often of the same Type as the ones used in the subsequent OPA stages; OPG occurs at those frequencies for which the parametric interaction is phase-matched. Since PSF is a stochastic process that starts from noise, it has the disadvantages of an inherently large shot-to-shot energy fluctuations of the seed and a poor spatial beam quality. In addition, the longitudinal position inside the nonlinear crystal at which amplification takes place, and thus the absolute timing of the seed pulse with respect to the pump, varies stochastically in time. The main advantage of OPG is the relatively modest pulse energy requirement, especially using crystals with very high effective nonlinearity such as periodically poled LN and SLT. This feature is particularly useful in OPAs which are driven by high-repetition-rate, low-pulse-energy oscillators. One final remark about OPG is that it often occurs as a parasitic effect, as it can be generated in the initial stage(s) of the OPA, in the absence of a proper seed, and then amplified in the subsequent stages, thus spoiling the properties (energy stability, spectral purity) of the amplified pulses.

WLC generation (Gaeta, 2000) is achieved by focusing the ultrashort pulse in a suitable transparent dielectric medium (typically a sapphire plate) displaying a third-order ($\chi^{(3)}$) nonlinearity. WLC is a complicated nonlinear optical process, involving the interplay and coupling of temporal (self-phase modulation, dispersion, self-steepening) and spatial (self-focusing, diffraction) effects, as well as ionization and plasma generation, leading to the formation of a filament and to dramatic spectral broadening. WLC generation is performed by focusing a fraction of the driving pulse (typically the FF off the laser, but sometimes also the SH or the TH) into the plate (with thickness from 1 to 10 mm); by adjusting the pulse energy (using a variable attenuator), the focusing conditions (by moving the plate in the focus of the beam) and the numerical aperture of the focused beam (often by a diaphragm placed in front of the focusing lens) a single filament WLC is formed. Typical spectral energy density of the WLC is 10–30 pJ/nm in the visible range. The WLC displays excellent shot-to-shot stability and diffraction-limited spatial beam quality, so that it is the ideal seed for an OPA, although with limited spectral energy density. Using 800 nm Ti:sapphire driving lasers, with pulse duration of 100 fs or shorter, the best medium for WLC generation is undoped sapphire which, due to its very high thermal conductivity and damage threshold, guarantees excellent stability and damage-free operation. Using 1030 nm Yb-based driving lasers, which generate longer pulses, sapphire is not a suitable material for WLC generation. Other media with lower bandgap, such as undoped YAG, allow the generation of stable WLC under these conditions Bradler *et al.* (2009). The WLC provides a very broadband seed; the duration of the OPA pulses will depend on the fraction of this broad bandwidth which can be amplified, i.e., on the gain bandwidth of the OPA crystal(s), which has been discussed in Section Phase Matching Bandwidth of an OPA.

Parametric Amplification

In the parametric amplification stage(s) energy is transferred from the pump to the seed pulses. In order to optimize this process, one should carefully adjust the pump-pulse (wavelength, energy, and duration) and the nonlinear interaction (crystal type, crystal length, and phase-matching Type) parameters.

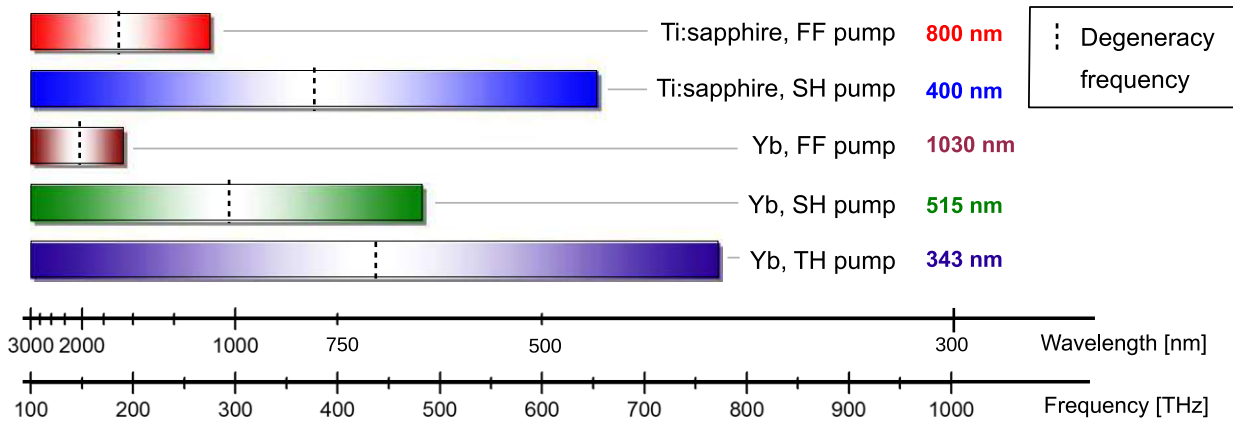


Fig. 4 Map of tuning ranges of an optical parametric amplifier (OPA) based on β -barium borate (BBO), whose infrared transparency is limited to 3 μm . The degeneracy point is at half the pump frequency.

Typically gain in an OPA is achieved in multiple stages, with the number dictated by the final pulse energy: a preamplifier (or signal amplifier, 1st stage) starting from the low-energy seed beam and displaying high gain (from 10^3 to 10^5) and one (2nd stage) or more power amplifiers, displaying relatively low gain (from 10 to 10^2). The multistage approach has several advantages:

1. it allows to compensate for the GVM arising between pump and signal pulses within each stage;
2. it enables to adjust the pump intensity, and thus the parametric gain, separately in the different stages. Most of the available pump energy should be used for the final stage, which is driven into saturation in order to efficiently convert the available pump energy and minimize shot-to-shot energy fluctuations of the amplified signal. The spot size of the beams progressively increases for the subsequent stages, in order to accommodate higher pulse energies while keeping comparable peak intensities.

Several criteria influence the choice of the pump wavelength for an OPA. The most important is the required tuning range, which depends on the pump wavelength, on the possibility to satisfy the phase matching condition and on the transparency range of the nonlinear crystal. Absorption is one of the most important factors limiting the tuning of an OPA; care must be taken in order to have both signal and idler frequencies in the transparency range of the nonlinear crystal(s).

Fig. 4 summarizes the OPA tuning ranges for the popular β -barium borate (BBO) crystal, when pumped by the FF and the SH of Ti:sapphire and by the FF/SH/TH of Yb. For this crystal and these pumping conditions wavelength tunability is between 390 nm and $\approx 3 \mu\text{m}$, limited by the pump pulse wavelength and by the onset of IR absorption in BBO at wavelengths longer than 3 μm . Of course other frequency ranges can be covered by nonlinear frequency conversion of the OPA output, for example, by SHG of or by SFG with the FF or the SH of the driving pulses. In particular, the mid-IR region (3–20 μm), which is very important for vibrational spectroscopy, can be covered by DFG between the signal and the idler of a 800 nm pumped OPA. By inspecting **Fig. 4**, one notices that for FF-pumped OPAs signal/idler tunability is limited to the IR. By frequency doubling an FF-pumped OPA, one can only generate signal wavelengths down to 550 nm, thus leaving a significant portion of the visible range uncovered. The visible region (450–700 nm), which is very important for spectroscopic applications, can be well covered by a SH-pumped Ti:sapphire OPA; on the other hand, for Yb lasers a visible OPA requires pumping with the TH, with an additional frequency conversion process which reduces the overall conversion efficiency.

The tuning range of OPAs in the visible is limited to ≈ 400 nm (when pumped by the TH of an Yb laser) by the energy of the pump photon; pumping with more energetic photons, such as, for example, the TH of Ti:sapphire at 266 nm, is difficult because of the onset of strong two-photon absorption in the nonlinear crystals at the intensities required for the OPA process; generation of tunable UV pulses is nevertheless possible by SHG of the visible OPA or SFG with the FF pulses. Long-wavelength tuning is limited in BBO to 3 μm by the onset of mid-IR absorption. Other crystals with higher refractive index display a more extended mid-IR transparency, allowing the direct generation of mid-IR idler pulses out to a wavelength of $\approx 5 \mu\text{m}$. Examples of these crystals are LiIO_3 , KNbO_3 , KTiOPO_4 (KTP) or its isomorphs, magnesium-oxide (MgO): LiNbO_3 , LN and SLT. These crystals are however not transparent in the 5–10 μm region (the fingerprint region), where many molecular vibrational transitions occur. There are crystals with extended mid-IR transparency, such as ZnGeP_2 (ZGP), GaSe, AgGaSe_2 and AgGaS_2 (AGS), but they all display absorption onsets in the visible (0.74 μm for ZGP, 0.65 μm for GaSe, 0.73 μm for AgGaSe_2 and 0.53 μm for AGS) so that they all suffer from two-photon absorption when pumped at 800 nm, preventing the direct generation of mid-IR pulses as the idler of an 800 nm pumped OPA for wavelengths longer than 5 μm . The situation improves slightly at 1030 nm wavelength, which allows pumping of the AGS crystal and direct generation of idler pulses tunable out to 7 μm . However, due to the large GVM between signal and idler, the amplified bandwidth is rather narrow, resulting in pulsewidths of 160 fs or longer. There is a current ongoing research effort in developing powerful sources of ultrashort pulses at 2 μm , based on Ho: or Tm:doped gain media, which would allow direct pumping of mid-IR OPAs based on the above mentioned crystals.

According to the previous discussion, the choice of the nonlinear OPA crystal(s) is hence determined by the balance of several parameters: the transparency range; the nonlinear coefficients; the nonlinear interaction parameters; the optical damage threshold;

the availability and price. Due to the combination of several of the above listed parameters (Nikogosyan, 1991), BBO is the crystal of choice for many OPA applications. It has a broad enough transparency range to allow amplification from 400 nm to 3 μm . Its nonlinear coefficient is moderately high, with $d_{\text{eff}} \approx 2$ pm/V for Type I OPA pumped at 400 nm. Bismuth triborate (BiB_3O_6 , or BIBO) is a biaxial crystal which possesses extremely large parametric gain bandwidth for degenerate collinear interaction when pumped at 800 nm. Its main advantage is the high second order nonlinear susceptibility ($d_{\text{eff}} = 3.2$ pm/V for Type-I SHG). Lithium triborate (LiB_3O_5 , LBO), on the other hand, despite its lower nonlinear coefficient, has high damage threshold, which makes it suited for the final power amplification stages. LN and SLT have IR transparency extended to 5.2 μm and 5.5 μm , respectively, which make them suited for amplification in the infrared. For both crystals, the highest element of the nonlinear tensor is d_{33} (27 pm/V for LN, and 21 pm/V for SLT), an extremely high nonlinearity which can be only exploited with Type-0 interaction. For this reason, they are typically periodically poled to obtain QPM. LN has low optical and photorefractive damage threshold, which can be increased by doping the crystal with MgO without affecting its high nonlinear coefficient. In power stages or booster amplifiers, PPSLT is preferred thanks to its higher damage threshold.

Pulse Compression

The third optional subsystem of an OPA is pulse compression, to compensate for the spectral phase accumulated by the amplified pulses, and achieve TL pulse duration. Typically, for visible/near-IR wavelengths, the OPA pulses are positively chirped due to material dispersion introduced by the seed generation process (e.g., by linear propagation of the WLC in the generation plate), by optical elements in the signal path (lenses, beam splitters, spectral filters ...) and by the OPA crystals. For narrow amplified pulse bandwidths (i.e., corresponding to TL pulsewidths of ≈ 50 fs or longer) the effects of such dispersion are negligible, so that no compression is necessary and the OPA pulses can be used as generated. For broader bandwidths, pulse compression is necessary to achieve the TL pulse duration. For moderately broad bandwidths, it is sufficient to correct the second-order dispersion, or group delay dispersion (GDD), while for broader bandwidths, and in particular for sub-10-fs TL pulses, it is also necessary to simultaneously correct for the third order dispersion (TOD).

Fig. 5 shows some of the several available optical systems which are able to provide negative dispersion in the visible/near-IR including grating pairs, prism pairs, chirped mirrors (CMs) and adaptive optical systems.

Grating pairs (see panel (a)) are not commonly used with OPAs due to their high losses and the unnecessarily large GDD values that they introduce. The simplest compressor consists of a Brewster-cut prism pair in a folded configuration (panel (b));

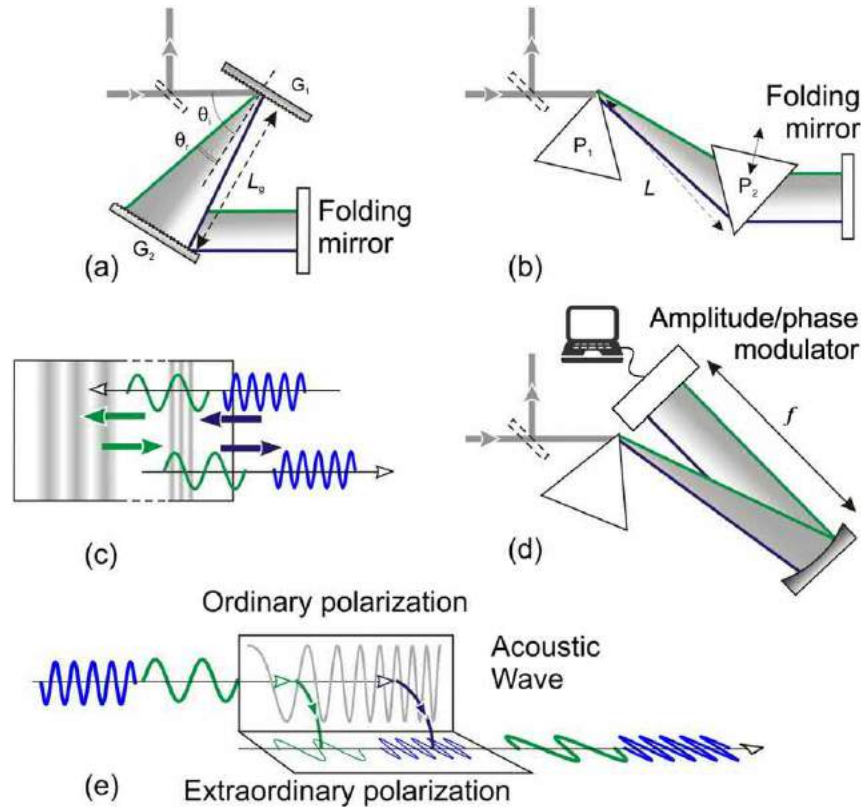


Fig. 5 Pulse-compression schemes. (a) Grating pair; (b) prism pair; (c) chirped dielectric mirror; (d) amplitude and phase-modulator in the Fourier plane of a $4f$ pulse shaper; and (e) acousto-optic programmable dispersive filter (AOPDF).

for a given prisms distance and glass insertion, it can be adjusted to introduce a negative GDD which compensates for the GDD of the OPA pulses, however it simultaneously introduces a large negative TOD, which cannot be independently controlled. Since both GDD and TOD are proportional to the prism tips separation, the ratio TOD/GDD is a characteristic of the prisms material and the wavelength, and can be minimized by choosing materials with low dispersion (such as fused silica, MgF_2 , or CaF_2). In this case, however, the prism distance required to achieve a given value of GDD increases. Since it can only correct for GDD but not also for TOD, a simple prism compressor is suitable for TL pulse durations down to ≈ 20 fs, but not for broader bandwidths. An additional disadvantage of the prism pair is the fact that the prism distance needs to be changed as the OPA wavelength is tuned. A better solution consists in multiple bounces on a pair of CMs (panel (c)), which can be designed to correct simultaneously for both GDD and TOD and ensure either full compression of the ultra-broadband spectrum generated by an OPA (see Section Ultra-Broadband OPAs) or pulse compression throughout the OPA tuning range. CMs have the advantage of high throughput and a particularly robust setup, introducing a dispersion that is essentially insensitive to misalignment and depends only on the number of bounces. CMs introduce a dispersion that can only be varied in discrete steps (e.g., $\text{GDD} \approx -50 \text{ fs}^2$ in the visible range), so that fine dispersion tuning requires the insertion of some variable amount of material on the beam path, typically a pair of fused silica wedges. Finally, pulse compression can be achieved by adaptive optical systems, which employ pulse shapers to provide an arbitrary control of the frequency-dependent spectral phase. As shown in Fig. 5(d), pulse shapers typically use the so-called 4- f arrangement, in which, after a dispersive element (a grating or a prism), a lens (or a spherical mirror) performs a spatial Fourier transform which converts the angular dispersion to a spatial separation at the back focal plane, where a phase (and/or intensity) modulator is located. Different phase modulators, based on liquid-crystals, acousto-optic modulators or deformable mirrors, can be employed. Alternatively, phase modulation can be achieved by an acousto-optic programmable dispersive filter ((AOPDF), panel (e)), in which the light wave interacts with a collinearly propagating acoustic wave.

Ultra-Broadband OPAs

Eq. (21) makes it clear that, in order to obtain broad phase matching bandwidths, one must achieve, for a given signal frequency $\bar{\omega}_1$, $\delta_{12}=0$, i.e., group velocity matching between signal and idler pulses. It will then become possible to amplify a broad bandwidth centered around $\bar{\omega}_1$. In an OPA using a collinear interaction geometry, the propagation direction inside the nonlinear crystal is selected to satisfy the phase-matching condition ($\Delta k=0$) for a given signal wavelength. In this configuration the signal and idler group velocities are in general not matched. Group velocity matching can be obtained in a Type I (or Type 0 in a periodically poled crystal) degenerate configuration, in which signal and idler have the same frequency ($\omega_1=\omega_2=\omega_3/2$) and the same polarization.

If the signal wavelength is tuned away from degeneracy, then the $\delta_{12}=0$ condition is generally not fulfilled in a collinear configuration, leading to narrow phase matching bandwidths. An additional degree of freedom can be introduced using a noncollinear geometry (Brida *et al.*, 2010; Gale *et al.*, 1995), in which pump and signal wave-vectors form an angle α (independent of signal wavelength) and the idler is emitted at an angle Ω with respect to the signal. In this case the phase matching condition is a vector equation, $\mathbf{k}_3=\mathbf{k}_1+\mathbf{k}_2$ that, when projected on directions parallel and perpendicular to the signal wave-vector, becomes

$$k_3 \cos \alpha = k_1 + k_2 \cos \Omega \quad (23a)$$

$$k_3 \sin \alpha = k_2 \sin \Omega \quad (23b)$$

Note that the angle Ω is not fixed, but it depends on the idler frequency according to Eq. (23b). If the signal frequency increases by $\Delta\omega$, the idler frequency decreases by $\Delta\omega$ and the wave-vector mismatches along the two directions parallel and perpendicular to the signal wave vector can be approximated, to the first order, as

$$\Delta k_{\text{par}} \cong -\frac{\partial k_1}{\partial \omega_1} \Delta \omega + \frac{\partial k_2}{\partial \omega_2} \cos \Omega \Delta \omega - k_2 \sin \Omega \frac{\partial \Omega}{\partial \omega_2} \Delta \omega \quad (24a)$$

$$\Delta k_{\text{perp}} \cong \frac{\partial k_2}{\partial \omega_2} \sin \Omega \Delta \omega + k_2 \cos \Omega \frac{\partial \Omega}{\partial \omega_2} \Delta \omega \quad (24b)$$

To achieve broadband phase matching, both Δk_{par} and Δk_{perp} must vanish. Upon multiplying Eq. (24a) by $\cos \Omega$ and Eq. (24b) by $\sin \Omega$ and adding the results, we get

$$\frac{\partial k_2}{\partial \omega_2} - \cos \Omega \frac{\partial k_1}{\partial \omega_1} = 0 \quad (25)$$

which is equivalent to

$$v_{g1} = v_{g2} \cos \Omega \quad (26)$$

Eq. (24) lends itself to a very simple geometrical interpretation (Riedle *et al.*, 2000): in a noncollinear configuration, broadband phase matching can be achieved for a signal-idler angle Ω such that the signal group velocity equals the projection of the idler group velocity along the signal direction. The so-called noncollinear optical parametric amplifier (NOPA), is a widely used device for the generation of few-optical-cycle pulses in the visible range. In a practical situation the pump-signal angle α is determined by the propagation direction of the seed beam, while the signal-idler angle Ω adjusts itself, according to Eq. (23b), to satisfy the

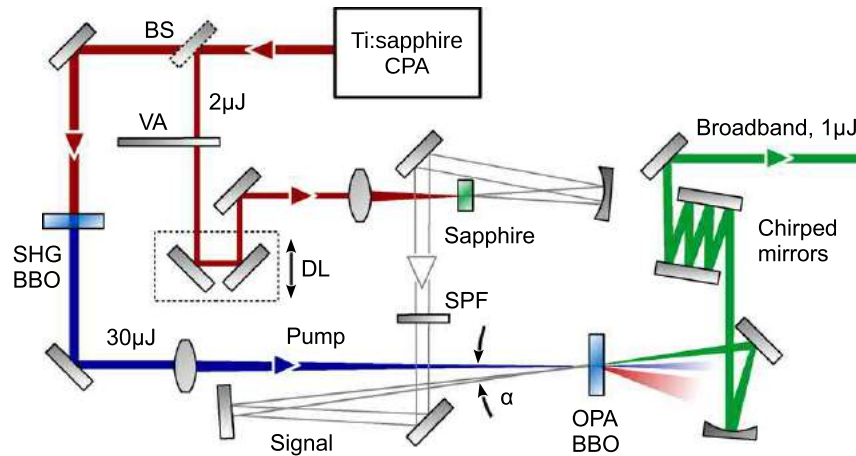


Fig. 6 Scheme of a β -barium borate (BBO)-based noncollinear optical parametric amplifier (NOPA) pumped at 400 nm. BS, beam splitter; CPA, chirped pulse amplification; DL, delay line; SHG, second harmonic generation; SPF, shortpass filter; VA, variable attenuator.

phase-matching condition. For this reason, the idler is emitted at a different angle for each wavelength, i.e., it is angularly dispersed and not easily usable.

In the following we will describe a typical visible NOPA design (Zavelani-Rossi *et al.*, 2001), the schematic of which is shown in Fig. 6. The system is pumped by an amplified Ti:sapphire laser generating 100-fs, 800-nm pulses at 1 kHz with energy up to 500 μ J. The energy is sufficient for simultaneously pumping several independent NOPAs. A fraction of the beam is used to generate the pump pulses at 400 nm by SHG in a 1-mm-thick BBO crystal; pulse energies up to 30 μ J are used to pump the first stage. Another small fraction of the beam, with energy of approximately 2 μ J, is focused into a 1-mm-thick sapphire plate to generate the single-filament WLC seed. The chirp of the visible portion of the white light is small and fairly linear with frequency. To avoid the introduction of additional chirp, only reflective optics are employed to guide the WLC to the amplification stage. Parametric gain is achieved in a 1-mm-thick BBO crystal, cut for type I phase matching ($\theta = 32$ degree), using a single-pass configuration to increase the gain bandwidth. The WLC seed is imaged by a spherical mirror in the BBO crystal, with a 100- μ m spot size matching that of the pump beam. A thin short-pass filter removes the strong residual 800-nm component from the WLC, preventing its parasitic amplification.

When the BBO crystal is illuminated by the pump pulse and aligned perpendicularly to the pump beam, it emits a strong off-axis PSF in the visible in the form of a cone with apex angle of ≈ 6.2 degree (corresponding to an angle of 3.82 degree inside the crystal); this is the direction for which the group velocities of signal and idler are matched and therefore the gain bandwidth is maximized. The visible cone gives a visual aid to the identification of the optimum condition for broadband generation, which is found when the pump-signal angle matches the cone apex angle. In this condition, for optimum pump-seed delay, an ultrabroad gain bandwidth that extends over most of the visible (500–750 nm) is observed. The amplified pulses from a single stage have energy of approximately 1–2 μ J; much higher energies, up to ≈ 300 μ J, can be obtained by a second amplification stage. After the gain stage the amplified pulses are collimated by a spherical mirror and sent to the compressor.

Several compressor schemes have been implemented for the visible NOPA. Simple prism pairs can correct the second but not third-order dispersion and thus can only compress the pulses down to 10–15 fs; sub-10-fs pulses can be achieved by using either prism-grating or prism-CMs combinations, as well as adaptive compressors based on deformable mirrors. It is also possible to use exclusively CMs (Zavelani-Rossi *et al.*, 2001), greatly simplifying the system design, allowing for compactness, insensitivity to misalignment and high day-to-day reproducibility, which are of great importance in practical applications.

Optical Parametric CPA

In a broadband OPA both the pump and the seed are typically femtosecond pulses generated by a Ti:sapphire or Yb laser system. It is generally difficult and expensive to scale the energy of femtosecond pump lasers; on the other hand, it is easier to generate energetic picosecond pulses (5–50 ps duration) exploiting well-established gain media such as Nd: or Yb:doped crystals. With a long pump pulse, in order to achieve efficient energy extraction, it is necessary to temporally overlap the seed pulse with the pump by first stretching the seed pulse and then, after the amplification step, recompressing it back to nearly TL duration. This scheme, which is very similar to the chirped pulse amplification (CPA) (Backus *et al.*, 1998) occurring in a real gain medium, is known as OPCPA (Dubietis *et al.*, 1992) and is considered as the most promising route for energy scaling of few-optical-cycle pulses. A typical scheme of an OPCPA is shown in Fig. 7. The key technical hurdle in OPCPA is the synchronization of the pump and seed pulses; it can be achieved either electronically, using a suitable circuitry, or optically, by using a part of the seed pulse spectrum to injection seed the pump laser. The optical approach allows a much lower pump-seed timing jitter but it is more challenging, since pump and seed are at different frequencies.

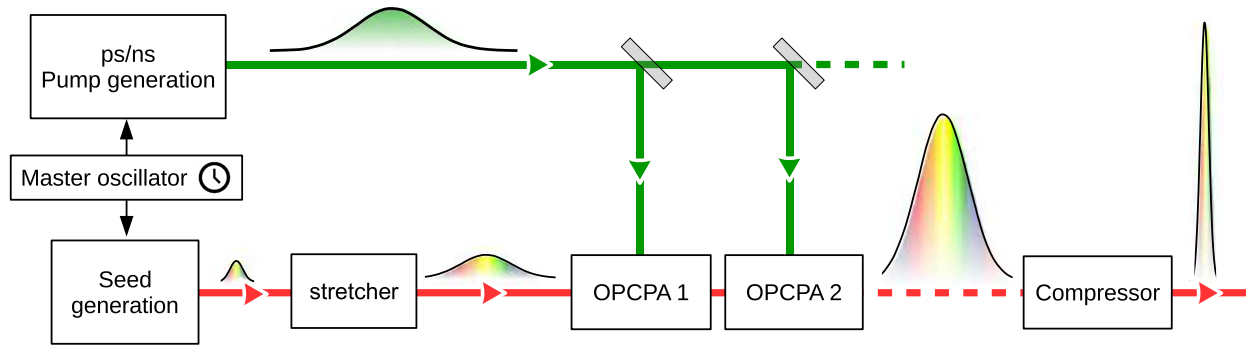


Fig. 7 General scheme of a multistage optical parametric chirped pulse amplifier (OPCPA). In this scheme the timing between the seed and the pump sources is optically provided by a master oscillator.

The OPA/OPCPA approaches offer some distinct advantages with respect to CPA for the generation of high peak power few-optical-cycle pulses:

1. The OPA/OPCPA has the capability of providing a high gain in a relatively short path, allowing a compact, tabletop amplifier setup, minimizing the linear and nonlinear phase distortions and ensuring an excellent temporal and spatial quality of the pulses.
2. With suitable selection of the phase-matching conditions (see Section Ultra-Broadband OPAs), the OPA process can provide gain bandwidths well in excess of those achievable with conventional amplifiers and can sustain pulse spectra corresponding to a TL duration of just a few cycles of the carrier frequency.
3. In an OPCPA, amplification occurs only during the pump pulse, so that amplified spontaneous emission and the consequent pre-pulse pedestal are greatly reduced.
4. Since the OPA/OPCPA is an instantaneous process of energy exchange between the interacting beams, no fraction of the pump photon energy is deposited in the medium; thus, if parasitic pump absorption can be neglected, thermal loading effects are absent, greatly reducing spatial aberrations of the beams and allowing to achieve much higher repetition rates avoiding the heat removal problems that limit the frequency scaling of CPA systems.

In addition to these clear advantages, the OPA/OPCPA concept also has some drawbacks:

1. an OPA/OPCPA requires pump pulses with short duration (a few picoseconds to tens of picoseconds), which are technologically challenging to generate, especially at high energy and/or high repetition rate; on the other hand the CPA, thanks to energy storage in the excited state, can use much longer nanosecond pump pulses (or even continuous wave pumping for materials with long excited state lifetime like Yb-doped crystals);
2. an OPA/OPCPA is very sensitive to the spatial and temporal quality of the pump, again putting a severe technological constraint on the pump laser.

Conclusions

Continuous progress in the development of light sources based on nonlinear optical effects for the generation of tunable ultrashort pulses ranging from the near-infrared throughout all the visible range provides very efficient tools to be exploited in time-resolved spectroscopic techniques and high field physics. Second-order nonlinear optical processes like OPA have demonstrated the capability of generating light pulses with durations down to few electric field cycles. Basic concepts and implementation schemes have been described, offering a comprehensive state of art tour in this rapidly evolving research field. Tunable few-cycle pulses constitute unprecedented tools for investigating ultrafast processes in many sectors related to material science and biology. Pump-probe techniques with broadband pulses and more recently two-dimensional spectroscopic techniques allow to follow in great detail complex dynamical processes.

References

- Backus, S., Dufree III, C.G., Murnane, M.M., Kapteyn, H.C., 1998. High power ultrafast lasers. *Review of Scientific Instruments* 69, 1207–1223.
- Baumgartner, R.A., Byer, R., 1979. Optical parametric amplification. *IEEE Journal of Quantum Electronics* 15, 432–444.
- Boyd, R.W., 2003. *Nonlinear Optics*. New York, NY: Academic Press.
- Bradler, M., Baum, P., Riedle, E., 2009. Femtosecond continuum generation in bulk laser host materials with sub- μJ pump pulses. *Applied Physics B* 97, 561–574.
- Brida, D., Manzoni, C., Cirri, G., *et al.*, 2010. Few-optical-cycle pulses tunable from the visible to the mid-infrared by optical parametric amplifiers. *Journal of Optics* 12, 013001.

- Cerullo, G., De Silvestri, S., 2003. Ultrafast optical parametric amplifiers. *Review of Scientific Instruments* 74, 1–8.
- Dmitriev, V.G., Gurzadyan, G.G., Nikogosyan, D.N., Lotsch, H.K.V., 1999. *Handbook of Nonlinear Optical Crystals*. Springer Series in Optical Sciences, vol. 64. Berlin: Springer.
- Dubietis, A., Jonušauskas, G., Piskarskas, A., 1992. Powerful femtosecond pulse generation by chirped and stretched pulse parametric amplification in BBO crystal. *Optics Communications* 88, 437–440.
- Gaeta, A.L., 2000. Catastrophic collapse of ultrashort pulses. *Physical Review Letters* 84, 3582–3585.
- Gale, G.M., Cavallari, M., Driscoll, T.J., Hache, F., 1995. Sub-20 fs tunable pulses in the visible from an 82 MHz optical parametric oscillator. *Optics Letters* 20, 1562–1564.
- Giordmaine, J.A., Miller, R.C., 1965. Tunable coherent parametric oscillation in LiNO_3 at optical frequencies. *Physical Review Letters* 14, 973–976.
- Harris, S.E., Oshman, M.K., Byer, R.L., 1967. Observation of tunable optical parametric fluorescence. *Physical Review Letters* 18, 732–734.
- Hum, D.S., Fejer, M.M., 2007. Quasi-phasematching. *C R Physique* 8, 180–198.
- Nikogosyan, D.N., 1991. Beta barium borate (BBO). *Applied Physics A* 52, 359–368.
- Riedle, E., Beutner, M., Lochbrunner, S., *et al.*, 2000. Generation of 10 to 50 fs pulses tunable through all of the visible and the NIR. *Applied Physics B* 71, 457–465.
- Zavelani-Rossi, M., Cerullo, G., De Silvestri, S., *et al.*, 2001. Pulse compression over a 170 THz bandwidth in the visible by use of only chirped mirrors. *Optics Letters* 26, 1155–1157.

Mode-Locked Lasers

Ladan Arissian and Jean-Claude Diels, University of New Mexico, Albuquerque, NM, United States

© 2018 Elsevier Inc. All rights reserved.

1	Introduction on Frequency Combs	302
2	The Mode-Locked Laser as a Frequency Comb Generator	303
2.1	Typical Laser Configurations	303
2.2	Mode-Locking as Orthodocny	303
2.3	Phase Modulation and Dispersion - the Soliton Operation	304
3	The Control Knobs for the Frequency Comb	305
3.1	Cavity Length Change	306
3.2	Repetition Rate and Group Velocity	306
3.3	Comb Control with an Intracavity Etalon	306
3.4	Locking the Repetition Rate	307
4	Intracavity Phase Interferometry	308
References		309

1 Introduction on Frequency Combs

A mode-locked laser is more than the typical source of ultrashort pulse; it is the source of the most accurate ruler in the frequency domain. It combines the properties of a narrow line single mode laser, with a femtosecond time resolution. The sketches of Fig. 1 summarize the properties of the output pulse train of a mode-locked laser, both in time and frequency domain. A simple description in time is that of a perfectly monochromatic wave, being sampled at regular time intervals τ_{rt} . There is in general no relation between the period of the optical wave $1/\nu$ and τ_{rt} . The phase of the harmonic wave at the “center” or “peak” of the temporal gate (pulse) is called the Carrier to Envelope Phase (CEP) (Fig. 1(a)). The difference between the CEP of two successive pulses $\Delta\phi_1$ and $\Delta\phi_2$ is constant over the (ideal) pulse train. The ratio $(\Delta\phi_2 - \Delta\phi_1)/(2\pi\tau_{rt}) = f_0$ is a frequency, called the carrier to frequency offset (CEO).

In the frequency domain, for an infinite pulse train, the laser spectrum consists of δ -functions equally spaced by $1/\tau_{rt}$, over a range of the order of the inverse of the gate (pulse) width (Fig. 1(b)). Extending this pulse train to close to zero frequency, one can show that f_0 corresponds to the first tooth of the extended comb [1].

While the CEO is a well defined measurable quantity, there is some fuzziness in the definition given for the CEP. Indeed, for a very long pulse compared to the carrier period $1/\nu$, it is at best difficult to pinpoint the “center” or “peak” of the pulse envelope as compared to a crest of the wave (Fig. 1(c)). For a very short pulse as in (Fig. 1(d)), in the frequency domain it is difficult to pinpoint which tooth of the comb is closest to the “center” or “peak” of the pulse spectrum. Clearly an alternate definition is desirable for the CEP. Techniques exist to measure the real electric field of a pulse in amplitude and phase, without a decomposition in “carrier” and “envelope”. A complex representation of the field is obtained by taking the Fourier transform of this real measured field,

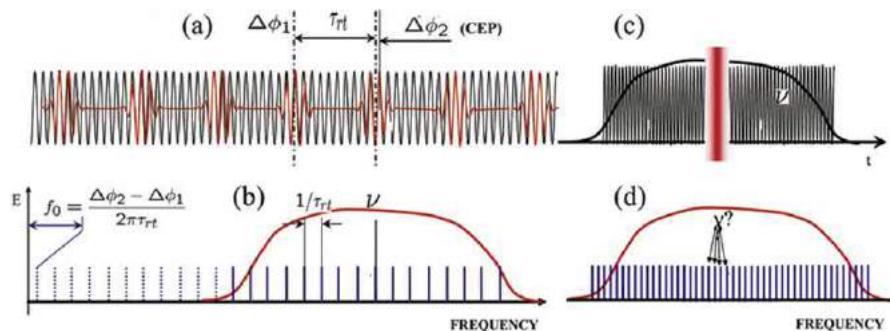


Fig. 1 (a) A set of pulses is periodically (period τ_{rt}) extracted from a cw wave. Each pulse envelope peaks for a particular phase of the cw wave, called Carrier to Envelope Phase (CEP). The ratio of CEP to repetition period τ_{rt} is the carrier to envelope offset (CEO). (b) The Fourier transform of (a) is a set of δ -functions equally spaced. The CEO is the frequency of the tooth closest to zero, in an extension of the comb. (c) For a long pulse, the exact location (within a fraction of the light period) of the center or peak of the envelope is difficult to define. (d) for an ultrashort pulse, the mode corresponding to the peak of the spectrum is equally difficult to define.

eliminating the negative frequency parts, before taking the inverse Fourier transform, which is now the “complex” electric field. A new definition of the CEP, independent of a decomposition in “carrier” and “envelope” of the optical pulses, is the difference between the value of the phase of the complex electric field at the peak of its amplitude, and the value of the phase taken at the peak of the real field [2,3].

It is often believed that the notions of CEP and CEO apply only to few cycle pulses. This is clearly disproved by experiments of intracavity phase interferometry performed with ps pulses [4]. In these experiments to be detailed in Section 4, the variation in CEO is used to perform phase detection with sensitivity better than 1 part in 10^8 .

2 The Mode-Locked Laser as a Frequency Comb Generator

2.1 Typical Laser Configurations

The typical laser consists in a linear (Fabry-Perot) or ring resonator in which a broadband gain medium is inserted. In mode-locked operation, one or more pulses are circulating in the cavity, generating a train of pulses through an output mirror, spaced by the cavity round-trip time. One essential element required to create a train of identical ultrashort pulses at the output is a phase modulator inducing a frequency modulation or chirp on the circulating pulse. In the most common Ti:sapphire laser, the gain crystal modulates the phase through its nonlinear index $n_2 I$ (I being the pulse intensity in the crystal). This creates a positive frequency sweep or chirp across the circulating pulse. If a saturable absorber is used with a resonance centered at an optical frequency above that of the circulating pulse, the saturation of the resonant dispersion associated with the homogeneously broadened line results in a downchirp [1]. Another essential element is a transparent optical system in which the light velocity is dependent on the optical frequency. This dispersive element can consist in a pair of prisms, as in Fig. 2, which can be tuned from a positive to a negative dispersion, dependent on the amount of glass traversed by the beam and the distance between prisms. Exact expressions for the dispersion associated with a pair of prisms can be found in reference [5]. Often the pair of prisms is replaced by mirrors of which the coating has been engineered to have dispersion (frequency dependent phase shift) of the desired sign and shape. The ratio of phase modulation to dispersion is a key element in having a mode-locked laser produce a perfect frequency comb, as will be detailed in Section 2.3.

In the case of a linear cavity (Fig. 2(a)), the laser consists in a gain section, a dispersive section (a pair of prisms in the example of the figure) and end-cavity mirrors. The various control knobs available to fix the parameters of the frequency comb (center of pulse spectrum, pulse duration, CEO and repetition rate) are detailed in Section 3. A slit, aperture or razor blade between the prism sequence and the end mirror determines the position of the spectral envelope of the individual pulse. The pulse duration is tuned by acting on the dispersion, translating one of the prism of the sequence, thus changing the amount of glass. The repetition rate can be adjusted independently of other parameters by tilting the end mirror after the prism sequence. The pump power provides a fine adjustment.

Cavity length adjustment with a piezo-element on one end mirror provides another tuning parameter for the comb CEO. The exact transfer function of all these controls can be found in reference [5]. While manipulation of dispersion and phase modulation is essential in the generation of a stable frequency comb, mode-locking requires some amplitude modulation to bring down the duration of noise fluctuation to the picosecond range before dispersive compression kicks in. These initial pulses can be provided by saturable absorbers, for instance multiple quantum wells used as end mirror of a linear cavity (Fig. 2(a)). The saturable absorber can have the additional function of locking the repetition rate (group velocity control) in the case of mode-locked lasers producing dual mode combs. It is then located in the middle of a linear cavity, or at 1/4 perimeter of the gain medium in a ring cavity (Fig. 2(b)). This type of mode-locked operation with two pulses circulating in the cavity is dealt with in Section 4.

2.2 Mode-Locking as Orthodancy

The resonance condition of any cavity is that the total phase shift $\sum k_i l_i$ experienced by the light in a round-trip time be a multiple of 2π . The sum is taken over all elements i of the cavity of index of refraction n_i and length l_i . An average index n_{av} can be defined

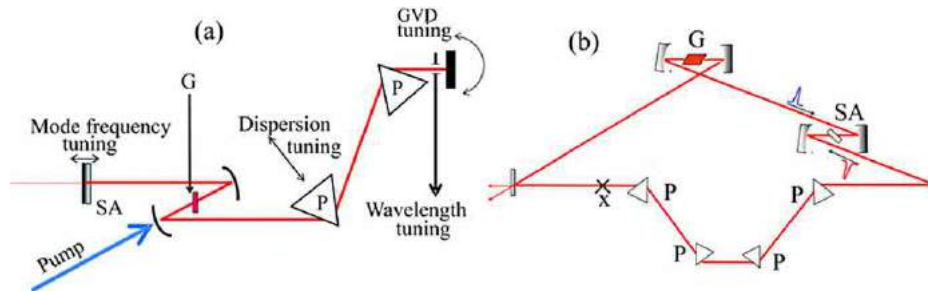


Fig. 2 Typical mode-locked laser configurations. SA: saturable absorber, P dispersive (Brewster angle) prism; G: gain medium (a) The linear laser where the different comb controls are indicated. (b) Typical bidirectional ring laser configuration, where two pulses circulate in opposite direction crossing in the saturable absorber (SA) and in position X.

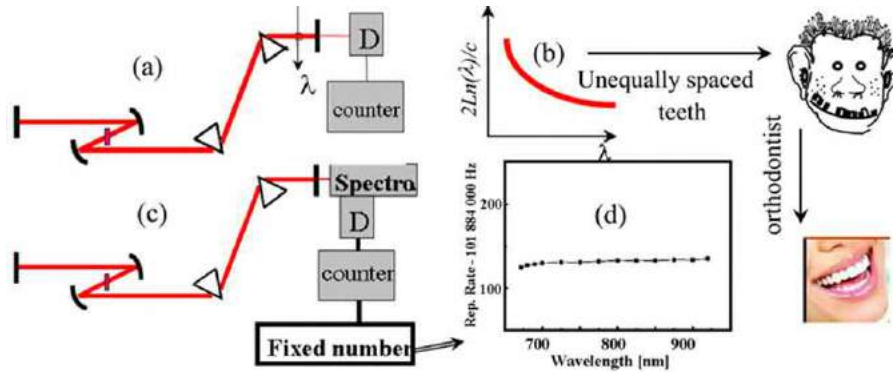


Fig. 3 (a) The aperture (or slit) at the output end of the laser is closed so as to let only a few modes oscillate [6,7]. A detector D records the mode beating, at the round-trip frequency of the cavity. That round-trip frequency, recorded as a function of the wavelength λ by a frequency counter, shows the variation of mode spacing across the gain spectrum (b). (c) The aperture is open, allowing mode-locked operation of the laser. A portion of the spectrum of the frequency comb is selected by a monochromator. The repetition rate is recorded by the same frequency counter as function of the wavelength (d), showing no variation across the spectrum. Locking the modes of the laser is the analogue of an orthodontist intervention on the comb.

by $\sum k_i l_i = k_{av} L = (2\pi/\lambda) n_{av} L = [\omega n_{av}/c] L$ where L is the cavity perimeter. The index of refraction n_{av} generally varies significantly over the frequency range corresponding to the bandwidth of a typical fs pulse. This implies that the modes spacing $n_{av} L/c$ varies across the bandwidth of the pulse. In the classical textbook describing mode-locking, it is said that the short pulse is the result of all the modes of the cavity emitting simultaneously and in phase.

This textbook description does not correspond to reality: it can be shown that the emission by a set of non-equidistant modes in phase does not result in an uniform pulse train [1]. Despite the non-uniformity of the spacing of the modes of the laser cavity, a simple experiment as sketched in Fig. 3 demonstrates the difference between the modes of the laser cavity and the teeth of the comb. Mode-locked operation is prevented by closing the aperture at the output end of the laser cavity (Fig. 3(a)). The wavelength of the laser can be tuned by translating the aperture in the plane of the figure. An RF spectrum analyzer (or frequency counter) can be used to record the variation of mode-beating frequency $1/\tau_r$ versus wavelength [6,7]. This variation is a measure of the average group velocity of the intracavity pulse $v_{av} = 2L/\tau_r$ where L is the cavity length (Fig. 3(b)). Opening or eliminating the aperture results in mode-locked operation (Fig. 3(c)). The teeth spacing of the comb can be measured across the output spectrum by selecting a wavelength range of interest with a spectrometer, and recording the round-trip frequency with a spectrum analyzer or a frequency counter (Fig. 3(d)). The very small variation of this number across the spectrum (< 10 Hz) over a 250 nm wavelength range reflects the thermal expansion of the laser cavity during the course of the experiment [8]. Measurements with stabilized lasers have shown the comb teeth spacing of the order of hundred of MHz to be constant within mHz over the full spectral width of the laser output [9].

2.3 Phase Modulation and Dispersion - the Soliton Operation

To a coarse approximation, the simplest model of steady state operation can be reduced to a cascade of phase modulation by Kerr nonlinearity followed by quadratic dispersion. As sketched in Fig. 4, a transparent medium of thickness d with a time dependent index of refraction $n(t)$ will phase modulate a pulse sent through it. In a plane wave approximation, we describe a pulse of optical frequency ω propagating in the z direction by an electric field $E(t, z) = \tilde{\mathcal{E}}(t, z) \exp[i(\omega t - kz)]$, where $\tilde{\mathcal{E}}(t, z) = \tilde{\mathcal{E}}(t, z) \exp(i\phi(t, z))$ (complex quantities are written here with a tilde). After propagating through an element with a time dependent index of refraction, the pulse will be phase modulated or “chirped” and have the complex envelope $\tilde{\mathcal{E}}(t, d) = \tilde{\mathcal{E}}(t, 0) \exp[-ik(t)d]$ where $k(t) = \omega n(t)/c = (2\pi/\lambda) n(t)$, c is the speed of light and λ the wavelength in vacuum. Most often, the time dependent index is due to the Kerr effect and is proportional to the light intensity $n(t) = n_2 I(t)$, resulting in an instantaneous frequency generally increasing with time or “upchirp”, $\phi(t) = [(\partial/\partial t)k(t)]d$ as sketched to the right of Fig. 4. In this temporal phase modulation process, the envelope shape $\mathcal{E}(t, z)$ remains unchanged.

Dispersion can be understood as phase modulation in the frequency domain. The input to a transparent medium with a frequency dependent index of refraction $n(\Omega)$ is the Fourier transform of the field defined as:

$$\tilde{E}(\Omega) = \mathcal{F}\{E(t)\} = \int_{-\infty}^{\infty} E(t) e^{-i\Omega t} dt = |\tilde{E}(\Omega)| e^{i\Phi(\Omega)} \quad (1)$$

In the definition (1), $|\tilde{E}(\Omega)|$ denotes the spectral amplitude and $\Phi(\Omega)$ is the spectral phase. If we refer to the central frequency of the pulse, $\Delta\Omega = \Omega - \omega$, the transformation of the pulse spectral envelope by propagation of a distance z through the dispersive medium is:

$$\tilde{\mathcal{E}}(\Delta\Omega, z) = \tilde{\mathcal{E}}(\Delta\Omega, 0) e^{-ik(\Delta\Omega, z)} \quad (2)$$

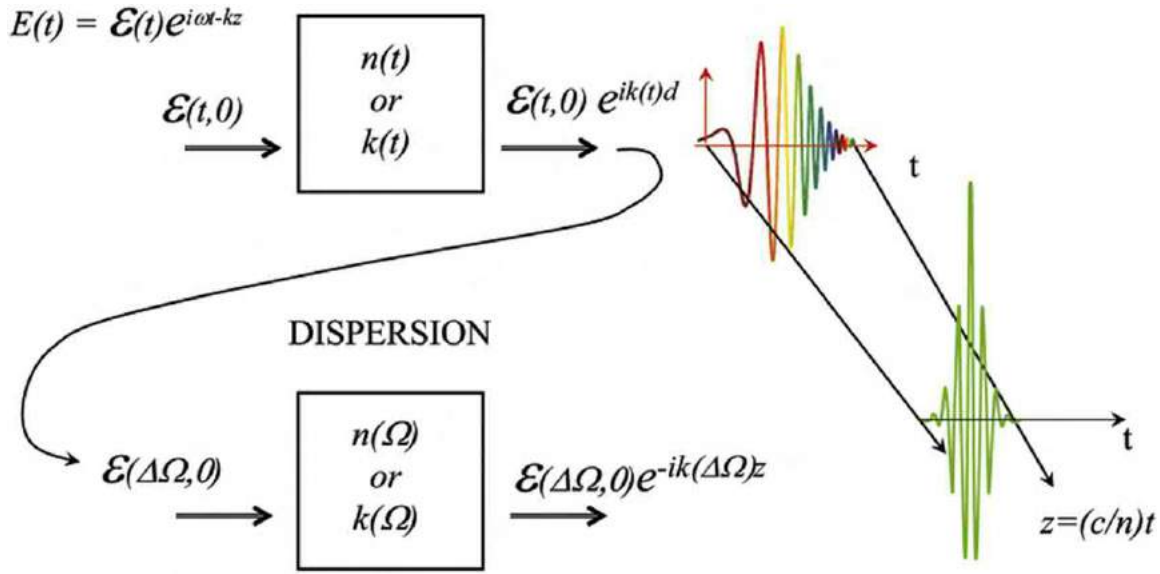


Fig. 4 Phase modulation and dispersion are the time-frequency correspondent of each other. In the time domain, phase modulation broadens the spectrum and introduce chirp (in general upchirp as sketched on the right) without affecting the pulse temporal profile. In the frequency domain, dispersion leaves the spectrum of the pulse unchanged. In the case of negative dispersion applied to the chirped pulse of the figure, the faster (higher frequency) leading edge catches up with the slower leading edge, leading to pulse compression.

In this spectral phase modulation process, the envelope shape $\tilde{\mathcal{E}}(\Delta\Omega, z)$ remains unchanged. If we expand the wavevector in series $k(\Delta\Omega) = k_0 + \Delta\Omega k' + \Delta\Omega^2 k''/2$ the first term k_0 corresponds to an irrelevant phase shift, the second term is simply the group delay, and the third term is the group velocity dispersion. As shown in the sketch on the right side of Fig. 4, if k'' is negative, the high frequency at the tail of the chirped pulse will propagate faster than the pulse front resulting in pulse compression. An exact balance is sought in mode-locked lasers between dispersion and Kerr phase modulation to achieve stable operation and minimum pulse duration.

Given the phase modulation and dispersion, one can derive a relation that predicts the values of pulse duration and intensity in steady state [3]. A question that remains is the nature of the mysterious orthodontist that transforms the unequal tooth spacing into a perfect comb with equally spaced teeth. The answer is in a frequency domain calculation similar to the one in time domain leading to solitons. In this calculation detailed in reference [3], the unequal mode spacing of the cavity is modeled by quadratic dispersion. The modification of the comb brought upon by self-phase modulation is calculated in the frequency domain. Imposing that the teeth of the comb are equally spaced leads to the condition between pulse duration τ , the cavity perimeter P ($= 2L$ in the case of a linear cavity), the length of the Kerr medium ℓ , the nonlinear index n_2 the peak intensity I_0 , the wavelength λ and the average dispersion of the cavity k''_{av} :

$$\tau^2 = \frac{\lambda |k''_{av}| P}{2\pi n_2 I_0 \ell} \quad (3)$$

It is remarkable this calculation leads to exactly the same prediction of pulse duration and intensity in steady state than the derivation of the soliton equation [3].

3 The Control Knobs for the Frequency Comb

While the CEO controls the frequency comb, it remains to be seen how this control is actuated, which depends on the specific goals to be achieved. If the aim is to create a frequency standard or an ultra-stable frequency comb, it is desirable to have coarse (usually slow) and fast (fine control) actuators on two parameters that define the comb — for instance the frequency of one mode, and the repetition rate.

Most of the controls easily accessible in a linear mode-locked laser are sketched in Fig. 2(a). Translating a slit or aperture after the prism sequence [or between the prism sequence in the case of the ring laser of Fig. 2(b)] provides a control of the central wavelength of the pulses. Translating a prism orthogonally to its base controls the pulse duration, through tuning of the cavity dispersion. A slow control of the cavity length is provided typically by mounting the end cavity mirror on a piezo-element. A fast control of the repetition rate is often achieved by inserting an electro-optic modulator in the pump beam, using the repetition rate dependence on pump power. Tilting the end mirror after the two prism sequence can be used for coarse (slow) control of the repetition rate. It has been shown that this control — effected with piezo-elements — offers a control of the comb repetition rate

(or mode spacing), without affecting the position of the central mode [5]. The latter repetition rate control combined with an acousto-optic modulator provides a near perfect set of orthogonal controls. Indeed, an acousto-optic modulator positioned outside the laser cavity can shift the comb frequency without affecting the mode spacing.

Rather than offering orthogonal control, most of the cavity parameters modify simultaneously both parameters of the comb, as illustrated below in the case of a change in cavity length.

3.1 Cavity Length Change

The easiest parameter to change in a mode-locked laser is the optical path between end mirrors (or the perimeter in the case of a ring cavity). This operation can be performed mechanically by putting an end mirror on a piezoelectric element in the case of a linear laser. In the case of a fiber laser, a portion of fiber will be wrapped around a piezoelectric cylinder.

There is no universal answer as to what is the result of a change in cavity length. One deduces easily from the sketch in Fig. 1 that the frequency ν_N of a mode N of the comb is:

$$\nu_N = f_0 + N \frac{1}{\tau_{rt}} \quad (4)$$

where f_0 is the CEO and the cavity round-trip time is $\tau_{rt} = 2L/v_{av}$. L is the length of the cavity (or $2L = P$ is the perimeter of a ring cavity), and v_{av} is the average velocity of the pulse circulating in the mode-locked laser cavity.

Let us imagine an adiabatic expansion of the resonator by an amount ΔL . We can expect that the number of modes N is unchanged. Assuming $f_0 = 0$ and is not affected by the cavity expansion (Fig. 1(b)), then the mode spacing (repetition rate) has to change. The accordion motion of the comb results in the following changes in mode frequency and repetition rate:

$$\Delta \nu_N = \nu_N \times \frac{\Delta L}{L} \quad (5)$$

$$\Delta \left(\frac{1}{\tau_{rt}} \right) = \frac{2\Delta L}{c} = \frac{\Delta P}{c} \quad (6)$$

where the second equality in Eq. (6) applies to a ring cavity. If we take as an example an elongation by a half wavelength $\lambda/2$ ($\lambda = 800$ nm) of a $\tau_{rt} = 10$ ns cavity, since $N = 4 \cdot 10^6$, the change in central mode frequency is $\approx 10^8$ Hz. The change in repetition rate is 50 Hz.

These considerations suggest that a change in intracavity optical path primarily translates the comb. The parameter that is modified by a particular control knob is not always obvious, and may depend on the way the control is applied. A different result is predicted in reference [10], where it is proposed to insert an electro-optic modulator (EOM) in a mode-locked cavity, regeneratively driven at the cavity repetition rate or a multiple thereof. The DC bias applied to the EOM would be used for fast control of the cavity length, or optical frequency of the comb. The intracavity pulse traversing the modulator experiences a positive, negative, or zero frequency shift to the pulse, depending on which part of the modulation waveform the pulse encountered. In soliton mode-locking, the cavity has negative dispersion $dk/d\omega < 0$. In the modulator of length ℓ , a positive frequency shift $\Delta\omega$ will cause a negative group delay τ_d , implying an increase of repetition rate. In this scenario, the phase of the wave applied to the EOM controls the repetition rate of the mode spacing of the comb, while the DC bias controls the mode position. A single actuator would thus be sufficient to stabilize the CEO f_0 and the teeth spacing of the comb.

3.2 Repetition Rate and Group Velocity

The mode spacing of the frequency comb — or repetition rate of the pulse train — is set by the average group velocity of the pulse circulating in the laser. It is generally taken for granted [11] that the group velocity v_g of a pulse circulating in a laser cavity is determined by the derivative of the k vector with respect to angular frequency:

$$v_g = \frac{1}{\frac{dk}{d\omega}} \quad (7)$$

where ω is the pulse carrier frequency. Contrary to the association of pulse envelope velocity and group velocity defined in Eq. (7), it is demonstrated in Ref. [12], that the envelope velocity of a pulse circulating in a mode-locked laser is generally not related to $dk/d\omega$, for a k vector averaged in the cavity, but to the gain and loss dynamics inside the laser. For instance, it is well known that the pulse envelope propagates at superluminal velocities in a saturable gain medium [13]. This effect contributes to the fine tuning of the group velocity by modulation of the pump power in a Ti:sapphire laser. In a saturable absorber, the envelope velocity will be decreased by elimination of the pulse leading edge in favor of the trailing edge. In mode-locked fiber lasers, the average envelope velocity is a combination of group velocity, modal velocity, gain velocity and absorber velocity.

3.3 Comb Control with an Intracavity Etalon

Intracavity Fabry-Perot have been used for decades to narrow down the laser spectrum of cw and pulsed lasers [14,15]. It is less known that they are an effective tool to control the parameters of a mode-locked laser. A uncoated intracavity glass etalon will tune

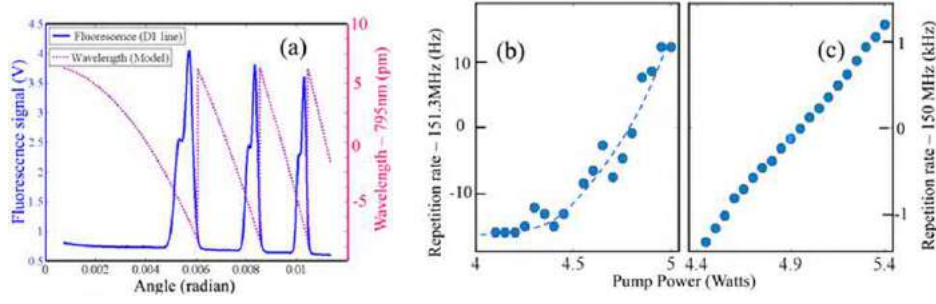


Fig. 5 Tuning a frequency comb by insertion of a fused silica etalon in a linear mode-locked laser cavity. (a) Solid line: fluorescence of ^{87}Rb irradiated by the frequency comb, as a function of the (internal) tilt angle of the etalon. The zero corresponds to normal incidence. Dashed line: calculated resonance wavelength of the etalon. The right ordinate indicates the difference (in pm) between the calculated wavelength and 795 nm. (b) Tuning of the repetition rate (mode-spacing) of the comb, as the pump power of the laser is ramped. The full tuning range is less than 30 Hz. (c) Tuning curve of the repetition rate (mode-spacing) after insertion of the etalon. The full tuning range is now 2.45 kHz over a pump power variation of 1 W.

not only the mode spacing, but also the pulse repetition rate [16]. Unlike the action of the slit in Fig. 2 or a birefringent filter, the etalon does not act on the average pulse frequency, but on the individual modes. As a demonstration, a 15.119 mm thickness uncoated fused-silica etalon is inserted in the linear cavity of Fig. 2(a). The output of the laser is sent to a cell containing ^{87}Rb , and the fluorescence of the vapor is monitored as the laser is tuned to the D1 line centered at 794.978851 nm. As the etalon is tilted away from the normal to the beam, the fluorescence of rubidium crosses successive resonances (Fig. 5(a)). The dotted line shows the wavelength corresponding to the etalon resonance condition. The 814.52 MHz hyperfine splitting of the upper state of the D1 transition is resolved in the structures of Fig. 5(a). Such a measurement indicates that, despite the low finesse of the etalon, precise mode-tuning is achieved. This property can be explained by the fact that the modes of the etalon are locked to those of the comb. Fig. 5(b) is a measurement of the change in repetition rate, in the absence of intracavity etalon, as the pump power is changed. The tuning range of the ≈ 150 MHz repetition rate is only 30 Hz for a pump power change of 1 Watt. If the fused silica etalon is inserted in the cavity, the tuning range increases to 2.45 kW, over the same 1 W pump power increase, as shown in Fig. 5(c). The insertion of an etalon inside the mode-locked cavity allows fine tuning over a large dynamic range of both the mode position (through the etalon angle) and the mode spacing (through the pump power). This property applies as well to linear lasers [16] as to ring lasers [12].

3.4 Locking the Repetition Rate

In most applications related to stabilization of frequency combs, the aim is to keep the CEO under tight control. There are however some applications where it is the CEO itself that is the free parameter being measured, while other characteristics of the comb (pulse duration, repetition rate, power) are to be kept constant. One application is intracavity phase interferometry, discussed in Section 4. In that technique, the mode locked laser is generating two correlated frequency combs, corresponding each to a different pulse circulating in the cavity. If the two combs have the same mode spacing, it is possible to interfere them. The interference between the two combs gives a signal at a frequency that is the difference between the two CEO f_{01} and f_{02} of each comb. Indeed, taking the difference of the frequency of corresponding modes (index N) of the comb:

$$\Delta\nu = \left(f_{02} + N \frac{1}{\tau_{rt,2}}\right) - \left(f_{01} + N \frac{1}{\tau_{rt,1}}\right) = f_{02} - f_{01} \quad (8)$$

since $\tau_{rt,2} = \tau_{rt,1}$ for the mode spacings to be equal. Note that the visibility of the beating between the two combs can be near 100%, since all pairs of modes contribute thanks to the property of the combs to have equal tooth spacing. The challenge to create this double comb is to lock the round-trip time of either pulse to the same value, and force the pulses to meet at the same location at each round-trip, *without perturbing* any other parameter of the comb. One obvious approach is to have both circulating pulses in the cavity created by gain switching, as is the case in optical parametric oscillators [17,18]. Unlike in inversion based lasers, in optical parametric oscillators the gain for a signal of intensity I_s at frequency ω_s is proportional to the product $I_s I_p$. It is therefore limited in duration to that of the pump pulse. In fs pulse synchronously pumped optical parametric oscillators, the pump laser is the clock that determines the repetition rate of either pulse that it generates in the optical parametric oscillators cavity.

Another method of locking the meeting point of two pulses in a mode-locked cavity is the saturable absorber. A saturable absorber (homogeneously broadened) is inserted either in a ring cavity, or in the middle of a linear cavity. Two pulses circulate in the cavity. The configuration of minimum losses is that where the two pulses meet in the absorber. It can indeed be shown that, for two pulses of equal intensity counter-propagating in the absorber, the effective saturation intensity is reduced by a factor 3, as compared to the saturation intensity for a single propagating pulse [1]. The mechanism by which the “colliding pulse” mode-locking is stable is explained in Fig. 6. Let us assume that the absorber, represented by the dot patterned rectangle, has moved to the right. The leading edge of pulse A will have no overlap with pulse B as it enters first the medium, and will be more attenuated

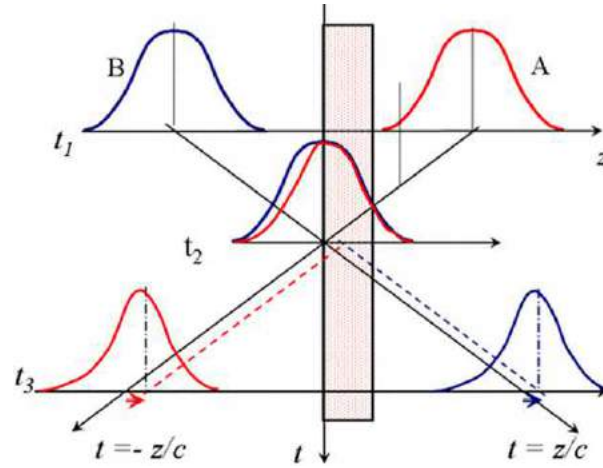


Fig. 6 Time sequence of two pulses *A* and *B* crossing in a saturable absorber. At the initial time t_1 , pulses *A* is closer to the saturable absorber than *B*. At time t_2 , the leading edge of *A* has passed the absorber, having been attenuated, while the leading edge of *B* overlapping with the trailing edge of *A* experiences less attenuation (mutual saturation).

than at the trailing edge (mutual saturation) where the two pulse overlap. Therefore, the center of gravity of pulse *A* will shift to the right, towards the absorber (dashed red line). The situation is opposite for pulse *B*, since pulse *A* will have left the absorber at the passage of the trailing edge of *B*. More attenuation of *B*'s trailing edge implies that its trajectory moves to the right (dashed blue line). The two dashed trajectories cross close to the center of the absorber, showing that, after successive round-trips, the pulse crossing point will be centered with the saturable absorber.

Saturable absorption is often the mechanism that sets the average group velocity in a discrete components laser, dominating the timing imposed by other elements. The situation can be different in other systems, such as fiber lasers. In contrast to observations in dye and solid state mode-locked lasers, bidirectional fiber lasers mode-locked by saturable absorber (single wall carbon nanotubes or CNT) produce pulses of different central wavelength, durations, and repetition rates [19]. This can be explained by the fact that fiber lasers cannot be adequately modeled by differential equations: the change in pulse characteristic per element is much larger than for discrete component lasers, which can result in a very large asymmetry in the evolution of the two pulses in the cavity.

4 Intracavity Phase Interferometry

Most applications of frequency combs involve fast and accurate stabilization electronics, to maintain the position of a tooth of a comb within a few Hz. Intracavity Phase Interferometry (IPI) is a technique that exploits the basic properties of mode-locking, to achieve phase detection better than 10^{-8} radian, without the need for stabilization. The IPI being referred to in this chapter is dealing with active (laser) resonators, not to be confused with "intracavity nonlinear spectroscopy" [20], in which actively stabilized ultra-high finesse passive Fabry-Perot cavities are used to transform a few ppm absorption into some % change in transmission. A broadband laser can be used as an amplitude sensor: the nonlinearity of laser close to threshold has been exploited to detect very small absorption [21]. More recently, it was realized that a mode-locked laser can be made an ultra-sensitive phase sensor [4]. A basic properties being exploited is that a laser is akin to a Fabry-Perot of infinite finesse. In addition, in a laser, a change in phase $\Delta\phi$ is converted into a change in frequency $\Delta\nu = \Delta\phi / (2\pi\tau_n)$, in contrast to passive interferometers where any change in phase is measured as a change in amplitude.

Implementation of IPI in the case of a ring laser is sketched in Fig. 7. Two pulses are counter-circulating in the ring cavity, with their average group velocity locked to the same value (using for instance a saturable absorber). At 1/4 cavity perimeter from their crossing (starting) point, one pulse passes through the gain medium, and the other has its phase modified by the sensor *S*. The latter is a generic term for the phase modification by the element to be measured, which could be an elongation, a change in index due to Faraday rotation, Kerr effect, air current, rotation etc... The laser outputs two frequency combs, which are recombined via a delay line. The interfering pulse trains recorded on a detector have a modulation at a frequency $\Delta\nu$ equal to the difference between the corresponding modes of order *m* of the comb. Since the combs have equal mode spacing over the spectrum, that difference is simply equal to the difference CEO frequency between the two combs: $\Delta\nu = f_{02} - f_{01}$. One would expect the bandwidth of the beat note $\Delta\nu$ to be equal to the sum of the bandwidths of the two CEOs. This is clearly incorrect: in an unstabilized laser such as considered in Fig. 2, the mode bandwidth is typical 1 MHz. But the measured bandwidth of the beat note $\Delta\nu$ is as small as 0.17 Hz [17]. The reason for the small noise in beat note is that the two output frequency combs are correlated. If the beat note is caused by a differential optical path ΔP seen by the two pulses, its frequency is simply the product of the light carrier frequency by the relative elongation: $\Delta\nu = \nu\Delta P/P$.

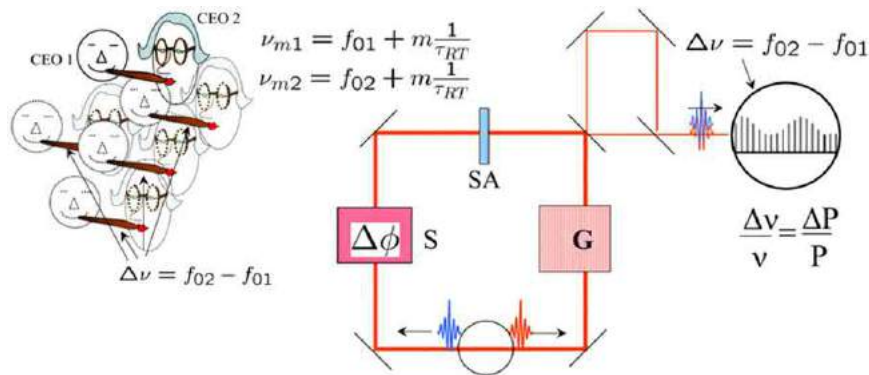


Fig. 7 Sketch of implementation of IPI with a ring cavity. Two pulses (red and blue) are circulating in opposite sense in the cavity, crossing in the circle at the bottom of the cavity, and in the saturable absorber SA. G is the gain medium, traversed alternatively at every half cavity round-trip by one of the pulses. S is the sample to be measured: an element that provides a differential phase shift proportional to the quantity to be measured. The two output pulse trains are overlapped through a delay line, and interfered on a detector. An oscilloscope will show the pulse train modulated at the “beat note” frequency $\Delta\nu$ equal to the difference between the CEOs of the two pulse trains. The CEO frequencies of this unstabilized laser are strongly fluctuated. However, because the two trains are correlated, these fluctuations are not seen in the beat frequency $\Delta\nu$.

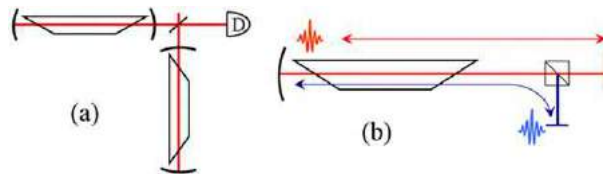


Fig. 8 (a) The active interferometer of Abramovici and Vager. Two gain sections are inserted in each branch of a Michelson-like interferometer. (b) IPI concept for a linear cavity [4]. The two pulses experience the same cavity and medium fluctuation with a half cavity round-trip time difference.

The small beat note sensitivity of IPI seems to contradict a theoretical proof by Abramovici and Vager [22] that active and passive cavity interferometers have comparable performances. The “active cavity interferometer” considered by Abramovici and Vager is as sketched in Fig. 8(a), with gain media in each branch of a Michelson-like interferometer. Clearly, the noise of each gain section is uncorrelated. In the linear implementation of IPI, there is only one gain medium in which each pulse circulates alternatively at several ns interval. For longer times (compared to the round-trip time) the two pulses see exactly the same condition, which explains why the noise is correlated.

References

- Diels, J.C., Rudolph, W., . Ultrashort laser pulse phenomena, second ed. Boston: Elsevier. ISBN 0-12-215492-4.
- Diels, J.C., Luo, X., Xu, X., Masuda, K., Arissian, L., . Definitions and control of the CEP and CEO. *Laser Physics* 20, 1038–1043.
- Arissian, L., Diels, J.C., . Investigation of carrier to envelope phase and repetition rate — fingerprints of mode-locked laser cavities. *Journal of Physics B: At. Mol. Opt. Phys* 42, 183001.
- Arissian, L., Diels, J.C., . Intracavity phase interferometry: Frequency combs inside a laser cavity. *Laser Photonics Rev* 8, 799–826.
- Arissian, L., Diels, J.C., . Carrier to envelope and dispersion control in a cavity with prism pairs. *Physical Review A* 75, 013814–013824.
- Knox, W.H., . Femtosecond intracavity dispersion measurements. In: Martin, J.L., Migus, A., Mourou, G.A., Zewail, A.H. (Eds.), *Ultrafast Phenomena VIII*. Berlin: Springer, pp. 192–193.
- Knox, W.H., . In situ measurement of complete intracavity dispersion in an operating Ti:Sapphire femtosecond laser. *Optics Letters* 17, 514–516.
- Jones, R.J., Diels, J.C., Jasapara, J., Rudolph, W., . Stabilization of the frequency, phase, and repetition rate of an ultra-short pulse train to a Fabry-Perot reference cavity. *Optics Communication* 175, 409–418.
- Udem, T., Reichert, J., Holzwarth, R., Hänsch, T., . Accurate measurement of large optical frequency differences with a mode-locked laser. *Optics Letters* 24, 881–883.
- Hudson, D.D., Holman, K.W., Jones, R.J., Cundiff, S.T., Ye, J., . Mode-locked fiber laser frequency-controlled with an intracavity electro-optic modulator. *Opt. Lett.* 30, 2948–2950.
- Yum, H.N., Salit, M., Yablon, J., Salit, K., Wang, Y., Shahriar, M.S., . Superluminal ring laser for hypersensitive sensing. *Optics Express* 18, 17658.
- Hendrie, J., Lenzner, M., Akhmiardakani, H., Diels, J.-C., Arissian, L., . Impact of resonant dispersion on the sensitivity of intracavity phase interferometry and laser gyros. *Optics Express* 24, 30402–304010.
- Basov, N.G., Ambartsumyan, R.V., Zuev, V.S., Kryukov, P.G., Letokhov, V.S., . Nonlinear amplifications of light pulses. *Soviet Physics JETP* 23, 16–22.
- Lecomte, C., Mainfray, G., Manus, C., Sanchez, F., . Laser temporal coherence effects on multiphoton ionization processes. *Phys. Rev. A* 11, 1009.
- Lompre, L.A., Mainfray, G., Manus, C., Thebault, J., . Multiphoton ionization of rare gases by a tunable-wavelength 30-psec laser pulse at 1.06 μm . *Physical Review A* 15, 1604–1612.
- Masuda, K., Hendrie, J., Diels, J.C., Arissian, L., . Envelope, group and phase velocities in a nested frequency comb. *Journal of Physics B* 49, 085402.

- Velten, A., Schmitt-Sody, A., Diels, J.C., . Precise intracavity phase measurement in an optical parametric oscillator with two pulses per cavity round-trip. *Optics Letters* 35, 1181–1183.
- Gowda, R., Nguyen, N., Diels, J.-C., Norwood, R., Peyghambarian, N., Kieu, K., . All-fiber bidirectional optical parametric oscillator for precision sensing. *Optics Letters* 40, 2033–2036.
- Zeng, C., Liu, X., Yun, L., . Bidirectional fiber soliton laser mode-locked by single-wall carbon nanotubes. *Optics Express* 21, 18937–18942.
- Ma, L.-S., Ye, J., Dube, P., Hall, J.L., . Ultrasensitive frequency-modulation spectroscopy enhanced by a high-finesse optical cavity. *Journal of the Optical Society of America B* 16, 2255.
- Hänsch, T.W., Schawlow, A.L., Toschek, P.E., . Ultrasensitive response of a cw dye laser to selective extinction. *IEEE Journal of Quantum Electronics* 8, 802–808.
- Abramovici, A., Vager, Z., . Comparison between active- and passive-cavity interferometers. *Physical Review A* 33, 3181–3184.

Few-Cycle and Attosecond Lasers

Francesca Calegari, Center for Free-Electron Laser Science, Hamburg, Germany; University of Hamburg, Hamburg, Germany; and Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

Caterina Vozzi, Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The tremendous advancements in the development of laser sources delivering few-optical-cycle pulses in the visible and near-infrared spectral ranges allowed to create even shorter light pulses, down to a few tens of attoseconds ($1 \text{ as} = 10^{-18} \text{ s}$), in the extreme ultraviolet (XUV) spectral region. This impressive progress in laser technology has given access to the electronic time-scale in matter (Calegari *et al.*, 2016). Attosecond pulses were first employed for the investigation of ultrafast electron dynamics in atomic systems, where for instance Auger decay, autoionization and photoemission delays have been measured in real time. In the past few years, attosecond science has successfully addressed physical process in more complex targets such as bio-relevant molecules and condensed matter.

In this article the main schemes for the generation and characterization of light pulses from few-femtosecond down to a few tens of attoseconds are described. The first part of the article introduces the femtosecond laser technology required for the generation of attosecond pulses. Section Hollow-Core Fiber Compression describes the hollow-core fiber approach as a post-compression stage for the generation of broadband laser pulses comprising only a few-optical cycles. A crucial requirement for the generation of isolated attosecond pulses with few-cycle pulses is a stable carrier-envelope-phase (CEP). Section Carrier-Envelope-Phase Stability provides a description of the most commonly used CEP stabilization techniques. The generation of tunable driving pulses in the infrared spectral region is briefly presented in Section OPA and OPCPA Technology, while the possibility to drive the process with single-cycle pulses is examined in Section Pulse Synthesis. An overview on the main techniques for the temporal characterization of such ultrashort driving pulses is presented in Section Ultrashort Pulse Characterization. The basis for the generation of attosecond pulses is then introduced, starting from the description of the high-order harmonic generation (HHG) process in Section High-Order Harmonic Generation. Techniques for gating HHG and isolate a single attosecond pulse from the attosecond pulse train are then presented in Section Gating Techniques. Section Attosecond Pulse Generation in the Water-Window Region discusses the possibility to extend the attosecond pulse generation process from the XUV to the soft-x energy region exploiting the driving laser sources described in Section OPA and OPCPA Technology. Finally a comprehensive description of all the temporal characterization methods for attosecond technology is reported in Section Attosecond Pulse Characterization.

Few-Cycle Lasers

Few-optical-cycle laser pulses with stabilized CEP are a fundamental prerequisite for the generation of attosecond pulses. Ti:sapphire lasers based on chirped pulse amplification (CPA) are the typical systems used to drive the HHG process. However, such laser systems routinely deliver 20 fs pulses and post compression techniques are required to shorten the pulse duration to sub-10 fs at 800-nm carrier wavelength. The most common technique for pulse compression of high-energy femtosecond pulses is based on propagation in a gas-filled hollow-core fiber in combination with ultra-broadband dispersion compensation. A promising alternative for scaling the generation of ultrashort laser pulses at the petawatt level is offered by the optical parametric chirped pulse amplification (OPCPA) technique, which combines optical parametric amplification (OPA) and CPA.

Hollow-Core Fiber Compression

In general, a post compression scheme for femtosecond pulses includes a phase modulator, which broadens the spectrum of the pulse, and a dispersion line, which compensates for the spectral phase thus making possible the achievement of near transform-limited pulse durations.

Self-phase modulation (SPM) in a Kerr medium produces an efficient spectral broadening. However, inhomogeneous spectral broadening affects the free space propagation of the pulses in a bulk medium because of the intensity profile of the beam. To overcome this limitation, guided propagation through a nonlinear medium can be used. Single-mode guiding line has to be employed otherwise intermodal dispersion distorts the pulse making its recompression impossible. This idea was first developed exploiting the propagation through optical fiber, where the nonlinear material is the fiber core itself, but the small diameter of single-mode optical fibers severely limited the maximum pulse energy to few tens of nJ in order to avoid material damage. Gas-filled hollow core fibers have been successfully demonstrated for pulse compression, overcoming this limitation. Rare gases are chosen as nonlinear medium, due to their high damage threshold and fast nonlinear response.

The hollow fiber consists of a fused silica capillary with an inner diameter of a few hundred microns. Light propagation in hollow fiber occurs by grazing incidence reflections at the dielectric inner surface of the fiber. The losses introduced by these reflections greatly limit the coupling of the incident radiation with high order modes for long enough fibers. Thus, unlike in total internal reflection, in

this propagation mechanism the radiative mode-dependent losses select one fundamental mode. By a proper choice of the hollow fiber inner diameter, coupling efficiency for Gaussian beams can exceed 98%. Equations that describe the propagation of the laser pulse through the gas filled hollow fiber are the same as for propagation in standard optical fibers. The fundamental mode sustained by the hollow fiber is the hybrid EH₁₁ mode, which corresponds to a truncated Bessel radial profile of the electric field.

The spectral broadening due to SPM in a hollow fiber of length L and radius a can be evaluated with the broadening factor F , which is defined as the ratio between the pulse bandwidth at the output of the fiber with respect to the same quantity at its input: $F = \Delta\omega/(\Delta\omega)_0$. A simple expression for this broadening factor can be derived as a function of the maximum phase shift $(\varphi)_m$ (Vozzi *et al.*, 2005).

$$F = \left(1 + \frac{4}{3\sqrt{3}} (\varphi)_m^2 \right)^{\frac{1}{2}} \quad (1)$$

where $(\varphi)_m = \gamma P_0 L_{eff}$, being γ the nonlinear coefficient of the hollow core fiber, P_0 the peak power of the pulse and $L_{eff} = [1 - e^{-\alpha L}]/\alpha$. $\frac{\alpha}{2}$ is the field attenuation constant of the fiber.

The nonlinear coefficient γ is given by the following relation:

$$\gamma = \frac{n_2 \omega_0}{c A_{eff}} \quad (2)$$

where n_2 is the nonlinear index parameter, ω_0 is the central frequency of the pulse, c is the speed of light and A_{eff} is the effective mode area of the fiber, given by $A_{eff} \cong 0.48 \pi a^2$. Thus for maximizing the broadening factor for a given incoming pulse with a certain energy and duration, one can increase the fiber length, increase the nonlinearity strength by increasing the gas pressure, or decrease the radius of the fiber. Several factors limit all these possibilities. The propagation losses of the fundamental mode, the distortion of the temporal pulse shape and some practical reasons limit the fiber length. Recently, new schemes based on a stretched flexible hollow fiber have been successfully implemented to overcome this last limitation. The pulse peak power should be lower than the critical power for self-focusing, in order to minimize the coupling of the fundamental transverse mode to the higher order modes which would introduce losses. The critical power for self-focusing is given by:

$$P_{crit} = \frac{\lambda_0^2}{2\kappa_2 p} \quad (3)$$

where κ_2 represents the ratio between the nonlinear coefficient and the gas pressure p and λ_0 is the central wavelength of the pulse.

A further limit on the fiber radius is set by the onset of ionization effects: The maximum pulse peak intensity, thus the minimum fiber radius a_{min} , must be chosen in order to have:

$$\frac{\Delta n_{kerr}}{\Delta n_p} \gg 1 \quad (4)$$

where $\Delta n_{kerr} = \kappa_2 p I$ is the refractive index variation due to the Kerr effect for a pulse of intensity I , while Δn_p is the plasma induced refractive index change, given by

$$\Delta n_p = \frac{\omega_p^2}{2\omega_0^2} \quad (5)$$

where $\omega_p = \sqrt{e^2 \rho_e / (m_e \epsilon_0)}$ is the plasma frequency, e and m_e are the electron charge and mass respectively and ρ_e is the free electron density in the gas.

The pulses emerging from the hollow core fiber have a group delay dispersion essentially due to the exit window of the cell that contains the fiber and to the air path after the fiber. In order to compensate for this group delay, chirped mirrors can be used, allowing compression down to sub 3-fs for mJ-energy pulses.

Carrier-Envelope-Phase Stability

The CEP ϕ of a light pulse is the phase delay between the highest half-cycle of the electric field associated to the pulse and the peak of the pulse envelope, as depicted in Fig. 1. Pulses propagating through dispersive media experience a CEP change due to the difference between the phase velocity and the group velocity. CEP control matters for few-optical-cycle pulses in all those phenomena that depend directly on the electric field, such as multiphoton absorption, above-threshold ionization and high-harmonic generation. In all these cases, the reproducibility of the electric waveform is essential. In particular, CEP stability is a fundamental prerequisite for the generation of reproducible isolated attosecond pulses. The CEP of standard laser sources is intrinsically unstable, but it is possible to stabilize it in either active or passive ways.

In mode-locked oscillators, the carrier propagates with the phase velocity and the envelope propagates with the group velocity, thus a systematic change in the CEP $\Delta\phi$ is introduced for each round trip. A slow drift also affects this phase change $\Delta\phi$ due to the change in the oscillator parameters. At the output of the oscillator cavity a sequence of pulses is emitted in such a way that a pulse with CEP ϕ is followed by a pulse with CEP $\phi + \Delta\phi + 2\pi$. It is then possible to obtain two pulses with the same CEP at the oscillator output by waiting $2\pi/\Delta\phi$ round trips. Thus, the CEP period can be defined as:

$$T_{CEP} = \frac{2\pi}{\Delta\phi} T_R \quad (6)$$

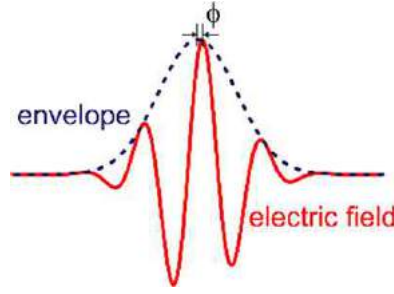


Fig. 1 Carrier-envelope-phase (CEP) ϕ of a light pulse.

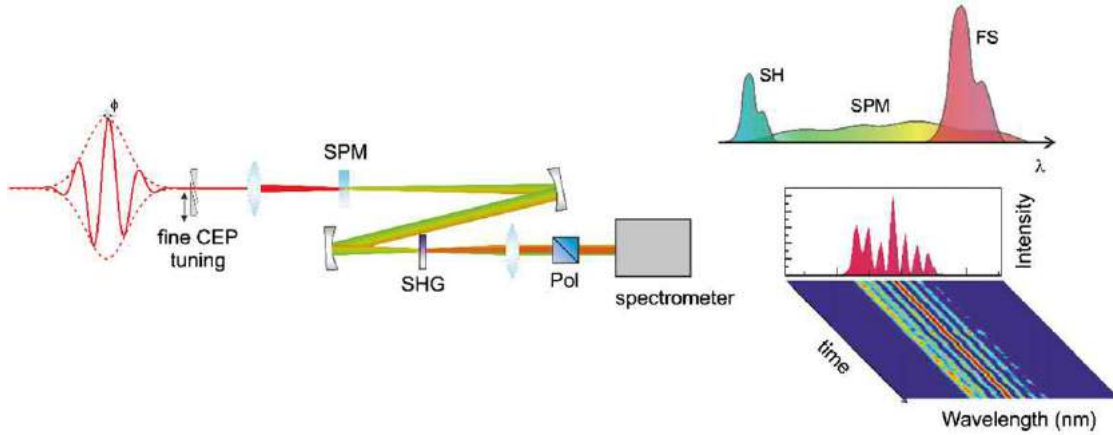


Fig. 2 Left panel: scheme of the interferometric measurement of the CEP drift by an f-to-2f interferometer. Right panel: spectral fringes. FS, fundamental spectrum; Pol, polarizer; SHG, second harmonic generation; SPM, self-phase modulation.

where, T_R is the period of a roundtrip in the laser cavity. The corresponding CEP frequency reads:

$$\omega_{CEP} = \frac{\Delta\phi}{2\pi} \omega_R \quad (7)$$

The last relation suggests that a train of pulses with the same value of CEP can be obtained by selecting the pulses at a frequency ω_{CEP} or at a fraction of it. The measurement of ω_{CEP} exploits the frequency comb emitted by the oscillator, which is the superposition of longitudinal modes with frequencies $\omega_n = \omega_{CEP} + n\omega_R$, with well-known metrology applications. In order to access the frequency ω_{CEP} , it is possible to measure the beating between the fundamental comb and its sum frequency in the spectral region where they overlap, assuming that the fundamental spectrum spans over an octave. In an f-to-2f interferometer, the fundamental comb with frequencies ω_n and a second harmonic comb at frequencies $\omega_{SH} = 2\omega_n = 2\omega_{CEP} + 2m\omega_R$ overlap and generate a beating for the modes with $n = 2m$ at the frequency ω_{CEP} . The frequency ω_{CEP} can then be stabilized by an active feedback on the oscillator, typically by changing the pump power with an acousto-optic modulator. In this way, one obtains a train of CEP-stable pulses at a fraction of the frequency ω_{CEP} that can be amplified in solid-state amplifiers or in optical parametric amplifiers.

In amplified systems with repetition rates of the order of a few kHz it is impossible to directly measure the comb frequencies. In this case, in order to characterize the CEP of the individual pulses it is possible to exploit interferometric techniques. Indeed the electric field associated with the pulse can be written as:

$$E(t) = A(t)e^{i(\omega t + \phi)} + c.c. = E_0(t)e^{i\phi} + c.c. = F\{\tilde{E}_0(\omega)\}e^{i\phi} + c.c. \quad (8)$$

If the pulse with fundamental frequency ω_0 and its m -th harmonic at frequency $m\omega_0$ overlap in some region of the pulse spectrum, in this spectral region an interference modulation appears

$$\cos[(m+1)\phi + \omega\tau + c] \quad (9)$$

where, τ is the delay between the two components at ω_0 and $m\omega_0$ and c is a constant. The shot-to-shot CEP shift can then be tracked by the direct observation of interference fringes. A typical setup for this kind of measurement is shown in Fig. 2. In this f-to-2f interferometer the fundamental pulse (F) is focused into a sapphire plate for generating white light with an octave spanning bandwidth. The fundamental pulse is then frequency doubled in a nonlinear crystal in such a way that the second harmonic (SH) spectrally overlaps the white light. These two spectral components are sent to a spectrometer through a polarizer. The intensity

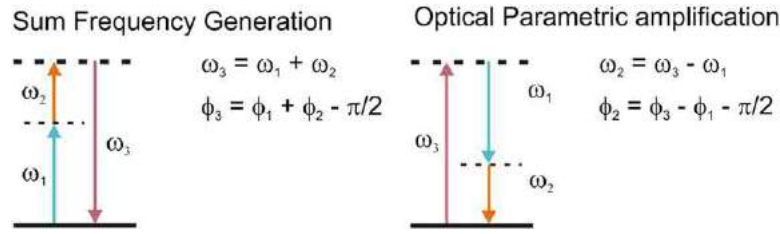


Fig. 3 Schematic view of how the CEP is affected by some common nonlinear phenomena.

measured by the spectrometer is modulated according to the following relation:

$$I(\omega) = I_F(\omega) + I_{SH}(\omega) + 2\sqrt{I_F(\omega)I_{SH}(\omega)}\cos(\omega\tau + \phi) \quad (10)$$

where τ is the delay between the fundamental pulse and its second harmonic. A drift in the CEP reflects in a shot-to-shot change of the position of the fringes in the interference pattern. This change can be used in a slow feedback loop for stabilizing slow-varying CEP noise in amplified systems. Commercially available amplified Ti:sapphire systems with active CEP stabilization show rms CEP fluctuations lower than 100 mrad. In order to measure the actual value of the CEP, more sophisticated instruments based on above threshold ionization (ATI) are necessary for the absolute calibration of the above mentioned interference trace.

An alternative approach to generate CEP stabilized pulses is the passive one, cancelling CEP fluctuations based on the nonlinear optical processes. The effect of these processes on the CEP can be considered as follows. In sum frequency generation (SFG), two pulses at frequency ω_1 and ω_2 and CEP ϕ_1 and ϕ_2 generate a third pulse at frequency $\omega_3 = \omega_1 + \omega_2$ with a CEP $\phi_3 = \phi_1 + \phi_2 + \frac{\pi}{2}$. Second harmonic generation is a particular case of SFG and thus from a pulse of frequency ω and CEP ϕ one obtains a pulse at frequency 2ω with CEP $2\phi + \frac{\pi}{2}$. In the difference frequency generation (DFG), two pulses at frequencies ω_1 and ω_2 and CEP ϕ_1 and ϕ_2 give a resulting frequency $\omega_3 = \omega_1 - \omega_2$ with CEP $\phi_3 = \phi_1 - \phi_2 - \frac{\pi}{2}$. Optical parametric amplification (OPA) is similar to DFG: the intense pump pulse at frequency ω_3 with CEP ϕ_3 amplifies the weak signal pulse at frequency ω_1 with CEP ϕ_1 and generate an idler pulse at frequency $\omega_2 = \omega_3 - \omega_1$. In the process the CEP of the signal is not affected, while the idler has a CEP $\phi_2 = \phi_3 - \phi_1 - \frac{\pi}{2}$ (Fig. 3). Self-phase modulation preserves the CEP and adds a constant phase shift of $\pi/2$.

Passive stabilization of the CEP is based on DFG between two pulses sharing the same CEP. The CEP of the DFG results as the difference between the phases of the two generating pulses, thus the shot-to-shot fluctuations of the CEP cancel. This approach allows the generation of CEP stable few cycle pulses in a broad spectral range from the visible to the mid-IR. This approach is all-optical and has the advantage of avoiding electronic feedback. Moreover, the train of CEP stable pulses is automatically generated at the same repetition rate of the driving field, without the need of picking up pulses at a fraction of the frequency ω_{CEP} . There are two possible configurations for implementing this method. In the *inter-pulse* configuration, two CEP-locked frequency-shifted pulses are synchronized by a delay line and then they are mixed in the nonlinear crystal for DFG. A drawback of this scheme is that the delay line can introduce CEP jitter due to mechanical instabilities. *Intra-pulse* schemes involve mixing between different frequencies of a single ultra-broadband pulse that generate the difference frequency. In this case, since the DFG occurs between spectral portions of the same pulse, the delay-induced CEP jitter is suppressed and the synchronization between the two components is automatically achieved by means of the control of the group delay of the pulse. A review of several laser systems implementing passive CEP stabilization is presented in Cerullo *et al.* (2010).

OPA and OPCPA Technology

Aside from the well established laser sources based on Ti:sapphire for HHG and attosecond pulse generation, innovative driving sources based on Optical Parametric Amplification (OPA) and Optical Parametric Chirped Pulse Amplification (OPCPA) have been proposed.

OPA is an excellent technique for the generation of high-energy, few-optical cycle pulses tunable in a wide spectral range. Passive stabilization of the CEP can be also implemented in OPAs. OPCPA-based sources promise to offer the possibility of generating ultrashort laser pulses with peak power in the PW range, overcoming some of the limitations on the scalability of OPA based lasers.

A review of recent developments in few-cycle OPAs and OPCPAs for HHG and attosecond generation is presented in Ciriolo *et al.* (2017).

Pulse Synthesis

Exploiting the above listed techniques, ultra-broadband pulses comprising just a few optical cycles are routinely produced. In order to achieve a pulse duration down to the single-cycle limit, coherent combination (or synthesis) of pulses having different colours is required.

One possible approach to achieve pulse synthesis consists of the generation of a super broadband spectrum (above one-octave) to be divided into several CEP-stable channels, which are individually compressed by custom-designed chirped mirrors (CMs) and then coherently recombined in a sub-cycle optical waveform. This passive approach is schematically shown in Fig. 4.

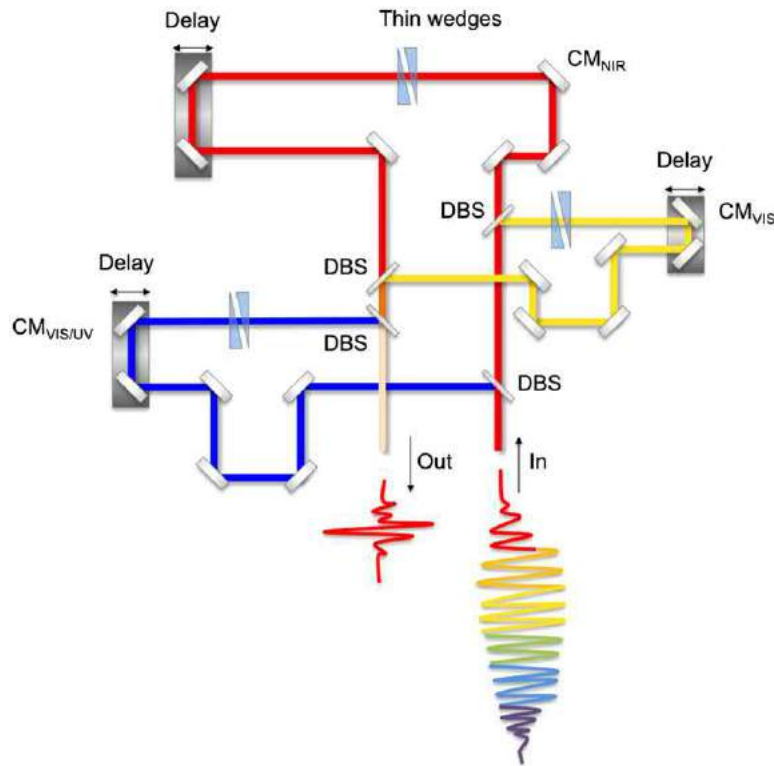


Fig. 4 Scheme of a three-channel light-field synthesizer. Dichroic beam splitters (DBS) are used to divide a supercontinuum into three channels: ultraviolet (UV), visible (VIS), near-infrared (NIR). Pulses of each channel are compressed using custom-designed chirped mirrors (CM). Fine-tuning of dispersion, CEP and delay is achieved with pairs of fused silica wedges and with delay stages in each channel, respectively.

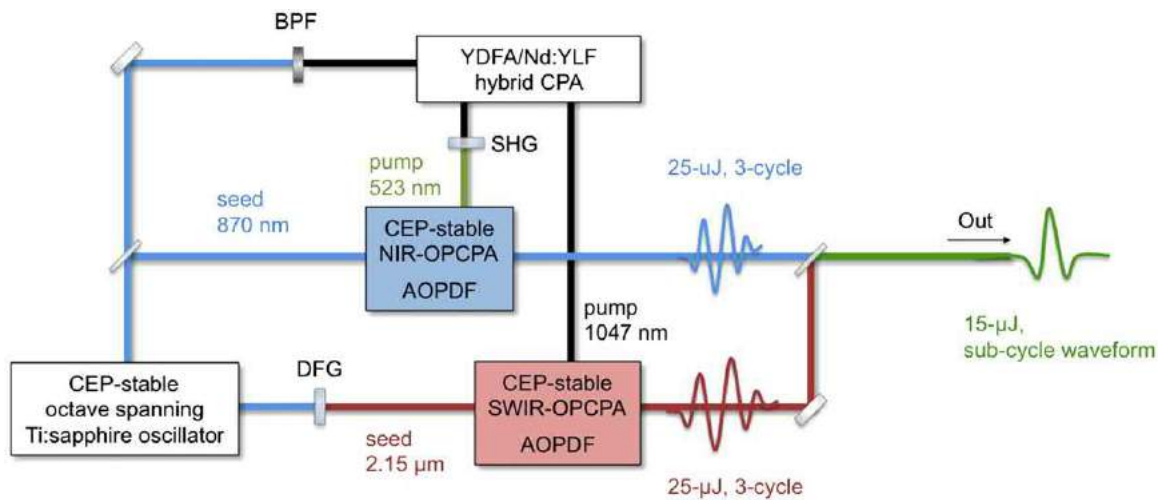


Fig. 5 Scheme of the OPCPA-based pulse synthesizer. A sub-cycle waveform spanning over 1.8 octaves is obtained by combining a NIR OPCPA and SWIR OPCPA. A common front-end is used for both OPCPAs. Waveform shaping is achieved by controlling the CEP of both channels. Control of the chirp is obtained with an acousto-optic programmable dispersive filter (AOPDF). BPF, band pass filter; DFG, difference frequency generation; SHG, second harmonic generation.

Coherent combination of ultra-broadband OPAs is another approach to single-cycle waveform synthesis. In this case, the synthesis is achieved by coherently combining two spectrally adjacent OPAs seeded by distinct portions of the same CEP-stable white light continuum. Energy scaling of this approach can be achieved by replacing the OPA with OPCPA technology. The design of an OPCPA-based pulse synthesizer is shown in Fig. 5. This design could be easily scaled to higher energies by implementing into the scheme higher-power pump lasers as for instance the new generation of cryo-Yb:YAG lasers.

A comprehensive review of the pulse synthesis methods is reported in [Manzoni et al. \(2015\)](#).

Ultrashort Pulse Characterization

The tremendous progress in the development of ultra-broadband pulses calls for a precise temporal characterization of these light transients. The time-dependent electric field of the laser pulse can be written as

$$E(t) = \text{Re} \left\{ \sqrt{I(t)} \exp(i\omega_0 t - i(\varphi)(t)) \right\} \quad (11)$$

where, $I(t)$ and $(\varphi)(t)$ are the time-dependent intensity and phase and ω_0 is the carrier frequency.

The main technique used to temporally characterize laser pulses is the autocorrelation. An autocorrelator is a device that involves splitting the pulse into two variably delayed replica, subsequently spatially overlapped in a second harmonic generation (SHG) crystal. The second harmonic signal acquired as a function of the time-delay between the two pulses provides the autocorrelation trace. In order to estimate $I(t)$ this method relies on a “guess” for the pulse shape, which is the main source of error. Moreover no information can be retrieved on the pulse phase $(\varphi)(t)$.

A large number of schemes have been developed over the past two decades to provide a better measurement of ultrashort laser pulses. Among them, the two most commonly used methods are the frequency-resolved optical gating (FROG) and the spectral phase interferometry for direct electric-field reconstruction (SPIDER), both providing a full characterization ($I(t), (\varphi)(t)$) of the laser pulse. FROG is an extension of the autocorrelation technique: when the second harmonic signal from an autocorrelator is acquired by a spectrometer, the autocorrelation becomes a two-dimensional spectrally-resolved trace (spectrogram). In general, the measured spectrogram can be expressed as:

$$S(\omega, \tau) = \left| \int E(t) g(t - \tau) \exp(-i\omega t) dt \right|^2 = \left| \int E_{\text{sig}}(t) \exp(-i\omega t) dt \right|^2 \quad (12)$$

where $g(t-\tau)$ is a gate function, which selects a portion of the electric field $E(t)$ around the delay τ . Extracting the laser field $E(t)$ from the measured FROG trace $S(\omega, \tau)$ is a two-dimensional phase retrieval problem, and iterative algorithms are required for finding a solution. Various versions of FROG exist, which rely on different nonlinear gating mechanisms.

Spectral interferometry is another approach to characterize a laser pulse. Ideally, the spectral interference between the pulse under test and a reference pulse with a well-known spectral phase should be used to extract the spectral phase of the unknown pulse. However, the reference pulse is usually not available. SPIDER is a technique introduced to overcome this limitation. The idea is to generate two upconverted spectra slightly shifted in frequency and to measure their spectral interference. The electric field of each individual pulse can be expressed as:

$$\begin{aligned} E_1(t) &= E(t) e^{i\omega_s t} \\ E_2(t) &= E(t - \tau) e^{i(\omega_s + \Omega)(t - \tau)} \end{aligned} \quad (13)$$

where ω_s and $\omega_s + \Omega$ are the two frequencies resulting from the non-linear mixing, Ω is called spectral shear, τ is the time delay between the two replicas and $E(t)$ is the electric field to be characterized. The measured spectrum can be written as:

$$S(\omega) = \left| \int_{-\infty}^{+\infty} (E_1(t) + E_2(t)) e^{-i\omega t} dt \right|^2 = S_{\text{DC}}(\omega) + S^+(\omega) e^{-i\omega\tau} + S^-(\omega) e^{i\omega\tau} \quad (14)$$

For a sufficiently large delay, the term $S^+(\omega) e^{-i\omega\tau}$ can be extracted and its phase can be uniquely determined:

$$\phi(\omega) = (\varphi)(\omega - \omega_s - \Omega) - (\varphi)(\omega - \omega_s) - \omega\tau \quad (15)$$

If the delay τ is precisely calibrated with an independent method, the linear phase $\omega\tau$ can be subtracted in [Eq. \(15\)](#) and the spectral phase of the unknown pulse can be determined by integration. A simple way to implement this scheme is shown in [Fig. 6](#).

A more detailed description of the ultrashort pulse characterization methods can be found in [Manzoni et al. \(2015\)](#).

Attosecond Technology

High-Order Harmonic Generation

High order harmonic generation (HHG) is a strongly nonlinear process occurring when atoms or molecules are exposed to laser radiation with intensity of the order of $I \sim 10^{14}$ W/cm². It represents a unique and table-top source of coherent XUV and soft X-ray radiation. This radiation has also very interesting temporal characteristics: it consists of a sequence of light bursts, with sub femtosecond duration.

The semi-classical three-step model ([Corkum, 1994](#)) gives a simple picture of the physical process leading to HHG, which is valid in the strong field regime and assuming a single active electron approximation. As sketched in [Fig. 7](#), the external laser field ionizes the outermost electron in the atom. The freed electron is then accelerated by the external field and brought back to the parent ion where it can recombine, giving rise to the emission of and XUV photon. The energy of the emitted photon depends on the kinetic energy of the recombining electron, thus on the trajectory followed by the electron in the external field. If we consider

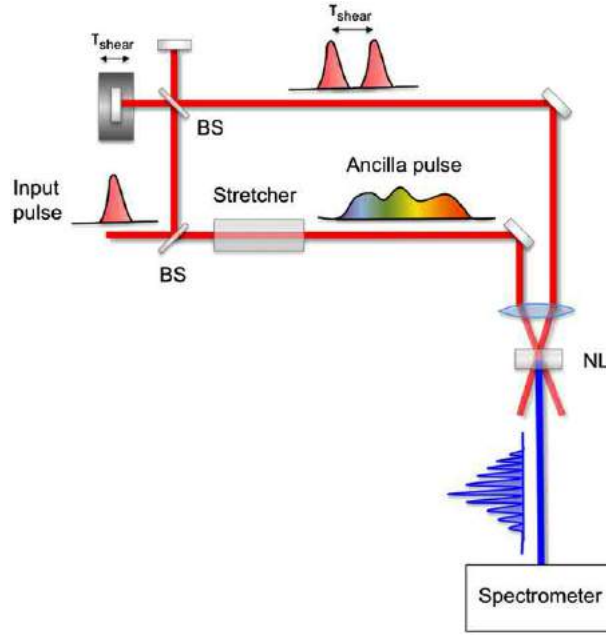


Fig. 6 Scheme of the SPIDER technique. Two delayed replicas of the input pulse are synchronized with two different frequency components of an ancilla pulse, separated by Ω . The upconversion process, occurring in a non-linear crystal (NL), frequency shifts the central frequency of the two pulses to ω_s and $\omega_s + \Omega$ respectively. A spectrometer detects the interference pattern. BS, beam splitter.

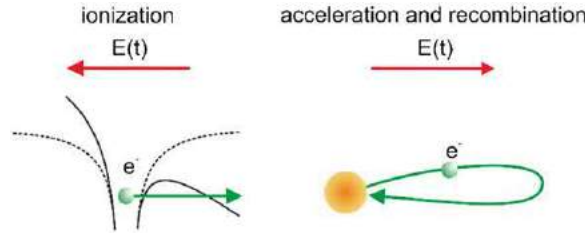


Fig. 7 Sketch of the HHG process as depicted in the three step model: on the left the active electron is ionized by the external laser field; on the right the electron is accelerated by the external field and can eventually recombine with the parent ion.

the interaction of an atom with a monochromatic electric field linearly polarized along the x direction $E(t) = E_0 \cos(\omega_0 t)$, the electron ionized at the instant t' , is accelerated by the external field and depending on the ionization time, it can be driven away or it returns to the parent ion. A photon of frequency $\omega = (I_p + E_k/\hbar)$ is emitted upon recombination, being I_p the atom ionization potential and E_k the kinetic energy gained by the electron in the continuum. The electron motion equation in the external field can be integrated assuming initial conditions $x(t') = 0$ and $v(t') = 0$ and for a given ionization time t' , it is possible to calculate the recombination time t . For some ionization times, the electron revisits the parent ion several times, but as a first approximation, the successive recollisions can be neglected due to the quantum diffusion of the electron wave function, which reduces the recombination probability. This process is repeated periodically every half cycle of the external field and, due to the inversion symmetry of the atom, only the ionization and recollision events within this time interval must be considered. As shown in Fig. 8, for any photon energy, there are two solutions of the motion equation characterized by the different values of the time by the electron in the continuum. The solutions associated to the shorter time travel times are called short trajectories, whilst the other ones are referred to as long trajectories. The maximum kinetic energy gained by the electron in the continuum turns out to be $(E_k)_{\max} = 3.17U_p$ where U_p is the ponderomotive energy given by

$$U_p = \frac{e^2 E_0^2}{4m_e \omega_0} \quad (16)$$

Since the laser pulse contains a few optical cycles, in the temporal domain the XUV emission corresponds to a sequence of XUV pulses separated in time by half-cycle of the driving field, as is depicted in Fig. 9. In the spectral domain, this emission corresponds to odd harmonics of the fundamental laser frequency. The harmonic spectrum contains two distinct regions. In the *plateau* region, the intensity of the harmonic is almost flat as a function of harmonic order. In the *cut-off* region, which comprises only a few harmonic orders, the harmonic yield decreases exponentially.

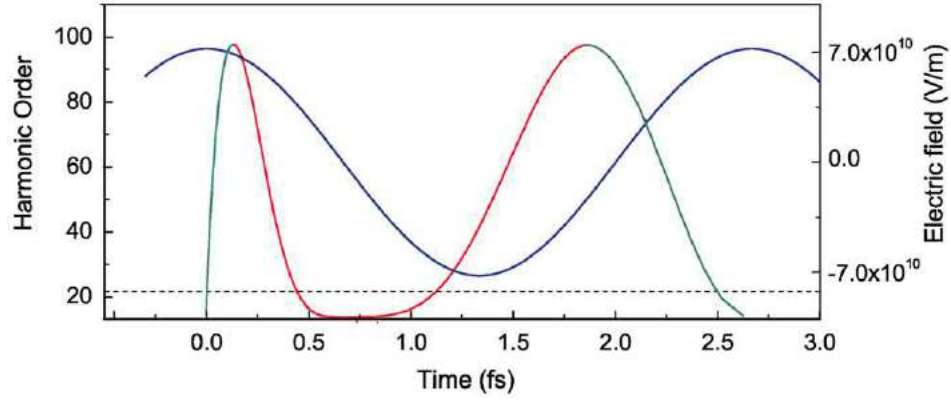


Fig. 8 Solution of the classical motion equation for the electron in a monochromatic external field of the form $E(t) = E_0 \cos(\omega_0 t)$ with $E_0 = 7.2 \cdot 10^{10}$ V/m and $\omega_0 = 2.36 \cdot 10^{15}$ rad/s. The electric field evolution is shown as a blue curve. The left-hand side of the curve individuates the ionization times t' for short (red) and long (green) electron trajectories. On the right-hand side of the curve, the recombination times for short (red) and long (green) electron trajectories are shown.

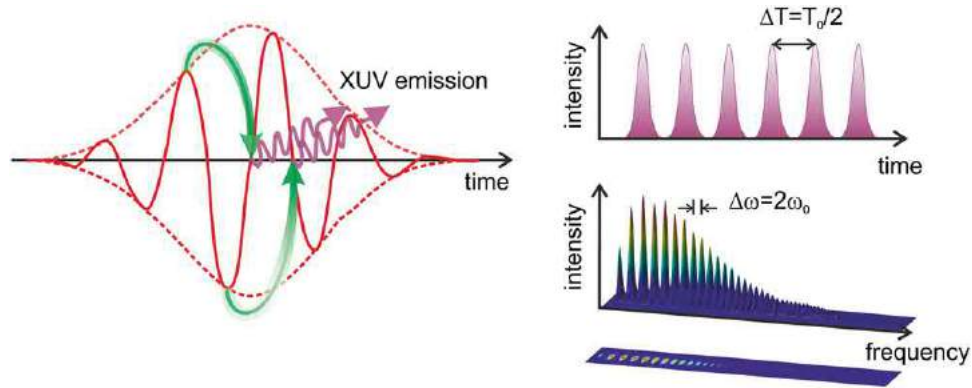


Fig. 9 Characteristics of harmonic emission. Left panel: the HHG process is repeated periodically every half cycle. Right panel: harmonic emission corresponds to a train of attosecond pulses in the temporal domain (upper figure) and to a sequence of odd harmonics of the fundamental driving frequency (lower figure).

The semi-classical model allows grasping most of the peculiarities of harmonic emission, nevertheless a full description of the laser-atom interaction requires a quantum model, such as the Lewenstein model (Lewenstein *et al.*, 1994). In this framework the harmonic emission spectrum is proportional to the Fourier transform of the atomic dipole moment:

$$x(t) = \langle \psi(t) | e \mathbf{r} \cdot \mathbf{E}(t) | \psi(t) \rangle \quad (17)$$

where $|\psi(t)\rangle$ is the solution of Schrödinger equation and $e \mathbf{r}$ is the dipole moment operator. A simple analytical expression for the atomic dipole moment can be derived with the assumptions of single active electron (SAE – the atom is considered as a hydrogen-like system and multiple ionization is neglected) and strong field approximation (SFA – the influence of the atomic Coulomb potential on the motion of the electron in the continuum is neglected and the external electric field has no influence on the atomic ground state wave function):

$$x(t) = i \int_{-\infty}^t dt' \int d^3p E(t') d_x^* [p - A(t)] d_x [p - A(t')] e^{-iS(p,t,t')} + c.c. \quad (18)$$

where atomic units have been used and $E(t')$ is the external electric field; p is the canonical momentum given by $p = v + A(t)$, v is the electron velocity, $A(t)$ is the vector potential associated to the external field, $d_x [p - A(t)]$ is the expectation value for the dipole moment transition between the atomic ground state and the continuum state associated to electron velocity $v = p - A(t)$, $S(p,t,t')$ is the quasi-classical action defined as:

$$S(p,t,t') = \int_{t'}^t dt'' \left(\frac{[p - A(t'')]^2}{2} + I_p \right) \quad (19)$$

which corresponds to the phase accumulated by the free electron during its propagation in the continuum.

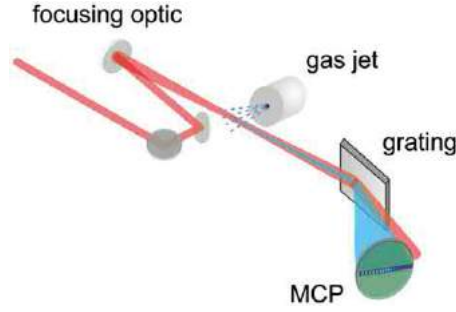


Fig. 10 Typical setup for harmonic generation and detection. MCP, micro channel plate.

The single atom harmonic spectrum can be calculated as:

$$\tilde{x}(\omega) \propto \int_0^{T_0} dt' x(t') e^{i\omega t'} \quad (20)$$

Thus the dipole moment at the time t is given by the contribution of all the electrons injected in the continuum at the time t' with a momentum p . $E(t') d_x[p - A(t')]$ is the probability that an electron is ionized by the external field at time t and appears in the continuum with a momentum p . The electrons accelerated by the external field accumulate a phase $S(p, t, t')$ and eventually recombine with parent ions at the time t with a probability $d_x^*[p - A(t)]$.

The relevant trajectories contributing to the Lewenstein integral can be selected with the saddle point approximation (SPA). The trajectories found with the three-step model correspond to the classical limit of the saddle-point quantum paths. The real part of such quantum paths is mostly similar to the classical electron trajectories while the imaginary part is related to the quantum process of tunnel ionization.

The harmonic spectrum results from the coherent superposition of single-atom contributions. Thus for the physical interpretation of experimental results the propagation of harmonic radiation in the atomic gas has to be taken into account. For maximizing the harmonic conversion efficiency, phase-matching conditions between the driving pulse and the co-propagating harmonic radiation must be fulfilled. Several factors contribute to phase matching in HHG, such as the phase $S(p, t, t')$ accumulated by the electron in the continuum, the time-dependent refractive index due to the plasma induced by the fundamental pulse and the geometrical phase factor related to the focusing geometry of the fundamental field (Guoy phase). In the case of a Gaussian driving pulse, when the generating medium is located around the laser focus – see Fig. 10 for a sketch of the typical experimental setup – the phase matching condition is fulfilled for harmonic radiation generated off-axis and the structure of harmonic beam is annular. On the other hand, if the atomic medium is located after the laser focus the harmonic beam is efficiently generated along the driving pulse propagation direction. Furthermore, as the dipole phase term depends on the quantum trajectory, phase-matching conditions can enhance or depress the contribution of short or long trajectories. For a Gaussian beam, when the atomic medium is placed after the laser focus, the contribution of the long quantum paths is suppressed while when the atomic medium is placed in correspondence of the laser focus the contribution of the long quantum paths increases.

The optimization of phase-matching conditions is essential for maximizing the HHG yield and several strategies have been demonstrated so far, such as loose focusing geometry, hollow waveguide for the generating medium or periodic modulation of the gas density.

Gating Techniques

For several time-resolved applications, it is required to generate a single attosecond pulse instead of an attosecond pulse train. Various schemes have been developed to confine the harmonic generation process to a single event. These schemes can be grouped into two main categories: amplitude gating and temporal gating. In the amplitude gating scheme the selection of a single attosecond pulse is done by spectrally filtering the cutoff energy region of the harmonic spectrum, where only the most intense half cycle of the driving pulse contributes to the emission (left panel of Fig. 11). This generation scheme requires the use of intense few-optical-cycle driving pulses, with controlled CEP. Isolated attosecond pulses with a temporal duration down to 80 as have been generated using this approach (Goulielmakis *et al.*, 2008).

In the case of temporal gating, HHG is confined to a single emission event by properly manipulating the driving electric field in time, without limitation in the generated XUV pulse bandwidth (right panel of Fig. 11). One possible temporal gating approach is called polarization gating (PG) and it takes advantage of the strong sensitivity of the HHG process on the degree of ellipticity of the fundamental field. If the polarization of the driving field is temporally manipulated from circular to linear and back to circular, the XUV emission is limited to the temporal window in which the driving pulse is linearly polarized. Top panel of Fig. 12 shows a possible scheme for the implementation of PG: a first quartz plate is used to separate the linearly polarized input pulse into two orthogonally polarized delayed pulses, so that the outgoing pulse polarization evolves from linear to circular, and back to linear. A zero-order broadband quarter waveplate is then used to obtain a pulse linearly polarized at its centre and circularly polarized in its

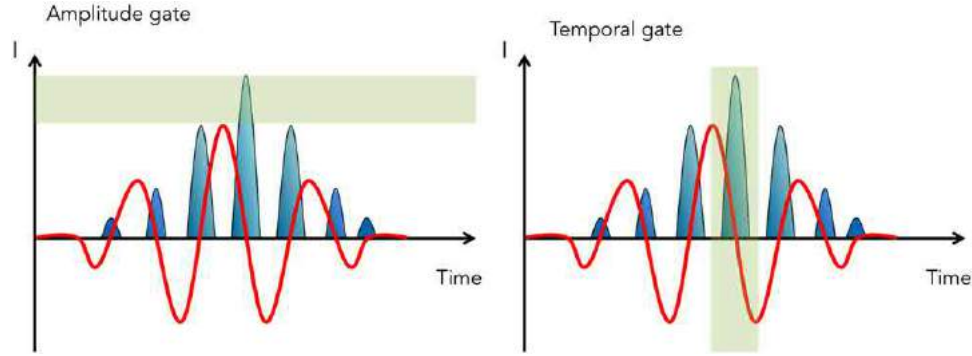


Fig. 11 Operating principle of the main gating schemes for the generation of isolated attosecond pulses. The red line represents the electric field of the driving pulse and the blue shaded area represents the corresponding ionization times. A single attosecond pulse can be selected either by spectrally filtering the cut-off energies of the harmonic spectrum (left panel) or by temporally gating the recombination process (right panel).

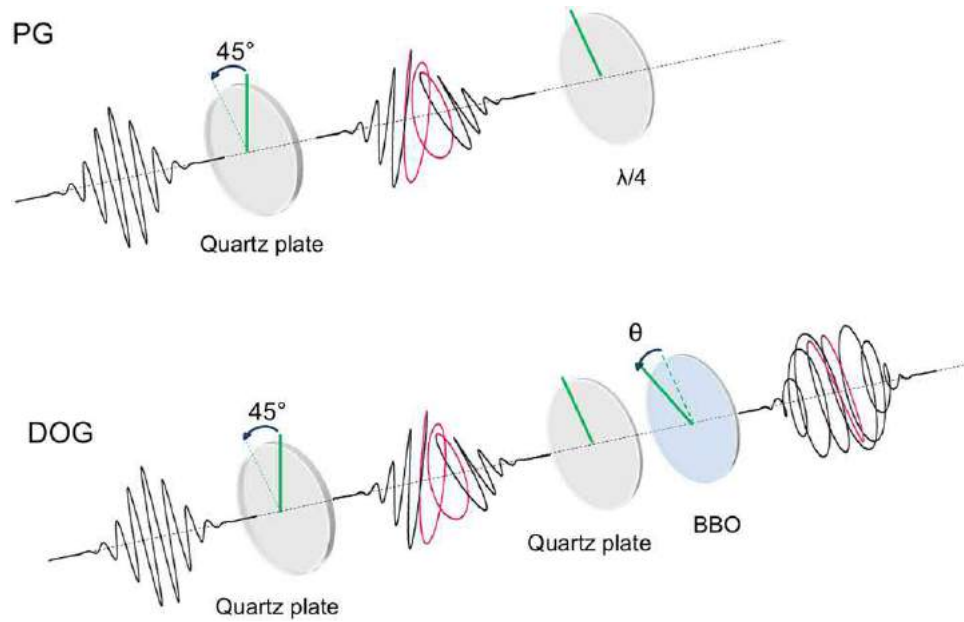


Fig. 12 Optical scheme for polarization gating (top panel) and for double optical gating (bottom panel).

wings. The duration of the temporal gate δt_g where the polarization is linear and HHG can occur can be estimated as:

$$\delta t_g = 0.3 \frac{\epsilon_{thr} \tau_p^2}{\tau_d} \quad (21)$$

where τ_d is the temporal delay between the two orthogonally polarized pulses at the output of the first quartz plate, τ_p is the pulse duration and ϵ_{thr} is a threshold value of ellipticity for which the harmonic signal is decreased to 50%. By using this approach isolated attosecond pulses as short as 130 as have been generated and characterized (Sansone *et al.*, 2006).

In the PG approach, the gate window should be approximately $T/2$, where T is the period of the driving field. By using Eq. (21), it is possible to demonstrate that such a narrow gate requires that $\tau_p = 2.5T$, which corresponds to less than 7 fs at 800 nm carrier wavelength. The double optical gating (DOG) technique is an extension of the PG method, which relaxes the requirement on the pulse duration. The idea is to add the second harmonic of the driving field to the fundamental pulse to increase the separation between two XUV emission events by a factor of two. This allows one to apply a one-optical-cycle gate instead of a half-cycle gate. The optical scheme for the DOG method is depicted in bottom panel of Fig. 12 and it differs from the PG scheme only for the thickness of the quartz plates and the presence of a BBO crystal for second harmonic generation. In this case, the second quartz plate and the BBO crystal act together as a quarter wave plate. The DOG technique has been successfully implemented to generate 67-as isolated pulses. A generalized version of the DOG (GDOG) also exists, allowing for the generation of isolated attosecond pulses from multi-cycles pulses to be achieved.

Another possible approach for temporally confining the XUV emission is dubbed ionization gating (IG) and it exploits high-intensity driving fields. For high excitation intensities and ultrashort pulse durations, the plasma density rapidly increases on the leading edge of the pulse, thus creating a single-atom sub-cycle ionization response that can be exploited for temporal gating. With

this technique, isolated attosecond pulses with pulse duration of 155 as and pulse energy on target of 2.1 nJ were generated in xenon (Ferrari *et al.*, 2009).

An alternative approach to temporal gating is spatial gating, in which the selection of a single attosecond pulse is obtained by spatially filtering the EUV radiation. The idea is to drive HHG using ultrashort pulses with a non-negligible pulse-front tilt, so that each attosecond pulse is emitted in a slightly different direction, which corresponds to the instantaneous direction of propagation of the driving field at the instant of generation. Assuming that the angular separation between two consecutive XUV pulses is larger than their divergence, in the farfield it is possible to spatially select each individual attosecond pulse. Since isolated attosecond pulses are emitted in different directions, the effect has been dubbed attosecond lighthouse. Angular separation between two consecutive attosecond pulses can be also achieved by focusing two non-collinear laser beams at the position of the gas target, the method is called noncollinear optical gating (NOG).

Finally it is worth mentioning that most of the gating techniques require CEP stability of the driving few-optical-cycle pulse. CEP-control can be also exploited to tailor the XUV emission: by changing the CEP value it is indeed possible to spectrally tune the XUV emission as well as to select the emission of an isolated attosecond pulse or a pair of attosecond pulses. A more detailed description, with appropriate references, to all the above-mentioned gating methods is reported in Calegari *et al.* (2016).

Attosecond Pulse Generation in the Water-Window Region

According to the three-step model, the harmonic cut-off scales proportionally to the ponderomotive energy, thus it increases with the driving wavelength squared: $\hbar\omega_{\text{cut-off}} \propto I\lambda^2$, where I and λ are the peak intensity and the wavelength of the driving field respectively. This result has been confirmed experimentally and it implies that mid-IR driving sources can be very promising for the generation of XUV photons with very high energy, since for the same peak intensity of 800-nm Ti:Sapphire lasers they allow a huge extension of the harmonic cut-off toward the X-rays.

Unfortunately for longer driving pulses the excursion time of the electron in the continuum increases and the probability of recombination with the parent ion strongly decreases due to the spatial spread of the electron wave function. For the single atom, the harmonic power spectrum yield decreases with the driving wavelength roughly as λ^{-6} .

Since the possibility of generating coherent radiation in the water window spectral region is very intriguing, specific generation geometries have been developed for mid-IR driving sources. In particular, the unfavourable scaling of the single-atom response can be compensated by a consistent increase of the generating medium gas pressure (Popmintchev *et al.*, 2012). In this way, a cutoff energy of 1.6 keV was observed in HHG driven by 3.9 μm driving pulses in a helium-filled capillary.

Another important feature of HHG is the attochirp of the emitted XUV bursts. This is the natural dispersion of photon energies along the temporal duration of a single XUV attosecond burst emitted during HHG: recalling the three-step model, different flight times (trajectories) of the electron correspond to different kinetic energies, hence, different photon energies are emitted at different times. The attochirp scales as λ^{-1} , thus it decreases for increasing the driving wavelength. Indeed the attochirp is proportional to the ratio between the period of the fundamental pulse and the harmonic cutoff energy. At constant intensity, the former scales as λ and the latter scales as λ^2 , thus the attochirp scales as the inverse of the driving wavelength. The scaling of the attochirp with the driving wavelength also implies that shorter attosecond pulses could be emitted when harmonics are driven by mid-IR sources.

Attosecond Pulse Characterization

Femtosecond metrology allows for the complete characterization of ultrashort light pulses in the visible or near-visible range. As mentioned above, such a complete characterization requires the use of at least one time-nonstationary filter, typically obtained with a nonlinear effect. Transferring this idea to attosecond metrology is exceedingly difficult primarily due to limitations imposed by the low XUV photon flux. For this reason, other approaches have been implemented for characterizing attosecond pulses, which can be mainly divided in two categories. In the first approach, called *ex situ*, the photoelectrons produced in a target medium by the attosecond pulse are perturbed by the presence of a synchronized laser field, while, the second approach, called *in situ*, is based on the gentle perturbation of the electron trajectories during the attosecond pulse generation process itself.

Among the *ex situ* techniques, the Reconstruction of Attosecond Beating By Interference of Two-photon Transitions (RABBITT) was the first proposed method (Paul *et al.*, 2001). In the RABBITT scheme, the XUV pulse train is synchronized with a weak IR laser field to produce electrons by photoionization of a gas target. As shown in Fig. 13, the XUV pulse train generate a photoelectron spectrum, which is a replica of the harmonic spectrum: the peaks are separated by $2\hbar\omega$, where ω is the carrier frequency of the IR field. When the IR field is properly synchronized with the XUV field, additional peaks at $\pm\hbar\omega$ are produced in the photoelectron spectrum, thus leading to the appearance of the so-called sidebands. These peaks are due to the absorption or emission of one (or more) IR photons together with the absorption of one XUV photon. Typically the IR intensity is chosen such that only one IR photon can be absorbed or emitted. As depicted in the left panel of Fig. 13, the same sideband can originate from two different paths: (i) the absorption of one XUV photon corresponding to harmonic $q + 1$ and the emission of one IR photon, or (ii) the absorption of one XUV photon corresponding to harmonic $q - 1$ plus one IR photon. The two indistinguishable paths lead to interference, and the amplitude of the sidebands (SB) is periodically modulated as a function of the delay τ between the XUV and the IR pulses as:

$$SB = A_f \cos \left(2\omega\tau - \Delta(\varphi) - \Delta(\varphi_f) \right) \quad (22)$$

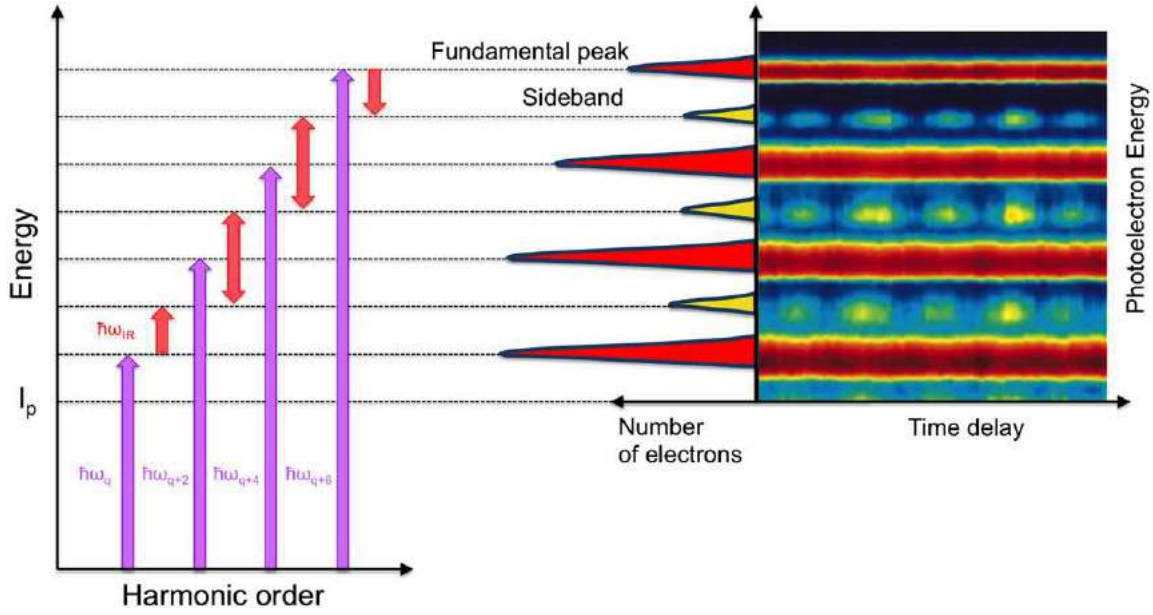


Fig. 13 RABBITT scheme. Left panel: odd order harmonics ionize the medium and create a photoelectron signal (red peaks). Further absorption/emission of an IR photon (red arrows) creates sideband peaks in the photoelectron spectrum (yellow peaks). Right panel: sideband amplitude oscillation at twice the frequency of the IR field.

where A_f is the amplitude, $\Delta(\varphi) = (\varphi)_{q+1} - (\varphi)_{q-1}$ is the phase difference between harmonics $q+1$ and $q-1$, and $\Delta(\varphi)_f$ is the intrinsic atomic phase difference between the matrix elements corresponding to photo-ionization from the $q+1$ and $q-1$ harmonics. This equation clearly indicates that the interference results in amplitude oscillations at twice the IR frequency upon changing the delay τ (see right panel of Fig. 13). If a suitable gas is used as a target, the intrinsic phase $\Delta(\varphi)_f$ can be calculated and from the time-delay scan the phase difference between two consecutive harmonics can be extracted.

In a typical optical scheme for RABBITT, a 800-nm laser is split in two arms: a portion of the beam is used to generate the harmonics in a noble-gas jet, while the remaining part of the IR beam is properly delayed and then collinearly focused together with the XUV beam in a second gas jet, producing photoionization. The resulting photoelectron spectrum is generally analysed with a time-of-flight (TOF) electron spectrometer as a function of the relative delay between the IR and XUV pulses.

As for trains of attosecond pulses, isolated attosecond pulses can be characterized using a method based on the measurement of the electrons photoionized by the XUV pulse in the co-presence of an IR field. The method is dubbed attosecond streak-camera (Mairesse and Quéré, 2005). The experimental setup is very similar to what is used for RABBITT, except that the IR pulse (streaking pulse) intensity in this case is low enough not to ionize atoms but high enough to impart substantial momentum to the photoelectrons liberated by the XUV pulse. If the dipole transition matrix element does not vary significantly (in phase and amplitude) over the XUV energy range, the photoelectron wave packet can be considered as a replica of the attosecond pulse. The streaking field modulates in time the phase of the electron wave packet as:

$$\Phi(t) = - \int_t^{+\infty} dt' \left[\vec{v} \cdot \vec{A}(t') + \vec{A}^2(t')/2 \right] = - \int_t^{+\infty} dt' U_p(t') + \frac{\sqrt{8WU_p}}{\omega} \cos \theta \cos \omega t - (U_p/2\omega) \sin 2\omega t \quad (23)$$

where ω is the frequency of the IR field, U_p is the ponderomotive potential of the IR pulse, θ is the angle between the electron velocity $\vec{v}$ and the vector potential $\vec{A}$, and $W = p^2/2$ is the kinetic energy of the emitted electron. As shown in bottom panel of Fig. 14, a modulation in the temporal phase corresponds to an energy shift of the photoelectron spectrum, which follows the shape of the vector potential of the IR pulse. Thus, by measuring the streaking effect on the photoelectron distribution for different time delays, it is possible to estimate the temporal duration of the XUV pulse.

The attosecond streak camera has a strong analogy to the FROG technique used to characterize femtosecond optical pulses. Indeed, the phase modulation induced by the IR streaking pulse can be seen as a phase gate allowing for a FROG-like measurement. The extension of FROG to the attosecond domain is called FROG-CRAB (FROG for Complete Reconstruction of Attosecond Bursts). The analogy with the FROG technique can be easily identified by comparing Eq. (12) with the streaking spectrogram measured in a given observation direction:

$$S(\vec{v}, \tau) = \left| \int \exp(i\Phi(t)) \vec{d}_{\vec{p}-\vec{A}(t)} \cdot \vec{E}_{XUV}(t-\tau) \exp(i(W+I_p)t) dt \right|^2 \quad (24)$$

where $\Phi(t)$ is the phase reported in Eq. (22), $\vec{d}$ is the dipole matrix element, I_p is the ionization potential of the medium, and $\vec{E}_{XUV}$ is the field to be characterized. The streaking spectrogram thus corresponds to a FROG trace where the temporal gate is a

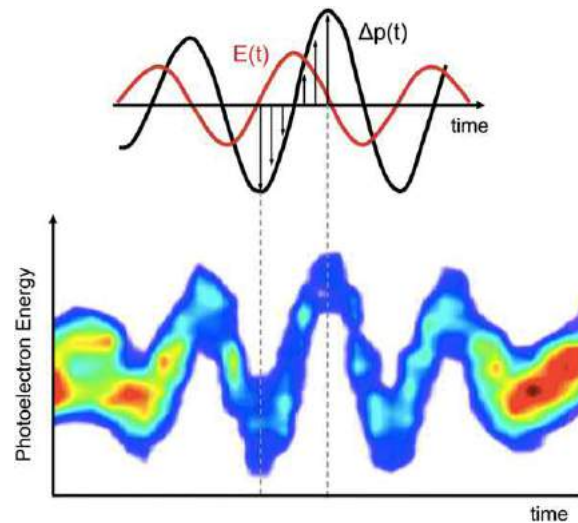


Fig. 14 Attosecond streak camera. Top panel: a photoelectron is released after XUV ionization and the IR probe pulse transfers to it an additional momentum Δp , which depends on the phase and amplitude of the IR electric field $E(t)$. Bottom panel: streaked photoelectron spectra acquired as a function of the XUV-IR time delay. The acquired spectrogram follows the shape of the vector potential of the IR pulse.

pure phase gate: $g(t) = \exp(i\Phi(t))$. Iterative inversion algorithms such as the very efficient Principal Component Generalized Projections Algorithm (PCGPA) can be used to extract from the CRAB trace the spectral phase of the attosecond pulse. Alternatively to the attosecond streak camera, the Phase Retrieval by Omega Oscillation Filtering (PROOF) can be used. This method takes advantage of the same approach as RABBITT, using a weak perturbing IR field. The reconstruction procedure does not require an iterative method and it does not rely on the central momentum approximation as in the case of FROG-CRAB, thus making the technique particularly suitable for characterizing ultra-broadband attosecond pulses (Chini *et al.*, 2010).

Finally, *in situ* techniques can be used for the temporal characterization of the attosecond pulse. These techniques rely on an all-optical method – solely based on the measurement of the XUV photon spectrum – for the reconstruction of the phase of the attosecond pulse, thus overcoming the technical challenge of measuring photoelectron spectra. The typical approach to implement an *in situ* characterization is to use a weak second-harmonic perturbing field to slightly alter the XUV generation mechanisms (Kim *et al.*, 2014). By implementing methods based on this idea, both trains and isolated attosecond pulses have been fully characterized.

See also: Attosecond Spectroscopy

References

- Calegari, F., *et al.*, 2016. *Journal of Physics B: Atomic, Molecular and Optical Physics* 49, 062001.
- Cerullo, G., *et al.*, 2010. *Laser and Photonics Reviews*. 1–29.
- Chini, M., *et al.*, 2010. *Optical Express* 18, 13006.
- Ciriolo, A.G., *et al.*, 2017. *Applied Science* 2017 (7), 265.
- Corkum, P., 1994. *Physical Review Letters* 71 (1993),
- Ferrari, F., *et al.*, 2009. *Nature Photonics* 4, 875.
- Goulielmakis, E., *et al.*, 2008. *Science* 320, 1614.
- Kim, K.T., *et al.*, 2014. *Nature Photonics* 8, 187.
- Lewenstein, M., *et al.*, 1994. *Physical Review A* 49, 2117.
- Mairesse, Y., Quéré, F., 2005. *Physical Review A* 71, 011401(R).
- Manzoni, C., *et al.*, 2015. *Laser Photonics Reviews* 9 (2), 129.
- Paul, P., *et al.*, 2001. *Science* 292, 1689.
- Popmintchev, T., *et al.*, 2012. *Science* 336, 1287.
- Sansone, G., *et al.*, 2006. *Science* 314, 443.
- Vozzi, C., *et al.*, 2005. *Applied Physics B* 80, 285.

Chirped Pulse Amplification

GA Mourou, University of Michigan, Ann Arbor, MI, USA

© 2018 Elsevier Inc. All rights reserved.

In the five years after the invention of the laser in 1960, tabletop lasers advanced in a series of technological leaps to reach a power of one gigawatt (10^9 W). For the next 20 years, progress was stymied and the maximum power of tabletop laser systems did not grow. The sole way to increase power was to build ever larger lasers. Trying to operate beyond the limiting intensity would create unwanted nonlinear effects in components of the laser, impairing the beam quality and even damaging the components. Only in 1985 was this optical damage problem circumvented with the introduction of chirped pulse amplification (CPA), a technique developed by the research group led by the author. Tabletop laser powers then leaped ahead by factors of 10^3 to 10^5 . CPA is used in all ultrashort pulse amplifiers from the very small, such as the fiber-based amplifier to the extremely large, such as the ones used in inertial confinement fusion, for fast ignition.

'Chirping' a signal or a wave means stretching it in time, arranging the frequencies from blue to red or vice versa. In chirped pulse amplification, the first step is to produce a short pulse with an oscillator and stretch it, usually to 10^3 to 10^5 times as long, by using a pair of diffraction gratings (see Fig. 1). This operation decreases the intensity of the pulse by the same amount. Standard laser amplification techniques can now be applied to increase the laser pulse energy by up to 10^3 to 10^5 times what we could achieve if we were not stretching the pulse. Finally, when the amplifier energy is fully extracted a sturdy device, such as a pair of diffraction gratings – sometimes in vacuum – recompresses the pulse to its original duration, by regrouping all the frequencies – increasing its power 10^3 to 10^5 times beyond the amplifier's limit. A typical example would begin with a seed pulse lasting 100 femtoseconds and having 0.2 nanojoule of energy. We stretch it by a factor of 10^4 to a nanosecond (reducing its power from about two kilowatts to 0.2 watt) and amplify it by ten orders of magnitude to two joules and two gigawatts. Recompressing the pulse to 100 femtoseconds increases the power to 20 terawatts. Without this method, sending the original two-kilowatt pulse through a tabletop amplifier would have destroyed the amplifier – unless we increased the amplifier's cross-sectional area 10^4 times and dispersed the beam across it. The CPA technique makes it possible to use conventional laser amplifiers and to stay below the onset of nonlinear effects.

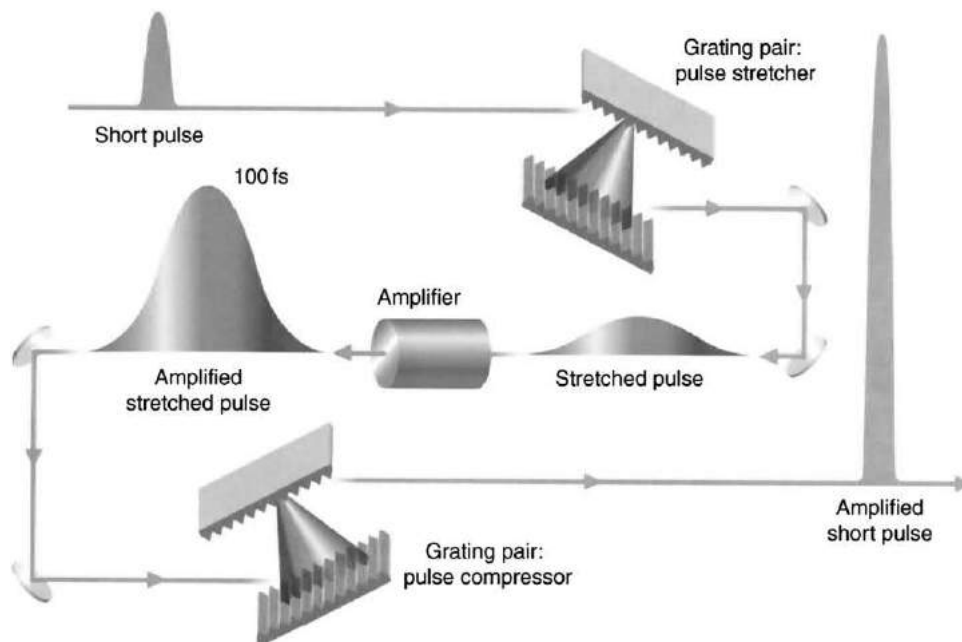


Fig. 1 The key to tabletop ultrahigh-intensity lasers is a technique called chirped pulse amplification. An initial short laser pulse is stretched out (chirped) by a factor of about 10^4 , for instance, by a pair of diffraction gratings. The stretched pulse's intensity is low, allowing it to be amplified by a small laser amplifier. A second pair of gratings recompresses the pulse, boosting it to 10^4 times the peak intensity that the amplifier could have withstood.

Carbon Dioxide Laser

CR Chatwin, University of Sussex, Brighton, UK

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

α The direct excitation of carbon dioxide (CO₂) ground state molecules (sec⁻¹)
 β The direct de-excitation of nitrogen (N₂) (sec⁻¹)
 γ The direct excitation of nitrogen (N₂) ground state molecules (sec⁻¹)
 η The direct de-excitation of carbon dioxide (CO₂) (sec⁻¹)
 λ wavelength (m)
 σ absorption coefficient (cm²)
 τ Electrical current pulse length (μ s)
 τ_{sp} Spontaneous emission life time of the upper laser level (sec)
 $A = (\tau_{sp}^{-1} = 0.213(\text{sec}^{-1}))$ The Einstein 'A' coefficient for the laser transition
 c velocity of light (cm sec⁻¹)
 C_c Coupling capacitance (nF)
CO₂ Carbon dioxide
 e subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium
 E Electric Field (V cm⁻¹)
 F_{CO_2} Fraction of the input power (IP) coupled into the excitation of the energy level $E_{00^0_1}$
 F_{N_2} Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$
 g gain (cm⁻¹)
 g_1 and g_2 energy level degeneracy's of levels 1 and 2
He Helium
 I beam irradiance (W cm⁻²)
 I_0 beam irradiance at the center of a Gaussian laser beam (W cm⁻²)

I_p Photon population density (photons cm⁻³)
 I_p Input current (A)
IP Electrical input power (W cm⁻³)
 J the rotational quantum number
 K_{51}, K_{15} Resonant energy transfer between the CO₂ (00⁰₁) and N₂ ($v=2$) energy levels proceeds via excited molecules colliding with ground state molecules (sec⁻¹)
 $K_{132}, K_{2131}, K_{2231}$ are vibration/vibration transfer rates (sec⁻¹) between energy levels 1 and 32; 21 and 31; 22 and 31, respectively (see Fig. 1)
 K_{320} is a vibration/translation transfer rate (sec⁻¹) between energy levels 32 and 0 (see Fig. 1)
 K_{sp}, A Spontaneous emission rate (sec⁻¹)
 L The distance between the back and the front mirrors, which have reflectivity's of R_B and R_F (cm)
 n molecular population (molecules cm⁻³)
N₂ Nitrogen
 P Pressure (Torr)
 P_{CO_2}, P_{He} and P_{N_2} The respective gas partial pressures (Torr)
 P_{in} Electrical input power (kW)
 r radius of laser beam (cm)
 R_B Back mirror reflectivity
 R_F Front mirror reflectivity
 t time (sec)
 T Temperature (deg K)
 T_0 the photon decay time (sec)
 w the radial distance where the power density is decreased to $1/e^2$ of its axial value

Introduction

This article gives a brief history of the development of the laser and goes on to describe the characteristics of the carbon dioxide laser and the molecular dynamics that permit it to operate at comparatively high power and efficiency. It is these commercially attractive features and its low cost that has led to its adoption as one of most popular industrial power beams. This outline also describes the main types of carbon dioxide laser and briefly discusses their characteristics and uses.

Brief History

In 1917 Albert Einstein developed the concept of stimulated emission which is the phenomenon used in lasers. In 1954 the MASER (Microwave Amplification by Stimulated Emission of Radiation) was the first device to use stimulated emission. In that year Townes and Schawlow suggested that stimulated emission could be used in the infrared and optical portions of the electromagnetic spectrum. The device was originally termed the optical maser, this term being dropped in favor of LASER, standing for: Light Amplification by Stimulated Emission of Radiation. Working against the wishes of his manager at Hughes Research Laboratories, the electrical engineer Ted Maiman created the first laser on the 15 May 1960. Maiman's flash lamp pumped ruby laser produced pulsed red electromagnetic radiation at a wavelength of 694.3 nm. During the most active period of laser systems discovery Bell Labs made a very significant contribution. In 1960, Ali Javan, William Bennet and Donald Herriot produced the first Helium Neon laser, which was the first continuous wave (CW) laser operating at 1.15 μ m. In 1961, Boyle and Nelson

developed a continuously operating Ruby laser and in 1962, Kumar Patel, Faust, McFarlane and Bennet discovered five noble gas lasers and lasers using oxygen mixtures. In 1964, C.K.N. Patel created the high-power carbon dioxide laser operating at 10.6 μm . In 1964, J.F. Geusic and R.G. Smith produced the first Nd:Yag laser using neodymium doped yttrium aluminum garnet crystals and operating at 1.06 μm .

Characteristics

Due to its operation between low lying vibrational energy states of the CO_2 molecule, the CO_2 laser has a high quantum efficiency, $\sim 40\%$, which makes it extremely attractive as a high-power industrial materials processing laser (1 to 20 kW), where energy and running costs are a major consideration. Due to the requirement for cooling to retain the population inversion, the efficiency of electrical pumping and optical losses – commercial systems have an overall efficiency of approximately 10%. Whilst this may seem low, for lasers this is still a high efficiency. The CO_2 laser is widely used in other fields, for example, surgical applications, remote sensing, and measurement. It emits infrared radiation with a wavelength that can range from 9 μm up to 11 μm . The laser transition may occur on one of two transitions: $(00^01) \rightarrow (10^00)$, $\lambda = 10.6 \mu\text{m}$; $(00^01) \rightarrow (02^00)$, $\lambda = 9.6 \mu\text{m}$, see Fig. 1. The 10.6 μm transition has the maximum probability of oscillation and gives the strongest output; hence, this is the usual wavelength of operation, although for specialist applications the laser can be forced to operate on the 9.6 μm line.

Fig. 1 illustrates an energy level diagram with four vibrational energy groupings that include all the significantly populated energy levels. The internal relaxation rates within these groups are considered to be infinitely fast when compared with the rate of energy transfer between these groups. In reality the internal relaxation rates are at least an order of magnitude greater than the rates between groups.

Excitation of the upper laser level is usually provided by an electrical glow discharge. However, gas dynamic lasers have been built where expanding a hot gas through a supersonic nozzle creates the population inversion; this creates a nonequilibrium region in the downstream gas stream with a large population inversion, which produces a very high-power output beam (135 kW – Avco Everett Research Lab). For some time the gas dynamic laser was seriously considered for use in the space-based Strategic Defence Initiative (SDI-USA). The gas mixture used in a CO_2 laser is usually a mixture of carbon dioxide, nitrogen, and helium. The proportions of these gases varies from one laser system to another, however, a typical mixture is 10%- CO_2 ; 10%- N_2 ; 80%-He. Helium plays a vital role in the operation of the CO_2 laser in that it maintains the population inversion by depopulating the lower laser level by nonradiative collision processes. Helium is also important for stabilization of the gas discharge; furthermore it greatly improves the thermal conductivity of the gas mixture, which assists in the removal of waste heat via heat exchangers.

Small quantities of other gases are often added to commercial systems in order to optimize particular performance characteristics or stabilize the gas discharge; for brevity we only concern ourselves here with this simple gas mixture.

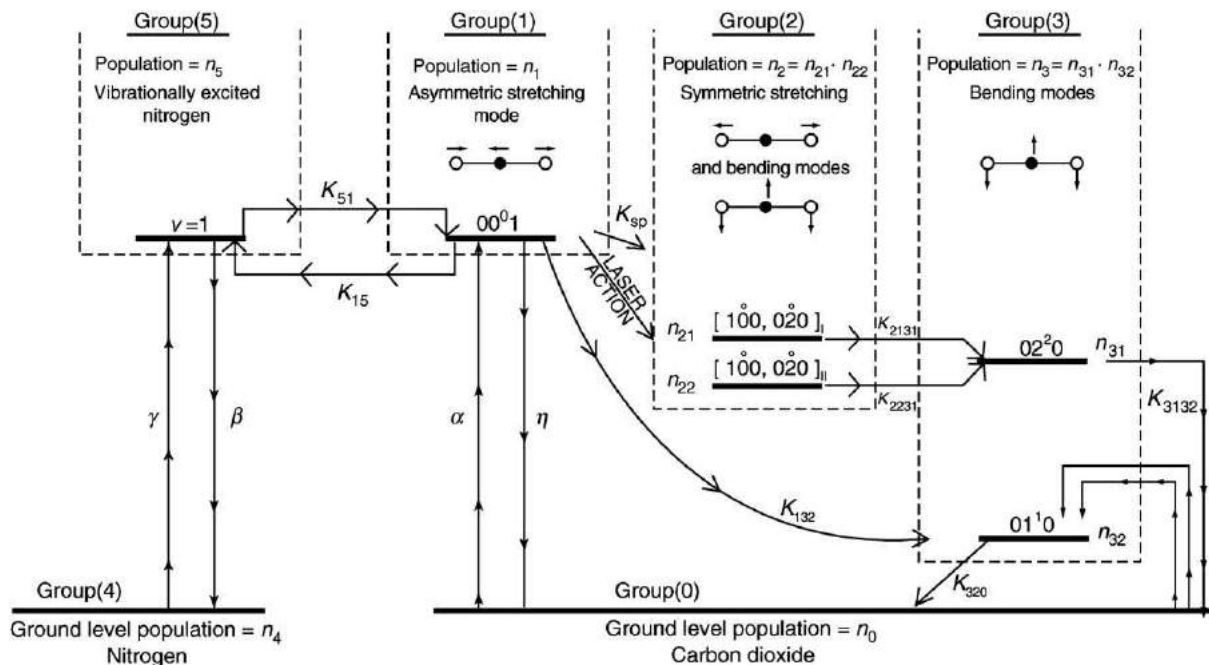


Fig. 1 Six level model used for the theoretical description of CO_2 laser action.

Molecular Dynamics

Direct Excitation and De-excitation

It is usual for excitation to be provided by an electrical glow discharge. The direct excitation of carbon dioxide (CO_2) and nitrogen (N_2) ground state molecules proceeds via inelastic collisions with fast electrons. The rates of kinetic energy transfer are α and γ , respectively, and are given by Eqs. (1) and (2):

$$\alpha = \frac{F_{\text{CO}_2} \times \text{IP}}{E_{00^0 1} \times n_0} (\text{sec}^{-1}) \quad (1)$$

$$\gamma = \frac{F_{\text{N}_2} \times \text{IP}}{E_{v=1} \times n_4} (\text{sec}^{-1}) \quad (2)$$

where:

F_{CO_2} = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{00^0 1}$, n_0 is the CO_2 ground level population density;

F_{N_2} = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$; and n_4 is the N_2 ground level population density.

The reverse process of the above occurs when molecules lose energy to the electrons and the electrons gain an equal amount of kinetic energy; the direct de-excitation rates are given by η and β , Eqs. (3) and (4), respectively:

$$\eta = \alpha \times \exp\left(\frac{E_{00^0 1}}{E_e}\right) (\text{sec}^{-1}) \quad (3)$$

$$\beta = \gamma \times \exp\left(\frac{E_{v=1}}{E_e}\right) (\text{sec}^{-1}) \quad (4)$$

where E_e is the average electron energy in the discharge.

E_e , F_{CO_2} and F_{N_2} are obtained by solution of the Boltzmann transport equation (BTE); the average electron energy can be optimized to maximize the efficiency (F_{CO_2} , F_{N_2}) with which electrical energy is utilized to create a population inversion. Hence, the discharge conditions required to maximize efficiency can be predicted from the transport equation. Fig. 2 shows one solution of the BTE for the electron energy distribution function.

Resonant Energy Transfer

Resonant energy transfer between the CO_2 ($00^0 1$) and N_2 $v=2$ energy levels (denoted 1 and 5 in Fig. 1) proceeds via excited molecules colliding with ground state molecules. A large percentage of the excitation of the upper laser level takes place via collisions between excited N_2 molecules and ground state CO_2 molecules. The generally accepted rate of this energy transfer is

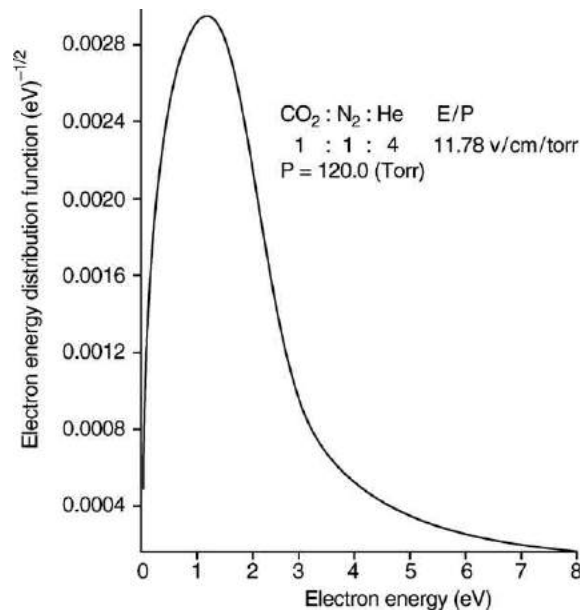


Fig. 2 Electron energy distribution function.

given by Eqs. (5) and (6):

$$K_{51} = 19\,000P_{\text{CO}_2}(\text{sec}^{-1}) \quad (5)$$

$$K_{15} = 19\,000P_{\text{N}_2}(\text{sec}^{-1}) \quad (6)$$

where P_{CO_2} and P_{N_2} are the respective gas partial pressures in Torr.

Hence, CO_2 molecules are excited into the upper laser level by both electron impact and impact with excited N_2 molecules. The contribution from N_2 molecules can be greater than 40% depending on the discharge conditions.

Collision-Induced Vibrational Relaxation of the Upper and Lower Laser Levels

The important vibrational relaxation processes are illustrated by Fig. 1 and can be evaluated from Eqs. (7–10); where the subscripts refer to the rate between energy levels 1 and 32; 21 and 31; 22 and 31; 32 and 0, respectively:

$$K_{132} = 367P_{\text{CO}_2} + 110P_{\text{N}_2} + 67P_{\text{He}}(\text{sec}^{-1}) \quad (7)$$

$$K_{2131} = 6 \times 10^5 P_{\text{CO}_2}(\text{sec}^{-1}) \quad (8)$$

$$K_{2231} = 5.15 \times 10^5 P_{\text{CO}_2}(\text{sec}^{-1}) \quad (9)$$

$$K_{320} = 200P_{\text{CO}_2} + 215P_{\text{N}_2} + 3270P_{\text{He}}(\text{sec}^{-1}) \quad (10)$$

K_{132} , K_{2131} , and K_{2231} are vibration/vibration transfer rates and K_{320} is a vibration/translation transfer rate. Note the important effect of helium on Eq. (10); helium plays a major role in depopulating the lower laser level, thus enhancing the population inversion. P_{He} is the partial pressure of helium in Torr.

Radiative Relaxation

Spontaneous radiative decay is not a major relaxation process in the CO_2 laser but it is responsible for starting laser action via spontaneous emission. The Einstein 'A' coefficient for the laser transition is given by Eq. [11]:

$$A = (\tau_{\text{sp}})^{-1} = 0.213(\text{sec}^{-1}) \quad (11)$$

Gain

The gain (g) is evaluated from the product of the absorption coefficient (σ) and the population inversion, Eq. (12):

$$g = \sigma \left(n_{00^0_1} - \frac{g_1}{g_2} n_{10^0_0} \right) \text{cm}^{-1} \quad (12)$$

For most commercial laser systems the absorption coefficient is that for high-pressure collision-broadening ($P > 5.2$ Torr) where the intensity distribution function describing the line shape is Lorentzian. The following expression describes the absorption coefficient, Eq. (13):

$$\sigma = \frac{692.5}{T n_{\text{CO}_2} \left(1 + 1.603 \frac{n_{\text{N}_2}}{n_{\text{CO}_2}} + 1.4846 \frac{n_{\text{He}}}{n_{\text{CO}_2}} \right)} (\text{cm}^2) \quad (13)$$

where T is the absolute temperature and n refers to the population density of the gas designated by the subscript.

This expression takes account of the different constituent molecular velocity distributions and different collision cross-sections for $\text{CO}_2 \rightarrow \text{CO}_2$, $\text{N}_2 \rightarrow \text{CO}_2$ and $\text{He} \rightarrow \text{CO}_2$ type collisions. Eq. (13) also takes account of the significant line broadening effect of helium.

Neglecting the unit change in rotational quantum number, the energy level degeneracies g_1 and g_2 may be dropped. $n_{10^0_0}$ is partitioned such that $n_{10^0_0} = 0.145n_2$ and Eq. (12) can be re-cast as Eq. (14):

$$g = \sigma(n_1 - 0.1452n_2) \text{cm}^{-1} \quad (14)$$

where n_1 and n_2 are the population densities of energy groups '1' and '2' respectively.

Stimulated Emission

Consider a laser oscillator with two plane mirrors, one placed at either end of the active gain medium, with one mirror partially transmitting (see Fig. 4). Laser action is initiated by spontaneous emission that happens to produce radiation whose direction is normal to the end mirrors and falls within the resonant modes of the optical resonator. The rate of change of photon population density (I_p within the laser cavity can be written as Eq. (15):

$$\frac{dI_p}{dt} = I_p c g - \frac{I_p}{T_0} \quad (15)$$

where the first term on the right-hand side accounts for the effect of stimulated emission and the second term represents the number of photons that decay out of the laser cavity, T_0 is the photon decay time, given by Eq. (16), and is defined as the average time a photon remains inside the laser cavity before being lost either through the laser output window or due to dispersion; if dispersion is ignored, I_p/T_0 is the laser output:

$$T_0 = \frac{2L}{c \log_e \left(\frac{1}{R_B R_F} \right)} \quad (16)$$

where L is the distance between the back and the front mirrors, which have reflectivities of R_B and R_F , respectively. The dominant laser emission occurs on a rotational–vibrational P branch transition $P(22)$, that is $(J=21) \rightarrow (J=22)$ line of the $(00^0 1) \rightarrow (10^0 0)$, $\lambda = 10.6 \mu\text{m}$ transition, where J is the rotational quantum number. The rotational level relaxation rate is so rapid that equilibrium is maintained between rotational levels so that they feed all their energy through the $P(22)$ transition. This model simply assumes constant intensity, basing laser performance on the performance of an average unit volume. By introducing the stimulated emission term into the molecular rate equations, which describe the rate of transfer of molecules between the various energy levels illustrated in Fig. 1, a set of molecular rate Eqs. (17–21) can be written that permit simulation of the performance of a carbon dioxide laser:

$$\frac{dn_1}{dt} = \alpha n_0 - \eta n_1 + K_{51} n_5 - K_{15} n_1 - K_{sp} n_1 - K_{132} \left(n_1 - \left[\frac{n_1}{n_{32}} \right]_e n_{32} \right) - I_p c g \quad (17)$$

$$\frac{dn_2}{dt} = K_{sp} n_1 + I_p c g - K_{2131} \left(n_{21} - \left[\frac{n_{21}}{n_{31}} \right]_e n_{31} \right) - K_{2231} \left(n_{22} - \left[\frac{n_{22}}{n_{31}} \right]_e n_{31} \right) \quad (18)$$

$$\frac{dn_3}{dt} = 2K_{2131} \left(n_{21} - \left[\frac{n_{21}}{n_{31}} \right]_e n_{31} \right) - 2K_{2231} \left(n_{22} - \left[\frac{n_{22}}{n_{31}} \right]_e n_{31} \right) + K_{132} \left(n_1 - \left[\frac{n_1}{n_{32}} \right]_e n_{32} \right) - K_{320} \left(n_{32} - \left[\frac{n_{32}}{n_0} \right]_e n_0 \right) \quad (19)$$

$$\frac{dn_5}{dt} = \gamma n_4 - \beta n_5 - K_{51} n_5 + K_{15} n_1 \quad (20)$$

$$\frac{dI_p}{dt} = I_p c g - \frac{I_p}{T_0} \quad (21)$$

The terms in square brackets ensure that the system maintains thermodynamic equilibrium; subscript ‘e’ refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium. The set of five simultaneous differential equations can be solved using a Runge–Kutta method. They can provide valuable performance prediction data that is helpful in optimizing laser design, especially when operated in the pulsed mode. Fig. 3(a) illustrates some simulation results for the transverse flow laser shown in Fig. 9. The results illustrate the effect of altering the gas mixture and how this can be used to control the gain switched spike that would result in unwelcome work piece plasma generation if allowed to become too large. Fig. 3(b) shows an experimental laser output pulse from the high pulse repetition frequency (prf=5kHz) laser illustrated in Fig. 9. This illustrates that even a quite basic physical model can give a good prediction of laser output performance.

Optical Resonator

Fig. 4 shows a simple schematic of an optical resonator. This simple optical system consists of two mirrors (the full reflector is often a water cooled gold coated copper mirror), which are aligned to be orthogonal to the optical axis that runs centrally along the length of the active gain medium in which there is a population inversion. The output coupler is a partial reflector (usually dielectrically coated zinc selenide – ZnSe—that may be edge cooled) so that some of the electromagnetic radiation can escape as an output beam. The ZnSe output coupler has a natural reflectivity of about 17% at each air–solid interface. For high power lasers (2 kW) 17% is sufficient for laser operation; however, depending on the required laser performance the inside surface is often given a reflective coating. The reflectivity of the inside face depends on the balance between the gain (Eq. (14)), the output power and the power stability requirements. The outside face of the partial reflector must be anti-reflection (AR) coated.

Spontaneous emission occurs within the active gain medium and radiates randomly in all directions; a fraction of this spontaneous emission will be in the same direction as the optical axis, perpendicular to the end mirrors, and will also fall into a resonant mode of the optical resonator. Spontaneous emission photons interact with CO_2 molecules in the excited upper laser level, excited state $(00^0 1)$, which stimulates these molecules to give up a quanta of vibrational energy as photons via the radiative transition $(00^0 1 \rightarrow 10^0 0)$, $\lambda = 10.6 \mu\text{m}$. The radiation given up will have exactly the same phase and direction as the stimulating radiation and thus will be coherent with it. The reverse process of absorption also occurs, but so long as there is a population inversion there will be a net positive output. This process is called light amplification by stimulated emission of radiation (LASER). The mirrors continue to redirect the photons parallel to the optical axis and so long as the population inversion is not depleted, more and more photons are stimulated by stimulated emission which dominates the process and also dominates the spontaneous emission, which is important to initiate laser action.

Light emitted by lasers contains several optical frequencies, which are a function of the different modes of the optical resonator; these are simply the standing wave patterns that can exist within the resonator structure. There are two types of resonator modes:

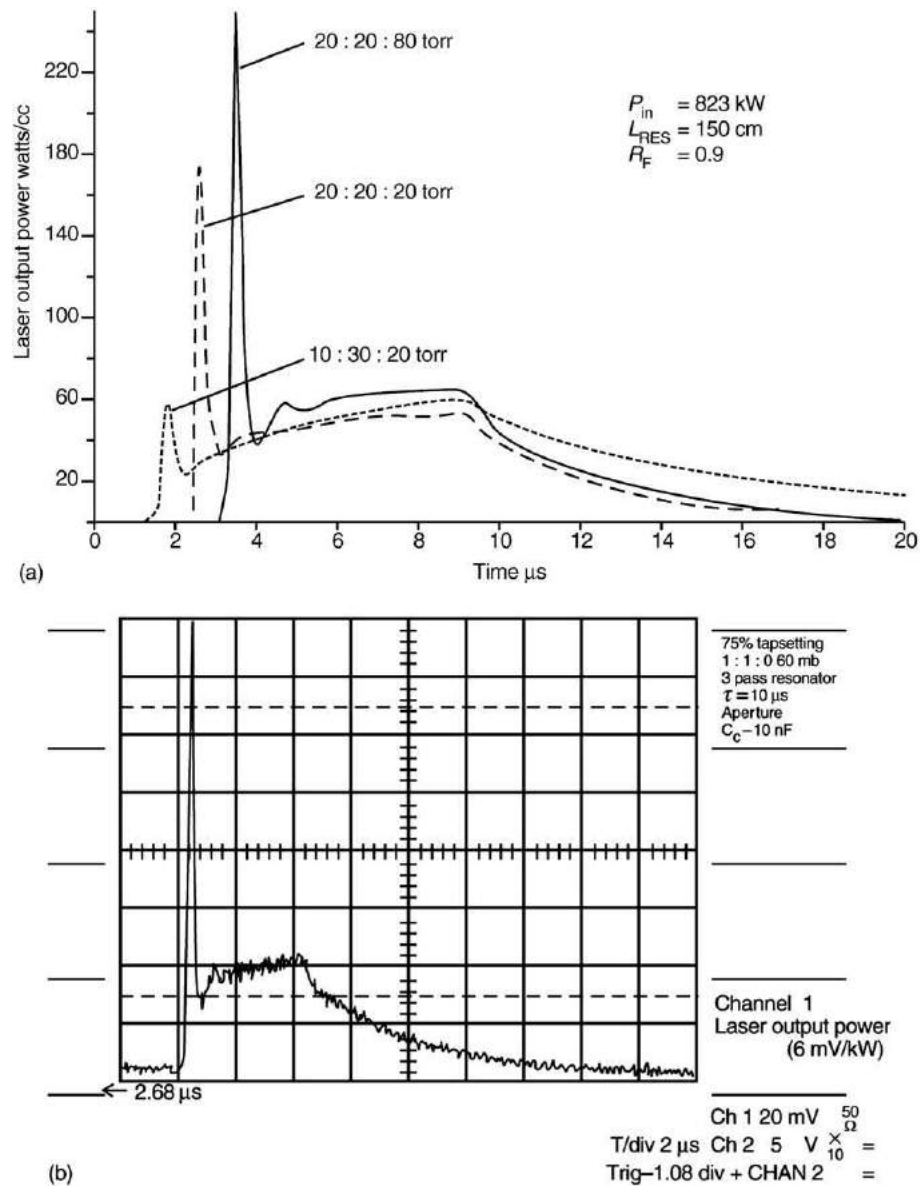


Fig. 3 (a) Predicted output pulses for transverse flow CO₂ laser for different gas mixtures, (b) Experimental output pulse from transverse flow CO₂ laser.

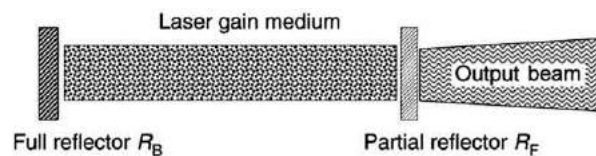


Fig. 4 Optical resonator.

longitudinal and transverse. Longitudinal modes differ from one another in their frequency of oscillation whereas transverse modes differ from one another in their oscillation frequency and field distribution in a plane orthogonal to the direction of propagation. Typically CO₂ lasers have a large number of longitudinal modes; in CO₂ laser applications these are of less interest than the transverse modes, which determine the transverse beam intensity and the nature of the beam when focused. In cylindrical coordinates the transverse modes are labelled TEM_{pl}, where subscript 'p' is the number of radial nodes and 'l' is the number of angular nodes. The lowest order mode is the TEM₀₀, which has a Gaussian-like intensity profile with its maximum on the beam axis. A light beam emitted from an optical resonator with a Gaussian profile is said to be operating in the 'fundamental

mode' or the TEM₀₀ mode. The decrease in irradiance (I) with distance ' r ' from the axis (I_0) of the Gaussian beam is described by Eq. (22):

$$I(r) = I_0 \exp(-2r^2/w^2) \quad (22)$$

where w is the radial distance, where the power density is decreased to $1/e^2$ of its axial value. Ideally a commercial laser should be capable of operation in the fundamental mode as, with few exceptions, this results in the best performance in applications. Laser cutting benefits from operation in the fundamental mode; however, welding or heat treatment applications may benefit from operation with higher-order modes. Output beams are usually controlled to be linearly or circularly polarized, depending upon the requirements of the application. For materials processing applications the laser beam is usually focused via a water cooled ZnSe lens or, for very high power lasers, a parabolic gold coated mirror. Welding applications will generally use a long focal length lens and cutting applications will use a short focal length, which generates a higher irradiance at the work piece than that necessary for welding. The beam delivery optics are usually incorporated into a nozzle assembly that can deliver cooling water and assist gases for cutting and anti-oxidizing shroud gases for welding or surface engineering applications.

Laser Configuration

CO₂ lasers are available in many different configurations and tend to be classified on the basis of their physical form and the gas flow arrangement, both of which greatly affect the output power available and beam quality. The main categories are: sealed off lasers, waveguide lasers, slow axial flow, fast axial flow, diffusion cooled, transverse flow, transversely excited atmospheric lasers, and gas dynamic lasers.

Sealed-Off and Waveguide Lasers

Depopulation of the lower laser level is via collision with the walls of the discharge tube, so the attainable output power scales with the length of the discharge column and not its diameter. Output powers are in the range 5 W to 250 W. Devices may be constructed from concentric glass tubes with the inner tube providing the discharge cavity and the outer tube acting to contain water-cooling of the inner discharge tube. The inner tube walls act as a heat sink for the discharge thermal energy (see Fig. 5). The DC electrical discharge is provided between a cathode and anode situated at either end of the discharge tube. A catalyst must be provided to ensure regeneration of CO₂ from CO. This may be accomplished by adding about 1% of H₂O to the gas mixture, or alternatively, recombination can be achieved via a hot (300 °C) Ni cathode, which acts as a catalyst. RF-excited all metal sealed-off tube systems can deliver lifetimes greater than 45 000 hours.

Diffusion-cooled slab laser technology will also deliver reliable sealed operation for 20 000 hrs. Excitation of the laser medium occurs via RF excitation between two water-cooled electrodes. The water-cooled electrodes dissipate (diffusion cooled) the heat generated in the gas discharge. An unstable optical resonator provides the output coupling for such a device (see Fig. 6). Output powers are in the range 5 W to 300 W and can be pulsed from 0 to 100 kHz. These lasers are widely used for marking, rapid prototyping, and cutting of nonmetals (paper, glass, plastics, ceramics) and metals.

Waveguide CO₂ lasers use small bore tubes (2–4 mm) made of BeO or SiO₂ where the laser radiation is guided by the tube walls. Due to the small tube diameter, a gas total pressure of 100 to 200 Torr is necessary, hence the gain per unit length is high. This type of laser will deliver 30 W of output power from a relatively short (50 cm long) compact sealed-off design; such a system is useful for microsurgery and scientific applications. Excitation can be provided from a longitudinal DC discharge or from an RF source that is transverse to the optical axis; RF excitation avoids the requirement for an anode and cathode and results in a much lower electrode voltage.

Slow Axial Flow Lasers

In slow flow lasers the gas mixture flows slowly through the laser cavity. This is done to remove the products of dissociation that will reduce laser efficiency or prevent it from operating at all, and the main contaminant is CO. The dissociated gases (mainly CO and O₂) can be recombined using a catalyst pack and then reused via continuous recirculation. Heat is removed via

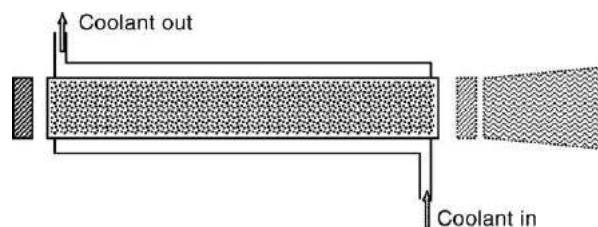


Fig. 5 Schematic of sealed-off CO₂ laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

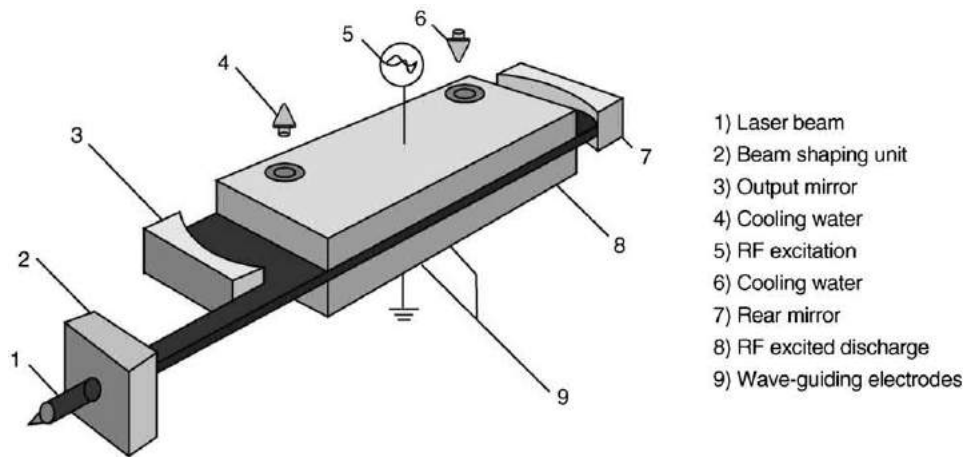


Fig. 6 Schematic for sealed-off slab laser and diffusion cooled laser with RF excitation (courtesy of Rofin).

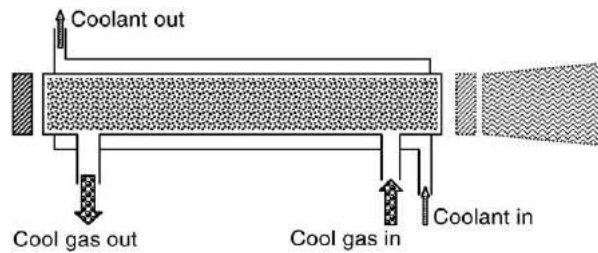


Fig. 7 Slow flow CO₂ laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

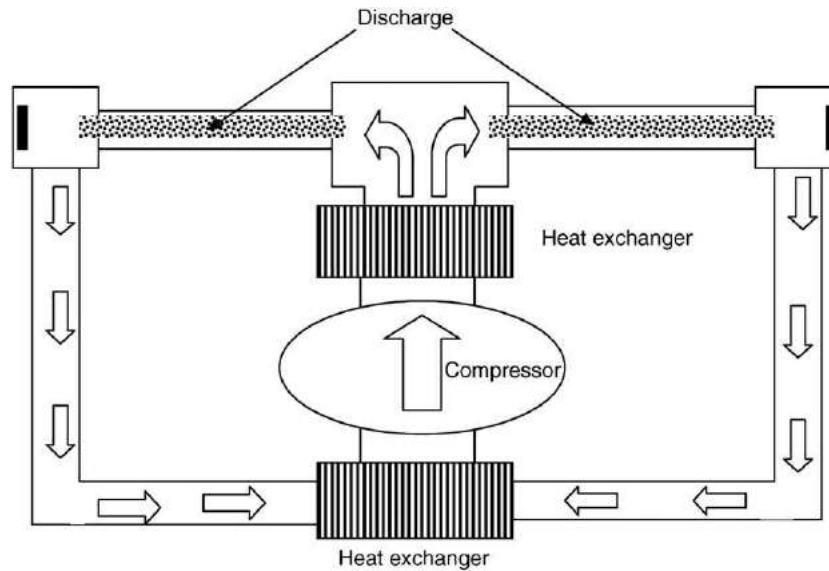


Fig. 8 Fast axial flow carbon dioxide laser.

diffusion through the walls of the tube containing the active gain medium. The tube is frequently made of Pyrex glass with a concentric outer tube to facilitate water-cooling of the laser cavity (see Fig. 7). Slow flow lasers operate in the power range 100 W to 1500 W, and tend to use a longitudinal DC electrical discharge which can be made to run continuously or pulsed if a thyatron switch is built into the power supply; alternatively, electrical power can be supplied via transverse RF excitation. The power scales with length, hence high power slow flow lasers have long cavities and require multiple cavity folds in order to reduce their physical size.

Fast Axial Flow Lasers

The fast axial flow laser, **Fig. 8**, can provide output powers from 1 kW to 20 kW; it is this configuration that dominates the use of CO₂ lasers for industrial applications. Industrial lasers are usually in the power range 2–4 kW. The output power from these devices scales with mass flow, hence the gas mixture is recycled through the laser discharge region at sonic or supersonic velocities. Historically this was achieved using Roots blowers to compress the gas upstream of the laser cavity. Roots compressors are inherently inefficient and the more advanced laser systems utilize turbine compressors, which deliver greater efficiency and better laser stability. Roots compressors can be a major source of vibration. With this arrangement heat exchangers are required to remove heat after the laser discharge region and also after the compressor stage, as the compression process heats up the laser gases. Catalyst packs are used to regenerate gases but some gas replacement is often required. These laser systems have short cavities and use folded stable resonator designs to achieve higher output powers with extremely high-quality beams that are particularly suitable for cutting applications. They also give excellent results when used for welding and surface treatments.

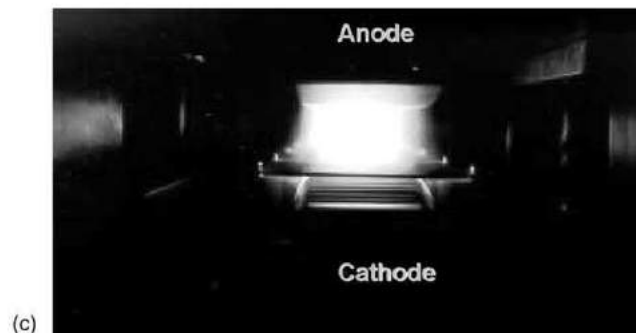
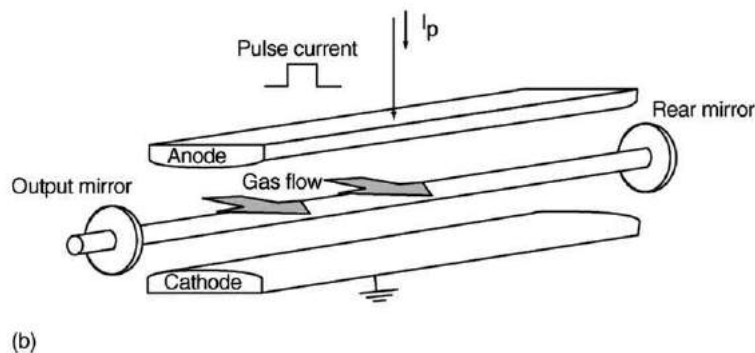
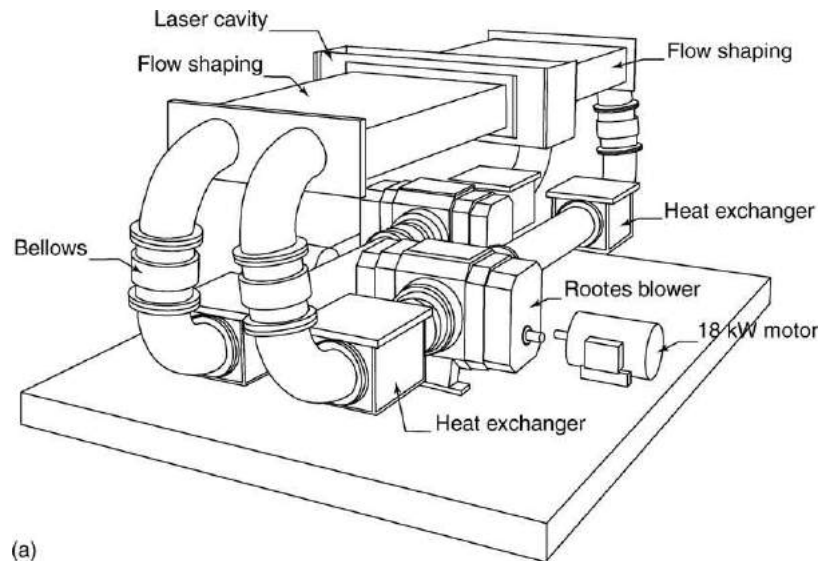


Fig. 9 (a) Transverse flow carbon dioxide laser gas recirculator, (b) Transverse flow carbon dioxide electrodes, (c) Transverse flow carbon dioxide gas discharge as seen from the output window.

Fast axial flow lasers can be excited by a longitudinal DC discharge or a transverse RF discharge. Both types of electrical excitation are common. For materials processing applications it is often important to be able to run a laser in continuous wave (CW) mode or as a high pulse repetition rate (prf) pulsed laser and to be able to switch between CW and pulsed in real time; for instance, laser cutting of accurate internal corners is difficult using CW operation but very easy using the pulsed mode of operation. Both methods of discharge excitation can provide this facility.

Diffusion Cooled Laser

The diffusion-cooled slab laser is RF excited and gives an extremely compact design capable of delivering 4.5 kW pulsed from 8 Hz to 5 kHz prf or CW with good beam quality (see Fig. 6). The optical resonator is formed by the front and rear mirrors and two parallel water cooled RF-electrodes. Diffusion cooling is provided by the RF-electrodes, removing the requirement for conventional gas recirculation via Rootes blowers or turbines. This design of laser results in a device with an extremely small footprint that has low maintenance and running costs. Applications include: cutting, welding, and surface engineering.

Fast Transverse Flow Laser

In the fast transverse flow laser (Fig. 9(a)) the gas flow, electrical discharge, and the output beam are at right angles to each other (Fig. 9(b)). The transverse discharge can be high voltage DC, RF, or pulsed up to 8 kHz (Fig. 9(c)). Very high output power per unit discharge length can be obtained with an optimal total pressure (P) of ~ 100 Torr; systems are available delivering 10 kW of output power, CW or pulsed (see Figs. 3(a) and (b)). The increase in total pressure requires a corresponding increase in the gas discharge electric field, E , as the ratio E/P must remain constant, since this determines the temperature of the discharge electrons, which have an optimum mean value (optimum energy distribution, Fig. 2) for efficient excitation of the population inversion. With this high value of electric field, a longitudinal-discharge arrangement is impractical (500 kV for a 1 m discharge length); hence, the discharge is applied perpendicular to the optical axis. Fast transverse flow gas lasers provided the first multikilowatt outputs but tend to be expensive to maintain and operate. In order to obtain a reasonable beam quality, the output coupling is often obtained using a multipass unstable resonator. As the population inversion is available over a wide rectangular cross-section, this is a disadvantage of this arrangement and beam quality is not as good as that obtainable from fast axial flow designs. For this reason this type of laser is suitable for a wide range of welding and surface treatment applications.

Transversely Excited Atmospheric (TEA) Pressure

If the gas total pressure is increased above ~ 100 Torr it is difficult to sustain a stable glow discharge, because above this pressure instabilities degenerate into arcs within the discharge volume. This problem can be overcome by pulse excitation; using sub-microsecond pulse duration, instabilities do not have sufficient time to develop; hence, the gas pressure can be increased above atmospheric pressure and the laser can be operated in a pulsed mode. In a mode locked format optical pulses shorter than 1 ns can be produced. This is called a TEA laser and with a transverse gas flow is capable of producing short high power pulses up to a few kHz repetition frequency. In order to prevent arc formation, TEA lasers usually employ ultraviolet or e-beam preionization of the gas discharge just prior to the main current pulse being applied via a thyatron switch. Output coupling is usually via an unstable resonator. TEA lasers are used for marking, remote sensing, range-finding, and scientific applications.

Conclusions

It is 40 years since Patel operated the first high power CO₂ laser. This led to the first generation of lasers which were quickly exploited for industrial laser materials processing, medical applications, defense, and scientific research applications; however, the first generation of lasers were quite unreliable and temperamental. After many design iterations, the CO₂ laser has now matured into a reliable, stable laser source available in many different geometries and power ranges. The low cost of ownership of the latest generation of CO₂ laser makes them a very attractive commercial proposition for many industrial and scientific applications. Commercial lasers incorporate many novel design features that are beyond the scope of this article and are often peculiar to the particular laser manufacturer. This includes gas additives and catalysts that may be required to stabilize the gas discharge of a particular laser design; it is this optimization of the laser design that has produced such reliable and controllable low-cost performance from the CO₂ laser.

Further Reading

- Anderson, J.D., 1976. *Gasdynamic Lasers: An Introduction*. New York: Academic Press.
- Chatwin, C.R., McDonald, D.W., Scott, B.F., 1991. Design of a High P.R.F. Carbon Dioxide Laser for Processing High Damage Threshold Materials. In: *Selected Papers on Laser Design*, SPIE Milestone Series., Washington: SPIE Optical Engineering Press.
- Cool, A.C., 1969. Power and gain characteristics of high speed flow lasers. *Journal of Applied Physics* 40 (9), 3563.
- Crafer, R.C., Gibson, A.F., Kent, M.J., Kimmit, M.F., 1969. Time-dependent processes on CO₂ laser amplifiers. *British Journal of Applied Physics* 2 (2), 183.

- Gerry, E.T., Leonard, A.D., 1966. Measurement of 10.6- μ CO₂ laser transition probability and optical broadening cross sections. *Applied Physics Letters* 8 (9), 227.
- Gondhalekar, A., Heckenberg, N.R., Holzhauer, E., 1975. The mechanism of single-frequency operation of the hybrid CO₂ laser. *IEEE Journal of Quantum Electronics* QE-11 (3), 103.
- Gordiets, B.F., Sobolev, N.N., Shelepin, L.A., 1968. Kinetics of physical processes in CO₂ lasers. *Soviet Physics JETP* 26 (5), 1039.
- Herzberg, G., 1945. *Molecular Spectra & Molecular Structure, Infra-red and Raman Spectra of Polyatomic Molecules*, vol. 2. New York: Van Nostrand.
- Hoffman, A.L., Vlases, G.C., 1972. A simplified model for predicting gain, saturation and pulse length for gas dynamic lasers. *IEEE Journal of Quantum Electronics* 8 (2), 46.
- Johnson, D.C., 1971. Excitation of an atmospheric pressure CO₂-N₂-He laser by capacitor discharges. *IEEE Journal of Quantum Electronics* Q.E.-7 (5), 185.
- Koechner, W., 1988. *Solid State Laser Engineering*. Berlin: Springer Verlag.
- Kogelnik, H., Li, T., 1966. Laser beams and resonators. *Applied Optics* 5, 1550–1567.
- Levine, A.K., De Maria, A.J., 1971. *Lasers*, vol. 3. New York: Marcel Dekkar, Chapter 3.
- Moeller, G., Ridgen, J.D., 1965. *Applied Physics Letters* 7, 274.
- Moore, C.B., Wood, R.E., Bei-Lok, H., Yardley, J.T., 1967. Vibrational energy transfer in CO₂ lasers. *Journal of Chemical Physics* 46, 4222.
- Patel, C.K.N., 1964. *Physics Review Letters* 12, 588.
- Siegman, A., 1986. *Lasers*. Mill Valley, California: University Science.
- Smith, K., Thomson, R.M., 1978. *Computer Modeling of Gas Lasers*. New York and London: Plenum Press.
- Sobolev, N.N., Sokolov, V.V., 1967. CO₂ lasers. *Soviet Physics USPEKHI* 10 (2), 153.
- Svelto, O., 1998. *Principles of Lasers*, 4th edn New York: Plenum.
- Tychinskii, V.P., 1967. Powerful lasers. *Soviet Physics USPEKHI* 10 (2), 131.
- Vlases, G.C., Money, W.M., 1972. Numerical modelling of pulsed electric CO₂ lasers. *Journal of Applied Physics* 43 (4), 1840.
- Wagner, W.G., Haus, H.A., Gustafson, K.T., 1968. High rate optical amplification. *IEEE Journal of Quantum Electronics* Q.E.-4, 287.
- Witteaman, W., 1987. *The CO₂ Laser*, Springer Series in Optical Sciences, vol. 53. .

Dye Lasers

FJ Duarte, Eastman Kodak Company, New York, NY, USA

A Costela, Consejo Superior de Investigaciones Cientificas, Madrid, Spain

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Background

Dye lasers are the original tunable lasers. Discovered in the mid-1960s these tunable sources of coherent radiation span the electromagnetic spectrum from the near-ultraviolet to the near-infrared (**Fig. 1**). Dye lasers spearheaded and sustained the revolution in atomic and molecular spectroscopy and have found use in many and diverse fields from medical to military applications. In addition to their extraordinary spectral versatility, dye lasers have been shown to oscillate from the femtosecond pulse domain to the continuous wave (cw) regime. For microsecond pulse emission, energies of up to hundreds of joules per pulse have been demonstrated. Further, operation at high pulsed repetition frequencies (prfs), in the multi-kHz regime, has provided average powers at kW levels. This unrivaled operational versatility is summarized in **Table 1**.

Dye lasers are excited by coherent optical energy from an excitation, or pump, laser or by optical energy from specially designed lamps called flashlamps. Recent advances in semiconductor laser technology have made it possible to construct very compact all-solid-state excitation sources that, coupled with new solid-state dye laser materials, should bring the opportunity to build compact tunable laser systems for the visible spectrum. Further, direct diode-laser pumping of solid-state dye lasers should prove even more advantageous to enable the development of fairly inexpensive tunable narrow-linewidth solid-state dye laser systems for spectroscopy and other applications requiring low powers. Work on electrically excited organic gain media might also provide new avenues for further progress.

The literature of dye lasers is very rich and many review articles have been written describing and discussing traditional dye lasers utilizing liquid gain media. In particular, the books *Dye Lasers*, *Dye Laser Principles*, *High Power Dye Lasers*, and *Tunable Lasers Handbook* provide excellent sources of authoritative and detailed description of the physics and technology involved. In this article we offer only a survey of the operational capabilities of the dye lasers using liquid gain media in order to examine with more attention the field of solid-state dye lasers.

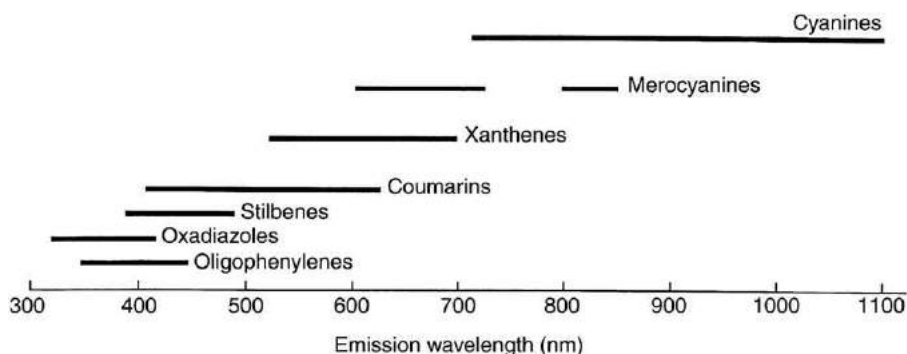


Fig. 1 Approximate wavelength span from the various classes of laser dye molecules. Reproduced with permission from Duarte FJ (1995)

Table 1 Emission characteristics of liquid dye lasers

Dye laser class	Spectral coverage ^a	Energy per pulse ^b	Pr ^b	Power ^b
Laser-pumped pulsed dye lasers	350–1100 nm	800 J ^c	13 kHz ^d	2.5 kW ^d
Flashlamp-pumped dye lasers	320–900 nm	400 J ^e	850 Hz ^f	1.2 kW ^f
CW dye lasers	370–1000 nm			43 W ^g

^aApproximate range.

^bRefers to maximum values for that particular emission parameter.

^cAchieved with an excimer-laser pumped coumarin dye laser by Tang and colleagues, in 1987.

^dAchieved with a multistage copper-vapor-laser pumped dye laser using rhodamine dye by Bass and colleagues, in 1992.

^eReported by Baltakov and colleagues, in 1974. Uses rhodamine 6G dye.

^fReported by Morton and Dragoo, in 1981. Uses coumarin 504 dye.

^gAchieved with an Ar⁺ laser pumped folded cavity dye laser using rhodamine 6G dye by Baving and colleagues, in 1982.

Brief History of Dye Lasers

1965: Quantum theory of dyes is discussed in the context of the maser (R. P. Feynman).
 1966: Dye lasers are discovered (P. P. Sorokin and J. R. Lankard; F. P. Schäfer and colleagues).
 1967: The flashlamp-pumped dye laser is discovered (P. P. Sorokin and J. R. Lankard; W. Schmidt and F. P. Schäfer).
 1967–1968: Solid-state dye lasers are discovered (B. H. Soffer and B. B. McFarland; O. G. Peterson and B. B. Snively).
 1968: Mode-locking, using saturable absorbers, is demonstrated in dye lasers (W. Schmidt and F. P. Schäfer).
 1970: The continuous-wave (cw) dye laser is discovered (O. G. Peterson, S. A. Tuccio, and B. B. Snively).
 1971: The distributed feedback dye laser is discovered (H. Kogelnik and C. V. Shank).
 1971–1975: Prismatic beam expansion in dye lasers is introduced (S. A. Myers; E. D. Stokes and colleagues; D. C. Hanna and colleagues).
 1972: Passive mode-locking is demonstrated in cw dye lasers (E. P. Ippen, C. V. Shank, and A. Dienes).
 1972: The first pulsed narrow-linewidth tunable dye laser is introduced (T. W. Hänsch).
 1973: Frequency stabilization of cw dye lasers is demonstrated (R. L. Barger, M. S. Sorem, and J. L. Hall).
 1976: Colliding-pulse-mode locking is introduced (I. S. Ruddock and D. J. Bradley).
 1977–1978: Grazing-incidence grating cavities are introduced (I. Shoshan and colleagues; M. G. Littman and H. J. Metcalf; S. Saikan).
 1978–1980: Multiple-prism grating cavities are introduced (T. Kasuya and colleagues; G. Klauminzer; F. J. Duarte and J. A. Piper).
 1981: Prism pre-expanded grazing-incidence grating oscillators are introduced (F. J. Duarte and J. A. Piper).
 1982: Generalized multiple-prism dispersion theory is introduced (F. J. Duarte and J. A. Piper).
 1983: Prismatic negative dispersion for pulse compression is introduced (W. Dietel, J. J. Fontaine, and J.-C. Diels).
 1987: Laser pulses as short as six femtoseconds are demonstrated (R. L. Fork, C. H. Brito Cruz, P. C. Becker, and C. V. Shank).
 1994: First narrow-linewidth solid-state dye laser oscillator (F. J. Duarte).
 1999–2000: Distributed feedback solid-state dye lasers are introduced (Wadsworth and colleagues; Zhu and colleagues).

Molecular Energy Levels

Dye molecules have large molecular weights and contain extended systems of conjugated double bonds. These molecules can be dissolved in an adequate organic solvent (such as ethanol, methanol, ethanol/water, and methanol/water) or incorporated into a solid matrix (organic, inorganic, or hybrid). These molecular gain media have a strong absorption generally in the visible and ultraviolet regions, and exhibit large fluorescence bandwidths covering the entire visible spectrum. The general energy level diagram of an organic dye is shown in Fig. 2. It consists of electronic singlet and triplet states with each electronic state containing a multitude of overlapping vibrational–rotational levels giving rise to broad continuous energy bands. Absorption of visible or ultraviolet pump light excites the molecules from the ground state S_0 into some rotational–vibrational level belonging to an upper excited singlet state, from where the molecules decay nonradiatively to the lowest vibrational level of the first excited singlet state S_1 on a picosecond time-scale. From S_1 the molecules can decay radiatively, with a radiative lifetime on the nanosecond time-scale, to a higher-lying vibrational–rotational level of S_0 . From this level they rapidly thermalize into the lowest vibrational–rotational levels of S_0 . Alternatively, from S_1 , the molecules can experience nonradiative relaxation either to the triplet state T_1 by an intersystem crossing process or to the ground state by an internal conversion process. If the intensity of the pumping radiation is high enough a population inversion between S_1 and S_0 may be attained and stimulated emission occurs. Internal conversion and intersystem crossing compete with the fluorescence decay mode of the molecule and therefore reduce the efficiency of the laser emission. The rate for internal conversion to the electronic ground state is usually negligibly small so that the most important loss process is intersystem crossing into T_1 that populates the lower metastable triplet state. Thus, absorption on the triplet–triplet allowed transitions could cause considerable losses if these absorption bands overlap the lasing band, inhibiting or even halting the lasing process. This triplet loss can be reduced by adding small quantities of appropriate chemicals that favor nonradiative transitions that shorten the effective lifetime of the T_1 level. For pulsed excitation with nanosecond pulses, the triplet–triplet absorption can be neglected because for a typical dye the intersystem crossing rate is not fast enough to build up an appreciable triplet population in the nanosecond time domain.

Dye molecules are large (a typical molecule incorporates 50 or more atoms) and are grouped into families with similar chemical structures. A survey of the major classes of laser dyes is given later. Solid-state laser dye gain media are also considered later.

Liquid Dye Lasers

Laser-Pumped Pulsed Dye Lasers

Laser-pumped dye lasers use a shorter wavelength, or higher frequency, pulsed laser as the excitation or pump source. Typical pump lasers for dye lasers are gas lasers such as the excimer, nitrogen, or copper lasers. One of the most widely used solid-state laser pumps is the frequency doubled Nd:YAG laser which emits at 532 nm.

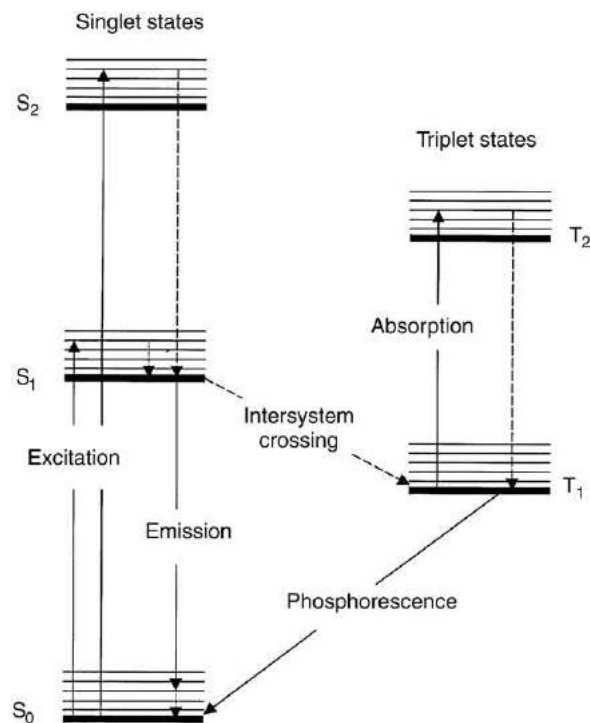


Fig. 2 Schematic energy level diagram for a dye molecule. Full lines: radiative transitions; dashed lines: nonradiative transitions; dotted lines: vibrational relaxation.

In a laser-pumped pulsed dye laser the active medium, or dye solution, is contained in an optical cell often made of quartz or fused silica, which provides an active region typically some 10 mm in length and a few mm in width. The active medium is then excited either longitudinally, or transversely, via a focusing lens using the pump laser. In the case of transverse excitation the pump laser is focused to a beam ~ 10 mm in width and ~ 0.1 mm in height. Longitudinal pumping requires focusing of the excitation beam to a diameter in the 0.1–0.15 mm range. For lasers operated at low prfs (a few pulses per second), the dye solution might be static. However, for high-prf operation (a few thousand pulses per second) the dye solution must be flowed at speeds of up to a few meters per second in order to dissipate the heat. A simple broadband optically pumped dye laser can be constructed using just the pump laser, the active medium, and two mirrors to form a resonator. In order to achieve tunable, narrow-linewidth, emission, a more sophisticated resonator must be employed. This is called a dispersive tunable oscillator and is depicted in Fig. 3. In a dispersive tunable oscillator the exit side of the cavity is comprised of a partial reflector, or an output coupler, and the other end of the resonator is composed of a multiple-prism grating assembly. It is the dispersive characteristics of this multiple-prism grating assembly and the dimensions of the emission beam produced at the gain medium that determine the tunability and the narrowness, or spectral purity, of the laser emission.

In order to selectively excite a single vibrational–rotational level of a molecule such as iodine (I_2), at room temperature, one needs a laser linewidth of $\Delta\nu \approx 1.5$ GHz (or $\Delta\lambda \approx 0.0017$ nm at $\lambda = 590$ nm). The hybrid multiple-prism near-grazing-incidence (HMPGI) grating dye laser oscillator illustrated in Fig. 4 yields laser linewidths in the $400 \text{ MHz} \leq \Delta\nu \leq 650 \text{ MHz}$ range at 4–5% conversion efficiencies whilst excited by a copper-vapor laser operating at a prf of 10 kHz. Pulse lengths are ~ 10 ns at full-width half-maximum (FWHM). The narrow-linewidth emission from these oscillators is said to be single-longitudinal-mode lasing because only one electromagnetic mode is allowed to oscillate.

The emission from oscillators of this class can be amplified many times by propagating the tunable narrow-linewidth laser beam through single-pass amplifier dye cells under the excitation of pump lasers. Such amplified laser emission can reach enormous average powers. Indeed, a copper-vapor-laser pumped dye laser system at the Lawrence Livermore National Laboratory (USA), designed for the laser isotope separation program, was reported to yield average powers in excess of 2.5 kW, at a prf of 13.2 kHz, at a better than 50% conversion efficiency as reported by Bass and colleagues in 1992. Besides high conversion efficiencies, excitation with copper-vapor lasers, at $\lambda = 510.554$ nm, has the advantage of inducing little photodegradation in the active medium thus allowing very long dye lifetimes.

Flashlamp-Pumped Dye Lasers

Flashlamps utilized in dye laser excitation emit at black-body temperatures in the 20 000 K range, thus yielding intense ultraviolet radiation centered around 200 nm. One further requirement for flashlamps, and their excitation circuits, is to deliver light pulses with a fast rise time. For some flashlamps this rise time can be less than a few nanoseconds.

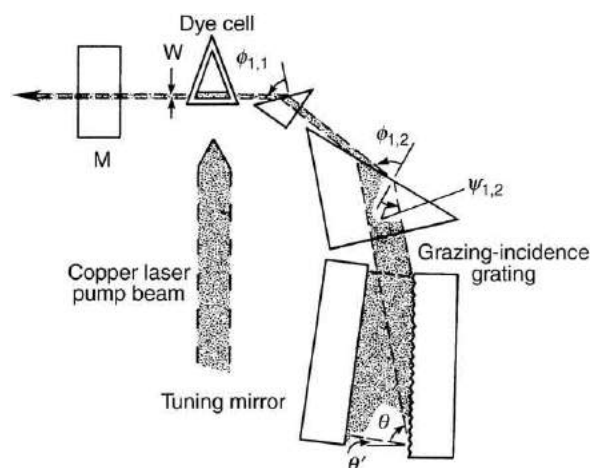


Fig. 3 Copper-vapor-laser pumped hybrid multiple-prism near grazing incidence (HMPGI) grating dye laser oscillator. Adapted with permission from Duarte FJ and Piper JA (1984) Narrow linewidth high prf copper laser-pumped dye-laser oscillators. *Applied Optics* 23: 1391–1394.

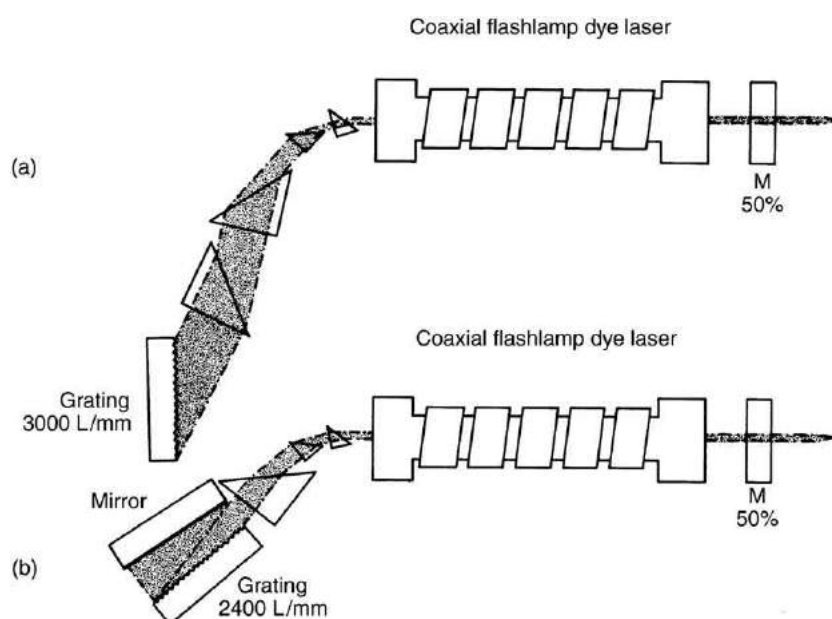


Fig. 4 Flashlamp-pumped multiple-prism grating oscillators. From Duarte FJ, Davenport WE, Ehrlich JJ and Taylor TS (1991) Ruggedized narrow-linewidth dispersive dye laser oscillator. *Optics Communications* 84: 310–316. Reproduced with permission from Elsevier.

Flashlamp-pumped dye lasers differ from laser-pumped pulsed dye lasers mainly in the pulse energies and pulse lengths attainable. This means that flashlamp-pumped dye lasers, using relatively large volumes of dye, can yield very large energy pulses. Excitation geometries use either coaxial lamps, with the dye flowing in a quartz cylinder at the center of the lamp, or two or more linear lamps arranged symmetrically around the quartz tube containing the dye solution. Using a relatively weak dye solution of rhodamine-6G (2.2×10^{-5} M), a coaxial lamp, and an active region defined by a quartz tube 6 cm in diameter and a length of 60 cm, Baltakov and colleagues, in 1974, reported energies of 400 J in pulses 25 μ s long at FWHM.

Flashlamp-pumped tunable narrow-linewidth dye laser oscillators described by Duarte and colleagues, in 1991, employ a cylindrical active region 6 mm in diameter and 17 cm in length. The dye solution is made of rhodamine 590 at a concentration of 1×10^{-5} M. This active region is excited by a coaxial flashlamp. Using multiple-prism grating architectures (see Fig. 4) these authors achieve a diffraction limited TEM₀₀ laser beam and laser linewidths of $\Delta\nu \approx 300$ MHz at pulsed energies in the 2–3 mJ range. The laser pulse duration is reported to be $\Delta t \approx 100$ ns. The laser emission from this class of multiple-prism grating oscillator is reported to be extremely stable. The tunable narrow-linewidth emission from these dispersive oscillators is either used directly in spectroscopic, or other scientific applications, or is utilized to inject large flashlamp-pumped dye laser amplifiers to obtain multi-joule pulse energies with the laser linewidth characteristics of the oscillator.

Continuous Wave Dye Lasers

CW dye lasers use dye flowing at linear speeds of up to 10 meters per second which are necessary to remove the excess heat and to quench the triplet states. In the original cavity reported by Peterson and colleagues, in 1970, a beam from an Ar^+ laser was focused on to an active region which is contained within the resonator. The resonator comprised dichroic mirrors that transmit the blue-green radiation of the pump laser and reflect the red emission from the dye molecules. Using a pump power of about 1 W, in a TEM_{00} laser beam, these authors reported a dye laser output of 30 mW. Subsequent designs replaced the dye cell with a dye jet, an introduced external mirror, and integrated dispersive elements in the cavity. Dispersive elements such as prisms and gratings are used to tune the wavelength output of the laser. Frequency-selective elements, such as etalons and other types of interferometers, are used to induce frequency narrowing of the tunable emission. Two typical cw dye laser cavity designs are described by Hollberg, in 1990, and are reproduced here in Fig. 5. The first design is a linear three-mirror folded cavity. The second one is an eight-shaped ring dye laser cavity comprised of mirrors M_1 , M_2 , M_3 , and M_4 . Linear cavities exhibit the effect of *spatial hole burning* which allows the cavity to lase in more than one longitudinal mode. This problem can be overcome in ring cavities (Fig. 5(b)) where the laser emission is in the form of a traveling wave.

Two aspects of cw dye lasers are worth emphasizing. One is the availability of relatively high powers in single longitudinal mode emission and the other is the demonstration of very stable laser oscillation. First, Johnston and colleagues, in 1982, reported 5.6 W of stabilized laser output in a single longitudinal mode at 593 nm at a conversion efficiency of 23%. In this work eleven dyes were used to span the spectrum continuously from ~ 400 nm to ~ 900 nm. In the area of laser stabilization and ultra narrow-linewidth oscillation it is worth mentioning the work of Hough and colleagues, in 1984, that achieved laser linewidths of less than 750 Hz employing an external reference cavity.

Ultrashort-Pulse Dye Lasers

Ultrashort-pulse, or femtosecond, dye lasers use the same type of technology as cw dye lasers configured to incorporate a saturable absorber region. One such configuration is the ring cavity depicted in Fig. 6. In this cavity the gain region is established between mirrors M_1 and M_2 whilst the saturable absorber is deployed in a counter-propagating arrangement. This arrangement is necessary to establish a collision between two counter-propagating pulses at the saturable absorber thus yielding what is known as colliding-pulse mode locking (CPM) as reported by Ruddock and Bradley, in 1976. This has the effect of creating a transient grating, due to interference, at the absorber thus shortening the pulse. Intracavity prisms were incorporated in order to introduce negative dispersion, by Dietel and colleagues in 1983, and thus subtract dispersion from the cavity and ultimately provide the compensation needed to produce femtosecond pulses.

The shortest pulse obtained from a dye laser, 6 fs, has been reported by Fork and colleagues, in 1987, using extra-cavity compression. In that experiment, a dye laser incorporating CPM and prismatic compensation was used to generate pulses that were amplified by a copper-vapor laser at a prf of 8 kHz. The amplified pulses, of a duration of 50 fs, were then propagated through two grating pairs and a four-prism sequence for further compression.

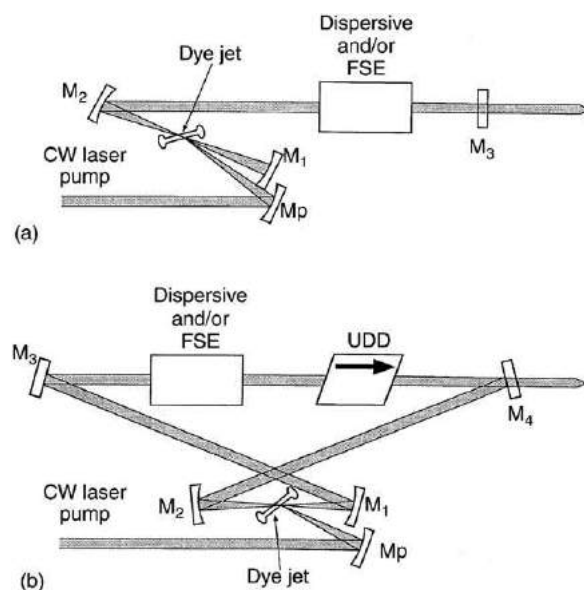


Fig. 5 CW laser cavities: (a) linear cavity and (b) ring cavity. Adapted from Hollberg LW (1990) CW dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 185–238. New York: Academic Press. Reproduced with permission from Elsevier.

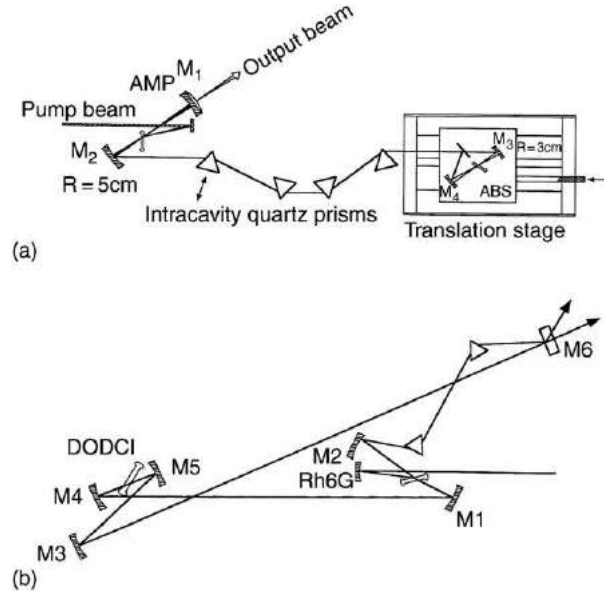


Fig. 6 Femtosecond dye laser cavities: (a) linear femtosecond cavity and (b) ring femtosecond cavity. Adapted from Diels J-C (1990) Femtosecond dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 41–132. New York: Academic Press. Reproduced with permission from Elsevier.

Solid-State Dye Laser Oscillators

In this section the principles of linewidth narrowing in dispersive resonators are outlined. Albeit the discussion focuses on multiple-prism grating solid-state dye laser oscillators, in particular, the physics is applicable to pulsed high-power dispersive tunable lasers in general.

Multiple-Prism Dispersion Grating Theory

The spectral linewidth in a dispersive optical system is given by

$$\Delta\lambda \approx \Delta\theta (\nabla_\lambda \theta)^{-1} \quad (1)$$

where $\Delta\theta$ is the light beam divergence, $\nabla_\lambda = \partial/\partial\lambda$, and $\nabla_\lambda \theta$ is the overall dispersion of the optics. This identity can be derived either from the principles of geometrical optics or from the principles of generalized interferometry as described by Duarte, in 1992.

The cumulative single-pass generalized multiple-prism dispersion at the m th prism, of a multiple-prism array as illustrated in Fig. 7, as given by Duarte and Piper, in 1982,

$$\nabla_\lambda \phi_{2,m} = H_{2,m} \nabla_\lambda n_m + (k_{1,m} k_{2,m})^{-1} \times \left(H_{1,m} \nabla_\lambda n_m \pm \nabla_\lambda \phi_{2,(m-1)} \right) \quad (2)$$

In this equation

$$k_{1,m} = \cos \psi_{1,m} / \cos \phi_{1,m} \quad (3a)$$

$$k_{2,m} = \cos \phi_{2,m} / \cos \psi_{2,m} \quad (3b)$$

$$H_{1,m} = \tan \phi_{1,m} / n_m \quad (4a)$$

$$H_{2,m} = \tan \phi_{2,m} / n_m \quad (4b)$$

Here, $k_{1,m}$ and $k_{2,m}$ represent the physical beam expansion experienced by the incident and the exit beams, respectively.

Eq. (2) indicates that $\nabla_\lambda \phi_{2,m}$, the cumulative dispersion at the m th prism, is a function of the geometry of the m th prism, the position of the light beam relative to this prism, the refractive index of the prism, and the cumulative dispersion up to the previous prism $\nabla_\lambda \phi_{2,(m-1)}$.

For an array of r identical isosceles, or equilateral, prisms arranged symmetrically, in an additive configuration, so that the angles of incidence and emergence are the same, the cumulative dispersion reduces to

$$\nabla_\lambda \phi_{2,r} = r \nabla_\lambda \phi_{2,1} \quad (5)$$

Under these circumstances the dispersions add in a simple and straightforward manner. For configurations incorporating right-angle prisms, the dispersions need to be handled mathematically in a more subtle form.

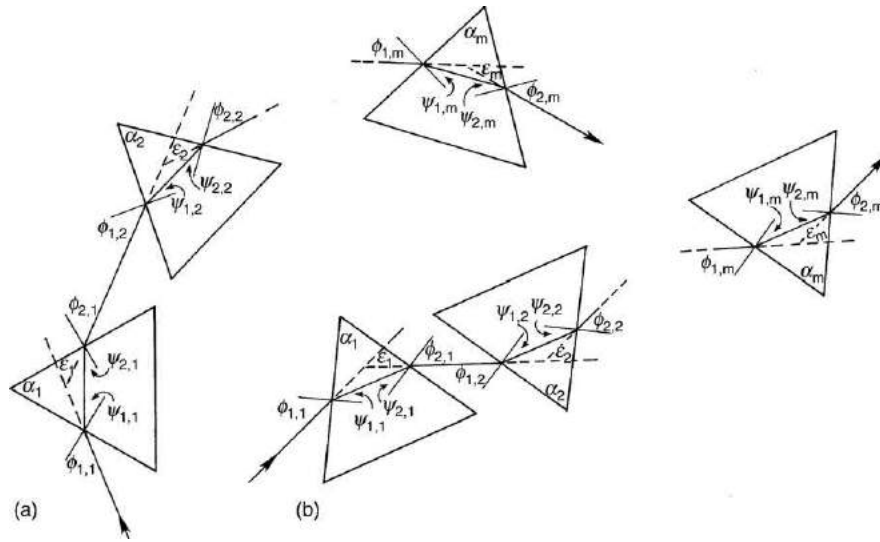


Fig. 7 Generalized multiple-prism arrays in (a) additive and (b) compensating configurations. From Duarte FJ (1990) *Narrow-linewidth pulsed dye laser oscillators*. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 133–183. New York: Academic Press. Reproduced with permission from Elsevier.

The generalized double-pass, or return-pass, dispersion for multiple-prism beam expanders was introduced by Duarte, in 1985:

$$\nabla_{\lambda}\Phi_P = 2M_1M_2 \sum_{m=1}^r (\pm 1)H_{1,m} \left(\prod_{j=1}^r k_{1,j} \prod_{j=1}^r k_{2,j} \right)^{-1} \nabla_{\lambda}n_m + 2 \sum_{m=1}^r (\pm 1)H_{2,m} \left(\prod_{j=1}^m k_{1,j} \prod_{j=1}^m k_{2,j} \right) \nabla_{\lambda}n_m \quad (6)$$

Here, M_1 and M_2 are the beam magnification factors given by

$$M_1 = \prod_{m=1}^r k_{1,m} \quad (7a)$$

$$M_2 = \prod_{m=1}^r k_{2,m} \quad (7b)$$

For a multiple prism expander designed for an orthogonal beam exit, and Brewster's angle of incidence, Eq. (6) reduces to the succinct expression given by Duarte, in 1990:

$$\nabla_{\lambda}\Phi_P = 2 \sum_{m=1}^r (\pm 1)(n_m)^{m-1} \nabla_{\lambda}n_m \quad (8)$$

Eq. (6) can be used to either quantify the overall dispersion of a given multiple-prism beam expander or to design a prismatic expander yielding zero dispersion, that is, $\nabla_{\lambda}\Phi_P=0$, at a given wavelength.

Physics and Architecture of Solid-State Dye-Laser Oscillators

The first high-performance narrow-linewidth tunable laser was introduced by Hänsch in 1972. This laser yielded a linewidth of $\Delta\nu \approx 2.5$ GHz (or $\Delta\lambda \approx 0.003$ nm at $\lambda \approx 600$ nm) in the absence of an intracavity etalon. Hänsch demonstrated that the laser linewidth from a tunable laser was narrowed significantly when the beam incident on the tuning grating was expanded using an astronomical telescope. The linewidth equation including the intracavity beam magnification factor can be written as

$$\Delta\lambda \approx \Delta\theta (M\nabla_{\lambda}\Theta_G)^{-1} \quad (9)$$

From this equation, it can be deduced that a narrow $\Delta\lambda$ is achieved by reducing $\Delta\theta$ and increasing the overall intracavity dispersion ($M\nabla_{\lambda}\Theta_G$). The intracavity dispersion is optimized by expanding the size of the intracavity beam incident on the diffractive surface of the tuning grating until it is totally illuminated.

Hänsch used a two-dimensional astronomical telescope to expand the intracavity beam incident on the diffraction grating. A simpler beam expansion method consists in the use of a single-prism beam expander as disclosed by several authors (Myers, 1971; Stokes and colleagues, 1972; Hanna and colleagues, 1975). An extension and improvement of this approach was the introduction of multiple-prism beam expanders as reported by Kasuya and colleagues, in 1978, Klauminzer, in 1978, and Duarte and Piper, in 1980. The main advantages of multiple-prism beam expanders, over two-dimensional telescopes, are simplicity, compactness, and the fact that the beam expansion is reduced from two dimensions to one dimension. Physically, as explained previously, prismatic beam expanders also introduce a dispersion component that is absent in the case of the astronomical telescope. Advantages of

multiple-prism beam expanders over single-prism beam expansion are higher transmission efficiency, lower amplified spontaneous emission levels, and the flexibility to either augment or reduce the prismatic dispersion.

In general, for a pulsed multiple-prism grating oscillator, Duarte and Piper, in 1984, showed that the return-pass dispersive linewidth is given by

$$\Delta\lambda = \Delta\theta_R(MR\nabla_\lambda\theta_G + R\nabla_\lambda\Phi_P)^{-1} \quad (10)$$

where R is the number of return-cavity passes. The grating dispersion in this equation, $\nabla_\lambda\theta_G$, can be either from a grating in Littrow or near grazing-incidence configuration. The multiple-return-pass equation for the beam divergence was given by Duarte, in 2001:

$$\Delta\theta_R = (\lambda/\pi w)(1 + (L_R/B_R)^2 + (A_R L_R/B_R)^2)^{1/2} \quad (11)$$

Here, $L_R = (\pi w^2/\lambda)$ is the Rayleigh length of the cavity, w is the beam waist at the gain region, while A_R and B_R are the corresponding multiple-return-pass elements derived from propagation matrices.

At present, very compact and optimized multiple-prism grating tunable laser oscillators are found in two basic cavity architectures. These are the multiple-prism Littrow (MPL) grating laser oscillator, reported by Duarte in 1999 and illustrated in Fig. 8, and the hybrid multiple-prism near grazing-incidence (HMPGI) grating laser oscillator, introduced by Duarte in 1997 and depicted in Fig. 9. In early MPL grating oscillators the individual prisms integrating the multiple-prism expander were deployed in an additive configuration thus adding the cumulative dispersion to that of the grating and thus contributing to the overall dispersion of the cavity. In subsequent architectures the prisms were deployed in compensating configurations so as to yield zero dispersion and thus allow the tuning characteristics of the cavity to be determined by the grating exclusively. In this approach the principal role of the multiple-prism array is to expand the beam incident on the grating thus augmenting significantly the overall dispersion of the cavity as described in Eqs. [9] or [10]. In this regard, it should be mentioned that beam magnification factors of up to 100, and beyond, have been reported in the literature. Using solid-state laser dye gain media, these MPL and HMPGI grating laser oscillators deliver tunable single-longitudinal-mode emission at laser linewidths $350 \text{ MHz} \leq \Delta\nu \leq 375 \text{ MHz}$ and pulse lengths in the 3–7 ns (FWHM) range. Long-pulse operation of this class of multiple-prism grating oscillators has been reported by Duarte and colleagues, in 1998. In these experiments laser linewidths of $\Delta\nu \approx 650 \text{ MHz}$ were achieved at pulse lengths in excess of 100 ns (FWHM) under flashlamp-pumped dye laser excitation.

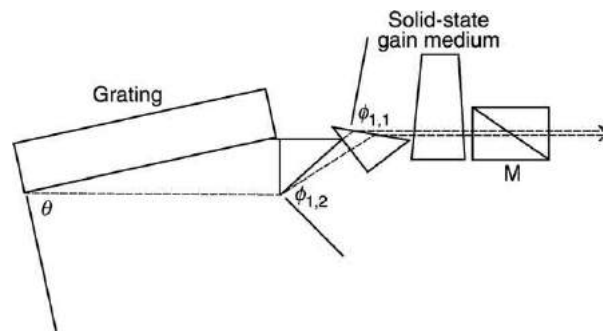


Fig. 8 MPL grating solid-state dye laser oscillator: optimized architecture. The physical dimensions of the optical components of this cavity are shown to scale. The length, of the solid-state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1999) Multiple-prism grating solid-state dye laser oscillator: optimized architecture. *Applied Optics* 38: 6347–6349.

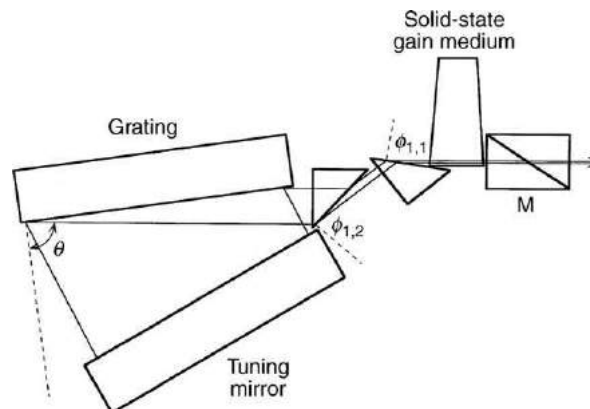


Fig. 9 HMPGI grating solid-state dye laser. Schematics to scale. The length of the solid state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1997) Multiple-prism near-grazing-incidence grating solid-state dye laser oscillator. *Optics and Laser Technology* 29: 513–516.

The dispersive cavity architectures described here have been used with a variety of laser gain media in the gas, the liquid, and the solid state. Applications to tunable semiconductor lasers have also been reported by Zorabedian, in 1992, Duarte, in 1993, and Fox and colleagues, in 1997.

Concepts important to MPL and HMPGI grating tunable laser oscillators include the emission of a single-transverse-mode (TEM_{00}) laser beam in a compact cavity, the use of multiple-prism arrays, the expansion of the intracavity beam incident on the grating, the control of the intracavity dispersion, and the quantification of the overall dispersion of the multiple-prism grating assembly via generalized dispersion equations. Sufficiently high intracavity dispersion leads to the achievement of return-pass dispersive linewidths close to the free-spectral range of the cavity. Under these circumstances single-longitudinal-mode lasing is readily achieved as a result of multipass effects.

A Note on the Cavity Linewidth Equation

So far we have described how the cavity linewidth equation

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda\theta)^{-1}$$

and the dispersion equations can be applied to achieve highly coherent, or very narrow-linewidth, emission. It should be noted that the same physics can be applied to achieve ultrashort, or femtosecond, laser pulses. It turns out that intracavity prisms can be configured to yield negative dispersion as described by Duarte and Piper, in 1982, Dietel and colleagues, in 1983, and Fork and colleagues in 1984. This negative dispersion can reduce significantly the overall dispersion of the cavity. Under those circumstances, Eq. (1) predicts broadband emission which according to the uncertainty principle, in the form of,

$$\Delta\nu\Delta t \approx 1 \quad (12)$$

can lead to very short pulse emission since $\Delta\nu$ and $\Delta\lambda$ are related by the identities

$$\Delta\lambda \approx \lambda^2/\Delta x \quad (13)$$

and

$$\Delta\nu \approx c/\Delta x \quad (14)$$

Distributed-Feedback Solid-State Dye Lasers

Recently, narrow-linewidth laser emission from solid-state dye lasers has also been obtained using a distributed-feedback (DFB) configuration by Wadsworth and colleagues, in 1999, and Zhu and colleagues, in 2000. As is well known, in a DFB laser no external cavity is required, the feedback being provided by Bragg reflection from a permanently or dynamically written grating structure within the gain medium as reported by Kogelnik and Shank, in 1971. By using a DFB laser configuration where the interference of two pump beams induces the required periodic modulation in the gain medium, laser emission with linewidth in the 0.01–0.06 nm range have been reported. Specifically, Wadsworth and colleagues reported a laser linewidth of 12 GHz (0.016 nm at 616 nm) using Perylene Red doped PMMA at a conversion efficiency of 20%.

It should also be mentioned that the DFB laser is also a dye laser development that has found wide and extensive applicability in semiconductor lasers employed in the telecommunications industry.

Laser Dye Survey

Cyanines

These are red and near-infrared dyes which present long conjugated methine chains (!CH=CH!) and are useful in the spectral range longer than 800 nm, where no other dyes compete with them. An important laser dye belonging to this class is 4-dicyanomethylene-2-methyl-6-(*p*-dimethylaminostyryl)-4H-pyran (DCM, Fig. 10(a)), with laser emission in the 600–700 nm range depending on the pump and solvent, liquid or solid, used.

Xanthenes

These dyes have a xanthene ring as the chromophore and are classified into rhodamines, incorporating amino radicals substituents, and fluoresceins, with hydroxyl (OH) radical substituents. They are generally very efficient and chemically stable and their emission covers the wavelength region from 500 to 700 nm.

Rhodamine dyes are the most important group of all laser materials, with rhodamine 6G (Rh6G, Fig. 10(b)), also called rhodamine 590 chloride, being probably the best known of all laser dyes. Rh6G exhibits an absorption peak, in ethanol, at 530 nm and a fluorescence peak at 556 nm, with laser emission typically in the 550–620 nm region. It has been demonstrated to lase efficiently both in liquid and solid solutions.

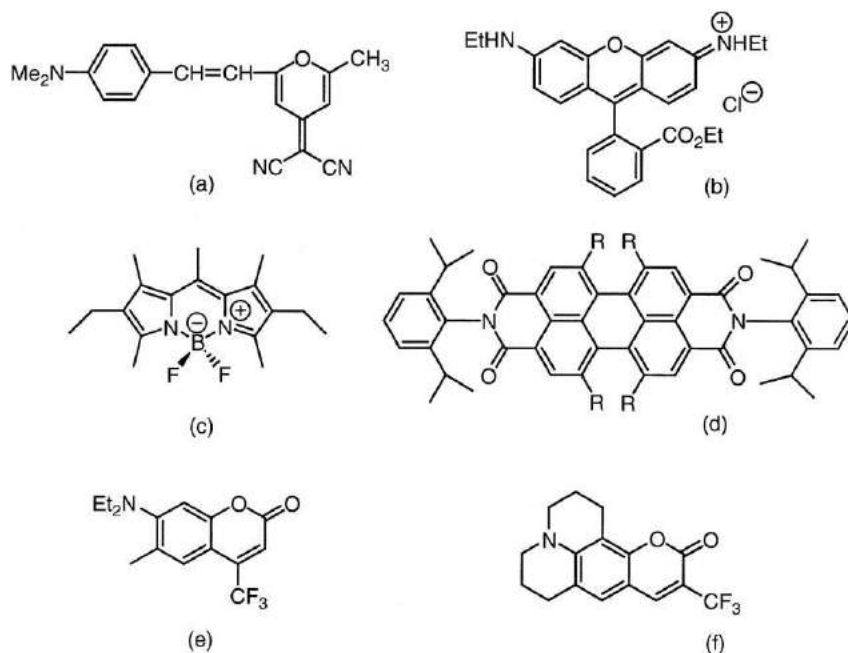


Fig. 10 Molecular structures of some common laser dyes: (a) DCM; (b) Rh6G; (c) PM567; (d) Perylene Orange ($R=H$) and Perylene Red ($R=C_4H_9O$); (e) Coumarin 307 (also known as Coumarin 503); (f) Coumarin 153 (also known as Coumarin 540A).

Pyrromethenes

Pyrromethene. BF_2 complexes are a new class of laser dyes synthesized and characterized more recently as reported by Shah and colleagues, in 1990, Pavlopoulos and colleagues, in 1990, and Boyer and colleagues, in 1993. These laser dyes exhibit reduced triplet-triplet absorption over their fluorescence and lasing spectral region while retaining a high quantum fluorescence yield. Depending on the substituents on the chromophore, these dyes present laser emission over the spectral region from the green/yellow to the red, competing with the rhodamine dyes and have been demonstrated to lase with good performance when incorporated into solid hosts. Unfortunately, they are relatively unstable because of the aromatic amine groups in their structure, which render them vulnerable to photochemical reactions with oxygen as indicated by Rahn and colleagues, in 1997. The molecular structure of a representative and well-known dye of this class, 1,3,5,7,8-pentamethyl-2,6-diethylpyrromethene-difluoroborate complex (pyrromethene 567, PM567) is shown in Fig. 10(c). Recently, analogs of dye PM567 substituted at position 8 with an acetoxypolymethylene linear chain or a polymerizable methacryloyloxy polymethylene chain have been developed. These new dipyrromethene. BF_2 complexes have demonstrated improved efficiency and better photostability than the parent compound both in liquid and solid state.

Conjugated Hydrocarbons

This class of organic compounds includes perylenes, stilbenes, and *p*-terphenyl. Perylene and perylimide dyes are large, nonionic, nonpolar molecules (Fig. 10(d)) characterized by their extreme photostability and negligible singlet-triplet transfer, as discussed by Seybold and Wagenblast and colleagues, in 1989, which have high quantum efficiency because of the absence of nonradiative relaxation. The perylene dyes exhibit limited solubility in conventional solvents but dissolve well in acetone, ethyl acetate, and methyl methacrylate, with emission wavelengths in the orange and red spectral regions.

Stilbene dyes are derivatives of uncyclic unsaturated hydrocarbons such as ethylene and butadiene, with emission wavelengths in the 400–500 nm range. Most of them end in phenyl radicals. Although they are chemically stable, their laser performance is inferior to that of coumarins. *p*-Terphenyl is a valuable and efficient *p*-oligophenylene dye with emission in the ultraviolet.

Coumarins

Coumarin derivatives are a popular family of laser dyes with emission in the blue-green region of the spectrum. Their structure is based on the coumarin ring (Fig. 10(e) and (f)) with different substituents that strongly affect its chemical characteristics and allow covering an emission spectral range between 420 and 580 nm. Some members of this class rank among the most efficient laser dyes known, but their chemical photobleaching is rapid compared with xanthene and pyrromethene dyes. An additional class of blue-green dyes is the tetramethyl derivatives of coumarin dyes introduced by Chen and colleagues, in 1988. The emission from

these dyes spans the spectrum in the 453–588 nm region. These coumarin analogs exhibit improved efficiency, higher solubility, and better lifetime characteristics than the parent compounds.

Azaquinolone Derivatives

The quinolone and azaquinolone derivatives have a structure similar to that of coumarins but with their laser emission range extended toward the blue by 20–30 nm. The quinolone dyes are used when laser emission in the 400–430 nm region is required. Azaquinolone derivatives, such as 7-dimethylamino-1-methyl-4-methoxy-8-azaquinolone-2 (LD 390), exhibit laser action below 400 nm.

Oxadiazole Derivatives

These are compounds which incorporate an axodiazole ring with aryl radical substituents and which lase in the 330–450 nm spectral region. Dye 2-(4-biphenyl)-5-(4-*t*-butylphenyl)-1,3,4-oxadiazole (PBD), with laser emission in the 355–390 nm region, belongs to this family.

Solid-State Laser Dye Matrices

Although some first attempts to incorporate dye molecules into solid matrices were already made in the early days of development of dye lasers, it was not until the late 1980s that host materials with the required properties of optical quality and high damage threshold to laser radiation began to be developed. In subsequent years, the synthesis of new high-performance dyes and the implementation of new ways of incorporating the organic molecules into the solid matrix resulted in significant advances towards the development of practical tunable solid-state dye lasers. A recent detailed review of the work done in this field has been given by Costela *et al.*

Inorganic glasses, transparent polymers, and inorganic–organic hybrid matrices have been successfully used as host matrices for laser dyes. Inorganic glasses offer good thermal and optical properties but present the difficulty that the high melting temperature employed in the traditional process of glass making would destroy the organic dye molecules. It was not until the development of the low-temperature sol-gel technique for the synthesis of glasses that this limitation was overcome. The sol-gel process, based on inorganic polymerization reactions performed at about room temperature and starting with metallo-organic compounds such as alkoxides or salts, as reported by Brinker and Scherrer, in 1989, provides a safe route for the preparation of rigid, transparent, inorganic matrix materials incorporating laser dyes. Disadvantages of these materials are the possible presence of impurities embedded in the matrix or the occurrence of interactions between the dispersed dye molecules and the inorganic structure (hydrogen bonds, van der Waals forces) which, in certain cases, can have a deleterious effect on the lasing characteristics of the material.

The use of matrices based on polymeric materials to incorporate organic dyes offers some technical and economical advantages. Transparent polymers exhibit high optical homogeneity, which is extremely important for narrow-linewidth oscillators, as explained by Duarte, in 1994, good chemical compatibility with organic dyes, and allow control over structure and chemical composition making possible the modification, in a controlled way, of relevant properties of these materials such as polarity, free volume, molecular weight, or viscoelasticity. Furthermore, the polymeric materials are amenable to inexpensive fabrication techniques, which facilitate miniaturization and design of integrated optical systems. The dye molecules can be either dissolved in the polymer or linked covalently to the polymeric chains.

In the early investigations on the use of solid polymer matrices for laser dyes, the main problem to be solved was that of the low resistance to laser radiation exhibited by the then existing materials. In the late 1980s and early 1990s new modified polymeric organic materials began to be developed with a laser-radiation-damage threshold comparable to that of inorganic glasses and crystals as reported by Gromov and colleagues, in 1985, and Dyumaev and colleagues, in 1992. This, together with the above-mentioned advantages, made the use of polymers in solid-state dye lasers both attractive and competitive.

An approach that intends to bring about materials in which the advantages of both inorganic glasses and polymers are preserved but the difficulties are avoided is that of using inorganic–organic hybrid matrices. These are silicate-based materials, with an inorganic Si–O–Si backbone, prepared from organosilane precursors by sol-gel processing in combination with organic cross-linking of polymerizable monomers as reported by Novak, in 1993, Sanchez and Ribot, in 1994, and Schubert, in 1995. In one procedure, laser dyes are mixed with organic monomers, which are then incorporated into the porous structure of a sol-gel inorganic matrix by immersing the bulk in the solution containing monomer and catalyst or photoinitiator as indicated by Reisfeld and Jorgensen, in 1991, and Bosch and colleagues, in 1996. Alternatively, hybrids can be obtained from organically modified silicon alkoxides, producing the so-called ORMOCERS (organically modified ceramics) or ORMOSILS (organically modified silanes) as reported by Reisfeld and Jorgensen, in 1991. An inconvenient aspect of these materials is the appearance of optical inhomogeneities in the medium due to the difference in the refractive index between the organic and inorganic phases as well as to the difference in density between monomer and polymer which causes stresses and the formation of optical defects. As a result, spatial inhomogeneities appear in the laser beam, thereby decreasing its quality as discussed by Duarte, in 1994.

Organic Hosts

The first polymeric materials with improved resistance to damage by laser radiation were obtained by Russian workers, who incorporated rhodamine dyes into modified poly(methyl methacrylate) (MPMMA), obtained by doping PMMA with low-molecular weight additives. Some years later, in 1995, Maslyukov and colleagues demonstrated lasing efficiencies in the range 40–60% with matrices of MPMMA doped with rhodamine dyes pumped longitudinally at 532 nm, with a useful lifetime (measured as a 50% efficiency drop) of 15 000 pulses at a prf of 3.33 Hz.

In 1995, Costela and colleagues used an approach based on adjusting the viscoelastic properties of the material by modifying the internal plasticization of the polymeric medium by copolymerization with appropriate monomers. Using dye Rh6G dissolved in a copolymer of 2-hydroxyethyl methacrylate (HEMA) and methyl methacrylate (MMA) and transversal pumping at 337 nm, they demonstrated laser action with an efficiency of 21% and useful lifetime of 4500 pulses (20 GJ/mol in terms of total input energy per mole of dye molecule when the output energy is down to 50% of its initial value). The useful lifetime increased to 12 000 pulses when the Rh6G chromophore was linked covalently to the polymeric chains as reported by Costela and colleagues, in 1996.

Comparative studies on the laser performance of Rh6G incorporated either in copolymers of HEMA and MMA or in MPMMA were carried out by Giffin and colleagues, in 1999, demonstrating higher efficiency of the MPMMA materials but superior normalized photostability (up to 240 GJ/mol) of the copolymer formulation.

Laser conversion efficiencies in the range 60–70% for longitudinal pumping at 532 nm were reported by Ahmad and colleagues, in 1999, with PM567 dissolved in PMMA modified with 1,4-diazobicyclo[2,2,2] octane (DABCO) or perylene additives. When using DABCO as an additive, a useful lifetime of up to 550 000 pulses, corresponding to a normalized photostability of 270 GJ/mol, was demonstrated at a 2 Hz prf.

Much lower efficiencies and photostabilities have been obtained with dyes emitting in the blue-green spectral region. Costela and colleagues, in 1996, performed some detailed studies on dyes coumarin 540A and coumarin 503 incorporated into methacrylate homopolymers and copolymers, and demonstrated efficiencies of at most 19% with useful lifetimes of up to 1200 pulses, at a 2 Hz prf, for transversal pumping at 337 nm. In 2000, Somasundaram and Ramalingam obtained useful lifetimes of 5240 and 1120 pulses, respectively, for coumarin 1 and coumarin 490 dyes incorporated into PMMA rods modified with ethyl alcohol, under transversal pumping at 337 nm and a prf of 1 Hz.

A detailed study of photo-physical parameters of R6G-doped HEMA-MMA gain media was performed by Holzer and colleagues, in 2000. Among the parameters determined by these authors are quantum yields, lifetimes, absorption cross-sections, and emission cross-sections. A similar study on the photophysical parameters of PM567 and two PM567 analogs incorporated into solid matrices of different acrylic copolymers and in corresponding mimetic liquid solutions was performed by Bergmann and colleagues, in 2001.

Preliminary experiments with rhodamine 6G-doped MPMMA at high prfs were conducted using a frequency-doubled Nd:YAG laser at 5 kHz by Duarte and colleagues, in 1996. In those experiments, involving longitudinal excitation, the pump laser was allowed to fuse a region at the incidence window of the gain medium so that a lens was formed. This lens and the exit window of the gain medium comprised a short unstable resonator that yielded broadband emission. In 2001, Costela and colleagues demonstrated lasing in rhodamine 6G- and pyrromethene-doped polymer matrices under copper-vapor-laser excitation at prfs in the 1.0–6.2 kHz range. In these experiments, the dye-doped solid-state matrix was rotated at 1200 rpm. Average powers of 290 mW were obtained at a prf of 1 kHz and a conversion efficiency of 37%. For a short period of time average powers of up to 1 W were recorded at a prf of 6.2 kHz. In subsequent experiments by Abedin and colleagues the prf was increased to 10 kHz by using as pump laser a diode-pumped, Q-switched, frequency doubled, solid state Nd:YLF laser. Initial average powers of 560 mW and 430 mW were obtained for R6G-doped and PM567-doped polymer matrices, respectively. Lasing efficiency was 16% for R6G and the useful lifetime was 6.6 min (or about 4.0 million shots).

Inorganic and Hybrid Hosts

In 1990, Knobbe and colleagues and McKiernan and colleagues, from the University of California at Los Angeles, incorporated different organic dyes, via the sol-gel techniques, into silicate (SiO_2), aluminosilicate ($\text{Al}_2\text{O}_3\text{-SiO}_2$), and ORMOSIL host matrices, and investigated the lasing performance of the resulting materials under transversal pumping. Lasing efficiencies of 25% and useful lifetimes of 2700 pulses at 1 Hz repetition rate were obtained from Rh6G incorporated into aluminosilicate gel and ORMOSIL matrices, respectively. When the dye in the ORMOSIL matrices was coumarin 153, the useful lifetime was still 2250 pulses. Further studies by Altman and colleagues, in 1991, demonstrated useful lifetimes of 11 000 pulses at a 30 Hz prf when rhodamine 6G perchlorate was incorporated into ORMOSIL matrices. The useful lifetime increased to 14 500 pulses when the dye used was rhodamine B.

When incorporated into xerogel matrices, pyrromethene dyes lase with efficiencies as high as the highest obtained in organic hosts but with much higher photostability: a useful lifetime of 500 000 pulses at 20 Hz repetition rate was obtained by Faloss and colleagues, in 1997, with PM597 using a pump energy of 1 mJ. The efficiency of PM597 dropped to 40% but its useful lifetime increased to over 1 000 000 pulses when oxygen-free samples were prepared, in agreement with previous results that documented the strong dependence of laser parameters of pyrromethene dyes on the presence of oxygen. A similar effect was observed with perylene dyes: deoxygenated xerogel samples of Perylene Red exhibited a useful lifetime of 250 000 pulses at a prf of 5 Hz and 1 mJ pump energy, to be compared with a useful lifetime of only 10 000 pulses obtained with samples prepared in normal conditions. Perylene Orange incorporated into polycom glass and pumped longitudinally at 532 nm gave an efficiency of 72% and a useful lifetime of 40 000 pulses at a prf of 1 Hz as reported by Rhan and King, in 1998.

A direct comparison study of laser performance of Rh6G, PM567, Perylene Red, and Perylene Orange in organic, inorganic, and hybrid hosts was carried out by Rahn and King in 1995. They found that the nonpolar perylene dyes had better performance in partially organic hosts, whereas the ionic rhodamine and pyrromethene dyes performed best in the inorganic sol-gel glass host. The most promising combinations of dye and host for efficiency and photostability were found to be Perylene Orange in polycom glass and Rh6G in sol-gel glass.

An all-solid-state optical configuration, where the pump was a laser diode array side-pumped, Q-switched, frequency-doubled, Nd:YAG slab laser, was demonstrated by Hu and colleagues, in 1998, with dye DCM incorporated into ORMOSIL matrices. A lasing efficiency of 18% at the wavelength of 621 nm was obtained, and 27 000 pulses were needed for the output energy to decrease by 10% of its initial value at 30 Hz repetition rate. After 50 000 pulses the output energy decreased by 90% but it could be recovered after waiting for a few minutes without pumping.

Violet and ultraviolet laser dyes have also been incorporated into sol-gel silica matrices and their lasing properties evaluated under transversal pumping by a number of authors. Although reasonable efficiencies have been obtained with some of these laser dyes, the useful lifetimes are well below 1000 pulses.

More recently, Costela and colleagues have demonstrated improved conversion efficiencies in dye-doped inorganic-organic matrices using pyrromethene 567 dye. The solid matrix in this case is poly-trimethylsilyl-methacrylate cross-linked with ethylene glycol and copolymerized with methyl methacrylate. Also, Duarte and James have reported on dye-doped polymer-nanoparticle laser media where the polymer is PMMA and the nanoparticles are made of silica. These authors report TEM₀₀ laser beam emission and improved dn/dT coefficients of the gain media that results in reduced $\Delta\theta$.

Dye Laser Applications

Dye lasers generated a renaissance in diverse applied fields such as isotope separation, medicine, photochemistry, remote sensing, and spectroscopy. Dye lasers have also been used in many experiments that have advanced the frontiers of fundamental physics.

Central to most applications of dye lasers is their capability of providing coherent narrow-linewidth radiation that can be tuned continuously from the near ultraviolet, throughout the visible, to the near infrared. These unique coherent and spectral properties have made dye lasers particularly useful in the field of high-resolution atomic and molecular spectroscopy. Dye lasers have been successfully and extensively used in absorption and fluorescence spectroscopy, Raman spectroscopy, selective excitations, and nonlinear spectroscopic techniques. Well documented is their use in photochemistry, that is, the chemistry of optically excited atoms and molecules, and in biology, studying biochemical reaction kinetics of biological molecules. By using nonlinear optical techniques, such as harmonic and sum frequency and difference frequency generation, the properties of dye lasers can be extended to the ultraviolet.

Medical applications of dye lasers include cancer photodynamic therapy, treatment of vascular lesions, and lithotripsy as described by Goldman in 1990.

The ability of dye lasers to provide tunable sub-GHz linewidths in the orange-red portion of the spectrum, at kW average powers, made them particularly suited to atomic vapor laser isotope separation of uranium as reported by Bass and colleagues, in 1992, Singh and colleagues, in 1994, and Nemoto and colleagues, in 1995. In this particular case the isotopic shift between the two uranium isotopes is a few GHz. Using dye lasers yielding $\Delta\nu \leq 1$ GHz the ^{235}U can be selectively excited in a multistep process by tuning the dye laser(s) to specific wavelengths, in the orange-red spectral region, compatible with a transition sequence leading to photoionization. Also, high-prf narrow-linewidth dye lasers, as described by Duarte and Piper, in 1984, are ideally suited for guide star applications in astronomy. This particular application requires a high-average power diffraction-limited laser beam at $\lambda \approx 589$ nm to propagate for some 95 km, above the surface of the Earth, and illuminate a layer of sodium atoms. Ground detection of the fluorescence from the sodium atoms allows measurements on the turbulence of the atmosphere. These measurements are then used to control the adaptive optics of the telescope thus compensating for the atmospheric distortions.

A range of industrial applications is also amenable to the wavelength agility and high-average power output characteristics of dye lasers. Furthermore, tunability coupled with narrow-linewidth emission, at large pulsed energies, make dye lasers useful in military applications such as directed energy and damage of optical sensors.

The Future of Dye Lasers

Predicting the future is rather risky. In the early 1990s, articles and numerous advertisements in commercial laser magazines predicted the demise and even the oblivion of the dye laser in a few years. It did not turn out this way. The high power and wavelength agility available from dye lasers have ensured their continued use as scientific and research tools in physics, chemistry, and medicine. In addition, the advent of narrow-linewidth solid-state dye lasers has sustained a healthy level of research and development in the field. Today, some 20 laboratories around the world are engaged in this activity.

The future of solid-state dye lasers depends largely on synthetic and manufacturing improvements. There is a need for strict control of the conditions of fabrication and a careful purification of all the compounds involved. In the case of polymers, in particular, stringent control of thermal conditions during the polymerization step is obligatory to ensure that adequate optical uniformity of the polymer matrix is achieved, and that intrinsic anisotropy developed during polymerization is minimized. From

an engineering perspective what are needed are dye-doped solid-state media with improved photostability characteristics and better thermal properties. By better thermal properties is meant better dn/dT factors. To a certain degree progress in this area has already been reported. As in the case of liquid dye lasers, the lifetime of the material can be improved by using pump lasers at compatible wavelengths. In this regard, direct diode laser pumping of dye-doped solid-state matrices should lead to very compact, long-lifetime, tunable lasers emitting throughout the visible spectrum. An example of such a device would be a green laser-diode-pumped rhodamine-6G doped MPMMA tunable laser configured in a multiple-prism grating architecture.

As far as liquid lasers are concerned, there is a need for efficient water-soluble dyes. Successful developments in this area were published in the literature with coumarin-analog dyes by Chen and colleagues, in 1988, and some encouraging results with rhodamine dyes have been reported by Ray and colleagues, in 2002. Efficient water-soluble dyes could lead to significant improvements and simplifications in high-power tunable lasers.

Further Reading

- Costela, A., García-Moreno, I., Sastre, R., 2001. In: Nalwa, H.S. (Ed.), *Handbook of Advanced Electronic and Photonic Materials and Devices*, vol. 7. San Diego: Academic Press, pp. 161–208.
- Diels, J.-C., Rudolph, W., 1996. *Ultrashort Laser Pulse Phenomena*. New York: Academic.
- Demtröder, W., 1995. *Laser Spectroscopy*, 2nd edn Berlin: Springer-Verlag.
- Duarte, F.J. (Ed.), 1991. *High Power Dye Lasers*. Berlin: Springer-Verlag.
- Duarte, F.J. (Ed.), 1995. *Tunable Lasers Handbook*. New York: Academic.
- Duarte, F.J. (Ed.), 1995. *Tunable Laser Applications*. New York: Marcel Dekker.
- Duarte, F.J., 2003. *Tunable Laser Optics*. New York: Elsevier Academic.
- Duarte, F.J., Hillman, L.W. (Eds.), 1990. *Dye Laser Principles*. New York: Academic Press.
- Maeda, M., 1984. *Laser Dyes*. New York: Academic Press.
- Schäfer, F.P. (Ed.), 1990. *Dye Lasers*, 3rd edn Berlin: Springer-Verlag.
- Radziemski, L.J., Solarz, R.W., Paisner, J.A. (Eds.), 1987. *Laser Spectroscopy and its Applications*. New York: Marcel Dekker.
- Weber, M.J., 2001. *Handbook of Lasers*. New York: CRC.

Introduction

The semiconductor edge emitting diode laser is a critical component in a wide variety of applications, including fiber optics telecommunications, optical data storage, and optical remote sensing. In this section, we describe the basic structures of edge emitting diode lasers and the physical mechanisms for converting electrical current into light. Laser waveguides and cavity resonators are also outlined. The power, efficiency, gain, loss, and threshold characteristics of laser diodes are presented along with the effects of temperature and modulation. Quantum well lasers are outlined and a description of grating coupled lasers is provided.

Like all forms of laser, the edge emitting diode laser is an oscillator and has three principal components; a mechanism for converting energy into light, a medium that has positive optical gain, and a mechanism for obtaining optical feedback. The basic edge emitting laser diode is shown schematically in Fig. 1. Current flow is vertical through a pn junction and light is emitted from the ends. The dimensions of the laser in Fig. 1 are distorted, in proportion to typical real laser diodes, to reveal more detail. In practical laser diodes, the width of the stripe contact is actually much smaller than shown and the thickness is smaller still. Typical dimensions are (stripe width $\times$ thickness $\times$ length) $2 \times 100 \times 1000 \mu\text{m}$. Light emission is generated only within a few micrometers of the surface, which means that 99.98% of the volume of this small device is inactive.

Energy conversion in diode lasers is provided by current flowing through a forward-biased pn junction. At the electrical junction, there is a high density of injected electrons and holes which can recombine in a direct energy gap semiconductor material to give an emitted photon. Charge neutrality requires equal numbers of injected electrons and holes. Fig. 2 shows a simplified energy versus momentum ($E-k$) diagrams for (a) direct, and (b) indirect semiconductors. In a direct energy gap material, such as GaAs, the energy minima have the same momentum vector, so recombination of an electron in the conduction band and a hole in the valence band takes place directly. In an indirect material, such as silicon, momentum conservation requires the participation of a third particle – a phonon. This third-order process is far less efficient than direct recombination and, thus far, too inefficient to support laser action.

Optical gain in direct energy gap materials is obtained when population inversion is reached. Population inversion is the condition where the probability for stimulated emission exceeds that of absorption. The density of excited electrons n , in the conduction band is given by

$$n = \int_{E_c}^{\infty} \rho_n(E) f_n(E) dE \text{ cm}^{-3} \quad (1)$$

where $\rho_n(E)$ is the conduction band density of states function and $f_n(E)$ is the Fermi occupancy function. A similar equation can be

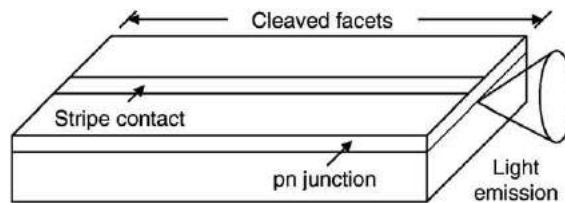


Fig. 1 Schematic drawing of an edge emitting laser diode.

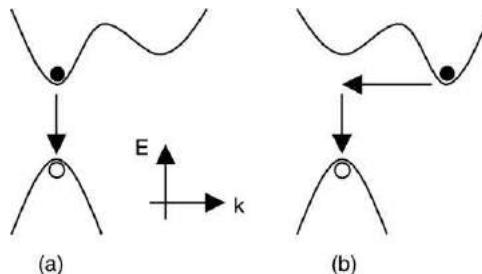


Fig. 2 Energy versus momentum for (a) direct, and (b) indirect semiconductors.

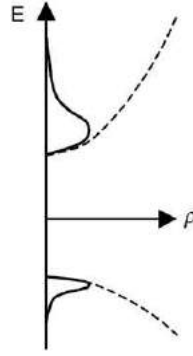


Fig. 3 Density of states versus energy for conduction and valence bands.

written for holes in the valence band. **Fig. 3** shows the density of states ρ versus energy as a dashed line for conduction and valence bands. The solid lines in **Fig. 3** are the $\rho(E)f(E)$ products, and the areas under the solid curves are the carrier densities n and p .

The Fermi functions are occupation probabilities based on the quasi-Fermi levels E_{Fn} and E_{Fp} , which are, in turn, based on the injected carrier densities, δn and δp . Population inversion implies that the difference between the quasi-Fermi levels must exceed the emission energy which, for semiconductors, must be greater than the bandgap energy:

$$E_{Fn} - E_{Fp} \geq \hbar\omega \sim E_g \quad (2)$$

Charge neutrality requires that $\delta n = \delta p$. Transparency is the point where these two conditions are just met and any additional injected current results in optical gain. The transparency current density J_o , is given by

$$J_o = \frac{qn_o L_z}{\tau_{\text{spn}}} \quad (3)$$

where n_o is the transparency carrier density ($\delta n = \delta p = n_o$ at transparency), L_z is the thickness of the optically active layer, and τ_{spn} is the spontaneous carrier lifetime in the material (typically 5 nsec).

A resonant cavity for optical feedback is obtained simply by taking advantage of the natural crystal formation in single crystal semiconductors. Most III-V compound semiconductors form naturally into a zincblende lattice structure. With the appropriate choice of crystal planes and surfaces, this lattice structure can be cleaved such that nearly perfect plane-parallel facets can be formed at arbitrary distances from each other. Since the refractive index of these materials is much larger (~ 3.5) than that of air, there is internal optical reflection for normal incidence of approximately 30%. This type of optical cavity is called a Fabry-Perot resonator.

Effective lasers of all forms make use of optical waveguides to optimize the overlap of the optical field with material gain and cavity resonance. The optical waveguide in an edge emitting laser is best described by considering it in terms of a transverse waveguide component, defined by the epitaxial growth of multiple thin heterostructure layers, and a lateral waveguide component, defined by conventional semiconductor device processing methods. One of the simplest, and yet most common, heterostructure designs is shown in **Fig. 4**. This five-layer separate confinement heterostructure (SCH) offers excellent carrier confinement in the active layer as well as a suitable transverse waveguide.

Clever choice of different materials for each layer results in the energy band structure, index of refraction profile, and optical field mode profile shown in **Fig. 5**. The outer confining layers are the thickest ($\sim 1 \mu\text{m}$) and have the largest energy bandgap and lowest refractive index. The inner barrier layers are thinner ($\sim 0.1 \mu\text{m}$), have intermediate bandgap energy and index of refraction, and serve both an optical role in defining the optical waveguide and an electronic role in confining carriers to the active layer. The active layer is the narrowest bandgap, highest index material, and is the only layer in the structure designed to provide optical gain. It may be a single layer or, as in the case of a multiple quantum well laser, a combination of layers. The bandgap energy, E , of this active layer plays a key role in determining the emission wavelength, λ , of the laser:

$$E = \frac{hc}{\lambda} \quad (4)$$

where h is Planck's constant and c is the speed of light.

The bandgap energy of the SCH structure is shown in **Fig. 5**. The lowest energy active layer collects injected electrons and holes and the carrier confinement provided by the inner barrier layers allows the high density of carriers necessary for population inversion and gain. The layers in the structure also have the refractive index profile shown in **Fig. 5** which forms a five-layer slab dielectric waveguide. With appropriate values for thicknesses and indices of refraction, this structure can be a very strong fundamental mode waveguide with the field profile shown. A key parameter of edge emitting diode lasers is Γ , the optical confinement factor. This parameter is the areal overlap of the optical field with the active layer, shown as dashed lines in the figure and can be as little as a few percent, even in very high performance lasers.

Lateral waveguiding in a diode laser can be accomplished in a variety of ways. A common example of an index guided laser is the ridge waveguide laser shown in **Fig. 6**. After the appropriate transverse waveguide heterostructure sample is grown,

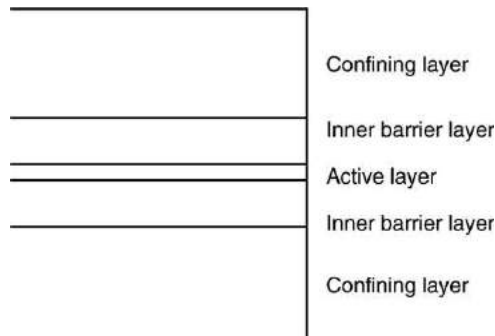


Fig. 4 Schematic cross-section of a typical five-layer separate confinement heterostructure (SCH) laser.

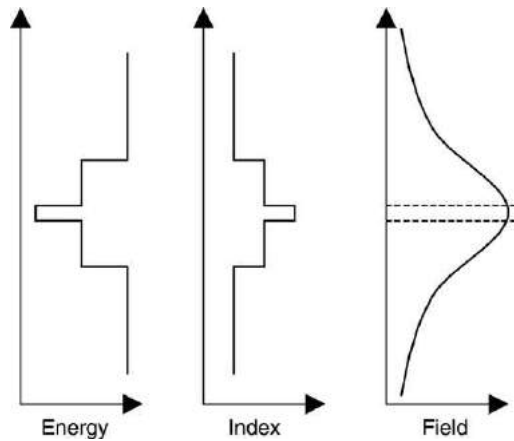


Fig. 5 Energy band structure, index of refraction profile, and optical field mode profile of the SCH laser of Fig. 4.

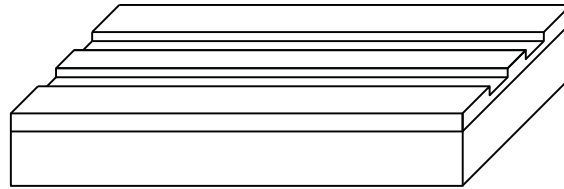


Fig. 6 Schematic diagram of a real index guided ridge waveguide diode laser.

conventional semiconductor processing methods are used to form parallel etched stripes from the surface near, but not through, the active layer. An oxide mask is patterned such that electrical contact is formed only to the center (core) region between the etched stripes, the core stripe being only a few microns wide. The lateral waveguide that results arises from the average index of refraction (effective index) in the etched regions outside the core being smaller than that of the core. Fig. 7 shows the asymmetric near field emission profile of the ridge waveguide laser and cross-sectional profiles of the laser emission and effective index of refraction.

The optical spectrum of an edge emitting diode laser below and above threshold is the superposition of the material gain of the active layer and the resonances provided by the Fabry–Perot resonator. The spectral variation of material gain with injected carrier density (drive current) is shown in Fig. 8. As the drive is increased, the intensity and energy of peak gain both increase.

The Fabry–Perot resonator has resonances every half wavelength, given by

$$L = \frac{m \lambda}{2 n} \quad (5)$$

where L is the cavity length, λ/n is the wavelength in the medium (n is the refractive index), and m is an integer. For normal length edge emitter cavities, these cavity modes are spaced only 1 or 2 Å apart. The result is optical spectra, near laser threshold and above, that look like the spectra of Fig. 9. Many cavity modes resonate up to threshold, while above threshold, the superlinear increase in emission intensity with drive current tends to greatly favor one or two of the Fabry–Perot longitudinal cavity modes.

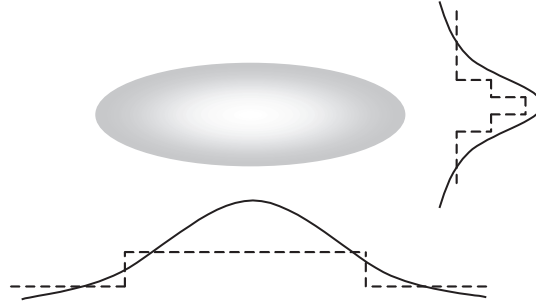


Fig. 7 Near field emission pattern, cross-sectional emission profiles (solid lines), and refractive index profiles (dashed lines) from an index guided diode laser.

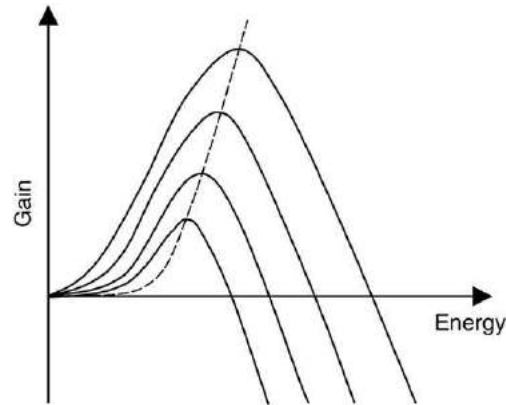


Fig. 8 Gain versus emission energy for four increasing values of carrier (current) density. Peak gain is shown as a dashed line.

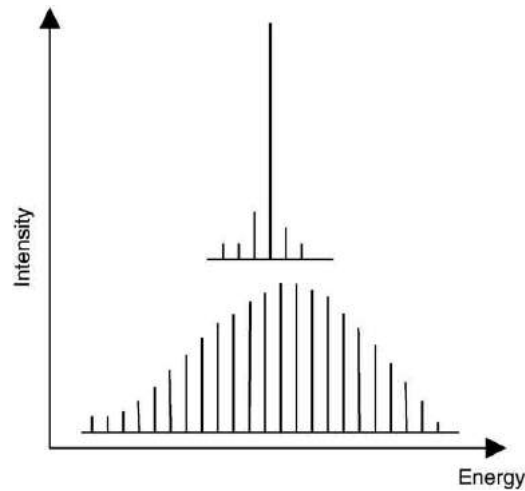


Fig. 9 Emission spectra from a Fabry-Perot edge emitting laser structure at, and above, laser threshold.

Laser threshold can be determined from a relatively simple analysis of the round trip gain and losses in the Fabry-Perot resonator. If the initial intensity is I_0 , after one round trip the intensity will be given by

$$I = I_0 R_1 R_2 e^{(g - \alpha_i) 2L} \quad (6)$$

where R_1 and R_2 are the reflectivities of the two facets, g is the optical gain, α_i are losses associated with optical absorption (typically $3\text{--}15\text{ cm}^{-1}$), and L is the cavity length. At laser threshold, $g = g_{\text{th}}$ and $I = I_0$, i.e., the round trip gain exactly equals the losses. The result is

$$g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \quad (7)$$

where the second term represents the mirror losses α_m . Of course, mirror losses are desirable in the sense that they represent useable output power. Actually, the equation above does not take into account the incomplete overlap of the optical mode with the gain in the active layer. The overlap is defined by the confinement factor Υ , described above, and the equation must be modified to become

$$\Gamma g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \quad (8)$$

For quantum well lasers, the material peak gain, shown in Fig. 8, as a function of drive current is given approximately by

$$g_t = \beta J_o \ln \frac{J}{J_o} \quad (9)$$

where J_o is the transparency current density. The threshold current density J_{th} , which is the current density where the gain exactly equals all losses, is given by

$$J_{th} = \frac{J_o}{\eta_i} \exp \frac{\alpha + \frac{1}{2L} \ln \frac{1}{R_1 R_2}}{\Gamma \beta J_o} \quad (10)$$

Note that an additional parameter η_i has appeared in this equation. Not all of the carriers injected by the drive current participate in optical processes; some are lost to other nonradiative processes. This is accounted for by including the internal quantum efficiency term η_i , which is typically at least 90%. The actual drive current, of course, must include the geometry of the device and is given by

$$I_{th} = wLJ_{th} \quad (11)$$

where L is the cavity length, and w is the effective width of the active volume. The effective width is basically the stripe width of the core but also includes current spreading and carrier diffusion under the stripe.

The power-current characteristic (P - I) for a typical edge emitting laser diode is shown in Fig. 10. As the drive current is increased from zero, the injected carrier densities increase and spontaneous recombination is observed. At some point the gain equals the internal absorption losses, and the material is transparent. Then, as the drive current is further increased, additional gain eventually equals the total losses, including mirror losses, and laser threshold is reached. Above threshold, nearly all additional injected carriers contribute to stimulated emission (laser output). The power generated internally is given by

$$P = wL(J - J_{th})\eta_i \frac{hv}{q} \quad (12)$$

where wL is the area and hv is the photon energy. Only a fraction of the power generated internally is extracted from the device

$$P_o = wL(J - J_{th})\eta_i \frac{hv}{q} \frac{\alpha_m}{\alpha_i + \alpha_m} \quad (13)$$

Clearly a useful design goal is to minimize the internal optical loss α_i . The slope of the P - I curve of Fig. 10 above threshold is described by an external differential quantum efficiency η_d , which is related to the internal quantum efficiency by

$$\eta_d = \eta_i \left[\frac{\alpha_m}{\alpha_i + \alpha_m} \right] \quad (14)$$

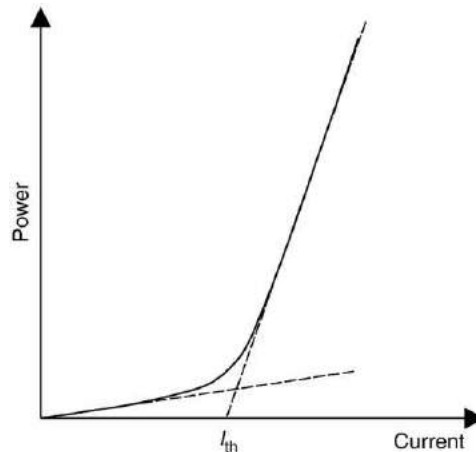


Fig. 10 Power versus current (P - I) curve for a typical edge emitting laser diode.

Other common parameters used to characterize the efficiency of laser diodes include the power conversion efficiency η_p

$$\eta_p = \frac{P_o}{V_F I} \quad (15)$$

and the wall plug efficiency η_w

$$\eta_w = \frac{P_o}{I} \quad (16)$$

The power conversion efficiency η_p , recognizes that the forward bias voltage V_F is larger than the photon energy by an amount associated with the series resistance of the laser diode. The wall plug efficiency η_w , has the unusual units of WA^{-1} and is simply a shorthand term that relates the common input unit of current to the common output unit of power. In high performance low power applications, the ideal device has minimum threshold current since the current used to reach threshold contributes little to light emission. In high power applications, the threshold current becomes insignificant and the critical parameter is external quantum efficiency.

Temperature effects may be important for edge emitting lasers, depending on a particular application. Usually, concerns related to temperature arise under conditions of high temperature operation ($T > 50^\circ\text{C}$), where laser thresholds rise and efficiencies fall. It became common early in the development of laser diodes to define a characteristic temperature T_o for laser threshold, which is given by

$$I_{th} = I_{tho} \exp\left(\frac{T}{T_o}\right) \quad (17)$$

where a high T_o is desirable. Unfortunately, this simple expression does not describe the temperature dependence very well over a wide range of temperatures, so the appropriate temperature range must also be specified. For applications where the emission wavelength of the laser is important, the change in wavelength with temperature $d\lambda/dT$ may also be specified. For conventional Fabry-Perot cavity lasers, $d\lambda/dT$ is typically $3\text{--}5 \text{ \AA}^\circ\text{C}^{-1}$.

In order for a diode laser to carry information, some form of modulation is required. One method for obtaining this is direct modulation of the drive current in a semiconductor laser. The transient and temporal behavior of lasers is governed by rate equations for carriers and photons which are necessarily coupled equations. The rate equation for carriers is given by

$$\frac{dn}{dt} = \frac{J}{qL_z} - \frac{n}{\tau_{sp}} - (c/n)\beta(n - n_o)\varphi(E) \quad (18)$$

where J/qL_z is the supply, n/τ_{sp} is the spontaneous emission rate, and the third term is the stimulated emission rate $(c/n)\beta(n - n_o)\varphi(E)$ where β is the gain coefficient and $\varphi(E)$ is the photon density. The rate equation for photons is given by

$$\frac{d\varphi}{dt} = (c/n)\beta(n - n_o)\varphi + \frac{\theta n}{\tau_{sp}} - \frac{\varphi}{\tau_p} \quad (19)$$

where θ is the fraction of the spontaneous emission that couples into the mode (a small number), and τ_p is the photon lifetime ($\sim 1 \text{ psec}$) in the cavity. These rate equations can be solved for specific diode laser materials and structures to yield a modulation frequency response. A typical small signal frequency response, at two output power levels, is shown in Fig. 11. The response is typically flat to frequencies above 1 GHz, rises to a peak value at some characteristic resonant frequency, and quickly rolls off. The characteristic small signal resonant frequency ω_r is given by

$$\omega_r^2 = \frac{(c/n)\beta\bar{\varphi}}{\tau_p} \quad (20)$$

where $\bar{\varphi}$ is the average photon density. Direct modulation of edge emitting laser diodes is limited by practicality to less than 10 GHz. This is in part because of the limits imposed by the resonant frequency and in part by chirp. Chirp is the frequency modulation that arises indirectly from direct current modulation. Modulation of the current modulates the carrier densities which,

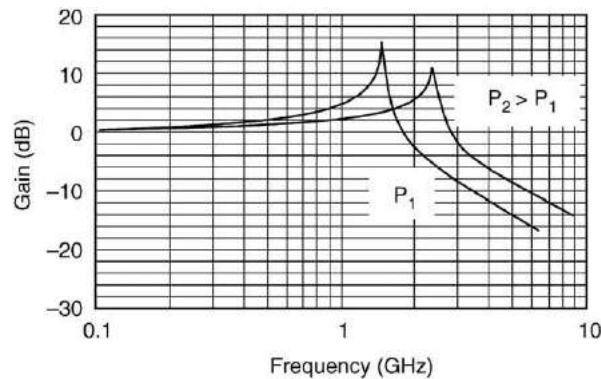


Fig. 11 Direct current modulation frequency response for an edge emitting diode laser at two output power levels.

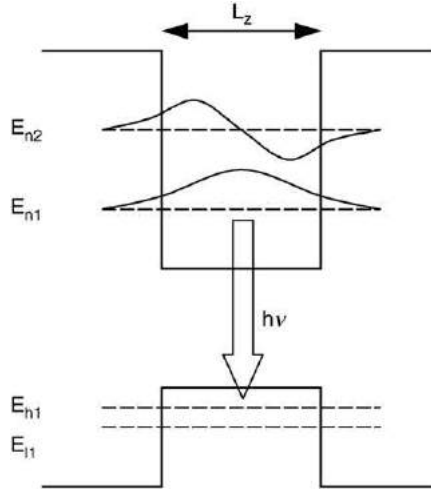


Fig. 12 Energy band diagram for a quantum well heterostructure.

in turn, results in modulation of the quasi-Fermi levels. The separation of the quasi-Fermi levels is the emission energy (wavelength). Most higher-speed lasers make use of external modulation schemes.

The discussion thus far has addressed aspects of semiconductor diode edge emitting lasers that are common to all types of diode lasers irrespective of the choice of materials or the details of the active layer structure. The materials of choice for efficient laser devices include a wide variety of III–V binary compounds and ternary or quaternary alloys. The particular choice is based on such issues as emission wavelength, a suitable heterostructure lattice match, and availability of high-quality substrates. For example, common low-power lasers for CD players ($\lambda \sim 820$ nm) are likely to be made on a GaAs substrate using a GaAs–Al_xGa_{1–x}As heterostructure. Lasers for fiberoptic telecommunications systems ($\lambda \sim 1.3$ – 1.5 μ m) are likely to be made on an InP substrate using an InP–In_xGa_{1–x}As_yP_{1–y} heterostructure. Red laser diodes ($\lambda \sim 650$ nm) are likely to be made on a GaAs substrate using a GaAs–In_xGa_{1–x}As_yP_{1–y} heterostructure. Blue and ultraviolet lasers, a relatively new technology compared to the others, make use of Al_xGa_{1–x}N–GaN–In_yGa_{1–y}N heterostructures formed on sapphire or silicon carbide substrates.

An important part of many diode laser structures is a strained layer. If a layer is thin enough, the strain arising from a modest amount of lattice mismatch can be accommodated elastically. In thicker layers, lattice mismatch results in an unacceptable number of dislocations that affect quantum efficiency, optical absorption losses, and, ultimately, long-term failure rates. Strained layer GaAs–In_xGa_{1–x}As lasers are a critical component in rare-earth doped fiber amplifiers.

The structure of the active layer in edge emitting laser diodes was originally a simple double heterostructure configuration with a single active layer of 500–1000 Å in thickness. In the 1970s however, advances in the art of growing semiconductor heterostructure materials led to growth of high-quality quantum well active layers. These structures, having thicknesses that are comparable to the electron wavelength in a semiconductor (< 200 Å), revolutionized semiconductor lasers, resulting in much lower threshold current densities, higher efficiencies, and a broader range of available emission wavelengths. The energy band diagram for a quantum well laser active region is shown in Fig. 12.

This structure is a physical realization of the particle-in-a-box problem in elementary quantum mechanics. The quantum well yields quantum states in the conduction band at discrete energy levels, with odd and even electron wavefunctions, and breaks the degeneracy in the valence band, resulting in separate energy states for light holes and heavy holes. The primary transition for recombination is from the $n=1$ electron state to the $h=1$ heavy hole state and takes place at a higher energy than the bulk bandgap energy. In addition, the density of states for quantum wells becomes a step-like function and optical gain is enhanced. The advantages in practical edge emitting lasers that are the result of one-dimensional quantization are such that virtually all present commercial laser diodes are quantum well lasers. These kinds of advantages are driving development of laser diodes with additional degrees of quantization, including two dimensions (quantum wires) and three dimensions (quantum dots).

The Fabry–Perot cavity resonator described here is remarkably efficient and relatively easy to fabricate, which makes it the cavity of choice for many applications. There are many applications, however, where the requirements for linewidth of the laser emission or the temperature sensitivity of the emission wavelength make the Fabry–Perot resonator less desirable. An important laser technology that addresses both of these concerns involves the use of a wavelength selective grating as an integral part of the laser cavity. A notable example of a grating coupled laser is distributed feedback laser shown schematically in Fig. 13(a). This is similar to the SCH cross-section of Fig. 4 with the addition of a Bragg grating above the active layer.

The period Λ , of the grating is chosen to fulfill the Bragg condition for m th-order coupling between forward- and backward-propagating waves, which is

$$\Lambda = \frac{m\lambda}{2n} \quad (21)$$

where λ/n is the wavelength in the medium. The index step between the materials on either side of the grating is small and the

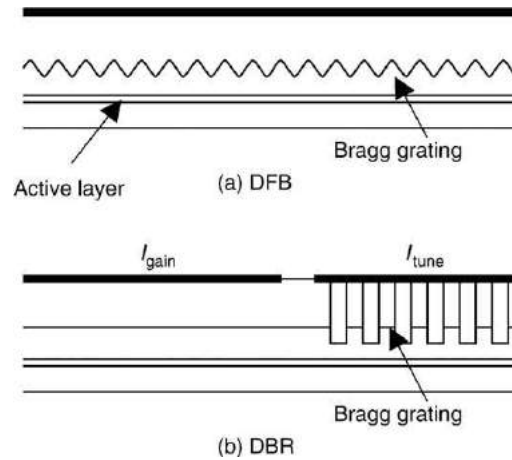


Fig. 13 Schematic diagram of distributed feedback (DFB) and distributed Bragg reflector (DBR) edge emitting laser diodes.

amount of reflection is also small. Since the grating extends throughout the length of the cavity, however, the overall effective reflectivity can be large. Another variant is the distributed Bragg reflector (DBR) laser resonator, shown in [Fig. 13\(b\)](#). This DBR laser has a higher index contrast, deeply etched surface grating located at one or both ends of the otherwise conventional SCH laser heterostructure. One of the advantages of this structure is the opportunity to add a contact for tuning the Bragg grating by current injection.

Both the DBR and DFB laser structures are particularly well suited for telecommunications or other applications where a very narrow single laser emission line is required. In addition, the emission wavelength temperature dependence for this type of laser is much smaller, typically $0.5 \text{ \AA}^\circ\text{C}^{-1}$.

Further Reading

- Agrawal, G.P. (Ed.), 1995. *Semiconductor Lasers Past, Present, and Future*. Woodbury, NY: American Institute of Physics.
- Agrawal, G.P., Dutta, N.K., . *Semiconductor Lasers*. New York: Van Nostrand Reinhold.
- Bhattacharya, P., 1994. *Semiconductor Optoelectronic Devices*. Upper Saddle River, NJ: Prentice-Hall.
- Botez, D., Scifres, D.R. (Eds.), 1994. *Diode Laser Arrays*. Cambridge, UK: Cambridge University Press.
- Chuang, S.L., 1995. *Physics of Optoelectronic Devices*. New York: John Wiley & Sons.
- Coleman, J.J., 1992. *Selected Papers on Semiconductor Diode Lasers*. Bellingham, WA: SPIE Optical Engineering Press.
- Einspruch, N.G., Frensley, W.R. (Eds.), 1994. *Heterostructures and Quantum Devices*. San Diego, CA: Academic Press.
- Iga, K., 1994. *Fundamentals of Laser Optics*. Plenum Press.
- Kapon, E., 1999. *Semiconductor Lasers I and II*. San Diego, CA: Academic Press.
- Nakamura, S., Fasol, G., 1997. *The Blue Laser Diode*. Berlin: Springer.
- Verdeyen, J.T., 1989. *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
- Zory Jr, P.S. (Ed.), 1993. *Quantum Well Lasers*. San Diego, CA: Academic Press.

Excimer Lasers

JJ Ewing, Ewing Technology Associates, Inc., Bellevue, WA, USA

© 2018 Elsevier Inc. All rights reserved.

Introduction

We present a summary of the fundamental operating principles of ultraviolet, excimer lasers. After a brief discussion of the economics and application motivation, the underlying physics and technology of these devices is described. Key issues and limitations in the scaling of these lasers are presented.

Background: Why Excimer Lasers?

Excimer lasers have become the most widely used source of moderate power pulsed ultraviolet sources in laser applications. The range of manufacturing applications is also large. Production of computer chips, using excimer lasers as the illumination source for lithography, has by far the largest commercial manufacturing impact. In the medical arena, excimer lasers are used extensively in what is becoming one of the world's most common surgical procedures, the reshaping of the lens to correct vision problems. Taken together, the annual production rate for these lasers is a relatively modest number compared to the production of semiconductor lasers. Current markets are in the range of \$0.4 billion per year (Fig. 1). More importantly, the unique wavelength, power, and pulse energy properties of the excimer laser enable a systems and medical procedure market, based on the excimer laser, to exceed \$3 billion per year. The current average sales price for these lasers is in the range of \$250 000 per unit, driven primarily by the production costs and reliability requirements of the lasers for lithography for chip manufacture. Relative to semiconductor diode lasers, excimer lasers have always been, and will continue to be, more expensive per unit device by a factor of order 10^4 . UV power and pulse energy, however, are considerably higher.

The fundamental properties of these lasers enable the photochemical and photophysical processes used in manufacturing and medicine. Short pulses of UV light can photo-ablate materials, being of use in medicine and micromachining. Short wavelengths enable photochemical processes. The ability to provide a narrow linewidth using appropriate resonators enables precise optical processes, such as lithography. We trace out some of the core underlying properties of these lasers that lead to the core utility. We also discuss the technological problems and solutions that have evolved to make these lasers so useful.

The data in Fig. 1 provide a current answer to the question of 'why excimer lasers?' At the dawn of the excimer era, the answer to this question was quite different. In the early 1970s, there were no powerful short wavelength lasers, although IR lasers were being scaled up to significant power levels. However, visible or UV lasers could in principle provide better propagation and focus to a very small spot size at large distances. Solid state lasers of the time offered very limited pulse repetition frequency and low

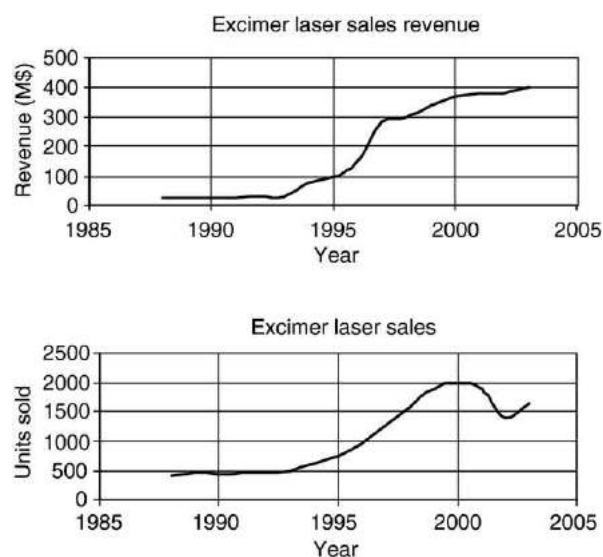


Fig. 1 The sales of excimer lasers have always been measured in small numbers relative to other, less-expensive lasers. For many years since their commercial release in 1976 the sales were to R&D workers in chemistry, physics and ultimately biology and medicine. The market for these specialty but most useful R&D lasers was set to fit within a typical researcher's annual capital budget, i.e., less than \$100K. As applications and procedures were developed and certified or approved, sales and average unit sales price increased considerably. Reliance on cyclical markets like semiconductor fabrication led to significant variations in production rates and annual revenues as can be seen.

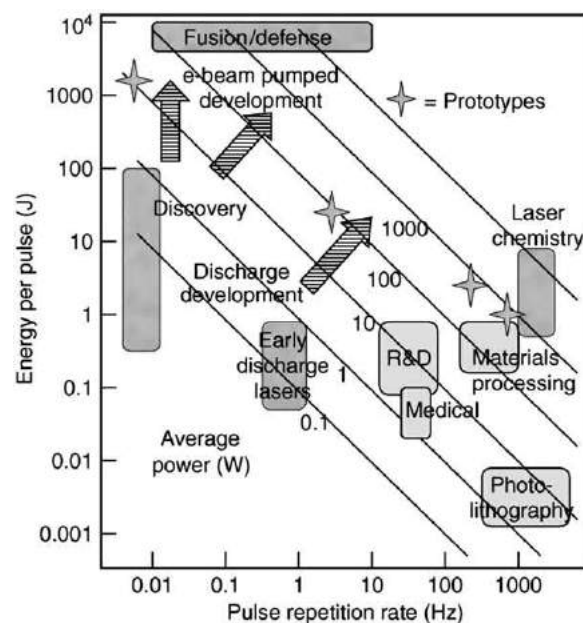


Fig. 2 The typical energy and pulse rate for certain applications and technologies using excimer lasers. For energy under a few J, a discharge laser is used. The earliest excimer discharge lasers were derivatives of CO₂ TEA (transversely excited atmospheric pressure) and produced pulse energy in the range of 100 mJ per pulse. Pulse rates of order 30 Hz were all that the early pulsed power technology could provide. Early research focused on generating high energy. Current markets are at modest UV pulse energy and high pulse rates for lithography and low pulse rates for medical applications.

efficiency. Diode pumping of solid state lasers, and diode arrays to excite such lasers, was a far removed development effort. As such, short-wavelength lasers were researched over a wide range of potential media and lasing wavelengths. The excimer concept was one of those candidates.

We provide, in [Fig. 2](#), a 'positioning' map showing the range of laser parameters that can be obtained with commercial or developmental excimer lasers. The map has coordinates of pulse energy and pulse rate, with diagonal lines expressing average power. We show where both development goals and current markets lay in a map. Current applications are shown by the boxes in the lower right-hand corner. Excimer lasers have been built with clear apertures in the range of 0.1 to 10⁴ cm², with single pulse energies covering a range of 10⁷. Lasers with large apertures require electron beam excitation. Discharge lasers correspond to the more modest pulse energy used for current applications. In general, for energies lower than ~5 J, the excitation method of choice is a self-sustained discharge, clearly providing growth potential for future uses, if required.

The applications envisaged in the early R&D days were in much higher energy uses, such as in laser weapons, laser fusion, and laser isotope separation. The perceived need for lasers for isotope separation, laser-induced chemistry, and blue-green laser-based communications from satellites, drove research in high pulse rate technology and reliability extension for discharge lasers. This research resulted in prototypes, with typical performance ranges as noted on the positioning map that well exceed current market requirements. However, the technology and underlying media physics developed in these early years made possible advances needed to serve the ultimate real market applications.

The very important application of UV lithography requires high pulse rates and high reliability. These two features differentiate the excimer laser used for lithography from the one used in the research laboratory. High pulse rates, over 2000 Hz in current practice, and the cost of stopping a computer chip production line for servicing, drive the reliability requirements toward 10¹⁰ shots per service interval. In contrast, the typical laboratory experiments are more often in the range of 100 Hz and, unlike a lithography production line, do not run 24 hours a day, 7 days a week. The other large market currently is in laser correction of myopia and other imperfections in the human eye. For this use the lasers are more akin to the commercial R&D style laser in terms of pulse rate and single pulse energy. Not that many shots are needed to ablate tissue to make the needed correction, and shot life is a straightforward requirement. However, the integration of these lasers into a certified and accepted medical procedure and an overall medical instrument drove the growth of this market.

The excimer concept, and the core technology used to excite these lasers, is applicable over a broad range of UV wavelengths, with utility at specific wavelength ranges from 351 nm in the near UV to 157 nm in the vacuum UV (VUV). There were many potential excimer candidates in the early R&D phase ([Table 1](#)). Some of these candidates can be highly efficient, >50% relative to deposited energy, at converting electrical power into UV or visible emission, and have found a parallel utility as sources for lamps, though with differing power deposition rates and configurations. The key point in the table is that these lasers are powerful and useful sources at very short wavelengths. For lithography, the wavelength of interest has consistently shifted to shorter and shorter wavelengths as the printed feature size decreased. Though numerous candidates were pursued, as shown in [Table 1](#), the most useful

Table 1 Key excimer wavelengths

Excimer emitter	λ (nm)	Comment
Ar ₂	126	Very short emission wavelength, deep in the Vacuum UV; inefficient as laser due to absorption, see Xe ₂ .
Kr ₂	146	Broad band emitter and early excimer laser with low laser efficiency. Very efficient converter of electric power into Vacuum UV light.
F ₂	157	Molecule itself is not an excimer, but uses lower level dissociation; candidate for next generation lithography.
Xe ₂	172	The first demonstrated excimer laser. Very efficient formation and emission but excited state absorption limits laser efficiency. Highly efficient as a lamp.
ArF	193	Workhorse for corneal surgery and lithography.
KrCl	222	Too weak compared to KrF and not as short a wavelength as ArF.
KrF	248	Best intrinsic laser efficiency; numerous early applications but found major market in lithography. Significant use in other materials processing.
XeI	254	Never a laser, but high formation efficiency and excellent for a lamp. Historically the first rare gas halide excimer whose emission was studied at high pressure.
XeBr	282	First rare gas halide to be shown as a laser; inefficient laser due to excited state absorption, but excellent fluorescent emitter; a choice for lamps.
Br ₂	292	Another halogen molecule with excimer like transitions that has lased but had no practical use.
XeCl	308	Optimum medium for laser discharge excitation.
Hg ₂	335	Hg vapor is very efficiently excited in discharges and forms excimers at high pressure. Subject of much early research, but as in the case of the rare gas excimers such as Xe ₂ suffer from excited state absorption. Despite the high formation efficiency, this excimer was never made to lase.
I ₂	342	Iodine UV molecular emission and lasing was first of the pure halogen excimer-like lasers demonstrated. This species served as kinetic prototype for the F ₂ 'honorary' excimer that has important use as a practical VUV source.
XeF	351, 353	This excimer laser transition terminates on a weakly bound lower level that dissociates rapidly, especially when heated. The focus for early defense-related laser development; as this wavelength propagates best of this class in the atmosphere.
XeF	480	This very broadband emission of XeF terminates on a different lower level than the UV band. Not as easily excited in a practical system, so has not had significant usage.
HgBr	502	A very efficient green laser that has many of the same kinetic features of the rare gas halide excimer lasers. But the need for moderately high temperature for the needed vapor pressure made this laser less than attractive for UV operation.
XeO	540	This is an excimer-like molecular transition on the side of the auroral lines of the O atom. The optical cross-section is quite low and as a result the laser never matured in practice.

lasers are those using rare gas halide molecules. These lasers are augmented by dye laser and Raman shifting to provide a wealth of useful wavelengths in the visible and UV.

All excimer lasers utilize molecular dissociation to remove the lower laser level population. This lower level dissociation is typically from an unbound or very weakly bound ground-state molecule. However, halogen molecules also provide laser action on excimer-like transitions, that terminate on a molecular excited state that dissociates quickly. The vacuum UV transition in F₂ is the most practical method for making very short wavelength laser light at significant power. Historically, the rare gas dimer molecules, such as Xe₂, were the first to show excimer laser action, although self absorption, low gain, and poor optics in the VUV limited their utility. Other excimer-like species, such as metal halides, metal rare gas continua on the edge of metal atom resonance lines, and rare gas oxides were also studied. The key differentiators of rare gas halides from other excimer-like candidates are strong binding in the excited state, lack of self absorption, and relatively high optical cross-section for the laser transition.

Excimer Laser Fundamentals

Excimer lasers utilize lower-level dissociation to create and sustain a population inversion. For example, in the first excimer laser demonstrated, Xe₂, the lasing species is one which is comprised of two atoms that do not form a stable molecule in the ground state, as xenon dimers in the ground state are not a stable molecule. In the lowest energy state, that with neither of the atoms electronically excited, the interaction between the atoms in a collision is primarily one of repulsion, save for a very weak 'van der Waals' attraction at larger internuclear separations. This is shown in the potential energy diagram of Fig. 3. If one of the atoms is excited, for example by an electric discharge, there can be a chemical binding in the electronically excited state of the molecule relative to the energy of one atom with an electron excited and one atom in the ground state. We use the shorthand notation * to denote an electronically excited atomic molecular species, e.g., Xe*. The excited dimer, excimer for short, will recombine rapidly at high pressure from atoms into the Xe₂* electronically excited molecule. Radiative lifetimes of the order 10 nsec to 1 μ sec are typical

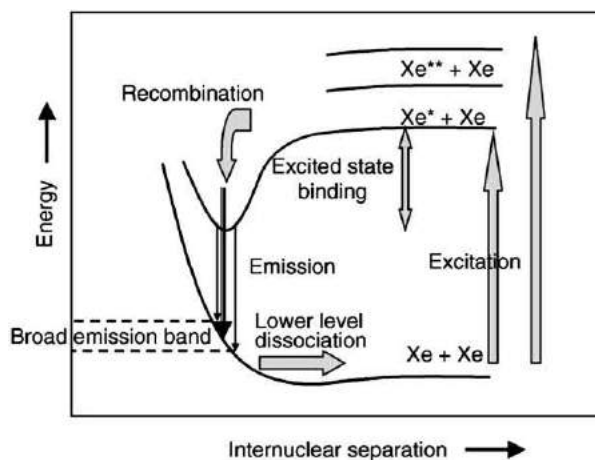


Fig. 3 A schematic of the potential energy curves for an excimer species such as Xe_2 . The excited molecule is bound relative to an excited atom and can radiate to a lower level that rapidly dissociates on psec timescales since the ground state is not bound, save for weak binding due to van der Waals forces. The emission band is intrinsically broad.

before photon emission returns the molecule to the lower level. Collisional quenching often reduces the lifetime below the radiative rate. For Xe_2^* excimers, the emission is in the vacuum ultraviolet (VUV).

There are a number of variations on this theme, as noted in Table 1. Species, other than rare gas pairs, can exhibit broadband emission. The excited state binding energy can vary significantly, changing the fraction of excited states that will form molecules. The shape of the lower potential energy curve can vary as well. From a semantic point of view heteronuclear, diatomic molecules, such as XeO , are not excimers, as they are not made of two of the same atoms. However, the more formal name ‘exciplex’ did not stick in the laser world; excimer laser being preferred. Large binding energy is more efficient for capturing excited atoms into excited excimer-type molecules in the limited time they have before emission. Early measurements of the efficiency of converting electrical energy deposited into fluorescent radiation showed that up to 50% of the energy input could result in VUV light emission. Lasers were not this efficient due to absorption.

The diagram shown in Fig. 3 is simplified because it does not show all of the excited states that exist for the excited molecule. There can be closely lying excited states that share the population and radiate at a lower rate, or perhaps absorb to higher levels, also not shown in Fig. 3. Such excited state absorption may terminate on states derived from higher atomic excited states, noted as Xe^{**} in the figure, or yield photo-ionization, producing the diatomic ion-molecule plus an electron, such as:



Such excited state absorption limited the efficiency for the VUV rare gas excimers, even though they have very high formation and fluorescence efficiency.

Rare gas halide excimer lasers evolved from ‘chemi-luminescence’ chemical kinetics studies which were looking at the reactive quenching of rare gas metastable excited states in flowing after glow experiments. Broadband emission in reactions with halogen molecules was observed in these experiments, via reactions such as:



At low pressure the emissions from molecules, such as XeCl^* , are broad in the emissions bandwidth. Shortly after these initial observations, the laser community began examination of these species in high-pressure, electron-beam excited mixtures. The results differed remarkably from the low-pressure experiments and those for the earlier excimers such as Xe_2^* . The spectrum was shifted well away from the VUV emission. Moreover, the emission was much sharper at high pressure, though still a continuum with a bandwidth of the order of 4 nm. A basis for understanding is sketched in the potential energy curves of Fig. 4. In this schematic we identify the rare gas atoms by Rg, excited rare gas atoms by Rg^* , and halogen atoms by X. Ions play a very important role in the binding and reaction chemistry, leading to excited laser molecules.

The large shift in wavelength from the VUV of rare gas dimer excimers, ~ 308 nm in XeCl^* versus 172 nm in Xe_2^* , is due to the fact that the binding in the excited state of the rare gas halide is considerably stronger. The sharp and intense continuum at high pressure ($> \sim 100$ torr total pressure) is due to the fact that the ‘sharp’ transition terminates on a ‘flat’ portion of the lower-level potential energy curve. Indeed, in XeCl and XeF , the sharp band laser transition terminates on a slightly bound portion of the lower-level potential energy curve. Both the sharp emissions and broad bands are observed, due to the fact that some transitions from the excited levels terminate on a second, more repulsive potential energy curve.

An understanding of the spectroscopy and kinetic processes in rare gas plus halogen mixtures is based on the recognition that the excited state of a rare gas halide molecule is very similar to the ground state of the virtually isoelectronic alkali halide. The binding energy and mechanism of the excited state is very close to that of the ground state of the most similar alkali halide. Reactions of excited state rare gases are very similar to those of alkali atoms. For example, Xe^* and Cs are remarkably similar, from

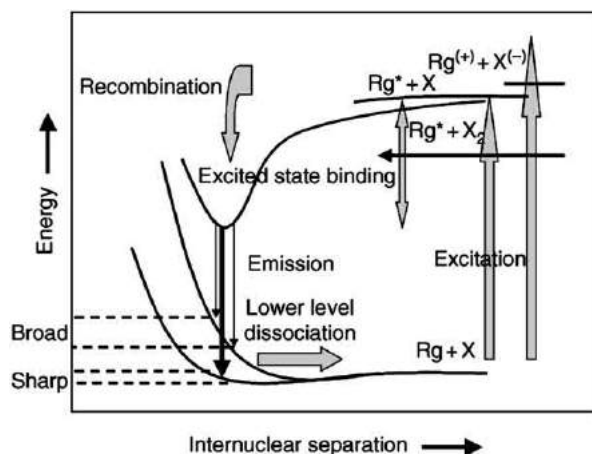


Fig. 4 The potential energy curves for rare gas halides are sketched here. Both relatively sharp continuum emission is observed along with broad bands corresponding to different lower levels. The excited ion pair states consist of 3 distinct levels, two that are very close in energy and one that is shifted up from the upper laser level (the B state) by the spin orbit splitting of the rare gas ion.

a chemistry and molecular physics point of view, as they have the same outer shell electron configuration and similar ionization potentials. The ionization potential of Xe^* is similar to that of cesium (Cs). Cs, and all the other alkali atoms, react rapidly with halogen molecules. Xenon metastable excited atoms react rapidly with halogen molecules. The alkali atoms all form ionically bonded ground states with halogens. The rare gas halides are effectively strongly bonded ion pairs, $\text{Xe}^{(+)}\text{X}^{(-)}$, that are very strongly bonded relative to the excited states that yield them. The difference between, for example, XeCl^* and CsCl , is that XeCl^* radiates in a few nanoseconds while CsCl is a stable ground-state molecule. The binding energy for these ionic bonded molecules is much greater than the more covalent type of bonding that is found in the rare gas excimer excited states. Thus XeI^* , which is only one electron different (in the I) than Xe_2^* , emits at 254 nm instead of 172 nm.

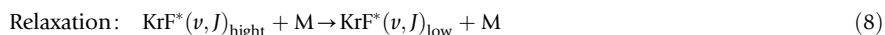
The first observed and most obvious connection to alkali atoms is in formation chemistry. Mostly, rare gas excited states react with halogen molecules rapidly. The rare gas halide laser analog reaction sequence is shown below, where some form of electric discharge excitation kicks the rare gas atom up to an excited state, so that it can react like an alkali atom, [Eq. \(3\)](#):



There are subtle and important cases where the neutral reaction, such as that shown in [Eqs. \(2\) or \(4\)](#), is not relevant as it is too slow compared to other processes. Note that the initial reaction does not yield the specific upper levels for the laser but states that are much higher up in the potential well of the excited rare gas halide excimer molecule. Further collisions with a buffer gas are needed to relax the states to the vibrational levels that are the upper levels for the laser transition. Such relaxation at high pressure can be fast, which leads the high-pressure spectrum to be much sharper than those first observed in flowing after-glow experiments.

The excited states of the rare gas halides are ion pairs bound together for their brief radiative lifetime. This opens up a totally different formation channel, indeed the one that often dominates the mechanism, ion-ion recombination. Long before excimer lasers were studied, it was well known that halogen molecules would react with 'cold' electrons (those with an effective temperature less than a few eV) in a process called attachment. The sequence leading to the upper laser level is shown below for the Kr/ F_2 system, but the process is generically identical in other rare gas halide mixtures.

KrF* formation kinetics via ion channel:



Finally, there is one other formation mechanism, displacement, that is unique to the rare gas halides. In this process, a heavier rare gas will displace a lighter rare gas ion in an excited state to form a lower-energy excited state:



In general, the most important excimer lasers tend to have all of these channels contributing to excited state formation for optimized mixtures.

These kinetic formation sequences express some important points regarding rare gas halide lasers. The source of the halogen atoms for the upper laser level disappears via both attachment reactions with electrons and by the reaction with excited states. The halogen atoms eventually recombine to make molecules, but at a rate too slow for sufficient continuous wave (cw) laser gain. For excimer-based lamps, however, laser gain is not a criterion and very efficient excimer lamps can be made. Another point to note is that forming the inversion requires transient species (and in some cases halogen fuels) that absorb the laser light. Net gain and extraction efficiency are a trade-off between pumping rate, an increase in which often forms more absorbers and absorption itself. **Table 2** provides a listing of some of the key absorbers. Finally, the rare gas halide excited states can be rapidly quenched in a variety of processes. More often than not, the halogen molecule and the electrons in the discharge react with the excited states of the rare gas halide, removing the contributors to gain. We show, in **Table 3**, a sampling of some of the types of formation and quenching reactions with their kinetic rate constant parameters. Note that one category of reaction, three-body quenching, is unique to excimer lasers. Note also that the three-body recombination of ions has a rate coefficient that is effectively pressure

Table 2 Absorbers in rare gas halide lasers

Species	XeF	XeCl	KrF	ArF
Laser λ (nm)	351, 353	308	248	193
F ₂ absorption σ (cm ²)	8×10^{-21}	NA	1.5×10^{-20}	
Cl ₂ absorption σ (cm ²)	NA	1.7×10^{-19}	NA	NA
HCl absorption σ (cm ²)	NA	Nil	NA	NA
F ⁽⁻⁾ absorption σ (cm ²)	2×10^{-18}	NA	5×10^{-18}	5×10^{-18}
Cl ⁽⁻⁾ absorption σ (cm ²)	NA	2×10^{-17}	NA	NA
Rg ₂ ⁽⁺⁾ diatomic ion absorption σ (cm ²) ^a	Ne: $\sim 10^{-18}$ Ar: $\sim 3 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-17}$	Ne: $\sim 10^{-17}$ Ar: $\sim 4 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-18}$	Ne: $\sim 2.5 \times 10^{-17}$ Ar: $\sim 2 \times 10^{-18}$ Kr $< 10^{-18}$	Ne: $\sim 10^{-18}$ Ar: NIL
Other transient absorbers	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$ – 10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$ – 10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$ – 10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$ – 10^{-17}
Other	Lower laser level absorbs unless heated	Weak lower level absorption $\sigma \sim 4 \times 10^{-16}$	Windows and dust can be an issue	Windows and dust can be an issue

^aNote that the triatomic rare gas halides of the form Rg₂X will exhibit similar absorption as the diatomic rare gas ions as the triatomic rare gas halides of this form are essentially ion pairs of an Rg₂⁽⁺⁾ ion and the halide ion.

Table 3 Typical formation and quenching reactions and rate coefficients

Formation	
Rare gas plus halogen molecule	$\text{Kr}^* + \text{F}_2 \rightarrow \text{KrF}^*(v, J) + \text{F}$ $k \sim 7 \times 10^{-10}$ cm ³ /sec; Branching ratio to KrF* ~ 1 $\text{Ar}^* + \text{F}_2 \rightarrow \text{ArF}^*(v, J) + \text{F}$ $k \sim 6 \times 10^{-10}$ cm ³ /sec; Branching ratio to ArF* $\sim 60\%$ $\text{Xe}^* + \text{F}_2 \rightarrow \text{XeF}^*(v, J) + \text{F}$ $k \sim 7 \times 10^{-10}$ cm ³ /sec; Branching ratio to XeF* ~ 1 $\text{Xe}^* + \text{HCl}(v=0) \rightarrow \text{Xe} + \text{H} + \text{Cl}$ $k \sim 6 \times 10^{-10}$ cm ³ /sec; Branching ratio to XeCl* ~ 0 $\text{Xe}^* + \text{HCl}(v=1) \rightarrow \text{XeCl}^*(v, J) + \text{H}$ $k \sim 2 \times 10^{-10}$ cm ³ /sec; Branching ratio to XeCl* ~ 1
Ion-ion recombination	$\text{Ar}^{(+)} + \text{F}^{(-)} + \text{Ar} \rightarrow \text{ArF}^*(v, J)$ $k \sim 1 \times 10^{-25}$ cm ⁶ /sec at 1 atm and below $k \sim 7.5 \times 10^{-26}$ cm ⁶ /sec at 2 atm; Note effective 3-body rate constant decreases further as pressure goes over ~ 2 atm $\text{Kr}^{(+)} + \text{F}^{(-)} + \text{Kr} \rightarrow \text{KrF}^*(v, J)$ $k \sim 7 \times 10^{-26}$ cm ⁶ /sec at 1 atm and below; rolls over at higher pressure $\text{ArF}^*(v, J) + \text{Kr} \rightarrow \text{KrF}^*(v, J) + \text{Ar}$ $k \sim 2 \times 10^{-10}$ cm ³ /sec; Branching ratio to KrF* ~ 1
Quenching	
Halogen molecule	$\text{XeCl}^* + \text{HCl} \rightarrow \text{Xe} + \text{Cl} + \text{HCl}$ $k \sim 5.6 \times 10^{-10}$ cm ³ /sec; $\text{KrF}^* + \text{F}_2 \rightarrow \text{Kr} + \text{F} + \text{F}$ $k \sim 6 \times 10^{-10}$ cm ³ /sec; All halogen molecule quenching have very high reaction rate constants
3-body inert gas	$\text{ArF}^* + \text{Ar} + \text{Ar} \rightarrow \text{Ar}_2\text{F}^* + \text{Ar}$ $k \sim 4 \times 10^{-31}$ cm ⁶ /sec $\text{KrF}^* + \text{Kr} + \text{Ar} \rightarrow \text{Kr}_2\text{F}^* + \text{Ar}$ $k \sim 6 \times 10^{-31}$ cm ⁶ /sec $\text{XeCl}^* + \text{Xe} + \text{Ne} \rightarrow \text{Xe}_2\text{Cl}^* + \text{Ne}$ $k \sim 1 \times 10^{-33}$ cm ⁶ /sec; 4×10^{-31} cm ⁶ /sec with Xe as 3rd body These reactions yield a triatomic excimer at lower energy that can not be recycled back to the desired laser states
Electrons	$\text{XeCl}^* + e \rightarrow \text{Xe} + \text{Cl} + e$ $k \sim 2 \times 10^{-7}$ cm ³ /sec In 2-body quenching by electrons, the charge of the electron interacts with the dipole of the rare gas halide ion pair at long range; above value is typical of others
2-body inert gas	$\text{KrF}^* + \text{Ar} \rightarrow \text{Ar} + \text{Kr} + \text{F}$ $k \sim 2 \times 10^{-12}$ cm ³ /sec $\text{XeCl}^* + \text{Ne} \rightarrow \text{Ne} + \text{Xe} + \text{Cl}$ $k \sim 3 \times 10^{-13}$ cm ³ /sec 2-body quenching by rare gases is typically slow and less important than the reactions noted above

dependent. At pressures above ~ 3 atm, the diffusion of ions toward each other is slowed and the effective recombination rate constant decreases.

Though the rare gas halide excited states can be made with near-unity quantum yield from the primary excitation, the intrinsic laser efficiency is not this high. The quantum efficiency is only $\sim 25\%$ (recall rare gas halides all emit at much longer wavelengths than rare gas excimers); there is inefficient extraction due to losses in the gas and windows. In practice there are also losses in coupling the power into the laser, and wasted excitation during the finite start-up time. The effect of losses is shown in Table 4 and Fig. 5 where the pathways are charted and losses identified. In KrF, it takes ~ 20 eV to make a rare gas ion in an electron beam excited mixture, somewhat less in an electric discharge. The photon release is ~ 5 eV; the quantum efficiency is only 25%. Early laser efficiency measurements (using electron beam excitation) showed laser efficiency in the best cases slightly in excess of 10%, relative to the deposited energy. The discrepancy relative to the quantum efficiency is due to losses in the medium. For the sample estimate in Table 4 and Fig. 5 we show approximate density of key species in the absorption chain, estimated on a steady-state approximation of the losses in a KrF laser for an excitation rate of order 1 MW/cc, typical of fast discharge lasers. The net small signal gain is of the order of 5%/cm. The key absorbers are the parent halogen molecule, F_2 , the fluoride ion $F^{(-)}$, and the rare gas excited state atoms. A detailed pulsed kinetics code will provide slightly different results due to transient kinetic effects and the intimate coupling of laser extraction to quenching and absorber dynamics. The ratio of gain to loss is usually in the range of 7 to 20, depending on the actual mixture used and the rare gas halide wavelength. The corresponding extraction efficiency is then limited to the range of $\sim 50\%$. The small signal gain, g_o , is related to the power deposition rate per unit volume, $P_{\text{deposited}}$, by Eq. (10).

Table 4 Typical ground and excited state densities shown for KrF at $P_{\text{in}} \sim 1 \text{ MW/cm}^3$

Species	Amount particles/cm ³	Absorption σ (cm) ²	Loss %/cm
F_2	$\sim 1.2 \times 10^{17}$	1.5×10^{-20}	0.18
Kr	$\sim 3 \times 10^{18}$	—	—
Buffer	4×10^{19}	—	—
$F^{(-)}$	$\sim 10^{14}$	5×10^{-18}	0.05
Rg^*, Rg^{**}	1×10^{15}	$\langle 10^{-18} \rangle$ depends on ratio of Rg^{**}/Rg^*	0.1
KrF (before dissociating)	$\sim 10^{12}$	2×10^{-16}	0.02
Rg_2F^*	8×10^{14} , Note this species density decreases as flux grows	1×10^{-18}	0.08
KrF* (B)	2.5×10^{14}	2×10^{-16}	5 (gain)

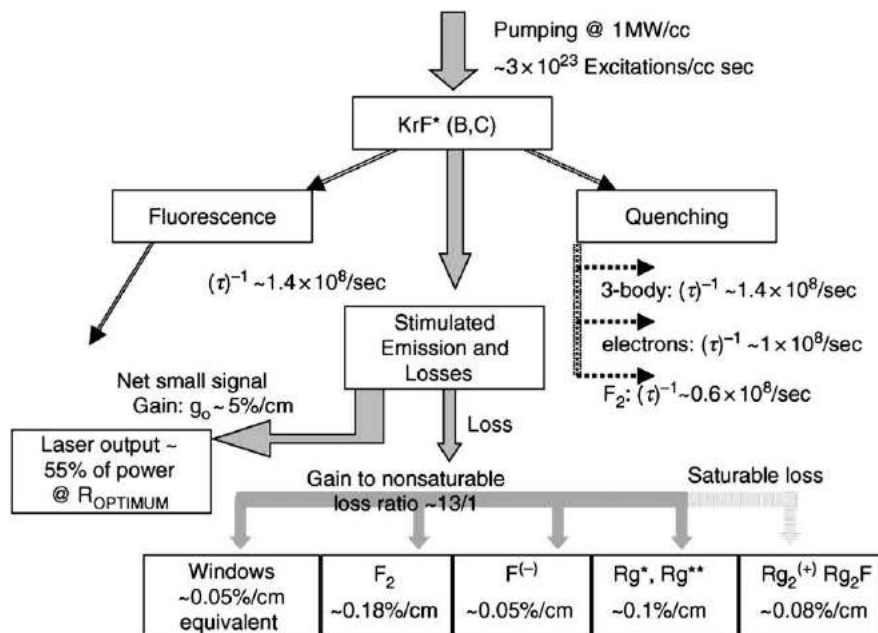


Fig. 5 The major decay paths and absorbers that impact the extraction of excited states of rare gas halide excimers. Formation is via the ion-ion recombination, excited atom reaction with halogens and displacement reactions, not shown. We note for a specific mixture of KrF the approximate values of the absorption by the transient and other species along with the decay rates for the major quenching processes. Extraction efficiency greater than 50% is quite rare.

$$P_{\text{deposited}} = g_0 E^* \left(\tau_{\text{upper}} \eta_{\text{pumping}} \eta_{\text{branching}} f \sigma \right)^{-1} \quad (10)$$

E^* is the energy per excitation, ~ 20 eV or 3.2×10^{-18} J/excitation, f is the fraction of excited molecules in the upper laser level, $\sim 40\%$ due to KrF* in higher vibrational states and the slowly radiating 'C' state. The laser cross-section is σ , and the η terms are efficiencies for getting from the primary excitation down to the upper laser level. The upper state lifetime, τ_{upper} , is the sum of the inverses of radiative and quenching lifetimes. When the laser cavity flux is sufficiently high, stimulated emission decreases the flow of power into the quenching processes.

The example shown here gives a ratio of net gain to the nonsaturating component of loss of 13/1. Of the excitations that become KrF*, only $\sim 55\%$ or so will be extracted in the best of cases. For other less ideal rare gas halide lasers, the gain to loss ratio, the extraction efficiency, and the intrinsic efficiency are all lower. Practical discharge lasers have lower practical wall plug efficiency due to a variety of factors, discussed below.

Discharge Technology

Practical discharge lasers use fast-pulse discharge excitation. In this scheme, an outgrowth of the CO₂ TEA laser, the gas is subjected to a rapid high-voltage pulse from a low-inductance pulse forming line or network, PFL or PFN. The rapid pulse serves to break down the gas, rendering it conductive with an electron density of $> \sim 10^{14}$ electrons/cm³. As the gas becomes conductive, the voltage drops and power is coupled into the medium for a short duration, typically less than 50 nsec. The discharge laser also requires some form of 'pre-ionization', usually provided by sparks or a corona-style discharge on the side of the discharge electrodes. The overall discharge process is intrinsically unstable, and pump pulse duration needs to be limited to avoid the discharge becoming an arc.

Moreover, problems with impedance matching of the PFN or PFL with the conductive gas lead to both inefficiency and extra energy that can go into post-discharge arcs, which both cause component failure and produce metal dust which can give rise to window losses. A typical discharge laser has an aperture of a few cm² and a gain length of order 80 cm and produces outputs in the range of 100 mJ to 1 J. The optical pulse duration tends to be ~ 20 nsec, though the electrical pulse is longer due to the finite time needed for the gain to build up and the stimulated emission process to turn spontaneously emitted photons into photons in the laser cavity.

The pre-ionizer provides an initial electron density on the order of 10^8 electrons/cm³. With spark pre-ionization or corona pre-ionization, it is difficult to make a uniform discharge over an aperture greater than a few cm². For larger apertures, one uses X-rays for pre-ionization. X-ray pre-ionized discharge excimer lasers have been scaled to energy well over 10 J per pulse and can yield pulse durations over 100 nsec. The complexity and expense of the X-ray approach, along with the lack of market for lasers in this size range, resulted in the X-ray approach remaining in the laboratory.

Aside from the kinetic issues of the finite time needed for excitation to channel into the excimer upper levels and the time needed for the laser to 'start up' from fluorescent photons, these lasers have reduced efficiency due to poor matching of the pump source to the discharge. Ideally, one would like to have a discharge such that when it is conductive the voltage needed to sustain the discharge is less than half of that needed to break the gas down. Regrettably for the rare gas halides (and in marked distinction to the CO₂ TEA laser) the voltage that the discharge sustains in a quasi-stable mode is $\sim 20\%$ of the breakdown voltage, perhaps even less, depending on the specific gases and electrode shapes.

Novel circuits, which separate the breakdown phase from the fully conductive phase, can readily double the efficiency of an excimer discharge laser, though this approach is not necessary in many applications. The extra energy that is not dissipated in the useful laser discharge often results in arcing or 'hot' discharge areas after the main laser pulse, leading to electrode sputtering, dust that finds its way to windows, and the need to service the laser after $\sim 10^8$ pulses. When one considers the efficiency of the power conditioning system, along with the finite kinetics of the laser start-up process in short pulses, and the mismatch of the laser impedance to the PFN/PFL impedance, lasers that may have 10% intrinsic efficiency in the medium in e-beam excitation may only have 2% to 3% wall plug efficiency in a practical discharge laser.

A key aspect of excimer technology is the need for fast electrical circuits to drive the power into the discharge. This puts a significant strain on the switch that is used to start the process. While spark gaps, or multichannel spark gaps, can handle the required current and current rate of rise, they do not present a practical alternative for high-pulse rate operation. Thyratrons and other switching systems, such as solid state switches, are more desirable than a spark gap, but do not offer the desired needed current rate of rise. To provide the pumping rate of order 1 MW/cm³ needed for enough gain to turn the laser on before the discharge becomes unstable, the current needs to rise to a value of the order of 20 kA in a time of ~ 20 nsec; the current rate of rise is $\sim 10^{12}$ A/sec. This rate of rise will limit the lifetime of the switch that drives the power into the laser gas. As a response to this need, magnetic switching and magnetic compression circuits were developed so that the lifetime of the switch can be radically increased. In the simplest cases, one stage of magnetic compression is used to make the current rate of rise to be within the range of a conventional thyatron switch as used in radar systems. Improved thyratrons and a simple magnetic assist also work. For truly long operating lifetimes, multiple stages of magnetic switching and compression are used, and with a transformer one can use a solid state diode as the start switch for the discharge circuit. A schematic of an excimer pulsed power circuit, that uses a magnetic assist with a thyatron and one stage of magnetic compression, is shown in Fig. 6. The magnetic switching approach can also be used to provide two separate circuits to gain high efficiency by having one circuit to break down the gas (the spiker) and a second,

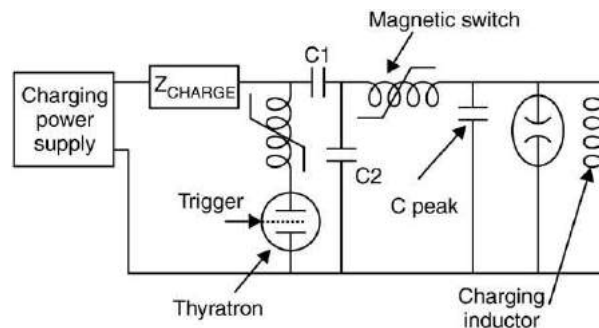


Fig. 6 One of the variants on magnetic switching circuits that have been used to enhance the lifetime of the primary start switch. A magnetic switch in the form of a simple 1 turn inductor with a saturable core magnetic material such as a ferrite allows the thyatron to turn on before major current flows. Other magnetic switches may be used in addition to peak up the voltage rate of rise on the laser head or to provide significant compression. For example one may have the major current flow in the thyatron taking place on the 500 nsec time scale while the voltage pulse on the discharge is in the 50 nsec duration. A small current leaks through and is accommodated by the charging inductor.

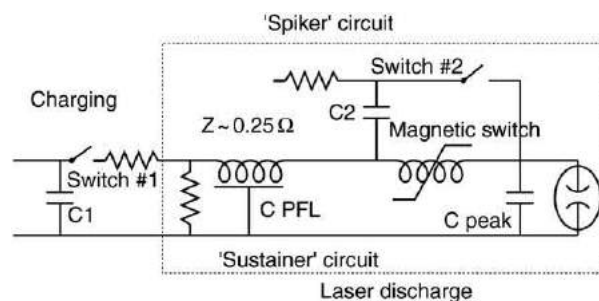


Fig. 7 For the ultimate in efficiency the circuit can be arranged to provide a leading edge spike that breaks the gas down using a high impedance 'spiker' and a separate circuit that is matched to the discharge impedance when it is fully conductive. High coupling efficiency reduces waste energy that can erode circuit and laser head components via late pulse arcs. The figure shows a magnetic switch as the isolation but the scheme can be implemented with a second laser head as the isolation switch, a rail gap (not appropriate for long life) or a diode at low voltage.

low-impedance 'sustainer' circuit, to provide the main power flow. In this approach, the laser itself can act as part of the circuit by holding off the sustaining voltage for a μ -sec or so during the charge cycle and before the spiker breaks down the gas. A schematic of such a circuit is shown in Fig. 7. Using this type of circuit, the efficiency of a XeCl laser can be over 4% relative to the wall plug. A wide variety of technological twists on this approach have been researched.

Excimer lasers present some clear differences in resonator design compared to other lasers. The apertures are relatively large, so one does not build a TEM₀₀ style cavity for good beam quality and expect to extract any major portion of the energy available. The typical output is divergent and multimode. When excimer lasers are excited by a short pulse discharge, ~ 30 nsec, the resulting gain duration is also short, ~ 20 nsec, limiting the number of passes a photon in a resonator can have in the gain. The gain duration of ~ 20 nsec only provides ~ 3.5 round trips of a photon in the cavity during the laser lifetime, resulting in the typical high-order multimode beam. Even with an unstable resonator, the gain duration is not long enough to collapse the beam into the lowest-order diffraction-limited output. If line narrowing is needed, and reasonable efficiency is required, an oscillator amplifier configuration is often used with the oscillator having an appropriate tuning resonator in the cavity. If both narrow spectral linewidth and excellent beam quality are needed one uses a seeded oscillator approach where the narrowband oscillator is used to seed an unstable resonator on the main amplifier. By injecting the seed into the unstable resonator cavity, a few nsec before the gain is turned on, the output of the seeded resonator can be locked in wavelength, bandwidth, and provide good beam quality. Long pulse excimer lasers, such as e-beam excited lasers or long pulse X-ray pre-ionized discharge devices can use conventional unstable resonator technology with good results.

Excimer lasers are gas lasers that run at both high pressure and high instantaneous power deposition rates. During a discharge pulse, much of the halogen bearing 'fuel' is burned out by the attachment and reactive processes. The large energy deposition per pulse means that pressure waves are generated. All of these combine with the characteristic device dimensions to require gas flow to recharge the laser mixture between pulses. For low pulse rates, less than ~ 200 Hz, the flow system need not be very sophisticated. One simply needs to flush the gas from the discharge region with a flow having a velocity of the order of 5 m/sec for a typical discharge device. The flow is transverse to the optical and discharge directions. At high pulse rates, the laser designer needs to consider flow in much more detail to minimize the power required to push the gas through the laser head. Circulating pressure waves need to be damped out so that the density in the discharge region is controlled at the time of the next pulse. Devices called acoustic dampers are placed in the sides of the flow loop to remove pressure pulses. A final subtlety occurs in providing gas flow

over the windows. The discharge, post pulse arcs, and reactions of the halogen source with the metal walls and impurities creates dust. When the dust coats the windows, the losses in the cavity increase, lowering efficiency. By providing a simple gas flow over the window, the dust problem can be ameliorated. For truly long life, high-reliability lasers, one needs to take special care in the selection of materials for electrodes and insulators and avoid contamination, by assembling in clean environments.

Further Reading

- Ballanti, S., Di Lazzaro, P., Flora, F., *et al.*, 1998. Ianus the 3-electrode laser. *Applied Physics B* 66, 401–406.
- Brau, C.A., 1978. Rare gas halogen lasers. In: Rhodes, C.K. (Ed.), *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer Verlag.
- Brau, C.A., Ewing, J.J., 1975. Emission spectra of XeBr, XeCl, XeF, and KrF. *Journal of Chemical Physics* 63, 4640–4647.
- Ewing, J.J., 2000. Excimer laser technology development. *IEEE Journal of Selected Topics in Quantum Electronics* 6 (6), 1061–1071.
- Levatter, J., Lin, S.C., 1980. Necessary conditions for the homogeneous formation of pulsed avalanche discharges at high pressure. *Journal of Applied Physics* 51, 210–222.
- Long, W.H., Plummer, M.J., Stappaerts, E., 1983. Efficient discharge pumping of an XeCl laser using a high voltage prepulse. *Applied Physics Letters* 43, 735–737.
- Rhodes, C.K. (Ed.), 1979. *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer-Verlag.
- Rokni, M., Mangano, J., Jacob, J., Hsia, J., 1978. Rare gas fluoride lasers. *IEEE Journal of Quantum Electronics* QE-14, 464–481.
- Smilanski, I., Byron, S., Burkes, T., 1982. Electrical excitation of an XeCl laser using magnetic pulse compression. *Applied Physics Letters* 40, 547–548.
- Taylor, R.S., Leopold, K.E., 1994. Magnetic-spiker excitation of gas discharge lasers. *Applied Physics B, Lasers and Optics* 59, 479–509.

Metal Vapor Lasers

DW Coutts, University of Oxford, Oxford, UK

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

k	Boltzmann constant [eV K^{-1}]
M^{+*}	Excited metal ion
N^*	Excited noble gas atom
ΔE	Energy difference [eV]
e^-	Electron
Gas flow	[mbar l min^{-1}]
R_1, R_2	Mirror curvatures [m]
M	Metal atom
N	Noble gas atom

N^+	Noble gas ion
Pulse duration	[ns]
Pulse repetition frequency	[kHz]
Q	Quality factor
T	Temperature [K]
D	Tube diameter [mm]
L	Tube length [m]
Wavelength	[nm], [μm]
M	Unstable resonator magnification

Introduction

Metal vapor lasers form a class of laser in which the active medium is a neutral or ionized metal vapor usually excited by an electric discharge. These lasers fall into two main subclasses, namely cyclic pulsed metal vapor lasers and continuous-wave metal ion lasers. Both types will be considered in this article, including basic design and construction, power supplies, operating characteristics (including principal wavelengths), and brief reference to their applications.

Self-Terminating Resonance-Metastable Pulsed Metal Vapor Lasers

The active medium in a self-terminating pulsed metal vapor laser consists of metal atoms or ions in the vapor phase usually as a minority species in an inert buffer gas such as neon or helium. Laser action occurs between a resonance upper laser level and a metastable lower laser level (Fig. 1). During a fast pulsed electric discharge (typically with a pulse duration of order 100 ns) the upper laser level is preferentially excited by electron impact excitation because it is strongly optically connected to the ground state (resonance transition) and hence has a large excitation cross-section. For a sufficiently large metal atom (or ion) density, the resonance radiation becomes optically trapped, thus greatly extending the lifetime of the upper laser level such that decay from the upper laser level is channelled through the emission of laser radiation to the metastable lower laser level. Lasing terminates when the electron temperature falls to a point such that preferential pumping to the upper laser level is no longer sustained, and the build-up of population in the metastable lower laser level destroys the population inversion. Therefore after each excitation pulse, the resulting excited species in the plasma (in particular the metastable lower laser levels which are quenched by collisions with cold electrons) must be allowed sufficient time to relax and the plasma must be allowed to partially recombine before applying the next excitation pulse. The relaxation times for self-terminating metal vapor lasers correspond to operating pulse repetition frequencies from 2 kHz to 200 kHz.

Many metal vapors can be made to lase in the resonance-metastable scheme and are listed together with their principal wavelengths and output powers in Table 1. The most important self-terminating pulsed metal vapor laser is the copper vapor laser and its variants which will be discussed in detail in the following sections. Of the other pulsed metal vapor lasers listed in Table 1,

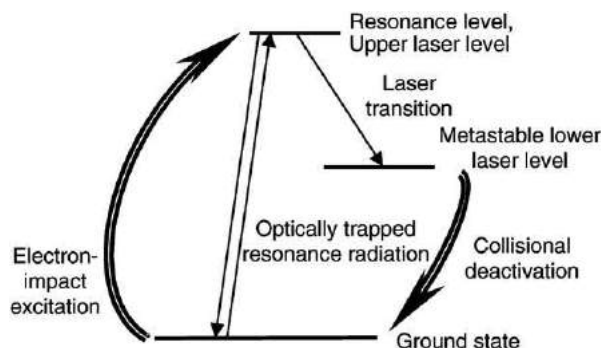
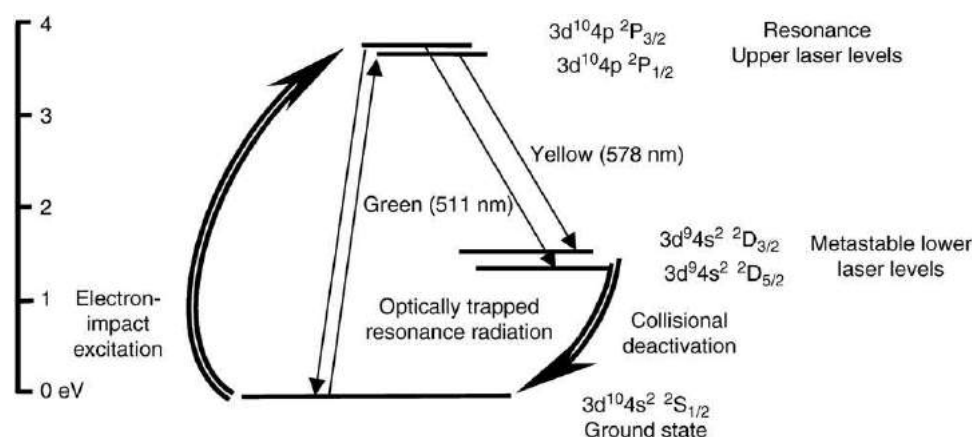


Fig. 1 Resonance-metastable energy levels for self-terminating metal vapor lasers.

Table 1 Principal self-terminating resonance-metastable metal vapor lasers

Metal	Principal wavelengths (nm)	Powers		Total efficiencies	Pulse repetition frequency (kHz)	Technological development
		Typical (W)	Maximum (W)			
Cu	510.55	2–70	2500 total	1%	4–40	Highly developed and commercially available
Au	578.2	1–50	20	0.13%	2–40	Commercially available
	627.8	1–8				
Ba	312	0.1–0.2	1.2	0.5%	1–8	Have been produced commercially, but largely experimental
	15002550	2–101	121.5		5–15	
Pb	1130	0.5	1.0	0.15%	10–30	Experimental
Mn	722.9		4.4			
	534.1 (> 50%)		12 total	0.32%	~10	Experimental
	1290					

**Fig. 2** Partial energy level scheme for the copper vapor laser.

only the gold vapor laser (principal wavelengths 627.8 nm and 312.3 nm), and the barium vapor laser which operates in the infrared (principal wavelengths 1.5 μm and 2.55 μm) have had any commercial success. All the self-terminating pulsed metal vapor lasers have essentially the same basic design and operating characteristics as exemplified by the copper vapor laser.

Copper Vapor Lasers

Copper vapor lasers (CVLs) are by far the most widespread of all the pulsed metal vapor lasers. **Fig. 2** shows the energy level scheme for copper. Lasing occurs simultaneously from the $^2P_{3/2}$ level to the $^2D_{5/2}$ level (510.55 nm) and from the $^2P_{1/2}$ level to the $^2D_{3/2}$ level (578.2 nm). Commercial devices are available with combined outputs of over 100 W at 510.55 nm and 578.2 nm (typically with a green-to-yellow power ratio of 2:1).

A typical copper vapor laser tube is shown in **Fig. 3**. High-purity copper pieces are placed at intervals along an alumina ceramic tube which typically has dimensions of 1–4 cm diameter and 1–2 m long. The alumina tube is surrounded by a solid fibrous alumina thermal insulator, and a glass or quartz vacuum envelope. Cylindrical electrodes made of copper or tantalum are located at each end of the plasma tube to provide a longitudinal discharge arrangement. The cylindrical electrodes and silica laser end windows are supported by water-cooled end pieces. The laser windows are usually tilted by a few degrees to prevent back reflections into the active medium. The laser head is contained within a water cooled metal tube to provide a coaxial current return for minimum laser head inductance. Typically a slow flow ($\sim 5 \text{ mbar l min}^{-1}$) of neon at a pressure of 20–80 mbar is used as the buffer gas with an approximately 1% H_2 additive to improve the afterglow plasma relaxation. The buffer gas provides a medium to operate the discharge when the laser is cold and slows diffusion of copper vapor out of the ends of the hot plasma tube. Typical copper fill times are of order 200–2000 hours (for 20–200 g copper load). Sealed-off units with lifetimes of order 1000 hours have been in production in Russia for many years.

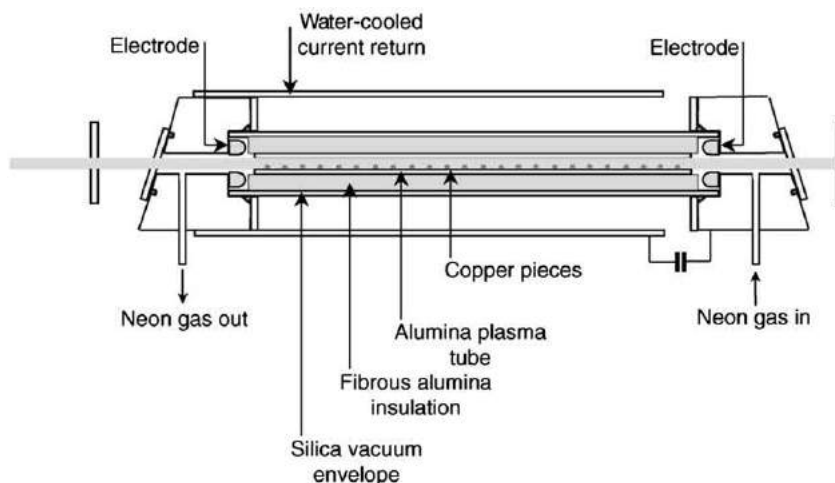


Fig. 3 Copper vapor laser tube construction.

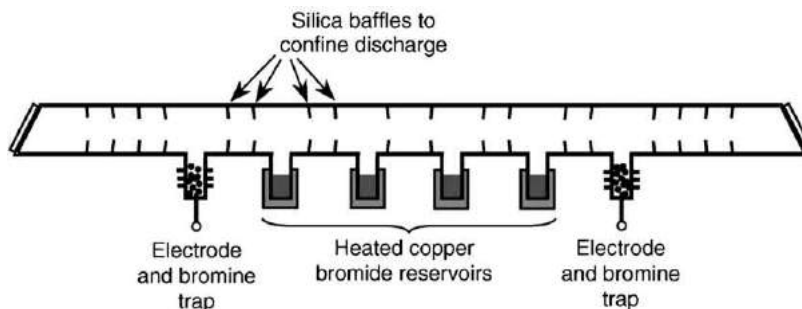


Fig. 4 Copper bromide laser tube construction.

During operation waste heat from the repetitively pulsed discharge heats the alumina tube up to approximately 1500°C at which point the vapor pressure of copper is of approximately 0.5 mbar which corresponds to the approximate density required for maximum laser output power. Typical warm-up times are therefore relatively long at around one hour to full power.

One way to circumvent the requirement for high temperatures required to produce sufficient copper density by evaporation of elemental copper (and hence also reduce warm-up times) is to use a copper salt with a low boiling point located in one or more side-arms of the laser tube (Fig. 4). Usually copper halides are used as the salt, with the copper bromide laser being the most successful. For the CuBr laser, a temperature of just 600°C is sufficient to produce the required Cu density by dissociation of CuBr vapor in the discharge. With the inclusion of 1–2% H₂ in the neon buffer gas, HBr is also formed in the CuBr laser, which has the additional benefit of improving recombination in the afterglow via dissociative attachment of free electrons: $\text{HBr} + e^- \rightarrow \text{H} + \text{Br}^-$, followed by ion neutralization: $\text{Br}^- + \text{Cu}^+ \rightarrow \text{Br} + \text{Cu}^*$. As a result of the lower operating temperature and kinetic advantages of HBr, CuBr lasers are typically twice as efficient (2–3%) as their elemental counterparts. Sealed-off CuBr systems with powers of order 10–20 W are commercially produced.

An alternative technique for reducing the operating temperature of elemental CVLs is to flow a buffer gas mixture consisting of ~5% HBr in neon at approximately 50 mbar l min⁻¹ and allow this to react with solid copper metal placed within the plasma tube at about 600 °C to produce CuBr vapor *in situ*. The so-called Cu HyBrID (hydrogen bromide in discharge) laser has the same advantages as the CuBr laser (e.g., up to 3% efficiency) but at the cost of requiring flowing highly toxic HBr in the buffer gas, a requirement which has so far prevented commercialization of Cu HyBrID technology.

The kinetic advantages of the hydrogen halide in the CuBr laser discharge can also be applied to a conventional elemental CVL through the addition of small partial pressure of HCl to the buffer gas in addition to the 1–2% H₂ additive. HCl is preferred to HBr as it is less likely to dissociate (the dissociation energy of HCl at 0.043 eV is less than HBr at 0.722 eV). Such kinetic enhancement leads to a doubling in average output power, a dramatic increase in beam quality through improved gain characteristics, and shifts the optimum pulse repetition frequency for kinetically enhanced CVLs (KE-CVLs) from 4–10 kHz up to 30–40 kHz.

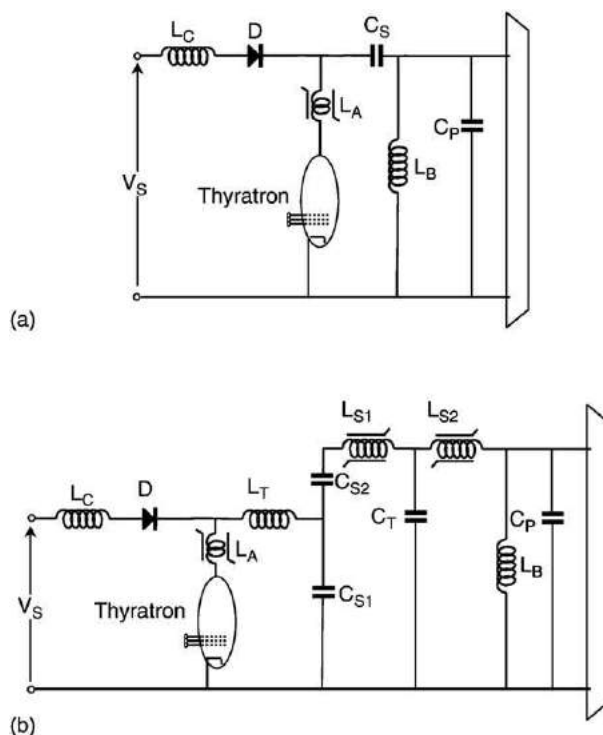


Fig. 5 Copper vapor laser excitation circuits. (a) Charge transfer circuit; (b) LC inversion circuit with magnetic pulse compression.

Preferential pumping of the upper laser levels requires an electron temperature in excess of the 2 eV range, hence high voltage (10–30 kV), high current (hundreds of A), short (75–150 ns) excitation pulses are required for efficient operation of copper vapor lasers. To generate such excitation pulses CVLs are typically operated with a power supply incorporating a high-voltage thyatron switch. In the most basic configuration, the charge-transfer circuit shown in Fig. 5(a), a dc high-voltage power supply resonantly charges a storage capacitor (C_S , typically a few nF) through a charging inductor L_C , a high-voltage diode and bypass inductor L_B , up to twice the supply voltage V_S in a time of order 100 μ s. When the thyatron is triggered, the storage capacitor discharges through the thyatron and the laser head on a time-scale of 100 ns. Note that during the fast discharge phase, the bypass inductor in parallel with the laser head can be considered to be an open circuit. A peaking capacitor C_P ($\sim 0.5 C_S$) is provided to increase the rate of rise of the voltage pulse across the laser head. Given the relatively high cost of thyatrons, more advanced circuits are now often used to extend the service lifetime of the thyatron to several thousand hours. In the more advanced circuit (Fig. 5(b)) an LC inversion scheme is used in combination with magnetic pulse compression techniques and operates as follows. Storage capacitors C_S are resonantly charged in parallel to twice the dc supply voltage V_S as before. When the thyatron is switched, the charge on C_{S1} inverts through the thyatron and transfer inductor L_T . This LC inversion drags the voltage on the top of C_{S2} down to $-4V_S$. When the voltage across the first saturable inductor L_{S1} reaches a maximum ($-4V_S$) L_{S1} saturates, and allows current to flow from the storage capacitors (now charged in series) to the transfer capacitor ($C_T = 0.5C_{S2}$) thereby transferring the charge from C_{S1} and C_{S2} to C_T in a time much less than the initial LC inversion time. At the moment when the charge on transfer capacitor C_T reaches a maximum (also $-4V_S$), L_{S2} saturates, and the transfer capacitor is discharged through the laser tube, again with a peaking capacitor to increase the voltage rise time. By using magnetic pulse compression the thyatron switched voltage can be reduced by 4 and the peak current similarly reduced (at the expense of increased current pulse duration) thereby greatly extending the thyatron lifetime. Note that in both circuits a 'magnetic assist' L_A saturable inductor is provided in series with the thyatron to delay the current pulse through the thyatron until after the thyatron has reached high conductivity thereby reducing power deposition in the thyatron.

Copper vapor lasers produce high average powers (2–100 W available commercially, with laboratory devices producing average powers of over 750 W) and have wall-plug efficiencies of approximately 1%. Copper vapor lasers also make excellent amplifiers due to their high gains, and the amplifiers can be chained together to produce average powers of several kW. Typical pulse repetition frequencies range from 4 to 20 kHz, with a maximum reported pulse repetition frequency of 250 kHz. An approximate scaling law states that for an elemental device with tube diameter D (mm) and length L (m) the average output power in watts will be of order $D \times L$. For example, a typical 25 W copper vapor laser will have a 25 mm diameter by 1 m long laser tube, and operate at 10 kHz corresponding to 2.5 mJ pulse energy and 50 kW peak power (50 ns pulse duration). Copper vapor lasers have very high single-pass gains (greater than 1000 for a 1 m long tube), large gain volumes and short gain durations (20–80 ns, sufficient for the intracavity laser light to make only a few round trips within the optical resonator). Maximum output power is therefore usually obtained in a highly 'multimode' (spatially incoherent) beam by using either a fully stable or a plane-plane resonator with a low

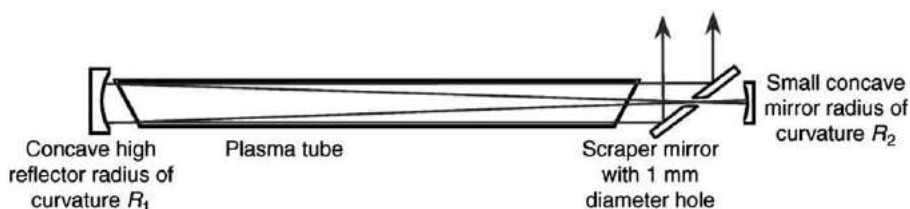


Fig. 6 Unstable resonator configuration often used to obtain high beam quality from copper vapor lasers.

reflectivity output coupler (usually Fresnel reflection from an uncoated optic is sufficient). To obtain higher beam quality a high magnification unstable resonator is required (**Fig. 6**). Fortunately, copper vapor lasers have sufficient gain to operate efficiently with unstable resonators with magnifications ($M = R_1/R_2$) up to 100 and beyond. Resonators with such high magnifications impose very tight geometric constraints on propagation of radiation on repeated round-trips within the resonator, such that after two round-trips the divergence is typically diffraction-limited. Approximately half the stable-resonator output power can therefore be obtained with near diffraction-limited beam quality by using an unstable resonator. Often, a small-scale oscillator is used in conjunction with a single power amplifier to produce high output power with diffraction-limited beam quality. Hyperfine splitting combined with Doppler broadening lead to an inhomogeneous linewidth for the laser transitions of order 8–10 GHz corresponding to a coherence length of order 3 cm.

The high beam quality and moderate peak power of CVLs allows efficient nonlinear frequency conversion to the UV by second harmonic generation (510.55 nm $\rightarrow$ 255.3 nm, 578.2 nm $\rightarrow$ 289.1 nm) and sum frequency generation (510.55 nm + 578.2 nm $\rightarrow$ 271.3 nm) using β -barium borate β -BaB₂O₄ (BBO) as the nonlinear medium. Typically average powers in excess of 1 W can be obtained at any of the three wavelengths from a nominally 20 W CVL. Powers up to 15 W have been obtained at 255 nm from high-power CVL master-oscillator power-amplifier systems using cesium lithium borate; CsLi B₆O₁₀ (CLBO) as the nonlinear crystal.

Key applications of CVLs include pumping of dye lasers (principally for laser isotope separation) and pumping Ti:sapphire lasers. Medically, the CVL yellow output is particularly useful for treatment of skin lesions such as port wine stain birth marks. CVLs are also excellent sources of short-pulse stroboscopic illumination for high-speed imaging of fast objects and fluid flows. The high beam quality, visible wavelength and high pulse repetition rate make CVLs very suited to precision laser micromachining of metals, ceramics and other hard materials. More recently, the second harmonic at 255 nm has proved to be an excellent source for writing Bragg gratings in optical fibers.

Afterglow Recombination Metal Vapor Lasers

Recombination of an ionized plasma in the afterglow of a discharge pulse provides a mechanism for achieving a population inversion and hence laser output. The two main afterglow recombination metal vapor lasers are the strontium ion vapor laser (430.5 nm and 416.2 nm) and calcium ion vapor laser (373.7 nm and 370.6 nm) whose output in the violet and UV spectral regions extends the spectral coverage of pulsed metal vapor lasers to shorter wavelengths.

A population inversion is produced by recombination pumping where doubly ionized Sr (or Ca) recombines to form singly ionized Sr (or Ca) in an excited state: $\text{Sr}^{++} + e^- + e^- \rightarrow \text{Sr}^{+*} + e^-$. Note that recombination rates for a doubly ionized species are much faster than for singly ionized species hence recombination lasers are usually metal ion lasers. Recombination (pumping) rates are also greatest in a cool dense plasma, hence helium is usually used as the buffer gas as it is a light atom hence promotes rapid collisional cooling of the electrons. Helium also has a much higher ionization potential than the alkaline-earth metals (including Sr and Ca) which ensures preferential ionization of the metal species (up to 90% may be doubly ionized).

Recombination lasers can operate where the energy level structure of an ion species can be considered to consist of two (or more) groups of closely spaced levels. In the afterglow of a pulsed discharge electron collisional mixing within each group of levels will maintain each group in thermodynamic equilibrium yielding a Boltzmann distribution of population within each group. If the difference in energy between the two groups of levels ($5\text{eV} \gg kT$) is sufficiently large then thermal equilibrium between the groups cannot be maintained via collisional processes. Given that recombination yields excited singly ionized species (usually with a flux from higher to lower ion levels), it is possible to achieve a population inversion between the lowest level of the upper group and the higher levels within the lower group. This is the mechanism for inversion in the Sr^+ (and analogous Ca^+) ion laser as indicated in the partial energy level scheme for Sr^+ shown in **Fig. 7**.

Strontium and calcium ion recombination lasers have similar construction to copper vapor lasers described above. The lower operating temperatures (500–800°C) mean that minimal or no thermal insulation is required for self-heated devices. For good tube lifetime BeO plasma tubes are required due to the reactivity of the metal vapors. Usually helium at a pressure of up to one atmosphere is used as the buffer gas. Typical output powers at 5 kHz pulse repetition frequency are of order 1 W ($\sim 0.1\%$ wall plug efficiency) at 430.5 nm from the He Sr^+ laser and 0.7 W at 373.7 nm from the He Ca^+ laser. For the high specific input power densities ($10\text{--}15\text{ W cm}^{-3}$) required for efficient lasing, overheating of the laser gas limits aperture scaling beyond 10–15 mm diameter. Slab laser geometries have been used successfully to aperture-scale strontium ion lasers. Scaling of the laser tube length

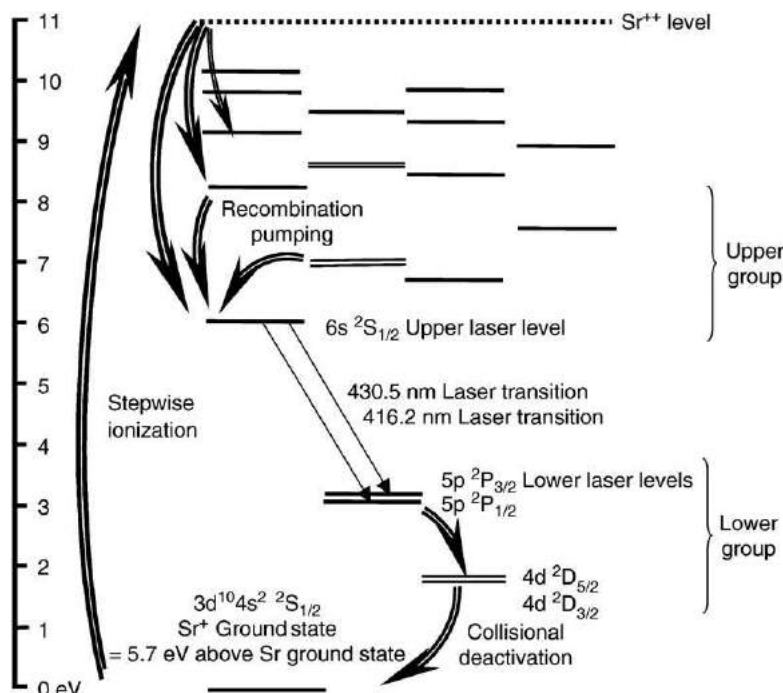


Fig. 7 Partial energy level scheme for the strontium ion laser.

beyond 0.5 m is not practical as achieving high enough excitation voltages for efficient lasing becomes problematic. Gain in strontium and calcium ion lasers is lower than in resonance-metastable metal vapor lasers such as the CVL, hence optimum output coupler reflectivity is approximately 70%. The pulse duration is also longer at around 200–500 ns resulting in moderate beam quality with plane-plane resonators.

For both strontium and calcium ion lasers the two principal transitions share upper laser levels and hence exhibit gain competition such that without wavelength-selective cavities usually only the longer wavelength of the pair is produced. With wavelength-selective cavities up to 60% (Sr^+) and 30% (Ca^+) of the normal power can be obtained at the shorter wavelength.

Many potential applications exist for strontium and calcium ion lasers, given their ultraviolet (UV)/violet wavelengths. Of particular importance is fluorescence spectroscopy in biology and forensics, treatment of neonatal jaundice, stereolithography, micromachining and exposing photoresists for integrated circuit manufacture. Despite these many potential applications, technical difficulties in power scaling mean that both strontium ion and calcium ion lasers have only limited commercial availability.

Continuous-Wave Metal Ion Lasers

In a discharge excited noble gas, there can be a large concentration of noble gas atoms in excited metastable states and noble gas ions in their ground states. These species can transfer their energy to a minority metal species (M) via two key processes: either charge transfer (Duffendack reactions) with the noble gas ions (N^+):



or Penning ionization with a noble gas atom in an excited metastable state (N^*):



In a continuous discharge in a mixture of a noble gas and a metal vapor, steady generation of excited metal ions via energy transfer processes can lead to a steady state population inversion on one or more pairs of levels in the metal ion and hence produce cw lasing.

Several hundred metal ion laser transitions have been observed to lase in a host of different metals. The most important such laser is the helium cadmium laser, which has by far the largest market volume by number of unit sales of all the metal vapor lasers. In a helium cadmium laser, Cd vapor is present at a concentration of about 1–2% in a helium buffer gas which is excited by a dc discharge. Excitation to the upper laser levels of Cd (Fig. 8) is primarily via Penning ionization collisions with $\text{He}^+ 2^3\text{S}_1$ metastable ions produced in the discharge. Excitation via electron-impact excitation from the ion ground state may also play an important role in establishing a population inversion in Cd^+ lasers. Population inversion can be sustained continuously because the $2^3\text{P}_{3/2}$ and $2^3\text{P}_{1/2}$ lower laser levels decay via strong resonance transitions to the $2^3\text{S}_{1/2}$ Cd^+ ground state, unlike in the

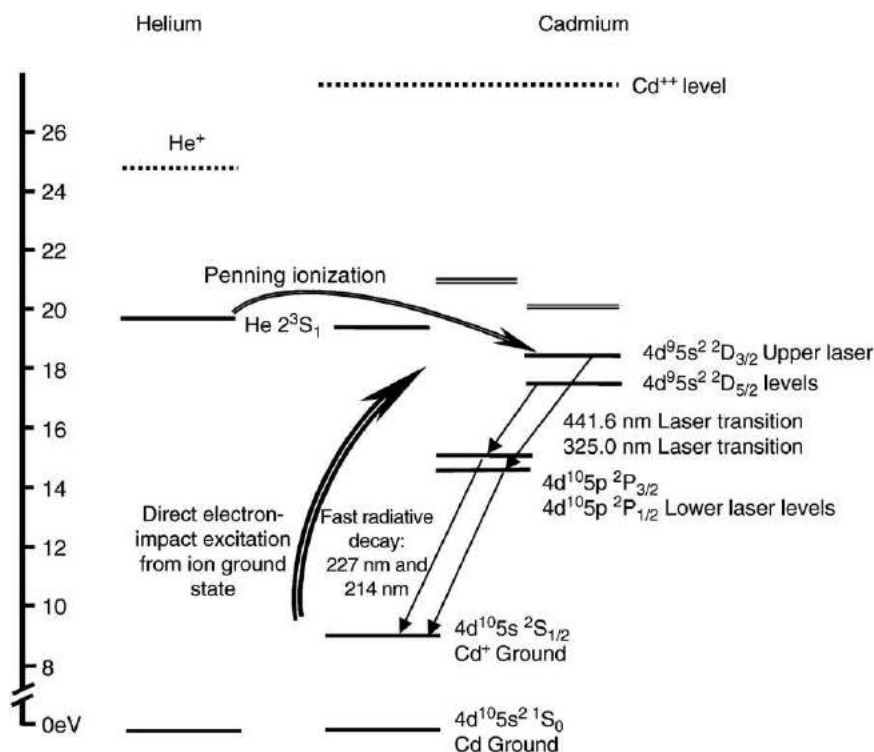


Fig. 8 Partial energy level diagram for helium and cadmium giving HeCd laser transitions.

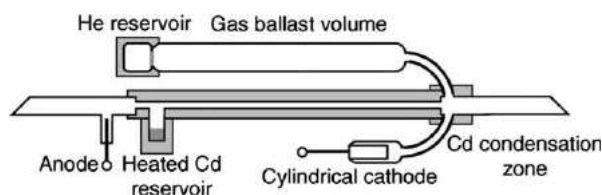


Fig. 9 Helium cadmium laser tube construction.

self-terminating metal vapor lasers. Two principal wavelengths can be produced, namely 441.6 nm (blue) and 325.0 nm (UV) with cw powers up to 200 mW and 50 mW available respectively from commercial devices.

Typical laser tube construction (Fig. 9) consists of a 1–3 mm diameter discharge channel typically 0.5 m long. A pin anode is used at one end of the laser tube together with a large-area cold cylindrical cathode located in a side-arm at the other end of the tube. Cadmium is transported to the main discharge tube from a heated Cd reservoir in a side-arm at 250–300 °C. With this source of atoms at the anode end of the laser a cataphoresis process (in which the positively charged Cd ions are propelled towards the cathode end of the tube by the longitudinal electric field) transports Cd into the discharge channel. Thermal insulation of the discharge tube ensures that it is kept hotter than the Cd reservoir to prevent condensation of Cd from blocking the tube bore. A large-diameter Cd condensation region is provided at the cathode end of the discharge channel. A large-volume side arm is also provided to act as a gas ballast for maintaining correct He pressure. As He is lost through sputtering and diffusion through the Pyrex glass tube walls, it is replenished from a high-pressure He reservoir by heating a permeable glass wall separating the reservoir from the ballast chamber which allows He to diffuse into the ballast chamber. Typical commercial sealed-off Cd lasers have operating lifetimes of several thousand hours. Overall laser construction is not that much more complex than a HeNe laser, hence unit costs are considerably lower than low-power argon ion lasers which also provide output in the blue. Usually Brewster angle windows are provided together with a high Q stable resonator (97–99% reflectivity output coupler) to provide polarized output.

With their blue and UV wavelengths and relatively low cost (compared to low-power argon ion lasers), HeCd lasers have found wide application in science, medicine and industry. Of particular relevance is their application for exposing photoresists where the blue wavelength provides a good match to the peak photosensitivity of photoresist materials. A further key application is in stereolithography where the UV wavelength is used to cure an epoxy resin. By scanning the UV beam in a raster pattern across the surface of a liquid epoxy, a solid three-dimensional object may be built up in successive layers.

Further Reading

- Little, C.E., 1999. Metal Vapor Lasers. Chichester, UK: John Wiley.
- Ivanov, I.G., Latush, E.L., Sem, M.F., 1996. Metal Vapor Ion Lasers, Kinetic Processes and Gas Discharges. Chichester, UK: John Wiley.
- Little CE and Sabotinov NV (eds) (1996) *Pulsed Metal Vapor Lasers*. Dordrecht, The Netherlands: Kluwer.
- Lyabin, N.A., Chursin, A.D., Ugol'nikov, S.A., Koroleva, M.E., Kazaryan, M.A., 2001. Development, production, and application of sealed-off copper and gold vapor lasers. *Quantum Electronics* 31, 191–202.
- Petrash, G.G. (Ed.), 1989. Metal Vapor and Metal Halide Vapor Lasers. Commack, NY: Nova Science Publishers.
- Silfvast, W.T., 1996. Laser Fundamentals. New York: Cambridge University Press.

Noble Gas Ion Lasers

WB Bridges, California Institute of Technology, Pasadena, CA, USA

© 2005 Elsevier Ltd. All rights reserved.

History

The argon ion laser was discovered in early 1964, and is still commercially available in 2004, with about \$70 million in annual sales 40 years after this discovery. The discovery was made independently and nearly simultaneously by four different groups; for three of the four, it was an accidental result of studying the excitation mechanisms in the mercury ion laser (historically, the first ion laser), which had been announced only months before. For more on the early years of ion laser research and development, see the articles listed in the Further Reading section at the end of this article.

The discovery was made with pulsed gas discharges, producing several wavelengths in the blue and green portions of the spectrum. Within months, continuous operation was demonstrated, as well as oscillation on many visible wavelengths in ionized krypton and xenon. Within a year, over 100 wavelengths were observed to oscillate in the ions of neon, argon, krypton, and xenon, spanning the spectrum from ultraviolet to infrared; oscillation was also obtained in the ions of other gases, for example, oxygen, nitrogen, and chlorine. A most complete listing of all wavelengths observed as gaseous ion lasers is given in the Laser Handbook cited in the Further Reading section. Despite the variety of materials and wavelengths demonstrated, however, it is the argon and krypton ion lasers that have received the most development and utilization.

Continuous ion lasers utilize high current density gas discharges, typically 50 A or more, and 2–5 mm in diameter. Gas pressures of 0.2 to 0.5 torr result in longitudinal electric fields of a few V/cm of discharge, so that the power dissipated in the discharge is typically 100 to 200 W/cm. Such high-power dissipation required major technology advances before long-lived practical lasers became available. Efficiencies have never been high, ranging from 0.01% to 0.2%. A typical modern ion laser may produce 10 W output at 20 kW input power from 440 V three-phase power lines and require 6–8 gallons/minute of cooling water. Smaller, air-cooled ion lasers, requiring 1 kW of input power from 110 V single-phase mains can produce 10–50 mW output power, albeit at even lower efficiency.

Theory of Operation

The strong blue and green lines of the argon ion laser originate from transitions between the 4p upper levels and 4s lower levels in singly ionized argon, as shown in Fig. 1. The 4s levels decay radiatively to the ion ground state. The strongest of these laser lines are listed in Table 1. The notation used for the energy levels is that of the L - S coupling model. The ion ground state electron configuration is $3s^23p^5(^2P_{3/2}^o)$. The inner ten electrons have the configuration $1s^22s^22p^6$, but this is usually omitted for brevity. The excited states shown in Fig. 1 result from coupling a 4p or 4s electron to a $3s^23p^4(^3P)$ core, with the resulting quantum numbers S (the net spin of the electrons), L (the net angular momentum of the electrons), and J (the angular momentum resulting from

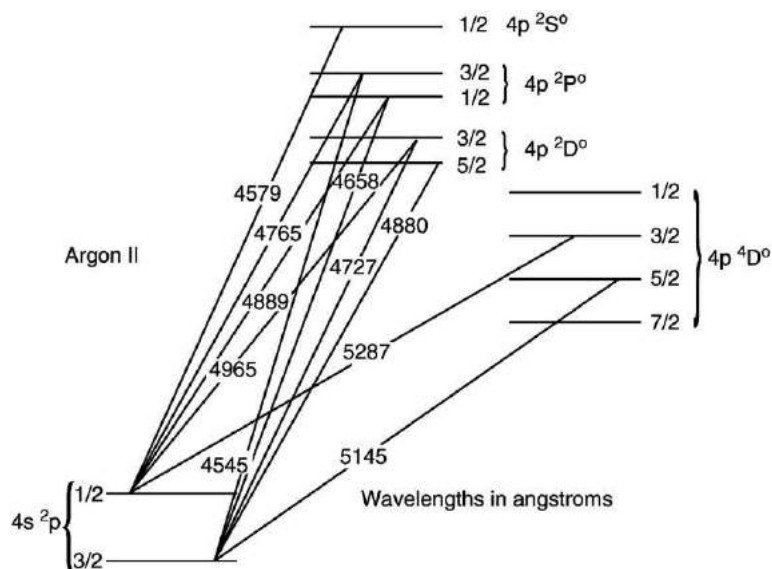


Fig. 1 4p and 4s doublet levels in singly ionized argon, showing the strongest blue and green laser transitions.

Table 1 Ar II laser blue-green wavelengths

Wavelength (nanometers)	Transition ^a (upper level) → (lower level)	Relative strength ^b (Watt)
454.505	$4p \ ^2P_{3/2}^o \rightarrow 4s \ ^2P_{3/2}$	0.8
457.935	$4p \ ^2S_{1/2}^o \rightarrow 4s \ ^2P_{1/2}$	1.5
460.956	$(^1D)4p \ ^2F_{7/2}^o \rightarrow (^1D)4s \ ^2D_{5/2}$	—
465.789	$4p \ ^2P_{1/2}^o \rightarrow 4s \ ^2P_{3/2}$	0.8
472.686	$4p \ ^2D_{3/2}^o \rightarrow 4s \ ^2P_{3/2}$	1.3
476.486	$4p \ ^2P_{3/2}^o \rightarrow 4s \ ^2P_{1/2}$	3.0
487.986	$4p \ ^2D_{5/2}^o \rightarrow 4s \ ^2P_{3/2}$	8.0
488.903	$4p \ ^2P_{1/2}^o \rightarrow 4s \ ^2P_{1/2}$	^c
496.507	$4p \ ^2D_{3/2}^o \rightarrow 4s \ ^2P_{1/2}$	3.0
501.716	$(^1D)4p \ ^2D_{5/2}^o \rightarrow 3d \ ^2D_{3/2}$	1.8
514.179	$(^1D)4p \ ^2F_{7/2}^o \rightarrow 3d \ ^2D_{5/2}$	^c
514.532	$4p \ ^4D_{5/2}^o \rightarrow 4s \ ^2P_{3/2}$	10
528.690	$4p \ ^4D_{3/2}^o \rightarrow 4s \ ^2P_{1/2}$	1.8

^aAll levels are denoted in L - S coupling with the (3P) core unless otherwise indicated. Odd parity is denoted by the superscript 'o'.

^bRelative strengths are given as power output from a commercial Spectra-Physics model 2080-25S argon ion laser.

^cThese lines may oscillate simultaneously with the nearby strong line, but are not resolved easily in the output beam.

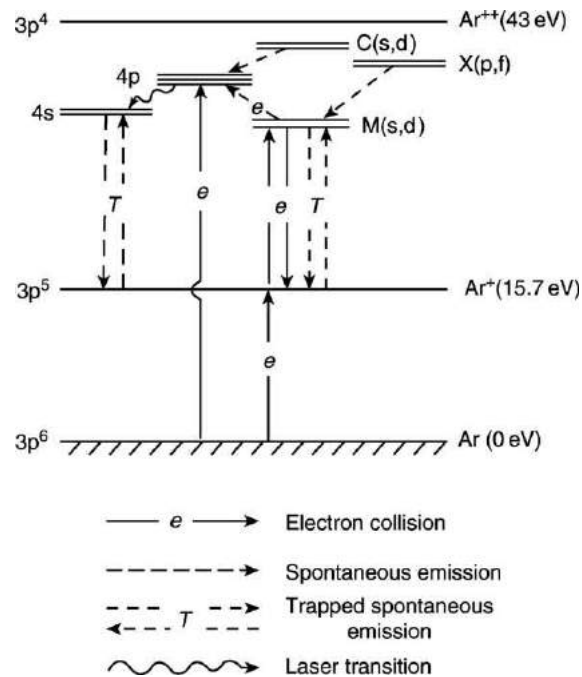


Fig. 2 Schematic representation of energy levels in neutral and singly ionized argon, indicating alternative pathways for excitation and de-excitation of the argon ion laser levels.

coupling S to L) are represented by $^{2S+1}L_J$, where $L=0,1,2,\dots$ is denoted $S, P, D, F,\dots$. The superscript 'o' denotes an odd level, while even levels omit a superscript. Note that some weaker transitions involving levels originating from the $3s^23p^4(^1D)$ core configuration also oscillate. Note also that the quantum mechanical selection rules for the L - S coupling model are not rigorously obeyed, although the stronger laser lines satisfy most or all of these rules. The selection rule $|\Delta J|=1$ or 0 , but not $J=0 \rightarrow J=0$ is always obeyed. All transitions shown in Fig. 1 and Table 1 belong to the second spectrum of argon, denoted Ar II. Lines originating from transitions in the neutral atom make up the first spectrum, Ar I; lines originating from transitions in doubly ionized argon are denoted Ar III, and so forth for even more highly ionized states.

Much work has been done to determine the mechanisms by which the inverted population is formed in the argon ion laser. Reviews of this extensive research are found in the Further Reading section. While a completely quantitative picture of argon ion laser operation is lacking to this day, the essential processes are known. Briefly, the $4p$ upper levels are populated by three pathways, as illustrated in Fig. 2:

- by electron collision with the $3p^6$ neutral ground state atoms. This 'sudden perturbation process' requires at least 37 eV electrons, and also singles out the $4p \ ^2P_{3/2}^o$ upper level, which implies that only the 476 and 455 nm lines would oscillate.

This is the behavior seen in pulsed discharges at very low pressure and very high axial electric field. This pathway probably contributes little to the 4p population under ordinary continuous wave (cw) operating conditions, however.

- (ii) by electron collision from the lowest-lying s and d states in the ion, denoted M(s,d) in Fig. 1. This only requires 3–4 eV electrons. These states have parity-allowed transitions to the ion ground state, but are made effectively metastable by radiation trapping (that is, there is a high probability that an emitted photon is re-absorbed by another ground state ion before it escapes the discharge region), or by requiring $|\Delta J|$ to be 2 to make the transition (forbidden by quantum selection rules). Thus, these levels are both created and destroyed primarily by electron collision, causing the population of the M(s,d) states to follow the population of the singly ionized ground state, which, in turn, is approximately proportional to the discharge current. Since a second electron collision is required to get from the M(s,d) states to the 4p states, a quadratic variation with current for the laser output power would be expected, and that is what is observed over some reasonable range of currents between threshold and saturation. Note that radiative decay from higher-lying opposite-parity p and f states, denoted X(p,f), can also contribute to the population of M(s,d), but the linear variation in population of the M(s,d) with discharge current is assured by electron collision creation and destruction.
- (iii) by radiative decay from higher-lying opposite-parity s and d states, denoted C(s,d). These states are populated by electron collision with the $3p^5$ ion ground states, and thus have populations that also vary quadratically with discharge current. The contribution of this cascade process has been measured to be 20% to 50% of the 4p upper laser level population.

Note that it is not possible to distinguish between processes (ii) and (iii) by the variation of output power with discharge current; both give the observed quadratic dependence.

The radiative lifetimes of the $4s^2P$ levels are sufficiently short to depopulate the lower laser levels by radiative decay. However, radiation trapping greatly lengthens this decay time, and a bottleneck can occur. In pulsed ion lasers, this is exhibited by the laser pulse terminating before the excitation current pulse ends, which would seem to preclude continuous operation. However, the intense discharge used in continuous operation heats the ions to the order of 2300 K, thus greatly Doppler broadening the absorption linewidth and reducing the magnitude of the absorption. Additionally, the ions are attracted to the discharge tube walls, so that the absorption spectrum is further broadened by the Doppler shift due to their wall-directed velocities. The plasma wall sheath gives about 20 V drop in potential from the discharge axis to the tube wall, so most ions hit the wall with 20 eV of energy, or about ten times their thermal velocity. A typical cw argon ion laser operates at ten times the gas pressure that is optimum for a pulsed laser, and thus the radiation trapping of the $4s \rightarrow 3p^5$ transitions is so severe that it may take several milliseconds after discharge initiation for the laser oscillation to begin.

As the discharge current is increased, the intensities of the blue and green lines of Ar II eventually saturate, and then decrease with further current. At these high currents, there is a buildup in the population of doubly ionized atoms, and some lines of Ar III can be made to oscillate with the appropriate ultraviolet mirrors. Table 2 lists the strongest of these lines, those that are available in the largest commercial lasers. Again, there is no quantitative model for the performance in terms of the discharge parameters, but the upper levels are assumed to be populated by processes analogous to those of the Ar II laser. At still higher currents, lines in Ar IV can be made to oscillate as well.

Table 2 Ultraviolet argon ion laser wavelengths

Wavelength (nanometers)	Spectrum	Transition ^a (upper level) → (lower level)	Relative strength ^b (Watt)
275.392	III	$(^2D^o)4p\ ^1D_2 \rightarrow (^2D^o)4s\ ^1D_2^o$	0.3
275.6	?	?	0.02
300.264	III	$(^2P^o)4p\ ^1P_1 \rightarrow (^2P^o)3d\ ^1D_2^o$	0.5
302.405	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)4s\ ^3P_2^o$	0.5
305.484	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)4s\ ^3P_1^o$	0.2
333.613	III	$(^2D^o)4p\ ^3F_4 \rightarrow (^2D^o)4s\ ^3D_3^o$	0.4
334.472	III	$(^2D^o)4p\ ^3F_3 \rightarrow (^2D^o)4s\ ^3D_2^o$	0.8
335.849	III	$(^2D^o)4p\ ^3F_2 \rightarrow (^2D^o)4s\ ^3D_1^o$	0.8
350.358	III	$(^2D^o)4p\ ^3D_2 \rightarrow (^2D^o)4s\ ^3D_2^o$	0.05
350.933	III	$(^4S^o)4p\ ^3P_0 \rightarrow (^4S^o)4s\ ^3S_1^o$	0.05
351.112	III	$(^4S^o)4p\ ^3P_2 \rightarrow (^4S^o)4s\ ^3S_1^o$	2.0
351.418	III	$(^4S^o)4p\ ^3P_1 \rightarrow (^4S^o)4s\ ^3S_1^o$	0.7
363.789	III	$(^2D^o)4p\ ^1F_3 \rightarrow (^2D^o)4s\ ^1D_2^o$	2.5
379.532	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)3d\ ^3P_2^o$	0.4
385.829	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)3d\ ^3P_1^o$	0.15
390.784	III	$(^2P^o)4p\ ^3D_1 \rightarrow (^2P^o)3d\ ^3P_0^o$	0.02
408.904	IV?	?	0.04
414.671	III	$(^2D^o)4p\ ^3P_2 \rightarrow (^2P^o)4s\ ^3P_2^o$	0.02
418.298	III	$(^2D^o)4p\ ^1P_1 \rightarrow (^2D^o)4s\ ^1D_2^o$	0.08

^aAll levels are denoted in $L-S$ coupling with the core shown in (). Odd parity is denoted by the superscript 'o'.

^bRelative strengths are given as power output from a commercial Spectra-Physics model 2085-25S argon ion laser.

Table 3 Krypton ion laser wavelengths

Wavelength (nanometers)	Spectrum	Transition ^a (upper level) → (lower level)	Relative strength ^b (Watt)
337.496	III	(² P ^o)5p ³ D ₃ → (² P ^o)5s ³ P ₂ ^o	–
350.742	III	(⁴ S ^o)5p ³ P ₂ → (⁴ S ^o)5s ³ S ₁ ^o	1.5
356.432	III	(⁴ S ^o)5p ³ P ₁ → (⁴ S ^o)5s ³ S ₁ ^o	0.5
406.737	III	(² D ^o)5p ¹ F ₃ → (² D ^o)5s ¹ D ₂ ^o	0.9
413.133	III	(⁴ S ^o)5p ⁵ P ₂ → (⁴ S ^o)5s ³ S ₁ ^o	1.8
415.444	III	(² D ^o)5p ³ F ₃ → (² D ^o)5s ¹ D ₂ ^o	0.3
422.658	III	(² D ^o)5p ³ F ₂ → (² D ^o)4d ³ D ₁ ^o	–
468.041	II	(³ P)5p ² S _{1/2} ^o → (³ P)5s ² P _{1/2}	0.5
476.243	II	(³ P)5p ² D _{3/2} ^o → (³ P)5s ² P _{1/2}	0.4
482.518	II	(³ P)5p ⁴ S _{3/2} ^o → (³ P)5s ² P _{1/2}	0.4
520.832	II	(³ P)5p ⁴ P _{3/2} ^o → (³ P)5s ⁴ P _{3/2}	–
530.865	II	(³ P)5p ⁴ P _{5/2} ^o → (³ P)5s ⁴ P _{3/2}	1.5
568.188	II	(³ P)5p ⁴ D _{5/2} ^o → (³ P)5s ² P _{3/2}	0.6
631.024	III	(² D ^o)5p ³ P ₂ → (² P ^o)4d ³ D ₁	0.2
647.088	II	(³ P)5p ⁴ P _{5/2} ^o → (³ P)5s ² P _{3/2}	3.0
676.442	II	(³ P)5p ⁴ P _{7/2} ^o → (³ P)5s ² P _{1/2}	0.9
752.546	II	(³ P)5p ⁴ P _{3/2} ^o → (³ P)5s ² P _{1/2}	1.2
793.141	II	(¹ D)5p ⁴ F _{7/2} ^o → (³ P)4d ² F _{5/2}	0.3
799.322	II	(³ P)5p ⁴ P _{3/2} ^o → (³ P)4d ⁴ D _{1/2}	

^aAll levels are denoted in *L–S* coupling with the core shown in (). Odd parity is denoted by the superscript 'o'.

^bRelative strengths are given as power output from a commercial Spectra-Physics model 2080RS ion laser.

Much less research has been done on neon, krypton, and xenon ion lasers, but it is a good assumption that the population and depopulation processes are the same in these lasers. **Table 3** lists both the Kr II and Kr III lines that are available from the largest commercial ion lasers. Oscillation on lines in still-higher ionization states in both krypton and xenon have been observed.

Operating Characteristics

A typical variation of output power with discharge current for an argon ion laser is shown in **Fig. 3**. This particular laser had a 4 mm diameter discharge in a water-cooled silica tube, 71 cm in length, with a 1 kG axial magnetic field. The parameter is the argon pressure in the tube before the discharge was struck. Note that no one curve is exactly quadratic, but that the envelope of the curves at different filling pressures is approximately quadratic. At such high discharge current densities (50 A in the 4 mm tube is approximately 400 A/cm²) there is substantial pumping of gas out of the small-bore discharge region. Indeed, a return path for this pumped gas must be provided from anode to cathode ends of the discharge to keep the discharge from self-extinguishing. The axial electric field in this discharge was 3–5 V/cm, so the input power was of the order of 10 to 20 kW, yielding an efficiency of less than 0.1%, an unfortunate characteristic of all ion lasers.

Technology

With such high input powers required in a small volume to produce several watts output, ion laser performance has improved from 1964 to the present only as new discharge technologies were introduced. The earliest laboratory argon ion lasers used thin-walled (≈ 1 mm wall thickness) fused silica discharge tubes, cooled by flowing water over the outside wall of the tube. The maximum input power per unit length of discharge was limited by thermal stresses in the silica walls caused by the temperature differential from inside to outside. Typically, ring-shaped cracks would cause the tube to fail catastrophically. Attempts to make metal–ceramic structures with alumina (Al₂O₃) discharge tubes to contain the plasma were made early on (1965) but were not successful. Such tubes invariably failed from fracture by thermal shock as the discharge was turned on. Later, successful metal–ceramic tubes were made with beryllia (BeO), which has a much higher thermal conductivity than silica or alumina and is much more resistant to thermal shock. Today, all of the lower power (less than 100 mW output) are made with BeO discharge tubes. Some ion lasers in the 0.5 to 1 W range are also made with water-cooled BeO discharge tubes.

A typical low-power air cooled argon ion laser is shown in **Fig. 4**. The large metal can (2) on the right end of the tube contains an impregnated-oxide hot cathode, heated directly by current through ceramic feed-through insulators (3). The small (≈ 1 mm diameter) discharge bore runs down the center of the BeO ceramic rod (1), and several smaller diameter gas return path holes run off-axis parallel to the discharge bore to provide the needed gas equalization between cathode can and anode region. A copper honeycomb cooler is brazed to the cathode can, two more to the outer wall of the BeO cylinder, and one to the anode (4) at the left end of the tube. The laser mirrors (5) are glass-fritted to the ends of the tube, forming a good vacuum seal. Note that in this small laser, the active discharge bore length is less than half of the overall length.

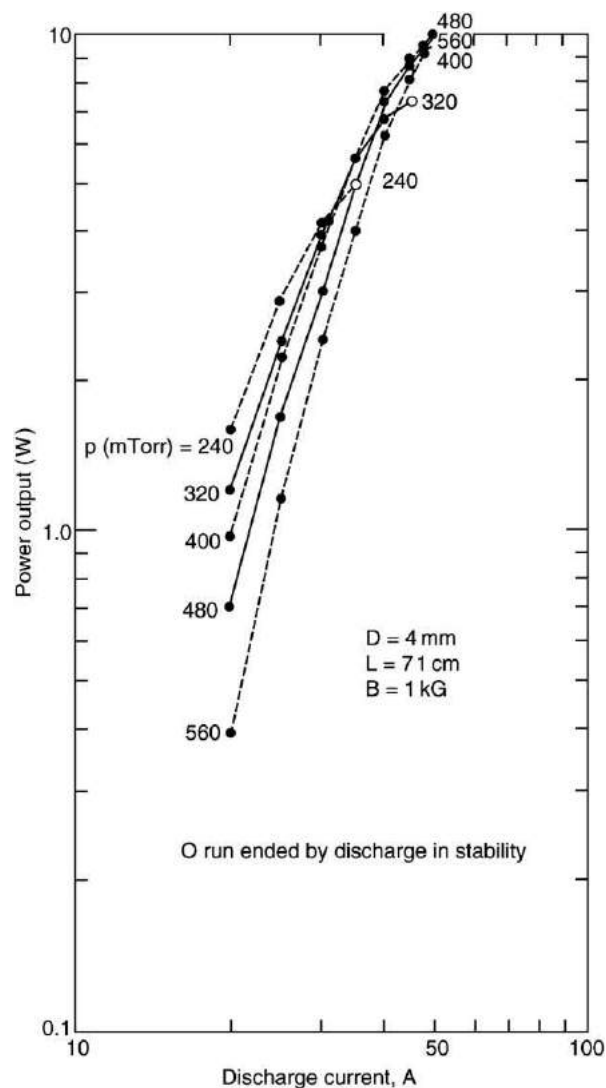


Fig. 3 Laser output power (summed over all the blue and green laser lines) versus discharge current for a laser discharge 4 mm in diameter and 71 cm long, with a 1 kilogauss longitudinal magnetic field. The parameter shown is the argon fill pressure in mTorr prior to striking the discharge. The envelope of the curves exhibits the quadratic variation of output power with discharge current.

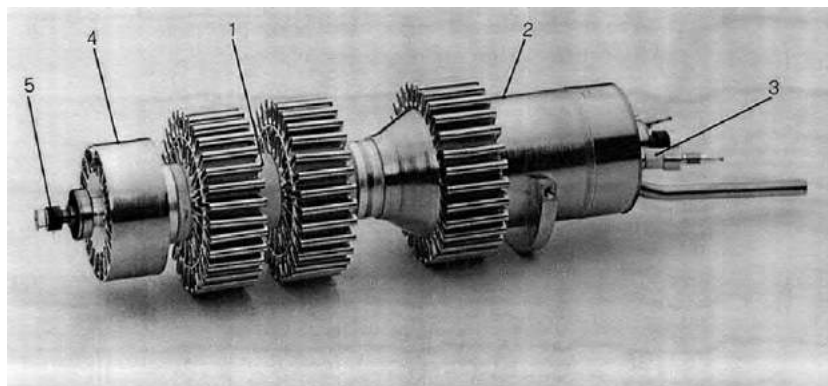


Fig. 4 A typical commercial low-power, air-cooled argon ion laser. The cathode can (2) is at the right and the beryllia bore (1) and anode (4) is at the left. The cathode current is supplied through ceramic vacuum feed-throughs (3). Mirrors (5) are glass fritted onto the ends of vacuum envelope. (Photo courtesy of JDS Uniphase Corp.)

The very early argon ion lasers were made with simple smooth dielectric tubes, allowing the continuous variation in voltage along the length of the tube required by the discharge longitudinal electric field. However, this variation can be step-wise at the discharge walls, and still be more or less smooth along the axis. Thus, the idea arose of using metal tube segments, insulated one from another, to form the discharge tube walls. The first version of this idea (1965) used short (≈ 1 cm) metal cylinders supported by metal disks and stacked inside a large diameter silica envelope, with each metal cylinder electrically isolated from the others. The tubes and disks were made of molybdenum, and were allowed to heat to incandescence, thus radiating several kilowatts of heat through the silica vacuum envelope to a water-cooled collector outside. While this eliminated the problems of thermal shock and poor thermal conductivity inherent in dielectric wall discharges, it made another problem painfully evident. The intense ion bombardment of the metal tube walls sputtered the wall material, eventually eroding the shape of the metal cylinders and depositing metal films on the insulating wall material, thus shorting one segment to another.

Many different combinations of materials and configurations were investigated in the 1960s and 1970s to find a structure that would offer good laser performance and long life. It was found that simple thin metal disks with a central hole would effectively confine the discharge to a small diameter (on the order of the hole size) if a longitudinal d-c magnetic field of the order of 1 kiloGauss were used. The spacing between disks can be as large as 2–4 discharge diameters. Of the metals, tungsten has the lowest sputtering yield for argon ions in the 20 eV range (which is approximately the energy they gain in falling to the wall across the discharge sheath potential difference). An even lower sputtering yield is exhibited by carbon, and graphite cylinders contained within a larger diameter silica or alumina tube were popular for a while for ion laser discharges. Unfortunately, graphite has a tendency to flake or powder, so such laser discharge tubes became contaminated with 'dust' which could eventually find its way to the optical windows of the tube. Beryllia and silica also sputter under argon ion bombardment, but with still lower yields than metals or carbon; however, their smaller thermal conductivities limit them to lower-power applications.

The material/configuration combination that has evolved for higher-power argon ion lasers today is to use a stack of thin tungsten disks with 2–3 mm diameter holes for the discharge. These disks, typically 1 cm in diameter, are brazed coaxially to a larger copper annulus (actually, a drawn cup with a 1 cm hole on its axis). A stack of these copper/tungsten structures is, in turn, brazed to the inside wall of a large diameter alumina vacuum envelope. The tungsten disks are exposed to and confine the discharge, while the copper cups conduct heat radially outward to the alumina tube wall, which, in turn, is cooled by fluid flow over its exterior. Thus, the discharge is in contact only with a low-sputtering material (tungsten) while the heat is removed by a high thermal conductivity material (copper). Details differ among manufacturers, but this 'cool disk' technology seems to have won out in the end. A photo of a half-sectioned disk/cup stacked assembly from a Coherent Innova™ ion laser is shown in Fig. 5. The coiled impregnated-tungsten cathode is also shown.

Sputtering of the discharge tube walls is not uniform along the length of the gas discharge. The small diameter region where the laser gain occurs is always joined to larger diameter regions containing cathode and anode electrodes (as shown in Fig. 5, for example). A plasma double sheath (that is, a localized increase in potential) forms across the discharge in the transition region

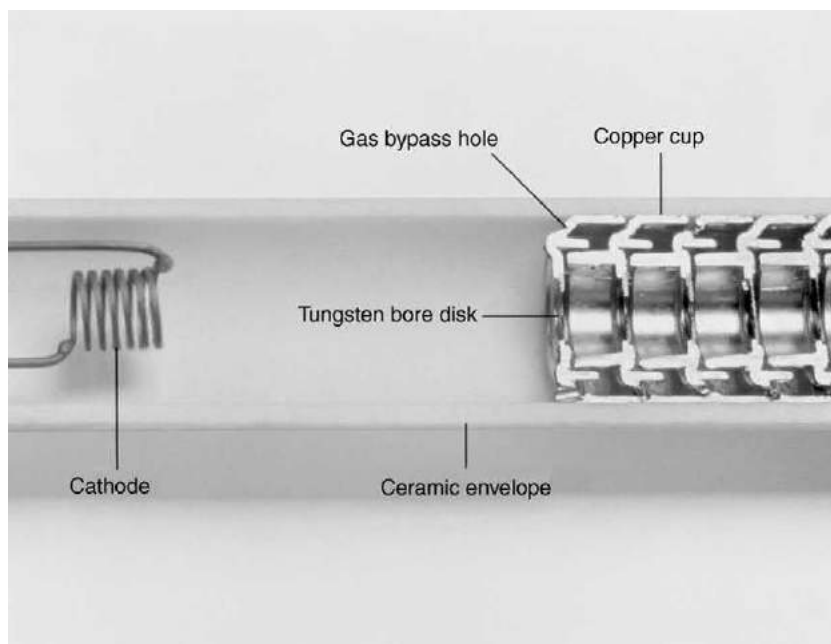


Fig. 5 Discharge bore structure of a typical commercial high-power argon ion laser, the Coherent Innova™. Copper cups are brazed to the inner wall of a ceramic envelope, which is cooled by liquid flow over its outer surface. A thin tungsten disk with a small hole defining the discharge path is brazed over a larger hole in the bottom of the copper cup. Additional small holes in the copper cup near the ceramic envelope provide a gas return path from cathode to anode. Also shown is the hot oxide-impregnated tungsten cathode. (Photo courtesy of Coherent, Inc.)

between large and small diameter regions (the discharge ‘throats,’ which may be abrupt or tapered). This sheath is required to satisfy the boundary conditions between the plasmas of different temperatures in the different regions. Such a double sheath imparts additional energy to the ions as they cross the sheath, perhaps an additional 20 eV. When these ions eventually hit the discharge walls near the location of the sheath, they have 40 eV of energy rather than the 20 eV from the normal wall sheath elsewhere in the small diameter discharge. Sputtering yield (number of sputtered wall atoms per incident ion) is exponentially dependent on ion energy in this low ion energy region, so the damage done to the wall in the vicinity of the ‘throat’ where the double sheath forms may be more than ten times that elsewhere in the small diameter bore region. This was the downfall of high-power operation of dielectric discharge bores, even BeO; while sputtering was acceptable elsewhere in the discharge, the amount of material removed in a small region near the discharge throat would cause catastrophic bore failure at that point. This localized increase in sputtering in the discharge throat is common to all ion lasers, including modern cooled tungsten disk tubes. Eventually, disks near the throat are eroded to larger diameters, and more material is deposited on the walls nearby. Attempts to minimize localized sputtering by tapering the throat walls or the confining magnetic field have proven unsuccessful; a localized double sheath always forms somewhere in the throat. Because of the asymmetry caused by ion flow, the double sheath is larger in amplitude in the cathode throat than the anode throat, so that the localized wall damage is larger at the cathode end of the discharge than at the anode end.

Another undesirable feature of sputtering is that the sputtered material ‘buries’ some argon atoms when it is deposited on a wall. This is the basis of the well-known Vac-Ion[®] vacuum pump. Thus, the operating pressure in a sealed laser discharge tube will drop during the course of operation. In low-power ion lasers, this problem is usually solved by making the gas reservoir volume large enough to satisfy the desired operating life (for example, the large cathode can in Fig. 4). In high-power ion lasers, the gas loss would result in unacceptable operating life even with a large reservoir at the fill pressure. Thus, most high-power ion lasers have a gas pressure measurement system and dual-valve arrangement connected to a small high-pressure reservoir to ‘burp’ gas into the active discharge periodically to keep the pressure within the operating range. Unfortunately, if a well-used ion laser is left inoperative for months, some of the ‘buried’ argon tends to leak back into the tube, with the gas pressure becoming higher than optimum. The pressure will gradually decrease to its optimum value with further operation of the discharge (in perhaps tens of hours). In the extreme case, the gas pressure can rise far enough so that the gas discharge will not strike, even at the maximum power supply voltage. Such a situation requires an external vacuum pump to remove the excess gas, usually a factory repair.

In addition to simple dc discharges, various other techniques have been used to excite ion lasers, primarily in a search for higher power, improved efficiency and longer operating life. Articles in the Further Reading section give references to these attempts. Radio-frequency excitation at 41 MHz was used in an inductively coupled discharge, with the laser bore and its gas return path forming a rectangular single turn of an air-core transformer. A commercial product using this technique was sold for a few years in the late 1960s. Since this was an ‘electrode-less’ discharge, ion laser lines in reactive gases such as chlorine could be made to oscillate, as well as the noble gases, without ‘poisoning’ the hot cathode used in conventional dc discharge lasers. A similar electrode-less discharge was demonstrated as a quasi-cw laser by using iron transformer cores and exciting the discharge with a 2.5 kHz square wave. Various microwave excitation configurations at 2.45 GHz and 9 GHz also resulted in ion laser oscillation. Techniques common to plasma fusion research were also studied. Argon ion laser oscillation was produced in Z-pinch and Θ -pinch discharges and also by high-energy (10–45 keV) electron beams. However, none of these latter techniques resulted in a commercial product, and all had efficiencies worse than the simple dc discharge lasers.

It is interesting that the highest-power output demonstrations were made in less than ten years after the discovery. Before 1970, 100 W output on the argon blue-green lines was demonstrated with dc discharges two meters in length. In 1970, a group in the Soviet Union reported 500 W blue-green output from a two-meter discharge with 250 kW of dc power input, an efficiency of 0.2%. Today, the highest output power argon ion laser offered for sale is 50 W output.

Manufacturers

More than 40 companies have manufactured ion lasers for sale over the past four decades. This field has now (2004), narrowed to the following:

Coherent, Inc.	http://www.coherentinc.com
INVERsion Ltd.	http://inversion.iae.nsk.su
JDS Uniphase	http://www.jdsu.com
Laser Physics, Inc.	http://www.laserphysics.com
Laser Technologies GmbH	http://www.lg-lasertechnologies.com
LASOS Lasertechnik GmbH	http://www.LASOS.com
Lexel Laser, Inc.	http://www.lexellaser.com
Melles Griot	http://lasers.mellesgriot.com
Spectra-Physics, Inc.	http://www.spectraphysics.com

Further Reading

Bridges, W.B., 1979. Atomic and ionic gas lasers. In: Marton, L., Tang, C.L. (Eds.), *Methods of Experimental Physics*, Vol 15: Quantum Electronics, Part A. New York: Academic Press, pp. 31–166.

- Bridges, W.B., 1982. Ionized gas lasers. In: Weber, M.J. (Ed.), *Handbook of Laser Science and Technology*, Vol. II, Gas Lasers. Boca Raton, FL: CRC Press, pp. 171–269.
- Bridges, W.B., 2000. Ion lasers – the early years. *IEEE Journal of Selected Topics in Quantum Electronics* 6, 885–898.
- Davis, C.C., King, T.A., 1975. Gaseous ion lasers. In: Goodwin, D.W. (Ed.), *Advances in Quantum Electronics*, vol. 3. New York: Academic Press, pp. 169–469.
- Dunn, M.H., Ross, J.N., 1976. The argon ion laser. In: Sanders, J.H., Stenholm, S. (Eds.), *Progress in Quantum Electronics*, vol. 4. New York: Pergamon, pp. 233–269.
- Weber, M.J., 2000. *Handbook of Laser Wavelengths*. Boca Raton, FL: CRC Press.

Planar Waveguide Lasers

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

a integer denoting number of laser modes
 E electric field
 k_z propagation constant for travel in z direction
 L_c length of laser cavity
 L_G length of grating
 m order of the grating, an integer
 n refractive index of the medium
 $\bar{n}$ modal effective refractive index
 R reflectance

V normalized frequency or the V parameter of the waveguide
 Δn_{eff} modulation of the effective index of the grating
 η confinement factor or fraction of the light confined in the region of interest
 λ wavelength of the lightwave
 λ_B resonant wavelength of a Bragg grating or Bragg wavelength
 Λ grating pitch
 Φ electric field distribution

Introduction

Lasers have evolved to become true marvels of technology. From their invention in the early 1960s, laser light has changed the way society operates, its application stretching to virtually every aspect of human activity. The versatility of lasers is derived from the fact that there are several different types, each with a unique set of advantages. Solid-state lasers, made from specific semi-conducting materials, are commonly used in the optical networks that underlie the greatest machine on Earth – the Internet. High power gas lasers, such as CO₂ lasers, are used in industrial machining, whereas rare-earth doped and dye lasers have found applications ranging from eye surgery to tattoo removal.

Planar waveguide lasers are a relatively new class of lasers made possible by advances in the field of optical waveguides. Over the past decade, with the aim to miniaturize optical components, planar waveguide technology has been developed to a high degree of sophistication. Likewise, Er and other rare-earth doping was studied extensively owing to its importance in all-optical amplification. These elements turned out to be the building blocks for the development of fiber lasers and subsequently planar waveguide lasers.

This chapter will discuss the basic concepts underlying this category of lasers, the variations therein, and the main characteristics. It is important to understand the basic features of waveguides and lasers, which is included as a prelude. We will then discuss planar waveguide lasers with emphasis on the essential elements: the laser materials, the various designs, their performance, and the resultant advantages as well as the applications.

Waveguides

A waveguide is a general conduit for efficient propagation of an electromagnetic field. It is usually defined by making a region of higher refractive index within a dielectric medium. This behaves as an attractive potential well that confines light in the form of bound modes, much like a potential well that confines electrons to a bound state. The propagating field is referred to as the lightwave and is governed by Maxwell's equation. For a weakly guiding approximation, as is the case when the refractive index contrast (Δn) is small, Maxwell's equation can be reduced to a scalar wave equation for the transverse electric field. For a given wavelength (λ) and polarization, the scalar wave equation in a nonmagnetic medium is

$$\nabla^2 E(x) + \frac{4\pi^2}{\lambda^2} n(x)^2 E(x) = 0 \quad (1)$$

where E is the electric field. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$ where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \Phi(x, y) + n(x, y)^2 \Phi(x, y) = (\bar{n})^2 \Phi(x, y) \quad (2)$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z \lambda / 2\pi$. The effective index dictates the occurrence of bound modes – these states occur when $\bar{n}$ is between n_{core} and n_{cladding} . For smaller values of $\bar{n}$ there is a continuum of unbound and radiation modes. These are leaky modes that are dispersed quickly over a short distance. Only bound modes are guided in the core of the waveguide structure.

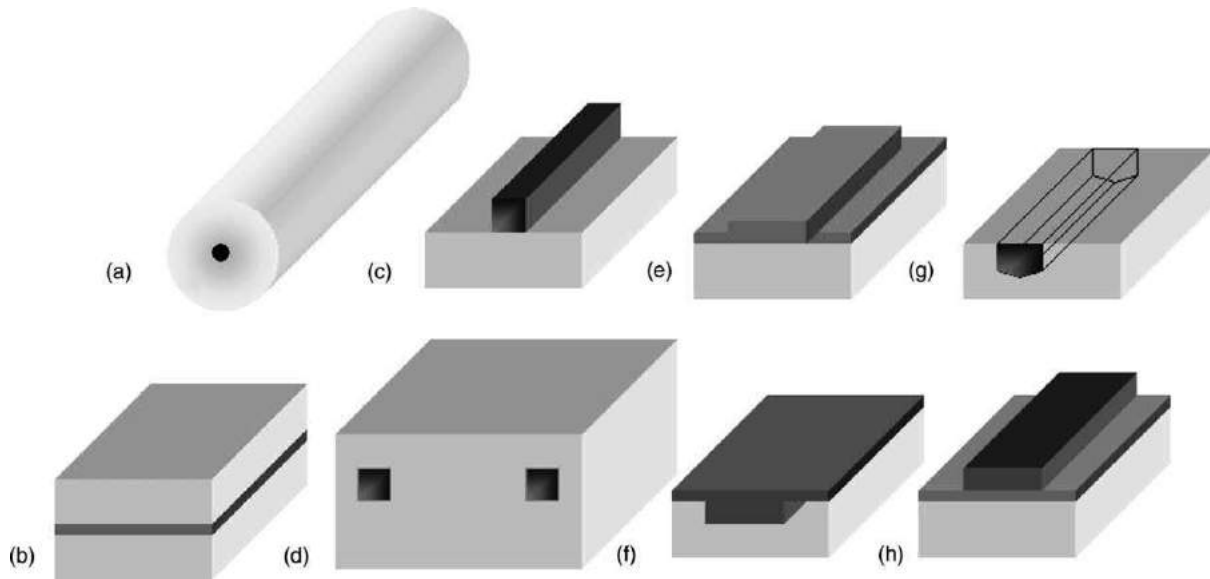


Fig. 1 Range of different waveguide designs, with the core represented by the darker regions: (a) Optical fiber; (b) slab waveguide; (c) channel waveguide; (d) embedded channel waveguides with a layer of the overcladding; (e) rib waveguide; (f) inverted rib; (g) ion diffused; and (h) strip loaded, comprising of a strip of a different higher index material (that serves to confine light) on a core layer.

Planar Waveguides

Waveguides formed on a flat substrate are called planar waveguides. These are typically made by stepwise deposition of films of dielectric materials (typically glass). The waveguide core is defined by one of several methods, the most common of which uses lithography and etching. In this case, a film of an appropriately higher index material is deposited on an 'under-cladding' film and then selectively etched to define a core. This is generally covered with a suitable 'upper-cladding' layer so that the index contrast is controlled in all directions. Other methods of forming the core include patterned ion exchange or ion implantation or increasing the index of the medium using a femtosecond laser, etc. In all these cases, the end result is the formation of an embedded higher-index region whose locus defines the optical path for the lightwave. A circuit made from these waveguides is known as planar lightwave circuits (PLC).

Waveguide cores may be designed in several different configurations, as shown in Fig. 1. These designs allow for different fabrication processes as well as the control over the modal properties of the lightwave. For a given index contrast, the number of modes traveling in the waveguide is primarily determined by the dimensions of the waveguide. The aspect ratio of the waveguide affects the propagation of the transverse electric (TE) and the transverse magnetic (TM) modes, due to different propagation constant k_z for each mode.

Lasers

The generation of laser light requires four basic conditions to be satisfied:

- a source capable of light emission;
- an energy supply source – typically electrical, optical, or chemical;
- population inversion to the excited state in the lasing medium; and
- absence of parasitic effects that would absorb the emitted light.

Laser light is often a result of either electronic or vibrational excitation. This is because these excitations involve transfer between energy states that coincide with the radiation of UV, visible, or infrared light. Population inversion is an important requirement in lasers since the emission is stimulated. This involves having a greater fraction of the atoms or molecules in the excited state so that a stimulating photon can cause a radiative decay to the lower energy level. The stimulated light from most lasing media is weak, which requires unacceptably long lengths to generate sufficient power. Instead, as shown in Fig. 2, lasers are made with two parallel mirrors with an axial light-generating source in between. This configuration is known as a resonant cavity, a resonator, or an oscillator. The back mirror is usually fully reflective but a partial front mirror allows a portion of the light in the cavity to pass through. The two mirrors cause the light to bounce to and fro, forcing it into making multiple passes through the cavity. Light emanating from this structure is thus amplified and is also coherent (that is the lightwaves are in phase with each other spatially and temporally). This has come to represent the signature output of a laser.

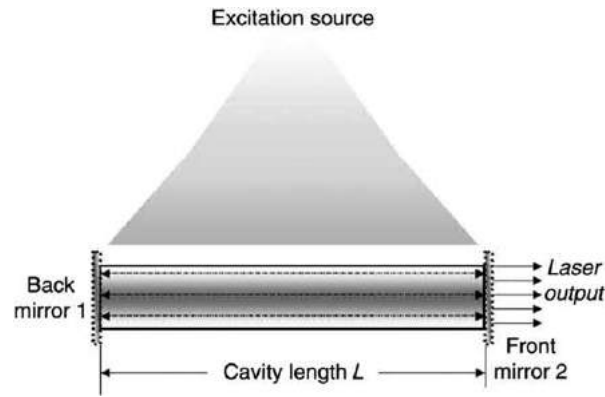


Fig. 2 Conceptual representation of a simple Fabry-Perot laser.

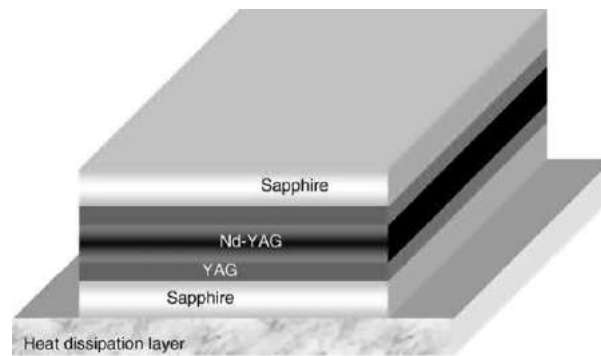


Fig. 3 End view of a planar waveguide form of a Nd-YAG laser made using sapphire (single crystal Al_2O_3) for cladding.

Planar Waveguide Lasers (PWL)

By way of an example, let us consider a solid state, rare-earth doped laser, such as an Nd-doped YAG laser. All rare-earth ions are able to fluoresce, some (e.g., Nd) more efficiently than others, and hence can be light emitters. These ions have to be dissolved (doped) into a solid host, say, a glass or a single crystal such as YAG (Yttrium Aluminum Garnet). The excitation of these ions is done optically with light of a suitably higher energy (or lower wavelength). When this 'pump' light is made incident on a rod of the rare earth doped material, it is absorbed by the rare earth ions and used to reach an excited electronic state. Amongst all the possible higher energy states, only certain ones are sustainable because the electron resides in that state for an acceptably long period of time (called the lifetime). Hence, population inversion can be more readily achieved in these longer lifetime states and stimulated processes such as amplification and lasing are readily possible. The laser light generated in the doped rod is emitted across its entire cross-section.

Planar waveguide lasers are designed to miniaturize the kinds of stand-alone solid state lasers describe above. In principle, a PWL is a laser where the light emitting resonant cavity is a planar waveguide. This topic of our interest is a relatively recent development that renders some unique advantages. These can be broadly classified into two categories: one relating to making efficient high-power lasers and the other, to the integration of lasers with other optical components.

An example of the first type is a Nd-YAG laser configured into a planar geometry so that the laser light is now contained within the waveguide. One design that has evolved to achieve this well is shown in Fig. 3. The light is guided by the waveguiding structure setup by the sapphire planes. Since both the excitation (or pump) and the lasing modes are well confined, high modal overlap is achieved. These PWLs can be edge-pumped or face-pumped and high powers (measured in the order of Watts) can be achieved.

In the above design, the cavity is defined by two parallel reflectors bracketing the light-generating medium and is known as a Fabry-Perot etalon. The distance between the two mirrors sets up a wavelength filtering mechanism whereby only resonant wavelengths are sustained. These are related to the intra-mirror distance by the simple equation shown below:

$$L_c = \frac{a\lambda}{2n} \quad (3)$$

where L_c is the length of the cavity, a is an integer, n is the refractive index of the medium, and λ is the wavelength in vacuum. As the cavity length increases, the number of possible resonant wavelengths increases and the spacing between wavelengths is

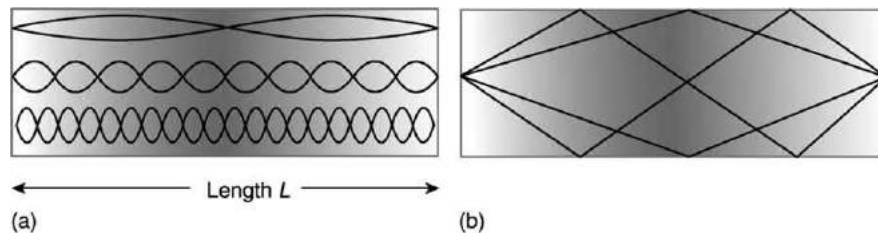


Fig. 4 Generation of longitudinal (a) and lateral modes (b) in a Fabry-Perot laser.

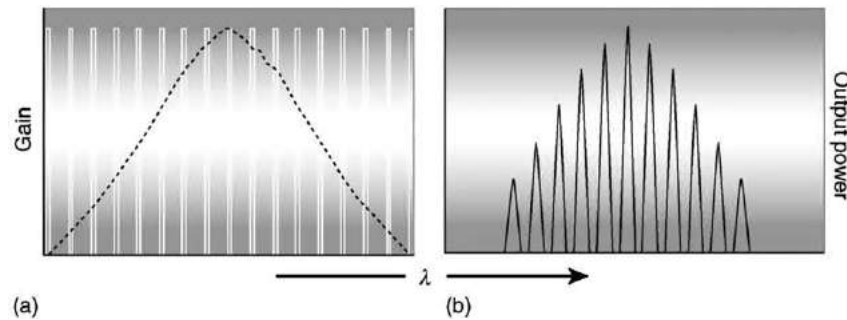


Fig. 5 The output of a Fabry-Perot laser: (b) is governed by the number of modes within the gain spectrum of the lasing medium shown by the dotted lines (a). The spacing between the modes can be altered changing the length of the cavity. The smallest mode in output spectrum (b) needs to have enough gain to overcome the lasing threshold.

reduced. These are called the longitudinal modes (**Fig. 4(a)**). Likewise, lateral modes (**Fig. 4(b)**) can be generated in the cavity as well, tending to make the operation of the laser multimoded, as shown in **Fig. 5**. This is characteristic of a Fabry-Perot laser, which can be overcome by embellishing the basic etalon design.

As stated above, another big advantage here is that the platform for fabricating PWLs can be chosen to be the same as to make PLCs. Waveguide comprising the laser can be continued further and configured into other desirable components, such as splitters, couplers, etc. Thus the laser does not have to be separately coupled to the waveguide. This is a major advantage as this connecting process is time-consuming and incurs loss of light. In the field of photonics, the ability to integrate active components with passive ones onto the same chip is considered to be the first step toward photonic chips, analogous to the integrated circuit chip that has revolutionized our world. This is the driver behind continued research and development of active elements such as PWLs that can be coupled with other actives, such as thin film electro-optic modulators or acousto-optic filters, etc.

Waveguide Laser Materials

The waveguiding structure is comprised of materials that are either dielectrics or semiconductors, both inorganic and organic. For their homogeneity and broad transmittance, high silica glasses are ideal materials for waveguiding. The chief advantages of high silica waveguides are easy index and mode matching to standard optical fibers, low propagation loss, and good durability. When crystalline materials are used, they have to be single-crystalline to avoid grain boundary scattering (for instance, Ti-diffused waveguide in single-crystal LiNbO_3). If semiconductors are used, the bandgap of the material needs to be sufficiently high so as to avoid absorption of the photons. Silicon waveguide (bandgap = 1.1 eV) on a silica cladding (SOI) is commonly used system in PLCs used for telecommunication applications.

The PWLs described above are fabricated in rare-earth doped glass waveguides. Waveguides have to be much shorter than fiber, so the challenge of incorporating the rare-earth ions at high concentration in the host material has to be overcome. Planar waveguide lasers are a compressed version fiber lasers that have the unique advantage that they can be directly integrated with other components on the same chip using very large-scale integration (VLSI)-style processing. The concentration of Er (or the lasing ion in general) in a PWL needs to be many times that in fiber amplifiers (~ 100 ppm). The need for high gain in a short length requires high pump powers and high lasing ion concentration, both of which lead to nonideal behavior. Since rare earth ions tend to have multiple excited electronic states, there can be absorption at the excited state level as well. When a higher energy state is separated from the upper emission level by the energy in one laser photon, the ions can undergo excited state absorption or ESA (**Fig. 6**). Rare-earth ions in close proximity tend to undergo cooperative upconversion processes, where two excited ions transfer energy between each other so that neither ion may fluoresce at the desired wavelength. These processes can negatively affect the useful gain of the system. Ensuring high spatial dispersion of the Er atoms is critical to minimize this effect. Special glass compositions, such as phosphate glasses, are used here to ensure that the Er is not phase separated or clustered. These glasses also

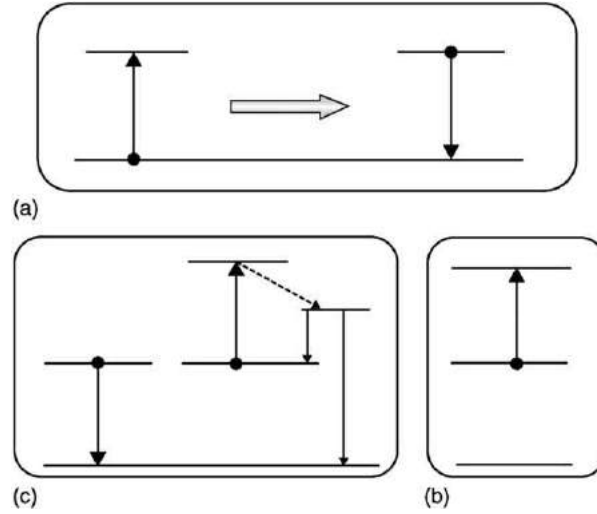


Fig. 6 The energy level diagram for a representative two level system. The inset (a) shows the excitation and emission process. Excited state absorption is shown in (b) and cooperative upconversion is shown in (c).

provide a high induced cross-section for gain. Glasses with low phonon energy are also desirable to minimize phonon-mediated nonradiative decays in the rare earths. Light emitting semiconductors and polymers containing fluorescent dyes have also been used to make PWLs.

Now that we have gained an overall perspective of PWLs, we can discuss different manifestations. The following discussion uses a Er doped system as a model since there is a substantial body of work on this system.

Distributed Bragg Reflector (DBR) PWL

In planar waveguide systems, it is difficult to construct efficient mirrors or reflective facets perpendicular to the waveguide. However, mirrors can be replaced by Bragg gratings, which be fabricated relatively easily. In-line Bragg gratings result in wavelength specific reflection as per Eq. (4):

$$\Lambda = \frac{m\lambda}{2n} \quad (4)$$

where Λ is the grating pitch, m is the order of the grating, and λ is the central Bragg wavelength. Gratings have another important benefit in that they ensure that unwanted frequencies outside the limited reflection spectrum are dispersed before they reach the lasing threshold.

A DBR laser is similar in principle to a Fabry–Perot design but uses a Bragg grating on either end of planar Er-doped waveguide. This is shown schematically in Fig. 7. Gratings can be made in a number of different ways. One popular way is to exploit the photosensitivity of glasses using 193 nm or 244 nm UV light to periodically alter the refractive index of a glass waveguide. Photosensitivity occurs by a combination of molecular changes (such as the formation of oxygen deficiency sites or color centers) as well as changes in the material strain in the region exposed to the UV light (Fig. 7(c)). Alternatively, gratings can also be formed using etching to make periodic grooves in the waveguide. These are called corrugated or surface relief gratings (Fig. 7(a) and 6(b)).

The back Grating 1 has a high reflectance of almost 100%, whereas Grating 2 is a weaker grating. The reflectance strength of Grating 2 is determined by taking into account the gain coefficient of the material and the output power required from the laser. The reflection intensity can be altered according to the following equation:

$$R = \tanh^2 \left[\frac{\pi L_G \Delta n_{\text{eff}} \eta(V)}{\lambda_B} \right] \quad (5)$$

where L_G is the grating length, Δn_{eff} is the index modulation of the grating, λ_B is the Bragg wavelength, $\eta(V)$ is the confinement factor, a function of the waveguide parameter V that represents the fraction of the integrated mode intensity contained in the core. Increasing the reflectance strength also widens the linewidth of reflected light. The overlap of the spectral width between the back (Grating 1) and the front grating (Grating 2) becomes the effective output window within the gain spectrum of the lasing material (Fig. 8).

Not unexpectedly, a DBR structure generates longitudinal and lateral modes. These modes are, however, limited to the overlapping section of the reflection spectrum of the two gratings, as opposed to the entire gain spectrum as in Fabry–Perot lasers (Fig. 9). Yet, this can lead to problems such as spectral hole burning and mode hopping, whereby the dominant mode hops from one wavelength to another. The separation between the modes needs to be made sufficiently high compared to the spectral overlap

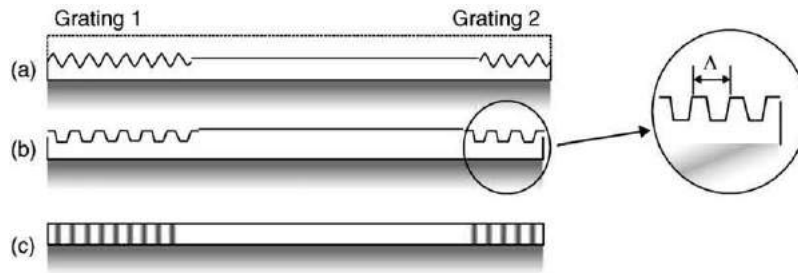


Fig. 7 DBR planar waveguide laser. The PWLs shown in (a) and (b) have sinusoidal and trapezoidal surface-relief gratings respectively etched in their cores. An in-line grating of the kind seen in (c) could be fabricated using the photosensitivity of the glass. The top cladding is shown only in (a).

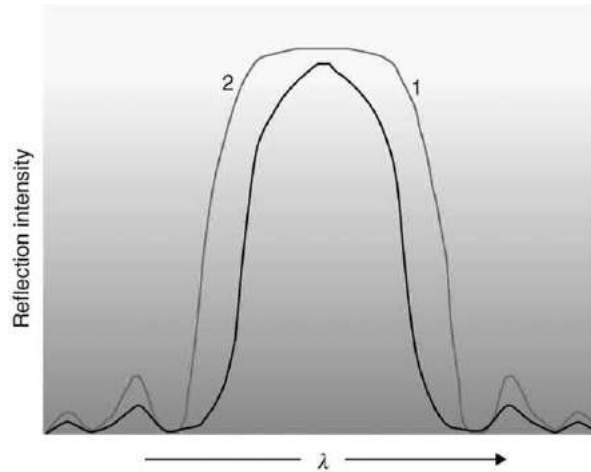


Fig. 8 The reflection spectrum of the back (1) and the front (2) gratings. As the intensity of the reflected light is made to increase, so does the linewidth. Hence, the narrower spectrum of partially reflecting grating 2 tends to dictate the operating wavelength range of a DBR.

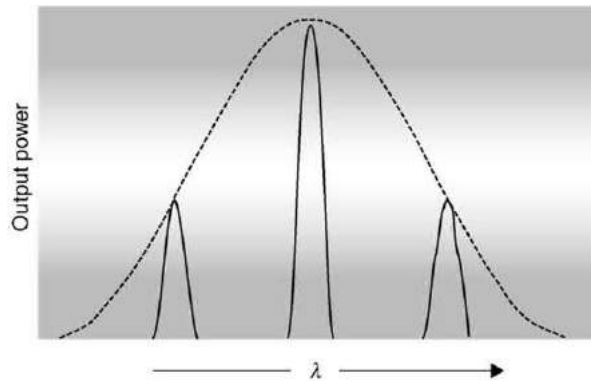


Fig. 9 A sample output from a DBR laser, shown in this case with 3 modes. The dotted line shows the effective overlap between the two gratings. The existence of the modes depends on several factors, in particular the pump intensity.

so that single-mode operation can be achieved. But in practice the length of the cavity (L_c in Eq. (3)) is determined by factors such as power and space and may not be independently altered to achieve this. However, the reflection spectrum of the back and the front gratings could be slightly shifted in opposite directions to realize a smaller overlap within which only one or a few modes may exist (Fig. 10). The occurrence of a single mode also depends on the gain profile of the lasing material and the optical pump power. In Er-doped PWLs, modes other than the primary one can be suppressed by using a pump power below a threshold limit and stable operation can be thus achieved.

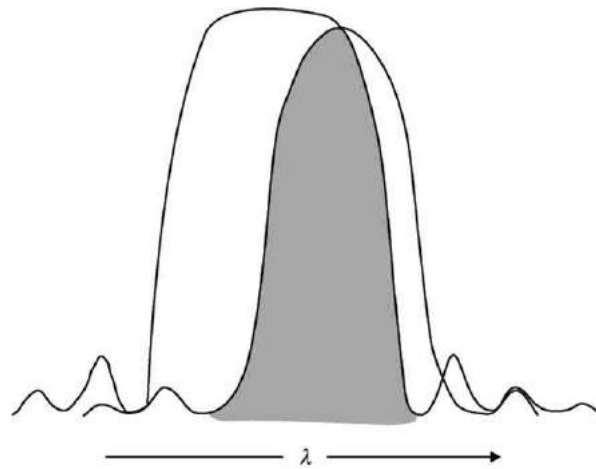


Fig. 10 Shifting the front and back gratings is one means of controlling the width of the overlap region, shown in gray. This way a single mode operation can be achieved even for a longer cavity with many modes.

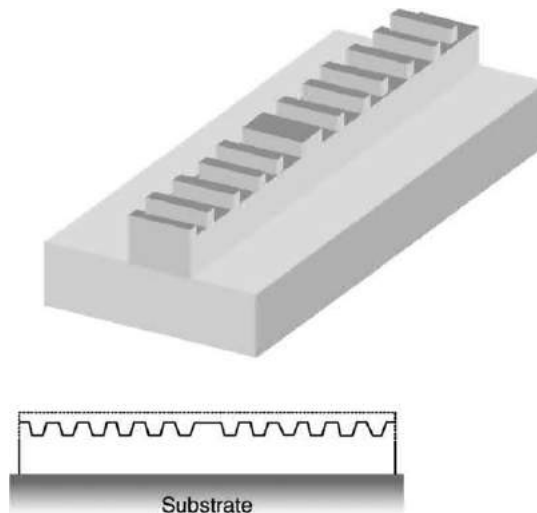


Fig. 11 A DFB planar waveguide laser.

Distributed Feedback (DFB) PWL

The modal problems listed above can be addressed by having a single, distributed grating that is placed along the entire length of the lasing waveguide itself (Fig. 11). Alternatively, the gratings may be also placed in the upper or the lower claddings to minimize loss of optical power due to scattering because of the longer grating. In this case, it is the evanescent portion of the light that interacts with the grating. In all cases, the periodicity of the grating leads to a condition for a single lasing wavelength based on constructive interference. The strongest mode occurs when the period Λ of the Bragg grating satisfies the primary Bragg condition (for first order grating or $m=1$) in Eq. (4). The device will also function if the grating pitch is equal to small integer multiples of $(\lambda/2\bar{n})$.

The grating needs to be made strong enough to generate sufficient feedback (reflections), which also widens the linewidth of the spectrum. To ensure a single narrow linewidth output, a quarter wave phase shift is introduced in the grating, as seen in Fig. 11. This phase shift creates a transmission fringe in the spectrum (Fig. 12) that significantly narrows the linewidth of the output signal. This design is considered to be standard for the DFB lasers of today. Fig. 13 shows a typical output of a DFB laser.

Any of the geometries seen in Fig. 1 can be used to make the waveguide for the laser. Due to processing constraints, rib (c) and ion diffused (d) configurations are commonly used. Waveguide losses, characteristically due to scattering from sidewall roughness, have to be low. Otherwise, this adds a significant overhead onto the pump power and lowers the output of the PWL. Birefringence caused by thermal expansion coefficient differences and excessive sidewall roughness also need to be minimized as they can substantially alter the loss of the TE mode compared to TM.

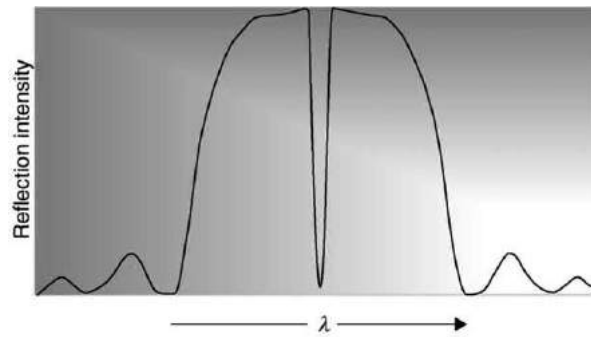


Fig. 12 The reflection spectrum of a grating with a central $\lambda/4$ shift in the pitch, showing the transmission fringe.

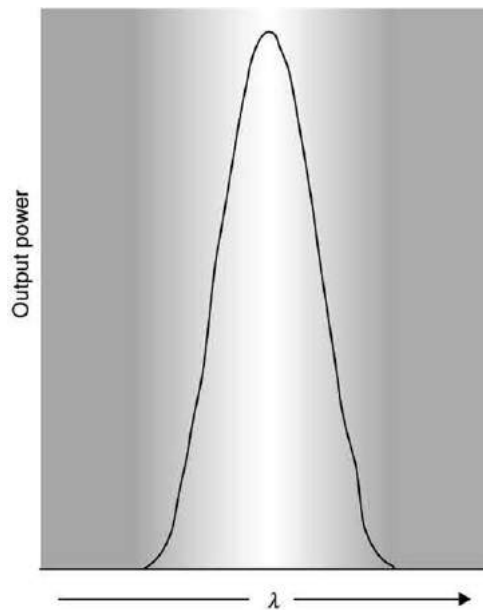


Fig. 13 The typical output of a DFB PWL. The linewidth can be made to be as narrow as a few hundred kHz.

Though their spectral properties are ideal as sources for long haul transmission, DFB lasers tend to be less powerful than the corresponding DBR designs. But the DBR PWLs are very sensitive to reflections that can cause a broadening of the laser linewidth or can act as a source of noise. In both cases, single mode lasers with very narrow linewidths in the kHz range with low radiance and noise, have been demonstrated. But neither is known to have been commercialized yet.

PWL Arrays

One of the unique advantages of the PWLs is that they are often processed using VLSI techniques. One important implication of this is the ability to make a PWL array where multiple lasers are formed simultaneously on the same chip, each with a unique output (**Fig. 14**). Such arrays are likely to be very useful in future dense wave division multiplexing (DWDM) optical network systems where as many as 128 channels (and more) are planned. Since packaging of individual lasers can be up to 75% of the total cost, making chips with multiple lasers is very attractive. The lasers may be DBRs or DFBs whose output wavelength is controlled by changing the grating parameters. To vary the individual wavelengths either the effective index and/or the pitch of the grating can be changed. Thus, any output spectrum can be designed and executed merely by making the appropriate lithography mask.

The gratings are often written holographically, where the wafer is exposed to an interference pattern caused by two coherent light beams. The wafer underneath can be sensitized with a photosensitive glass layer or a film a photoresist. It is not efficient to do multiple holographic exposures to individually change each of the gratings in the array. One way of doing this with a single holographic exposure is by constructing the waveguides at different relative angles. Then the component of the index modulation, resolved along the axis of the waveguide, would vary for each waveguide depending on the relative waveguide angle (**Fig. 15**); this would lead to differences in $\bar{n}$. Alternatively, the waveguide array can be made of the same height but of different widths. This is

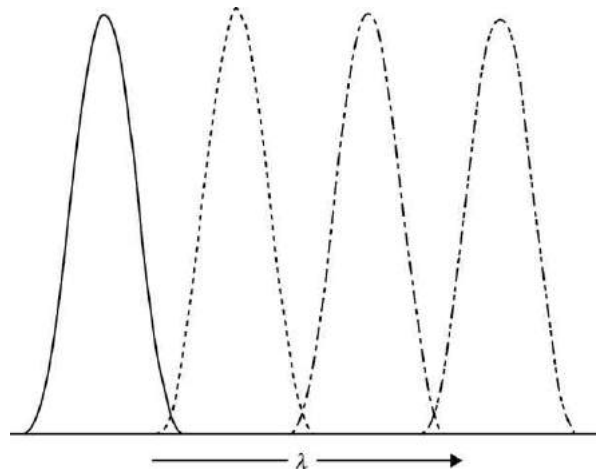


Fig. 14 Idealized output from a 4-PWL array.

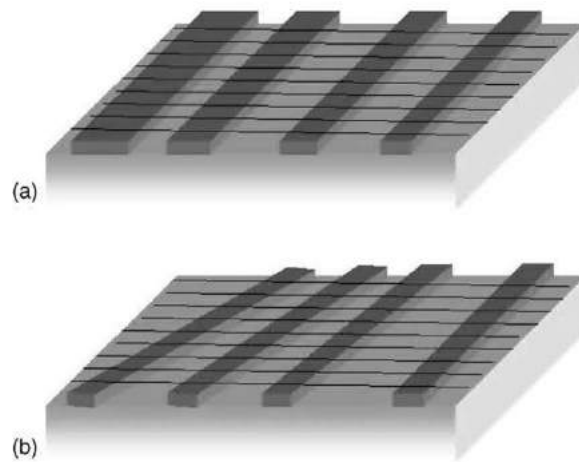


Fig. 15 Two different means of making planar waveguide laser array with a single holographic exposure to write the gratings. The darker rectangular blocks represent the core channels or ribs. In (a) the width and hence the effective index of the waveguide is varied whereas the relative angle between the waveguides is altered in (b). The lines represent the direction of the gratings, which is orthogonal only to the left most waveguide in (b).

easily done since waveguide height is usually deposited uniformly but the width is a function of the mask layout. Here the waveguide and the grating are orthogonal to each other and the variation in wavelengths is achieved by the fact that the effective propagation constant for each of the waveguides is different.

Arrays of PWLs have been made on a silicon substrate using deposition of an Er containing glass that is subsequently patterned into waveguide, and ultimately laser, arrays. Another approach has been to start with a phosphate glass as the host and the substrate. A phosphate glass is a superior host for rare-earth ions such as Er since the sensitization efficiency is nearly unity and much high doping levels are possible before concentration quenching effects set in. Waveguides are created by ion-exchange by immersion in a hot bath containing suitable ions (Na for Li, for instance). When starting with a uniformly doped glass, one issue in the assimilation of the laser array with other components is the fact that Er-doped region does not start and end with the waveguide laser section. Recently, the Er has been deposited selectively in the laser waveguide in phosphate glass using methods such as ion implantation, bringing the concept of integration closer to reality.

Pumping

One of the key issues here is the optical pumping of the rare-earth PWLs to generate acceptably high output power. Individual diode lasers or an array of diode lasers all on a single chip (called a diode bar) have been utilized as pumps. Both single and multimode diode lasers are available and have been used. There are pros and cons for either choice: whereas single mode lasers are less powerful (in the range of few 100 mW), it is difficult to convert the multimoded output of their counterparts to useful power.

The choice is determined by the waveguide design, which needs to be optimized for high overlap of the lasing mode with the pump. Carefully tapered waveguides can be used to distribute the pump power from a diode bar to the various lasers in an array without significant loss.

PWL Stability

Stable operation of a PWL is one of the big challenges. In DBR PWLs, it is seen that there is a threshold for pump power only below which single moded operation can be achieved. Additionally, any PWL die needs to be made immune to ambient temperature changes. These changes can affect many of the critical output parameters of the PWL: wavelength, linewidth, power and in the case of an array, the wavelength separation. This occurs primarily due to the thermal shift in the grating pitch as well as the thermo-optic effect. It is overcome by supplanting the laser die onto a heat spreader chip made from conducting material such as Cu or BeO, followed by a thermoelectric cooler. Ironically, the laser's output wavelength and linewidth are fine-tuned and 'fixed' using temperature to overcome any drifts due to processing errors or other reasons.

Reliability is another important issue that needs to be addressed, especially for telecommunication applications where these types of components have to pass high quality standards such as the Telecordia qualification in the USA.

Other PWL Systems

Many of the PWLs to date have been developed using Er doping, which emits light in the range of interest for telecommunications (the C-band and even the L-band). Here, Yb can be added as a co-dopant to improve power conversion efficiency. Other fluorescing ions have also been used: Nd, Tm, Ho, and Cr. Several research teams have successfully demonstrated Nd-YAG PWLs, in particular. Ridge waveguide lasers have been made using semiconductor materials InGaAs/InP heterojunctions in both Fabry-Perot and DFB configurations. These systems are electrically pumped. DFB PWLs have also been made using dye doped polymers.

It must be mentioned that there are other ways to use waveguides to make lasers. The primary example is a configuration where the waveguide resonator is external to the active region. Designs, such as ring resonator lasers and an external cavity lasers, fall in this category.

Future research and development of PWLs would undoubtedly utilize photonic crystal structures. These structures have a periodic fluctuation in the index of refraction in two or three dimensions. Owing to the resultant Bragg reflections, specific allowed and forbidden states are created which can be used to control the confinement of light. Hitherto, these structures have been used to form mirrors or hexagonal ring resonators. These more sophisticated laser designs are likely to enable new functions such as tunability of the output spectrum.

Summary

Planar waveguide lasers are a special class of laser where light is confined to a waveguide. They have distinctive advantages that include high optical gains, low laser thresholds, narrow linewidths in the kHz range, and optimal thermal management. Significantly, they afford a platform that can be readily expanded to include an array of lasers, each with a unique output, and can ultimately be combined with multiple optical components on the same chip. Such compact integrated devices are bound to take photonics to a new high level of capability.

Further Reading

- Carroll, J.E., Whiteaway, J.E.A., Plumb, R.G.S., Plumb, D., 1998. Distributed Feedback Semiconductor Lasers. London: IEE Publishing.
- Ghafouri-Shiraz, H., 2003. Principles of Semiconductor Laser Diodes and Amplifiers: Analysis and Transmission Line Laser. London: Imperial College.
- Hecht, J., 2001. Understanding Lasers: An Entry Level Guide. New York: Wiley Interscience.
- Lee, J.R., Baker, H.J., Friel, J.G., Hilton, G.J., Hall, D.R., 2002. High-average-power Nd:YAG planar waveguide laser, face-pumped by 10 laser diode bars. *Optical Letters* 27, 524–526.
- Milonni PW and Eberly JH (1988) *Lasers*. New York: John Wiley and Sons.
- Saleh, B.E.A., Teich, M.C., 1991. Fundamentals of Photonics. New York: John Wiley & Sons.
- Shepherd, D.P., Hettrick, S.J., Li, C., *et al.*, 2001. High-power planar dielectric waveguide lasers. *Journal of Physics D: Applied Physics* 34, 2420–2432.
- Siegman, A.E., 1986. Lasers. Harendon, Virginia: University Science Books.
- Svelto, O. (Ed.), 1998. The Principles of Lasers. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Veasey, D.L., Funk, D.S., Sanford, A., 1999. Arrays of distributed-Bragg-reflector waveguide lasers at 1536 nm in Yb/Er codoped phosphate glass. *Applied Physics Letters* 74 (6), 789–791.
- Young, M., 2001. Optics and Lasers: Including Fibers and Optical Waveguides. New York: Springer-Verlag New York, Inc.

Up-Conversion Lasers

A Brenier, University of Lyon, Villeurbanne, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Under photoluminescence excitation, a luminescent center located inside a transparent material usually emits at a longer wavelength than the excitation. This is because of the existence of de-excitation losses inside the material. Thus, efficient down-conversion lasers can be built, for example, the commercial $\text{Nd}^{3+}:\text{Y}_3\text{Al}_5\text{O}_{12}$ laser which works at 1064 nm, under usual pumping near 800 nm. More surprisingly, it has been demonstrated that fluorescence can also be emitted at shorter wavelength than the excitation. This is possible if the losses are reduced and overcome by so-called 'up conversion mechanisms'. Generally speaking, a laser emitting at a shorter wavelength than the excitation is classified as an up-conversion laser.

These lasers are able to generate coherent light in the red, green, blue, and uv spectral ranges. A point of practical importance is that some of them can be pumped in the near-infrared range (around 800 nm or 980 nm) by commercially available laser diodes. They can be based on all-solid-state technology and share its advantages: compactness of devices, low maintenance, and efficiency. They are needed for a variety of applications: color displays, high density optical data storage, laser printing, biotechnology, submarine communications, gas-laser replacement, etc.

Because the emitted photons have higher energy than the excitation ones, more than one pump photon is needed to generate a single laser photon, so the physical processes involved in the up-conversion laser are essentially nonlinear. Two kinds of laser devices can be considered. In the first one, the excitation of the luminescent center up-migrates through the electronic levels and reaches the initial laser level. This step is due to some luminescence mechanisms. Then, in a second step, lasing at a short wavelength occurs through a transition towards the final laser level. In the second kind of device, the pump excitation activates a down-conversion lasing, but the laser wave is maintained inside the cavity, cannot escape, and is up-converted in short wavelengths by the second-order nonlinear susceptibility of the crystal host. This requires bifunctional nonlinear optical and laser crystals. The device is thus a self-frequency doubling laser or a self-sum frequency mixing laser. If the nonlinearity is not an intrinsic property of the laser crystal but is due to a second nonlinear optical crystal located inside or outside the cavity, the device operates intracavity or extracavity up conversion.

Up-Conversion from Luminescence Mechanisms

The numerous $2S+1L_J$ energy levels ($4f^n$ configuration) of the trivalent rare earth ions, inserted in crystals or glasses, offer many possibilities for up conversion, particularly from long-lived intermediate levels that can be populated with infrared pumping. Crystal field induced transitions between them are comprised of sharp lines, because the $4f^n$ electrons are protected from external electric fields by the $5s^2p^6$ electrons. Their fluorescence spectra can exhibit cross-sections high enough (10^{-19} cm^2) to lead to laser operation in the visible range. For up-conversion purposes, the most popular ions are Er^{3+} , Pr^{3+} , Tm^{3+} , Ho^{3+} , and Nd^{3+} . Up conversion in Er^{3+} was used for the first time in 1959, to detect infrared radiation.

Energy Transfers

Two ions in close proximity interact electrostatically (Coulomb interaction) and can exchange energy. The ion which gives energy is called the sensitizer S or donor, and the energy receiving ion is the activator A or acceptor. Two examples discussed below are shown in Fig. 1. The most typical case of such nonradiative energy transfer in trivalent rare earth ions in insulating materials is due to electrical dipole-dipole interaction. For the latter, the transition probability $W(\text{s}^{-1})$ takes the form:

$$W = \frac{\hbar^4 c^4}{4\pi n^4 R^6} \frac{Q_A}{\tau_{\text{rad}}} \int \frac{f_A(E)f_S(E)}{E^4} dE \quad (1)$$

where R is the S-A distance, E is the photon energy of the transition operated by each ion, n the refractive index, and Q_A is the integrated absorption cross-section of the A-transition:

$$Q_A = \int \sigma_A(E) dE, \quad f_A = \frac{\sigma_A}{Q_A} \quad (2)$$

In Eq. (1), τ_{rad} is the radiative emission probability of S for the involved transition calculated from the emission probability A_S :

$$\frac{1}{\tau_{\text{rad}}} = \int A_S(E) dE, \quad f_S = \tau_{\text{rad}} A_S \quad (3)$$

The integral in Eq. (1) is called the overlap integral and shows that the transfer is efficient if the luminescence spectrum of S is resonant with the absorption spectrum of A (corresponding to the S and A transitions involved in the nonradiative transfer).

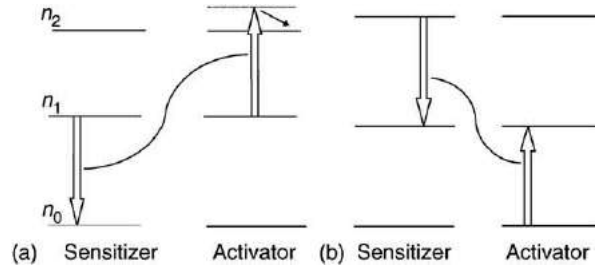
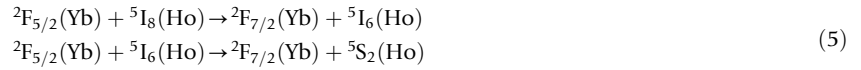
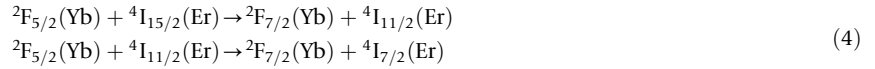


Fig. 1 (a) Up-conversion cross-relaxation energy transfer. (b) Down-conversion cross-relaxation energy transfer. The example in (a) is phonon assisted, the example in (b) is resonant.

The transitions are often not in perfect resonance but remain efficient enough to be used in practice. Then the energy mismatch is compensated by phonons, and the energy transfer is said to be phonon-assisted.

The energy transfer visualized in **Fig. 1(a)** is an up-conversion cross-relaxation. It is often met in the rare earth ions cited above when they are sufficiently concentrated in the material. The S and A ions can be of the same or of different chemical species. Successive energy transfers can promote the activator from its ground state towards higher energy levels, up to the one from which lasing can occur. This is the so-called 'Addition de Photons par Transferts d'Energie', sometimes appointed 'APTE'. Yb^{3+} is the most used ion as sensitizer and leads to the following transfers in examples of Er^{3+} or Ho^{3+} co-doping:



If several sensitizers give their energy simultaneously to a single activator, promoting it directly from its ground state towards a high energy level without intermediate levels, the sensitization is said to be 'cooperative'.

The energy transfer shown in **Fig. 1(b)** is a down-conversion cross-relaxation. It is the inverse of the previous one. At first sight it is, of course, undesirable in an up-conversion laser. However, we notice that one ion in the upper level will lead to two ions in the intermediate level. This fact can be exploited in special conditions (see below) and be beneficial for up conversion.

The dynamical study of the energy transfers, starting from the microscopic point of view contained in **Eq. (1)**, is very complicated because of the random locations of the ions or because of other processes, such as energy migration among sensitizers and back transfers. A popular method leading to useful predictions in practice is the rate equation analysis dealing with average parameters. The time evolution of the population densities n_i of the various levels i are described by first-order coupled equations, solved with adequate initial conditions and including the pumping rate. As an example, the rate equations for levels 1 and 2 in **Fig. 1(a)** take the form:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} - 2kn_1^2 + \frac{\beta_{21}}{\tau_2}n_2 + W \quad (6)$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + kn_1^2 \quad (7)$$

$$n_0 + n_1 + n_2 = 1 \quad (8)$$

where $\tau_{1,2}$ are the lifetimes of levels 1 and 2, k is the up-conversion transfer rate, β_{21} is the branching ratio for the $2 \rightarrow 1$ transition, and W is the pump rate (not shown in **Fig. 1**).

Sequential Two-Photon Absorption

Pumping the ground state of a luminescent ion (**Fig. 2(a)**) with a wave at angular frequency ω , leads to population with density n_1 of an excited state in a first step. Because the rare earth ions have numerous energy levels, it is possible for the pump to also be resonant with the transition from level 1 towards a higher energy level 2, and is further absorbed. Such a process leading to n_2 up-conversion population is a sequential two-photon absorption. If there is no $1 \rightarrow 2$ resonant transition for the ω frequency, it is possible to use a second pump wave at a different angular frequency ω' resonant with the $1 \rightarrow 2$ transition (**Fig. 2(b)**). Of course, in practice, it is advantageous when a single pump beam is required for up conversion.

Let us show the relevant terms of the rate equation analysis describing the process in **Fig. 2(a)**:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0n_0 - I\sigma_1n_1 + \frac{\beta_{21}}{\tau_2}n_2 \quad (9)$$

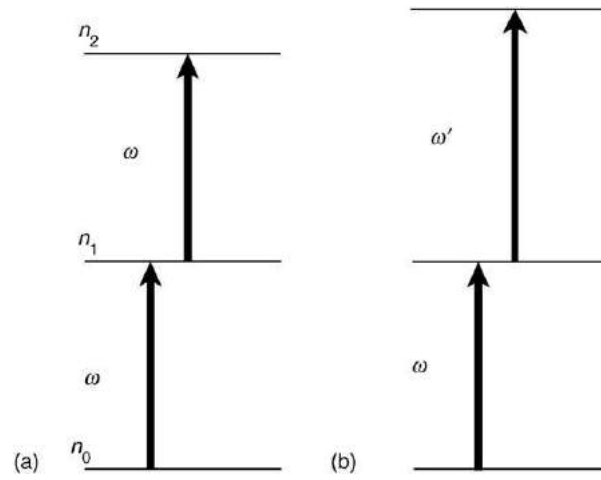


Fig. 2 Two-photon absorption up-conversion.

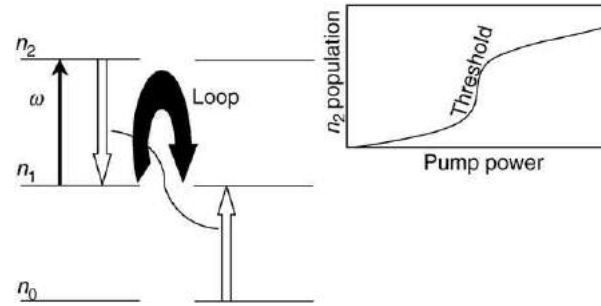


Fig. 3 Photon avalanche up-conversion.

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 \quad (10)$$

$$n_0 + n_1 + n_2 = 1 \quad (11)$$

where I is the photon density of the pump (photons $\text{s}^{-1} \text{cm}^{-2}$) and σ_0 and σ_1 are the absorption cross-sections (cm^2) of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. The n_2 population density shows a nonlinear I dependence, most often close to I^2 .

Photon Avalanche

In rare earth doped materials, an up-conversion emission, intense enough for visible lasing, can result from a more complex mechanism. In this case, absorption of the pump is not resonant with a transition from the ground state (weak ground-state absorption), but on the contrary, the photon pump energy matches the gap between two excited states (high excited state absorption). These levels are labeled 1 and 2 in **Fig. 3**. Level 1 generally has a rather long lifetime in order to accumulate population and level 2 is the initial up-conversion laser level. When the pump light is turned on, the populations of levels 1 and 2 first grow slowly, and then can accelerate exponentially. This is generally referred to as 'photon avalanche'. The term 'looping mechanism' also applies, because we can see in **Fig. 3** that the excited state absorption $1 \rightarrow 2$ is followed by a cross-relaxation energy transfer (quantum efficiency up to two) that returns the system back to level 1. So after one loop, the n_1 population density has been amplified, and is further increased after successive loops. This pumping scheme was discovered in 1979, with $\text{LaCl}_3:\text{Pr}^{3+}$ crystal used as an infrared photon counter.

The rate equation analysis applied to the three-level system in **Fig. 3** leads to the following time evolutions:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 + 2kn_0 n_2 \quad (12)$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 - kn_0 n_2 \quad (13)$$

$$n_0 + n_1 + n_2 = 1 \quad (14)$$

where the parameters have the same meaning as in **Eqs. (6)–(11)**.

The steady state limit of the populations at infinite time, upon continuous excitation starting at $t=0$, can be obtained from canceling the time derivative in the left-hand side of the equation. Let us point out the interesting result. Two different dynamical regimes and analytical expressions for $n_2(\infty)$ are obtained. In the usual case of low pumping rate in the ground state and if the parameters of the systems satisfy: $k < \frac{1-\beta_{21}}{\tau_2}$, then the first regime is observed, whatever the pumping rate $I\sigma_1$ in the excited state.

If the parameters satisfy:

$$k > \frac{1-\beta_{21}}{\tau_2} \quad (15)$$

the first regime is still observed at low pumping rate $I\sigma_1$ in the excited state, more precisely at pumping rate low enough such that

$$I\sigma_1 < \frac{\left(k + \frac{1}{\tau_2}\right) \frac{1}{\tau_1}}{k - \frac{(1-\beta_{21})}{\tau_2}}$$

But interestingly, the second regime, the so-called photon avalanche, is obtained at high pumping rate $I\sigma_1$ in the excited state, that is to say if:

$$I\sigma_1 > \frac{\left(k + \frac{1}{\tau_2}\right) \frac{1}{\tau_1}}{k - \frac{(1-\beta_{21})}{\tau_2}} \quad (16)$$

So the right-hand side of Eq. (16) appears to be a pump threshold separating two dynamical regimes when the condition in Eq. (15) is satisfied, as the insert in Fig. 3 shows.

Another way to distinguish between the two regimes is given by linearization of the system of Eqs. (12)–(14) ($n_0 \cong 1$). Then the description will only be valid close to $t=0$. We know from standard mathematical methods that the solutions of Eqs. (12)–(14) are then a combination of $\exp(\alpha t)$ terms. If the threshold condition in Eq. (16) is satisfied, a parameter α in the combination becomes positive, so the population of level 2 increases exponentially (photon avalanche) at early times and the dynamics of the system is unstable.

The physical meaning of Eqs. (15) and (16) is clarified by introducing the yield η_{CR} of the cross-relaxation process:

$$\eta_{CR} = \frac{k}{k + 1/\tau_2} \quad (17)$$

and the yield η of the $2 \rightarrow 1$ de-excitation:

$$\eta = \frac{b_{21}/\tau_2}{k + 1/\tau_2} \quad (18)$$

Then, the threshold condition in Eq. (16) can be written as:

$$\left(\frac{I\sigma_1}{I\sigma_1 + 1/\tau_1}\right)(2\eta_{CR} + \eta) > 1 \quad (19)$$

Starting with one ion in level 1, the first parenthesis in Eq. (19) is the yield with which it will be promoted to level 2, and the second parenthesis is the yield of the backward path $2 \rightarrow 1$. So, the left-hand side of Eq. (19) is the new number of ions after a looping path $1 \rightarrow 2 \rightarrow 1$, which has to be higher than 1 for the avalanche to occur. If we consider a lot of successive loops, the origin of the exponential behavior is clear. The role of the first ion in level 1 (and the weak pumping rate in the ground state $I\sigma_0$) is limited to initialize the process. Note that the value of the first parenthesis in Eq. (19) is less than 1, so the second parenthesis has to be higher than 1. This condition is equivalent to Eq. (15).

Materials: The Energy Gap Law

As we can see from the above mechanisms, up conversion needs intermediate energy levels and is not efficient if these cannot accumulate populations due to a short lifetime. The spontaneous rate of de-excitation of an electronic level is the sum of a radiative part W^{rad} and of a nonradiative part W^{nrad} . The radiative part can be obtained through the absorption cross-section corresponding to the transition or from the Judd–Ofelt theory which also exploits the absorption spectrum. The nonradiative rate is then obtained by subtracting W^{rad} from the measured fluorescence decay rate of the level. Based on the measurements in many hosts for the different trivalent rare earth ions, at different temperatures, the nonradiative decay rate of a J-level towards the next lower J'-level was found to have the form:

$$W_{J'J}^{\text{nrad}} = B \exp(-\beta \Delta E_{J'J}) \left(1 - \exp\left(\frac{-\omega_{\text{max}}}{kT}\right)\right)^{-p} \quad (20)$$

where $\Delta E_{J'J}$ is the energy gap between the two levels, $p = \Delta E_{J'J}/\omega_{\text{max}}$, ω_{max} being the energy of an effective optical phonon of the host. The values of B , β and ω_{max} are found by fitting with experimental data. An example of such a fit is represented in Fig. 4.

Eq. (20) is the well-known and familiar 'energy gap law'. We have represented it for several oxide and nonoxide hosts at $T=300$ K in Fig. 5. It is clear that chloride, bromide, fluoride will be favorable hosts for up conversion due to their low phonon energy (LaBr₃: $\omega_{\text{max}}=175$ cm⁻¹, LiYF₄: $\omega_{\text{max}}=400$ cm⁻¹, Y₃Al₅O₁₂: $\omega_{\text{max}}=700$ cm⁻¹, silicate glass: $\omega_{\text{max}}=1100$ cm⁻¹). Let us mention also a fluorozirconate glass, ZBLAN, widely used in up-conversion fiber lasers.

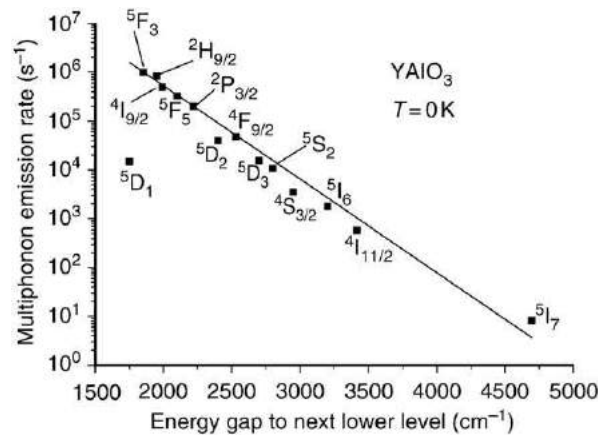


Fig. 4 Energy-gap dependence of nonradiative decay rates in YAlO_3 .

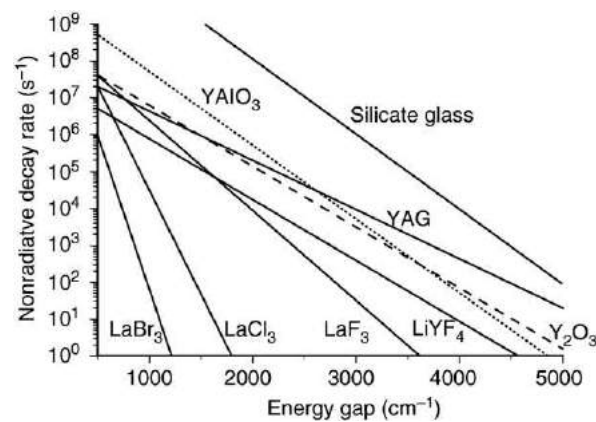


Fig. 5 Energy-gap law for different hosts at 300 K.

Table 1 Relevant terms for frequency up-conversion

Optical process	Term	New wave
Linear		
Refractive index, absorption, stimulated emission	$\chi^{(1)}(-\omega; \omega)$	
Second order nonlinear		
Second harmonic generation	$\chi^{(2)}(-\omega_3; \omega_1, \omega_1)$	$\omega_3 = 2\omega_1$
Sum-frequency mixing	$\chi^{(2)}(-\omega_3; \omega_1, \pm \omega_2)$	$\omega_3 = \omega_1 + \omega_2$

Up-Conversion from Second-Order Optical Nonlinearity

The inter-atomic electric field acting on the electrons inside a medium has a magnitude of $\sim 10^8$ V/cm and derives from a potential that is anharmonic. The electric field E of an electromagnetic wave propagating through the medium drives the electrons beyond the quadratic region of the potential in the case of high E . So, the electron response and the associated polarization P take the form of a function of E :

$$P(E) = \epsilon_0(\chi^{(1)} : E + \chi^{(2)} : E^2 + \chi^{(3)} : E^3 + \dots) \quad (21)$$

The different terms in Eq. (21) are responsible for many optical phenomena because in Maxwell electromagnetic theory, the polarization is the source of the waves. We have selected in Table 1 the terms relevant for linear effects and for the purpose of frequency up conversions: second-harmonic generation (SHG) and sum-frequency mixing, due to second-order terms. It should be noticed that the latter have a nonzero value only in noncentrosymmetric materials.

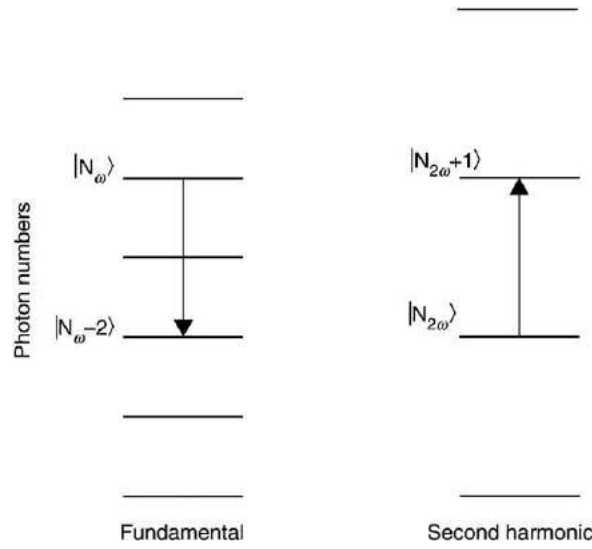


Fig. 6 Up-conversion from second harmonic generation.

Fig. 6 visualizes the up conversion from SHG: two photons of the fundamental field at angular frequency ω_1 are annihilated while one photon at twice the angular frequency is created. A similar picture could be drawn for sum-frequency up conversion: SHG is the degenerate case where $\omega_1 = \omega_2$.

The second-order nonlinear processes have an efficiency given by an effective coefficient d_{eff} given hereafter:

$$d_{\text{eff}} = \sum_{ijk} e_i(\omega_3) d_{ijk} e_j(\omega_1) e_k(\omega_2) \quad (22)$$

where $e_i(\omega_l)$ is the i th component of the unit vector of the electric field of the wave l at angular frequency ω_l and $d_{ijk} = (1/2)\chi_{ijk}$.

Moreover, the wave propagation equations in the slow varying envelope approximation, given hereafter for the E_3 electric field of the sum-frequency mixing wave:

$$\frac{dE_3}{dz} = \frac{i\omega_3 d_{\text{eff}}}{n_3 c} E_1 E_2 \exp(i(k_1 + k_2 - k_3)z) \text{ (with } k_j = \omega_j n_j / c) \quad (23)$$

impose that the frequency conversion is efficient in practice, only if the following phase matching condition is satisfied:

$$\omega_1 n_{1\uparrow}(\theta, \phi) + \omega_2 n_{2\uparrow}(\theta, \phi) = \omega_3 n_{3\downarrow}(\theta, \phi) \quad (24)$$

where $n_{i\uparrow}$ and $n_{i\downarrow}$ are respectively the upper and lower refractive index in the direction of propagation (θ, ϕ) (n_i in [Eq. \(23\)](#) are the same with a simplified notation). [Eq. \(24\)](#) is restricted here for simplicity to collinear type I (waves 1 and 2 and the same polarization) and is the expression of photon momentum conservation. It is usually achieved from the birefringence of the crystals. When this is not possible, another technique can be used, namely quasi-phase matching, which involves reversing periodically the sign of the nonlinear optical coefficient.

In bifunctional crystals, the laser effect and the $\chi^{(2)}$ interaction occur simultaneously inside the same crystal. In the case of the self-frequency doubling laser, the angular frequency $\omega_1 = \omega_2$ in [Eq. \(24\)](#) is that of the fundamental laser wave. In the case of the self-sum frequency mixing laser, ω_1 corresponds to the fundamental laser wave and ω_2 to the pump wave. As we can see from [Eqs. \(22\)–\(24\)](#), the polarization of the laser emission is crucial in order for the device to work. So, the laser stimulated emission cross-section σ for propagation in the phase matching direction (θ, ϕ) in the adequate polarization has to be evaluated. This can be performed with the formula:

$$\sigma(\theta, \phi) = \frac{\sigma_X \sigma_Y \sigma_Z}{\sqrt{\sigma_Y^2 \sigma_Z^2 e_X^2 + \sigma_X^2 \sigma_Z^2 e_Y^2 + \sigma_X^2 \sigma_Y^2 e_Z^2}} \quad (25)$$

where $\sigma_{X, Y, Z}$ are the emission cross-sections for X, Y, Z -polarizations and $e_{X, Y, Z}$ are the components of the unit vector or the electric field of the laser wave.

To summarize, the conditions necessary for a crystal to be a bi-functional material are: (i) it must be noncentrosymmetric; (ii) it must accept fluorescent doping (usually with Nd^{3+} or Yb^{3+}); and (iii) it must be phase matchable for its laser emission. In practice, only a few crystals satisfy these requirements and have been tried with some success. The main ones are LiNbO_3 , LaBGeO_5 , $\text{Ba}_2\text{NaNb}_5\text{O}_{15}$, β' - $\text{Gd}_2(\text{MoO}_4)_3$, $\text{YAl}_3(\text{BO}_3)_4$, $\text{GdAl}_3(\text{BO}_3)_4$, $\text{CaY}_4(\text{BO}_3)_3\text{O}$, and $\text{CaGd}_4(\text{BO}_3)_3\text{O}$.

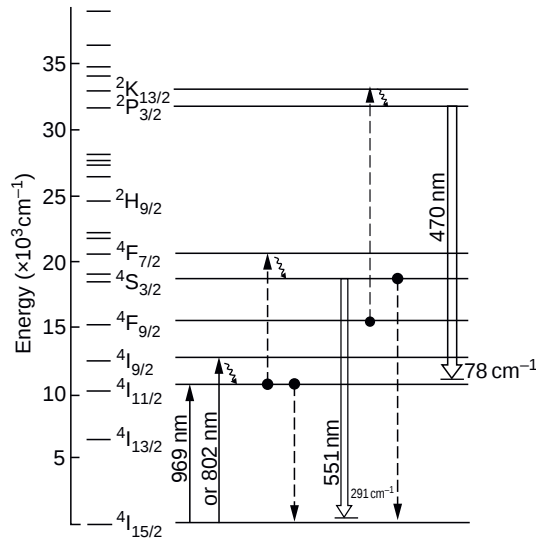


Fig. 7 Energy level diagram of $\text{LiYF}_4:\text{Er}^{3+}$. The dashed lines indicate the energy transfers responsible for the green and blue up-conversion lasing. Reproduced with permission of American Institute of Physics from Macfarlane RM, Tong F, Silversmith AJ and Lenth W (1988) Violet CW neodymium upconversion lasers. *Applied Physics Letters* 52(16): 1300.

Up-Conversion Lasers Based on Luminescence Mechanisms in Materials

Energy Transfers

Up-conversion lasers from energy transfers have the advantage of using a single pump laser. This latter populates firstly a long-lived intermediate level. A typical example, based on $\text{LiYF}_4:\text{Er}^{3+}$ crystal, is provided by Fig. 7.

The pump at 802 nm from a GaAlAs diode laser, or at 969 nm, populates efficiently at the 8 ms lifetime $^4\text{I}_{11/2}$ level. The green laser, at 551 nm, corresponds to the $^4\text{S}_{3/2} \rightarrow ^4\text{I}_{15/2}$ transition and it can be shown that energy transfer is an essential part of the mechanism by stopping abruptly the pump laser: green lasing continues to occur for several hundred microseconds because the $^4\text{S}_{3/2}$ level is still fed (an excited-state absorption would stop immediately). The involved energy transfer is:



and leads to a 20 mW laser threshold and 2.3 mW green output power at 50 K temperature upon 95 mW of incident power.

By using mirrors with high transmission in the green range but having high reflectivity at blue wavelengths, blue laser emission at 470 nm can be sustained, corresponding to the $^2\text{P}_{3/2} \rightarrow ^4\text{I}_{11/2}$ transition. It originates from a mechanism involving three up-conversion energy transfers:



The illustration in Fig. 7 shows up-conversion energy transfers between ions of the same chemical species, but co-doping the laser host with two ions of different chemical species provides the opportunity of separating their roles: one species is devoted to pump absorption (this is the sensitizer) and the other one is devoted to lasing (this is the activator). Due to the development of efficient semiconductor InGaAs diodes emitting near 980 nm, the most widely used sensitizer is Yb^{3+} . Table 2 provides a list of several up-conversion lasers based on energy transfer mechanisms, operating at room temperature. We can see that among the most efficient lasers, are those based on rare-earth ions doped glass fibers. The reason is that in fibers, the pump and laser waves remain confined inside a core (typically 5 μm diameter) and have high energy densities over a long length (tenths of cms), which is favorable for all up-conversion mechanisms.

Sequential Two-Photon Absorption

An example of an up-conversion laser pumped by a two-photon absorption mechanism is represented in Fig. 8. The material is $\text{LaF}_3:\text{Nd}^{3+}$ and violet lasing at 380 nm corresponds to the $^4\text{D}_{3/2} \rightarrow ^4\text{I}_{11/2}$ transition.

Transition from the ground state up to the $^4\text{F}_{5/2}$ level is firstly obtained from an infrared beam at 790 nm in order to feed the $^4\text{F}_{3/2}$ intermediate level after a fast nonradiative de-excitation. The $^4\text{F}_{3/2}$ level has a rather long lifetime: 700 μs , so it accumulates

Table 2 Main room temperature up-conversion lasers operating from energy transfers

Laser material	Laser wavelength (nm)	Pump wavelength (nm)	Output power
BaY _{1.4} Yb _{0.59} Ho _{0.01} F ₈	670	1540 + 1054	1%
BaYYb _{0.998} Tm _{0.002} F ₈	649	1054	
BaYYb _{0.99} Tm _{0.01} F ₈	799–649–510–455	960	
BaYYb _{0.99} Tm _{0.01} F ₈	348	960	37 mW
LiYF ₄ :Er(1%):Yb(3%)	551	966	
LiY _{0.89} Yb _{0.1} Tm _{0.01} F ₄	810–792	969	
LiY _{0.89} Yb _{0.1} Tm _{0.01} F ₄	650	969	80 mW
LiKYF ₅ :Er (1%)	550	808	5 mW
Fiber:Yb:Pr	635	849	150 mW
Fiber:Yb:Pr	635	1016	20 mW
Fiber:Yb:Pr	521	833	6.2 mW
Fiber:Yb:Pr	635	860	0.7 mW
Fiber:Yb:Pr	635	860	4 mW

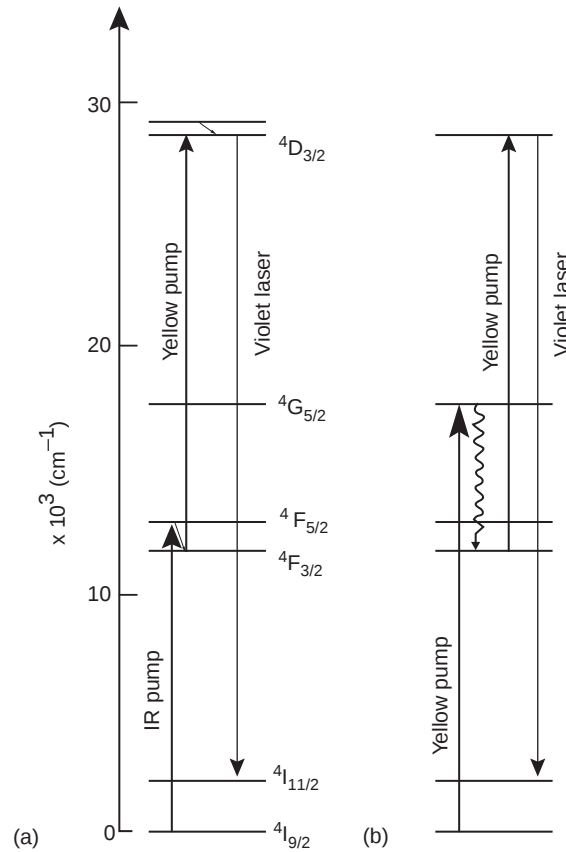


Fig. 8 Simplified energy level diagram of LaF₃:Nd³⁺ and up-conversion lasing from two-photon absorption. (a) Two different pump wavelengths, (b) doubly resonant single pump wavelength. Reproduced with permission of American Institute of Physics from Lenth W, Silversmith AJ and Macfarlane RM (1988) Green infrared pumped erbium up conversion lasers. In: TAM AC, Gole JL and Stwalley WC (eds) *Advances in Laser Science – III*. AIP Conference Proceedings n°172: 8–12.

enough population that a second yellow pumping at 591 nm (Fig. 8(a)) can efficiently populate the $4D_{3/2}$ initial laser level. With 110 mW of infrared pump and 300 mW of yellow pump, 12 mW of violet output were obtained at 20 K temperature.

A simpler device is obtained if a single-pump beam is used. This is the case in Fig. 8(b), where a yellow pump at 578 nm provides both ground state and excited state absorption. This doubly resonant scheme is less efficient in LaF₃:Nd³⁺ than the previous one but in Fig. 9, we show another example: LiYF₄:Er³⁺, working at room temperature. The lasing $4S_{3/2} \rightarrow 4I_{15/2}$ transition at 551 nm delivers 20 mW upon 1600 mW pumping at 974 nm.

The simplified Er³⁺ level scheme of Fig. 9 contains four main levels with population densities: $n_0 = [4I_{15/2}]$, $n_1 = [4I_{13/2}]$, $n_2 = [4I_{11/2}]$, $n_3 = [4S_{3/2}]$. The rate equations of populations of this system can be written and solved because all the implied parameters (lifetimes, branching ratios, ground, and excited states absorption cross-sections) have been measured in this material.

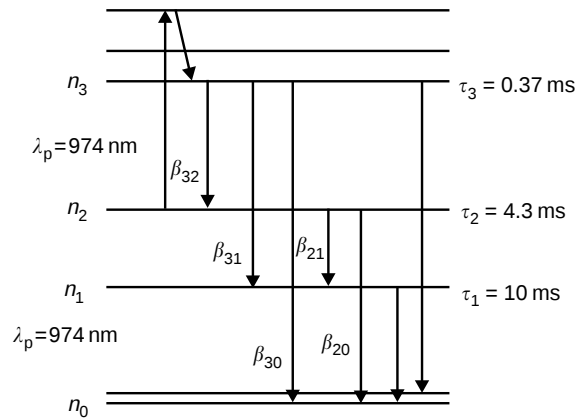


Fig. 9 Simplified energy level diagram of $\text{LiYF}_4:\text{Er}^{3+}$ and up-conversion lasing from two-photon absorption (doubly resonant). Reproduced with permission of Institute of Physics Publishing from Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*, Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing, 19.

Table 3 Main room temperature up-conversion lasers operating from two-photon absorption

Laser material	Laser wavelength (nm)	Pump wavelength (nm)	Output power
$\text{Y}_3\text{Al}_5\text{O}_{12}:\text{Er}$ 1%	561	647 + 810	
$\text{LiYF}_4:\text{Er}$ 1%	551	647 + 810	0.95 mJ
$\text{LiYF}_4:\text{Er}$ 1%	551	810	40 mW
$\text{LiYF}_4:\text{Er}$ 1%	551	974	45 mW
$\text{LiYF}_4:\text{Er}$	551	974	20 mW
$\text{KYF}_4:\text{Er}$ 1%	562	647 + 810	0.95 mJ
$\text{LiLuF}_4:\text{Er}$	552	974	70 mW
$\text{LiLuF}_4:\text{Er}$	552	970	213 mW
$\text{LiYF}_4:\text{Tm}$ 1%	453–450	781 + 649	0.2 mJ
Fiber:Tm	480	1120	57 mW
Fiber:Tm	480	1100–1180	45 mW
Fiber:Tm	455	645 + 1064	3 mW
Fiber:Tm	803–816	1064	1.2 W
Fiber:Tm	482	1123	120 mW
Fiber:Yb:Tm	650	1120	
Fiber:Ho	540–553	643	38 mW
Fiber:Ho	544–549	645	20 mW
Fiber:Er	548	800	15 mW
Fiber:Er	546	801	23 mW
Fiber:Er	544	971	12 mW
Fiber:Er	543	800	
Fiber:Pr	635	1010 + 835	180 mW
Fiber:Pr	605	1010 + 835	30 mW
Fiber:Pr	635	1020 + 840	54 mW
Fiber:Pr	520	1020 + 840	20 mW
Fiber:Pr	491	1020 + 840	7 mW
Fiber:Nd	381	590	74 μW
Fiber:Nd	412	590	500 μW

Predictive models of the up-conversion green laser are successful, matching experimental measurements, and confirm that the doubly resonant mechanism is realistic at low doping concentration.

Table 3 summarizes the performances of the main up-conversion lasers based on sequential photon absorption and working at room temperature.

Photon Avalanche

Photon avalanche up-conversion was observed mainly in Nd^{3+} , Tm^{3+} , Er^{3+} , and Pr^{3+} doped materials. As an illustration of this mechanism, let us consider the case of the continuous wave $\text{Pr}^{3+}-\text{Yb}^{3+}$ -doped fluorozirconate fiber laser (3000 ppm Pr,

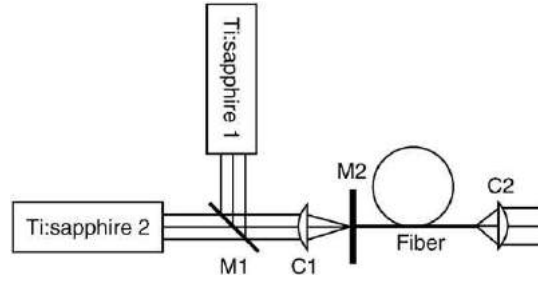


Fig. 10 Experimental set-up for Yb-Pr-doped fiber up-conversion laser. M1: dichroic mirror, C1, C2: lenses, M2 dichroic input mirror. Reproduced with permission of Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

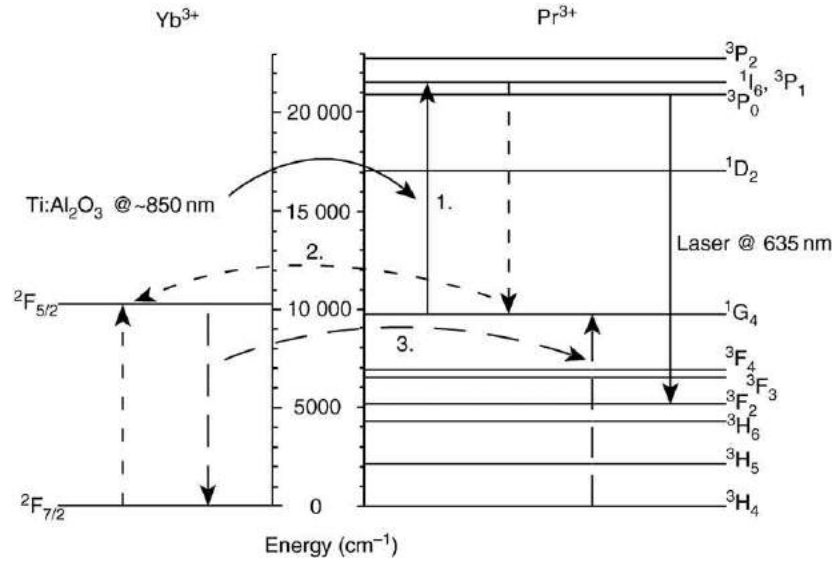


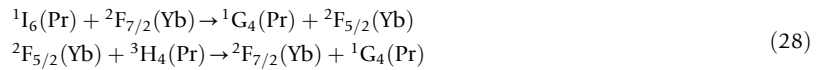
Fig. 11 Energy level diagram of fluorozirconate: Yb³⁺:Pr³⁺ fiber. The dashed lines indicate the energy transfers. Reproduced with permission of the Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

20 000 ppm Yb). With two Ti:sapphire pump lasers operating at 852 and 826 nm (Fig. 10), an output power at 635 nm as high as 1.02 W was obtained with an incident pump power of 5.51 W (19% slope efficiency).

This remarkable result can be explained quantitatively by modeling the photon avalanche with a five-level diagram and restricted in Fig. 11 to a single laser pump at 850 nm. The laser emission corresponds to the $^3P_0 \rightarrow ^3F_2$ transition. The five relevant levels have the population densities: $n_0 = [^3H_4(\text{Pr})]$, $n_1 = [^1G_4(\text{Pr})]$, $n_2 = [^3P_0(\text{Pr})]$, $n_3 = [^2F_{7/2}(\text{Yb})]$, $n_4 = [^2F_{5/2}(\text{Yb})]$.

The two components of the mechanism are:

- (i) the excited state absorption of the pump by the spin allowed transition: $^1G_4 \rightarrow ^1I_6$
- (ii) the process leading to the doubling of the 1G_4 population, in this instance through two energy transfers:



The rate equation analysis predicts that the avalanche threshold occurs at about 1 W pump and the long fluorescence rise time is somewhat shortened at higher pump power, as observed experimentally.

Table 4 summarizes the performances of the main up-conversion lasers based on photon avalanche mechanisms.

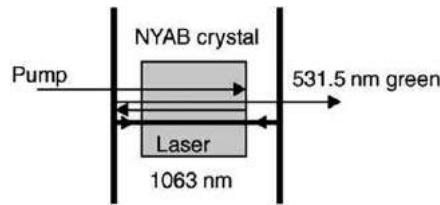
Up-Conversion Lasers Based on Bi-Functional Crystals

The laser stimulated emission in a bi-functional crystal was first obtained in 1969 with Tm³⁺ in LiNbO₃ but the most used ion is, of course, Nd³⁺ working in the two channels:



Table 4 Main room temperature up-conversion lasers operating from photon avalanche

Laser material	Laser wavelength (nm)	Pump wavelength (nm)	Output power
BaY ₂ F ₈ :Yb:Pr	607.5	822	55 mW
BaY ₂ F ₈ :Yb:Pr	638.7	841	26 mW
LiY _{0.89} Yb _{0.1} Pr _{0.01} F ₄	720	830	1%
LiY _{0.89} Yb _{0.1} Pr _{0.01} F ₄	639.5	830	
Fiber:Yb:Pr	635–637	780–880	300 mW
Fiber:Yb:Pr	605–622	780–880	44 mW
Fiber:Yb:Pr	517–540	780–820	20 mW
Fiber:Yb:Pr	491–493	780–880	4 mW
Fiber:Yb:Pr	635	850	1.02 W
Fiber:Yb:Pr	635	850	2.06 W
Fiber:Yb:Pr	520	850	0.3 W

**Fig. 12** Scheme of a self-frequency doubling laser. Note that in reality the three beams are superimposed.

$$^4F_{3/2} \rightarrow ^4I_{13/2} \quad (30)$$

and more recently Yb³⁺ working in the channel:

$$^2F_{5/2} \rightarrow ^2F_{7/2} \quad (31)$$

Channels in Eqs. (29) and (31) operate around 1060 nm wavelength and the channel in Eq. (30) near 1338 nm.

The Self-Frequency Doubling Laser

A typical laser scheme is shown in Fig. 12.

The laser beam cannot escape from the cavity and dichroic mirrors are used. The input mirror has high transmission at the pump wavelength and is highly reflective at laser- and second-harmonic wavelengths. The output mirror is highly reflective at laser wavelength and has high transmission for second harmonic. Channels in Eqs. (29) and (31) generate green light near 530 nm and the channel in Eq. (30) generates red light near 669 nm.

The most exploited bi-functional crystal is YAl₃(BO₃)₄:Nd³⁺, well-known as NYAB. It is a negative uniaxial trigonal crystal (extraordinary index lower than the ordinary one) and its laser emission is easily observed in ordinary polarization with a high cross-section: 2×10^{-19} cm² at 1063 nm. Type I phase matching occurs at a polar angle $\theta = 30.7^\circ$ and 26.8° for channels in Eqs. (29) and (30) respectively. The effective nonlinear optical coefficient d_{eff} is close to 1.4 pm V^{-1} at azimuthal angle $\phi = 0^\circ$.

The Yb³⁺ ion has some advantages over the Nd³⁺ ion due to its very simple energy level scheme: there is no excited state absorption at the laser wavelength, no up-conversion losses, no concentration quenching, and no absorption in the green. The small Stokes shift between pump and laser emission reduces the thermal loading of the material during laser operation. A self-doubling laser based on YAl₃(BO₃)₄:Yb³⁺ of crystal has produced 1.1 W green power upon 11 W diode pumping.

The drawback of YAl₃(BO₃)₄ is that it is rather difficult to grow because it is not congruent. This difficulty was overcome by the discovery at Ecole Nationale Supérieure de Chimie de Paris of the rare-earth calcium oxoborate CaGd₄(BO₃)₃O which can be grown to a large size by the Czochralski method. It is a monoclinic biaxial crystal and doped with Nd³⁺, it was firstly exploited (channel in Eq. (29)) in the XY principal plane at $\theta = 90^\circ$, $\phi = 46^\circ$ with ($d_{\text{eff}} = 0.5 \text{ pm V}^{-1}$). It was soon recognized that the optimum phase matching direction ($d_{\text{eff}} = 1.68 \text{ pm V}^{-1}$) occurred out of the principal planes, in the direction $\theta = 66.8^\circ$, $\phi = 132.6^\circ$, and high green power can be obtained: 225 mW under 1.56 W pump.

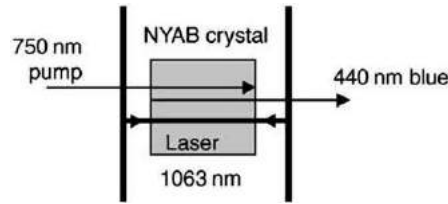
Table 5 summarizes the main self-frequency doubling results obtained with the channel in Eq. (29).

The Self-Sum Frequency Mixing Laser

The example in Fig. 13 gives the main features of such a laser. The input mirror has high transmission at the pump wavelength and both mirrors are highly reflective at laser wavelength. The output mirror has high transmission at sum frequency mixing wavelength.

Table 5 Main results of Nd^{3+} and Yb^{3+} green self-frequency doubling lasers near 530 nm (Eqs. (29) and (31))

Crystal	Input power (mW)	Output power (mW)	Pumping
$\text{LiNbO}_3:\text{MgO}:\text{Nd}$	215	1	Dye laser
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	870	10	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	280	3	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	400	69	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	1380	51	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	369	35	Laser diode
$\text{LiNbO}_3:\text{MgO}:\text{Nd}$	850	18	Dye laser
$\text{LiNbO}_3:\text{Sc}_2\text{O}_3:\text{Nd}$	65	0.14	Ti:sapphire laser
$\text{LiNbO}_3:\text{MgO}:\text{Nd}$	100	0.2	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	1600	225	Laser diode
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	2200	450	Ti:sapphire laser
$(\text{Y},\text{Lu})\text{Al}_3(\text{BO}_3)_4:\text{Nd}$	880	24	Laser diode
$\text{LiNbO}_3:\text{ZnO}:\text{Nd}$	430	0.65	Ti:sapphire laser
$\text{Ba}_2\text{NaNb}_5\text{O}_{15}:\text{Nd}$	270	46	Ti:sapphire laser
$\text{CaY}_4(\text{BO}_3)_3\text{O}:\text{Nd}$	900	62	Laser diode
$\text{CaGd}_4(\text{BO}_3)_3\text{O}:\text{Nd}$	1600	192	Ti:sapphire laser
$\text{CaGd}_4(\text{BO}_3)_3\text{O}:\text{Nd}$	1250	115	Laser diode
$\text{CaGd}_4(\text{BO}_3)_3\text{O}:\text{Nd}$	1560	225	Ti:sapphire laser
$\text{YAl}_3(\text{BO}_3)_4:\text{Yb}$	11 000	1100	Laser diode
$\text{GdAl}_3(\text{BO}_3)_4:\text{Nd}$	2.8 mJ/pulse	0.12 mJ/pulse	Pulsed dye laser

**Fig. 13** Scheme of a self-sum frequency mixing laser. Note that in reality the three beams are superimposed.**Table 6** Main results of Nd^{3+} self-sum frequency mixing lasers based on Eq. (29)

Crystal	Pump wavelength (nm)	Generated wavelength (nm)	Output power
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	740–760	436–443	0.16 mJ/pulse
$\text{GdAl}_3(\text{BO}_3)_4:\text{Nd}$	740–760	436–443	0.403 mJ/pulse
$\text{CaGd}_4(\text{BO}_3)_3\text{O}:\text{Nd}$	811	465	1 mW
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	585–600	377–383	0.25 mJ/pulse
$\text{GdAl}_3(\text{BO}_3)_4:\text{Nd}$	587–597	378–382	0.105 mJ/pulse
$\text{YAl}_3(\text{BO}_3)_4:\text{Nd}$	488, 515	330, 380	0.2 mW

The laser wave at angular frequency ω_1 in Table 1 corresponds to the channel in Eqs. (29) or (30) in the case of Nd^{3+} . The wave at angular frequency ω_2 in Table 1 has two different roles: first it excites the Nd^{3+} laser center and secondly its nonabsorbed part is up converted by the second-order nonlinear process. ω_2 (corresponding to the wavelength λ_2) is then chosen to match the main Nd^{3+} absorption lines:

$$\begin{aligned}
 &^4\text{I}_{9/2} \rightarrow ^4\text{G}_{5/2} - ^2\text{G}_{7/2} (\lambda_2 = 590 \text{ nm}) \\
 &^4\text{I}_{9/2} \rightarrow ^4\text{F}_{7/2} - ^4\text{S}_{3/2} (\lambda_2 = 750 \text{ nm}) \\
 &^4\text{I}_{9/2} \rightarrow ^4\text{F}_{5/2} - ^2\text{H}_{9/2} (\lambda = 800 \text{ nm}) \\
 &^4\text{I}_{9/2} \rightarrow ^4\text{F}_{3/2} (\lambda_2 = 800 \text{ nm})
 \end{aligned} \tag{32}$$

The most efficient self-frequency mixing lasers based on channels in Eqs. (29)–(31) are gathered in Table 6.

Further Reading

- Brenier, A., 2000. The self-doubling and summing lasers: overview and modeling. *Journal of Luminescence* 91, 121–132.
- Butcher, P.N., Cotter, D., 1990. *The Elements of Nonlinear Optics*. Cambridge: Cambridge University Press.
- Huber, G., 1999. Visible cw solid-state lasers. In *Advances in Lasers and Applications*. Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing.
- Joubert, M.F., 1999. Photon avalanche up-conversion in rare earth laser materials. *Optical Materials* 11, 181–203.
- Kaminskii, A.A., 1996. *Crystalline Lasers: Physical Processes and Operating Schemes*. CRC Press, Inc.
- Lenth, W., Macfarlane, R.M., 1992. Upconversion lasers. *Optics & Photonics News*. 3: 8–15.
- Powell, R.C., 1988. *Physics of Solid-State Laser Materials*. New York: Springer-Verlag.
- Risk, W.P., Gosnell, T., Nurmikko, A.V., 2003. *Compact Blue-Green Lasers*. Cambridge: Cambridge University Press.
- Weber, M.J., 2001. *Handbook of Lasers*. CRC Press.
- Yariv, A., 1995. *Quantum Electronics*, Third Edition John Wiley & Sons.

Thin Disk Lasers

Mikhail Larionov, Dausinger + Giesen GmbH, Stuttgart, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

A thin disk laser (TDL) is a kind of solid state lasers, which uses a laser medium with a small thickness in the direction of the laser beam. One disk surface is cooled resulting in a heat flow, which is mainly parallel to the laser beam (Fig. 1). The term “TDL” using disk, rarer “disc,” arises from the German word “Scheibenlaser” and is used in many publications, firstly in Giesen *et al.* (1994). The name “disk” presumes that the disk is circular shaped. The shape is indeed often used, but is not necessarily required. The disk is used in a laser as an amplifying mirror and is therefore often referred to as an active mirror.

The size of the laser disk and its thickness depend on many parameters. Typically, disks have thicknesses between 100 and 200 μm and diameters between 5 and 25 mm. Some pictures of laser disks mounted on heat sinks are shown in Fig. 2.

The idea of the TDL evolved in the early 1990s as the answer to the question: “How can the laser output power be increased, while preserving a good beam quality?” The beam quality is a measure of the beam focusability. The output beam quality of lasers often deteriorates at higher output power due to thermal issues. Part of the excitation energy is transferred into heat and deposited in the gain medium. The heat flow in the gain medium causes a temperature gradient, which, consequently, produces an optical parameters gradient in the gain medium.

In Fig. 3, left, a typical temperature distribution in a rod-shaped laser medium cooled from its cylindrical circumference is shown. This distribution results in a strong effective focusing lens of the rod. This lens is not completely spherical and disturbs the beam quality. Additionally, depolarization due to mechanical stress caused by thermal expansion produces losses for the laser beam, limiting the achievable output power. In order to avoid these limitations, cooling has to be improved by reduction of the material thickness. This is especially important, since all laser media have relatively low heat conductivity. Even for the best candidates with a heat conductivity of $10 \text{ W m}^{-1} \text{ K}^{-1}$ the temperature increase over 1 mm of the laser material for a heat flow of 5 MW m^{-2} (which is typical for TDL) is 500K. In order to avoid damage of the laser medium the material thickness has to be reduced, either by reducing the thickness, thereby leading to the thin disk geometry or slab geometry, or by decreasing the diameter, thereby leading to a fiber geometry. Basically all of these geometries allow for a high beam quality: mode-selective light guiding in the fiber ensures a good output beam quality in a properly designed fiber. The thin disk geometry reduces distortions by

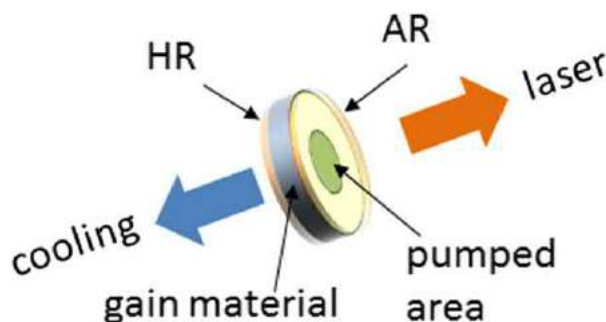


Fig. 1 Thin disk laser (TDL) geometry. AR, antireflective coating; HR, highly reflective coating.

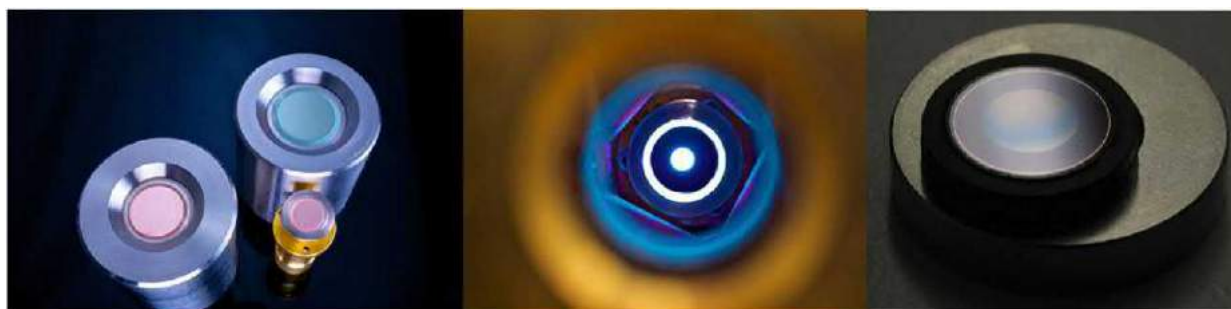


Fig. 2 Examples of disks with different coating, mounted on various heat sinks (left); disk in operation (middle); disk on a structured heat sink (right). Johannes Trbola and Dausinger + Giesen GmbH.

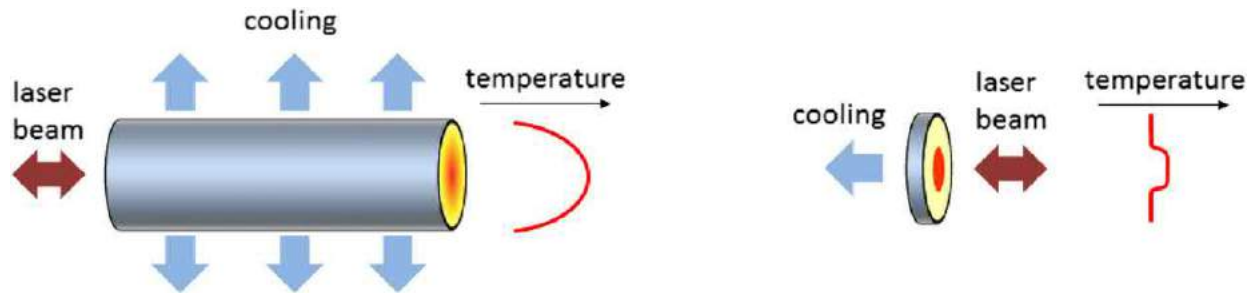


Fig. 3 Temperature distribution in a rod cooled at its cylinder surface (left) and in a disk cooled at one of its plane surfaces (right).



Fig. 4 Left: one of the first thin-disk laser (TDL) setups, demonstrated an output power of 100 W. This setup provides eight passes of the pump power radiation through the thin disk (courtesy of IFSW). Right picture shows two disk modules: the smaller one is designed for 1 kW pump power and the larger one – for 30 kW. Both modules provide 24 or 48 passes of pump radiation through the laser disk, depending on the pump beam quality. Dausinger + Giesen GmbH.

a strong reduction of the material thickness and the slab laser geometry allows for high-power, high beam quality output, however with some spatial anisotropy due to the different thicknesses in the cooling and the laser beam propagation directions.

The invention of the TDLs is closely connected to the ytterbium (Yb) laser ion. In the early 1990s the Nd:YAG gain material was broadly used in the flash-lamp pumped lasers. The advantages of using ytterbium were discussed, but only the availability of laser diodes with sufficient power and narrow spectrum enabled use of Yb in high-power lasers at that time. Ytterbium has several advantages over Nd: lower heat generation, absence of parasitic effects like up-conversion, cross-relaxation and concentration quenching. However, realization of this potential demands for a high excitation density, which is only possible with a sufficient cooling.

In 1992 a group of Dr. Giesen at the University of Stuttgart, IFSW institute invented the main principles of the TDL and started research on it. In 1993 IFSW together with DLR (German aerospace center) have applied for a patent, protecting the basic principles of TDL (Brauch *et al.*, 1995). One of the first TDL setups used at that time is shown in Fig. 4.

On the left photograph of the figure a disk holder with two water connections for cooling is shown. Rightward from the disk holder a set of three spherical mirrors and one flat mirror can be seen, reflecting the pump beam four times on the disk. The pump beam originates from an array of seven fiber bundles, placed in the upper right corner.

Support from the German ministry of science and education (BMBF) allowed a fast development, increase of the output power and commercialization of TDLs. Several companies in Germany and outside of Germany bring on the market laser products, based on the thin disk geometry, now. Nowadays, the TDL is a mature technology broadly used in material processing with continuous-wave and pulsed lasers, medicine, diagnostics, etc. In the research field TDL seems to be one of the most promising approaches for enabling of the next generation of research lasers (Fattahi *et al.*, 2014), providing powerful femto- and attosecond pulses. The TDL technology is regarded as enabling the third-generation femtosecond technology after dye (first generation) and titanium:sapphire lasers (second generation).

Principle

The disk is coated highly reflective (HR) on the cooled side and serves as an active mirror in the laser setup, i.e., a mirror amplifying the incident beam within two passes through the laser active medium. The opposite side of the disk is typically antireflective (AR) coated for both, laser and pump wavelengths (Fig. 1). In this geometry, the heat generated in the laser is

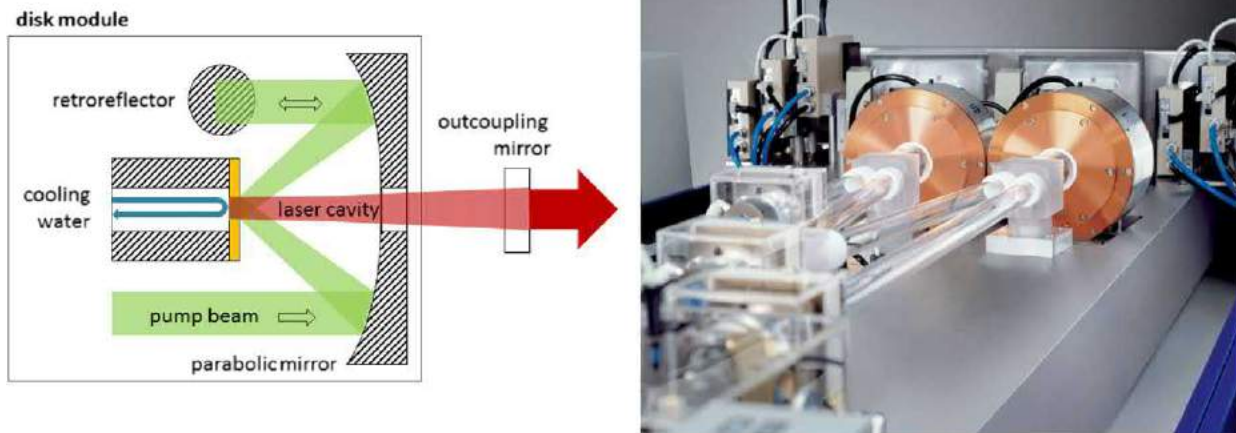


Fig. 5 Typical setup of a thin-disk laser (TDL) (left). Practical realization of a TDL using two disk modules in one laser cavity (right). TRUMPF group.

transported to the coolant on the shortest possible way. The heat flow is parallel to the laser beam. Thus, in the first approximation the radial temperature profile in the disk is nearly flat-top shaped, minimizing the thermal lens and other effects connected to the radial temperature distribution (like stress-induced depolarization).

A typical laser setup is shown in Fig. 5. The laser disk is fixed on a water cooled mount. The pump light is reflected on the disk several times in order to enhance absorption, using a parabolic mirror and several folding mirrors (retroreflectors). This setup is called disk module, pump module, pump chamber, or pump cavity. A laser of the company Trumpf GmbH & Co KG implementing two disk modules, where both disks are used as folding mirrors is shown on the right side of the figure. The cavity beam is encapsulated in transparent tubes.

Pumping Schemes

A thin laser medium provides low absorption and low gain. For the typically used disks absorption is in the range of 10–20%. This is not sufficient for an efficient laser operation. Therefore different multipass schemes are used for recycling the transmitted pump radiation, propagating it through the laser medium as often as possible.

Laser diodes are the most commonly used pump source of the TDL. Most of the schemes are using 4 f or Rayleigh imaging of the pump spot on the disk again and again to shape the same pump spot for each pass of the pump radiation. The simple 2 f-imaging used in the first TDL (as in Fig. 4) leads to an increase of the pump beam divergence from pass to pass, limiting the number of passes. In order to simplify the multipass setup a single parabolic mirror instead of many imaging lenses is typically used. Multiple passes are produced with different configurations of folding mirrors, placed near the focal plane of the parabolic mirror. Several patents dealing with this topic were issued.

An interesting idea is the usage of only two reflectors for a setup with an arbitrary number of absorbing passes through the laser medium. Each reflector consists of a mirror pair, with mirrors arranged at an angle of 90 degree between the mirrors. This idea is illustrated in Fig. 6. A collimated pump beam enters the setup between the reflectors and hits the parabolic mirror at position 1. The parabolic mirror focuses the beam onto the thin disk. The not absorbed part propagates to the position 2 on the parabolic mirror, is re-collimated there and propagates further passing the positions 7–12, where the propagation direction is reversed and the residual pump beam propagates the same way back to position 1 and exits the setup. In total the pump beam completes six reflections on the disk, passing 12 times through the laser medium. Changing of the angle between the symmetry axes of both reflectors changes the number of pump passes.

High pump power is typically available with a low beam quality only. Designing the pump optics for low-beam-quality pump radiation requires usage of larger optics and a reduction of the number of pump passes. In Fig. 4 on the right is a comparison of a pump optics designed for a pump power of 1 kW and a pump optics designed for a pump power of 30 kW.

An arrangement of the pump reflections on the parabolic mirror in two rings or even more complicated patterns can be used. Up to 48 passes of pump radiation in a commercially available multipass optics have been demonstrated. In general, properly designed pump optics can accept a rather low beam quality of the pump radiation. TDL efficiently converts this “low-quality” radiation to radiation with a high beam quality.

For low power lens systems can be also applied.

An obvious way to improve absorption is pumping of the disk through its thin side. This method is often referred as “radial” or “side” pumping. In this case a long way of the pump radiation in the disk plane, which is increased by zig-zag propagation due to total internal reflection, allows efficient absorption of the pump light. This scheme was proposed at the very beginning of the TDL

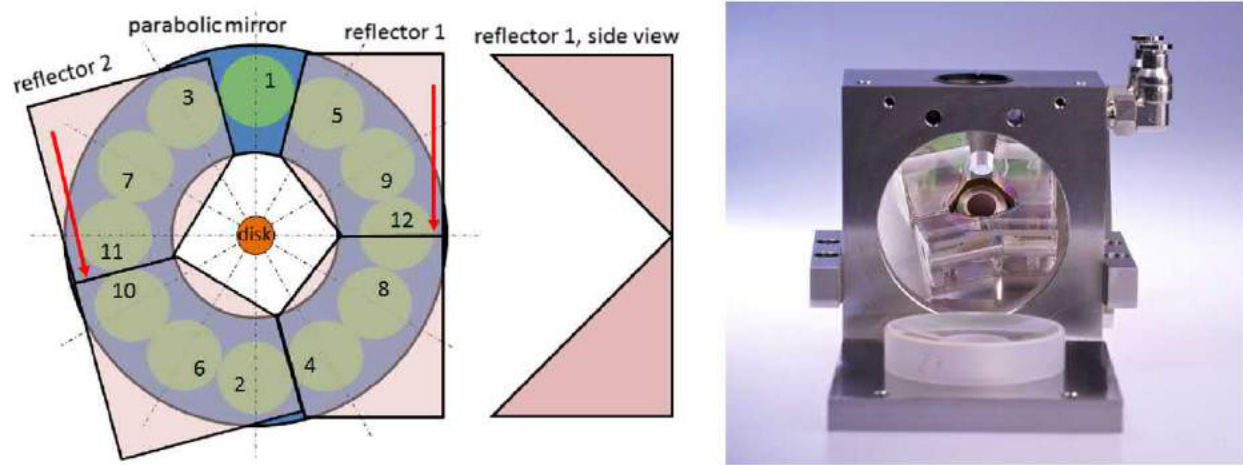


Fig. 6 Left: Principle of a pumping multipass setup, providing 12 reflections on the thin disk or 24 absorption passes of the pump radiation through the laser material. Beam positions on the parabolic mirror are shown with circles and numbered in the same order as beam passes them. Right: Picture shows a practical realization of this setup. TDM1.0, Source: Dausinger + Giesen GmbH.

development, but has not yet found any broad industrial application. Its main drawback is the nonuniformity of the absorbed power in the disk and its dependence on the changes of every emitter.

Gain

Typically the small signal gain of thin disks is low and does not exceed 30% (some exceptions are described in the next chapter). In order to deal with the low gain of the thin laser disk, special techniques have been developed. Multiple reflections on the laser disk in one round trip in the laser cavity or amplifier allow multiplying the gain. For example, using of the laser disk as a folding mirror increases the gain by a factor of 2 compared to the usage as an end mirror.

Assuming a laser disk gain of 8%, which is a typical value for > 10 kHz laser operation, 30 reflections on the disk are necessary to achieve a total amplification factor of 10. For an amplification of 1,000,000 the number of reflections is 180. While the first can be realized with a mirror arrangement (multipass amplifier) the second one needs regenerative amplification, where the number of amplifier bounces is defined by a number of round trips in the amplifier cavity.

Mounting

The laser disk itself does not have sufficient mechanical strength to be directly cooled. Therefore, the thin disk is typically mounted on a mechanically stable carrier, which often works also as a heat spreader. Mounting to the carrier is the key technology for TDL. The requirements to the mounting are tough: low thermal resistance, plane or spherical shape of the disk after the mounting within sub 100 nm, high stiffness of the resulting sandwich, withstanding mechanical stress in laser operation, homogeneous mechanical and thermal contact, manufacturability, cost, etc. Several techniques (overview in, e.g., [Larionov, 2009](#)) have been broadly used in the last years. First disks were attached to copper heat sinks with a layer of indium. This technique is well suited for the laboratory use, but reaches its limitations for high-power and/or long-term operation due to low creeping resistance and mechanical fatigue of indium. Soldering and gluing of the laser disks were consequently developed and are the mostly used mounting technologies at the moment. Especially, gluing, which allows usage of diamond as a heat sink is a well-established technique for high-power operation. New techniques like diffusion bonding or some kind of “chemically activated” bonding, epitaxial growth – have been also proposed for mounting.

Amplified Spontaneous Emission

ASE plays an important role for TDL. A photon, which is emitted in the pumped volume of the thin disk, propagates at some angle to the disk axes. A large fraction of the photons does not escape from the disk due to total internal reflection and is guided to the edge. For example, for Yb:YAG this fraction is 84%. On its zig-zag way through the pumped volume the photon is amplified with a coefficient, which is much higher than the gain for laser radiation, since the way of the fluorescence photon in the gain medium is much longer. At the edge the photon can be scattered or reflected back in direction of the pumped volume. In this case it gathers even more amplification. This process reduces effective gain of the thin disk for high pump power densities and/or for large pump spots. Typical ways for reduction of ASE is usage of a thick (compared to the thin-disk) undoped cap at the top of the disk, usage of thicker disks, beveling of the disk edge, introducing absorption at the disk edge by co-doping or attaching of an absorbing material.

Recently, several groups published research on cryogenically cooled TDLs with an undoped cap. The undoped cap is used to reduce ASE. In this special architecture the laser disk can be quite thick (up to 1 mm). The laser disk is optically contacted to a thick undoped cap on the side opposite to cooling and cooled down to cryogenic temperatures (usually liquid nitrogen). The thickness of the undoped cap is even larger than that of the disk, allowing an efficient suppression of ASE, because most fluorescence photons pass the gain medium only once and propagate further in the undoped cap. This is a boundary case between thin-disk and bulk laser geometry since the heat flow is not one-dimensional. This leads to a stronger thermal lensing and depolarization compared to the thinner disks. This disadvantage can be mitigated by operation at low temperature.

Scaling

The output power or energy of a TDL can be increased by increasing the pumped area and/or by increasing the number of the used laser disks. Whereas scaling via the disk number is limited at least by the disk damage at high power or energy density, scaling via pumped area is not limited by temperature or mechanical stress as long as pump and laser power/energy densities remain constant. Practically, there are some issues limiting enlarging of the laser disk. Single crystals are available up to a diameter of roughly 100 mm. Sintering allows production of polycrystalline material (often called laser ceramics) with any dimensions and quality comparable to that of the single crystal.

Mounting of a larger thin disk is a technological challenge. Also challenging is designing and setup of the laser cavity for large disks. Once the problems are solved, only ASE limits the achievable energy and power. Several approaches were developed to estimate the limitations of the TDL geometry. An output power of 1 MW is feasible with an optical efficiency of 50% and a pump spot diameter of 20 cm, if the cavity internal losses remain low $L < 0.25\%$. For pulsed laser operation the expected losses are higher. For a loss of 4% an energy of 8.3 J can be extracted. For repetition rates below $1/\tau$ (where τ is fluorescence lifetime) the use of pulsed pumping allows even higher output energies. Lower losses would also allow extraction of pulses with higher energy.

Performance in Different Operation Modes

Continuous Wave

The typical transmission of the output coupling mirror of a TDL in cw operation is only a few percent. The power circulating in the laser cavity is typically 10–50 times higher than the output power. Therefore, the TDL design is sensitive to losses. The typical loss amounts to about 0.1%, requiring a very good mirror coating and a clean environment. At the moment the highest multimode power from one laser disk is 10 kW, as demonstrated by [Schad et al. \(2016\)](#).

Fundamental mode operation is typically achieved by adjusting the mode size on the laser disk to the size of the pumped area. The pumped area provides an effective aperture, which cuts off higher transversal cavity modes. In general, cavity design and its sensitivity to thermal lensing and misalignment scales with r_{mode}^2 . Therefore, high-power fundamental mode operation was demonstrated later than high-power operation in multimode. The highest power of 4 kW was demonstrated.

In ([Speiser, 2016](#)) a way of scaling the output power to the multi-10 kW range is proposed. It includes using a high-power oscillator with a good beam quality and further amplification in a chain of multipass amplifiers. An output power of 12 kW with an M^2 of 10 has been generated using this approach.

Mostly, due to low gain of the thin disk, stable cavities are used for operation of TDLs. Unstable cavities provide high outcoupling ratio and cannot be readily used. The company Boeing used multiple disks to overcome this limitation. A record output power of 25 kW or even 30 kW according to the company's homepage with a "nearly diffraction limited" beam quality was demonstrated. Since the fundamental mode of an unstable cavity is not equal to fundamental Gaussian mode of a stable cavity, it is difficult to compare the beam quality of this result to other results in terms of beam quality.

Single-frequency

Oscillations in a TDL cavity can be limited to one longitudinal mode with suitable frequency-selective elements, allowing generation of high-power, spectrally narrow radiation. A Lyot filter together with an etalon ensure oscillations of only one longitudinal mode. The high laser power circulating in the laser cavity heats the frequency-selective elements used in transmission. A very promising approach is therefore the usage of reflective elements, like a highly efficient dielectric diffraction grating.

A typically negative effect of "spatial hole burning" reduces gain and laser efficiency in a laser operation with only one longitudinal cavity mode. This effect works in TDL also as an additional etalon, allowing to narrowly select one mode with a small number of filtering elements.

Frequency-converted

TDLs offer an interesting opportunity for intra-cavity frequency conversion. Since the optimum outcoupling of a continuous-wave cavity is in the range of a few percent, only this amount of circulating radiation has to be converted for efficient operation. This allows the usage of nonlinear frequency-converting crystals at moderate power-densities, still providing high laser efficiency. The "green problem" causing output power instability, can be solved, operating the cavity in one longitudinal mode or at least with a sufficiently narrow spectrum. As for single frequency operation, using of an efficient reflective diffraction grating is beneficial

for high power output. 403 W of continuous-wave green radiation was generated using this approach, recently. A few-Watt laser based on this approach is built commercially in high numbers.

Q-Switch and Cavity Dumping

Modulation of the cavity losses allows pulsed operation of the TDL. Typically acoustooptical or electrooptical modulators are used. In Q-switched operation the losses are modulated between low loss and high loss. The output pulse length is defined by the cavity length and laser dynamics. Typical pulse lengths are between 0.2 μs and several microseconds. An output power of 190 W with a nearly diffraction-limited beam quality and pulse energy of maximum 18 mJ has been demonstrated.

In cavity-dumped operation, the losses of the cavity are modulated between roughly 0 and 1. Reduction of the losses to 0 for pulse build-up allows a fast increase of cavity internal power. Fast switching of the losses to 1 dumps the cavity power in one round trip and provides at the output a pulse with a pulse duration defined by the cavity round trip time. Typical pulse durations in this regime are 10–50 ns. Shorter pulse durations are hard to achieve at high power, due to the limitations of the available cavity length. An output power of 850 W with a beam quality of 20 mm $\times$ mrad and a pulse energy of up to 80 mJ is commercially available.

Frequency-converted

Modulation of the cavity losses can be combined with intracavity frequency conversion. The low gain of the thin disk allows efficient operation with a low conversion efficiency and thus at a moderate power density in the nonlinear conversion crystal. This allows high-power output with a pulse duration of several 100 ns. The pulse duration is longer than that of cavity-dumped lasers, because conversion extracts power from the laser cavity in multiple round trips and not in one round trip. Lasers based on this principle with an output power of 300 W and a pulse energy of maximum 7.5 mJ are commercially available.

Regenerative and Multipass Amplification

Shorter pulse durations can be obtained with amplifiers. A regenerative amplifier theoretically allows for an unlimited number of reflections on the laser disk. Amplification factors of up to 10^{11} have been experimentally demonstrated. A robust design with a small number of optical components made this principle commercially successful. The mode size on the laser disk is not limited and can be enlarged for higher pulse energies, avoiding damage of the optical components. For amplification of spectrally broad picosecond and subpicosecond pulses chirped-pulse amplification is used. For picosecond pulse duration an output powers of 220 W at 1 kHz, 300 W at 10 kHz and 1.1 kW at 100 kHz have been experimentally demonstrated.

Another interesting application is the amplification of spectrally narrow nanosecond or sub-nanosecond pulses. The output of the laser repeats temporal and spectral characteristics of the seed pulse and can be used for applications requiring high energy and high power pulses with narrow spectrum.

Gain narrowing during amplification can be compensated by nonlinear spectrum broadening in the electro-optical switch or another optical element used in transmission. Pulse duration of an amplifier using Yb:YAG is typically slightly below 1 ps. Using spectrum broadening a pulse duration of well below 300 fs has been shown.

For generation of even higher pulse energy or power the electro-optical switch may become the limiting factor. Therefore, multipass amplifiers, reflecting the signal beam on the laser disk repeatedly by means of turning mirrors, have been developed. Different concepts of beam propagation, including 4 f imaging and not imaging systems, have been built up and evaluated. For pulse amplification optimized for high output power an output power of 1.4 kW with a pulse energy of 6.3 mJ and picosecond pulse duration and, recently, an output power of 1.9 kW has been demonstrated. For high-energy operation an output energy of 1 J at a repetition rate of 100 Hz was achieved with an amplifier chain consisting of a regenerative amplifier and two multipass stages. Rather thick disks with a thickness of 0.75 mm and pulsed pumping have been used, allowing efficient ASE suppression and thermal management. Combination of both goals: high power and high energy remain challenging.

Mode Locking

Mode locking of TDLs allows the highest output-power and energy achieved in the mode-locked regime when compared with other lasers. The reason for this is a very low nonlinearity of the laser gain medium combined with the principal power-scalability of the TDL.

Typically passive mode-locking is realized using semiconductor saturable absorbing mirror (SESAM) or Kerr-lensing. Since the nonlinearity of the laser medium is small, the nonlinearity of other optical components and even of the air in the laser cavity becomes the critical factor for increasing of the output pulse energy using SESAM. This problem can be solved by operation in vacuum or in helium atmosphere or by increase of the output coupling rate, which decreases the energy of the circulating pulses in the cavity. Kerr lens mode-locking produces shorter pulses.

Materials

Yb:YAG

Yttrium aluminum garnet (YAG, $\text{Y}_3\text{Al}_5\text{O}_{12}$), where some yttrium ions are substituted with ytterbium ions (doping) is the mostly used and mature material for TDL. YAG has a heat conductivity of $11 \text{ W m}^{-1} \text{ K}^{-1}$. Doping decreases heat conductivity to, for

example, $6 \text{ Wm}^{-1} \text{ K}^{-1}$ for 10% doping concentration. Any doping concentration from 0 to 100% can be manufactured. Typically a concentration around 10% is used for thin disks.

The ytterbium ion has a very simple energy structure, consisting only of two multiplets, which are further split into seven energy levels by crystal field. The heat generation is mainly caused by relaxation in the multiplets, resulting in the quantum defect. At high excitation densities, usual for TDLs, not completely understood parasitic processes provide additional decay channels involving several excited ions.

There is a variety of other materials used in TDL. The reason for using another material is either enhancing of power capability and efficiency or demand for a different spectrum – a broader one for generation of shorter pulses.

Power capability of a laser material depends mainly on its heat conductivity. Therefore, several materials with a heat conductivity higher than that of YAG were proposed for usage in TDL. Among them only Yb:LuAG (lutetium aluminum garnet) found industrial application. It has a very similar structure and parameters as YAG, just Y is substituted by Lu. Since Lu ion has a larger ion radius than Y, doping with Yb does not decrease heat conductivity that much. Sesquioxides are cubic crystals with a formula X_2O_3 . Especially Lu_2O_3 , Sc_2O_3 and Y_2O_3 have been examined in TDL. The heat conductivity of a doped Lu_2O_3 ($12 \text{ Wm}^{-1} \text{ K}^{-1}$) is higher than that of doped YAG. Additionally, mixed oxides can be manufactured, allowing to broaden the gain spectrum. The main drawback impeding the usage of sesquioxides, so far, is their high melting temperature, making crystal growth difficult. Polycrystalline material (laser ceramics) does not have this drawback.

There are plenty of materials, which were examined in TDL in order to get a broader gain spectrum, maintaining amplification and generation of pulses shorter than possible with Yb:YAG. As a rule of thumb the emission cross-section is inverse proportional to the gain bandwidth. Thus, materials with broader gain bandwidth typically provide lower gain. Broader gain spectrum is mostly caused by some disorder in the crystal structure – anisotropy of presence of different crystal places for Yb ion. Such materials are typically more difficult to handle and to manufacture. Some of interesting candidates are tungstates ($\text{KY}(\text{WO}_4)_2$), calcium fluoride (CaF_2), CALGO (CaGdAlO_4). Only tungstates are used for industrial lasers, so far.

Nd Doping

In general, Neodymium has a more complicated energy structure allowing various parasitic decay channels, compared to Yb and a higher quantum defect. Therefore, the heat generation is much higher, limiting the output power. Due to the large radius of the Nd ion, maximum doping concentration is limited for most host materials, either reducing the laser efficiency or requiring thicker disks, which in turn limits the output power. Nd:YAG and YVO (YVO₄) are used in TDL, in order to get access to various laser wavelengths offered by the energy structure of Nd and to use them after frequency conversion. Different visible colors from blue to red can be produced using this approach. The typical output power is limited to a few dozen Watt.

Ho Doping

Holmium is used to get access to the “eye-safe” wavelength of 2.1 μm . Continuous-wave operation with an output power of 20 W and Q-switched operation with an output power of 9 W have been demonstrated. Ho can be pumped at 1.9 μm , providing a low quantum defect, which is beneficial for high-power operation. The principal possibility of pumping with laser diodes, which is a prerequisite for its commercial usage, has been demonstrated. However, availability and parameters of the available laser diodes are still not sufficient. Once this issue is solved, Holmium can be an interesting candidate even for high-power TDLs. At the moment Tm lasers are often used for pumping of Ho TDL.

Tm Doping

Emission wavelength of Thulium is 1.9 μm and is also “eye-safe.” An output power of 21 W and a tuning range from 1899 to 1927 nm was demonstrated using Tm:LiLuF₄.

Thulium is often used as a co-doping to provide a pumping wavelength at 790 nm, which is easily accessible with laser diodes. However in TDL this scheme has not been demonstrated to work well.

Cr Doping

Chromium doped ZnSe has been demonstrated to achieve an output power of 5 W at 2.4 μm . The laser was tunable over more than 100 nm from 2.23 to 2.37 μm . The same pumping options as for Ho are available.

Semiconductor

TDL geometry is also used in vertical-external-cavity surface-emitting lasers (VECSELs), which are optically pumped semiconductor lasers (OPSLs). A semiconductor chip including an end Bragg mirror is grown to provide the required parameters, mounted on a heat sink and pumped from the opposite side. This geometry is often referred to as semiconductor disk laser (SDL), underlining its connection to TDL geometry. Using a grown semiconductor structure as a laser medium provides a very interesting possibility to engineer the laser wavelength. Intracavity frequency doubling allows to access the visible spectrum range. The inherent “green

problem” is not present in SDL design allowing for a very simple cavity design. Such lasers are commercially available in a variety of wavelengths with an output power of up to 20 W.

References

- Brauch, U., Giesen, A., Voss, A., Wittig, K., 1995. Laser Amplifier System. Idea of Thin-Disk Laser, European Patent EP 0 632 551 B1.
- Fattahi, H., Barros, H.G., Gorjan, M., *et al.*, 2014. Third-generation femtosecond technology. *Optica* 1, 45–63.
- Giesen, A., Hügel, H., Voss, A., *et al.*, 1994. Scalable concept for diode-pumped high-power solid-state-lasers. *Applied Physics B* 58, 365–378.
- Larionov, M., 2009. Kontaktierung und Charakterisierung von Kristallen für Scheibenlaser. Munich: Herbert Utz Verlag.
- Schad, S., Gottwald, T., Kuhn, V., *et al.*, 2016. Recent development of disk lasers at TRUMPF. In: *Proceedings of the SPIE*. 9726, Solid State Lasers XXV: Technology and Devices.
- Speiser, J., 2016. Thin disk lasers: History and prospects. In: *Proceedings of SPIE* 9893, Laser Sources and Applications III, 98930 L.

Further Reading

- Contag, K., Erhard, S., Giesen, A., *et al.*, 2000. Laser Amplification System. Pumping of Multiple Disks. International Patent WO 00/08726 A3.
- Erhard, S., Giesen, A., Stewen, C., 2002. Laser Amplifier System. Pump Optics With Two Reflectors. European Patent EP1252687 B1.
- Erhard, S., Giesen, A., Karzewski, M., Stewen, C., Voss, A., 2000. Laser Amplification System. Pump Optics With Parabolic Mirror. International Patent WO 00/08728.
- Hartke, R., Baev, V., Seger, K., *et al.*, 2008. Experimental study of the output dynamics of intracavity frequency doubled optically pumped semiconductor disk lasers. *Applied Physics Letters* 92, 101107.

Relevant Websites

- www.dausinger-giesen.de
Dausinger + Giesen GmbH.
- www.jenoptik.com
Jenoptik.
- www.trumpf.com
Trumpf.
- www.rp-photonics.de
RP Photonics.

Microchip Lasers

John J Zayhowski, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, United States

© 2017 Elsevier Inc. All rights reserved.

Introduction

Microchip lasers are a rich family of solid-state lasers defined by their small size, robust integration, reliability, and potential for low-cost mass production. In its simplest embodiment, a microchip laser is a monolithic device that consists of a small piece of solid-state gain medium polished flat and parallel on two opposing sides. Dielectric cavity mirrors are deposited directly on the gain medium and the laser is pumped with a diode laser, either directly, as shown in Fig. 1, or via an optical fiber. In other embodiments, two or more materials are joined together before being polished and dielectrically coated to form a composite-cavity microchip laser with added functionality or enhanced mode properties. For example, the additional material could be an electro-optic material to enable frequency tuning of the laser, a nonlinear material for intracavity wavelength conversion, a saturable absorber to Q switch the device, or a passive transparent material to increase both the transverse mode area and the output power of the laser.

Continuous-wave (CW) microchip lasers cover a wide range of wavelengths, often operate single frequency in a near-ideal mode, and can provide a modest amount of tunability. Q-switched microchip lasers provide the shortest output pulses of any Q-switched solid-state laser, with peak powers up to several hundred kilowatts. The average output power of microchip lasers is typically in the range from several tens to several hundreds of milliwatts.

Background

Most solid-state lasers are built from discrete optical components that must be carefully assembled and critically aligned. Laser assembly is typically performed by trained technicians, and is time consuming and expensive. As a result, the cost of most solid-state lasers makes them unattractive for a wide range of applications for which they would otherwise be well suited. Additional characteristics that have historically impeded the widespread use of solid-state lasers include their size and reputation for being fragile and unreliable. This was even truer in the early 1980s, the infancy of microchip laser development, than it is today. Microchip lasers were developed to overcome these limitations – cost, size, robustness, and reliability – and thereby become viable components for a variety of large-volume applications. The term “microchip laser” was coined at MIT Lincoln Laboratory in the early 1980s to draw an analogy between this new class of lasers and semiconductor electronic microchips with their inherent small size, reliability, and low-cost mass production.

Consider the simplest of all possible microchip lasers, a small piece of solid-state gain medium polished flat and parallel on two opposing sides, with dielectric cavity mirrors deposited directly onto the polished faces, as shown in Fig. 1. Fabrication of the laser starts with a large boule of gain material, such as Nd:YAG. The boule is sliced into wafers about 0.5-mm thick. The wafers are then polished and dielectrically coated before they are diced into 1-mm-square pieces, with each piece being a complete laser. One boule can produce thousands of lasers, and throughout the fabrication process the lasers never need to be handled independently.

To complete a microchip laser system, the laser must be coupled to a pump source. Microchip lasers are pumped with semiconductor diode lasers. The 1980s were a time of rapid development in diode lasers. The amount of power that was available from commercial diode lasers was rapidly increasing and the cost per watt of output power was quickly decreasing, with projections of extremely inexpensive, high-power diode lasers in the near future. As a result, diode lasers fit nicely into the picture of low-cost microchip laser systems. To keep the cost of the system low, it is important that the coupling of the diode to the microchip laser be performed inexpensively. The use of a flat-flat laser cavity eliminates any critical alignment between the diode and the laser and makes the assembly of the system quick and simple, with the potential for inexpensive automation.

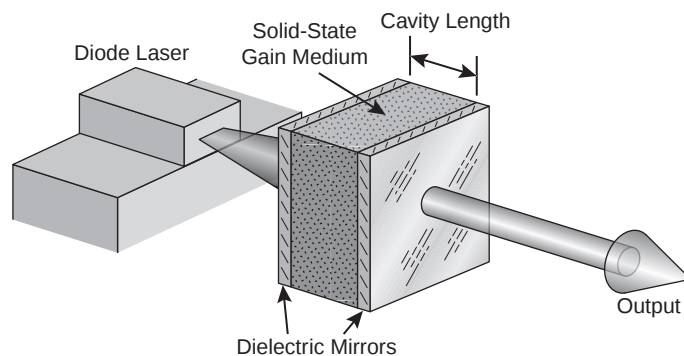


Fig. 1 Illustration of a monolithic microchip laser.

The use of simple, small, monolithic, mass-produced solid-state lasers noncritically coupled to low-cost semiconductor diode pump lasers gives microchip laser systems their defining characteristics – small size, robust integration of components, reliability, and the potential for low-cost mass production – and differentiates them from other miniature laser systems that are designed and constructed using more conventional techniques.

Since their early development, microchip lasers have evolved into a rich family of devices with capabilities that often exceed those of conventional lasers. The first microchip lasers developed operated CW, and a variety of CW microchip lasers were quickly demonstrated, covering a wide range of wavelengths. Some of the early applications required a modest amount of tunability, and several tuning mechanisms were incorporated into the laser structure. It was not long before researchers realized that the short cavity lengths inherent to microchip lasers gave them tremendous potential as pulsed devices. This led to the development of actively Q-switched microchip lasers, quickly followed by the most successful variation of microchip laser, the passively Q-switched microchip laser, which has demonstrated the shortest output pulse of any Q-switched solid-state laser.

Overview

Monolithic Microchip Lasers

The earliest microchip lasers were monolithic devices based on optical transitions in the Nd^{3+} ion near 0.94, 1.06, or 1.32 μm , using a variety of gain media including Nd:YAG, NPP, LNP, Nd:GSGG, Nd:YVO₄, Nd:LaMgAl₁₁O₁₉, Nd:YCeAG, Nd:YLF, Nd_xY_{1-x}Al₃(BO₃)₄, Nd:La₂O₂S, Nd:MgO:LiNbO₃, and Cr,Nd:YAG. It was not long, however, before other active ions were investigated and monolithic microchip lasers were constructed at a wide variety of wavelengths. These include Cr:LiSAF microchip lasers operating in the 0.8- to 1.0- μm spectral region, Yb:YAG microchip lasers at 1.03 μm , Yb,Er:glass microchip lasers at 1.5 μm , Tm and Ho microchip lasers operating near 2 μm , Cr-doped chalcogenides microchip lasers near 2.3 and 2.5 μm , and Er-doped microchip lasers at 3 μm . Of the gain media demonstrated in microchip lasers, Nd:YAG and Nd:YVO₄ are the most commonly used.

All of the lasers listed above operate in a near-ideal fundamental transverse mode, and several operate single frequency (in a single longitudinal mode) with a narrow linewidth. The Nd_xY_{1-x}Al₃(BO₃)₄ microchip laser has another very interesting characteristic. In addition to being a good gain medium, Nd_xY_{1-x}Al₃(BO₃)₄ has a second-order nonlinearity that allows it to perform harmonic generation, and the microchip laser is self frequency doubled to produce green output at 531 nm. The Nd:MgO:LiNbO₃ microchip laser is electro-optically tunable. The doubly doped Cr,Nd:YAG laser is the only one of the lasers listed above that is not CW – it is self passively Q switched.

Composite-Cavity Microchip Lasers

The search for multifunctional gain media – gain media that simultaneously act as harmonic converters, electro-optic material, or passive Q switches – has led to several interesting monolithic devices. However, the multifunctional media are often less than ideal in one or more of their functions, and/or are difficult or expensive to grow. Higher-performance devices can be obtained by combining two or more specialized materials within the same composite cavity. In composite-cavity microchip lasers the constituent materials are bonded together to form a quasi-monolithic device, with dielectric mirrors deposited (or bonded) on the outer surfaces.

Several issues must be addressed in the design of composite-cavity microchip lasers. Different materials have different refractive indices and different thermal-expansion coefficients. The optical, mechanical, and thermal properties of the material interface must be dealt with in a way that satisfies the optical requirements of the laser cavity, is robust enough for the intended application, and is cost-effective. Otherwise, the less-than-ideal performance of a monolithic device, if one can be constructed, may be the best solution.

Composite cavities have been successfully employed in coupled-cavity microchip lasers to obtain single-frequency operation from gain media with extremely broad gain bandwidths, harmonically converted green and blue CW microchip lasers, electro-optically tunable microchip lasers, actively Q-switched microchip lasers, passively Q-switched microchip lasers, and frequency-converted passively Q-switched microchip lasers.

Pumping

Microchip lasers are typically pumped by semiconductor diode lasers. The simplest configuration places the microchip laser in close proximity to the output facet of the pump diode with no intervening optics. As the amount of pump power increases, the diameter of the oscillating mode in a microchip laser usually decreases and, for this proximity-coupled configuration, the typically large divergence of the diode output overfills the oscillating-mode volume resulting in inefficient operation or multi-transverse-mode oscillation. The situation can be improved by putting a lens between the diode and the microchip cavity, which is common practice in moderate- or high-power microchip laser systems (output powers in excess of several tens of milliwatts).

An alternative configuration for pumping microchip lasers uses fiber-coupled diode lasers. This configuration decouples the diode-laser system from the microchip cavity and offers several practical advantages. It separates the engineering of the pump subsystem from the microchip cavity; it makes it easy to independently control the temperature of the diodes and the microchip

cavity; and it facilitates an extremely compact laser head that can fit into small spaces, coupled to the rest of the system by a single, flexible optical fiber. High-power microchip lasers are often pumped by diode-laser arrays. In this case, fiber coupling serves the additional function of shaping the combined diode-laser output.

Transverse Mode Definition

Most microchip lasers use a flat–flat cavity design. The eigenmodes of a flat–flat cavity are plane waves, yet microchip lasers typically operate in a near-ideal, fundamental transverse mode with a well-defined mode radius. The mode-defining mechanisms vary depending on the gain medium, other media within a composite-cavity device, and pump power.

For many microchip lasers the transverse mode is determined primarily by thermal effects. When a microchip laser is longitudinally pumped, the pump beam deposits heat. In materials with a positive change in refractive index with temperature, such as Nd:YAG, this creates a thermal lens that stabilizes the cavity and defines the transverse mode. When the cavity mirrors are deposited directly on the gain medium the heat also induces curvature of the mirrors. For materials with a positive thermal-expansion coefficient this contributes to the stabilization of the transverse mode.

In three-level or quasi-three-level lasers there can be significant absorption of the oscillating light in unpumped regions of the gain medium. This creates a radially dependent loss that can restrict the transverse dimensions of the lasing mode. The mere absence of gain in the unpumped regions of the gain medium can have a similar effect. In passively Q-switched devices bleaching of the saturable absorber results in a dynamic aperture that opens as the Q-switched pulse forms and is an important mode-defining mechanism.

Optical gain provides dispersion. Gain-related index guiding has been shown to play an important role in Nd:YVO₄ microchip lasers, and can lead to interesting effects such as self Q switching.

Parallelism between the cavity mirrors can be critical for the creation of a circularly symmetric fundamental transverse mode in a flat–flat cavity. At high pump powers the mode-defining mechanisms in microchip lasers are usually thermal, and the requirement on parallelism is relaxed as the power of the laser increases. The requirement is less severe in microchip lasers based on three-level gain media or employing a saturable absorber.

Methods consistent with low-cost mass fabrication have been developed to put curved mirrors on microchip lasers. Curved mirrors can stabilize the transverse mode of the laser when the mechanisms discussed above are not strong enough to do so. They can reduce the threshold of CW microchip lasers and make low-power operation more consistent from device to device. For a large variety of gain media they are not needed, especially when the laser is pumped with medium- or high-power diodes (typically more than several tens of milliwatts).

Typical mode radii of microchip lasers fall in the range from 20 to 150 μm , depending on the gain medium, pump power, cavity length, and several other factors. To ensure oscillation in the fundamental transverse mode, that mode must use most of the gain available to the laser. If the radius of the fundamental mode is much smaller than the radius of the pumped region of the gain medium, higher-order transverse modes will oscillate.

Spectral Properties

Single-Frequency Operation

As the cavity length of a laser decreases, its free spectral range increases. The original microchip-laser concept included making the laser cavity sufficiently short that its free spectral range is comparable to the gain bandwidth of the gain medium. Robust single-frequency operation can then be obtained as long as one of the cavity modes falls near the peak of the gain profile. This requires precise, sub-wavelength control of the cavity's optical length (length times refractive index), and is often accomplished through thermal control of the cavity.

Typically, each pulse of a Q-switched microchip laser is single frequency. However, at high repetition rates (when the interpulse period is short compared to the relaxation time of the gain medium) consecutive pulses may correspond to different longitudinal modes.

Fundamental Linewidth

One contribution to the spectral width of all lasers is the coupling of spontaneous emission to the oscillating mode. This gives rise to the Schawlow–Townes linewidth, which has a Lorentzian power spectrum with a width that scales inversely with the output power of the laser. In microchip lasers, thermal fluctuations of the cavity length at a constant temperature can result in a much larger fundamental linewidth. These fluctuations result in a Gaussian power spectrum that, for monolithic devices, scales as the inverse square root of the oscillating-mode volume. Monolithic CW Nd:YAG microchip lasers with a cavity length of ~ 1 mm have a Gaussian spectral profile with a linewidth of several kilohertz, with spectral tails corresponding to a Lorentzian contribution of only a few hertz.

Single-frequency Q-switched microchip lasers have a Fourier-transform-limited optical spectrum that is much broader than the fundamental linewidth of CW devices.

Frequency Tuning

Microchip lasers are frequency tuned by changing the cavity's optical length. The optical length can be changed using a variety of techniques including thermal tuning, stress tuning, and electro-optic tuning. Pump-power modulation represents a special case of thermal tuning. Each of these techniques allows continuous frequency modulation of a single longitudinal cavity mode.

Changing the temperature of an element in a laser cavity, or the entire cavity, is often the simplest way to tune a laser. A change in temperature results in a change to both the physical length of the component and its refractive index.

The cavity modes of a resonator also tune as elements within the cavity are squeezed. Squeezing transverse to the resonator's optical axis results in an elongation of the material along the axis. Superimposed on this is the stress-optic effect. In crystals with cubic symmetry, the stress-optic effect can split the frequency degeneracy of orthogonally polarized optical modes. For squeezing along the optical axis of the cavity there is a compression of the squeezed elements and the frequency degeneracy of orthogonally polarized modes remains unchanged. By using a piezoelectric transducer to squeeze a monolithic Nd:YAG microchip laser, researchers demonstrated tuning at modulation frequencies up to 20 MHz, although nonresonant response was limited to ~ 80 kHz. The nonresonant tuning response was 300 kHz V^{-1} .

Many applications require high rates of frequency tuning that can only be achieved electro-optically. For high-sensitivity tuning it is desirable to fill the cavity with as large a fraction of electro-optic material as possible. However, it is often still important to keep the total cavity length as short as possible, to ensure single-frequency operation, to maximize the tuning range, and to minimize the response time. Composite-cavity electro-optically tuned Nd:YAG/LiNbO₃ microchip lasers have been continuously tuned over a 30-GHz range with a tuning sensitivity of $\sim 14 \text{ MHz V}^{-1}$. The tuning response was relatively flat for tuning rates from DC to 1.3 GHz. Monolithic Nd:MgO:LiNbO₃ and Nd:LiNbO₃ electro-optically tuned microchip lasers have also been demonstrated.

Changes in pump power induce frequency changes in the output of solid-state lasers. As the pump power increases more heat is deposited in the gain medium, causing the temperature to rise and changing both the refractive index and length. Because frequency tuning via pump-power modulation relies on thermal effects, it is often thought to be too slow for many applications. In addition, modulating the pump power has the undesirable effect of changing the amplitude of the laser output. However, for microchip lasers significant frequency modulation can be obtained at relatively high modulation rates with little associated amplitude modulation. For example, pump-power modulation of a 1.32- μm microchip laser has been used to obtain 10-MHz frequency modulation at a 1-kHz rate and 1-MHz frequency modulation at a 10-kHz rate, with an associated amplitude modulation of less than 5%. This technique has been employed to phase lock two microchip lasers and introduced less than 0.1% amplitude modulation on the slave laser. When it can be used, pump-power modulation has advantages over other frequency-modulation techniques since it requires very little power, no high-voltage electronics, no special mechanical fixturing, and no additional intracavity elements.

Polarization Control

Linear polarization is easily achieved in microchip lasers that employ an anisotropic gain medium, but can be problematic for monolithic microchip lasers with isotropic gain media. For lasers with isotropic gain media, the polarization degeneracy of the gain medium can often be removed by applying uniaxial transverse stress or asymmetric heat sinking. In the absence of any other polarizing mechanism, the polarization of the pump light, or asymmetry in its transverse mode profile, may determine the polarization of the laser output. In either of these cases, the polarization selectivity can be weak and the polarization of the laser may be sensitive to perturbations, including external feedback.

In passively Q-switched microchip lasers that employ an isotropic gain medium it is often the properties of the saturable absorber that determine the polarization of the laser.

Pulsed Operation

Pulsed output has been obtained from microchip lasers using a variety of techniques, including active Q switching, passive Q switching, and mode locking.

The shortest output pulse that can be obtained from a Q-switched laser has a full-width at half maximum of 8.1 times the round-trip time of light in the laser cavity divided by the natural logarithm of the round-trip gain. Since microchip lasers are physically very short, they have very short round-trip times. This leads to the possibility of producing very short Q-switched output pulses.

Active Q Switching

Actively Q-switched microchip lasers can be realized using the coupled-cavity configuration illustrated in [Fig. 2](#). An etalon containing an electro-optic element (LiTaO₃) serves as a variable-reflectivity output coupler for a gain cavity defined by the two reflective surfaces adjacent to the gain medium (Nd:YAG). The reflectivity of the etalon for the potential lasing frequencies of the device is controlled by a voltage applied to a pair of electrodes in contact with the electro-optic material.

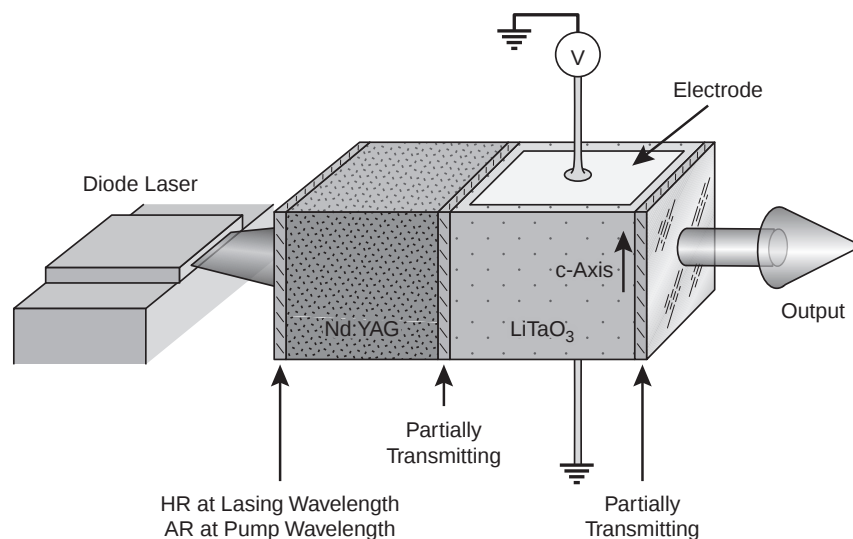


Fig. 2 Coupled-cavity electro-optically Q-switched microchip laser. AR, antireflective; HR, highly reflective.

Coupled-cavity electro-optically Q-switched microchip lasers have demonstrated the shortest Q-switched pulses obtained from any actively Q-switched solid-state laser to date. A coupled-cavity Nd:YAG device, pumped with a 500-mW diode laser, produced 270-ps pulses with a pulse energy of 6.8 μJ at a repetition rate of 5 kHz. A coupled-cavity Nd:YVO₄ microchip laser pumped at the same power level produced 115-ps pulses with a pulse energy of 12 μJ at a 1-kHz repetition rate. Pumped slightly harder, with a pump power of 1.2 W, a Nd:YVO₄ microchip laser demonstrated 8.8-ns pulses at pulse repetition rates as high as 2.25 MHz.

Passive Q Switching

Passively Q-switched microchip lasers contain a gain medium and a saturable absorber, as shown in the top of Fig. 3. When the gain medium absorbs sufficient pump energy that the gain in the laser cavity is greater than the loss, an intracavity optical field forms. The optical field saturates the loss of the saturable absorber and the gain of the gain medium. When the materials are chosen properly, the loss of the saturable absorber saturates more quickly than the gain of the gain medium and the laser will Q switch, generating a short, intense pulse of light without the need for high-voltage or high-speed electronics.

In addition to simplicity of implementation, the advantages of a passively Q-switched laser include the generation of pulses with a well-defined energy and duration. The pulse energy and duration are determined by the design of the laser cavity and the material properties of the gain medium and passive Q switch. Passively Q-switched microchip lasers have demonstrated pulse energies and pulse widths with stabilities of better than 1 part in 10^4 . This is achieved at the expense of pulse-to-pulse timing jitter, caused primarily by fluctuations in the pump source.

To manage thermal effects, high-power passively Q-switched microchip lasers are often pulse pumped. A common implementation uses an external clock to turn the pump diodes on and a signal generated by the Q-switched output pulse to turn them off. By using high-power pump diodes at a low duty cycle, larger amounts of energy can be stored in the gain medium of the laser with reduced thermal loading. This results in a larger oscillating-mode diameter and more energetic pulses.

Q switching with bulk saturable absorbers

The most commonly used bulk saturable absorber for passive Q switching of microchip lasers is Cr⁴⁺:YAG. It has been used to Q switch Nd:YAG microchip lasers operating at 946 nm, 1.064 μm , and 1.074 μm ; Nd:YVO₄ microchip lasers operating at 1.064 μm ; Nd:GdVO₄ microchip lasers operating at 1.062 μm ; and Yb:YAG microchip lasers operating at 1.03 μm .

The combination of Cr⁴⁺:YAG and a doped YAG gain medium, such as Nd:YAG, is particularly attractive from the point of view of an extremely robust device. Since both materials use the same host crystal, YAG, they can be diffusion bonded to each other in a way that blurs the distinction between a monolithic and a composite-cavity device. Both materials have the same thermal and mechanical properties and the same refractive index, and the bond between them can be sufficiently strong that the composite device acts in all ways as if it were a single crystal. An early, low-power, diffusion-bonded Nd:YAG/Cr⁴⁺:YAG microchip laser is shown in the bottom of Fig. 3.

As an alternative to diffusion bonding, Nd:YAG can be epitaxially grown on Cr⁴⁺:YAG and vice versa, using nonequilibrium growth techniques that can produce Nd or Cr⁴⁺ concentrations that could not otherwise be achieved. Cr⁴⁺:Nd:YAG can also be grown as a single crystal, although this approach precludes some of the device optimization that can be achieved when the two materials are physically distinct.

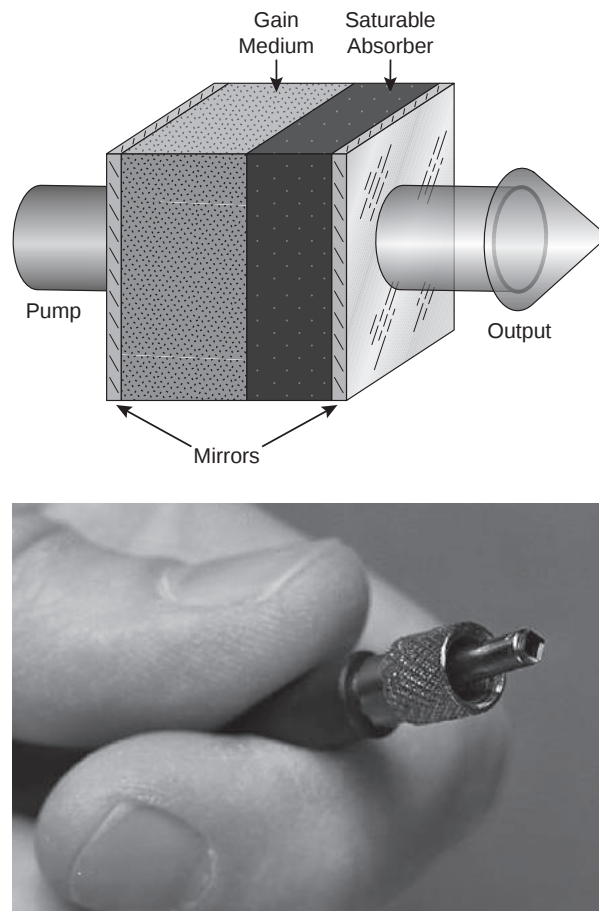


Fig. 3 Composite-cavity Nd:YAG/Cr⁴⁺:YAG passively Q-switched microchip laser: (top) schematic and (bottom) photograph of laser bonded to ferrule of pump fiber.

Typically, Nd:YAG/Cr⁴⁺:YAG passively Q-switched microchip lasers are operated with an output coupling approximately equal to the single-pass loss of the unsaturated saturable absorber. The minimum pulse width that can be obtained is limited by the average inversion density that can be achieved within the oscillating-mode volume, pump-induced bleaching of the saturable absorber, and a reduction in the gain cross section caused by heating of the gain medium as the laser is pumped harder. Passively Q-switched Nd:YAG/Cr⁴⁺:YAG microchip lasers typically produce pulses with full-widths at half maximum between 300 ps and ~1 ns, and pulses as short as 150 ps have been demonstrated.

The maximum pulse energy that can be obtained from passively Q-switched Nd:YAG/Cr⁴⁺:YAG microchip lasers is limited by the amount of stored energy the fundamental mode can access, and is strongly influenced by thermal effects. Devices pumped with a 1-W diode laser can generate 1.064- μ m pulses with energies up to ~15 μ J and peak powers up to ~30 kW, at pulse repetition rates of ~10 kHz. Less energetic pulses, with lower peak powers, have been demonstrated at repetition rates up to 110 kHz. By pulse pumping microchip lasers with high-power diode-laser arrays, pulse energies up to ~30 μ J with peak powers up to ~100 kW are achieved at repetition rates of ~10 kHz; pulse energies of ~250 μ J with peak powers up to ~500 kW are obtained at repetition rates of ~1 kHz; and pulse energies of several millijoules with peak powers of several megawatts are achieved at repetition rates up to ~100 Hz.

The combination of Yb:YAG and Cr⁴⁺:YAG has the same potential as Nd:YAG and Cr⁴⁺:YAG for quasi-monolithic and monolithic integration. The longer upper-state lifetime of Yb:YAG is attractive for low-repetition-rate systems since it allows the gain medium to accumulate energy for a longer time, making it possible to use lower-power, less-expensive pump diodes.

To date, the best results in the eye-safe spectral region were obtained from passively Q-switched microchip lasers that use Yb,Er:glass as the gain medium and Co²⁺:MgAl₂O₄ (Co²⁺:MALO) as the saturable absorber. When pumped with a 1-W 975-nm diode laser, these devices have produced output pulses of ~5-ns duration at repetition rates up to 25 kHz, with peak powers as high as 1.6 kW and average output powers up to 150 mW. At low duty cycles, pulses as short as 880 ps, pulse energies as high as 110 μ J, and peak powers in excess of 35 kW have been demonstrated.

Other combinations of gain medium and bulk solid-state saturable absorber have been used in Q-switched microchip lasers operating at a variety of wavelengths. Cr⁴⁺:YAG has been used with numerous gain media at wavelengths between 0.91 and 1.08 μ m. The Cr⁵⁺-vanadates are a relatively new family of saturable absorbers that are used in the same spectral region. Although lasers based on them have not yet demonstrated the short pulse durations or high peak powers that are obtained with Cr⁴⁺:YAG,

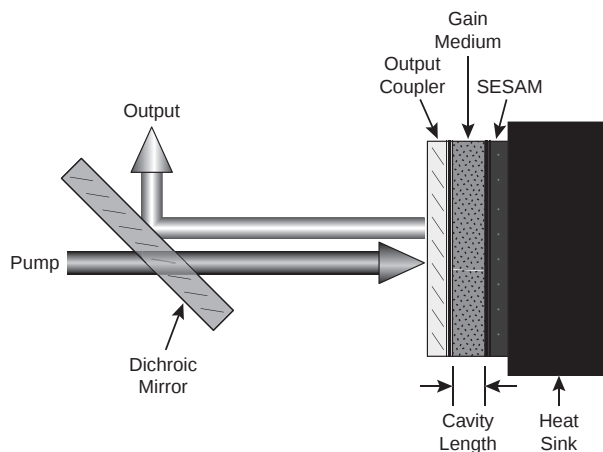


Fig. 4 Typical configuration for a passively Q-switched microchip laser employing a semiconductor saturable-absorber mirror (SESAM) Q switch.

they offer the potential for quasi-monolithic and monolithic integration with vanadate gain media. V^{3+} :YAG has been used as a saturable absorber over the spectral range from 0.93 to 1.44 μm . It is most useful in the long-wavelength portion of this range, where Cr^{4+} :YAG is not an option. Co^{2+} : $\text{LaMgAl}_{11}\text{O}_{19}$ (Co^{2+} :LMA) and Co^{2+} :MALO extend the coverage of solid-state saturable absorbers into the eye-safe spectral region.

Bulk semiconductor saturable absorbers have been used to Q switch miniature solid-state lasers at wavelengths from 1 to 2 μm , but have a much lower damage threshold than the solid-state saturable absorbers discussed above and are therefore rarely used. They also have very different thermal and mechanical properties than solid-state gain media, making robust integration of miniature devices challenging.

Q switching with semiconductor saturable-absorber mirrors

Semiconductor saturable-absorber mirrors (SESAMs) contain quantum-well saturable absorbers. When high-intensity light at the proper wavelength is incident on the mirror the absorption of the quantum wells saturates and the reflectivity of the mirror increases. In a passively Q-switched microchip laser employing a SESAM, the SESAM is used as one of the cavity mirrors.

When a SESAM is used to Q switch a microchip laser, the physical length of the saturable-absorber region of the microchip cavity is small and its contribution to the round-trip time of light in the laser cavity is negligible, resulting in the shortest possible Q-switched pulses. SESAMs have extremely short upper-state lifetimes, which allow Q-switched lasers employing them to operate at very high pulse repetition rates. One of the main limitations of SESAMs is their relatively low damage threshold.

A practical consideration when using SESAMs is that they typically cannot be used as input or output couplers. As a result, the output coupler is usually on the pump-side face of the laser, as shown in Fig. 4. Also, the thermal-expansion coefficients of SESAMs are not well matched to solid-state gain media, making robust bonding of the gain medium to the saturable absorber challenging.

As a result of their advantages and limitation, SESAMs are most attractive in applications requiring short, low-energy pulses, and where requirements on the system's robustness can be relaxed. In this regime, they can be operated at very high repetition rates, can produce extremely short pulses, and can be engineered to work with gain media at many different wavelengths.

$\text{Nd}:\text{YVO}_4$ microchip lasers passively Q switched with SESAMs have produced 1.064- μm output pulses as short as 16 ps, and have been pulsed at repetition rates up to 7 MHz. SESAMs have also been used to Q switch microchip lasers operating near 1.03, 1.34, and 1.5 μm . The largest pulse energy reported for a passively Q-switched microchip laser using a SESAM is 4 μJ , with tens to hundreds of nanojoules being more typical.

Mode locking

Although many microchip lasers are designed to operate in a single longitudinal mode, they need not be. Optical-domain generation of millimeter-wave signals for fiber radio led to the development of monolithic electro-optically mode-locked 1.085- μm $\text{Nd}:\text{LiNbO}_3$ microchip lasers with pulse widths as short as 18.6 ps and repetition rates up to 20 GHz. In this application, the pulse repetition rate of the laser is the radio carrier frequency.

Frequency Conversion and Amplification

Nonlinear frequency generation is an important adjunct to microchip laser technology. The high peak powers obtained from Q-switched microchip lasers have enabled a variety of miniature nonlinear optical devices. Harmonic generation, frequency

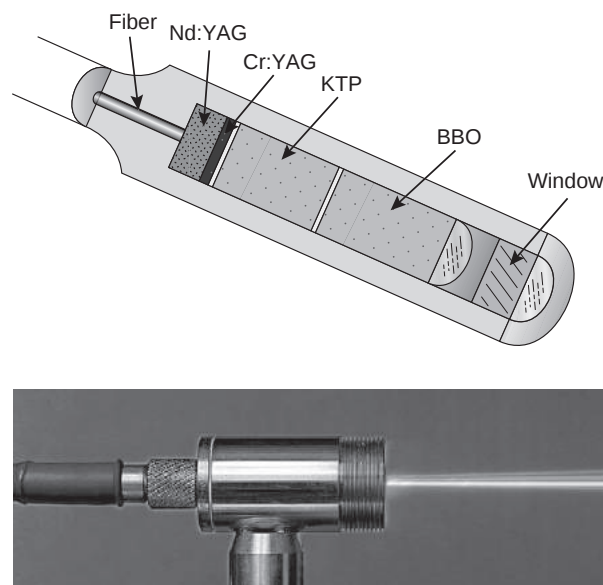


Fig. 5 Fiber-coupled 266-nm frequency-quadrupled passively Q-switched microchip laser: (top) schematic and (bottom) photograph of working device mounted on 12.7-mm-dia. post.

mixing, parametric conversion, and stimulated Raman scattering have been used with passively Q-switched microchip lasers for frequency conversion to wavelengths covering the entire spectrum from 213 nm to 8.1 μm in extremely compact optical systems.

Many applications for microchip lasers require harmonic conversion of the laser's infrared output. Because of its high peak intensity, the output of even low-average-power Nd:YAG/Cr⁴⁺:YAG passively Q-switched microchip lasers can be efficiently harmonically converted by placing the appropriate nonlinear crystals near the output facet of the laser with no intervening optics, as shown in Fig. 5. With this approach, a 1-W-pumped 1.064- μm Nd:YAG/Cr⁴⁺:YAG passively Q-switched microchip laser has been frequency converted to produce 7 μJ of second-harmonic (532-nm green), 1.5 μJ of third-harmonic (355-nm UV), 1.5 μJ of fourth-harmonic (266-nm UV), and 50 nJ of fifth-harmonic (213-nm UV) light at a typical pulse repetition rate of 10 kHz.

As an alternative to nonlinear frequency generation, the output of passively Q-switched microchip lasers and its harmonics have been used as a pump to gain switch miniature lasers in the infrared, visible, and ultraviolet portions of the spectrum. The high peak powers of Q-switched microchip lasers have also been exploited in fibers to generate white-light continua, and sub-200-fs pulses through a process that involves self-phase modulation, spectral filtering, and pulse compression.

For many applications the output power, or energy, produced directly by a microchip laser is sufficient. For numerous others, some amplification is required. A wide variety of amplifiers have been used in conjunction with microchip lasers, in systems that take advantage of the waveforms and near-ideal mode properties that they produce.

Applications

Within their range of capabilities, microchip lasers are the simplest, most compact, and most robust implementation of solid-state lasers. CW microchip lasers face strong competition from diode lasers and fiber lasers. Diode lasers are more efficient, smaller, and simpler, and are available at a greater variety of wavelengths. Fiber lasers also tend to be more efficient and can produce higher output powers in a fundamental transverse mode. To compete against either of these technologies, CW microchip lasers need to find a niche application that exploits their unique spectral properties. Nonetheless, the CW microchip geometry has established itself as a testing ground for newly developed gain media.

On the other hand, Q-switched microchip lasers provide capabilities that cannot be matched by semiconductor devices or fiber lasers. Semiconductor diode lasers have a very limited capacity to store energy, and their facets are damaged at very modest optical intensities. Fiber lasers have much longer cavities, which prevent them from producing short Q-switched pulses and make them susceptible to undesirable nonlinear effects at peak powers well below those demonstrated with Q-switched microchip lasers. Fiber amplifiers can be used to increase the peak power from pulsed semiconductor diodes, but this requires several stages of amplification, results in more complicated systems, and is still limited by fiber nonlinearities. Additionally, the amplifier output has interpulse amplified spontaneous emission that may be detrimental for some applications.

As a result of their small size, robust construction, reliability, and relatively low cost, coupled with their ability to produce energetic, diffraction-limited, Fourier-transform-limited, sub-nanosecond pulses, passively Q-switched microchip lasers have been embraced for applications in high-resolution time-of-flight three-dimensional imaging. Because their high peak output intensity allows for efficient nonlinear generation of ultraviolet light in very compact and reliable formats, they have also been well accepted

in the field of ultraviolet fluorescence spectroscopy, and are an integral part of numerous fielded spectroscopic instruments. In addition, passively Q-switched microchip lasers have made inroads in laser scribing and marking, laser-induced breakdown spectroscopy, and most recently laser ignition. The continued development of the technology will be driven by applications.

Further Reading

- Zayhowski, J.J., 2013. Microchip lasers. In: Denker, B., Shklovsky, E. (Eds.), *Handbook of Solid-State Lasers: Materials, Systems and Applications*. Cambridge: Woodhead Publishing (Chapter 14).
- Zayhowski, J.J., Welford, D., Harrison, J., 2007. Miniature solid-state lasers. In: Gupta, M.C., Ballato, J. (Eds.), *The Handbook of Photonics*, second ed. Boca Raton, FL: CRC Press (Chapter 10).

Supercontinuum Generation

James R Taylor, Imperial College London, London, United Kingdom

© 2017 Elsevier Inc. All rights reserved.

Introduction

Although nonlinear optical effects had been investigated in the 19th century, for example, Kerr in 1875 examined induced birefringence in isotropic media as a result of an applied DC electric field, showing it to be proportional to the field intensity, while 20 years later, Pockels demonstrated that the birefringence induced in piezoelectric crystals linearly depended on the applied DC field, however, it was not until just after the invention of the laser in 1960, particularly in its early pulsed or relaxation oscillation manifestations, that the electric field associated with the optical pulse was sufficiently intense to allow nonlinear optical conversion of the incident laser signal to be observed directly. Within a few years of the demonstration of the ruby laser by Maiman, the techniques of Q-switching and mode-locking were reported, which allowed the controlled generation of nanosecond pulses using the former method and picoseconds with the latter. Subsequently, over the following three decades, innovative mode-locking schemes were devised in various families of lasers to routinely allow the generation of controlled pulse durations of only a few femtoseconds, noting that in the visible around 600 nm a single optical cycle is approximately 2 fs.

Consequently, for relatively modest pulse energies, it is possible to generate peak powers in the focal region of a conventional convex lens exceeding a terawatt per square centimeter, with corresponding electric field strength over 10 MV per centimeter. In a transparent dielectric medium, electric field strengths of such magnitude give rise to a nonlinear response, in that the induced polarization has to be described by an expansion that includes higher order terms, such that

$$P = \epsilon_0 \left(\chi^{(1)} E + \chi^{(2)} E^2 + \chi^{(3)} E^3 + \dots \right)$$

where P is the polarization, ϵ_0 is the permittivity of free space, and $\chi^{(n)}$ is the n th order susceptibility. The first term on the right hand side represents the linear response of the medium, which until 1960 was the most commonly encountered case, where the electric field strength E was relatively low and higher order terms could be neglected.

If one examines the second term in E^2 and considers an input optical field varying as $\sin \omega t$, where ω is the angular frequency of the radiation, an expansion of this term will give rise to components varying as $\sin 2\omega t$, a nonlinear response, leading to frequency doubling or so called second harmonic generation. This was the first laser-generated nonlinear process observed, using a simple focused pump scheme in crystal quartz. In media with a center of symmetry, the second order term vanishes to zero and second harmonic generation is generally not observed. The process of second harmonic generation is also not particularly relevant in the contribution to supercontinuum generation, where the more important effects arise from the third order term. Driven by at a fundamental field at $\sin \omega t$, the third order term gives rise to a component response at $\sin 3\omega t$, generating the third harmonic. Once again, this is of little impact for supercontinuum generation, however, other nonlinear contributions arise from the third order term, such as the optical Kerr effect or intensity dependent refractive index, four-wave mixing, modulational instability, stimulated Raman scattering, and stimulated Brillouin scattering, although the latter also plays a minimal role in supercontinuum generation. Through a judicious choice of system parameters, such as pump wavelength, pump duration, bandwidth, system dispersion, and power, a particular nonlinear process may be chosen to dominate or alternatively many of the processes can be selected to occur simultaneously.

Spectral Broadening and Early Supercontinuum Sources

In early realizations of pulsed, Q-switched and mode-locked solid state lasers, anomalous spectral broadening of the output, much greater than the expected gain linewidth of the transition, was widely reported upon. This was also observed in oscillator–amplifier configurations and in particular where catastrophic optical damage to the gain medium was simultaneously recorded. Spectral broadening was also observed through focusing in highly nonlinear liquids, such as CS_2 . It was theoretically proposed that the nonlinear refractive index, both in the temporal and the spatial regimes, was responsible for these observations. If one considers the overall refractive index n_{tot}

$$n_{\text{tot}} = n_0 + n_2 I(x, t)$$

where n_0 is the refractive index as the intensity of the propagating field approaches zero, the low power case that is more commonly encountered and n_2 is the space and time intensity $I(x, t)$ dependent refractive index. For silica n_2 is approximately $3.2 \times 10^{-20} \text{ m}^2 \text{ W}^{-1}$, while in CS_2 it is approximately two orders of magnitude greater. In the spatial domain if one considers a Gaussian transverse mode profile propagating in a nonlinear medium, the intensity at the center is greater than in the wings. Consequently the total refractive index is greater centrally than in the wings of the mode, such that the wings will travel faster than the center. This causes the mode to collapse in on itself – to self-focus. It was observed that significant spectral broadening was observed beyond a critical power where this catastrophic collapse led to the formation of self-trapped filaments. Of course, these self-trapped filaments were highly irreproducible in space and in time, as were the associated spectra.

In the time domain, the accumulated phase difference on propagating over a distance L of a pulse of intensity $I(t)$ is given by $n_2 k L I(t)$, where k is the wavenumber. Since a frequency change is simply the negative time derivative of phase and as n_2 , k , and L are constants, the frequency shift is simply proportional to the negative time derivative of the pulse intensity, while assuming that the time response of the third order nonlinearity is ultrafast ($\sim$ fs) compared to the incident pulses, as was first explained by Shimizu and labeled self-phase modulation (SPM). As a consequence of SPM, the front of a pulse is frequency down shifted, i.e., red shifted, while the rear is frequency up (blue) shifted and a quasi-linear frequency shift exists throughout the central region of the pulse. The spectral broadening experienced is proportional to the propagation length and the intensity. In the time domain, SPM plays a vital role in the generation of temporal optical solitons, where in the anomalously dispersive regime for a specific power a balance can exist between dispersion and nonlinearity such that a pulse can propagate without dispersive temporal broadening. In optical fiber, solitons play one of the most important roles in supercontinuum generation.

Despite many reports of anomalous spectral broadening in laser systems, the first recognized study and report of supercontinuum generation, although not thus designated, was carried out by Alfano and Shapiro in 1970 by focusing the frequency doubled output of a pulsed, mode-locked Nd laser into bulk samples of BK7 glass at a power density of $\sim 1 \text{ GW cm}^{-2}$. The resulting beam filamentation gave rise to a white light spectrum that covered the range from 400 to 700 nm, which was substantially greater in extent than anything previously reported. Four-wave mixing was proposed as the principal formation process. Although very simple in the experimental configuration deployed, because of the uncontrollability of the formation mechanism, the generated spectra were not particularly reproducible. Stable supercontinua covering the range from 190 to 1600 nm were achieved by focusing the output of an amplified 80 fs dye laser at 627 nm to a power density of 10^{13} – $10^{14} \text{ W cm}^{-2}$ into a 500 μm thick jet stream of ethylene glycol, however, the energy density in the extremes of supercontinuum were up to five orders of magnitude lower than that around the pump and the overall average power was low. Other liquids such as water, carbon tetrachloride, and various fluorocarbons have also been used. In bulk samples, self-focusing and SPM have been proposed and identified as the primary spectral broadening mechanisms, however, the process is somewhat complicated by the formation of an optical shock generated at the rear of the pump due to spatial and temporal focusing as well as self-steepening. Visible supercontinuum have also been generated by focusing amplified femtosecond dye laser pulses into cells of high pressure gases, such as Xe, H_2 , and N_2 and in millimeter lengths of optical fibers, however, the large footprint of the pump laser/amplifier schemes together with bulk lens coupling to the samples restricted their application primarily to research laboratories.

Modeling of Supercontinuum Generation

The formation of optical solitons plays an underpinning role in supercontinuum generation. Essential for the generation process is the balance that arises between linear and nonlinear effects. This can happen both in the spatial domain, where the balance is achieved between self-focusing and diffraction or in the temporal domain where it is between SPM and dispersion. In optical fiber-based systems, temporal soliton formation is the dominant process, while, for example, in early experimental schemes where supercontinuum generation occurred through the focusing of a pulsed laser into a cell containing a nonlinear liquid or into bulk material, spatial soliton formation principally contributed. These simple systems can be described by the nonlinear Schrödinger equation (NLSE). In the case of a complex optical pulse envelope $U(x,t)$ propagating in a fiber, normalized such that $|U(z,\tau)|^2 = P(z,\tau)$, where $P(z,\tau)$ is the instantaneous power in Watts, z is the position in the fiber, and τ is a co-moving time frame, the NLSE can be written as

$$\partial_z U = -i \frac{\beta_2}{2} \partial_\tau^2 U + i\gamma |U|^2 U$$

where $\beta_2(\omega) = \partial^2 \beta / \partial \omega^2$ and $\beta(\omega)$, the axial wavenumber, $= n_{\text{eff}} \omega / c$, where n_{eff} is the effective modal refractive index, ω the angular frequency and c the speed of light. The nonlinear coefficient $\gamma = n_2 \omega / c A_{\text{eff}}$ (where n_2 is the nonlinear refractive index and A_{eff} the effective modal area). The first term on the right hand side of the NLSE above relates to the dispersive term and the second term is the nonlinear balancing contribution and the equation gives rise to solutions of the form

$$U(z, \tau) = \sqrt{P_0} \operatorname{sech}\left(\frac{\tau}{\tau_0}\right) \exp\left(i\gamma \frac{P_0}{2} z\right)$$

where the soliton peak power $P_0 = |\beta_2| / \gamma \tau_0^2$ and τ_0 is the soliton duration. For powers launched into a fiber where $P = N^2 P_0$, where N is an integer, high order soliton behavior is observed. In the early stages of propagation this leads to a rapid temporal compression, however, as the simple NLSE description is unable to adequately describe the evolution of such broad bandwidth pulses, additional terms are needed. For example, the effects of high order dispersion need to be included as excessive bandwidth is encountered, while broad bandwidth operation also gives rise to the need to consider Raman scattering effects, in particular self-Raman interaction or the so called soliton self-frequency shift, where short wavelength components in a pulse can provide Raman gain for the longer wavelengths. Historically, each of these modifications to consider each effect were introduced separately, however, all effectively reduce to the same equation, the generalized nonlinear Schrödinger equation (GNLSE), which is valid for pulse durations down to a couple of optical cycles. In the frequency domain it is given by

$$\partial_z \tilde{U} = i \left(\beta(\omega) - \frac{\omega}{v_{\text{ref}}} \right) \tilde{U} + i \frac{n_2 \omega}{c A_{\text{eff}}} \mathcal{F}[(R \ast |U|^2) U]$$

where $\tilde{U}(z, \omega) = \mathcal{F}[U(z, t)]$ and $\mathcal{F}$ is the Fourier Transform, v_{ref} a chosen reference velocity and $R(t)$ is the nonlinear response function. For silica glass $R(t)$ takes the form $R(t) = (1 - f_r)\delta(\tau) + f_r h_R(t)$, an instantaneous Kerr contribution and a delayed Raman response described by h_R . This GNLSE includes the full modal dispersion, the Raman effect and the effect of self-steepening, which leads to optical shock formation. The equation can also be modified to include, for example, the frequency dependence of the effective area of the fiber deployed or third harmonic generation.

Another vitally important ingredient in the modeling of supercontinuum generation is the inclusion of the contribution of noise, this is particularly relevant when concerned with the process of modulational instability. Most commonly this is managed through the addition of quantum noise fluctuations to the initial conditions, through the addition of a field with a spectral power equivalent of one photon per mode with random phase. In addition the noise on the actual pump source needs also to be included and this is of particular relevance to cw pumped systems.

Supercontinuum Generation in Optical Fiber

Optical fiber is an ideal medium for the study of nonlinear optics. In bulk samples, the nonlinear interaction length is limited by the confocal parameter of the focusing lens, which is the order of a millimeter in the visible for a 10 μm waist. Driven by the needs of telecommunications, silica-based glass fibers have been produced with losses better than 0.2 dB km⁻¹. Consequently, signals can propagate tens of kilometers, confined to core diameters of a few microns without significant reduction in the signal intensity. Neglecting processes that require phase matching, since nonlinearity is a power times length process, in-fiber generation can have an enhancement of up to 10⁶–10⁷ compared to that in the bulk, or equivalently the peak power requirement to observe an effect is reduced by this factor and nonlinear effects can be observed at modest pump power levels.

A nitrogen laser pumped broad band dye laser (~ 15 nm) was used as the fundamental source for the first supercontinuum generation in optical fiber as reported by Lin and Stolen (1976). A power of 1 kW in a 10 ns pulse around 434 nm launched into a 20 m long multimode 7 μm diameter fiber generated a “white light” spectrum extending from 434 to 614 nm. The dominant contributing process was cascaded, stimulated Raman scattering and consequently the generated supercontinuum was only to the long or Stokes wavelength side of the pump. By 1978, the more common approach of using a pulsed Nd:YAG laser was used to produce a supercontinuum that extended from about 700 to 2100 nm, using 50 kW, 20 ns pulses from a Q-switched system to excite 315 m of multimoded fiber with a 33 μm core diameter. Again cascaded Raman generation was the principal generation mechanism. Fig. 1 shows a typical supercontinuum generated in a standard 7 μm core silica fiber, with a dispersion zero around 1310 nm, using 80 ps pulses from a mode-locked Nd:YAG laser at 1060 nm.

The cascade of Raman orders is clearly seen in Fig. 1 evolving sequentially from the pump at 1060 nm in this example, from first to fourth labeled 1S through to 4S, respectively. SPM and cross phase modulation from the pump and between the Raman orders can give rise to additional spectral broadening but the individual orders are clearly resolved, each separated by approximately 440 cm⁻¹. Beyond the dispersion zero of the fiber around 1310 nm, the fourth Stokes component is seen to evolve into a continuum. This soliton-Raman continuum is made up of randomly distributed femtosecond soliton pulses, with the evolution proceeding via modulational instability. Closely related to soliton generation and four-wave mixing, modulational instability results from the interplay of anomalous dispersion and the intensity dependent refractive index. Amplitude or phase modulations on an effective continuous wave background, for example, the femtosecond amplified spontaneous emission (ASE) noise fluctuations on a tens of picosecond pulse, exhibit an exponential growth rate that is accompanied by a spectral sideband evolution at a frequency separation from the carrier that is proportional to the optical pump power. For exponential growth, the upper and

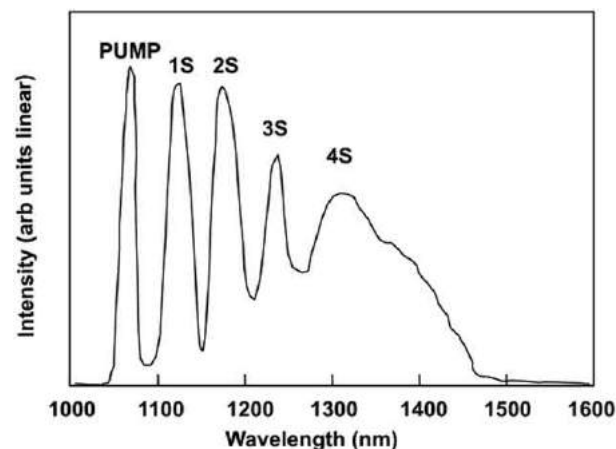


Fig. 1 Representative supercontinuum spectrum generated in a standard telecommunications fiber pumped by the 80 ps, mode-locked pulses from a Nd:YAG laser. Reproduced from Alfano, R.R. *Supercontinuum Laser Source the Ultimate White Light*, third ed. New York, NY: Springer. ISBN 978-1-4939-3324-2.

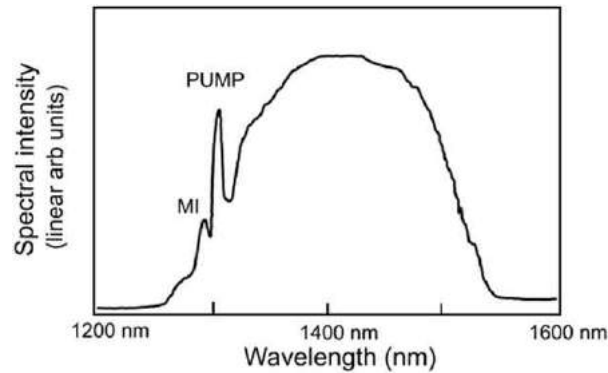


Fig. 2 Modulational instability initiated high average power soliton-Raman supercontinuum generated in 200 m of standard silica fiber at an average power of 1 W. Reproduced from IEEE Journal of Quantum Electronics 24, 332–340 (1988).

lower sideband frequency separation from the carrier must be less than a critical frequency given by $(4\gamma P_0/\beta_2)^{1/2}$, where P_0 is the incident power, γ is the nonlinear coefficient and β_2 is the group velocity dispersion parameter, while the maximum sideband growth occurs at a frequency separation of $(2\gamma P_0/\beta_2)^{1/2}$. The process is usually self-starting and from system noise but can also be seeded using an external signal lying within the gain bandwidth. Modulational instability plays an underpinning role in the initiation, amplification, and evolution of femtosecond pulses, which with subsequent self-Raman interaction and collisions form the basis of supercontinuum generation under picosecond, long pulse, or even cw pumping.

The role of modulational instability can be seen in Fig. 2, which is a soliton-Raman continuum generated from a cw mode-locked Nd:YAG laser generating 100 ps pulses at 1318 nm, with a launched power of approximately 1 W (~ 100 W peak power) in 200 m of silica fiber with a dispersion zero at 1290 nm. Significant depletion of the pump is observed and the anti-Stokes component of the originating modulational instability signal around 1295 nm is apparent. Following the rapid evolution of femtosecond solitons, the so-generated pulses exhibit self-Raman interaction (soliton self-frequency shift), where the bandwidth of the short pulse soliton is sufficient for the short wavelength component to provide significant Raman gain for the long wavelength components. This gives rise to a continuous long wavelength spectral shift with propagation length and one that also increases with pump power. The effects of noise amplification and soliton-soliton collisions also contribute to the spectral broadening and to the random temporal soliton structure of the output. On spectral filtering such a continuum, pulses as short as 50–80 fs can be selected but with large temporal jitter.

It should also be noted that only significant continuum generation is observed to the long (Stokes wavelength side) of the pump.

Solitons and Photonic Crystal Fiber

Although proposed by Hasegawa in 1973 and most certainly they were generated in the fiber-based supercontinua published by Lin *et al.* (1978), the optical soliton was not demonstrated unequivocally until 1980 by Mollenauer and colleagues in a series of innovative experiments that characterized their unique properties. The reason for the delay was two-fold. Only by the late 1970s were adequate lengths of low loss single mode optical fiber available and since the minimum dispersion zero wavelength of conventional silica-based fiber is 1270 nm, significant effort had been placed on the development of tuneable picosecond pulse sources to instigate soliton operation above that wavelength in the anomalously dispersive regime.

By the mid-1980s many laboratories were directing their efforts toward fiber-based sources of ultrashort pulses utilizing soliton effects for extreme compression. The power P_0 , required to establish a fundamental soliton that will propagate over its characteristic nonlinear length without change in the pulse shape and duration is fixed and proportional to $A_{\text{eff}}\lambda^3 D/\tau^2$, where A_{eff} is the effective core area of the fiber, λ the wavelength of the radiation, D the group delay dispersion at that wavelength, and τ the pulse duration. If one launches a pulse power $N^2 P_0$, where N is an integer, a high order soliton is established, which is a nonlinear superposition of N fundamental solitons. It has been demonstrated both theoretically and experimentally that on propagation, pulse narrowing and periodic splitting and recombination takes place as a result of the periodic interference between these N solitons. An optimal compression ratio of $4.1N$ is achieved with a corresponding increase in the associated spectral width. For the lower orders, for example, $N=2$ or 3 , etc., periodic breathing of the compression and restoration of the original pulse width is observed on propagation, however, for relatively high order solitons, after extreme compression, instability occurs through self-Raman effects and higher order dispersion perturbations and break-up of the high order soliton into numerous single solitons occurs. Originally termed colored soliton generation, the process has been renamed soliton fission. Using higher order soliton compression, pulses as short as 18 fs, only four optical cycles, have been generated centered around 1318 nm with an associated spectral half intensity width of about 100 nm and baseline width around 300 nm.

It was not possible to observe soliton effects in conventional silica-based fibers using the more common short pulse sources, such as the Nd-doped glass and crystal lasers, Yb fiber lasers, or the Ti:sapphire laser. This was to dramatically change with the

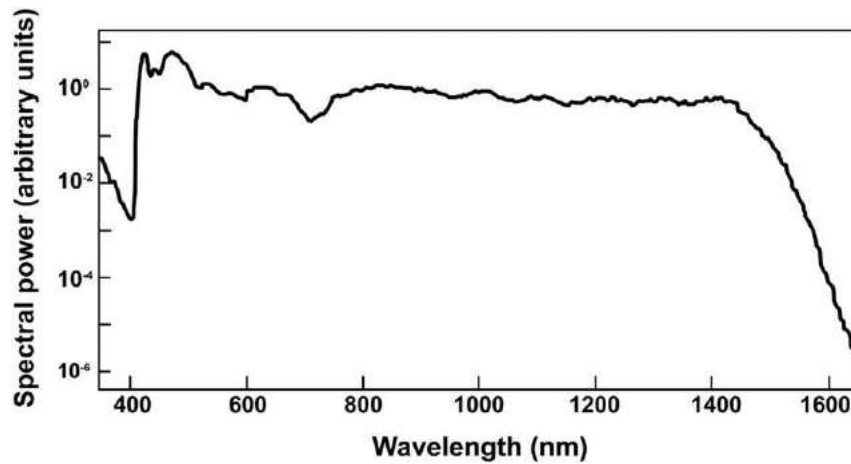


Fig. 3 Supercontinuum generated in a photonic crystal fiber with a dispersion zero at 770 nm, pumped by the 8 kW, 100 fs pulses of a Ti:sapphire laser at 790 nm. After Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25, 25–27.

introduction of the photonic crystal fiber (PCF) by the Russell group at the University of Bath in 1996. Through the adjustment of the pitch and diameter of the crystalline-like hole structure around the core, it became possible to precisely engineer the dispersion zero wavelength even into the visible spectral region. Enhanced nonlinearity can also be achieved through the controlled manufacture of smaller effective mode field diameters, while fibers can also be made to operate with single mode characteristic throughout extended spectral regions from the visible to the near infrared, in fact covering the complete window of transmission of silica. This culminated in 2000 in a report by Ranka and colleagues on the generation of a relatively modest average power supercontinuum extending approximately from 400 to 1600 nm (see Fig. 3) from a 75 cm length of photonic crystal fiber with a dispersion zero at 770 nm, pumped by the 8 kW, 100 fs 790 nm pulses from a mode-locked Ti:sapphire laser. This result was to rejuvenate research into supercontinuum generation, lead to renewed studies of nonlinear optics in fiber and ultimately led to the development of an extremely successful commercial product.

Femtosecond Pulse Pumped Supercontinua

A remarkable feature of the femtosecond laser pumped supercontinuum shown in Fig. 3 is the relative flatness of spectral intensity, however, it must be remembered that this results from the temporal integration of many millions of spectra generated by the individual pulses of the pumping mode-locked signal. Also of particular note is the complete depletion of the pump signal around 790 nm and importantly the extension of the generated supercontinuum to the short wavelength side of the pump, effectively covering the complete visible region.

The principal mechanism initiating the generation process is the temporal compression of the high order soliton that takes place over a length scale given approximately by $\tau_0^2/N\beta_0$, where N is the soliton order, τ_0 the duration of the input soliton, and β_0 the dispersion parameter. This is accompanied by inherent spectral broadening. In Fig. 4, which is a theoretical simulation of the spectral evolution with length of an input, $N \sim 9$, 10 kW peak power, 50 fs soliton at 835 nm propagating in a PCF with a dispersion zero in the region of 780 nm, the result of the rapid temporal compression is evidenced within an initial 1 cm of propagation.

Pumping in the region of anomalous dispersion and following rapid compression, the soliton spectrum ingresses the normal dispersion regime and dispersive waves feed off. Beyond the point of maximum compression the system is affected by system instabilities that inhibit the breathing and periodic restoration of the high order soliton. Modulational instability, higher order dispersion and the soliton self-frequency shift result in the spectral and temporal fragmentation into numerous solitons, with the associated spectrum being further complicated by the effects of soliton collisions, four-wave mixing, for example, between the soliton and the dispersive wave, and cross phase modulation. Despite the apparent complications of the generation process, the physical processes and dynamics of formation are very well understood with theoretically predicted spectra agreeing exceptionally well with experimental measurement.

Cross phase modulation between solitons and dispersive waves in the region of the dispersion zero induces a chirp on the dispersive wave, with the rear of the dispersive wave being frequency up shifted and as a consequence of propagation in the normal dispersion regime is de-accelerated and retarded. The short pulse soliton on the other hand experiences the soliton self-frequency shift, frequency down shifting via the Raman gain process. As operation is in the anomalously dispersive regime the soliton too is de-accelerated, consequently overlapping temporally with the up shifted dispersive wave. This process of dispersive wave generation and trapping by solitons can lead to significant power transfer to the dispersive wave and is the principal

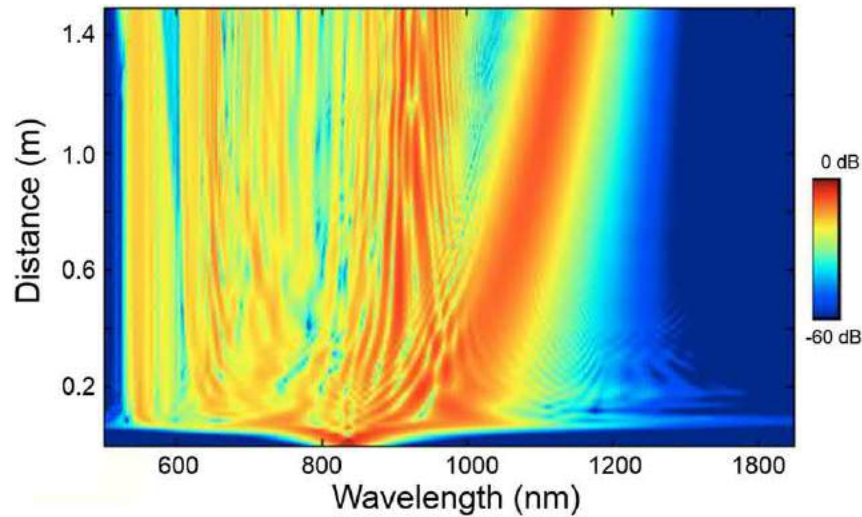


Fig. 4 Simulation of supercontinuum evolution with fiber length through high order femtosecond soliton compression and subsequent soliton fission. Courtesy J.C. Travers unpublished.

mechanism in the short wavelength extension of supercontinua. In **Fig. 4**, the result of soliton self-frequency shift is apparent. Following fragmentation of the highly compressed pulse after approximately 1 cm, a continuous long wavelength evolution with length can be seen. In many fiber structures, in the anomalously dispersive regime the group velocity dispersion increases with increasing wavelength. As a consequence, the required soliton power increases ($P_0 \propto D/\tau_0^2$) and the soliton accommodates for its now inherently low power by increasing in duration, however, with increased duration the effect of the soliton self-frequency shift is reduced ($\Delta\nu \propto \tau_0^{-4}$). Thus, an increasing anomalous dispersion with wavelength will inhibit the soliton self-frequency shift, long wavelength extension and consequently short wavelength extension of the supercontinuum.

Following rapid compression and soliton fission, signal noise can have a dramatic influence on the seeding of the modulational instability process. If intense femtosecond scale structures are present in the noise signature these can be rapidly amplified to soliton powers and can enhance the long wavelength self-frequency shifted edge of the supercontinuum, with the shortest most intense seeded signals giving the greatest spectral shift. Noise will strongly affect the evolving spectrum and considerable variation occurs in the spectra of the individual supercontinua generated by each individual pulse of the exciting mode-locked train – although the subtlety of this is lost in the integrated spectra as is normally recorded (see **Fig. 3**). This role of noise in the seeding of the long wavelength component of the supercontinuum has been likened to the generation of “rogue waves” in hydrodynamics.

Beyond the point of optimum pulse compression in high order soliton driven supercontinuum generation, the effects of higher order dispersion, self-Raman interaction, soliton fission, modulational instability, soliton–soliton collisions, and noise lead to spectral and temporal instability and spectral coherence is lost. To minimize the effects of noise and instability it is best to utilize fiber lengths cut to the optimal high order soliton compression length. Soliton effects can of course be totally negated by operating in a region of normal dispersion, such that the process of SPM dominates the spectral broadening process, however, for equivalent pump powers, the spectral coverage is generally reduced compared to that attainable with solitons.

Picosecond Pulse Pumped Supercontinua

Although providing adequate spectral coverage under femtosecond solid state laser pumping, the necessity of bulk coupling into PCF using lenses does not lend the technique to high stability in a relatively hostile environment and diminishes potential commercial impact and routine application. In addition, the use of femtosecond pulses also inhibits the average power scaling of the spectral power density of the generated supercontinua, since nonlinearity and dispersion in the amplifiers affects the efficiency of the amplification process. With the development of picosecond, master oscillator power fiber amplifier (MOPFA) configurations, primarily based upon picosecond pulse seeded Yb-doped schemes operating around 1060 nm, power scaling to hundreds of watts average power in compact, all-fiber geometries was possible, while it was also possible to completely fiber integrate the MOPFA and the PCF to allow high stability and reproducibility. Initial configurations utilizing average pump powers of a few watts allowed a minimum spectral power density of 1 mW nm^{-1} to be achieved throughout the spectral region 500–2000 nm. Since then, technological improvements in pump source technology have allowed the supercontinuum spectral power density to be increased to greater than 100 mW nm^{-1} . **Fig. 5** shows the supercontinuum generated in 30 m of PCF with a zero dispersion wavelength at 1040 nm pumped by 10 ps, 10 kW pulses at 1058 nm from an Yb fiber laser system.

Since the fundamental soliton power is proportional to τ^{-2} , where τ is the pulse duration, it may be thought that the process of high order soliton fission would play an important role in the development of the associated supercontinuum, however, the

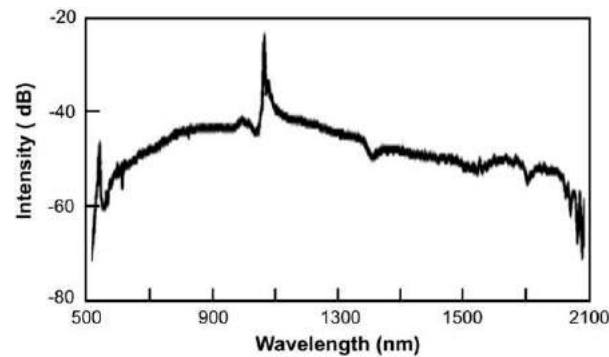


Fig. 5 Experimentally measured supercontinuum generated with 10 ps pulse pumping at a peak power of 10 kW in 30 m of photonic crystal fiber (PCF) with a zero dispersion wavelength of 1040 nm.

associated soliton period or nonlinear length scales as τ^2 and is many orders of magnitude greater than the few meters generally employed to observe substantial supercontinuum generation, hence soliton effects are not initially observed. For picosecond pumping in the region of the dispersion zero wavelength, modulational instability and four-wave mixing provide the principal initiation mechanisms. Modulational instability leads to the rapid break up of input picosecond scaled pulses into femtosecond scale subpulses, which are rapidly amplified to soliton powers. The modulational instability process can be significantly enhanced if the input pulses exhibit amplitude or phase fluctuations on their profile, which lie within the gain bandwidth of the modulational instability process and this noise can play an important role in the seeding of modulational instability and enhanced evolution of soliton structures. Subsequent four-wave mixing and soliton–soliton collisions lead to further spectral broadening, while the soliton self-frequency shift and soliton–dispersive wave trapping gives enhanced long and short wavelength extension respectively, as was described previously.

In early experimental realizations using pumping around 1060 nm in fibers with a dispersion zero wavelength around 1040 nm it was observed that the short wavelength extension of the supercontinua did not extend below about 550 nm, see, for example, [Fig. 5](#). This characteristic, initially and surprisingly, was relatively independent of the pump power and or the fiber length employed, but can be explained simply by imagining the requirement to match the group velocity of the short and long wavelength components of the supercontinuum in the soliton–dispersive wave interaction. To the short wavelength side, dispersion is dominated by the host material, while to the long wavelength side the waveguide plays a dominant role. A group velocity match is achieved straddling the dispersion zero at a fixed pair of wavelengths. As a femtosecond scaled soliton experiences the soliton self-frequency shift it moves to a longer wavelength, trapping with it the dispersive wave component that is locked to the corresponding short wavelength with a matching group velocity. In silica-based fibers, the waveguide and material losses increase as the wavelength approaches 2 μm and as the loss increases soliton propagation can no longer be sustained because of energy loss and associated temporal broadening. Therefore it is the limitation on the long wavelength operation that through the group velocity matching condition restricts the short wavelength extent of the trapped dispersive waves in the supercontinuum spectrum. [Fig. 6](#) shows a theoretical simulation of a spectrogram – the spectral distribution with time of a supercontinuum generated after 0.5 m of fiber by a 10 kW peak power 3 ps pulse at 1060 nm launched just above the dispersion zero wavelength 1040 nm. Discrete soliton structures are observed in the anomalously dispersive regime together with the corresponding trapped dispersive wave structures in the normally dispersive regime. In the temporal domain the overall format is noise-like with numerous randomly distributed intense soliton structures, which are also clearly identifiable in the spectrogram. The overall spectrum, which extends from about 550 to 2000 nm and is the result of a single input picosecond pulse is noisy and structured. Only with the integration of many pulses from the driving mode-locked laser source, does a supercontinuum take on the smooth structure representative of [Fig. 5](#), which is more commonly encountered.

Several techniques have been introduced to extend supercontinua spectra below 500 nm using 1 μm pumping. The first employed two concatenated PCFs. The first with a zero dispersion around 1040 nm and the second around 780 nm. When the continuum in the first fiber extends to the region around 780 nm the radiation around these wavelengths can successfully act as the pump in the second fiber, generating soliton-like structures around the dispersion zero and the process of soliton–dispersive wave trapping extends the spectrum. Since the shift in the zero dispersion of the second fiber is offset to shorter wavelengths the group velocity matching is similarly offset allowing matching to substantially shorter wavelengths. Pumping the fiber with a short zero dispersion wavelength alone would not give the same effect as the pump is too far away and any generated soliton structures would not give rise to dispersive waves in the normal dispersion regime. In such a case Raman would dominate the generation process, giving rise to a continuum that extended solely to the Stokes side of the pump pulse.

The technique was further extended using a length of tapered PCF, such that a continuous shift of the dispersion zero was produced over the length of the fiber. Such a structure enhances the group velocity matching condition, in addition, the tapered structure also leads to a decreasing group velocity with length. This leads to a de-acceleration of the soliton, equivalent to the action of the soliton self-frequency shift which is essential for soliton–dispersive wave interaction. Consequently, the tapered structures enhance the short wavelength achievement and supercontinuum generation has been demonstrated in 1.5 m long tapers

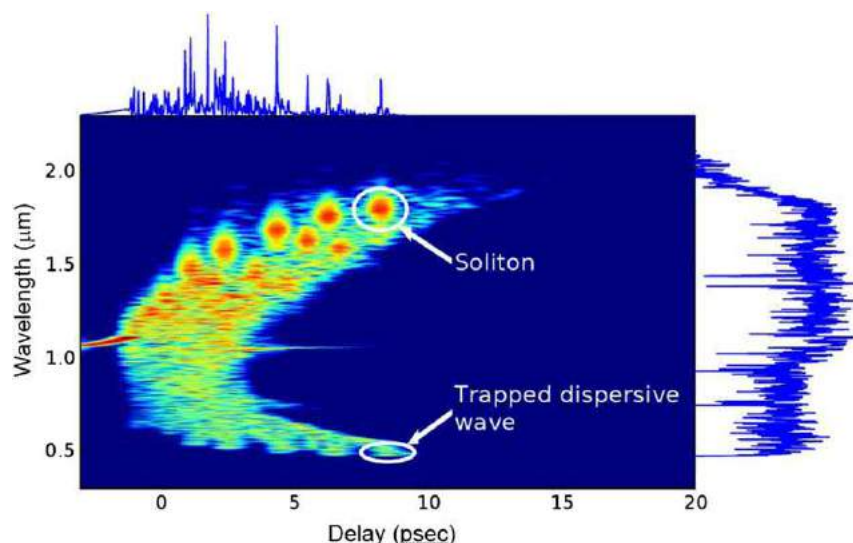


Fig. 6 Theoretical simulation of a spectrogram of a supercontinuum generated in 0.5 m of photonic crystal fiber (PCF) with a zero dispersion at 1040 nm pumped by 10 kW, 3 ps pulses at 1060 nm. The temporal structure of the output is shown above and the generated single shot supercontinuum spectrum is illustrated on the right. Also highlighted is the process of soliton–dispersive wave trapping.

pumped at 1060 nm from 320 to 2300 nm, effectively covering the complete window of transmission of silica fiber, with up to 2 mW nm⁻¹ in the ultraviolet.

Similar operation can also be obtained in conventional PCF structures by modifying the waveguide structure, such that core resembles a few micron silica strand surrounded by air, by producing a PCF with a large hole pitch and a high air fill fraction. Operation from below 400 to 2500 nm has been shown in single constant core fibers.

Continuous Wave Pumped Supercontinua

With advances in broad stripe semiconductor pump laser technology, double clad doped fiber manufacture and innovative procedures for multimode combination of pumps, single transverse mode operation of cw Yb-fiber lasers at the 10 kW level has been demonstrated, however, rather modest average powers are quite sufficient to surprisingly observe supercontinuum generation in relatively short lengths of both photonic crystal and conventionally structured silica-based fibers. The process proceeds through the mechanism of modulational instability and consequently, the pump radiation wavelength must be in the region of anomalous dispersion. In the temporal domain, the intensity profile of the nominally cw pump radiation exhibits a noise-like structure arising from mode beating of the large number of randomly phased modes lying under the lasing spectral profile of the pump. The shortest temporal feature is given by the inverse of the frequency bandwidth of the laser radiation and this noise can enhance or seed the modulational instability process as has been described above. Through this mechanism the amplitude or phase perturbations rapidly evolve into fundamental soliton structures of femtosecond duration. Once established, these further exhibit the soliton self-frequency shift, extending the long wavelength extent of the supercontinuum. As soliton dynamics play a pivotal role in the overall generation process, it can be envisaged that the bandwidth, or related temporal structure, of the pump source is an important consideration in the generation process. For example, with a narrow bandwidth, the related temporal structure will be long. This means that the associated soliton lengths can be excessive and so soliton characteristics will not be established in the short fiber lengths utilized for supercontinuum generation, while long duration solitons would also not exhibit efficient self-Raman interaction. Likewise, for broad spectra, the associated structures are short, consequently, the required powers for soliton generation are excessively high and in this case solitons would be difficult to establish in the fibers. Empirically it can be argued and it has been shown theoretically that an optimum bandwidth exists for cw pumped supercontinuum generation.

The overall generation process is very similar to that using long picosecond pulse pumping. When the pump wavelength is far from the zero dispersion wavelength the generated supercontinuum is dominated by soliton-Raman shaping and the soliton self-frequency shift, with the generated radiation solely lying at Stokes wavelengths from the pump. This can be seen in trace 2 of Fig. 7, which shows the continuum generated by a 170 W cw signal at 1060 nm in a PCF with a dispersion zero wavelength at 830 nm and pump depletion is also evident. A spectral power density of 100 mW nm⁻¹ was possible and the supercontinuum extended to above 2000 nm, but no short wavelength components were generated.

Similar to picosecond pulse pumping, when the generated short pulse soliton structures arising from the modulational instability process are in the region of the dispersion zero wavelength ingress to the normally dispersive region and the action

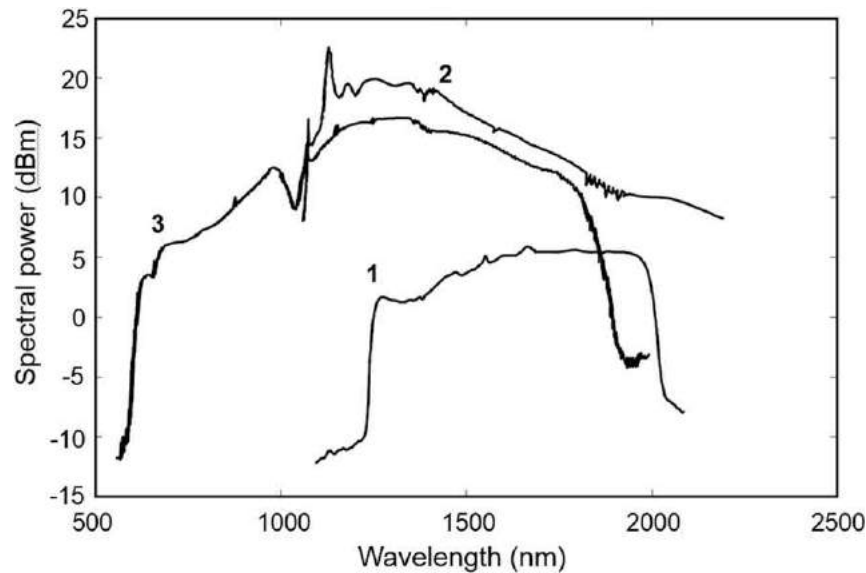


Fig. 7 Representative supercontinua generated under continuous wave (cw) pumping. (1) Conventionally structured highly nonlinear fiber pumped at 1.55 μm . (2) Photonic crystal fiber (PCF) with a dispersion zero at 830 nm pumped by 170 W at 1060 nm. (3) PCF with a dispersion zero at 1050 nm pumped by 230 W at 1060 nm.

of soliton–dispersive wave coupling leads to significant expansion of the supercontinua to shorter wavelengths. This process under cw pumping was first demonstrated in a 600 m length of conventionally structured highly nonlinear silica fiber with an effective core diameter of 2.1 μm and a zero dispersion wavelength of 1553 nm, which was excited by an Er-doped fiber-based ASE source at 1560 nm, amplified to 10 W average power. Plot 1 in Fig. 7 shows the generated supercontinuum, which is marked by its relative spectral flatness, which is the result of the integration of the noise-like spectra, depletion of the pump and spectral expansion to wavelengths shorter than the pump. By employing a similar technique only pumping a 50 m length of a PCF with a zero dispersion at 1050 nm pumped by 230 W at 1060 nm, extension to the visible region was achieved with cw pumped supercontinua, as represented by plot 3 in Fig. 7. Through optimizing fiber structures and employing tapered geometries it has been possible to achieve operation down to 470 nm and with considerably lower pump power requirement of only 40 W under cw pumping.

A characteristic of the CW pumped supercontinua is the smoothness of generated spectra and this again is a result of the integration of the highly noisy spectra. This can be seen in Fig. 8 which is a theoretical simulation of a spectrogram generated after 25 m propagation in a PCF with a zero dispersion at 1050 nm, pumped by 170 W at 1060 nm from an Yb fiber laser. In the figure, the one to one correspondence of the intense self-frequency shifted solitons to long wavelengths and the trapped dispersive waves to shorter wavelength can be clearly seen. The resultant spectrum shown to the right extends from 600 to 1800 nm, exhibits a relatively smooth structure while in the time domain, the output shown at the top of the figure exhibits a noise-like structure dominated by intense, randomly positioned solitons of varying duration and intensity. Power scaling has allowed spectral power densities in excess of 100 mW nm^{-1} to be obtained from cw pumped systems.

Wavelength Extension

The introduction of PCF and its integration with MOPFA assemblies underpinned the renaissance in both experimental and theoretical studies of supercontinuum generation, which subsequently led to the commercial development of rugged compact configurations that have found diverse application. To date, silica-based fibers have dominated the commercial product such that the spectral range from 320 to 2400 nm is covered, effectively the transmission window of silica. The average power capability has also been scaled such that in excess of 100 W has been achieved with both picosecond and continuous wave pumping. For many applications in spectroscopy, extension, particularly to the mid-infrared is desirable.

One possibility is to use fibers that are heavily doped with germanium. Beyond 2000 nm germanium oxide-based fibers exhibit a lower loss than silica, although it still is quite substantial, exceeding in the range of 100 dB km^{-1} , however, germanium oxide also exhibits a Raman gain coefficient that is nearly one order of magnitude greater than silica. As has been shown, the long wavelength extension in supercontinuum generation is a consequence of self-Raman interaction – the so called soliton self-frequency shift. As a result of the higher Raman gain coefficient, shorter gain lengths can be used, hence propagation losses are also reduced. Using pumping from a modestly powered sub-picosecond Tm-fiber laser-chirp pulse amplifier configuration, operation beyond 3000 nm has been achieved in these relatively robust Ge doped ($\sim 75\% \text{ GeO}_2$) fibers. In order to further extend sources to the interesting mid-infrared “molecular fingerprint” region, it is essential to make use of the soft glasses, such as the fluorides or the

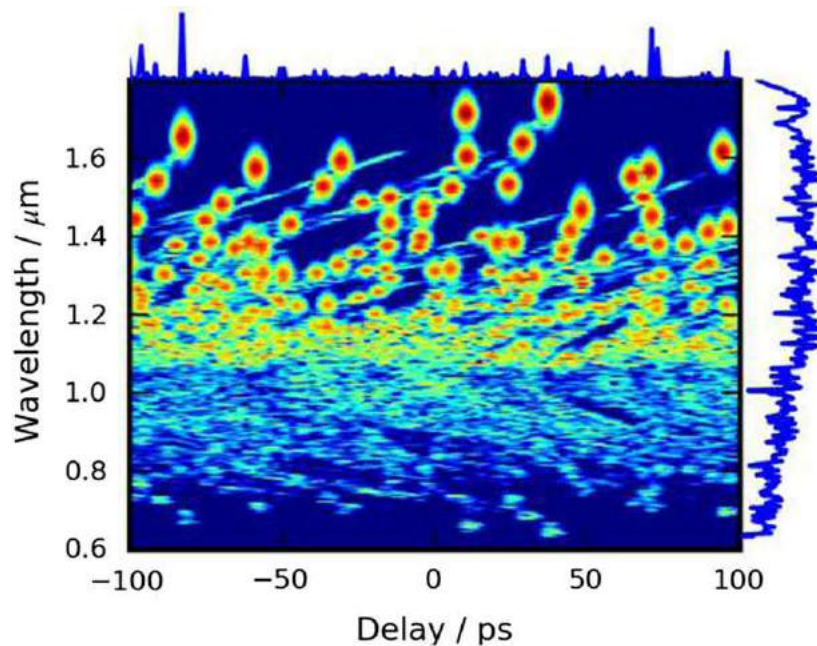


Fig. 8 Simulated spectrogram of a cw pumped supercontinuum generated after 25 m propagation in a photonic crystal fiber with a dispersion zero at 1050 nm pumped by 170 W at 1060 nm. A snapshot of the temporal output is shown to the top, while the associated supercontinuum is shown on the right.

chalcogenides, while to operate in the extreme ultraviolet, gas-filled, hollow-core fiber structures need to be utilized as the nonlinear medium.

The manufacture of single mode fluoride fibers of conventional structure is a well-established technology, with a material transmission window extending from 200 to 8000 nm, however, waveguide losses can significantly reduce the upper wavelength operational limit. In early work $\text{ZrF}_4\text{-BaF}_2\text{-LaF}_3\text{-AlF}_3\text{-NaF}$ (ZBLAN) fiber pumped by the 1.6 MW, 900 fs pulses from a mode-locked Er-fiber laser at 1550 nm allowed supercontinuum generation from around 1400 to 3400 nm. Average power scaling has been undertaken with in excess of 10 W average power obtained in a fluoride fiber-based supercontinuum extending from 800 to 4000 nm. Through pumping with peak powers of 50 MW, obtained from a femtosecond optical parametric amplifier system at 1450 nm and a 2 cm length of ZBLAN fiber, spectral coverage from 350 to 6280 nm has been achieved. In the past few years, fluoride-based PCF has been produced using the stack and draw technique rather than by extrusion which had previously been used for the soft glasses and this has allowed submicron structures to be utilized to enhance the power density, permitting spectral coverage in supercontinua extending from 400 to 2400 nm in 4.3 cm fiber samples when pumped at 1040 nm, with the potential for power dependent further spectral extension.

The soft glasses exhibit nonlinear coefficients that are many orders of magnitude greater than silica, consequently for similar peak power pump pulses substantially shorter lengths of fiber are required to establish continuum generation. Generally, the soft glass fibers have not been demonstrated to handle the high operational average powers that their silica-based analogues have, which generally favors pumping with femtosecond pulses and consequently short fiber lengths. Most commonly, these host fibers have been manufactured by extrusion techniques. An 8-mm-long tellurite fiber pumped by a femtosecond optical parametric oscillator at 1550 nm generated a continuum from 789 to 4870 nm, while a 4 cm SF6 PCF pumped by 60 fs pulses at 1060 nm produced a spectrum spanning from 600 to 1450 nm.

Chalcogenide fibers have successfully demonstrated their potential for operation in the 2–12 μm range. Conventionally structured As-S fiber pumped by 100 fs pulses at 2500 nm exhibited operation from 2000 to 3600 nm. In a 9 cm length of suspended core chalcogenide fiber pumped at 3500 nm, a supercontinuum extending from 2000 to 6000 nm was reported. Through pumping at 6.3 μm with the 100 fs pulses derived through difference frequency generation of pump and signal pulses derived from an optical parametric amplifier, albeit at an average power of 0.75 mW, a supercontinuum extending from 1.4 μm to 13.3 μm has been achieved in an 85-mm-long step index chalcogenide fiber. Research is on-going to simplify the pump and generation process, as well as to scale the average power levels. Chalcogenide rib waveguides have also been successfully employed, generating spectra covering the 2–10 μm range.

In the ultraviolet, silica PCF has permitted the generation of 280 nm radiation in a supercontinuum, however, practical application and long term operation is prevented by photo-damage of the glass. ZBLAN PCF, however, has allowed operation down to 200 nm without any reported short wavelength radiation induced damage problems. To generate supercontinua below this, the glass path has to be removed and gas used as the nonlinear medium. Most commonly PCF with Kagomé structure is used with hollow-core diameters in the 10–50 μm range. The associated waveguide dispersion of these structures is anomalous with a

relatively low dispersion slope. Since the dispersion of the gas fill at several atmospheres pressure is normal, the net dispersion can be tuned and the position of the zero dispersion wavelength selected through variation of the gas pressure. As a result, a small anomalous dispersion can be obtained in the visible/near infrared, and a zero dispersion in the near-UV. High peak power pumping can be achieved without optical damage problems and high order soliton dynamics observed under femtosecond pumping with amplified Ti:Sapphire laser systems. Over a 15 cm length of a 28 μm diameter Kagomé-PCF filled with hydrogen at 5 bar pressure, 2.5 μJ , 35 fs pulses at 800 nm generated a supercontinuum extending from 124 to 1200 nm. In He, at 28 bar pressure, operation down to 110 nm has been observed.

To date, however, only the compact visible/near infrared supercontinuum PCF-based sources, pumped primarily by Yb-MOPFA configurations have had widespread commercial success and have underpinned a broad, turn-key applications base.

See also: Nonlinear Optics

References

- Lin, C., Nguyen, V.T., French, W.G., 1978. Wideband near IR continuum (0.7–2.1 μm) generated in low loss optical fibers. *Electronics Letters* 14, 822–823.
 Lin, C., Stolen, R.H., 1976. New nanosecond continuum for excited-state spectroscopy. *Applied Physics Letters* 28, 216–218.

Further Reading

- Agrawal, G.P., 2001. *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
 Agrawal, G.P., 2008. *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
 Alfano, R.R. (Ed.), 2016. *The Supercontinuum Laser Source – The Ultimate White Light*, third ed. New York, NY: Springer.
 Alfano, R.R., Shapiro, S.L., 1970. Emission in the region 4000 Å to 7000 Å via four-photon coupling in glass. *Physical Review Letters* 24, 584–587.
 Dudley, J.M., Genty, G., Coen, S., 2006. Supercontinuum generation in photonic crystal fiber. *Reviews of Modern Physics* 78, 1135–1184.
 Dudley, J.M., Taylor, J.R. (Eds.), 2010. *Supercontinuum Generation in Optical Fibers*. Cambridge: Cambridge University Press.
 Genty, G., Coen, S., Dudley, J.M., 2007. Fiber supercontinuum sources. *Journal of the Optical Society of America B* 24, 1771–1785.
 Mollenauer, L.F., Gordon, J.P., 2006. *Solitons in Optical Fibers*. San Diego, CA: Academic Press.
 Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25, 25–27.
 Rulkov, A.B., Vyatkin, M.V., Popov, S.V., Taylor, J.R., Gapontsev, V.P., 2005. High brightness picosecond all-fiber generation in 525–1800 nm range with picosecond Yb pumping. *Optics Express* 13, 2377–2381.

Infrared Transition Metal Solid-State Lasers

Kenneth L Schepler, University of Central Florida, Orlando, FL, United States

© 2017 Elsevier Inc. All rights reserved.

Motivation

Broadband tunability of laser sources is of interest for a number of applications where broadband emission, wavelength selection, or wavelength scanning is needed. The uncertainty principle states that the product of the pulse width and bandwidth ($\Delta t \Delta \nu$) of a laser pulse has a minimum value. This requires that ultrashort pulses have broad bandwidth. Thus Ti^{3+} transition metal ions doped into a sapphire crystal have an emission bandwidth of 128 THz which could, in theory, be modelocked (ML) to produce a 3.4 fs pulse; 5 fs pulses have been demonstrated. Active remote sensing requires spectral scans over large frequency ranges and provides high signal to noise ratios with detection of multiple molecules of interest. Molecules with carbon–ligand bonds have vibration–rotation transition energies that correspond to wavelengths throughout the infrared spectral region. The atmosphere also has broad windows of infrared transmission (Fig. 1) as well as molecular constituents like water, carbon dioxide, carbon monoxide, etc. with optical transitions in the infrared region. Thus broadband infrared laser sources are needed for atmospheric remote sensing, molecular spectroscopy, medical sensing/detection of biomolecules, surgery, dentistry, gas leak detection, target designation/recognition, countermeasures to heat seeking missiles, and many others.

Early Transition Metal Lasers

The first demonstrated laser (1960), a ruby laser, employed a radiative transition between two electronic energy levels in Cr^{3+} , a transition metal ion doped into alumina (Al_2O_3). However, this transition metal laser had a fixed (not tunable) wavelength of 693.4 nm. There were also some early investigations of Ni^{2+} (1963) and Co^{2+} (1964) as lasing ions but operation required cryogenic cooling and tuning performance was poor due to excited state absorption (ESA). But the early transition metal lasers actually turned out to be exceptions to the rule that transition metal lasers have broadband tunability. This untunable nature of, particularly, the well-known ruby laser may have considerably delayed the development of broadly tunable transition metal lasers, such as alexandrite (1979) and Ti^{3+} -doped alumina (1982). So, where does the tunability of transition metal ions come from? After all, the rare earth lasers (e.g., Nd, Er, Tm, and Ho) have sharp energy transitions with quite limited laser tunability. The

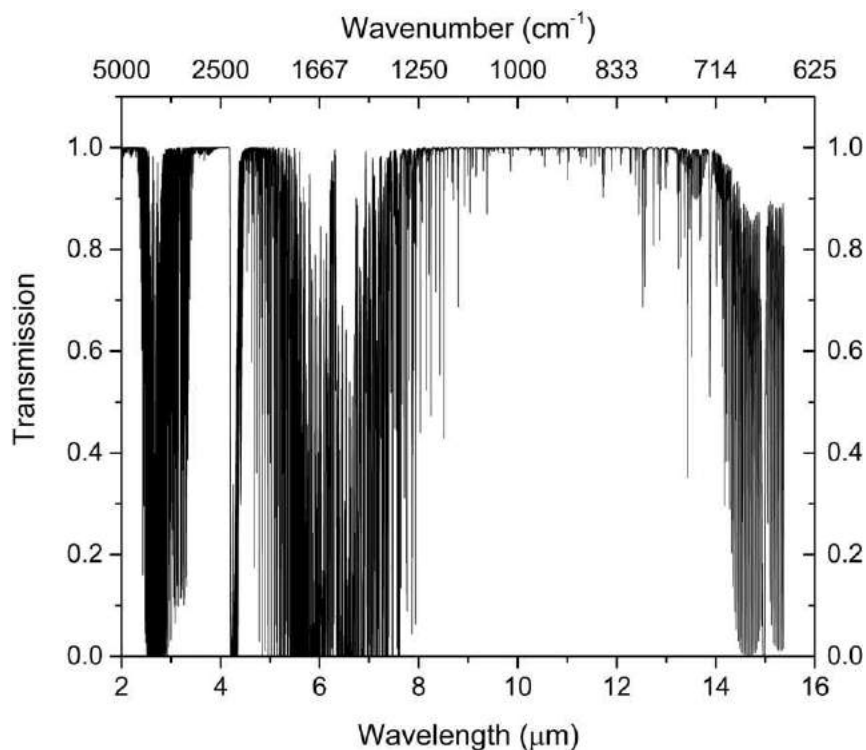


Fig. 1 Transmission through 10 m of mid-latitude summer atmosphere using HiTRAN on the Web.

difference lies in the nature of the energy levels involved. The partially filled electron orbitals in rare earths are f-shell orbitals, while the partially filled electron orbitals in transition metals are d-shell orbitals.

Transition Metal Spectroscopy

Crystal Field Theory

The IUPAC definition of a transition metal is an element whose atom has a partially filled d sub-shell, or which can give rise to cations with an incomplete d sub-shell. The d-shell electrons in a transition metal ion located in a crystalline solid are exposed to the influences of surrounding atoms and their electrons. Thus, for example, a Cr^{3+} ion substituting for an Al^{3+} ion in alumina (Al_2O_3), has six O^{2-} ions octahedrally arranged around it. These six negative charges form a crystal field which breaks the degeneracy of the d-electrons of an isolated chromium ion. Thus we now have the possibility of energy transitions (excitation, spontaneous emission, and stimulated emission) between d-shell orbitals. The symmetry of the surrounding ions and their proximities determine the strength of the crystal field. The crystal field and the nature of the transition metal ion determine the splitting of the d orbitals. However, the crystal field felt by f-shell electrons in rare earth ions is very weak. This is due to the fact that a filled s shell extends beyond the f shell and shields the f electrons from external fields. In general, for rare earths, Coulomb splitting of electronic energy levels is the strongest effect followed by spin-orbit splitting with crystal field being by far the weakest. Thus rare earth ions like Nd^{3+} have energy level splitting's that differ very little from host to host, for example, Nd:YAG (1064 nm), Nd:YVO₄ (1064 nm), Nd:YLF (1047, 1053 nm), and Nd:GSGG (1061 nm). Here, YAG (yttrium aluminum garnet [$\text{Y}_3\text{Al}_5\text{O}_{12}$]), yttrium vanadate [YVO_4], YLF (yttrium lithium fluoride [LiYF_4]), and GSGG (gadolinium scandium gallium garnet [$\text{Gd}_3\text{Sc}_2\text{Ga}_3\text{O}_{12}$]) are common crystalline hosts for rare earth ion dopants.

A simple example is given below to illustrate the concept of crystal field theory splitting of degenerate d-shell orbitals. Let us consider a single 3d electron in an octahedral field, for example, a Ti^{3+} ion substituted for Al^{3+} in Al_2O_3 . The degeneracy of the free ion is $(2s+1)(2l+1) = (2(1/2)+1)$ spin states $\times (2(2)+1)$ orbital states = 10. Let us now see how the orbital degeneracy is partially removed by the crystal field (Fig. 2).

The electrostatic potential from ligand ions approximated as point charges along the x direction is

$$V_x = \frac{-Ze}{4\pi\epsilon_0} \left[\frac{1}{\sqrt{r^2 + a^2 - 2ax}} + \frac{1}{\sqrt{r^2 + a^2 + 2ax}} \right]$$

where $r^2 = x^2 + y^2 + z^2$ is the position of the 3d electron and Ze is the charge of each ligand ion. We have similar terms for V_y and V_z . The total potential is $V = V_x + V_y + V_z$. After expanding in terms up to the 6th degree we have

$$V(x, y, z) = \frac{-Ze}{4\pi\epsilon_0} \left\{ \frac{6}{a} - \frac{35}{4a^5} \left[x^4 + y^4 + z^4 - \frac{3}{5}r^4 \right] - \frac{21}{2a^7} \left[x^6 + y^6 + z^6 + \frac{15}{4}(x^2y^4 + x^2z^4 + y^2x^4 + y^2z^4 + z^2x^4 + z^2y^4) - \frac{15}{14}r^6 \right] \right\}$$

The Hamiltonian describing the energy of this system in terms of spherical harmonics $Y_l^m(\theta, \phi)$, ($-l \leq m \leq l$) and neglecting terms greater than r^4 is

$$H_c^{O_h}(\mathbf{r}) = -eV = \frac{Ze^2}{4\pi\epsilon_0} \left\{ \frac{6}{a} + \frac{7r^4}{2a^5} \left[Y_0^4(\theta, \phi) + \sqrt{\frac{5}{14}}(Y_4^4(\theta, \phi) + Y_{-4}^4(\theta, \phi)) \right] \right\}$$

Let us define $D \equiv \frac{1}{4\pi\epsilon_0} \frac{35Ze^2}{4a^5}$ and $q \equiv \frac{2}{105} \langle r^4 \rangle$. Then using the properties of spherical harmonics the $\langle 3dm_l | H | 3dm_l \rangle$ matrix elements (where m_l is the orbital angular momentum quantum number) are: as shown in Fig. 3(a).

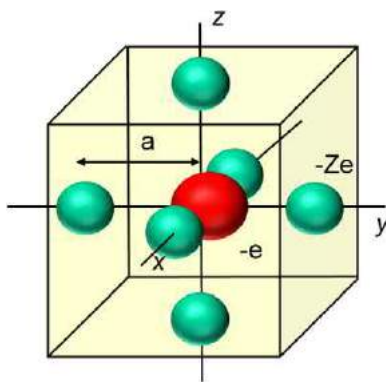


Fig. 2 Crystal field diagram for octahedral symmetry. Six negative ions (green) each with charge $-Ze$ are located on the faces of a cube each a distance a from the central transition metal ion (red) with a single d-shell electron.

Obviously, $|3d-1\rangle$, $|3d0\rangle$, and $|3d1\rangle$ are eigenstates since only diagonal matrix elements in Fig. 3(a) are nonzero. But the $m_l = \pm 2$ states are mixed. After diagonalizing the $m_l = \pm 2$ states we get degeneracy $g=3$ eigenstates (labeled T_2 in group theory terminology) with energy $-4Dq$ and $g=2$ eigenstates (labeled E) with energy $6Dq$. Thus as shown in Fig. 3(b) the octahedral crystal field has split the 10 degenerate electronic energy levels by $10Dq = \Delta_o$ with four levels in the upper level and six in the lower level. (Remember that the two electron spin states double the total degeneracy of each orbital degeneracy.) The calculated energy splitting makes sense physically when we look at the orbital orientations. (see Fig. 4). The T_2 orbitals have lower energy because they avoid the negative ions. The E orbitals have higher energy because they overlap neighboring negative ions.

At this point we will generalize the above discussion to state that a crystal field splits degenerate d orbital electronic energy levels into a set of multiple levels. Details for a specific transition metal ion depend upon how many d-level electrons are present and the strength and symmetry of the crystal field. For the case of Ti^{3+} there is only one electron in the 10d orbitals (d^1) and the crystal field energy levels are as just calculated. For other transition metal ions the number of electrons in d orbitals is different. For example, Cr^{2+} has four electrons in d orbitals (d^4) which is daunting to calculate until we realize that four of the five orbitals will be singly occupied (spin pairing is avoided) and we have one empty orbital. Thus the crystal field splitting is that of a single hole (a positive charge) and the T_2 and E sets of levels switch places but the size of the splitting for a given crystal field symmetry does not change. Details of crystal field splitting have been calculated for each of the d^q configurations ($q=1-9$) of the transition metal d orbital electrons and can be conveniently displayed in Tanabe–Sugano energy level diagrams. These useful diagrams can be found in Henderson and Bartram (2000).

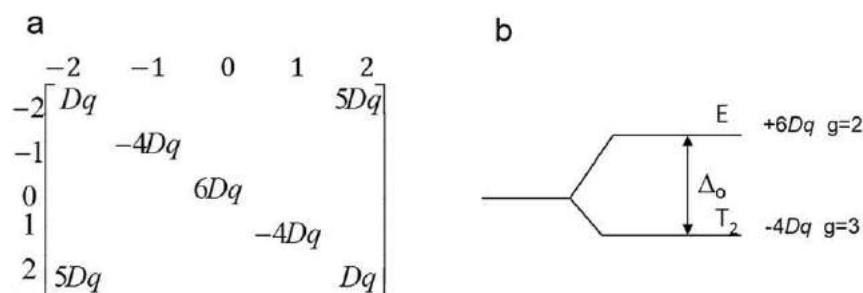


Fig. 3 (a) Crystal field Hamiltonian matrix elements for a single d-shell electron in octahedral symmetry with m_l eigenstates. (b) Crystal field splitting of the d-shell orbitals into a ground state triplet and an excited state doublet.

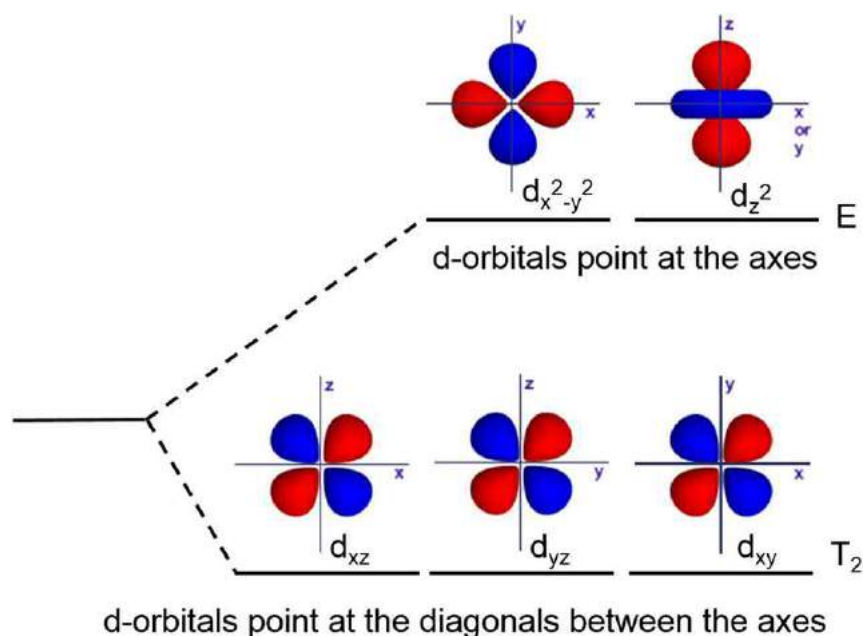


Fig. 4 Shapes of d orbitals and energy level splitting for octahedral crystal fields into two higher energy orbitals that point toward the surrounding negative charges and three lower energy orbitals that point diagonally between them.

Single Configuration Coordinate Model

But we have not yet explained the origin of broadband laser tunability of transition metal lasers. A crystal field not only splits the energies of the d orbitals but also couples the electronic motion to the motion (vibrations) of the surrounding crystal lattice ions. We can still assume that the electrons move much faster than the ions and that they adiabatically adjust to the varying ionic positions. This allows us to treat the full Hamiltonian of the electronic energies as a sum of the Hamiltonians for the electronic motions and the ionic vibrations. The ions can be modeled as masses connected to each other with springs, i.e., harmonic oscillators. Now instead of concentrating on the complex motion of many ions we can simplify things by considering the characteristic normal modes of vibration of the central transition metal ion and the nearest shell of surrounding ions. Each mode is a pattern of motion in which all parts of the system move sinusoidally with the same frequency and with a fixed phase relation. Each mode has a harmonic oscillator energy level structure of $\hbar\Omega(n + \frac{1}{2})$ where Ω is the vibrational frequency, n is the number of phonons (vibrational quanta), and $\hbar$ is Planck's constant. In any real crystal system, many normal modes of vibration must be taken into account. However, we will make the simplifying assumption that we can confine our attention to one representative mode only. It is conceptually useful to consider this the breathing mode where the surrounding shell of negative ions are moving in phase toward and away from the stationary central ion with the separation distance labeled as Q . In the single configuration coordinate (SCC) model, Q oscillates about its equilibrium value Q_0 . Each electronic energy level has its own set of phonon sublevels spaced by energy $\hbar\Omega$ and Q_0 values that are typically not equal due to different electron–lattice couplings; for example, consider how the two T_2 and E electronic levels have very different interactions with the surrounding electron clouds (Fig. 4).

A generalized example of the SCC energy curves is shown below (Fig. 5) for two electronic levels each of which has multiple vibrational sublevels. Each sublevel is an eigenstate of the system (an approximation that assumes electrons move much faster than the ions and that they adiabatically adjust to the varying ionic positions). The wavefunctions of the states can be represented as a product of an electron wavefunction ψ and a harmonic oscillator wavefunction χ . Probability of a photon absorption transition from electron–vibration state $|\psi_a\chi_n\rangle$ to electron–vibration state $|\psi_b\chi_m\rangle$ is proportional to:

$$|\langle\psi_b\chi_b(m)|\mu|\psi_a\chi_a(n)\rangle|^2$$

where μ is the appropriate electronic dipole operator. The harmonic oscillator states, χ , do not depend upon μ so we can separate the matrix element evaluation (Condon approximation) The transition probability is then:

$$W_{an-bm} = P_{ab} |\langle\chi_b(m)|\chi_a(n)\rangle|^2$$

where P_{ab} is the purely electronic transition probability and is the same for all n and m . The χ 's are the harmonic oscillator eigenfunctions, but defined with different equilibrium points for electronic levels a and b so the overlap integrals (matrix elements) are in general not zero for $m \neq n$.

Notice that if electronic levels a and b do have the same Q_0 equilibrium point, then the m th and n th χ eigenfunctions are orthogonal unless $m=n$ so the matrix element $|\langle\chi_b(m)|\chi_a(n)\rangle|^2 = \delta_{nm}$. We have uncoupled the electrons from the lattice vibrations. Rare earth ions act this way because the f electronic orbitals are shielded from the crystal field of the ligand ions.

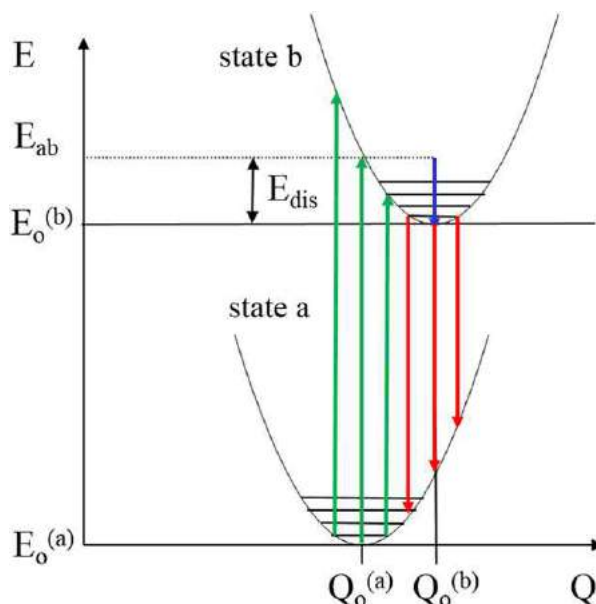


Fig. 5 Single configuration coordinate model vibronic energy levels. The green up-arrows are absorption transitions and the red down-arrows are emission transitions.

Classically we can think of the harmonic oscillator nuclei spending most of their time at the extrema of their oscillations so transitions take place primarily from one parabolic curve to another (Franck–Condon principle). Quantum mechanically one calculates the overlap integral of the product of the initial and final oscillator wavefunctions. These wavefunctions vary rapidly between the parabolic end points again resulting in transitions primarily happening at energies corresponding to going from one parabolic curve to the other. A closed form solution of the overlap integral is

$$|\langle \chi_a(m) | \chi_b(n) \rangle|^2 = \exp[-S/2] \sqrt{n!/m!} (\sqrt{S})^{m-n} L_n^{m-n}(S)$$

L_n^{m-n} is an associated Laguerre polynomial and S is the Huang–Rhys factor, a dimensionless constant characterizing the strength of the electron–lattice coupling.

$$S \equiv \frac{E_{\text{dis}}}{\hbar\Omega}$$

E_{dis} is defined in Fig. 5. Note that E_{dis} and thus S increase quadratically as the Q offset of the upper and lower parabolas increases.

At temperature $T=0\text{K}$ only the $n=0$ vibrational state is occupied so the matrix elements become

$$F_m(0) = |\langle \chi_b(m) | \langle \chi_a(0) \rangle|^2 = \frac{\exp[-S] S^m}{m!}$$

and the absorption bandshape is

$$I_{ab}(E) = I_o \sum_m \frac{\exp[-S] S^m}{m!} \delta(E_{bm} - E_{a0} - E)$$

Since $\sum_m |\langle \chi_b(m) | \langle \chi_a(0) \rangle|^2 = 1$ the intensity of the full band is I_o and is independent of S . This also means that the total intensity is independent of temperature. Intensity of the zero-phonon line ($m=0$, $n=0$, i.e., energy difference between the two bottom levels of the two parabolas in the SCC model diagram) is $I_o e^{-S}$. If $S=0$, then all of the intensity is contained in the zero-phonon line and there is no lateral displacement of the harmonic oscillator parabolae, i.e., $Q_o^{(a)} = Q_o^{(b)}$. As S increases, the intensity in the zero-phonon line decreases, but this is compensated by the appearance of vibrational sidebands observed at energies spaced by $\hbar\Omega$. The larger S is, the broader the absorption spectrum becomes; similarly emission is also broadened as S increases. These trends are shown in Fig. 6 for several values of S . S values are typically in the 5–10 range for transition metal ions.

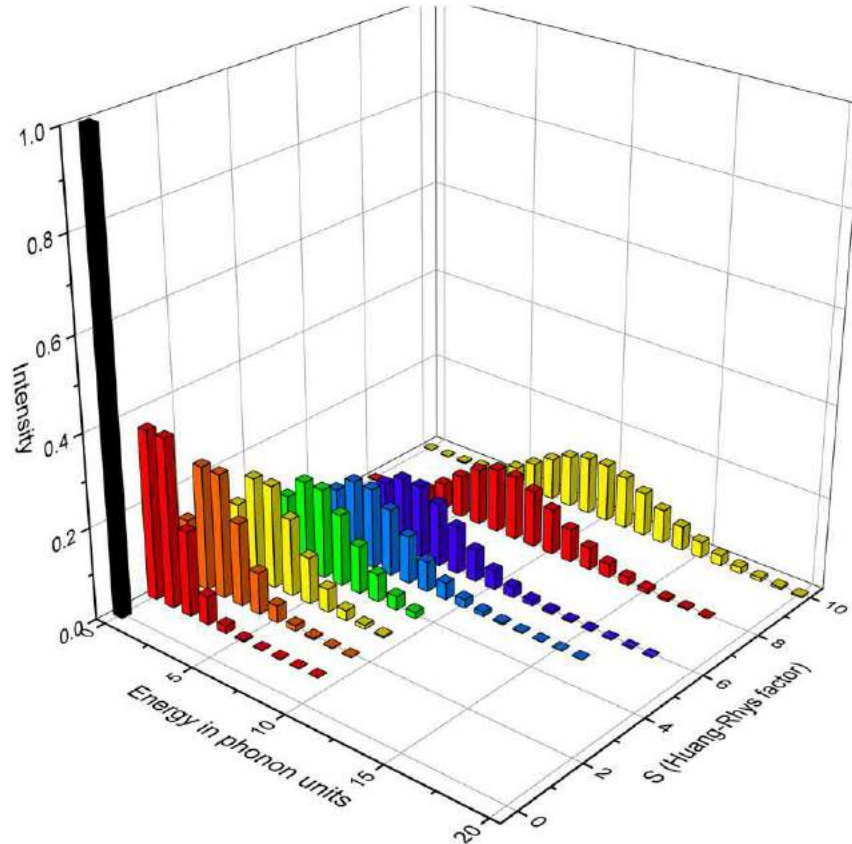


Fig. 6 Absorption line intensities for Huang–Rhys values in the $S=0$ –10 range. The energy of the absorbed photon is given in phonon units ($\hbar\Omega$).

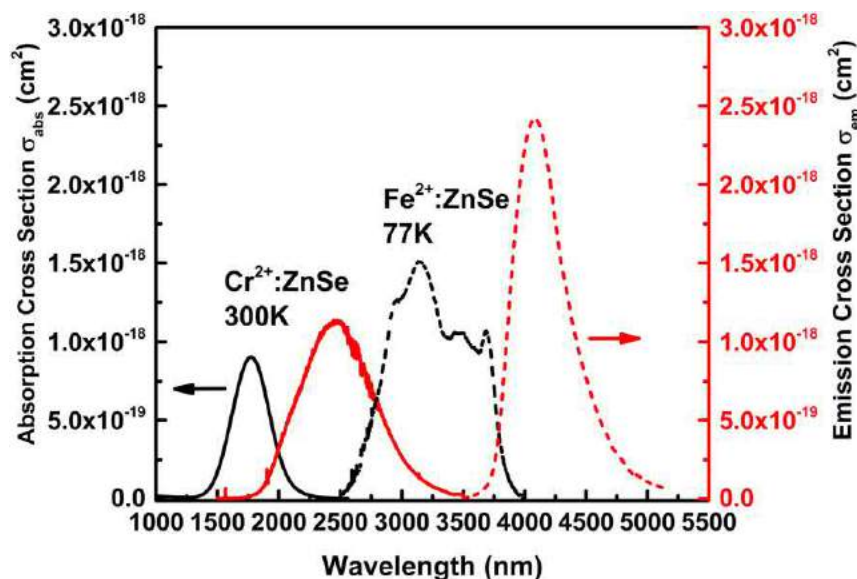


Fig. 7 Emission (red) and absorption (black) cross sections of Cr^{2+} (solid lines) and Fe^{2+} (dashed lines) doped into ZnSe.

The δ -function bandshape given above is not valid for a real system. A typical crystal lattice has many modes of vibration, not just one; this smoothes out the stick spectrum of Fig. 6 into a broad lineshape. There is one exception: all phonon modes have the same zero-phonon energy so the zero-phonon line remains a sharp Lorentzian line. However, increasing temperature further broadens the individual peaks into a single featureless peak hiding the zero-phonon line.

Note that we have finally come to the explanation of broadband tunable lasing of transition metal ions doped into solids. Vibronic (vibrational and electronic level) transitions combined with different vibrational center points for different electronic levels result in broad bands of absorption and emission. As an example Fig. 7 shows the measured absorption and emission spectra of Cr^{2+} and Fe^{2+} transition metal ions doped into ZnSe.

The Cr^{2+} Revolution

While ruby was technically the first transition metal laser, alexandrite ($\text{Cr}^{3+}:\text{BeAl}_2\text{O}_4$) was the first (1978) tunable transition metal laser. Shortly afterwards, $\text{Ti}^{3+}:\text{Al}_2\text{O}_3$ was discovered (1982) with tunability in the 700–1000 nm region and a few years later < 50 fs Kerr-lens modelocking was demonstrated (1990) leading to the ubiquitous use of this material for ultrashort pulse research and applications. In 1996 (DeLoach *et al.*, 1996) Cr^{2+} doped II–VI semiconductors were introduced as a new class of laser materials. (II Refers to 2^+ ions in the Zn column of the periodic table (Group 12) and VI refers to 2^- ions in the oxygen column (Group 16).) In the past 20 years Cr^{2+} and more recently Fe^{2+} lasers have become the Ti:sapphire lasers of the mid-IR (2–6 μm) spectral region. A large variety of II–VI semiconductor materials have been shown to work as transition metal hosts. These materials have a 2^+ Group 12 cation (such as Zn or Cd) surrounded by 4 Group 16 anions arranged at the corners of a tetrahedron. The tetrahedral crystal field, like the octahedral field discussed above, splits the d orbitals into the same E and T_2 sets of orbitals but for a single d electron the E levels are now lower than the T_2 levels. For tetrahedral symmetry the E orbitals avoid the surrounding anions, while the T_2 orbitals point toward the surrounding anions. However, Cr^{2+} is a d^4 configuration (four electrons in d orbitals acting as a single hole since five electrons would be a half-filled shell) so we must perform one more sign change bringing us back to the original case where the T_2 levels are lower than the E levels. Because there are fewer ligands (four instead of six) for the tetrahedral symmetry case, the crystal field splitting is also smaller, i.e., $\Delta_t = (4/9)\Delta_o$. This smaller crystal field splitting shifts optical absorption and emission transitions into the infrared region of the spectrum in contrast to the shorter wavelength transitions of $\text{Ti}^{3+}:\text{sapphire}$ which has octahedral crystal field symmetry. A number of TM:II–VI lasing ion – host combinations have been investigated including Cr^{2+} , Ni^{2+} , Co^{2+} , and Fe^{2+} as TM lasing ions and ZnSe, ZnS, ZnTe, CdSe, and alloys, such as $\text{Cd}_x\text{Mn}_{1-x}\text{Te}$ as host materials. ZnSe and ZnS have proven to be the best host materials and Cr^{2+} and Fe^{2+} the best transition metal lasing ions. Reasons for this down-selection are discussed in the Transition Metal Laser Advantages and Issues section.

Cr^{2+} Laser Advances

Table 1 shows examples of demonstrated Cr^{2+} laser performance. Modes of operation include continuous wave (CW), gain switched (GS), and ML. Q-switched operation is not practical because of the short E-level radiative lifetime of a few microseconds but GS pulses over 1 J have been demonstrated. Tuning over the 2–3.3 μm range and modelocking of ultrashort pulses as short as 26 fs have

Table 1 Cr²⁺ laser performance

Gain medium	Repetition rate	Energy	Wavelength/ range (μm)	Pulse width	Power (ave.) (W)	Temperature (K)	Regime	References
Cr:ZnS (PC)	80 MHz	90 nJ	2.3	26 fs	2	RT	ML	Vasilyev <i>et al.</i> (2016)
Cr:ZnSe (PC)			2.4		140	RT	CW	Moskalev <i>et al.</i> (2016)
Cr:ZnSe (PC)	0.7 kHz linewidth		2.3		5.5	RT	CW	Mirov <i>et al.</i> (2015b)
Cr:ZnSe (PC)	0.1 Hz	1.1 J	2.65	7 ms	0.110	RT	GS	Fedorov <i>et al.</i> (2012)
Cr:ZnS (PC)			1.96–3.20		0.01–0.70	RT	CW	Sorokin <i>et al.</i> (2010)
Cr:ZnSe (PC)			1.97–3.35		0.01–0.55	RT	CW	Sorokin <i>et al.</i> (2010)
Cr:ZnSe (WG)			2.08–2.78		0.01–0.12	RT	CW	MacDonald <i>et al.</i> (2015)
Cr:ZnSe (WG)			2.5		1.7	RT	CW	Berry <i>et al.</i> (2013)
Cr:CdSe (SC)	1 kHz	0.5 mJ	2.6/2.3–3.4	40 ns	0.50/35	RT	GS	McKay <i>et al.</i> (1999, 2002)

Abbreviations: CW, continuous wave; GS, gain switched; ML, modelocked; PC, polycrystalline; RT, room temperature; SC, single crystal; WG, waveguide.

Table 2 Fe²⁺ laser advances

Gain medium	Repetition rate (Hz)	Energy	Wavelength/ range (μm)	Pulse width	Power (ave.) (W)	Temperature (K)	Regime	References
Fe:ZnSe	Single pulse	0.3 mJ	3.95–5.05	60 ns		RT	GS	Fedorov <i>et al.</i> (2006)
Fe:ZnTe	Single pulse	0.18 mJ	4.35–5.45	40 ns		RT	GS	Frolov <i>et al.</i> (2011)
Fe:ZnSe			4.1		1.6	77	CW	Mirov <i>et al.</i> (2015,b)
Fe:ZnSe	850,000	600 nJ	4.04	64 ns	0.515	77	QS	Evans <i>et al.</i> (2014)
Fe:ZnSe	100	350 mJ	4.15	150 μs	35	77	GS	Mirov <i>et al.</i> (2015a)
Fe:ZnSe	Single pulse	42 mJ	4.5	750 μs		RT	GS	Frolov <i>et al.</i> (2013)
Fe:ZnS	Single pulse	3.2 J	3.6/3.4–4.2	750 μs		85	GS	Frolov <i>et al.</i> (2015)
Fe:ZnSe (WG)			4.1		0.049	77	CW	Lancaster <i>et al.</i> (2015)

Abbreviations: CW, continuous wave; GS, gain switched; QS, Q-switched; RT, room temperature; WG, waveguide.

been demonstrated. The current CW power record is 140 W. Cr²⁺ lasers are typically optically pumped in the 1.6–2.0 μm absorption band with Er³⁺ or Tm²⁺ fiber lasers although Raman-shifted emission from 1 μm Nd lasers can also be used. Energy transfer from the ZnSe conduction band to the Cr²⁺ ⁵E excited level has been demonstrated but transfer efficiency is poor.

Fe²⁺ Laser Advances

Fe²⁺ ions can also be doped into II–VI semiconductors. Again the crystal symmetry at the cation substitution site is tetrahedral. However, Fe²⁺ has a d⁶ configuration, i.e., a half-filled d shell plus one orbital occupied by a single electron of opposite spin. Thus we can treat Fe²⁺ as having a single electron in a tetrahedral field, resulting in a Δ_t splitting with the E levels lower in energy than the T₂ levels. However, the actual situation is more complicated. We have neither considered spin–orbit coupling up to this point nor have we considered the Jahn–Teller effect (degeneracy of a ground state will be removed by crystal field distortion). These effects are not required (because they are small) to describe the observed spectra of Ti³⁺ and Cr²⁺ ions. However, in Fe²⁺ ions the relative strengths of crystal field and spin–orbit effects are not well known and details of observed spectra are not clearly explained. Still the general concept of an SCC model is quite adequate to explain laser operation of this ion. Fe²⁺ demonstrated laser performance is shown in Table 2.

Fe²⁺ lasers have two important differences from Cr²⁺ lasers. First the energy splitting between the upper and lower levels is less for Fe²⁺ ions so the absorption peak is shifted from 1.8 to 3.2 μm and the emission peak is shifted from 2.4 to 4.1 μm (3.7–5.0 μm tuning). Optical pumping of Fe²⁺ ions is typically accomplished using Er³⁺ lasers operating at 3 μm. Cr²⁺ lasers tuned to the 3 μm region can also be used. Secondly, the lifetime of the Fe²⁺ upper lasing level is much more affected by temperature than the lifetime of the Cr²⁺ upper lasing level. The result is that Fe²⁺ room temperature (RT) CW lasing operation is not practical. However, it is possible to achieve GS operation at RT but at low efficiency since the lifetime has been reduced from 50 μs at low temperatures to 360 ns at RT.

Transition Metal Laser Advantages and Issues

The vibronic nature of absorption and emission energy transitions allow for broadband absorption and emission spectra as already shown in Fig. 7. The broadband emission provides the broadband tuning and ultrashort pulses characteristic of most

transition metal lasers. However, transition metal lasers do have some important disadvantages that must be mitigated. Ti^{3+} is an example. Ti^{3+} doped into alumina (Al_2O_3) works well as a RT, broadly tunable laser material. However, Ti^{3+} doped into YAlO_3 has not shown laser operation even though it exhibits bright yellow broadband fluorescence. The reason is thought to be ESA from the ${}^2\text{E}$ metastable level to the conduction band of YAlO_3 (band gap 7.1 eV). Stimulated emission gain cannot overcome ESA loss. Ti^{3+} ESA does not take place in alumina because the crystal field splitting is smaller and the alumina band gap (8.8 eV) is larger. Ti^{3+} ions have also been doped into lower crystal field hosts, such as YAG and Gd-Sc-Al garnet (GSAG). The lower crystal field results in significant nonradiative relaxation of the excited ${}^2\text{E}$ level back to the ${}^2\text{T}_2$ ground state. Thus efficient lasing at RT for these materials is not likely other than gain-switch operation by sub-microsecond pump pulses.

ESA: Co^{2+} Yes, Cr^{2+} No

Co^{2+} [d^7] and Cr^{2+} [d^4] are illustrative examples of transition metal ions that do and do not exhibit ESA. ESA directly competes with lasing gain so an ESA cross section similar in magnitude to the emission cross section is the death knell for efficient lasing.

Tanabe–Sugano diagrams are useful for describing the spectroscopy of transition metal ions. The horizontal-axis of a Tanabe–Sugano diagram is expressed in terms of the splitting parameter $10Dq$, or Δ , divided by the Racah parameter B . The vertical-axis is in terms of energy, E , also scaled by B . The Tanabe–Sugano diagrams for [d^7] (Fig. 8) and [d^4] (Fig. 9) energy levels in tetrahedral symmetry are shown below. Co^{2+} is a [d^7] transition metal ion and has a ${}^4\text{T}_2$ first excited state with broadband emission suitable for mid-IR lasing. However, the ${}^4\text{T}_1$ next higher level is at a location which approximately matches the emission energy so absorption from the ${}^4\text{T}_2$ level to the ${}^4\text{T}_1$ could be a problem. Details would depend upon how spin–orbit coupling and Jahn–Teller effects would split these levels but the experimental fact is that no Co^{2+} :II–VI laser with tetrahedral site symmetry has yet been demonstrated.

The first excited state for Cr^{2+} as shown in Fig. 9 is a ${}^5\text{E}$ state (spin multiplicity of 5). There are numerous higher lying levels so initially one might think that ESA could be a problem for Cr^{2+} in tetrahedral symmetry as well. However, all of the upper lying levels have different spin multiplicities (3 and 1) and since transitions between different total spins are forbidden, ESA should not occur and it has not been observed experimentally. The result is a robust lasing ion.

Nonradiative Relaxation

A good laser material should only be able to relax from the excited state back to the ground state through radiative emission (stimulated or spontaneous). Any nonradiative relaxation mechanism competes with the stimulated emission needed for lasing. In the case of transition metal ions, level crossing between upper and lower electronic level parabolas can occur due to offset of the different Q_0 equilibrium points. The diagram in (Fig. 10) shows a hypothetical case where this occurs. At low temperature only the lowest vibronic level of the upper electronic level is occupied after optical excitation. But at higher temperatures when $kT \approx \Delta E$, higher vibronic levels will be occupied. The probability of crossover from the upper electronic level to the lower electronic level

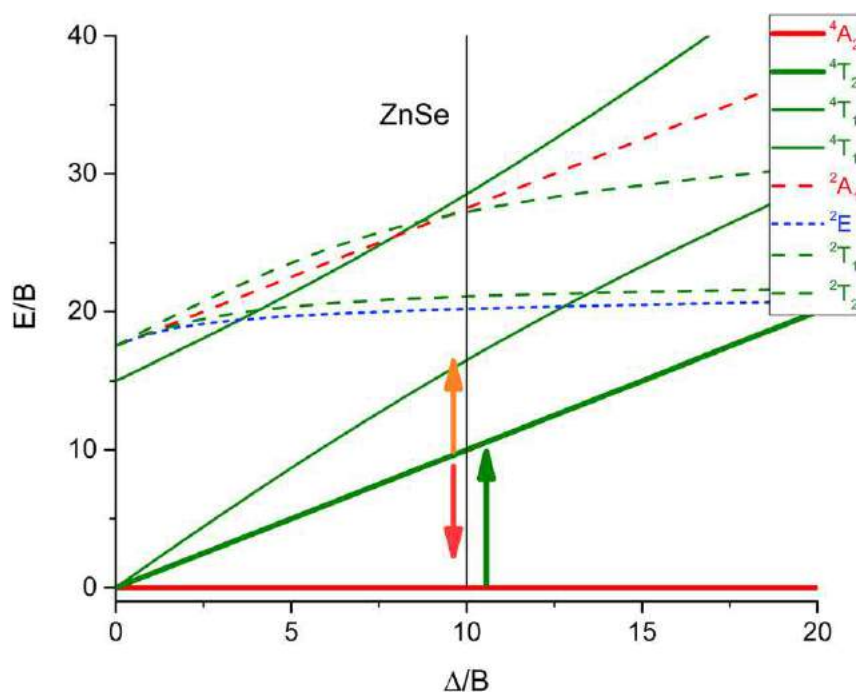


Fig. 8 Tanabe–Sugano diagram for Co^{2+} in tetrahedral symmetry. Solid-line curves have same spin as the ground state; dashed curves do not.

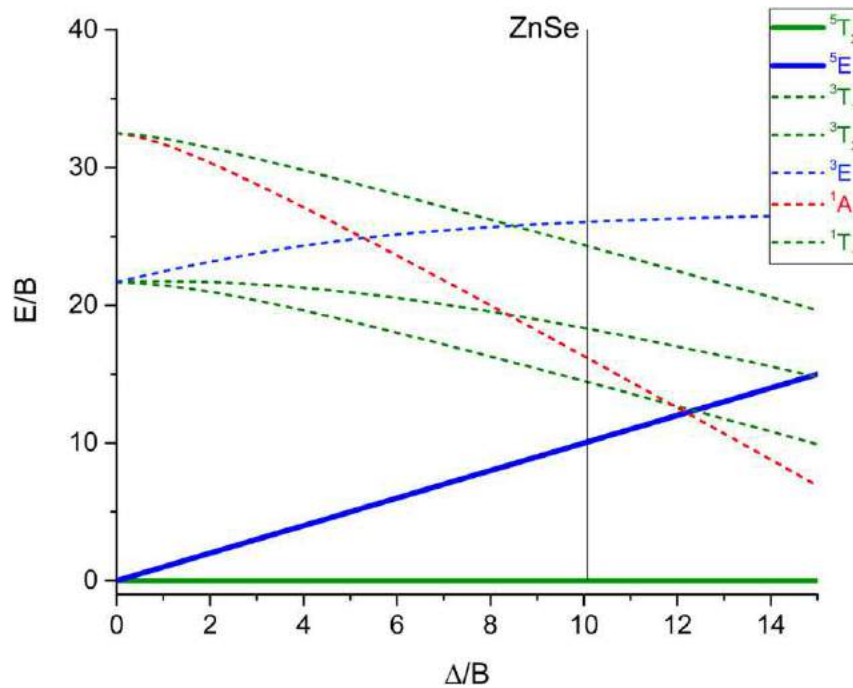


Fig. 9 Tanabe–Sugano diagram for Cr^{2+} in tetrahedral symmetry. Solid-line curves have same spin as the ground state; dashed curves do not.

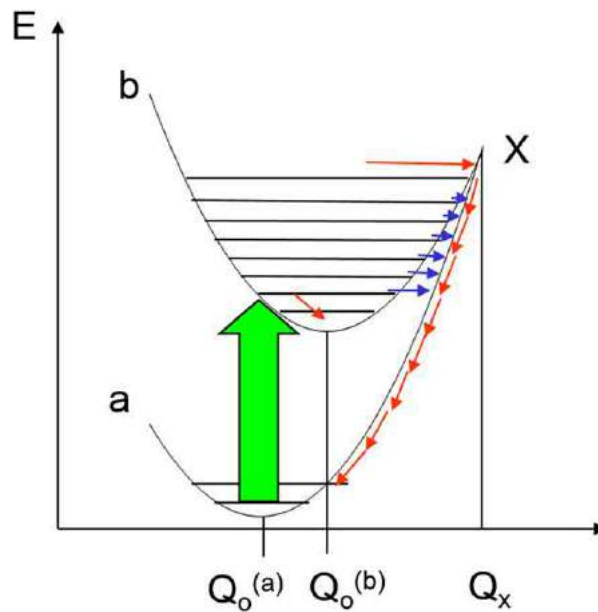


Fig. 10 Single configuration coordinate diagram of nonradiative relaxation. X is the classical energy crossover point. The blue lines represent tunneling from the b electronic level to the a electronic level. Phonon emission relaxes the a state back to thermal equilibrium.

without emission of a photon increases with temperature. Once the crossover has occurred, the vibronic excitation can relax quickly by emitting phonons, essentially sliding down the lower electronic parabola. Thus as temperature increases nonradiative relaxation becomes stronger to the point that a useful population inversion and lasing are not possible. The thermal relaxation path can be described by a rate constant N and an activation energy ΔE .

$$W_{\text{nr}} = N \exp \left[-\frac{\Delta E}{kT} \right]$$

Usually, the calculated activation energy based upon our SCC model is too high compared to the measured effect. The reason is that we have not allowed for quantum mechanical tunneling. Because the wavefunctions extend beyond the edge of the parabola there is a finite probability that tunneling from one parabola to another can take place. This makes the classical activation energy ΔE too high. A more correct way is to calculate the transition probability, which involves the overlap integral for the upper and lower vibrational wavefunctions and add the results for all overlapping pairs. This calculation has been done and yields a nonradiative relaxation rate $W_{\text{tunneling}} = N$ (the electronic part of the wavefunction overlap) $\times$ a vibrational exponential decay term that increases with temperature and has an S (Huang–Rhys) factor in the exponential like the radiative relaxation rate.

$$W_{\text{tunnel}} = N \exp[-S(2m+1)] \sum_{j=0}^{\infty} \frac{(S\langle m \rangle)^j (S\langle m+1 \rangle)^{p+j}}{j!(p+j)!}$$

N is typically $\sim 10^{13} \text{ s}^{-1}$. p is the number of phonons it takes to equal the zero-phonon energy gap between the electronic energy levels involved and $\langle m \rangle$ is the mean thermal occupancy.

$$\langle m \rangle = \left[\exp\left(\frac{\hbar\omega}{kT}\right) - 1 \right]^{-1}$$

This equation agrees quite well with the magnitude and temperature dependence of transition metal ion lifetimes. Experimentally one measures the excited state lifetime $\tau = 1/W$ and total relaxation rate $W = W_{\text{rad}} + W_{\text{nonrad}}$ or $1/\tau = 1/\tau_{\text{rad}} + 1/\tau_{\text{nonrad}}$. At low temperatures the relaxation is primarily radiative so we can determine the radiative lifetime by cooling our sample to the point where lifetime no longer depends upon temperature. As temperature increases we will reach a point where nonradiative relaxation becomes dominant and quenches emission. As emission becomes more quenched, lasing threshold increases and we reach a point where laser operation is no longer practical.

The details of transition metal ion excited level lifetimes versus temperature depend on both the ion and the host material. For example, Ti^{3+} has only a small amount of nonradiative relaxation at RT in sapphire ($W_{\text{nonrad}} = W_{\text{rad}}$ at 350K) but in the lower crystal field GdScGa-garnet, the nonradiative component is completely dominant at RT and $W_{\text{nonrad}} = W_{\text{rad}}$ already at 125K. The Cr^{2+} $W_{\text{nonrad}} = W_{\text{rad}}$ temperature is 350K in ZnSe but 300K in CdSe. Fe^{2+} is interesting as a transition metal laser ion because of its operation in the 3.5–5 μm regions. However, it has a $W_{\text{nonrad}} = W_{\text{rad}}$ temperature of 175K which requires cryogenic cooling for CW laser operation.

Nonradiative relaxation also becomes an issue for laser power scaling. As the laser power generated increases so does the heat load in the laser material. Active cooling is used to mitigate heating of the laser material but temperatures inside the laser crystal will increase as pump power increases regardless of how “good” the heat sink is. Rising temperature increases the nonradiative relaxation rate making lasing even less efficient, in turn leading to even more heating. This can result in a runaway situation where lasing will suddenly stop altogether.

Thermal Lensing

In addition to thermal quenching of emission, laser materials can, in general, experience thermal lensing. The refractive index of a material changes with temperature. The thermo-optic coefficient (dn/dT) of YAG at 1064 nm is $+9.1 \times 10^{-6} \text{ K}^{-1}$. But dn/dT is much higher for II–VI semiconductors, for example, ZnSe $dn/dT = 62 \times 10^{-6} \text{ K}^{-1}$ at 3.39 μm , 7x higher. Extracting heat from the sides of the laser material sets up a thermal gradient, which in turn establishes an index gradient, i.e., a lens in the material. The lensing power gets stronger as the pumping gets stronger. Once the thermal lens becomes strong enough to destabilize the laser resonator, lasing power will fail to increase with increasing pump power or completely stop. Techniques to mitigate this effect, such as thin disk cooling have been investigated. In thin disk cooling a thin slab of laser material is mounted directly to a heat sink. Heat then flows parallel to the resonating laser beam and the radial index gradient should be much less. The technique works to some degree. However, this requires a complex multi-pass pump beam setup to achieve efficient pump absorption in such a thin material. Also, intense pumping of the small volume of the disk leads to a temperature rise in the disk resulting in runaway thermal quenching of the emission.

Concentration Effects

Increasing the Cr^{2+} doping concentration to reduce the number of pump passes required for a thin disk to efficiently absorb the pump power is not an option due to concentration quenching of the emission. Doping higher than a particular level causes rapid nonradiative relaxation.

Many laser materials have an optimum lasing ion concentration. This effect is seen in Nd:YAG at concentrations of a few percent, a level where random substitution of Nd ions results in a significant fraction of them being at nearest neighbor locations. Neighboring ions in a crystal lattice can, in many cases, quench emission. However, in Cr-doped II–VI materials concentration quenching of emission is seen at much lower levels; the fluorescence lifetime is reduced by 50% at the 0.1% level in ZnSe. Interestingly, Fe^{2+} ions do not display strong quenching at such low concentration levels (Myoung *et al.*, 2012). The reason for Cr concentration quenching at low levels is unknown but may be a tendency of Cr ions to preferentially diffuse into a crystal at nearest neighbor lattice points.

Mitigating Thermal Effects

A spinning disk has recently been used to scale Cr:ZnSe laser power from ~ 20 to 140 W CW (Moskalev *et al.*, 2016). Rotation of the disk effectively increases the heated volume thereby reducing the temperature and the thermal gradient in the gain region. By the time the material has rotated back into the resonator beam, the heat has diffused away and convection has cooled the material back to its starting temperature. This allows considerable scaling of laser power but with the complication of using moving parts that must be very flat over the entire excitation ring to prevent beam pointing wobble.

Another technique that has been used to mitigate thermal effects is lasing in a fiber or waveguide (WG) configuration. Waveguiding can be designed to be resistant to thermal gradients in the material. Also, because the gain can be distributed over a long distance the temperature rise in the material can be reduced. Finally, heat sinking can be brought quite close to the heated region along the entire length of the WG fiber including total immersion in a flowing coolant. The technical issue here is fabrication of WGs or fibers with low loss. At this time WGs have been fabricated in Cr:ZnSe and Fe:ZnSe using laser inscription, a technique that uses focused ultrashort laser pulses to write refractive index changes into the laser material in a pattern that provides guiding. Another technique uses high pressure chemical deposition of Cr:ZnSe on the walls of a hollow silica fiber core to form a high index, laser active guiding region. Lasers have been demonstrated using both of these techniques but powers thus far have been limited to a few watts. Passive losses need to be reduced to achieve high-power operation. Current devices using both techniques have losses of ~ 1 dB/cm. Waveguiding devices may actually be of more practical use where flexibility of design and compact robust packaging are needed.

Summary

Transition metals have played a key role in the development of tunable solid-state lasers. The broadband emission spectra of transition metals enabled by coupling to lattice vibrations provides both broadband tunability and ultrashort pulse generation. Transition metal based laser devices operating in the visible, near-IR, and more recently the mid-IR are now commercially available. The emphasis in the future is likely to be the incorporation of these lasers into infrared applications where broad tunability, high pulse energy, high average power or ultrashort pulses are needed.

References

- Berry, P.A., MacDonald, J.R., Beecher, S.J., *et al.*, 2013. Fabrication and power scaling of a 1.7 W Cr:ZnSe waveguide laser. *Optical Materials Express* 3, 1250–1258.
- DeLoach, L.D., Page, R.H., Wilke, G.D., Payne, S.A., Krupke, W.F., 1996. Transition metal-doped zinc chalcogenides: Spectroscopy and laser demonstration of a new class of gain media. *IEEE Journal of Quantum Electronics* 32, 885–895.
- Evans, J.W., Berry, P.A., Schepler, K.L., 2014. A passively Q-switched, CW-pumped Fe:ZnSe laser. *IEEE Journal of Quantum Electronics* 50, 204–209.
- Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2011. Laser radiation tunable within the range of 4.35–5.45 μm in a ZnTe crystal doped with Fe^{2+} ions. *Journal of Russian Laser Research* 32, 528–536.
- Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2013. Study of a 2-J pulsed Fe:ZnSe 4- μm laser. *Laser Physics Letters* 10, 125001.
- Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2015. 3J Pulsed Fe:ZnS laser tunable from 3.44 to 4.19 μm . *Laser Physics Letters* 12, 055001.
- Fedorov, V., Mirov, M.S., Mirov, S., *et al.*, 2012. Compact 1J mid-IR Cr:ZnSe laser. In: *Frontiers in Optics 2012/Laser Science XXVIII*, Paper FW6B.9.
- Fedorov, V.V., Mirov, S.B., Gallian, A., *et al.*, 2006. 3.77–5.05-mm Tunable solid-state lasers based on Fe^{2+} -doped ZnSe crystals operating at low and room temperatures. *IEEE Journal of Quantum Electronics* 42, 907–917.
- Henderson, B., Bartram, R.H., 2000. *Crystal Field Engineering of Solid State Laser Materials*. Cambridge: Cambridge University Press.
- Lancaster, A., Cook, G., McDaniel, S.A., *et al.*, 2015. Fe:ZnSe channel waveguide laser operating at 4122 nm. In: *CLEO, OSA Technical Digest*, San Jose, CA, Paper SM2F.5.
- MacDonald, J.R., Beecher, S.J., Lancaster, A., *et al.*, 2015. Ultrabroad mid-infrared tunable Cr:ZnSe channel waveguide laser. *IEEE Journal of Selected Topics in Quantum Electronics* 21, 375–379.
- McKay, J., Schepler, K.L., Catella, G.C., 1999. Efficient grating-tuned mid-infrared Cr^{2+} : CdSe laser. *Optics Letters* 24, 1575–1577.
- McKay, J.B., Roh, W.B., Schepler, K.L., 2002. Extended mid-IR tuning of a Cr^{2+} : CdSe laser. In: *Fermann, M.E., Marshall, L.R. (Eds.), Proceedings of Advanced Solid-State Lasers*, Paper WA7.
- Mirov, S., Fedorov, V., Martyshkin, D., *et al.*, 2015a. High average power Fe:ZnSe and Cr:ZnSe mid-IR solid state lasers. In: *Advanced Solid-State Lasers*, Optical Society of America, Berlin, Paper AW4A.1.
- Mirov, S.B., Fedorov, V.V., Martyshkin, D., *et al.*, 2015b. Progress in mid-IR lasers based on Cr and Fe-doped II–VI chalcogenides. *IEEE Journal of Selected Topics in Quantum Electronics* 21, 1601719.
- Moskalev, I., Mirov, S., Mirov, M., *et al.*, 2016. 140 W Cr:ZnSe laser system. *Optics Express* 24, 21090–21104.
- Myoung, N., Fedorov, V.V., Mirov, S.B., Wenger, L.E., 2012. Temperature and concentration quenching of mid-IR photoluminescence in iron doped ZnSe and ZnS laser crystals. *Journal of Luminescence* 132, 600–606.
- Sorokin, E., Sorokina, I.T., Mirov, M.S., *et al.*, 2010. Ultrabroad continuous-wave tuning of ceramic Cr:ZnSe and Cr:ZnS lasers. In: *Advanced Solid-State Photonics*, Optical Society of America, Paper AMC2.
- Vasilyev, S., Moskalev, I., Mirov, M., *et al.*, 2016. Recent breakthroughs in solid-state mid-IR laser technology. *Laser Technik Journal* 13, 24–27.

Ultraviolet (UV) lasers, lasers with emission wavelengths shorter than 400 nm, have industrial as well as scientific applications including biochemical sensing, material processing, photolithography, and metrology. Because of their short wavelengths, UV light allows small features to be observed or patterned. The state-of-art semiconductor manufacturing process employs ArF excimer lasers at 193.4 nm to pattern (and fabricate) transistors as small as 10 nm scale (as of 2016) on silicon wafers, for example.

The energy of the photon is inversely proportional to its wavelength, therefore UV photons has higher energies compared to those in visible region. This makes their interaction with materials stronger, often leading to optical damage relatively easily. Materials transparent to shorter wavelength are limited as larger bandgap is needed to transmit photons with higher energy. One of the aspects of UV lasers is that they are required to have relatively narrow spectral width from chromatic aberration's point of view; most glass materials, which are limited in choices for transparency in the UV region, have large wavelength dependence at short wavelengths, as it gets close to the absorption edge of the material.

In this article, different types of UV laser sources are reviewed, and the techniques to generate UV light from longer wavelengths are explained below.

There are several types of UV lasers:

1. Excimer lasers such as XeCl at 308 nm, KrF at 248 nm, and ArF at 193.4 nm, which are currently in industrial use. Other excimer lasers include F₂ at 157 nm and Ar₂ at 126 nm.
2. Gas lasers, such as He–Cd at 325 nm, Ar⁺ at 364 nm or 351 nm. They are usually relatively large and bulky, inefficient, often need water cooling, and require fairly frequent maintenance (replacement of tubes). He–Cd lasers at 325 nm may be used in the manufacturing process of volume Bragg grating (VBG) devices by interferometric exposure, due to its matching sensitivity of photo-thermal refractive (PTR) glasses to that wavelength. It should be noted that the coherence length of such lasers to be used for interferometric exposure should be fairly long (narrow spectral width). With recent development of new types of lasers such as harmonic generation, these lasers are beginning to be replaced. Harmonically converted Ar⁺ lasers at either 257 nm (second harmonic generation (SHG) of 514.5 nm) as well as 244 nm (SHG of 488 nm), by placing the nonlinear crystal inside the laser resonator, has been used in industry; semiconductor inspection as well as fiber Bragg grating fabrication.
3. Semiconductor lasers using AlGaIn below 400 nm. Diode lasers are attractive from application point of view, for their potential low cost, efficiency, and compactness. With the bandgap of AlN being around 6 eV ($\lambda \sim 200$ nm), AlGaIn laser diodes (LDs) have the potential of emitting in the deep ultraviolet (DUV) region, which usually mean wavelengths shorter than 300 nm. It was in the late 1990s when the first blue-violet GaN diode laser was realized, and with the commercial demand in the optical disk, the display, as well as the lighting applications, the development of GaN LDs and light emitting diodes (LEDs) had been intensive to date. The demand for UV diode lasers are emerging, but not as strong as those for consumer applications, besides it is more challenging to fabricate reliable devices. As of this writing in 2016, the shortest wavelength that is commercially available from diode laser is 370 nm from Nichia with 70 mW output power. The process development of the material growth as well as the fabrication toward UV LDs at shorter wavelength has been a challenge, and only a few demonstrations with shorter emission wavelengths were made to date. With the electrical current injection pumping, laser diode grown on sapphire substrate with output wavelength as short as 336 nm had been reported (Yoshida *et al.*, 2008). With optical pumping, which requires another laser at a shorter wavelength, AlGaIn lasers grown on bulk AlN substrate have been demonstrated at wavelength down to 237 nm under ArF laser pumping (Kneissl and Rass, 2015). For the detailed principles of semiconductor lasers in general.
4. Solid-state lasers, using transitions in cerium ions in 280–290 nm. Cerium-doped laser host crystals, such as LiSrAlF₆ (Ce:LiSAF) or LiCaAlF₆ (Ce:LiCAF), had been studied since mid-1990s (Marshall *et al.*, 1994; Dubinskii *et al.*, 1993). Cerium ions in these host crystals can be pumped with the forth harmonic of neodymium lasers near 266 nm, and emit around 290 nm, tunable between approximately 280 nm and 320 nm. Tunable emission in this spectral range can be useful in chemical and biological sensing. Emission cross section is as high as of the order of 10^{-18} cm², so the optical gain can be high. Its lifetime is rather short, of the order of tens of nanoseconds, so they are better suited for pulsed pumping. Nevertheless, chopped-CW operation was demonstrated with a Ce:LiCAF laser.
5. Nonlinear frequency conversion from visible and/or infrared lasers. With the exception of excimer lasers in industrial applications, harmonically converted infrared lasers are the most common UV lasers. Common infrared lasers at 1064 nm including Nd:YAG, Nd:YVO₄, or fiber lasers can be converted to 355 nm by their third harmonic generation (THG), 266 nm by the fourth harmonic generation (4HG), or 213 nm by the fifth harmonic generation (5HG). Nonlinear materials that are transparent for these wavelengths and allows for phasematching are fairly limited; LBO is transparent to 160 nm but phasematches for THG, but not direct SHG at 266 nm. It can, however, phasematches for 266 nm generation by the sum-frequency mixing (SFM) between the fundamental and THG. Beta-barium-borate, β -BaB₂O₄, BBO is transparent to 189 nm, phasematches for the direct SHG down to 205 nm. It phasematches for 213 nm generation via both SFM between 1064 and 266 nm and SFM between

Table 1 Some of the nonlinear optical crystals and their properties

	LBO	BBO	CLBO	KDP	KBBF
Transparency	160 nm	180 nm	189 nm	200 nm	155 nm
Phasematched second harmonic generation (SHG)	278 nm	205 nm	238 nm	259 nm	162 nm
d_{eff} for SHG at 266 nm	N/A	1.54 pm/V	0.78 pm/v	0.46 pm/v	0.36 pm/v

532 and 355 nm. Cesium–lithium borate, $\text{CsLiB}_6\text{O}_{10}$, CLBO is transparent to 180 nm, and phasematches for direct SHG down to 238 nm as well as SFM at 213 nm via 1064 nm and 266 nm but not 532 and 355 nm. Potassium dihydrogen phosphate, KH_2PO_4 , KDP is transparent to 200 nm, can phasematch down to 259 nm for direct SHG. Potassium difluoro-diberylo-borate, $\text{KBe}_2\text{BO}_3\text{F}_2$, KBBF is transparent to 155 nm, phasematches for direct SHG at 162 nm, but is difficult to grow and fabricate into polished devices and not commercially available as of 2016. Some of the properties of these materials are summarized in **Table 1**. They have effective nonlinear optical coefficients of just a few pm/V or less, so with realistic size of the crystal of a few centimeter or less, the normalized conversion efficiency is of the order of 10^{-4} W^{-1} , meaning 1 W at the harmonic is generated with the fundamental of 100 W. When one needs very large pulse energies, crystals of large aperture are required to avoid optical damage caused by the high peak power. For that purpose, KDP crystals are developed, and are grown in aqueous solution in the size of multiple tens of kilogram of single crystals to yield SHG/SFM devices with a clear aperture of more than 30 cm. These are being used in institutes including National Ignition Facility at Lawrence Livermore National Laboratory with the goal to generate the THG pulses 1.8 MJ of pulse energy and 500 TW of peak power.

The range of wavelengths that can be phasematched can be extended by way of SFM. Besides the harmonics of common 1064 nm lasers, the eighth harmonic of 1547 nm from erbium-doped fiber amplifier was generated by the SFM between the fundamental at 1547 nm and the seventh harmonic at 221 nm to generate 193.4 nm (Ohtsuki *et al.*, 2000), corresponding to the ArF wavelength, for the application of semiconductor inspection.

Since the efficiency of nonlinear frequency conversion is higher with higher input powers, pulsed lasers are well suited for efficient generation by taking advantage of high peak powers, opening opportunities for pulsed UV lasers for material processing. Via-hole drilling is one of the emerging application for THG sources in 340–360 nm, with a few tens of watts of average power with >10 kHz repetition rates, beginning to replace CO_2 lasers. Motivated by the industrial applications with more energy-efficient processes, harmonic conversion, such as THG at 355 nm, 40 W at 266 nm (Nishioka *et al.*, 2003) or 5HG at 213 nm are being developed and some noteworthy results were reported; 103 W at 355 nm (Rajesh *et al.*, 2008) using CBO, 27.9 W at 266 nm (Katsura *et al.*, 2007) using CLBO, and 10.2 W at 213 nm (Katsura *et al.*, 2007) at 10 kHz using CLBO.

As the single-pass harmonic conversion efficiency will not be high for continuous-wave (CW) lasers, the input power at the fundamental must be enhanced. Waveguide devices would maintain high intensity at the fundamental, but such devices for UV generation have not been hugely successful so far. Another technique to enhance the power at the fundamental to incident on the nonlinear crystal is the use of optical resonator. Like intracavity-doubled Ar^+ laser, a nonlinear crystal can be placed in a laser oscillator cavity of a visible laser, generating the UV output. Solid-state counterpart, recently mostly pumped with near-infrared semiconductor lasers, emits in the infrared, so there is not many visible solid-state lasers. For one, Pr:YLF lasers had been pumped with semiconductor lasers to oscillate at 522 nm, and a nonlinear crystal in the laser cavity had generated the DUV output at 261 nm (Ostroumov and Seelert, 2008).

- Harmonic conversion in external resonators. When a nonlinear crystal cannot be placed in the CW laser resonator, such as frequency doubled source or a laser diode, or a MOPA structure, the nonlinear conversion efficiency can still be enhanced by placing the nonlinear crystal in an external cavity which is in resonance with the incoming fundamental input. The first concept of external resonant frequency doubler was introduced by Ashkin *et al.* (1966). In this first investigation, the resonance of fundamental and that of harmonic are considered. In practice, fundamental-resonant doublers are more commonly used, as the enhancement in the efficiency can be square of the enhancement factor of the resonator, and in 1987, an efficient external resonant doubler was demonstrated to generate 532 nm using a monolithic $\text{MgO}:\text{LiNbO}_3$ external resonator with 13% conversion efficiency (Kozlovsky *et al.*, 1987).

Fig. 1 shows the conceptual schematic of such external resonator. The input should typically be single frequency, although there were demonstrations of externally resonant doubler with CW multi-longitudinal mode lasers or modelocked lasers. In this section, the principle of external doubling resonator is explained with the assumption of input fundamental being single-frequency. We further assume that the external resonator is a unidirectional ring as schematically depicted in **Fig. 1**.

Let us consider the electric field of the circulating power just inside the input coupler. Following Siegman's convention (Siegman, 1986), we write the phase of the reflection as in phase and the transmission from partial mirrors as shifted by 90 degrees. For the sake of simplicity, the magnitude of electric field is represented by the square root of the optical power. Noting that the SHG acts as a loss to the fundamental in the resonator,

$$it_i\sqrt{P_{\text{in}}} + r_i\sqrt{(1-\delta)(1-\gamma P_{\text{circ}})}e^{i\phi}\sqrt{P_{\text{circ}}} = \sqrt{P_{\text{circ}}}$$

where r_i and t_i are the field reflectivity and the field transmission of the input coupler, ϕ is the phase shift for one round-trip, δ is the passive loss of the resonator, which is the sum of all the losses such as residual transmission of the high reflector or

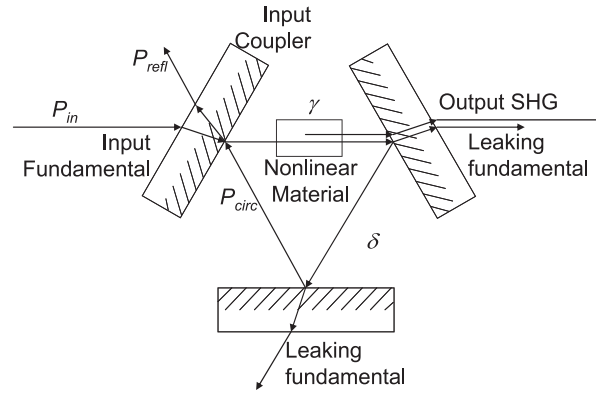


Fig. 1 Schematic of an optical resonator with a nonlinear crystal inside.

scattering/absorption of any optical components/surfaces, and g is the normalized nonlinear conversion efficiency which relates the generated power at the harmonic and the incident power at the fundamental by

$$P_{SH} = \gamma P_{circ}^2$$

Therefore,

$$\sqrt{P_{circ}} = \frac{it_i \sqrt{P_{in}}}{1 - r_i \sqrt{(1 - \delta)(1 - \gamma P_{circ})} e^{i\phi}}$$

For all practical purposes, the term $\delta \gamma P_{circ}$ in the square root in the denominator is much smaller than 1, and can be ignored. The circulating power at the fundamental in the resonator is written as, assuming the resonant condition ($\phi = 0$),

$$P_{circ} = \frac{(1 - R_i) P_{in}}{(1 - \sqrt{R_i} (1 - \delta - \gamma P_{circ}))^2}$$

where r_i^2 is rewritten as R_i as the power reflectivity. It is necessary to numerically solve the equation above to find the circulating power in the resonator. When we look at the reflection from the resonator at the fundamental, it can be written as

$$P_{refl} = P_{in} \left(\frac{\sqrt{R_i} - \sqrt{(1 - \delta - \gamma P_{circ})}}{1 - \sqrt{(1 - \delta - \gamma P_{circ})}} \right)^2$$

This indicates that the reflection becomes zero when $R_i = (1 - \delta - \gamma P_{circ})$, or $1 - R_i = \delta + \gamma P_{circ}$. It means the entire power incident onto the resonator is coupled into it when the transmission of the input coupler equals the total loss of the resonator. This condition is called impedance matching. When the transmission of the input coupler is higher/lower than the total loss of the resonator (including the depletion of the fundamental by the harmonic generation, which depends on the power at the fundamental), it is called “overcoupling/undercoupling.”

For example, 1 W of fundamental incident on the resonator with a loss of 0.5%, normalized conversion efficiency of $1 \times 10^{-4} \text{ W}^{-1}$, input coupler with 2% transmission would generate 0.569 W of harmonic with 75.5 W of fundamental circulating in the resonator. In this condition, only 53 mW is reflected from the resonator, meaning nearly 95% of the input power is coupled into the resonator. The optimum input coupling in this condition is found to be $R_i = 98.7\%$, therefore the input coupler $R_i = 98\%$ is slightly overcoupling. Single-pass interaction with the same fundamental power would yield 0.1 mW, therefore this resonator is providing more than 5000 times enhancement of the harmonic power. The enhancement factor, the ratio between the circulating fundamental inside the resonator and the incident power onto the resonator, is a good indication of the performance of the resonant doublers. Some examples are shown in Fig. 2 for different values of δ and R_i . Evidently the lower the resonator loss, the higher the enhancement factor, therefore the low-loss optical components are critical to the performance of the resonant doublers.

Examples of impedance matching are shown in Fig. 3. In this example, normalized efficiency of $1 \times 10^{-4} \text{ W}^{-1}$ and 1 W of input power is assumed. For different values of resonator loss, the impedance matching occurs at different input couplings.

The resulting output powers at the harmonic are shown in Fig. 4 for different resonator optical losses. The optimum output is obtained at the impedance-matched coupling. As was indicated in the previous example, hundreds of milliwatt at the harmonic can be obtained from 1 W of fundamental.

As one can see in Fig. 4 as well as 2, it is evident that small optical loss of the resonator is critical for the efficient conversion in the CW mode, in which single-pass conversion efficiency is not high. Such optical losses come from resonator mirrors, the bulk scattering or absorption of nonlinear crystal, its surfaces, or any components in the resonator. Proper quality control of these components and characterization is important for practical devices.

As shown, external resonant doublers are powerful tool to efficiently generate second harmonic in CW in which the efficiency would be very low in single-pass interactions. This approach had been used in the UV generation for industrial as well as scientific applications. Here are some of the examples: 80 mW of 257 nm output was demonstrated by the second harmonic of argon ion

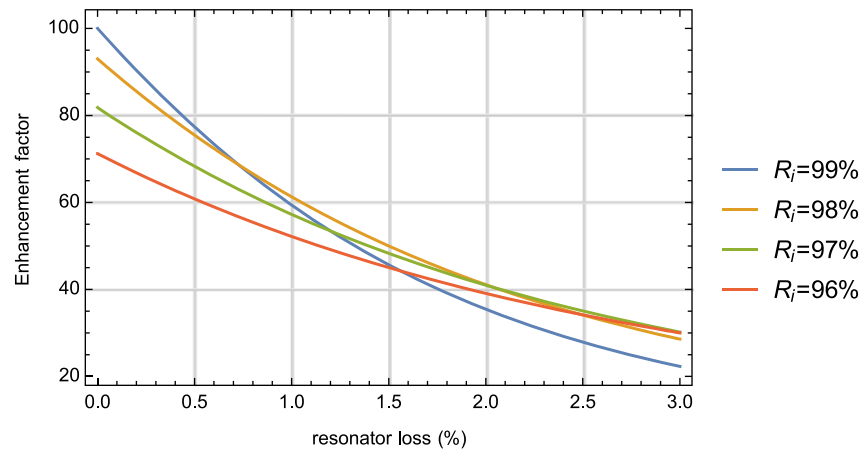


Fig. 2 Enhancement factors as the function of resonator loss.

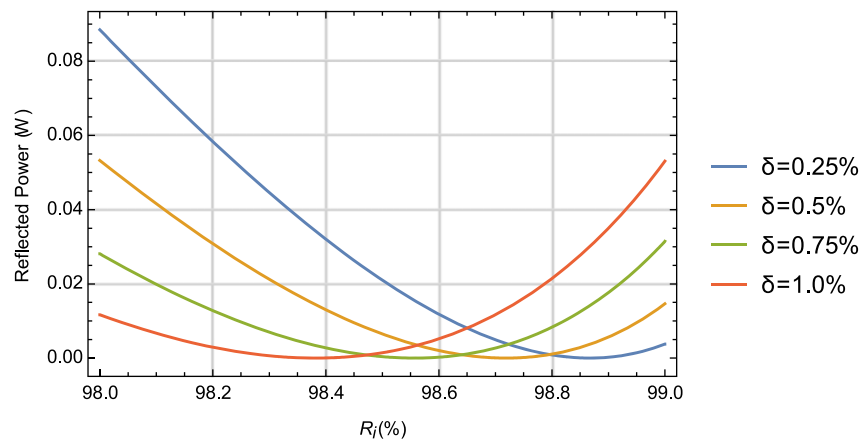


Fig. 3 Reflected power at the fundamental as the function of input coupler.

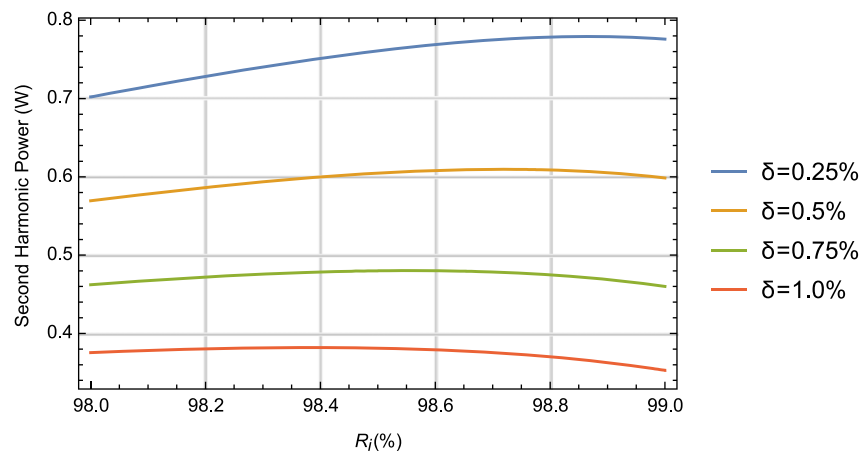


Fig. 4 Second harmonic output power as the function of input coupler.

laser in an external cavity for spectroscopy applications (Bergquist *et al.*, 1982). Fourth harmonic of 1064 nm near infrared laser for industrial applications was demonstrated with up to 12 W of CW output at 266 nm with 24 W of infrared power (Sudmeyer *et al.*, 2008), giving 50% efficiency from IR to DUV conversion efficiency. Owing to its continuous nature, and being single frequency giving very long coherence length, these lasers can be used for metrology or for material processing including interferometric lithography. UV output at tailored wavelengths were used for spectroscopic applications, such as optical lattice clock

(Katori *et al.*, 2003). Trapping/cooling laser for mercury atoms were demonstrated the fourth harmonic of an Yb:YAG disk laser with >750 mW of output power at 254 nm (Scheid *et al.*, 2007), or by the fourth harmonic of an optically-pumped semiconductor laser (OPSL) with more than 100 mW at the same wavelength (Paul *et al.*, 2011) were demonstrated. Trapping/cooling laser for cadmium atoms at 229 nm was demonstrated by the fourth harmonic of an OPSL with 0.56 W of output, which was used to resolve all the stable isotopes of cadmium atoms for the first time (Kaneda *et al.*, 2016).

In this section, we went over only harmonic conversion, but SFM technique is also available to extend the phasematching limit of a material toward the shorter wavelength. One of the examples is the SFM between 234 and 1110 nm to generate 193.4 nm in CW (Sakuma *et al.*, 2015), which is compatible with ArF lasers, using CLBO crystals, which phasematches for 238 nm or longer for direct SHG.

Conclusion

In this article, different types of UV lasers were reviewed, and practical aspects of nonlinear frequency conversion into the UV region were explained. A few borate crystals that are capable of harmonic conversion into the UV were raised and reviewed. Application of external resonator for harmonic conversion is reviewed and its principle is explained. As described, UV lasers carry its usefulness in applications including scientific as well as manufacturing, especially semiconductors with small feature sizes. With the demand driven by Moore's law, development of UV lasers seems to continue toward shorter wavelengths and higher powers. Laser sources near 200 nm, some in sub-200 nm region, are actively developed and their powers are of the order of watts for the applications in semiconductor inspection, >tens of watts for the exposure tools. Material processing also seems to be moving toward shorter wavelengths, and multiple watts to tens of watts of UV lasers are developed and deployed in the field.

References

- Ashkin, A., Boyd, G.D., Dziedzic, J.M., 1966. Resonant optical second harmonic generation and mixing. *IEEE Journal of Quantum Electronics* 2 (6), 109–124.
- Bergquist, J.C., Hemmati, H., Itano, W.M., 1982. High power second harmonic generation of 257 nm radiation in an external ring cavity. *Optics Communications* 43 (6), 437–442.
- Dubinskii, M.A., Semashko, V.V., Naumov, A.K., *et al.*, 1993. Ce^{3+} -doped Colquiriite. *Journal of Modern Optics* 40 (1), 1–5.
- Kaneda, Y., Yarborough, J.M., Merzlyak, Y., *et al.*, 2016. Continuous-wave, single-frequency 229 nm laser source for laser cooling of cadmium atoms. *Optics Letters* 41 (4), 705–708.
- Katori, H., Takamoto, M., Pal'chikov, V.G., *et al.*, 2003. Ultrastable optical clock with neutral atoms in an engineered light shift trap. *Physical Review Letters* 91 (17), 173005.
- Katsura, T., Kojima, T., Kurosawa, M., *et al.*, 2007. High-power, high-repetition UV beam generation with an all-solid-state laser. In: *CLEO/Europe and IQEC 2007 Conference Digest*. Munich: Optical Society of America, Munich, p. CA5_3.
- Kneissl, M., Rass, J., 2015. *III-Nitride Ultraviolet Emitters: Technology and Applications*. Springer, p. 208.
- Kozlovsky, W.J., Nabors, C.D., Byer, R.L., 1987. Second-harmonic generation of a continuous-wave diode-pumped Nd:YAG laser using an externally resonant cavity. *Optics Letters* 12 (12), 1014–1016.
- Marshall, C.D., Speth, J.A., Payne, S.A., *et al.*, 1994. Ultraviolet laser emission properties of Ce^{3+} -doped LiSrAlF_6 and LiCaAlF_6 . *Journal of the Optical Society of America B* 11 (10), 2054–2065.
- Nishioka, M., Fukumoto, S., Kawamura, F., *et al.*, 2003. Improvement of laser-induced damage tolerance in $\text{CsLiB}_6\text{O}_{10}$ for high-power UV laser source. In: *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference*. Baltimore, MA: Optical Society of America.
- Ostrovomov, V., Seelert, W., 2008. 1 W of 261 nm CW generation in a $\text{Pr}^{3+}:\text{LiYF}_4$ laser pumped by an optically pumped semiconductor laser at 479 nm. In: *Proceedings of SPIE, Solid State Lasers XVII: Technology and Devices*, San Jose, CA, pp. 68711K-687114.
- Ohtsuki, T., Kitano, H., Kawai, H., *et al.*, 2000. Efficient 193 nm generation by eighth harmonic of Er^{3+} -doped fiber amplifier. In: *Conference on Lasers and Electro-Optics*, paper CPD9.
- Paul, J., Kaneda, Y., Wang, T.-L., *et al.*, 2011. Doppler-free spectroscopy of mercury at 253.7 nm using a high-power, frequency-quadrupled, optically pumped external-cavity semiconductor laser. *Optics Letters* 36 (1), 61–63.
- Rajesh, D., Yoshimura, M., Eiro, T., *et al.*, 2008. UV laser-induced damage tolerance measurements of CsB_3O_5 crystals and its application for UV light generation. *Optical Materials* 31, 461–463.
- Sakuma, J., Kaneda, Y., Oka, N., *et al.*, 2015. Continuous-wave 193.4 nm laser with 120 mW output power. *Optics Letters* 40 (23), 5590–5593.
- Scheid, M., Markert, F., Walz, J., *et al.*, 2007. 750 mW Continuous-wave solid-state deep ultraviolet laser source at the 253.7 nm transition in mercury. *Optics Letters* 32 (8), 955–957.
- Siegman, A., 1986. *Lasers*. California: University Science Books, Mill Valley.
- Sudmeyer, T., Imai, Y., Masuda, H., *et al.*, 2008. Efficient 2nd and 4th harmonic generation of a single-frequency, continuous-wave fiber amplifier. *Optics Express* 16 (3), 1546–1551.
- Yoshida, H., Yamashita, Y., Kuwabara, M., *et al.*, 2008. Demonstration of an ultraviolet 336 nm AlGaIn multiple-quantum-well laser diode. *Applied Physics Letters* 93 (24), 241106.

Further Reading

Wernicke, T., Martens, M., Kuhn, C. *et al.*, 2015. Challenges for AlGaIn Based UV Laser Diodes. DM2D.4.

Single-Frequency Lasers

Xiushan Zhu, University of Arizona, Tucson AZ, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Single-frequency lasers are a category of lasers operating on only one longitudinal mode. Single longitudinal mode operation of a laser is achieved by inserting filters into the cavity to suppress the oscillation of all the other mode except the mode with the lowest threshold. The cavity of a single-frequency laser usually consists of a gain medium, an output coupler, one or two filters, and a cavity mirror, as shown in Fig. 1. Since a single frequency laser is susceptible to destabilization from backward reflection, Faraday isolator is always used after the output coupler to ensure high performance operation of single longitudinal mode. When a laser operates with only one longitudinal mode, it offers excellent features including ultra-narrow spectral linewidth, extreme stability in intensity and phase, and ultra-long coherence length, which are highly required for a variety of applications. A single-frequency laser can usually provide an amount of wavelength tunability, which depends on the bandwidth of the gain medium and the cavity design. Continuous-wave (CW) and Q-switched single-frequency lasers have been achieved with almost all types of current lasers. The average output power of single-frequency lasers can be elevated to hundreds of watts or even kW level by use of current laser amplifier technologies. Single-frequency lasers, both in CW and pulsed operation, have found a variety of applications in interferometric sensing, coherent light detection and ranging (LIDAR), coherent optical communication, high resolution spectroscopy, optical data storage, optical cooling and trapping, microwave photonics, and laser nonlinear frequency conversions.

Background

In their classical paper proposing an “optical maser”, single frequency operation of a laser was predicted by Schawlow and Townes (1958). However, single longitudinal mode operation was not observed in the first laser action in ruby (Maiman, 1960). Nevertheless, owing to the narrow gain bandwidth of gas gain medium, single-frequency operation of He-Ne laser was demonstrated by Kogelnik and Patel (1962) by using a resonant reflector acting as a spectral filter as well at one end of the resonator. From then on, different types of longitudinal mode selection schemes were used to achieve single longitudinal mode operation of a laser. Fabry-Perot etalons, for example, have been extensively used as interferometric filters in single-frequency solid-state lasers since Collins and White were the first to place tilted etalons inside the cavity (Collins and White, 1963). A single frequency laser can emit quasi-monochromatic radiation with very narrow linewidth and low noises, which are advantageous for diverse scientific investigations and engineering applications. For instance, intense research on single-frequency lasers resulted in the great discovery of the “Lamb dip” in 1963 (McFarlane *et al.*, 1963).

Spatial Hole Burning and Multimode Operation

Generally, the line-shape of stimulated emission cross-section of an ideal homogeneously broadened laser transition is fixed and the same for all atoms in the laser medium. Thus, the macroscopic gain of the laser medium at a given frequency depends only on the population inversion and the gain profile is just the constant atomic line-shape multiplied with a scalar factor, which is proportional to the population inversion. Therefore, in theory, a laser with an ideal gain profile always oscillates with one single longitudinal mode, which has a net gain of unity, i.e. under steady-state conditions the gain of this lasing mode exactly equals its losses, and all other modes experience a net gain of less than unity and will not lase. As shown in Fig. 2, when the population inversion is small, the net gains of all modes are less than the losses and all of the modes are below the threshold. As the population inversion increases, the net gains of all modes increase correspondingly and that of mode m equals the losses giving a net gain of unity. Therefore, only mode m reaches the threshold and starts to lase and all other modes are still below threshold. In this case, the laser operates on single longitudinal mode.

However, different from the prediction of Schawlow and Townes, people found that a laser usually tends to operate on multiple longitudinal modes when there is no mode selection element inside the laser cavity. The most significant effect leading to multi-longitudinal-mode operation in homogeneously broadened lasers is spatial inhomogeneity of the gain, especially “spatial

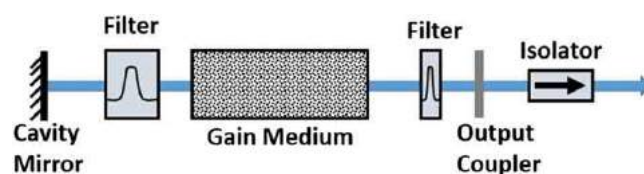


Fig. 1 Basic configuration of a single-frequency laser.

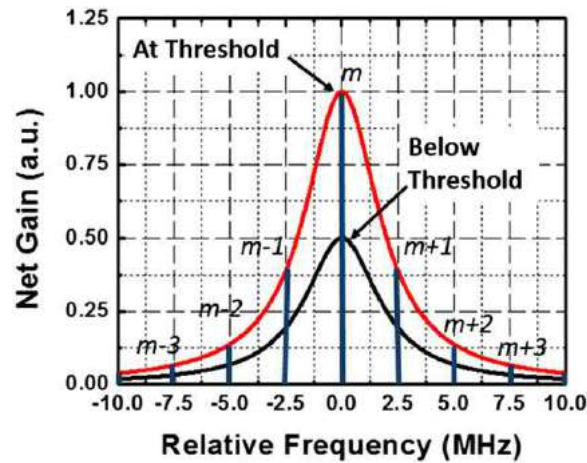


Fig. 2 Gain profile of an ideal homogeneously broadened laser transition and the net gain of corresponding longitudinal modes.

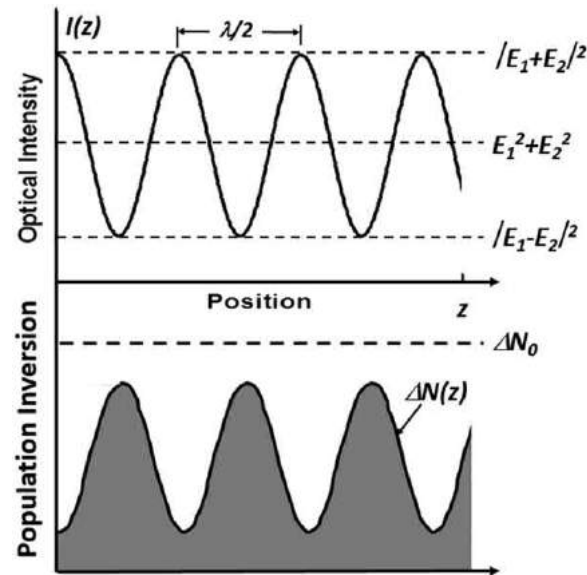


Fig. 3 Standing wave and the corresponding population inversion versus position in a laser gain medium.

hole burning” caused by the standing wave of the laser in the gain medium (Tang *et al.*, 1963). As shown in Fig. 3, in a linear cavity resonator, the forward (E_1) and the backward (E_2) propagating laser beams form a standing wave which depletes the gain much more strongly at the peaks of the standing wave than at the nodes. As a result, the population inversion becomes much larger at the nodes than at the peaks of the standing wave, thus leading to a spatial inhomogeneity of the gain. Since the bandwidth of the laser transition is normally much broader than the mode spacing, some adjacent modes will start to lase when these modes have their peaks near the peak positions of the population inversion and obtain a net gain equal to the losses as well. Therefore, due to the spatial hole burning, in a typical laser a few modes always oscillate simultaneously even if the gain of the adjacent modes is slightly smaller than that of the center mode. When a laser operates on multiple longitudinal modes, on the time scale of the round trip time (inverse of the mode spacing), the intensities of these modes change chaotically due to the beating between these lasing modes. On a time scale much longer than the cavity round trip time, the laser output power becomes fairly stable because it is the sum of the intensities of these lasing modes. Therefore, the intensity noise of a single-frequency laser is much lower than that of a laser operating on multiple longitudinal modes.

Single Mode Operation with Eliminated Spatial Hole Burning

There are basically two fundamental routes to achieve single-frequency operation of a laser. Because spatial hole burning is identified as the essential reason for multi-longitudinal-mode operation of a laser, the straightforward solution is to avoid spatial hole burning by preventing standing waves in the resonator. The alternative solution is to suppress the adjacent modes by

employing different mode-selection schemes to compensate or overwhelm the gain difference between the center mode and the adjacent modes caused by spatial hole burning.

The two most widely used methods to avoid the formation of standing waves are: the travelling wave laser and the “twisted-mode” laser. In the first paper explaining the role of “spatial hole burning”, Tang *et al.* (1963) suggested to develop travelling wave laser, such as unidirectional ring laser, to achieve the single frequency operation of a laser without using any mode-selection elements. In a unidirectional ring laser, as shown in Fig. 4(a), the light is forced to travel in one direction only by a Faraday isolator or an acoustic optical modulator and consequently standing waves are not formed. In 1965, Evtuhov and Siegman (1965) demonstrated that standing waves could be avoided by a “twisted-mode” technique, in which the laser medium is placed between two quarter-wave plates and the two counter-propagation waves are transformed into circular polarized light before entering the laser medium as shown in Fig. 4(b). The two circularly polarized waves travelling in the opposite direction do not interfere with each other if they have the same sense and therefore standing waves are not formed inside the gain medium and the gain is depleted uniformly. However, “twisted-mode” technique is sensitive to stress-induced birefringence caused by thermal load and mechanical stress. In addition, it cannot be used with birefringent laser materials.

Single Mode Operation with Mode Suppression

Although the adjacent modes and the center mode may experience the same gain of unity because of the spatial hole burning, single longitudinal mode operation can still be achieved if all longitudinal modes except the lasing mode are sufficiently suppressed by using different mode selection mechanisms.

The common approach is to incorporate an interferometric filter into the cavity to introduce much higher loss to all other modes than the lasing mode. Etalons acting as Fabry-perot filters have been widely used to achieve single-frequency operation of a laser. However, this approach requires the lasing mode to be at the minimum of the loss curve. Active locking techniques have to be employed to ensure stable single frequency operation. An alternative approach is to increase the mode spacing and reduce the gains of adjacent modes to values that are still smaller than unity even the effect of spatial hole burning is added. In a short cavity, the mode spacing becomes larger than the spectral width of a gain profile and thus single-frequency laser can be achieved. Sometimes, when the gain profile of a laser transition is broad or the bandwidth of the interferometric filter is larger than the mode spacing, both approaches have to be used to achieve single frequency operation as shown in Fig. 5.

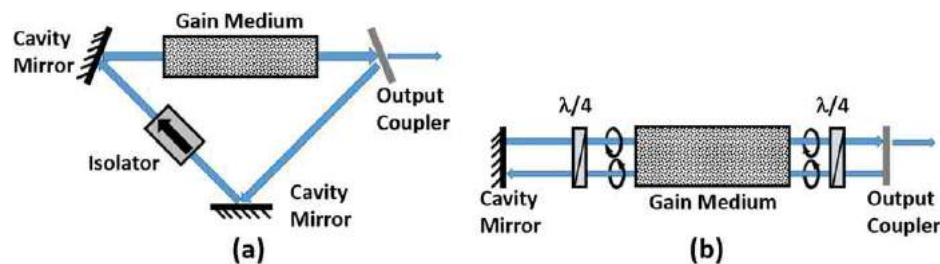


Fig. 4 (a) Configuration of a unidirectional ring cavity laser and (b) configuration of a twisted-mode laser.

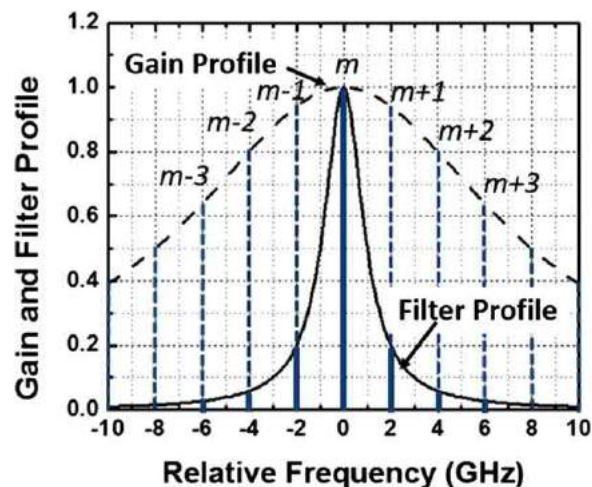


Fig. 5 Single-frequency operation of a laser achieved by using an interferometric filter and a short cavity with mode spacing of 2 GHz.

It should be noted that unidirectional ring laser and twisted-mode laser can effectively eliminate spatial hole burning and achieve single frequency operation with a homogeneously broadened gain medium. The two techniques, however, cannot be independently used to achieve single-frequency laser operation in some laser materials which have broad gain profiles or have both homogeneously and inhomogeneously broadened laser transitions. In this case, the two techniques have to be combined with the techniques of interferometric filters and short cavity to achieve single-frequency laser.

It should be noted that, due to the Guoy phase shift, a multi-transverse-mode laser doesn't operate with single-frequency even if the laser design meets the conditions for single longitudinal mode operation. Guoy phase shift depends on the order of the transverse modes and thus different transverse modes in a stable laser cavity have different oscillation frequencies. Therefore, a single-frequency laser has to be operated on single transverse mode as well.

Overview

Single longitudinal mode operation has been achieved in almost all types of current lasers by using different techniques discussed above or their combinations. The single frequency laser sources most available today are short cavity gas (i.e. He-Ne) lasers, ring cavity dye lasers, ring cavity or short cavity or twisted-mode solid-state lasers, ring cavity fiber lasers, semiconductor or fiber distributed feedback (DFB) lasers, semiconductor or fiber distributed Bragg reflector (DBR) lasers, and external cavity diode lasers.

Owing to the narrow gain bandwidth of gas, it is easy to develop a single frequency gas laser with a short length cavity (~ 20 cm or less). Ring cavity dye lasers can also easily operate on single longitudinal mode because of their homogeneously broadened laser transitions. Single-frequency dye lasers have the advantage of widely tunable wavelength ranges. However, gas and dye lasers are not widely used today because they are not efficient, reliable, robust, and compact as current solid-state lasers, semiconductor lasers, and fiber lasers.

Single frequency operation in solid-state lasers have been achieved with various techniques, such as microchip lasers, lasers inserted with etalons, twisted-mode lasers, and nonplanar ring oscillators (NPRO). A microchip laser usually has such a short cavity (less than 1 mm) that the free spectral range is much larger than the gain bandwidth and thus robust single-longitudinal-mode operation can be easily achieved (Kane and Byer, 1985). The techniques of using etalon filters and twisted-mode have already been discussed in detailed in Section 2. A NPRO is a unidirectional ring laser with special monolithic design in which the laser beam circulates in a single laser crystal (Zayhowski and Mooradian, 1989).

Nonplanar Ring Oscillator

As shown in Fig. 6, the resonator of an NPRO laser is a laser crystal cut at specific shapes enabling the unidirectional circulation of the laser beam via partial reflection of the dielectric mirror coated on the front face and total internal reflection on all other internal faces. The dielectric mirror coated on the front face is highly transmissive for the pump light and partially transmissive for the laser light and thus serves as the output coupler mirror of the resonator.

Unidirectional propagation of the laser light is obtained because the cavity loss of one oscillation direction is larger than that of the other one. Since the resonator is nonplanar, the polarization direction of the laser light is rotated slightly after each round trip. The two polarization rotations of one oscillation direction cancel partly, resulting in a smaller cavity loss when the laser light hits the front face of the crystal because the reflectivity of dielectric coating is polarization dependent. The other oscillation direction, however, experiences a larger cavity loss and is thus firmly suppressed. The discrimination between the two oscillation directions can be enhanced by attaching a small magnet to the laser crystal to increase the polarization rotation via the Faraday effect. Because a NPRO operates with travelling wave and the round trip length is short (< 10 cm), no spatial hole burning and large free spectral range make it very easy to obtain stable single-frequency operation. The relative large free spectral range also allows mode-hop free wavelength tuning over several gigahertz. The wavelength tuning can be achieved by applying a piezoelectric transducer on the laser crystal, incorporating an electro-optic crystal in the resonator, changing the crystal temperature, or adjusting the pump power.

NPROs have been developed with Yb: YAG at $1.03\ \mu\text{m}$, Nd: YAG at $1.06\ \mu\text{m}$, Er: YAG at $1.645\ \mu\text{m}$, Tm: YAG at $2.01\ \mu\text{m}$, and Ho: YAG at $2.09\ \mu\text{m}$. Because NPROs are monolithic and robust and have low optical losses, their laser noises are very small and

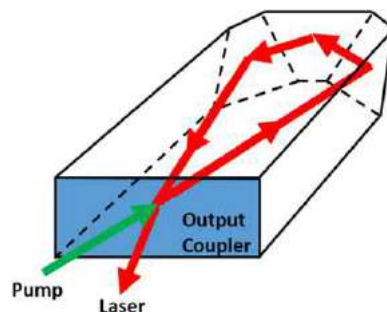


Fig. 6 Configuration of a nonplanar ring oscillator.

their typical linewidths can be a few kilohertz. Most importantly, the output power of a single-frequency NPRO can be up to several watts. Therefore, NPROs are often used as master oscillator for high-power single-frequency master oscillator and power amplifier laser system.

Because one of the prerequisites of single frequency laser is that its transverse mode must be single mode, complex resonator cavity design and fabrication and careful alignment are required to obtain single frequency operation in solid-state lasers, gas lasers, and dye lasers. Using single-mode waveguide, such as fiber and semiconductor, is therefore much easier to achieve stable and robust single-frequency operation of a laser. Fiber lasers and semiconductor lasers have become key workhorses for single frequency laser sources due to their compact sizes and robust single-longitudinal mode operation.

Ring Cavity Fiber Lasers

Fiber lasers have the advantages of compactness, reliability, cost-effective, alignment-free, and maintenance-free. Single-frequency fiber lasers combining the advantages of fiber lasers and single-frequency lasers have been thoroughly investigated and extensively used for various applications. However, a conventional fiber laser generally has meters of fiber length with linear cavity configuration and thus cannot generate single-frequency laser output due to the spatial hole burning in the gain medium as discussed above. This problem is usually solved by use of either a unidirectional ring cavity combined with a narrow-band filter or a very short (a few centimeters long) linear cavity combined with narrow-band fiber Bragg gratings capable of selecting a single longitudinal mode despite the presence of spatial hole burning.

Because the long fiber cavity can provide high gain and the potential for wide wavelength tunability by incorporating a wavelength tunable component into the ring cavity, the unidirectional fiber ring laser is very attractive for single-frequency laser development with high output power and wide wavelength tunable range. The lack of a spatial hole-burning effect in a travelling-wave field renders a ring-cavity design more suitable for single-frequency operation of a fiber laser with lower phase noise and a correspondingly smaller linewidth compared to linear cavity configuration. The configuration of a conventional single-frequency ring cavity fiber laser is shown in Fig. 7. The gain fiber is pumped by a pump laser source via a wavelength division multiplexer (WDM). Unidirectional propagation of the laser in the fiber ring is achieved with an isolator and the laser is coupled out from the fiber ring by a fiber coupler. Single-frequency operation of a fiber ring laser can also be achieved by other techniques such as using a passive multiple ring cavity or a compound ring resonator (Yeh *et al.*, 2006), using two cascaded Fabry-Perot filters (Park *et al.*, 1991), incorporating two or three fiber Bragg gratings (Guy *et al.*, 1995), and utilizing a saturable-absorber-based auto-tracking filter (Yeh and Chi, 2005). However, unidirectional ring lasers need expensive components such as Faraday isolators and TTFs and still have a tendency to mode hop that may have to be eliminated with an additional narrow-band optical filter in the fiber ring cavity.

Distributed Feed-back Lasers

As discussed above, single-frequency operation of a linear-cavity laser can be achieved with very short length cavity and narrow spectral filters. Distributed feedback (DFB) laser is a typical type of single-frequency laser in which the active region is periodically structured as a diffraction grating (Kogelnik and Shank, 1972). The active region thus not only provides optical gain for the laser but also optical feedback and mode selection for the laser.

DFB lasers either semiconductor lasers or fiber lasers have a typical configuration as shown in Fig. 8(a). The whole resonator consists of a periodic structure, which acts as a diffraction grating and contains a gain medium, and a highly reflective mirror.

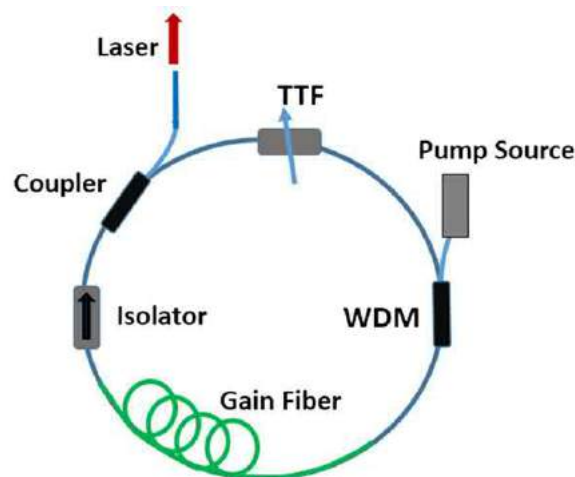


Fig. 7 Configuration of a single-frequency ring-cavity fiber laser.

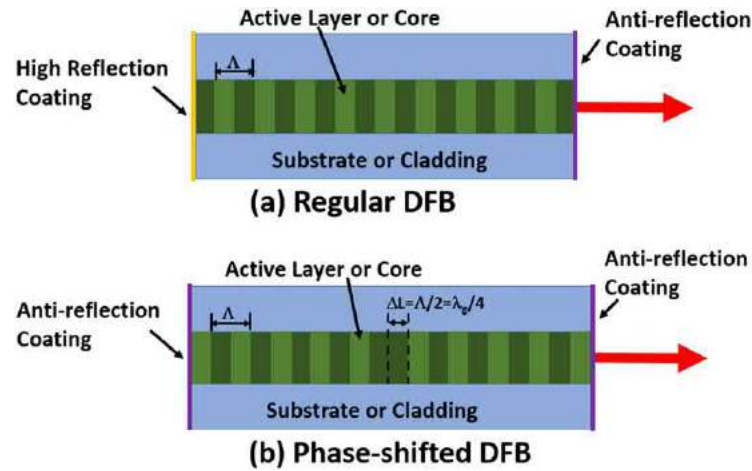


Fig. 8 Configuration of (a) regular DFB and (b) phase-shifted DFB single-frequency lasers.

The highly reflective mirror and the diffraction grating form the optical cavity. The diffraction grating is inscribed in the gain medium so as to reflect only a narrow band of wavelengths, and thus allow single longitudinal mode operation of the laser. The lasing wavelength for a DFB laser is approximately equal to the Bragg wavelength, which is defined by:

$$\lambda = 2n_{\text{eff}}\Lambda \quad (1)$$

where λ is the wavelength, n_{eff} is the effective refractive index of the guided laser mode, and Λ is the grating period. Because the effective refractive index and grating period are generally temperature dependent, altering the temperature of a DFB laser can change the wavelength selection of the grating and thus the wavelength of the laser output, producing a wavelength tunable laser. For a DFB laser either fiber or diode-based, the temperature tunability is about 0.01 nm/k. Altering the pump current of a DFB semiconductor laser will also lead to wavelength tuning.

An alternative approach is a phase-shifted DFB laser with a configuration as shown in Fig. 8(b). Typically, the periodic structure is made with a phase shift in its middle. This could be a single quarter-wave shift at the center of the cavity, or multiple smaller shifts distributed in the cavity. Due to the large free spectral range and ultra-narrow spectral filtering of a phase-shifted DFB laser, robust single-frequency operation is often easily achieved with excellent wavelength reproducibility and narrow linewidth.

For a DFB fiber laser, the distributed reflection occurs in a fiber Bragg grating, typically with a length of a few millimeters or centimeters. Efficient pump absorption can be achieved with a highly doped fiber such as phosphate glass fiber and the output power can be tens of milliwatts or over 100 mW. DFB fiber laser has excellent compactness and robustness leading to a low intensity and phase noise level, i.e., also a narrow linewidth. The linewidth of a DFB fiber laser is typically less than 100kHz.

Semiconductor DFB lasers can be built with an integrated grating structure, e.g. a corrugated waveguide, in a very compact package. Semiconductor DFB lasers are available for emission in different spectral regions (0.8–2.8 μm). Typical output powers are some tens of mW. The linewidth is typically a few MHz to several hundred MHz.

Distributed Bragg Reflector Lasers

Distributed Bragg reflector (DBR) laser is another type of single-frequency laser based on linear cavity design. DBR lasers either semiconductor lasers or fiber lasers have typical configurations as shown in Fig. 9. Different from a DFB laser, where the whole active medium is embedded in a single distributed reflector structure, a DBR laser has at least one DBR outside the gain medium. A resonant cavity is defined by a highly reflective DBR or thin-film mirror on one end, and a low reflectivity cleaved exit facet or DBR on the other end. The DBR mirror is designed to reflect only a single longitudinal mode. As a result, single-frequency operation of the laser is obtained.

DBR and DFB lasers have distinct operational characteristics caused by their different locations of the feedback elements. Because the grating of a DFB is distributed all along the gain region, the grating and gain region experience similar conditions as the device is tuned with current and temperature. The DFB can exhibit a continuous tuning range of 2 nm or more. However, over a sufficiently long current or temperature range, the emission wavelength will suddenly jump to a longer wavelength, leaving a gap in the tuning range. Because the DBR laser has a passive grating region, its tuning characteristic of the DBR laser is different from that of the DFB laser. Increasing current in the gain region causes a red shift in laser output due to heating. However, the reflectivity curve of the passive grating does not change and the grating will experience loss of reflectivity at the longer wavelengths. As a result, the wavelength characteristic of a DBR laser will repeat itself with increasing temperature or current, and no gaps will occur in the tuning.

The DBR diode laser is fabricated with a monolithic, single mode ridge waveguide running the entire length of the device (Dupuis and Dapkus, 1978). The wavelength tunable range of a DBR diode laser is approximately a 2 nm range by changing

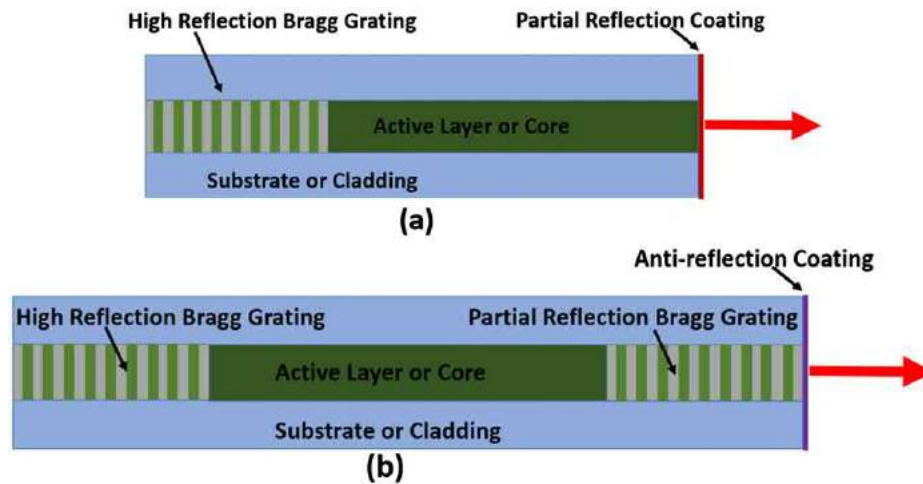


Fig. 9 Configuration of single-frequency DBR lasers with (a) one Bragg grating and (b) two Bragg gratings.

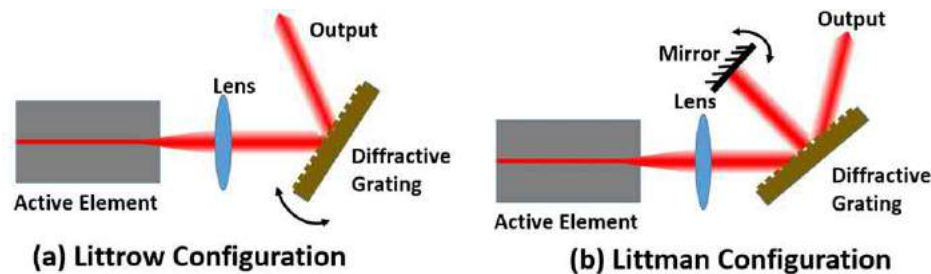


Fig. 10 Configuration of external cavity single-frequency lasers.

current or temperature. The temperature coefficient is approximately 0.07 nm/K, and the current coefficient is approximately 0.003 nm/mA. The typical linewidth of a DBR laser is less than 10 MHz. The output power typically can run up to several hundred milliwatts.

DBR fiber lasers have the advantages of a compact all-fiber design, easy and accurate wavelength selection during production, a large potential wavelength tuning range without mode-hopping through strain or temperature tuning, and narrow linewidth. The temperature coefficient is approximately 0.01 nm/K. A typical linewidth of a DBR fiber laser is less than 10 kHz. The output power typically can run up to several watts with a double-clad phosphate fiber laser (Polynkin *et al.*, 2005; Schulzgen *et al.*, 2006).

External Cavity Lasers

The wavelength tunable range of a DFB or DBR diode laser is typically a few nm. Broad wavelength tunable range (> 100 nm) can be achieved with external cavity diode lasers (ECDLs), which usually use bulky grating and free-space optics that is compatible with most standard diode lasers. As shown in Fig. 10, a lens collimates the output beam of the diode, which is then incident upon a grating. The grating provides optical feedback for a very narrow spectral light and is used to select the stabilized output wavelength. Single-frequency operation of the ECDLs with high side mode suppression ratio (SMSR > 45 dB) and very broad wavelength tunability can be obtained with proper optical design allowing only a single longitudinal mode to lase. Fiber Bragg grating (FBG) can be used to replace the free-space optics and grating to make ECDLs very compact and robust. However, the wavelength tunability of FBG-based ECDLs is usually not as broad as that of free-space ECDLs due to the limited wavelength tunable range of fiber Bragg grating. In addition to excellent wavelength tunability, compared to DFB and DBR diode lasers, ECDLs have much narrower linewidth (< 1 MHz) due to the relatively long cavity. The ECDL can have a large tuning range but is often prone to mode hops, which are very dependent on the mechanical design as well as the quality of the antireflection (AR) coating on the laser diode.

FBG-based ECDLs have a very simple configuration that the light from the diode is coupled into the fiber of FBG directly and the wavelength tuning is achieved by heating or mechanically stretching and compressing the FBG. Free-space ECDLs have two different configurations: Littrow (Fleming and Mooradian, 1981) and Littman (Liu and Littman, 1981). The common Littrow configuration is shown in Fig. 10(a). The first-order diffracted beam provides optical feedback to the laser diode and the wavelength tuning is realized by rotating the diffraction grating. A disadvantage of Littrow configuration is that the direction of the

output beam is also changed. The Littman configuration is shown in Fig. 10(b). The grating orientation is fixed, and an additional mirror is used to reflect the first-order beam back to the laser diode. The wavelength tuning is achieved by rotating that mirror. This configuration offers a fixed direction of the output beam. In addition, the Littman configuration intends to exhibit a smaller linewidth than Littrow configuration because the wavelength-dependent diffraction occurs twice per resonator round trip in the Littman configuration. However, the output power of a Littman laser is less than that of a Littrow laser because the zero-order reflection of the beam reflected by the tuning mirror is lost.

Characterization

Single-frequency lasers exhibit excellent features of low noises and narrow spectral linewidth. Relative intensity noise (RIN), phase noise or frequency noise, and spectral linewidth are key parameters defining the performance of a single-frequency laser.

Relative Intensity Noise

RIN is a typical parameter to characterize the optical power fluctuation of a laser and can be expressed as

$$RIN(t) = \frac{\langle \delta P(t)^2 \rangle}{\langle P(t) \rangle^2} \quad (2)$$

where $\delta P(t)$ is the power fluctuation of the laser and $\langle P(t) \rangle$ is the average power. RIN can also be described with a power spectral density depending on the noise frequency f and be calculated as the Fourier transform of the autocorrelation function of the normalized power fluctuations:

$$RIN(f) = \frac{2}{\langle P(t) \rangle^2} \int_{-\infty}^{+\infty} \langle \delta P(t) \delta P(t + \tau) \rangle e^{i2\pi f \tau} d\tau \quad (3)$$

RIN of a laser is often measured in the electrical domain with a photodiode and analyzed with an electronic spectrum analyzer (ESA) using a setup as shown in Fig. 11(a). The optical power of the laser is directly detected by an optical detector. The RIN is simply the ratio of the DC photocurrent to the AC noise on the detector output and then displayed on the ESA. The electric AC noise is proportional to optical power squared and hence RIN is usually presented as relative fluctuation in the square of the optical power specified in dB/Hz.

Intensity noise of a laser results partially from quantum noise associated with laser gain and resonator losses and partially from vibrations and thermal issues of the cavity, fluctuations in the laser gain medium, and the excess noise of the pump source. The RIN of a single-frequency laser is typically independent of laser power and is shown in Fig. 11(b). RIN typically has a peak at the relaxation oscillation frequency of the laser and then decreases with the increased frequency until it converges to the shot noise level. Since the relaxation oscillation is related to the emission transition of the gain medium, the relaxation oscillation frequency of DFB or DBR diode lasers is usually larger than 1 GHz while that of solid-state lasers and fiber lasers is typically in a frequency range between a few tens kHz and a few MHz. The relaxation frequency peak of RIN can usually be suppressed with a close-loop servo system. The RIN with suppressed relaxation oscillation frequency peak is shown in the inset of Fig. 11(b).

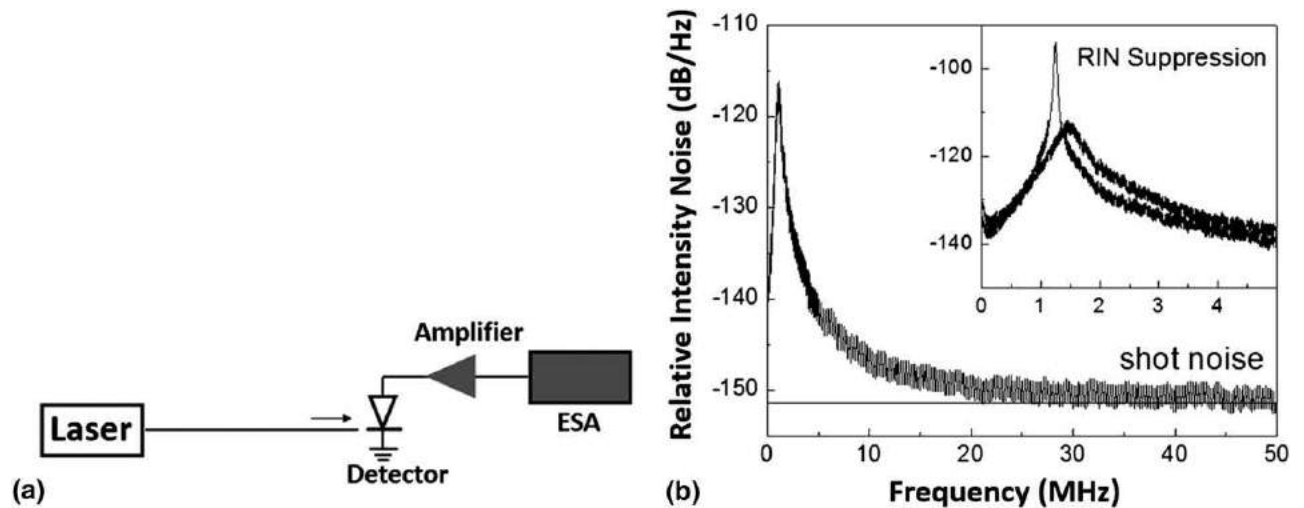


Fig. 11 (a) Configuration of setup for RIN measurement; (b) typical RIN of a single-frequency DBR fiber laser at 1.55 μm . Inset: relaxation oscillation frequency peak is suppressed.

Phase or Frequency Noise

Phase noise is the random deviation of the oscillation phase from a purely linear phase evolution, which results in random fluctuation of the instantaneous frequency of the single-frequency laser, namely frequency noise. Hence the output of a single-frequency laser is not a perfectly monochromatic but with a finite linewidth. Phase noise is mainly caused by quantum noise, cavity noise and intensity noise. The power spectral density of single-sided phase noise includes three contributions: white, flicker and random walk noises, and can be written as

$$S_\phi(f) = \frac{\delta(\nu)}{\pi f^2} + \frac{k_f}{f^3} + \frac{k_r}{f^4} \quad (4)$$

where $\delta(\nu)$ is the Lorentzian spectral linewidth of the laser, k_f and k_r are constants that give the strength of flicker frequency noise and random walk frequency noise, respectively. The power spectral density of frequency noise is directly related to that of the phase noise by the equation

$$S_{(\nu)}(f) = f^2 S_\phi(f) \quad (5)$$

Phase noise or frequency noise is generally directly measured by frequency discriminators, which convert frequency fluctuation to intensity fluctuation and then measure the power spectral density, $S_\nu(f)$, of the optical frequency fluctuations by detecting the intensity variations. Frequency discriminator is usually constructed with a configuration of interferometer, such as a Michelson interferometer, a Mach-Zehnder interferometer or a Fabry-Perot interferometer. A typical frequency discriminator based on an all-fiber Michelson interferometer is shown in Fig. 12(a). The single-frequency laser is launched into one left port of a 2×2 50/50 fiber coupler and split into the two arms of the fiber coupler. At the two right ports of the fiber coupler, two mirrors are used to reflect the laser back into the fiber coupler. The output at the other left port is detected by an optical detector and analyzed by a dynamic spectrum analyzer (DSA). The measured frequency noise of single-frequency DBR fiber lasers at 1550 and 1060 nm are shown in Fig. 12(b). The frequency noise typically decreases with the increased frequency and exhibits some spikes related to various noises. Generally, a single-frequency laser with a lower frequency noise has a narrower linewidth. However, frequency discriminator does not directly yield linewidth. The fundamental laser linewidth can be determined from the power spectral density of the optical frequency. Direct measurement of the spectral linewidth of a single-frequency laser should be accomplished with heterodyne methods.

Spectral Linewidth

Due to quantum mechanical fluctuations, cavity vibrations and fluctuations, gain medium fluctuations, and other laser noises, a single-frequency laser is not perfectly monochromatic and has finite spectral linewidths. The spectral linewidth of a single frequency laser is the full width at half maximum (FWHM) of the optical spectrum. When a laser only has quantum fluctuations, it has the Schawlow-Townes linewidth,

$$\Delta\nu = \frac{4\pi h\nu(\Delta\nu)^2}{P_{out}} \quad (6)$$

where $h\nu$ is the photon energy, $\Delta\nu$ is the resonator bandwidth, and P_{out} is the output power. Spontaneous emission of excited atoms and ions is a major quantum noise and makes the laser output has a finite spectral width. Nevertheless, the spectral linewidth of a single-frequency cannot be measured with typical grating-based optical spectrum analyzers or scanning filter methods. Frequency discriminator, heterodyne detection, and self-heterodyne detection have to be used for high resolution linewidth measurement. The spectral linewidth of a single-frequency laser cannot be directly obtained with frequency

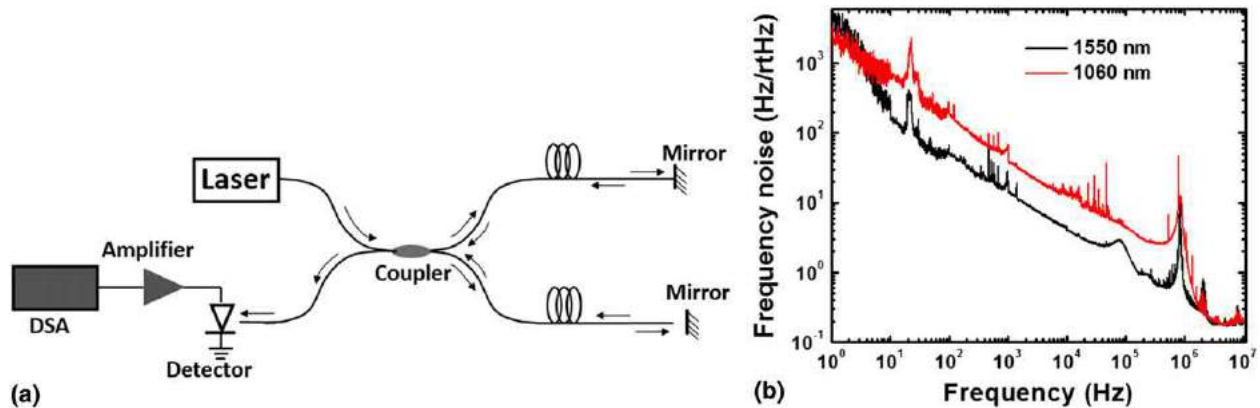


Fig. 12 (a) Configuration of the setup for frequency noise measurement and (b) typical frequency noise of single-frequency DBR fiber lasers at 1550 and 1060 nm.

discriminator but can be deduced from the data of frequency noise. Heterodyne and delayed self-heterodyne detection methods capable of measuring the spectral linewidth of laser directly have been extensively used.

Heterodyne detection

The basic configuration of heterodyne detection is shown in Fig. 13. A very stable single-frequency laser is used as the reference, namely local oscillator laser. The single-frequency laser under test is the signal laser. The central frequency of the local oscillator laser must be tuned close to that of the signal laser to allow the beat frequency to fall within the bandwidth of typical detection electronics. The local oscillator laser and the signal laser under test are launched into the two input ports of a 2×2 50/50 fiber coupler, respectively. An optical spectrum analyzer (OSA) and a photodetector are connected to the two output ports of the fiber coupler, respectively. The OSA is used for coarse wavelength tuning. The photodetector is used to detect the interference beat note and converts it to an electrical tone displaying on an electrical spectrum analyzer.

Heterodyne detection method can not only detect the spectral linewidth of a single-frequency laser, but also measure its optical power spectrum. This method is the only technique that is capable of characterizing non symmetrical spectral lineshape. The key component required for this method is a stable, narrow linewidth reference laser. The linewidth of the reference laser must be narrower than or at least comparable to that of the laser source to be measured in order to achieve reasonable measurement accuracy. However, this method becomes inaccurate for extremely narrow linewidth measurements since the characterization of the reference laser itself is very difficult. The applicability of this method is also limited by the bandwidth of the photodetector limiting the measurement frequency range to at most tens of gigahertz vicinity of the reference laser central frequency.

Delayed self-heterodyne detection

Delayed self-heterodyne detection doesn't need a separate local oscillator and the spectral linewidth measurement is based on the interference between the laser under test and a delayed replica of itself. The basic configuration of delayed self-heterodyne detection technique for spectral linewidth measurement is shown in Fig. 14(a). The basic idea of this technique is to convert the optical phase or frequency fluctuations of the laser into variations of light intensity in a Mach-Zehnder type interferometer. The light of the laser under test is split into the two arms of the interferometer by a 50/50 fiber coupler. An acousto-optic modulator is used on one arm to shift the spectrum up in frequency by Ω in order to reject the detected DC signal in the photodiode and to allow the use of a standard RF spectrum analyzer to measure the spectrum of the photocurrent fluctuations. On the other arm a long piece of optical fiber is used to achieve the optical delay that is longer than the coherence length of the laser under test. The optical fields from the two arms are mixed by another 50/50 fiber coupler and the interference signal is detected with a fast photodiode. The laser linewidth is then deduced from the recorded power spectrum of the fluctuations of the photocurrent.

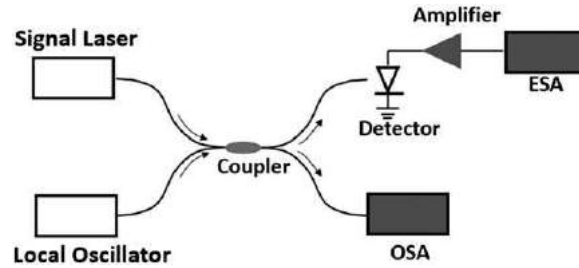


Fig. 13 Configuration of the setup for heterodyne detection.

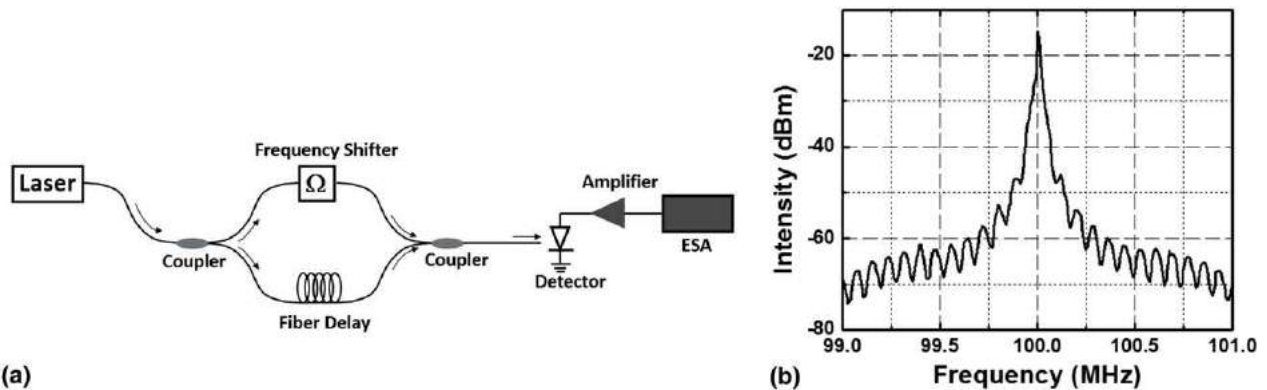


Fig. 14 (a) Configuration of the setup for self-delayed heterodyne detection; (b) typical spectral linewidth measurement result of self-delayed heterodyne detection for a single-frequency DBR fiber laser at 1550 nm.

A typical linewidth measurement result of a single-frequency DBR fiber laser is shown in Fig. 14(b). Because the linewidth of this laser is very narrow and the optical delay fiber is comparable to its coherent length, the measurement result has some ripples there. Generally, an optical delay length of at least 5 times longer than the laser coherence length was suggested for a direct laser linewidth measurement from the spectrum. For very narrow linewidths, however, the required length for the optical delay line may become unpractically long. Therefore self-delayed heterodyne method is only suited for direct measurement of laser linewidths on the order of or above 10 kHz. Nevertheless, the linewidth of a single-frequency laser with a coherence length longer than the optical delay length can be estimated from -20 dB bandwidth of the power spectrum.

Compared with heterodyne detection, the delayed self-heterodyne method doesn't need a local oscillator. Since the local oscillator signal in these measurements is provided by the laser under test, slow drift in wavelength is usually tolerable. Hence delayed self-heterodyne detection is inherently self-calibrated and capable of measuring a large range of laser frequencies where the fiber loss is tolerable.

Applications

Single-frequency lasers have found a variety of applications where long coherence length or narrow spectrum linewidth is essential. Those applications include high resolution spectroscopy, coherent light detection and ranging (LIDAR), remote sensing, laser nonlinear frequency conversions, coherent optical communication, optical data storage, optical cooling and trapping, and microwave photonics.

High resolution spectroscopy highly depends on the linewidth of the laser source. For instance, the absorption linewidth of a gas at standard temperature and pressure is in the range of a few GHz, so the linewidth and stability of the detection laser should be less than 10 MHz for high precision measurements of the absorption lineshape and strength. Single-frequency lasers with inherent narrow linewidth and high intensity stability are excellent laser sources for high resolution spectroscopy.

Single-frequency lasers are in great demand for LIDARs, which can be used to remotely sense properties of targets including the distance, the speed, the particle size and the concentration. Due to the long coherence length of the laser, these properties can be detected by measuring the intensity or frequency shift of returned pulses scattered by the target. LIDARs have been used for scientific meteorological applications and for commercial applications such as aviation safety control and weather forecasting. LIDARs can also be used to remotely monitor the physiological status of vegetation, the decline of forest, and the density of fish population.

Because the beating between longitudinal modes doesn't occur in a single-frequency laser, its output power is extremely stable, which is especially advantageous for nonlinear processes such as harmonic generation or optical parametric oscillator and optical data storage where a low intensity noise is required.

Single-frequency laser sources are also attractive for resonant enhancement cavities for high efficiency nonlinear frequency conversions and for coherent beam combining. Both applications require narrow-linewidth lasers with high stability.

References

- Collins, S.A., White, G.R., 1963. Interferometer laser mode selector. *Applied Optics* 2, 448–449.
- Dupuis, R.D., Dapkus, E.P., 1978. Room-temperature operation of distributed-Bragg-confinement $\text{Ga}_{1-x}\text{Al}_x\text{As}$ -GaAs lasers grown by metalorganic chemical vapor deposition. *Applies Physics Letters* 33, 68–70.
- Evtuhov, V., Siegman, A., 1965. A 'twisted-mode' technique for obtaining axially uniform energy density in a laser cavity. *Applied Optics* 4, 142–143.
- Fleming, M., Mooradian, A., 1981. Spectral characteristics of external-cavity controlled semiconductor lasers. *IEEE Journal of Quantum Electronics* 17, 44–59.
- Guy, M.J., Taylor, J.R., Kashyap, R., 1995. Single-frequency erbium fiber ring laser with intracavity phase-shifted fiber Bragg grating narrowband filter. *Electronics Letters* 31, 1924–1925.
- Kane, T.J., Byer, R.L., 1985. Monolithic, unidirectional single-mode ring laser. *Optics Letters* 10, 65–67.
- Kogelnik, H., Patel, C.K.N., 1962. Mode suppression and single frequency operation in gaseous optical masers. *Proceedings of the IRE* 50, 2365–2366.
- Kogelnik, H., Shank, C.V., 1972. Coupled-wave theory of distributed feedback lasers. *Journal of Applied Physics* 43, 2327.
- Liu, K., Littman, M.G., 1981. Novel geometry for single-mode scanning of tunable lasers. *Optics Letters* 6, 117–118.
- Maiman, T.H., 1960. Stimulated optical radiation in Ruby. *Nature* 187, 493–494.
- McFarlane, R.A., Bennett, W.R., Lamb, W.E., 1963. Single mode tuning dip in the power output of an He-Ne optical maser. *Applied Physics Letters* 2, 189–190.
- Park, N., Dawson, J.W., Vahala, K.J., 1991. All fiber, low threshold, widely tunable single-frequency, erbium-doped fiber ring laser with a tandem fiber Fabry-Perot filter. *Applied Physics Letter* 59, 2369–2371.
- Polynkin, A., Polynkin, P., Mansuripur, M., Peyghambarian, N., 2005. Single-frequency fiber ring laser with 1 W output power at 1.5 μm . *Optics Express* 13, 3179–3184.
- Schawlow, A.L., Townes, C.H., 1958. Infrared and optical masers. *Physical Review* 112, 1940–1949.
- Schulzgen, A., Li, L., Temyanko, V.L., Suzuki, S., Moloney, J.V., Peyghambarian, N., 2006. Single-frequency fiber oscillator with watt-level output power using photonic crystal phosphate glass fiber. *Optics Express* 14, 7087–7092.
- Tang, C.L., Statz, H., deMars, C., 1963. Spectral output and spiking behavior of solid-state lasers. *Journal of Applied Physics* 34, 2289–2295.
- Yeh, C.H., Huang, T.T., Chien, H.C., Ko, C.H., Chi, S., 2006. Tunable S-band erbium-doped triple-ring laser with single-longitudinal-mode operation. *Optics Express* 15, 382–386.
- Yeh, C., Chi, S., 2005. A broadband fiber ring laser technique with stable and tunable signal-frequency operation. *Optics Express* 13, 5240–5244.
- Zayhowski, J.J., Mooradian, A., 1989. Single-frequency microchip Nd laser. *Optics Letters* 14, 24–26.

Semiconductor Lasers

Stephan W Koch, Philipps University of Marburg, Marburg, Germany
Martin R Hofmann, Ruhr University Bochum, Bochum, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The optical emission from semiconductor lasers arises from the radiative recombination of charge carrier pairs, i.e., electrons and holes in the active area of the device. The conduction-band electron fills the valence-band hole by simultaneously transferring the energy difference to the light field, so-called radiative recombination. Hence, the frequency of the laser light is mainly determined by the bandgap of the semiconductor laser material. In order to achieve lasing, one needs carrier inversion, i.e., sufficiently many electrons in the conduction band such that the light absorption probability is more than compensated by the probability of emission.

A large variety of semiconductor materials has been shown to be suited for semiconductor lasers. The most stringent requirement for the material is that it has a direct bandgap. With other words, the maximum of the valence band and the minimum of the conduction band have to be at the same position in k -space (in the center of the Brillouin zone). This condition excludes the elemental semiconductors Silicon and Germanium from being used for semiconductor lasers.

In order to achieve carrier inversion, it is necessary to pump the semiconductor laser, i.e., to excite sufficiently many electrons from the valence band into the conduction band. Even though this pumping in semiconductors, as in other solid state laser materials, can be done optically, the large technological impact of semiconductor lasers is at least in part due to the possibility of electrical pumping with a few tens of milli-Amperes at voltages of a few Volts.

The original version of such a semiconductor laser device is shown in Fig. 1. The laser structure consists of a GaAs-based p-n junction which is biased in forward direction. Electrons are injected from the n-region and holes from the p-region, respectively. Due to the diode characteristics of the p-n junction, semiconductor lasers are often also called laser diodes. In the device shown in Fig. 1 the laser mirrors are formed by the cleaved facets of the semiconductor crystal.

Because of the high refractive index of GaAs, the semiconductor-air interface provides a reflectivity of about 32%, which is sufficient for laser emission. However, the efficiency of this early type of lasers, the so-called homojunction device, was very poor because of a low carrier density in the active area and because of a weak overlap between the inverted region and the optical mode. Considerable improvement toward room-temperature continuous wave (CW) operation was achieved by the introduction of the so-called double heterostructure. Its principle is shown in Fig. 2.

The basic idea behind the double-heterostructure laser is that the active material (e.g., GaAs) is embedded in another semiconductor with a slightly larger bandgap (e.g., (AlGa)As). This causes a potential well in which electrons and holes are confined in order to achieve high carrier densities in the active area. Moreover, the smaller refractive index of the surrounding high-bandgap material helps to guide the optical field within the region of highest inversion. Very often, additional ridge-shaped lateral structuring is used to form a complete optical waveguide.

Even more advanced laser structures can be fabricated since modern crystal growth techniques, such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD), became available. These techniques allow the grower not only to determine the composition of the semiconductors with remarkable precision, but also to define the shape virtually on an atomic scale. In particular, it is now possible to grow microstructures so small that their electronic and optical properties deviate substantially from those of the corresponding bulk materials.

So-called quantum-well (QW) structures, in particular, turned out to be superior to bulk materials in many aspects and are currently the active medium in most commercial semiconductor lasers. The gain medium in QW laser structures consists of one or more films with a thickness of several atomic monolayers. These are embedded in barrier material with a larger bandgap that serves to confine the carriers into the wells. QWs are examples of structures where quantum confinement restricts the free carrier motion to the two dimensions within the plane of the films since the well thickness is comparable to or less than the thermal wavelength of the carriers. Hence, one speaks of quasi-two-dimensional (2D) structures where no free motion perpendicular to the films is

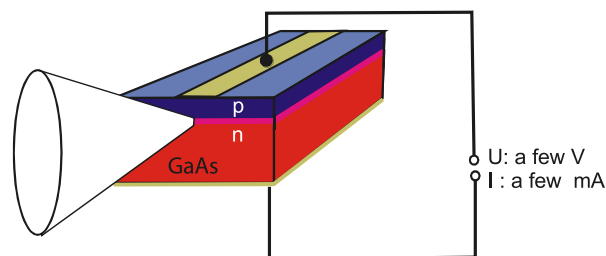


Fig. 1 Operation principle of semiconductor laser.

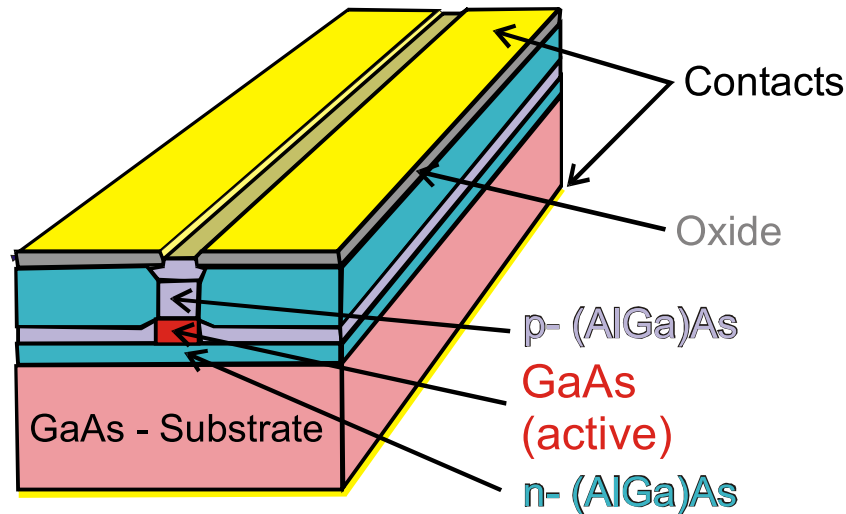


Fig. 2 Double-heterostructure laser diode.

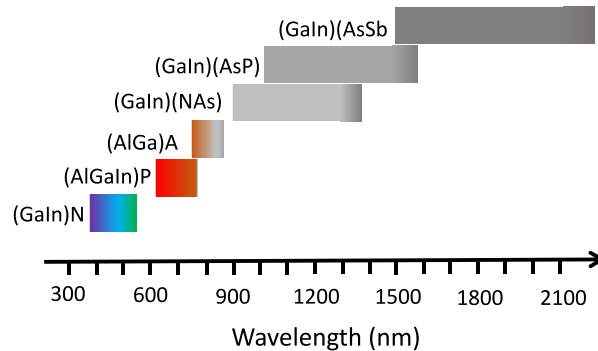


Fig. 3 Typical diode laser materials and their emission wavelength ranges.

possible. The corresponding energy states of the carriers are quantized, i.e., only discrete values are allowed for that part of the carrier kinetic energy which is related to the confined direction (Haug and Koch, 2009).

Such quantum confinement effects are not restricted to one dimension as in QWs, but can also be obtained in two and three dimensions, such as in quantum wires and quantum dots, respectively. Even though wire and dot structures can be grown with various degrees of precision, their use as semiconductor laser gain medium is still an emerging field. However, especially quantum dot structures hold great promise for the future, since they can be grown in large numbers using comparable techniques already used for QWs (Bimberg *et al.*, 1998). The possibility of layer growth under strained conditions, such as QWs, between a barrier material with a slightly different atomic lattice constant, also enriches the flexibility to design the laser gain medium with the desired emission wavelength.

In principle, the emission wavelength of a semiconductor laser can be roughly designed by the choice of the material. Different semiconductors have different gaps between valence band and conduction band and the bandgap energy determines the spectral range of the emission. By fabricating semiconductor alloys as, for example, (GaIn)As or (AlGa)As, one can tune the bandgap and thus the spectral range of the emission. Presently, a large variety of different materials is available, for example, covering the red ((AlGaIn)P), the near infrared ((AlGa)As, (GaIn)As), the telecom range ((GaIn)(AsP)), infrared (GaIn)(AsSb), and even the blue-ultraviolet range ((GaIn)N). In recent years, the progress in material growth has even enabled the realization of green laser diodes based on the (GaIn)N material system. Fig. 3 gives an overview of the most common commercial diode laser materials and their emission wavelengths.

Great care has to be taken that the crystal used for a semiconductor laser has almost perfect quality. In detail, dislocations in the crystal have to be avoided as well as any kinds of defects since they act as nonradiative recombination centers. Nonradiative recombination competes with the radiative recombination that is essential for the laser process. Nonradiative recombination effects include Auger-recombination and recombination via defects. Auger-recombination is the process where an electron and a hole recombine and completely transfer the energy difference to another carrier (electron or hole), which is excited to a higher state. This is an intrinsic effect that depends mainly on the band structure and the relevant values of the transition probabilities

(transition matrix elements). The probability for Auger-recombination is temperature and carrier density dependent and can be minimized, for example, by properly cooling the device and by keeping the carrier densities as low as possible.

Nonradiative recombination via defects is an extrinsic effect that depends on the material quality and can thus be minimized by optimized growth. The goal to avoid dislocations, however, introduces severe limitations for the choice of materials for laser structures. The growth processes require “host” crystals, so-called substrates, on which the real structure can be grown. As standard substrates for optoelectronic applications, GaAs, InP, SiC, and sapphire wafers are available in sufficient quality. The layers grown by epitaxy assume the crystal structure of the substrate which implies that the lattice constants of substrate and of the layers grown on top should be very similar. If the lattice constants do not agree, i.e., if there is structural mismatch, a certain amount of strain develops that increases with increasing lattice mismatch. Too much strain leads to the formation of dislocations, which are detrimental for laser applications. The control of the dislocation density was, for example, one of the critical issues for the wide-gap III–V materials, such as GaN, because no substrate was available that matched the GaN lattice constant. A small value of strain, however, is often even desirable as long as it does not cause dislocations. Strain influences the band structure of the material in a predictable way and can therefore be used to tailor the laser emission properties.

Searching a laser material for a certain desired wavelength range thus implies two important requirements that in some cases may only be simultaneously met with quaternary semiconductor alloys (e.g., (GaIn)(AsP) on InP-substrates): first, the bandgap of the material has to fit the desired photon energy and second, and more severe, a substrate with the right lattice constant has to be found. This second condition excludes many material compositions from laser applications. In many cases, both conditions can. Presently, this problem introduces enormous challenges in the growing area of optoelectronic/electronic integrated circuits. The growing demand on higher data transfer rates in modern communication systems motivates the use of optical rather than electrical connections even on a microchip level. This requires the embedding of laser diodes into electronic circuits, which are based on silicon technology. Since silicon as an indirect semiconductor is not an appropriate material for semiconductor lasers, alternative technologies have to be implemented to include laser diodes into electronic circuits. This may either be realized with large technological effort by fusion of electronic and optoelectronic wafers or by developing new laser materials that have a lattice constant compatible with that of silicon as, for example, Ga(NAsP).

Physics of the Gain Medium

From a material physics point of view, the important structural aspects of a certain laser material are summarized in the band structure of the gain medium, i.e., in the quantum mechanically allowed energy states of the material electrons. The calculation of the band structure of a particular gain medium is therefore a crucial aspect of semiconductor laser modeling (Chuang, 1995; Chow and Koch, 1999). Additionally, however, one has to understand the relevant properties of the excited electron–hole plasma in the laser’s active region.

In its ground state, i.e., without excitation and at very low temperatures, a perfect semiconductor is an insulator where no electrons are in the conduction-band states. The presence of a population inversion, i.e., occupied conduction-band and empty valence-band states in a certain range of frequencies is, however, a requirement for obtaining optical gain and light amplification from a semiconductor. Sufficiently strong pumping leads to the generation of an electron–hole plasma, where the term “electron” specifically refers to conduction-band electrons and “holes” is a short-hand term for the missing electrons in the originally full valence bands. Holes are also quasiparticles, just as the electrons in a solid are quasiparticles, which have an effective mass that is considerably different from the free-electron mass in vacuum and which is determined by the band structure, i.e., the interaction with all the ions of the solid. Thus the properties of the missing electrons, such as a positive charge for the missing negative electron charge, a spin opposite to that of the missing electron, etc., are attributed to the holes. As a consequence of their electrical charges, the excited electrons and holes interact via the Coulomb potential which is repulsive for equal charges (electron–electron, hole–hole) and attractive for opposite charges (electron–hole). Furthermore, electrons and holes obey the Fermi–Dirac statistics and the Pauli exclusion principle not allowing two Fermions to occupy the same quantum state. The consequences of the Pauli principle are often referred to as “band filling effects” in the physics of semiconductor lasers.

The theory therefore has to account not only for the band structure but also for this band filling, i.e., the detailed population of the electron and hole states. Taking into account only band structure and band filling leads to a so-called free-carrier theory. However, the carrier interactions and the exclusion principle render the electron–hole plasma in a semiconductor gain medium an interacting many-body system where the properties of any one carrier are influenced by all the other carriers in their respective quantum states. Generally, advanced many-body techniques are needed to systematically analyze such an interacting electron–hole plasma (Haug and Koch, 2009) and to compute the resulting optical properties, such as gain, absorption or refractive index, all of which are changing in a characteristic way with the changing electron–hole density.

It is often helpful to investigate the characteristic timescales and the most direct consequences of the various interaction effects even though the consequences of the different many-body effects are in general not additive. The fastest interaction under typical laser conditions is the carrier–carrier Coulomb scattering that acts on scales of femto- to picoseconds to establish Fermi–Dirac distributions of the carriers within their bands. These distributions are often referred to as “quasi-equilibrium distributions” since they describe electrons and holes in the conduction and valence bands, which may be characterized by an “electronic temperature” or “plasma temperature.” The plasma temperature is a measure for the mean kinetic energy of the respective carrier ensemble and relaxes toward the lattice temperature as a consequence of electron–phonon (=quantized lattice vibration) coupling, i.e., the carriers can emit or absorb phonons. The coupling to longitudinal optical phonons in polar materials is especially strong with

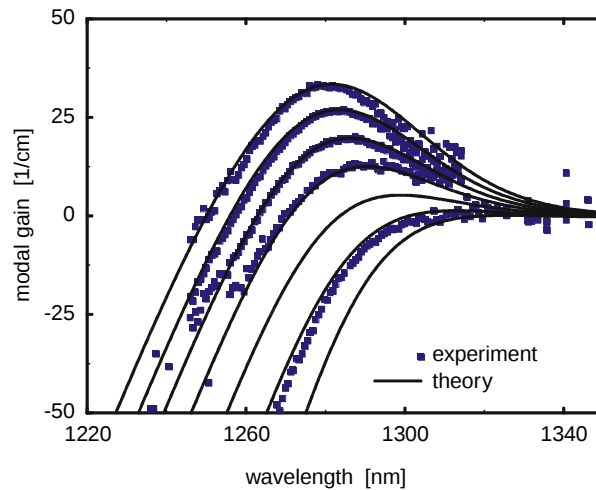


Fig. 4 Comparison of experimental and theoretical gain spectra of a (GaIn)(NAs)/GaAs laser.

scattering times in the picosecond range. The corresponding times for the electron-acoustic phonon scattering are in the nano-second range. Generally, the carrier-phonon interaction provides an exchange of kinetic energy of the carriers with the crystal lattice in addition to the relaxation of the initial nonequilibrium distribution. Deviations between plasma and lattice temperature may occur, especially when lasers are subject to strong pumping or rapid dynamic changes.

In addition to leading to rapid carrier scattering the Coulomb interaction directly influences the laser gain spectrum. Gain, i.e., negative absorption, occurs in the spectral region where we have population inversion, in other words, where the net probability for a test photon to be absorbed is lower than the probability to be amplified by stimulated electron-hole recombination. Energetically, the gain region is bounded from below by the effective (or renormalized) semiconductor bandgap and from above by the electron-hole chemical potential. Here, the chemical potential is the energy at which neither absorption nor amplification occurs, i.e., where the semiconductor medium is effectively transparent.

The effective bandgap energy for elevated carrier densities is significantly below the fundamental absorption edge of an unexcited semiconductor. This “bandgap renormalization” is a consequence of the fact that the Pauli exclusion principle reduces the repulsive Coulomb interaction between carriers of equal charge. Furthermore, the Coulomb screening, i.e., the density dependent weakening of the effective Coulomb interaction potential contributes to the reduction of the bandgap.

Another many-body effect originates in the Coulomb attraction between an electron and a hole, and leads to an excitonic or Coulomb enhancement of the interband transition probability. All these effects contribute to the detailed characteristics of the optical spectra of a semiconductor gain medium. **Fig. 4** shows an example of calculated spectra.

The example demonstrates that the microscopic theory correctly accounts for the changes in the gain spectrum with changing carrier density by means of a consistent treatment of band structure, band filling, and many-body Coulomb effects.

The Resonator

Besides a gain medium, a laser needs feedback for the emitted photons in order to allow for the positive feedback process required for amplified-stimulated emission. Hence, the inverted medium has to be embedded into a resonator. One of the conceptually simplest versions is a so-called Fabry-Perot resonator, which consists of two parallel plane mirrors. A double-heterostructure Fabry-Perot laser is shown in **Fig. 2**. For semiconductor lasers, such Fabry-Perot resonators, are very easy to fabricate: by cleaving the semiconductor crystal perpendicular to the optical waveguide. The cleaved edges provide a reflectivity of about 32% due to the high refractive index of the semiconductor. This is sufficient to provide enough feedback for laser operation.

However, despite its conceptual simplicity, the cleaved edge Fabry-Perot concept also introduces a number of problems. First, the light output occurs on two sides of the laser, whereas for most applications all the output power is desired to be concentrated from one facet only. For practical applications, this problem can be solved by depositing coatings onto the cleaved facets of the structure. For example, deposition of a 10% reflectivity coating on one facet and of a 90% reflectivity coating on the other facet concentrates the laser emission almost completely to the facet of lower reflectivity. Antireflection coatings with reflectivities below 10^{-4} are used when laser diodes are coupled to external cavities, for example, in tunable lasers.

A second problem with Fabry-Perot lasers is that the emission wavelength is often not well defined. A Fabry-Perot resonator of length L (optical length, nL) has transmission maxima – so-called modes – at all multiples of $c/2nL$, where c is the speed of light. That means that for typical devices with a length of 500 μm , hundreds of modes are present even within the limited gain bandwidth of a semiconductor. These modes often compete with each other resulting in multimode emission, which may become particularly complex and uncontrollable when the laser is modulated for high-speed operation.

This problem can be overcome in so-called distributed feedback (DFB) lasers. In these devices, the refractive index is periodically modulated along the waveguide. In a simplified picture, this periodic pattern causes multiple reflections at different locations within the waveguide. For certain wavelengths, these multiple reflections interfere constructively to provide enough “distributed” feedback for laser operation. With the DFB concept single-mode emission can be realized even for high-speed modulation of the lasers. Alternatively, the region of DFB can be separated from the region of amplification in so-called distributed Bragg reflector (DBR) lasers. These are devices with multiple sections where one section is responsible for the optical gain and another section (the DBR-section) is responsible for the optical feedback.

So far, we have discussed so-called edge-emitting lasers, which means that the light emission is in the plane of the active area of the device. The lengths of these lasers are typically hundreds of microns and thus correspond to a few hundreds optical wavelengths. This causes one of the major disadvantages of edge emitters the multimode emission. Another disadvantage is the poor elliptic beam profile due to diffraction at the small rectangular output aperture. Finally, the fabrication procedure of edge emitters is complex since it requires multiple steps of cleaving before it is even possible to test the laser.

An alternative semiconductor laser concept was realized in the early 1990s: the vertical-cavity surface-emitting laser (VCSEL). A typical VCSEL structure is shown in Fig. 5. In VCSELs, the light emission is parallel to the growth direction, i.e., perpendicular to the epitaxial layers of the structure. The cavity length of a VCSEL is in the order of a few (1–5) multiples of half the wavelength of the emitted light. These small dimensions imply that the interaction length between active medium and light field in the resonator is very short. Therefore, the mirrors of a VCSEL have to be of very good quality: reflectivities of more than 99% are required in most configurations.

Very high reflectivities can be achieved with dielectric multilayer structures, so-called Bragg reflectors. Bragg reflectors consist of a series of pairs of layers of different refractive index, where each layer has the thickness of a quarter of the emission wavelength. Most preferably, the Bragg reflectors are epitaxially grown together with the active area of the device in only one growth process.

Further technological challenges arise from the need of electrical pumping of the VCSEL. Thus a p–n junction has to be implemented and sufficient strategies to effectively inject charge carriers into the active region have to be developed. Consequently practical electrically pumped VCSEL structures have an even more complex architecture than the schematic structure shown in Fig. 5 (Michalzik, 2013).

The small VCSEL resonators with Bragg reflectors at both ends of the cavity are called microcavities. They are designed such that only one longitudinal mode is present in the region of the semiconductor gain spectrum. Accordingly, VCSELs are well suited to provide single-mode characteristics. Moreover, the aperture for light emission can be chosen round with a diameter of a few micrometers so that diffraction is weak and the beam profile is almost perfect.

VCSELs can be tested on wafer without further processing and can be arranged in 2D arrays. A major disadvantage of VCSELs is their temperature dependence. The bandgap of the active material and the cavity mode vary differently with temperature so that the cavity mode shifts away from the region of maximum gain when the temperature is varied. This severely limits the temperature range of operation for VCSELs.

In addition to VCSELs, current semiconductor laser development also focusses on *vertical external cavity surface emitting lasers* (VECSELs) also known as *semiconductor disk lasers*, schematically shown in Fig. 6. Here, the top Bragg mirror of the VECSEL is replaced by an external mirror (see Fig. 5) in order to obtain an extended external cavity. This novel class of microlasers is particularly attractive for several reasons. First of all, a wide variety of semiconductor materials can be used as the gain medium allowing for the realization of a broad range of emission wavelengths. Moreover, the external cavity allows for the inclusion of nonlinear elements. For example, by inserting a properly designed nonlinear crystal, one can reach new frequencies via intracavity frequency doubling (Kuznetsov *et al.*, 1999; Keller and Tropper, 2006).

As in VECSELs, also in VECSELs round-trip gain is provided by relatively few QWs such that the total material gain is rather small in comparison to a typical edge-emitting semiconductor laser. Therefore, the gain medium, i.e., the active mirror in Fig. 6 and the lasing mode have to be optimally aligned under the desired VECSEL operational conditions.

One of the intriguing challenges in the VECSEL design is that one typically does not know a priori what the internal operation conditions will be. In particular, the effective temperature of the lasing medium is unknown since the QWs suffer from significant

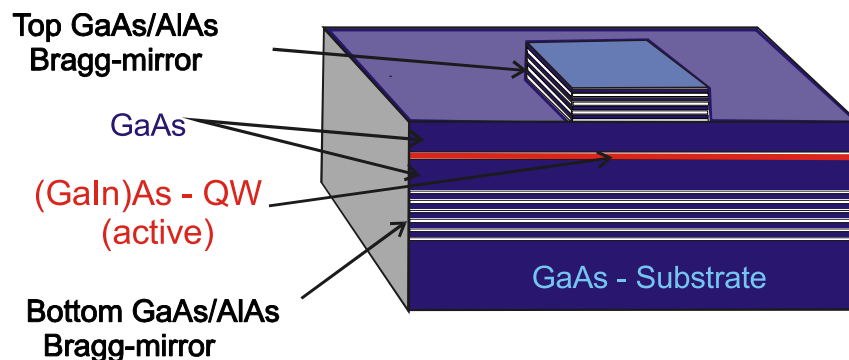


Fig. 5 Vertical-cavity surface-emitting laser (VCSEL).

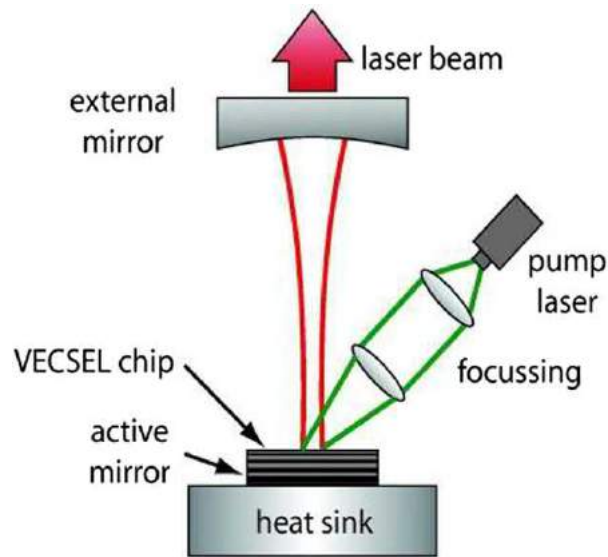


Fig. 6 Vertical external cavity surface emitting lasers (VECSELs) configuration showing the VECSEL chip (active mirror) consisting of a multiple quantum-well (QW) structure grown on a distributed Bragg reflector (DBR). The structure is excited with a pump laser which is focused on the QWs. Lasing in the VECSEL system is achieved by using an external high reflectivity mirror (output coupler) to form a closed cavity. The generated heat is removed by a heat sink typically consisting of a diamond heat-spreader on a copper block.

heating effects. These thermal problems result both from the fundamental impossibility to fully convert the optical pump power into lasing output and from the nonradiative carrier recombination processes. In particular, the intrinsic and thus unavoidable Auger losses lead to a significant QW heating. As countermeasure, one includes heat dissipating elements in the VECSEL structure. However, as we can see in Fig. 6, this heat sink typically provides only remote cooling of the QWs. Thus the effective gain medium temperature is determined by the detailed balance between heat generation and dissipation. To make matters worse, this temperature will change with the operating conditions.

To account for all these effects in the VECSEL design, detailed microscopic modeling is needed. A combined effort of modeling and dedicated growth of VECSEL structures has recently lead to the generation of record output powers in the regime of more than 100 W CW emission (Wang *et al.*, 2012).

Semiconductor Laser Dynamics

The dynamical behavior of semiconductor lasers is rather complex because of the coupled electron–hole and light dynamics in the active material. Like other laser materials, semiconductor lasers exhibit a typical threshold behavior when the pump intensity is increased. For weak pumping, the emission is low and spectrally broad: the device operates below threshold and emits (amplified) spontaneous emission. At threshold, the optical gain compensates the losses in the resonator (i.e., internal losses and mirror losses) and lasing action starts. In the optimum case the output power increases linearly with the injection current above threshold. The derivative of this curve is called differential quantum efficiency. The differential quantum efficiency can reach values close to unity and even the overall efficiency of the diode laser, i.e., the conversion rate from electrical power into optical power can be as high as 60%. Ideally, the laser emission is single mode above threshold. However, as discussed above, edge-emitting diode lasers tend to emit on multiple longitudinal modes and the competition of these modes can lead to a complex dynamical behavior, which is drastically enhanced as soon as the laser receives small amounts of optical feedback. Semiconductor lasers are extremely sensitive to feedback and tend to show dynamically unstable, often even chaotic behavior under certain feedback conditions.

As other lasers, also semiconductor lasers respond with relaxation oscillations when they are suddenly switched on or disturbed in stationary operation. Relaxation oscillations, i.e., the damped periodic response of the total carrier density and the output intensity as a dynamically coupled system, are important for many aspects of the semiconductor laser dynamics. The relaxation oscillation frequency determines the maximum modulation frequency of diode lasers. This frequency is strongly governed by the differential gain in the active material, i.e., by the variation of the optical gain with carrier density.

An important application of relaxation oscillations is short pulse generation by gain switching of laser diodes. The idea of gain switching is to inject a short and intense current pulse or optical pump pulse into the laser structure. This pulse initiates relaxation oscillations but it is so short that already the second oscillation maximum is not amplified such that only one intense pulse is emitted. The shortest pulse widths achieved with this scheme so far are in the few picoseconds range. For short pulse generation, a further complexity of the diode laser comes into play. The emitted pulses are observed to be strongly chirped, i.e., the center frequency changes during the pulse. The origin of this complex self phase modulation is the strong coupling of real and imaginary

part of the susceptibility in the semiconductor. When the gain (which is basically given by the imaginary part of the susceptibility) is changing during a pulse emission, the refractive index (basically the real part of the susceptibility) is also changing considerably. This leads to a self phase modulation when a strong optical pulse is emitted. This strong phase-amplitude coupling is also responsible for the high feedback sensitivity of diode lasers, for complex spatiotemporal dynamics (e.g., self focussing, filamentation), and for a considerable broadening of the emission linewidth. The phase-amplitude coupling is often quantified with the so-called linewidth enhancement factor α . This parameter, however, is a function of carrier density, photon wavelength, and temperature (Chow and Koch, 1999; Osinski and Buus, 1987).

Diode lasers and VECSELs are also well suited for the generation or amplification of extremely short pulses with durations far below 1 ps. The concept used for such extremely short pulse generation is called modelocking. The idea of modelocking is to synchronize the longitudinal modes of the laser such that their superposition is a sequence of short pulses. The most successful concept to achieve modelocking of diode lasers is to introduce a saturable absorber into the laser cavity and to additionally modulate the injection current synchronously to the cavity round-trip frequency (Delfyett *et al.*, 1994). A saturable absorber exhibits a nonlinear absorption characteristics: the absorption is high for weak intensities and weak for high intensities so that high peak power pulses are supported. Such absorbers can be integrated as a separate section into multiple-section diode lasers, which have already been successfully used for sub-picosecond pulse generation. Furthermore, in such multiple-section devices electro-absorption modulators, passive waveguide sections, and tunable DBR-segments for wavelength tuning can be integrated, too.

However, when sub-picosecond pulses are generated or amplified in semiconductor laser amplifiers the dynamics of the pulse generation or pulse amplification is strongly governed by complex nonequilibrium carrier dynamics in the semiconductor (Hofmann, 1999). The most prominent effects are spectral hole burning and carrier heating. Spectral hole burning occurs when an intense laser field removes carriers from a certain area in k -space faster than carrier-carrier scattering can compensate for. This leads to nonthermal carrier distributions and a gain suppression in certain energy regimes. As mentioned above, a nonthermal carrier distribution relaxes quickly (100 fs) into a thermal (quasi-equilibrium) distribution when the intense laser field is switched off but this carrier distribution may have a plasma temperature much higher than the lattice temperature. This effect which is referred to as carrier heating also leads to a transient reduction of the gain on a few picoseconds timescale until the carrier distributions have cooled down to the lattice temperature by interactions with phonons. These ultrafast effects are of course particularly important for amplification and generation of short pulses but they generally influence the whole dynamical behavior of diode lasers and VECSELs (Kilen *et al.*, 2016).

Applications

Semiconductor laser diodes are already an essential part of our daily life as they are key components in numerous important applications. To name a few prominent examples, laser diodes act as light sources in glass fiber transmission systems, which are the basis of the internet. They are used to read out information in CD, DVD, and Blue ray DVD systems. Further, well-established applications include bar code scanners, laser pointers, laser printers, or the optical computer mouse. The recent development of high power laser diodes also enables applications in material processing. The high flexibility of the laser diode materials in terms of emission wavelengths makes them also suitable for high-end scientific applications in spectroscopy, for example. The rapid development of laser diodes with new and improved specifications will continuously open further application fields as, for example, compact laser displays with high brilliance making use of the recently established availability of efficient red, blue, and green laser diodes.

References

- Bimberg, D., Grundmann, M., Ledentsov, N.N., 1998. Quantum Dot Heterostructures. New York, NY: Wiley.
- Chow, W.W., Koch, S.W., 1999. Semiconductor Laser Fundamentals: Physics of the Gain Materials. Berlin: Springer-Verlag.
- Chuang, S.L., 1995. Physics of Optoelectronic Devices. New York, NY: Wiley.
- Delfyett, P.J., Dienes, A., Heritage, J.P., Hong, M.Y., Chang, Y.H., 1994. Femtosecond hybrid mode-locked semiconductor laser and amplifier dynamics. *Applied Physics B* 58, 183–195.
- Haug, H., Koch, S.W., 2009. Quantum Theory of the Optical and Electronic Properties of Semiconductors, fifth ed. Singapore: World Scientific Publisher.
- Hofmann, M., 1999. Gain and emission dynamics of semiconductor lasers. *Recent Research Developments in Applied Physics* 2, 269–290.
- Keller, U., Tropper, A.C., 2006. Passively modelocked surface-emitting semiconductor lasers. *Physics Reports* 429, 67.
- Kilen, I., Koch, S.W., Hader, J., Moloney, J.V., 2016. Fully microscopic modeling of mode locking in microcavity lasers. *Journal of the Optical Society of America B* 33, 75.
- Kuznetsov, M., Hakimi, F., Sprague, S., Moodardian, A., 1999. Design and characteristics of high-power (> 0.5-W CW) diode-pumped vertical-external-cavity surface-emitting semiconductor lasers with circular TEM00 beams. *IEEE Journal of Selected Topics in Quantum Electronics* 5, 561.
- Michalzick, R., 2013. VCSELs Fundamentals, Technology and Applications of Vertical-Cavity Surface-Emitting Lasers. Berlin: Springer-Verlag.
- Osinski, M., Buus, J., 1987. Linewidth broadening factor in semiconductor lasers – An overview. *IEEE Journal of Selected Topics in Quantum Electronics* 23, 9–29.
- Wang, T.L., Heinen, B., Hader, J., *et al.*, 2012. Quantum design strategy pushes high-power vertical-external-cavity surface-emitting lasers beyond 100 W. *Laser & Photonics Reviews* 6, L12–L14.

Further Reading

- Koch, S.W., Jahnke, F., Chow, W.W., 1995. Physics of semiconductor microcavity lasers, review article. *Semiconductor Science and Technology* 10, 739–751.



Encyclopedia of Modern Optics

Second Edition

Editors in Chief

Bob Guenther
Duncan Steel

**ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION**

ENCYCLOPEDIA OF MODERN OPTICS SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 3

Displays ■ Biomedical: Tissue Measurements ■ Imaging (Non-Medical)



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 3

Editorial Board	v
List of Contributors for Volume 3	ix
Editors in Chief	xi
Introduction	xiii

VOLUME 3

Displays – Introduction	<i>Han-Ping D Shieh</i>	1
Liquid Crystal Physics and Materials	<i>Huang-Ming Philip</i>	8
TFT Materials and Devices	<i>Po-Tsun Lin</i>	12
Color Formation of LCD – Spatial and Temporal	<i>Fang-Cheng Lin</i>	17
LCD Components	<i>Fang-Cheng Lin</i>	25
Projection Displays – LCD/MEMS/LCoS Based	<i>Fleming Chuang</i>	32
3D Displays, Stereoscopic/Autostereoscopic 3D	<i>Yi-Pai Huang and Chun-Ho Chen</i>	44
Wearable Displays	<i>Paul Yang</i>	51
View Angles of Liquid Crystal Displays	<i>Huang-Ming Philip Chen</i>	59
Organic Light-Emitting Diode (OLED)	<i>Chin H (Fred) Chen, Wen-Shi Wen, and Chin-Ti Chen</i>	64
Optical Characteristics of Display Devices	<i>Han-Ping D Shieh</i>	70
Reflective Display Technologies	<i>Han-Ping D Shieh</i>	79
In Situ Optical Tissue Diagnostics/Miniaturized Optoelectronic Sensors for Tissue Diagnostics	<i>Seung Yup Lee, Yooree Grace Chung, and Mary-Ann Mycek</i>	86
In Situ Optical Tissue Diagnostics/Laser Speckle-Based Spectroscopy Techniques in Biomedicine	<i>Karthik Vishwanath and Sara Zanfardino</i>	95
Photodynamic Therapy and Photobiomodulation: Can All Diseases be Treated with Light?	<i>Michael R Hamblin</i>	100
Resolution and Multiple Scattering in Imaging	<i>Edwin A Marengo, Zambrano-Nunez Maytee, and Edson S Galagarza</i>	136
Coherent Diffractive Imaging	<i>Lei Tian</i>	146
Nonuniqueness in Imaging	<i>Greg Gbur</i>	156
Phase Space and Imaging	<i>Markus Testorf</i>	164
Information Theory in Imaging	<i>FO Huck and CL Fales</i>	175
Inverse Problems and Computational Imaging	<i>M Bertero and P Boccacci</i>	187
Adaptive Optics	<i>C. Pernechele</i>	197
Phase Conjugation and Image Correction	<i>EN Leith</i>	205
Hyperspectral Imaging	<i>ML Huebschman, RA Schultz, and HR Garner</i>	211
Imaging Through Scattering Media	<i>AC Boccara</i>	219

Infrared Imaging	<i>K Krapels and RG Driggers</i>	229
Interferometric Imaging	<i>DL Marks</i>	241
Multiplex Imaging	<i>A Lacourt</i>	247
Photon Density Wave Imaging	<i>V Toronov</i>	256
Three-Dimensional Field Transformations	<i>R Piestun</i>	262
Volume Holographic Imaging	<i>C Barbastathis</i>	267
Wavefront Sensors and Control (Imaging Through Turbulence)	<i>CL Matson</i>	272

LIST OF CONTRIBUTORS FOR VOLUME 3

G Barbastathis
Massachusetts Institute of Technology, Cambridge, MA, USA

M Bertero
University of Genova, Genova, Italy

P Boccacci
University of Genova, Genova, Italy

AC Boccara
Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

Chin H (Fred) Chen
Shine Materials Technology Co., Ltd., Kaohsiung, Taiwan, R.O.C.

Chin-Ti Chen
Chemistry Institute, Academia Sinica, Taipei, Taiwan, R.O.C.

Chun-Ho Chen
National Chiao Tung University, Hsinchu, Taiwan

Fleming Chuang
Coretronic Corporation, Taiwan, Republic of China

Yooree Grace Chung
University of Michigan, Ann Arbor, MI, United States

RG Driggers
US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

CL Fales
NASA Langley Research Center, Hampton, VA, USA

Edson S Galagarza
Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

HR Garner
University of Texas, Dallas, TX, USA

Greg Gbur
University of North Carolina at Charlotte, Charlotte, NC, United States

Michael R Hamblin
Massachusetts General Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States; and Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, United States

Yi-Pai Huang
National Chiao Tung University, Hsinchu, Taiwan

FO Huck
NASA Langley Research Center, Hampton, VA, USA

ML Huebschman
University of Texas, Dallas, TX, USA

K Krapels
US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

A Lacourt
Université de Franche-Comté, Besancon, France

Seung Yup Lee
Georgia Institute of Technology/Emory University, Atlanta, GA, United States

EN Leith
University of Michigan, Ann Arbor, MI, USA

Fang-Cheng Lin
National Chiao Tung University, Hsinchu, Taiwan

Po-Tsun Liu
National Chiao Tung University, Hsinchu, Taiwan

Edwin A Marengo
Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

DL Marks
University of Illinois at Urbana-Champaign, Urbana, IL, USA

CL Matson
Air Force Research Laboratory, Kirtland, NM, USA

Zambrano-Nunez Maytee
Technological University of Panama, Panama City, Panama

Mary-Ann Mycek
University of Michigan, Ann Arbor, MI, United States

C Pernechele
Astronomical Observatory of Padova, Padova, Italy

Huang-Ming Philip
National Chiao Tung University, Hsinchu, Taiwan

Huang-Ming Philip Chen
National Chiao Tung University, Hsinchu, Taiwan

R Piestun

University of Colorado, Boulder, CO, USA

RA Schultz

University of Texas, Dallas, TX, USA

Han-Ping D Shieh

National Chiao Tung University, Hsinchu, Taiwan

Markus Testorf

Thayer School of Engineering, Hanover, NH, United States

Lei Tian

Boston University, Boston, United States

V Toronov

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Karthik Vishwanath

Miami University, Oxford, OH, United States

Wen-Shi Wen

Shine Materials Technology Co., Ltd., Kaohsiung, Taiwan, R.O.C.

Paul Yang

Sun Yat-Sen University, Guangzhou, China

Sara Zanfardino

Miami University, Oxford, OH, United States

EDITORS IN CHIEF



Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.



Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Displays – Introduction

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

ACR	Ambient contrast ratio	LC	Laser crystallization
a-IGZO	Amorphous In-Ga-Zn-O	LCD	Liquid crystal display
ALD	Atomic layer deposition	LCOS	Liquid crystal on silicon
AOM	Acoustic optical modulator	LUMO	Lowest unoccupied molecular orbital
AR	Augmented reality	LPD	Local-primary-desaturation
BCE	Back channel etched	MILC	Metal-induced lateral crystallization
BEF	Brightness enhancement film	MIP	Memory-in-pixel
BRDF	Bidirectional reflectance distribution function	OCB	Optically compensated bend
CCFLs	Cold cathode fluorescence lamps	OLED	Organic light emitting diode
CCT	Correlated color temperature	OTFTs	Organic TFTs
CGL	Charge-generation layer	PEN	Polyethylene naphthalate
ChLCD	Cholesteric liquid crystal display	PECVD	Plasma-enhanced chemical vapor deposition
CIE	Commission International de l'Eclairage	PLD	Pulse laser deposition
CMFs	Color matching functions	PPI	Pixels per inch
CMYK	Cyan, magenta, yellow, black	PMMA	Polymethylmethacrylate
CRT	Cathode ray tube	PWM	Pulse width modulation
DBEF	Double brightness enhancement film	PVP	Poly-vinylphenyle
DLP	Digital light processing	PVA	Polyvinyl alcohol
DMD	Digital mirror display	QD	Quantum-dot
EIL	Electron injection layer	QR-LPD	Quick response liquid powder display
EL	Electroluminescence	R2R	Roll-to-roll
EPD	Electrophoretic display	SLM	Spatial light modulator
ESR	Enhanced specular reflector	SmC*	Chiral-smectic C
EWD	Electro-wetting display	SOI	Silicon-on-insulator
FOLED	Flexible OLED	SOP	System-on-panel
FoV	Field of view	SPC	Solid phase crystallization
FPD	Flat panel display	SRAM	Static random access memory
FSC	Field sequential color	TAC	Triacetyl cellulose
HIL	Hole injection layer	TFT	Thin film transistor
HOMO	Highest occupied molecular orbital	TTA	Triplet-triplet annihilation
IES	Internal enhancement layer	VA	Vertical alignment
IMOD	Interferometric modulator display	VFD	Vacuum fluorescent display
IPS	In-plane switching	VR	Virtual reality

A display can be defined as an object that transfers information to the viewer and/or an output unit that gives a visual representation of data. Displays can be divided into two types: emissive and non-emissive. By emissive we mean that the display surface gives off light. In the non-emissive, the display or a small portion of the display acts as a shutter or spatial filter which also can be reflective in nature. In the non-emissive mode, the display can pass light from a source in back of or in front of the shutter.

The dominant displays in the 20th century were Cathode Ray Tube (CRT). Since then, a non-emissive display – active matrix Thin Film Transistor (TFT) Liquid Crystal Displays (LCD), and potentially an emissive one-Active matrix OLED are dominated. Virtually all other types/forms of displays only have very niche applications or just disappeared due to cost-performance, form factor, health concerns, and others.

Emissive displays include the CRT, projection CRT, vacuum fluorescent display (VFD), field emissive display (FED), electroluminescent (EL) display, light emitting diode (LED), organic light emitting diode (OLED), plasma, Thin Film EL (TFEL) Displays, etc.

Most of emissive displays use certain types of emissive materials called phosphors. The phosphors are called cathodoluminescent, photoluminescent, and electroluminescent, depending whether the excitation action was due to the interaction of the phosphor with cathode rays, photons, or low-voltage electrons. Cathodoluminescent phosphors are used in CRTs – projection CRTs, VFDs, and FEDs. EL displays can use inorganic phosphors or organic phosphors such as those used in OLEDs. Under photoluminescence displays, we have “plasma displays,” where the excitation of the plasma in a localized region gives rise to UV photons which strike the phosphor screen in that region.

Under the heading of “shutter-type” we can consider LCDs that consist of AMLCDs, super twisted nematic liquid crystal displays (STN LCDs) and bistable LCDs. Another shutter-type is the liquid crystal on silicon (LCOS) display and the movable digital mirror display (DMD), and the digital light processing (DLP) projecting engines. Another spatial display type is both normal and dynamic inks. A dynamic ink can be a “polarizable” ink display in which a two shade (light/dark) ink sphere element can rotate under the presence of an electric field to show the light or dark side. The DMD movable mirrors reflect the digital information to the lens system and the unwanted rays are reflected away from the lens to a black light capture area.

In the 21st century, virtually all displays are in the form of “flat,” a very thin in thickness, a typical display system includes display panel and associated electronics to drive the display panel, timing controller, Gamma reference for color correction, signal conversion, and others as shown in Fig. 1. Within flat panel displays (FPD), their characteristics can be divided by luminance, driving method, display driving panel materials, processing temperature, non-emissive (liquid crystal), and emissive materials, as shown in Fig. 2.

Cathode Ray Tube

The CRT is a display that has been in use for over 100 years (the Braun CRT tube was developed in 1898). The CRT uses a phosphor screen which is bombarded by energetic electrons and emits visible light radiation. The single electron gun CRT was used for years in early monochrome television sets and computer displays. Monochrome green, amber, and white phosphors have been used in monitor tubes. CRT is not a flat panel display, due to the thickness, weight, difficult to scale up in size and most serious one of up to tens of hundreds driving voltage, CRTs have been totally displaced by FPD since the beginning of the 21st century.

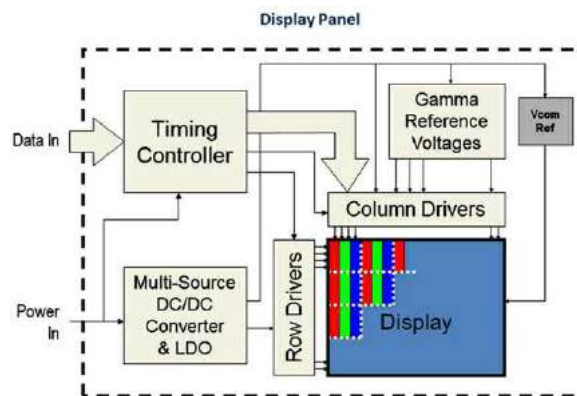


Fig. 1 A display system.

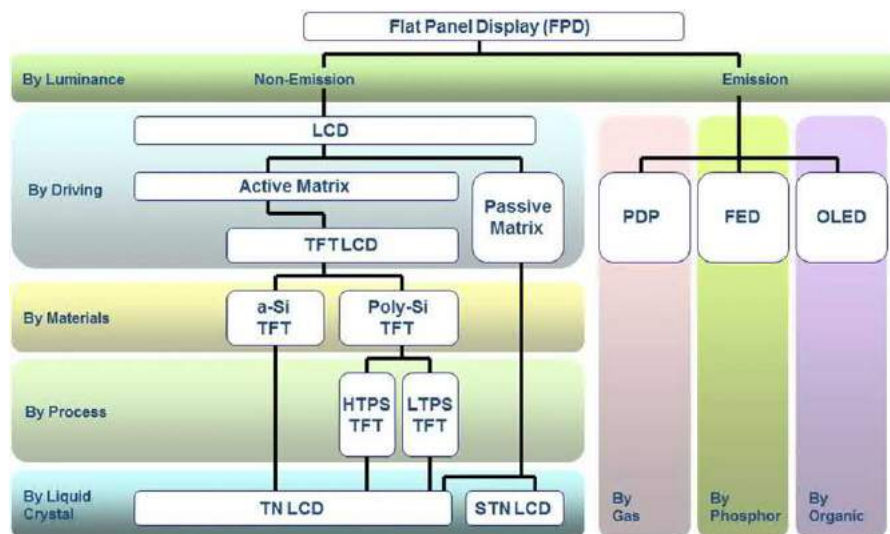


Fig. 2 Flat panel display (FPD) family and their characteristics.

Color CRT

The more ubiquitous method of obtaining a color display is the device called the shadow mask CRT (and the similar Trinitron CRT). This color CRT is designed in such a manner that through the use of an aperture mask, (seen in Fig. 3) the electrons from the red electron gun strike the red phosphor and likewise electrons from the green and blue electron guns strike the green and blue phosphors.

The combination allows the viewer to see a color image. This color separation action is seen in Fig. 4. Although Fig. 4 shows a delta tri-color design, there are also in-line systems in which the mask consists of an array of slots which are somewhat rectangular in nature. In this array, many slots can be on a line, or a mask made can be made of a series of wires as used in the Trinitron design.

The disadvantage of the shadow mask technique is that the mask can capture as much as 80% of the electrons from the electron guns. This yields lower screen luminance and lower contrast due to scattered electrons. As the screen size gets larger from 19 to 36 in. or more, additional temperature compensation methods must be used to reduce the effects of mask heating. Another method is using mask materials with lower coefficients of expansion such as INVAR. The phosphors used in the shadow mask color CRT are the red $Y_2O_3:S:Eu$ or, in some cases, the $Y_2O_3:Euphosphor$. The green phosphor is $ZnS:Cu$ and the blue phosphor is $ZnS:Ag$. Scale-up to size up to 40 in. as driving voltage needed in over 20,000 V. Shielding of high energy radiation also makes large size CRT very bulky and heavy in weight.

Vacuum Fluorescent Display

The VFD can be thought of as a low-voltage (12–40 V) CRT. Just as in the CRT, electrons from a cathode are controlled by a grid which allows them to be accelerated and strike the phosphor screen anode. In the operation of the VFD, electrons are emitted by the filament. The filament is a very thin tungsten wire coated with a tri-carbonate oxide (barium, strontium, and calcium

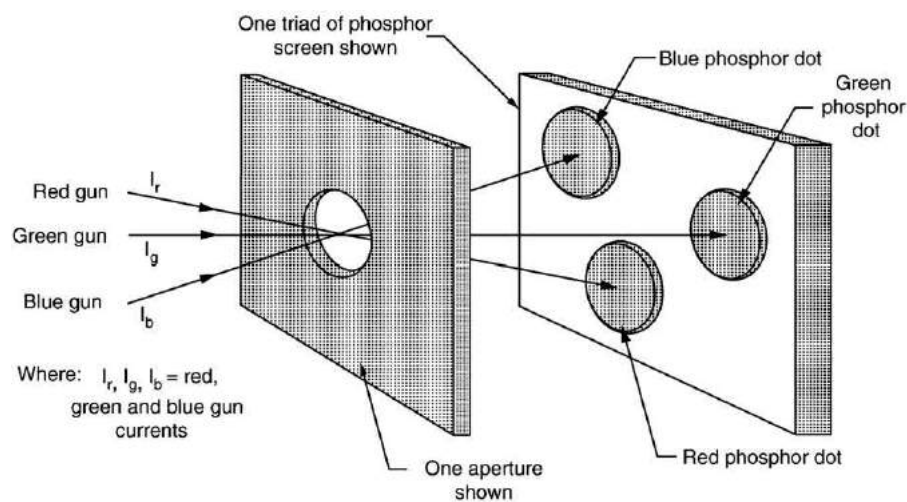


Fig. 3 The action of a shadow mask.



Fig. 4 A typical HDTV color CRT. Reproduced from Donofrio, R.L., 2002. The future of direct-view CRTs. SID Information Display Magazine 6, 10. Permission for reprint courtesy of the Society for Information Display.

carbonates) that are the same constituents of the CRT oxide cathode. By the application of a positive or negative voltage on this grid, electrons can either be cut off or accelerated to the phosphor on the anode. The VFD display is used in the auto industry and aging of the phosphor is very small, achieving 100,000 h to half of the initial luminance.

Field Emissive Displays

The FED is similar to the CRT and the VFD in that electrons from a source bombard a phosphor screen. In both the CRT and the VFD thermionic electrons are used. However, in the FED the electrons arise from field emission. Field emission devices create electrons through quantum mechanical tunneling from the emitter surface into the vacuum by the action of a large electrical field. This process is governed by the Fowler–Nordheim equation. The emitters can be of the Molybdenum Spindt type and also the poly-diamond or diamond-like carbon (DLC) coated emitters. We have the field enhancement with a sharp tip and reduced work function with the DLC. There has been renewed interest in FEDs, with the development of the “carbon nanotube” as an electron field emitting source. The carbon nanotube is a fullerene material, which is a distinct form of carbon. Fig. 5 shows that for this basic FED structure, the electrons from the Spindt emitter are controlled by the voltages of the anode, control electrode and emitter electrode.

Plasma Displays

Plasma displays used for television and information signs are photoluminescent displays. As shown in Fig. 6, the excitation of the plasma gas made of elements such as He, Ne, Ar, and Xe give rise to a glow discharge and the creation of vacuum ultraviolet photons that strike the screen and cause it to luminesce. In construction, the display is similar to the FED in that it is a matrix

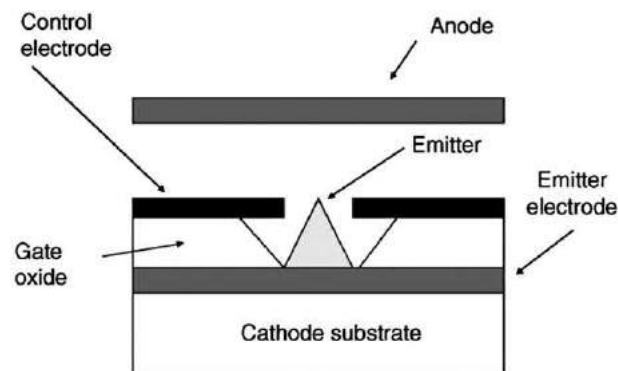


Fig. 5 Filed emission display cross-section. Reproduced from Green, P., Haven, D., 1998. Fundamentals of emissive displays. SID Short Course S-3. Permission for reprint courtesy of the Society for Information Display.

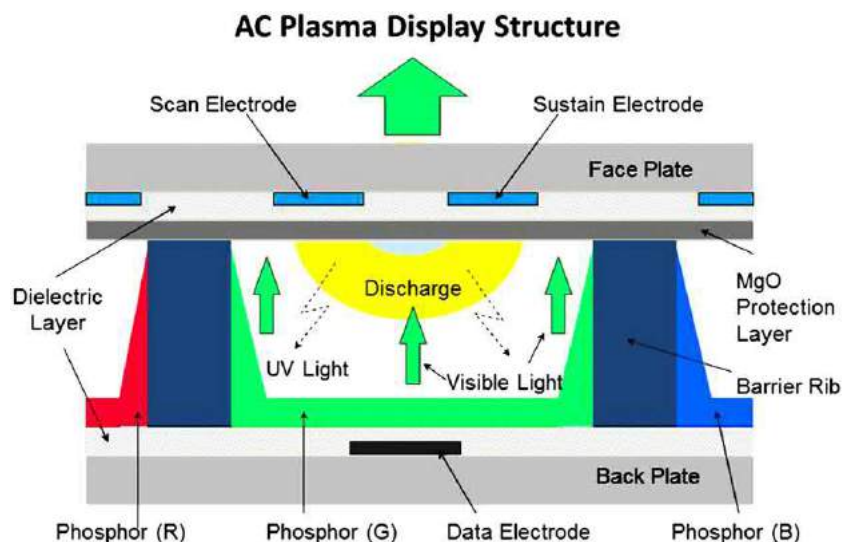


Fig. 6 A cross sectional view of typical AL plasma display.

addressed device. The AC design appears to be the more successful. Typical PDP phosphors for the triad are $(Y,Gd)BO_3:Eu$ for the red, $Zn_2SiO_4:Mn$ for the green, and $BaMgAl_{10}O_{17}:Eu$ (called BAM for short) for the blue. Fig. 6 below shows the basic structure of an AC plasma display. One phosphor element is shown with barrier ribs, address and scan electrodes. The discharge of the plasma takes place in the region of the phosphor and excites the photoluminescent phosphor, which is shaped (in this example) along the side of the barrier ribs and over the address electrode.

Electroluminescence

Electroluminescence is the generation of (non-thermal) light by use of an electric field. High electric field EL units emit light through the impact of high energy electrons in luminescent centers in a semi-conducting material.

Inorganic LED Displays

Light-emitting diodes are being used as backlights for LCD displays, replacing some of the fluorescent light backlighting methods. Using a matrix array row and column connectors, we can digitally select a given color LED and pattern in an LED array and generate a color image. Red, green, and blue-UV/blue LEDs are now available, with an increasing demand for white LEDs.

Organic LEDs

The basic process of light generation in OLEDs is the formation of a P-N junction. Electrons injected from the N-type material and holes from the P-type material recombine at the junction and light is generated, as shown in Fig. 7. Additionally, OLEDs can be made on flexible (F) substrates and are called FOLEDs. The use of a transparent (T) cathode gives rise to a top emission OLED or a TOLED.

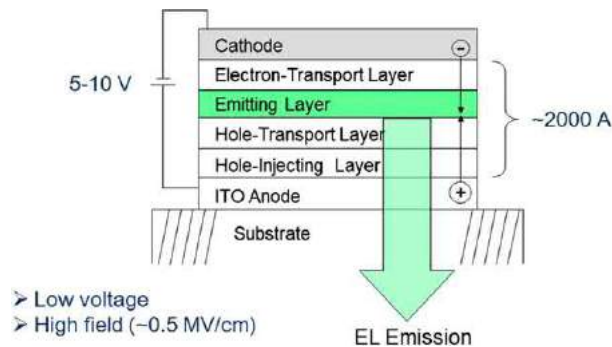


Fig. 7 A typical OLED configuration.

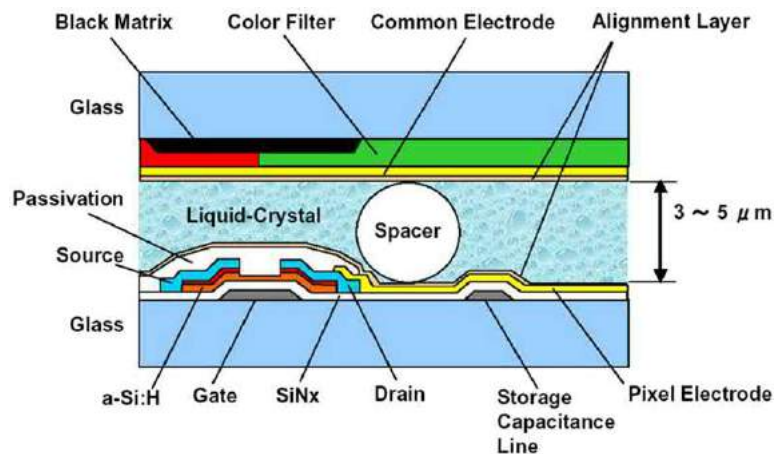


Fig. 8 A typical cross view of a TFT LCD panel.

Non-Emissive Displays – Active Matrix LCD – TFT

The active matrix liquid crystal display (AMLCD) device consists of a series of liquid crystal cells or windows whose characteristics can be controlled by an applied voltage to pass or restrict the passage of light through the windows via the external polarizers and the polarizing action of the LCD cells. AMLCD's are the dominated displays in the 21st century used almost in all area of mobile, laptop, monitors, TVs, and automotive displays (for driving information and entertainment). When the simple LCD is used without a transistor matrix, it is termed a passive display rather than an active one.

In order to show a color image, the LCD is designed to have three or more cells clustered per color pixel. The following (Fig. 8) is the transmission profiles of each of these red, green, and blue filters. The backlight used is a fluorescent lamp, but nowadays are white LED's or RGBW LED arrays.

Reflective LCD

For applications where there is suitable external light to view the display, no internal light source is used. Fig. 8 typical color LCD filter transmission characteristics. This reflective LCD consists of a polarizer and retarder plates stacked on the front side of the cell.

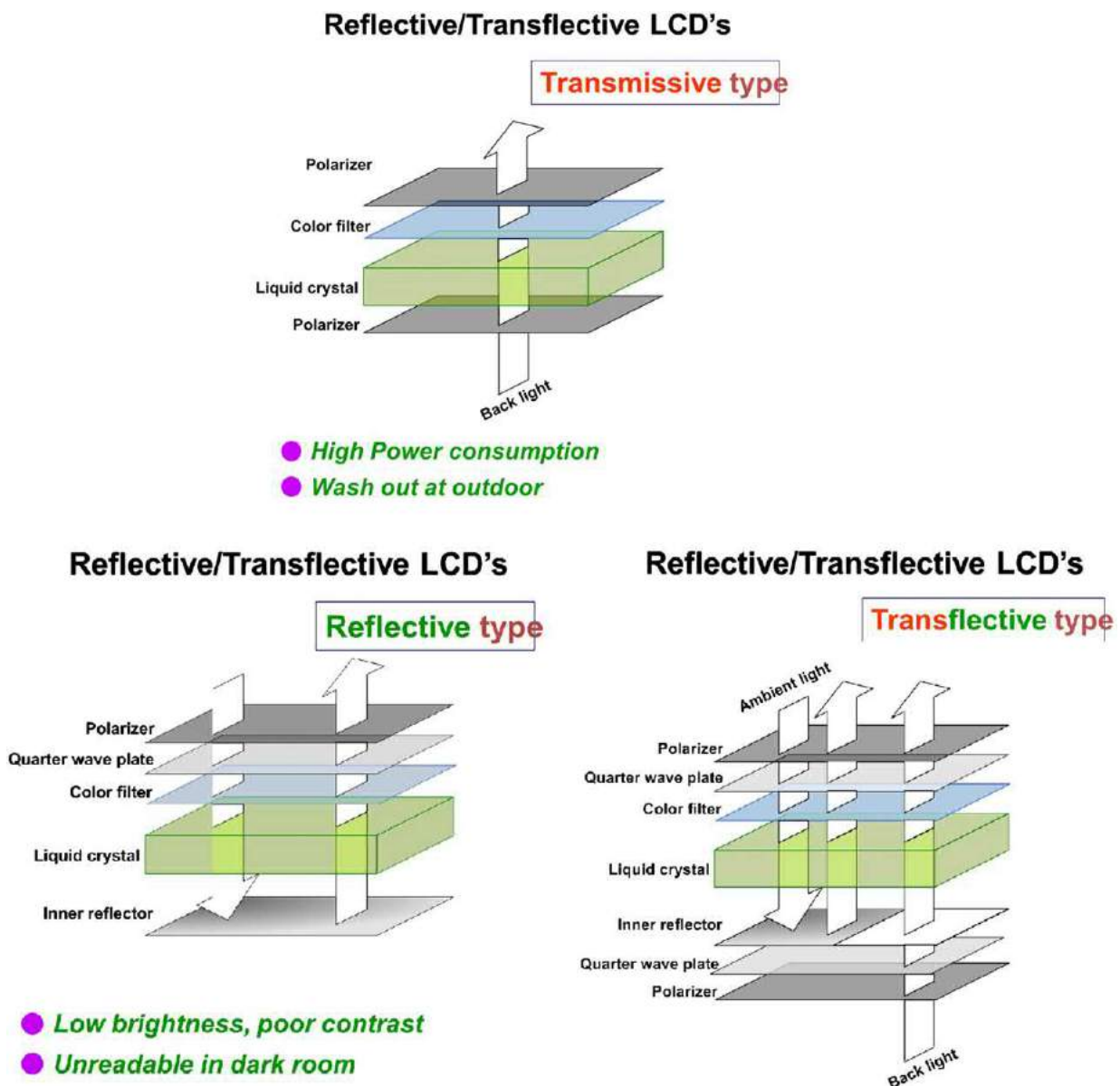


Fig. 9 Schematics of transmissive, reflective and transflective displays.

The cell with spacers contains the liquid crystal material and on the bottom of the cell is located a reflecting surface. It has been found that a dimpled reflecting surface makes a diffuse reflecting surface ([Fig. 9](#)).

Transflective LCD

The importance of the transflective device is that it has better contrast under high ambient illumination than the standard transmission back lighted LCD and fair contrast for low ambient illumination. The construction of this device consists of front and rear circular polarizers, back light, reflective elements, and novel filters. The back light is in the rear of the display. However, like the reflective LCD, internal reflectors are used on each pixel. The filters are designed spatially to be more transmissive for the reflective light than for the back light. Additionally the cell gap for the reflective light is smaller than the transmitted light region because the reflected light passes the cell twice whereas the transmitted light passes the cell once.

Further Reading

- de Gennes, P.G., Prost, J., 1993. *The Physics of Liquid Crystals*. Oxford: Clarendon Press.
- Fairchild, M.D., 2005. *Color Appearance Models*, second ed. John Wiley & Sons.
- Hecht, E., 1998. *Optics*, third ed. Addison Wesley Longman, Inc.
- Holliman, H., 2005. *3D Display System*.
- Lee, J.-H., Liu, D.N., Wu, S.-T., 2008. *Introduction to Flat Panel Displays*. 20. John Wiley & Sons.
- Shunsuke, Kobayashi, Mikoshiba, Shigeo, Lim, Sungkyoo (Eds.), 2009. *LCD Backlights* 21. John Wiley & Sons.
- Tsujimura, Takatoshi, 2017. *OLED Display Fundamentals and Applications*. John Wiley & Sons.
- Wu, S.T., Yang, D.K., 2001. *Reflective Liquid Crystal Displays*. SID Series in Displays Technology. Wiley and Sons Ltd.
- Yeh, Pochi, Gu, Claire, 2010. *Optics of Liquid Crystal Displays*. vol. 67. John Wiley & Sons.
- Yue Kuo, 2004. *Thin Film Transistors – Materials and Processes*. Volume 1: Amorphous Silicon Thin Film Transistors. Boston, MA: Kluwer Academic Publishers.

Liquid Crystal Physics and Materials

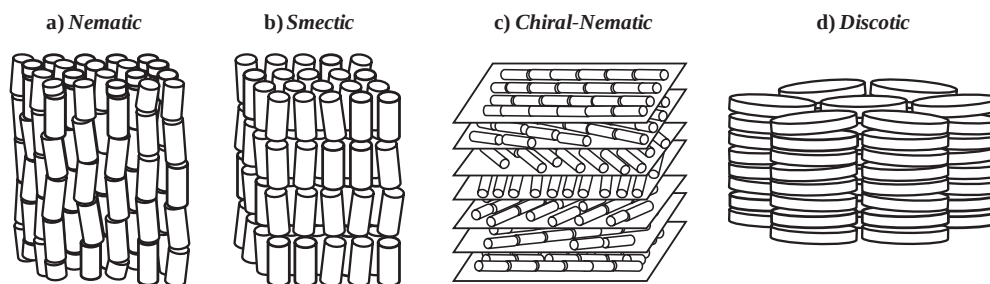
Huang-Ming Philip, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Liquid crystals (LCs) possess a state of matter that has properties in between isotropic liquid and solid crystal phases. They can flow like a liquid, but simultaneously hold crystal-like molecular orientations. LC phases can be induced by various factors: the temperature variations, the level of solvent concentration, and the ratios of organic-inorganic composition. Three different types of LC phases are most mentioned in the literature: thermotropic, lyotropic, and metallotropic LC phases. The thermotropic LC phases exhibit a liquid-crystal phase transition as the temperature fluctuates. The lyotropic phases, on the other hand, show phase transitions as a function of both temperature and solvent concentration by amphiphilic molecules. The third type, metallotropic LCs, consist of organic and inorganic hybrid molecular system, which retain an LC transition that is relied not only on temperature variation and solvent concentration, but also on the ratio of organic-inorganic composition.

Thermotropic Liquid Crystal Phases

Thermotropic LCs indicate their corresponding mesophases at a certain temperature range. Two distinct characters, monotropic and enantiotropic mesophases, can be found during the thermal treatment. The monotropic LC mesophase is observed when the mesophase is shown upon cooling from the isotropic liquids. In contrast, the enantiotropic LC mesophase is defined when the mesophase appears in both heating from crystalline solid and cooling from isotropic liquid (de Gennes and Prost, 1993). Many thermotropic LCs exhibit not merely a single phase, but a variety of phases as temperature changed. For instance, a rod-like LC molecule (also called mesogen) may exhibit different smectic phases followed by the nematic phase as temperature increased on heating process. Various LC phases can be distinguished by their optical properties, or X-ray crystallography. The distinct textures, or defect domains under polarized optical microscope, are usually the fingerprint to trace their respective phases. X-ray crystallography is a useful tool to probe into the details of different smectic or discotic liquid crystal phases. In nematic phase, the amorphous character, unfortunately, can be observed in the small angle X-ray spectrum. Four different phases of thermotropic LCs are listed as followed.



Nematic Phase

Nematic phase has been widely studied and applied among other LC phases. The word “nematic” originates from the Greek ($\nu\eta\mu\alpha$ (nema), which means thread-like textures observing from the topological defects under polarized optical microscope. It was previously called Schlieren defects or disclination lines. The dark threads are line singularities which connect either two- or four-fold brushes point defects. In particular, only nematics have the distinct two-fold brushes. In the nematic phase, the calamitic (or rod-like) molecules can be self-assembled into a long-range directional order with their molecular long axes roughly parallel to each other without any positional order. As a result, the molecules can flow like a liquid with center mass positions of molecules randomly distributed, yet still keep their long-range directional order. In general, the mesogen is a rod-like molecule consists of rigid and flexible components. The rigid part can be aromatic rings, cyclic rings, or π -bonding structures. The flexible component is mostly aliphatic chain with different carbon chain length. The polar functional groups, such as nitrile (cyano-functional group, -CN) or fluoro- (-F), isothiocyanate (-NCS), can be attached in different positions to alter dielectric constants of mesogens. The optical axis in most nematics is uniaxial, except that a few nematic LCs with special molecular design are biaxial nematics. The nematic LCs can be aligned by an external magnetic or electric field. The alignment layer provided uniaxial molecular alignment in current LCDs is made by polyimide. The order parameter (S) is designed to determine the quality of LC molecular alignment and is usually defined based on the average of the second Legendre polynomial: $S = 1/2 [3 \langle \cos^2 \theta \rangle - 1]$, where θ is the angle between the liquid crystal molecular axis and the local director (i.e. preferred direction in liquid crystal media). Specifically, $S=0$ represents a completely random and isotropic sample; and $S=1$ is a perfectly aligned sample. S is in the order of 0.3–0.8 in general liquid crystal samples. For current LCDs applications, S needs to be at 0.6–0.7. The order parameter can be measured experimentally in a

number of ways, such as diamagnetism, birefringence, UV-vis spectroscopy, Raman scattering, NMR and EPR, etc. The UV-vis or IR linear dichroism would be the most effective tool in general.

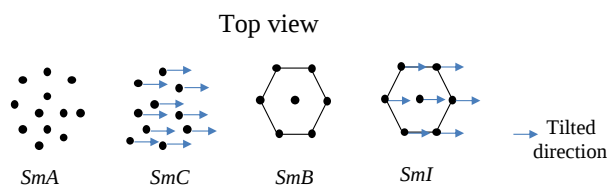
The theoretical treatments of LCs are mostly based on nematic LCs. Among them, the elastic continuum theory is a powerful method which is widely known for modeling LC devices under electric field. It discusses perturbations to a well-oriented continuum sample with strong boundary condition of parallel substrates. The distortions of the LC molecules in the continuum sample are described by the Frank free energy density. The response of the LC molecules can then be decomposed into three elastic constant terms, as the twist, splay, and bend distortions. The Maier-Saupe mean field theory provides a simple fluid nematic-isotropic phase transition model. Other models, such as Onsager hard-rod model and McMillan's model, address the Smectic A to nematic phase transition issues.

Nematic LCs are widely used as light modulators in the optical devices. Twist nematic LC device was first patented in 1970, which opened a gateway to the modern multi-billion LCD industry. The bulk molecular orientation following the direction of applied electric field is based on the dielectric constant of nematic LCs. In particular, the director of a positive dielectric constant material follows the applied electric field orientation, whereas the negative dielectric constant material stays perpendicular to the applied electric field. The in-plane-switching (IPS) or vertical alignment (VA) modes are developed for display applications, based on positive and negative dielectric constant LCs, respectively. Nevertheless, both positive and negative dielectric LCs can be utilized in the latest advance IPS mode (FFS mode) by manipulating the molecular alignment directions. The devices are noted as p-FFS or n-FFS mode to distinguish the difference. The p-FFS mode offers several attractive properties, such as high transmittance ($>90\%$ transmittance at 5 V), low operation voltage, fast response time (12 ms at room temperature and 30 ms at -20°C), and no gray scale inversion. These advantages make p-FFS particularly desirable for outdoor applications. The n-FFS offers high transmittance over 93% in 1-dimensional (1D) pixel pattern and 85% in a 2D pixel pattern. The response time is slower than p-FFS mode (20 ms at room temperature). There is noticeable gray scale inversion which can be suppressed by 2D pixel pattern. The Gamma curve is single curve for the R, G, B color which make it easier for the driving design (Chen *et al.*, 2015).

Nematic liquid crystal mixtures are mostly designed for display applications. Eutectic LC mixtures are prepared in order to accommodate wide temperature range to cover the device's application. The first commercial LC mixture E7, for instance, is made of four different molecules (3 biphenyl nitrile compounds and 1 terphenyl nitrile compound) in a fixed ratio. None of these four compounds has melting point below 24°C . The E7 mixture possesses nematic mesophase in the range of $-10 \sim 60.5^{\circ}\text{C}$. The lower temperature at -10°C indicates the eutectic point of this E7 mixture. For commercial nematic LC mixtures, the lower temperature can be reached around $-40 \sim -50^{\circ}\text{C}$ and the T_{ni} (nematic to isotropic temperature) is often varied from 80 to 120°C for display applications. In addition to application temperature, dielectric constant ($\Delta\epsilon$), birefringence (Δn), and rotational viscosity (γ_1), are the most important parameters to identify a proper LC for desired LC devices. In general, active-matrix (AM) MVA requires LC with $\Delta\epsilon = -3 \sim -5$, $\Delta n = 0.08 \sim 0.12$, and the rotational viscosity around 100 cp or lower. Conversely, the AM IPS mode requires the LC with $\Delta\epsilon = 7 \sim 11$, $\Delta n = 0.09 \sim 0.12$, and the rotational viscosity around 70 cp or lower. The faster rising and decay time could be achieved by using a thinner cell gap because the response time is at the ratio of cell gap squared ($\tau \propto d^2$). The response time can also be accelerated by warming up the device to elevated temperature, or by applying special driving voltage schemes (overdrive scheme). As the increasing demand of fast response time, Figure-of-Merit (FoM) grows to be one of the key parameters to evaluate the LC materials. The FoM is defined as $K(\Delta n)^2/\gamma_1$, where K is the elastic constant; Δn is the LC birefringence, and γ_1 is the rotational viscosity. High birefringence liquid crystal mixtures with low rotational viscosity are necessary for the high FoM LCs. They can be applied not only in the visible region, but in the infrared, GHz, or THz ranges. Gauza, *et al.* pointed out the importance of UV stability for the high birefringence liquid crystals. Mostly, the high birefringence liquid crystals consist of the rigid components, such as terphenyl or tolane moieties, which can enlarge its refractive index, but is susceptible to UV irradiation. Dąbrowski unveiled the material design and the corresponding birefringence (Dąbrowski *et al.*, 2013). The highest birefringence ($\Delta n = 0.79$) can be found in the LC molecule containing dibenzothiophene moiety. In addition, the highest Δn of isothiocyanate derivatives may reach 0.7 at 589 nm. As for nematic mixtures, the birefringence ($\Delta n = 0.5$) of LC mixtures can be well-prepared.

Smectic Phases

The smectic phases possess well-defined layers which are able to slide over one another as fluid. The word "smectic" originates from the Latin word "smecticus", meaning a soap-like texture. The smectic liquid crystals have positional order along with one long-range directional order. Since the layer structure is well-defined, any molecular tilt or distribution inside the layered structure is distinct to each other. For example, the molecules in Smectic A phase are oriented along the layer normal. In the same layer structure with molecules tilted away from the layer normal is called the Smectic C phase. Interestingly, the molecules self-aligned into a hexagonal packing and oriented along the layer normal is categorized as the Smectic B phase. As a result, various degrees of positional and directional orders are able to form different smectic phases.



The vast majority of LC research has been focused on chiral Smectic C for ferroelectric liquid crystals or scattering type Smectic A devices. The superior materials of Smectic LCs in the organic electronic devices have remained neglected for a while. Recently, the mobility of $13.9 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ has been reported from a solution prepared highly ordered Smectic E (SmE) mesophase in a bottom-gate bottom-contact FETs (Hsieh and Chen, 2015). Their intrinsic charge carrier mobility in the self-organized tilt orientation (SmE or SmI) is much superior to that of amorphous materials by two to three orders of magnitude.

Chiral Smectic and Nematic Phases

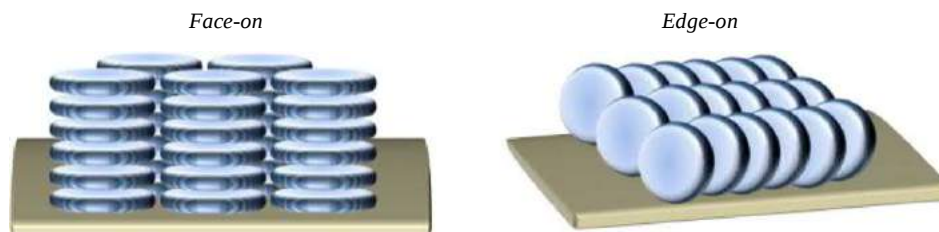
The chiral phases can be found in both nematic and smectic mesophases. The chiral-nematic phase exhibits chirality that was often called as the cholesteric phase as it was first observed from the cholesterol derivatives by Friedrich Reinitzer in 1888, and then confirmed in the publication by Otto Lehmann in 1889. This chiral-nematic phase exhibits a finite twisting angle between adjacent molecules perpendicular to the molecular director. The resulting helical layer structure is formed along the layer normal. Consisting of a helical stack of quasi-nematic layers, a molecular self-organization of chiral nematic structure is characterized by its helical pitch length p , and handedness. The pitch length p is defined as the distance over which the director rotates by 360° ; while the handedness describes the direction in which twisting of the nematic director occurs from one to another layer. The selective reflection in normal direction is described as $\lambda_R = p(n_e + n_o)/2$, in which n_e and n_o are the extraordinary and ordinary refractive indices, respectively. Consequently, the pitch and handedness of chiral-nematic liquid crystal systems can be manipulated to exploit unique optical properties, such as Bragg reflection, circularly polarized light, and low-threshold laser emission. The properties mentioned above can be utilized in a number of optical applications (Kopp *et al.*, 1998). Chiral-nematic LCs can be treated as 1-D photonic crystal when they aligned in the helical structure. Furthermore, Blue Phases (BPs) are found in the temperature range between chiral-nematic and isotropic liquid phases. They have three-dimensional cubic lattice structures assembled by the disclination line which is formed by crossing double helix twist cylinder structures. Coates and Gray (1973) first reported these thermodynamic phases in 1973. The theoretical predictions of the BPs hold icosahedral symmetry similar to quasicrystals were shown in 1981 (Seideman, 1990). The stabilization of blue phases over a temperature range of more than 60 K with room temperature (260–326 K) has been demonstrated in 2002 (Kikuchi *et al.*, 2002). The BPs stabilized at room temperature, that allows electro-optical switching with response times in the order of microsecond, opens the new path in fast light modulators or tunable photonic crystals for applications. The first Blue Phase mode LCD panel was developed in May 2008. Several drawbacks of blue phase LC for display applications were found in the high driving voltage, hysteresis, and light leakage in the dark state. Wu *et al.* has proposed a new electrode design to drastically reduce the driving voltage (Yoon *et al.*, 2010). A 12.7-in. LCD with 240 Hz frame rate color sequential display device was later realized by AUO in 2017 (Huang *et al.*, 2017). Its driving voltage was reduced to 15 V for single TFT addressing attributed to the protruded IPS electrodes structure. To solve hysteresis issue, Hsieh and Chen proposed using the fluoroacrylates monomers to ease the blue phase platelets texture formation under fast cooling and short annealing time before UV irradiation in 2015. The hysteresis was successfully suppressed and minimized (ΔV less than 0.8 V) in the PSBPLCs mixtures containing various fluoroacrylates (Hsieh and Chen, 2015). Although numerous issues of BP LCD remain unsolved, its exciting features, such as sub-millisecond gray-to-gray response time, the alignment-free, optically isotropic voltage-off state, and large cell gap tolerance, etc. continue to attract scientists to explore new possibilities.

Chiral-smectic C (SmC*) phase exhibits a spontaneous electric polarization which was discovered by Meyer in the mid-1970s. The first surface-stabilized SmC* ferroelectric liquid crystal was realized by Clark and Lagerwall at the beginning of the 1980s (Clark and Lagerwall, 1980). Since then, the outstanding polar effects with sub-million second response time in chiral-smectic phases have drawn great attention in research, as well as seen in the current commercial micro-display products. The ferroelectric liquid crystals were studied extensively in various aspects, with much attention turning toward display application. FLC device is usually considered as bi-stable optical element traditionally. However, the continuous-director-rotation mode (CDR mode; or so-called half-V shape mode) FLC type does not have bi-stable character, which has the advantage of gaining undisturbed transmittance from temperature variation due to the suppression of SmA phase (Chen and Lin, 2009). The antiferroelectric phases are first called SmC* subphases because of their unclear properties and the smaller temperature range of stability. This phase becomes clear when the tilted molecular direction alternates from one layer to another. The fact that the molecular assembly structure is referred to antclinic is to distinguish from the synclinic normal SmC phase, in which the director tilts in the same direction in adjacent layers. In addition to chiral smectic C, chiral smectic A can be tilted when electric field is applied to the system. The phenomenon is termed the “electroclinic effect” (Meyer and Pelcovits, 2002). The diffuse cone and the non-correlation models explain its field independence of birefringence. However, none of these models can fully describe the formation and temperature effects.

Discotic Phases

The first discotic liquid crystals (DLCs), benzene-hexa-*n*-alkanoates, were reported by Chandrasekhar *et al.* in the late 1970s. All kinds of discotic mesogens and metallomesogens have been reported and summarized (Chandrasekhar, 1993). In particular, disk-shape molecules orient themselves into a layer-like fashion known as columnar mesophases. The columns may be organized into rectangular or hexagonal arrays based on tilted structures of peripheral flexible chains attached to the aromatic cores. They are able to form nematic mesophase when the disk-shape mesogens do not stack into columns. Supramolecular organization of discotic

molecules can be controlled to give the macroscopic orientations either in planar uniaxial (face-on) or homeotropic alignment (edge-on) for different optoelectronic device applications as often seen in fast photoconductors for xerography, organic photovoltaics, organic light-emitting diodes, and chemical sensors. The negative birefringence property is unique in discotic liquid crystals. The hybrid orientation of discotic nematics film is applicable as optical compensation film improves the viewing angle of liquid crystal displays.



Discotic liquid crystals are novel organic semiconductor materials in the new generation. Interestingly, the mobility measured from FET transistors revealed that the higher values were found in crystalline phase right below discotic phase (van de Craats and Warman, 2001). A few critical reviews in 2007 have summarized the important aspects of recent researches in molecular design concepts, supramolecular structure, ordered thin films preparation, and electronic devices fabrication (Sergeyev *et al.*, 2007; Shimizu *et al.*, 2007).

Summary

The design and synthesis work of LCs were initially developed in depth by the Vorländer after Lehmann's report in 1889 and many others thereafter. The liquid crystals physics and materials still remained unpopular in early 20th century. Until the various device applications were demonstrated in the 1960s, the demands of novel LC materials and devices were sharply increased. The numerous varieties of physical and optical effects in liquid crystals have attracted many scientists and researchers to study the characteristics of LCs and apply them to various aspects of our lives. The fascinating anisotropic molecular self-assembly characteristics of liquid crystals have inspired many to manipulate their phases by energy, functional groups design, and surface alignments. The current interests in smectic liquid crystal research are focusing on de Vries/anti-de Vries type materials, tilting transitions without defect generation, and high-tilted anti-ferroelectric liquid crystals with perfect dark state and defect-free devices. In addition to these material developments, various functionalities can be integrated into different molecular assembly of LC mediums, such as organic solar cells, bio-sensors, chemical sensors, organic light-emitted diodes, etc. (Tyagi *et al.*, 2014; Popov *et al.*, 2016). It is worthy to further pursue the diverse liquid crystal systems from natural world and unfold many innovative possibilities in the future.

References

- Chandrasekhar, S., 1993. Discotic liquid crystals: A brief review. *Liq. Cryst.* 14, 3–14.
- Chen, H., Gao, Y., Wu, S.-T., 2015. n-FFS vs. p-FFS: Who wins? In: *SID Symposium Digest*, vol. 46, pp. 735–738.
- Chen, H.-M.P., Lin, C.-W., 2009. Free alignment defect, low driving voltage of half-V ferroelectric liquid crystal device. *Appl. Phys. Lett.* 95, 083501.
- Clark, N.A., Lagerwall, S.T., 1980. Submicrosecond bistable electrooptic switching in liquid crystals. *Appl. Phys. Lett.* 36, 899–901.
- Coates, D., Gray, G.W., 1973. Optical studies of the amorphous liquid-cholesteric liquid crystal transition: The "blue phase". *Phys. Lett.* 45, 115–116.
- Dąbrowski, R., Kula, P., Herman, J., 2013. High birefringence liquid crystals. *Crystals* 3, 443–482.
- de Gennes, P.G., Prost, J., 1993. *The Physics of Liquid Crystals*. Oxford: Clarendon Press.
- Hsieh, P.-J., Chen, H.-M.P., 2015. Hysteresis-free polymer stabilized blue phase liquid crystals comprising low surface tension monomers. *Liq. Cryst.* 42, 216–221.
- Huang, Y., Chen, H., Tan, G., *et al.*, 2017. New blue-phase liquid crystal optimized for color-sequential displays. In: *SID Symposium Digest*, vol. 48, pp. 486–489.
- Kikuchi, H., Yokota, M., Hisakado, Y., Yang, H., Kajiyama, T., 2002. Polymer-stabilized liquid crystal blue phases. *Nat. Mater.* 1, 64–68.
- Kopp, V.I., Fan, B., Vithana, H.K.M., Genack, A.Z., 1998. Low threshold lasing at the edge of a photonic stop band in cholesteric liquid crystals. *Opt. Lett.* 23, 1707–1709.
- Meyer, R.B., Pelcovits, R.A., 2002. Electroclinic effect and modulated phases in smectic liquid crystals. *Phys. Rev. E* 65, 061704.
- Popov, P., Honaker, L.W., Kooijman, E.E., Mann, E.K., Jákli, A.I., 2016. A liquid crystal biosensor for specific detection of antigens. *Sens. Bio-Sens. Res.* 8, 31–35.
- Seideman, T., 1990. The liquid-crystalline blue phases. *Rep. Prog. Phys.* 53, 659–705.
- Sergeyev, S., Pisula, W., Geerts, Y.H., 2007. Discotic liquid crystals: A new generation of organic semiconductors. *Chem. Soc. Rev.* 36, 1902–1929.
- Shimizu, Y., Oikawa, K., Nakayama, K.-i., Guillon, D., 2007. Mesophase semiconductors in field effect transistors. *J. Mater. Chem.* 17, 4223–4229.
- Tyagi, M., Chandran, A., Joshi, T., *et al.*, 2014. Self assembled monolayer based liquid crystal biosensor for free cholesterol detection. *Appl. Phys. Lett.* 104, 154104.
- van de Craats, A.M., Warman, J.M., 2001. The core-size effect on the mobility of charge in discotic liquid crystalline materials. *Adv. Mater.* 13, 130–133.
- Yoon, S., Kim, M., Kim, M.S., *et al.*, 2010. Optimisation of electrode structure to improve the electro-optic characteristics of liquid crystal display based on the Kerr effect. *Liq. Cryst.* 37, 201–208.

Further Reading

- Iino, H., Usui, T., Hanna, J., 2015. Liquid crystals for organic thin-film transistors. *Nat. Comm.* 6, 6828–7828.

TFT Materials and Devices

Po-Tsun Liu, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Active-matrix (AM) displays have already accumulated unbeatable influence and created a brand-new life style in recent 20 years. In the active-matrix display operation scheme, thin film transistor (TFT) backplane technology is essential for LCD, OLED, eEPD, and even flexible displays. Taking AMLCD as an example, in each display pixel the liquid crystal layer functions as a light shutter, in which the switch is controlled by the TFT. TFT devices play a role of conducting electric current to charge the LC capacitor and storage capacitor. When the TFT device is operated at “on” state, the pixel signal voltage is delivered and charging the LC capacitors during tens of micron seconds (μs). With TFT operated at “off” state, pixel signal voltage will be stored in the LC capacitors and prevent it from being leaky for several mini seconds (ms). For AMOLED operation, TFT acts as an electric current driver to light “on” and “off” the OLED pixel besides a switch.

A TFT device structure comprises several material layers, including metallic gate electrode, gate dielectric film, semiconductor channel, and source/drain electrodes, which is similar to metal-oxide-semiconductor field effect transistor (MOSFET). In the turn-on operation, charged carriers (electron or hole) induced by gate bias voltage are conductive in the semiconductor channel and move from source to drain terminal under the influence of drain-to-source voltage. The carrier mobility is thereby dependent on the semiconductor channel layer, which can be classified into inorganic and organic materials. The inorganic semiconductor materials widely cover hydrogenated amorphous silicon (a-Si:H), polycrystalline silicon (poly-Si), and amorphous oxide semiconductors (AOSs), while the organic ones mainly include small or large molecular polymers according to the present technical development (Table 1). As the TFT device structure, there are three types of structures commercially fabricated: (1) bottom-gate back channel etch (BCE) structure, (2) bottom-gate etch stopper (ES) structure and (3) top gate (TG) structure (Fig. 1). The selected steps of TFT fabrication processes are schematically depicted as followed (Fig. 2). For realistic applications in flat-panel displays

Table 1 Comparison of TFT devices with different semiconductor layers

Characteristic	a-Si TFT	poly-Si TFT	TAOS TFT (transparent amorphous oxide semiconductor)	Organic TFT
Stability in ambient	Excellent	Excellent	Excellent	Poor
Uniformity	Good	Vth Dispersion	Good	Vth Dispersion
Performance	$\mu_n = 0.7$ ($\text{cm}^2/\text{V}\cdot\text{s}$)	$\mu_n = 200$ ($\text{cm}^2/\text{V}\cdot\text{s}$) $\mu_p = 100$ ($\text{cm}^2/\text{V}\cdot\text{s}$)	$\mu_n = 1\text{--}100$ ($\text{cm}^2/\text{V}\cdot\text{s}$)	$\mu_p = 0.1\text{--}5$ ($\text{cm}^2/\text{V}\cdot\text{s}$)
Design/Device	NMOS	CMOS	NMOS	PMOS
Off current	low	high	lowest	fair
Processing temperature	Low	High	Low	Low
Flexibility	X	X	0	0
Transparent	X	X	0	0
Process	3–5 masks	7–9 masks (ELC + Implantation)	3–5 masks	3–5 masks

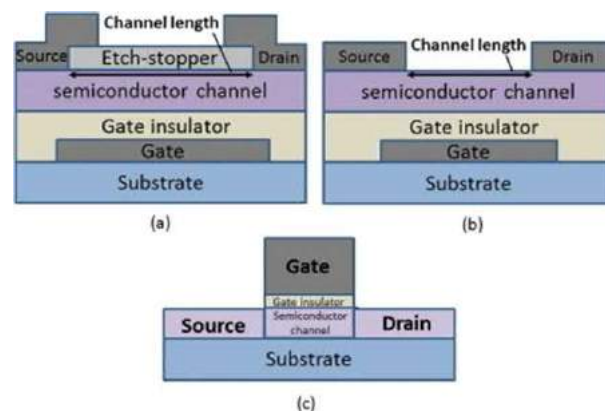


Fig. 1 TFT devices with (a) bottom-gate ES, (b) bottom-gate BCE, and (c) top-gate (TG) structure.

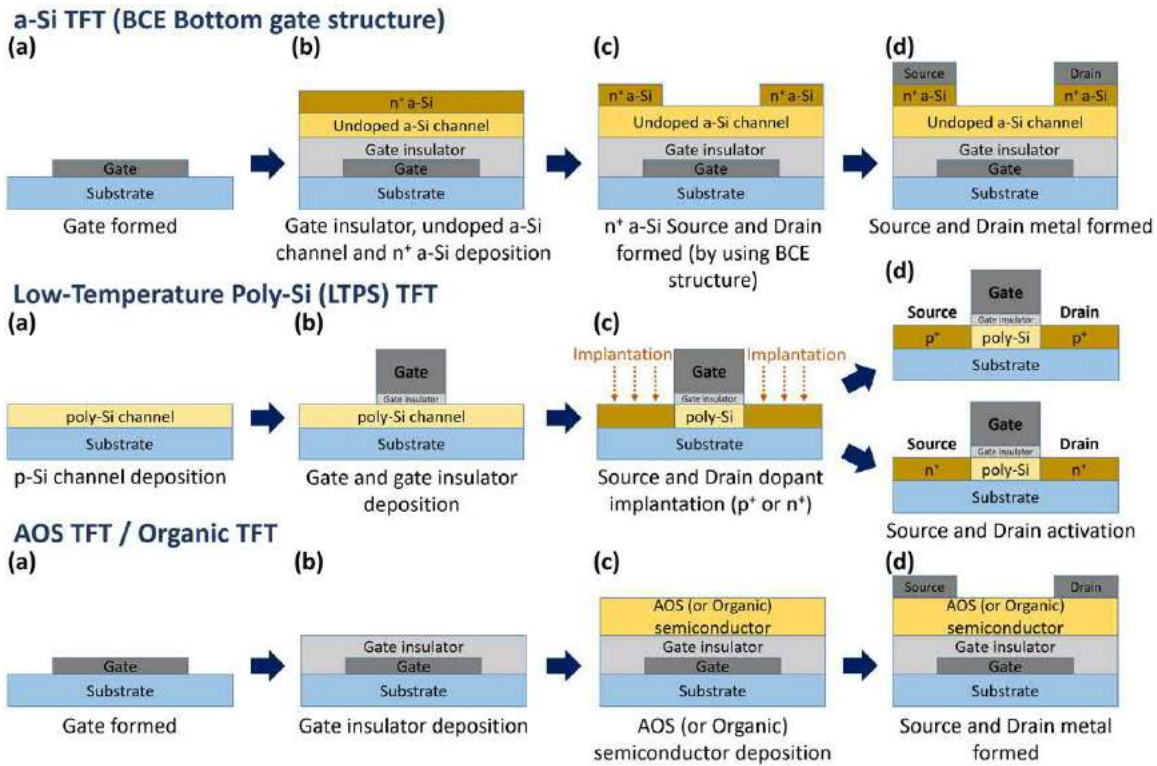


Fig. 2 Selected steps of fabrication processes for a-Si TFT, LTPS TFT, and AOS TFT/Organic TFT.

(FPDs), the TFT device has to meet several requirements, such as high turn-on current (drain current I_D), high field-effect mobility μ_{FE} (cm^2/Vs), low threshold voltage (V_{th}), low leakage current, low leakage current from gate terminal, low parasitic (overlap) capacitance, and low photosensitivity to ambient/backlight.

a-Si:H Thin Film and a-Si:H TFT

In the present FPD industry, the engineering of a-Si:H TFT is suitable for large-area and high-resolution AMLCD products. The plasma-enhanced chemical vapor deposition (PECVD) reactor is widely used in the deposition of a-Si:H thin film layer with the reactants of SiH_4 and H_2 gases around 300°C for TFT device fabrication. The network of a-Si:H thin film lacks long-range arrangement order so that bonding deviations occurs and perturbs the energy of sp^3 bonding and anti-bonding states in a specific co-valent bond compared to its crystalline case (*c*-Si). Therefore, a-Si:H film is full of a great deal of dangling bonds and weak Si-Si bonding associated with band tail electron states in energy band gap, leading to charge-trapping centers. Dangling bonds with electron energy states near Fermi level are the major contributor of these defects. Hydrogen is the most effective passivation agent for dangling bonds. Also, the uneven distribution of donor-like and acceptor-like trap states in energy band gap causes the slightly n-type behavior for an undoped a-Si:H film, which can be observed experimentally (Yue Kuo, 2004). The field-effect electron mobility is theoretically 1, because of the scattering effects of these band tail states.

There are two types of a-Si:H TFT devices adopted commercially in the present FPDs: the bottom-gate (inverted staggered) TFT with etch stop (ES) layer and backchannel etched (BCE) structure (Ban *et al.*, 1996; Maeda *et al.*, 1992). The BCE a-Si:H TFT reduces one photo-mask step for array fabrication compared to the ES a-Si:H TFT. Also, the gate insulator SiN_x , undoped a-Si:H channel, and n^+ a-Si:H layer as source/drain regions are deposited sequentially without unloading processes. However, the BCE a-Si:H TFT needs a thicker a-Si:H layer to have enough process margin for complete removal of n^+ a-Si:H on the top of the a-Si:H channel. Light illumination on a-Si:H will give rise to the decrease of both the dark conductivity and photoconductivity while the degradation can be recovered by a subsequent thermal annealing at around 200°C . In addition, the gate-to-source bias voltage applied to a-Si:H TFT will cause the increase of threshold voltage with increasing time. With a subsequent thermal annealing at 200°C for a few hours, the degraded characteristics of TFT can be restored to its original state (Yue Kuo, 2004; Street, 1991). Although a-Si:H TFT suffers several issues on electrical stability, the a-Si:H TFT array is nowadays a popular backplane technology in the large-area AMLCDs due to the features of a-Si:H film uniformity and low-cost fabrication.

Poly-Si Thin Film and Poly-Si TFT

Polycrystalline silicon thin-film transistors (poly-Si TFTs) have been widely used in many potential applications including pixel driving elements of active matrix organic light emitting diode (AM-OLED), and integrated peripheral driving circuits and addressing elements of active-matrix liquid crystal displays (AMLCDs) (Walker *et al.*, 2003; Shimoda *et al.*, 1998; Lewis *et al.*, 1990; Oshima and Morozumi, 1989). Recently, poly-Si TFTs are very attractive for system on top of the panel (SOP) as devices performances improve further. The degree of circuit integration will continue to increase as device performances improve further. The entire system will include memories, such as SRAM and nonvolatile FLSAH Memories, solar cells, and touch sensors as well as driver circuits for AMLCDs (Oh *et al.*, 2000; Young *et al.*, 1996; Cao *et al.*, 1994). Poly-Si TFTs have the potential advantages of silicon-on-insulator (SOI) MOSFETs such as simple fabrication process, good device-to-device isolation, high circuit density, and high device performance as well as the possibility to be applied in vertical 3-D integration. Poly-Si TFTs technology has been receiving more attention because it is a promising mean of achieving 3-D integration, which has been utilized in various 3-D circuits (Hayashi *et al.*, 1996; Kuriyama *et al.*, 1998; Wang *et al.*, 2000). A grain boundary is the interface between two grains, or crystallites, in a polycrystalline material. However, grain boundaries of poly-Si material cause trap and tail states which put strong influences on device characteristics including an increase in threshold voltage (V_{TH}), decrease in mobility, increase in off-state leakage current, decrease in on-state current, poor subthreshold characteristics and degradation in device reliability (Matore, 1984; Levinson *et al.*, 1982). The trap states in the grain boundaries is thought to be associated with the presence of dangling and strained bounds (Lewis *et al.*, 1989). To solve these problems, some crystallization methods have been introduced to enlarge the grain size. Poly-Si thin film with large crystalline grains can be obtained using a variety of techniques: solid phase crystallization (SPC) (Hatalis and Greve, 1988), laser crystallization (LC) (Smith *et al.*, 1997) and metal-induced lateral crystallization (MILC) (Lee *et al.*, 1995; Lee and Joo, 1996; Jin *et al.*, 1998) of amorphous silicon (*a*-Si). Because crystallization of poly-Si channel plays a main role in carrier mobility and uniformity of poly-Si channel, a robust crystallization method is required for poly-Si TFTs in future various applications.

SPC is a classical method of achieving poly-Si by furnace at 600°C for 24 h. SPC produces a large-grain poly-Si, but its crystallization temperature is too high for the ordinary glass substrate. Laser crystallization (LC) method is popular for low temperature poly-Si (LTPS) TFTs. A new low temperature (500°C) crystallization method for poly-Si TFTs was developed: Metal-Induced Lateral Crystallization (MILC). The *a*-Si film in the channel area of a TFT was laterally crystallized from the source/drain area, on which an ultrathin nickel layer was deposited before annealing. The enhanced crystallization rate at this low temperature has been attributed to the silicate formation of NiSi_2 . This low-temperature budget process with low metal contamination may be used to overcome problems associated with the SPC approach (Lewis *et al.*, 1989). Further improvement in the poly-Si TFTs performances has been recently obtained by the introduction of excimer laser crystallization (ELC), leading to electron field-effect mobility $> 100 \text{ cm}^2/\text{V s}$. Excimer lasers emit in the UV region (output wavelengths 193 nm, 248 nm and 308 nm for ArF, KrF and XeCl gas mixtures, respectively) with a short pulse duration (10–30 ns). The combination in Si of strong optical absorption of the UV light and small heat diffusion length during the laser pulse implies that high temperatures can be developed in the Si-surface region, causing melting, without appreciable heating ($< 400^\circ\text{C}$) of the substrate. This makes the ELC-process compatible with glass substrates, one of the major advantages of this technique. Such achievements have finally opened the door to the fabrication of complex CMOS circuits based on poly-Si TFTs and some major display manufacturers have already on the market AMLCDs with integrated polysilicon TFTs drivers (Smith *et al.*, 1997). The technologies of poly-Si TFTs will become more and more important in future applications including 3D integration of active devices and SOP. More researches investigate the related new technologies and underlying mechanisms in thin-film devices with scaling down dimension are worthy to study.

AOS Thin Film and AOS TFT

With the high demand for high-resolution image ($> 2000 \times 4000$ pixels) and fast-frame rate ($> 120 \text{ Hz}$), *a*-Si:H TFT is not an appropriate device for next generation FPD displays due to its low mobility ($< 1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) and high photosensitivity (low band gap about 1.7 eV). On the other hand, poly-Si TFT which has high mobility ($> 100 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) suffers from a problem with the variation of electrical properties due to grain boundary. Amorphous oxide semiconductors (AOS) including InGaZnO (Nomura *et al.*, 2004), InZnSnO (Fuh *et al.*, 2014), AlZnSnO (Teng *et al.*, 2013), InZnO (Liu *et al.*, 2009) and InWO (Liu *et al.*, 2016a), are materials with high electron mobility larger than $10 \text{ cm}^2/\text{Vs}$ and high transmission about 80% in visible portion of the electromagnetic spectrum even when their films are deposited at room temperature. The high transparency of AOSs enables the realization of emerging transparent display technology, in which users can be allowed to view what is shown on a glass video screen while still being able to see through it. Among AOS materials, amorphous In-Ga-Zn-O (*a*-IGZO) has been attractive. For the In_2O_3 - Ga_2O_3 -ZnO ternary system, the incorporation of cations with large ionic valance such as Ga^{3+} to high conductive oxides such as In_2O_3 and ZnO is effective to control the carrier concentration due to their strong metal-oxygen bonds (Nomura *et al.*, 2004). The mobility of InGaZnO₄ is primary determined by the In_2O_3 content fraction because only In^{3+} meets the electron configuration criterion $(n-1)d^{10}ns^0$ ($n \geq 4$) of heavy post transition metal cation for ionic AOS (IAOS) among the three cations. The large ionic valence ion of Ga^{3+} combines with high conductive oxides such as In_2O_3 and ZnO to control the carrier concentration effectively because of the strong metal-oxygen bonds. Ga^{3+} also suppresses carrier generation via oxygen vacancy formation because Ga ion forms stronger bond with oxygen than Zn and In ions (Nomura *et al.*, 2004).

AOS thin films can be formed by a variety of thin film deposition technologies, including sputtering, pulse laser deposition (PLD), and solution-based deposition methods (Fortunato *et al.*, 2012). For the large-area manufacturing, the sputtering technology is favorable. Three types of AOS TFT device aforementioned can be adopted for FPD applications. There are many factors for the electrical stability of AOS TFT devices, such as the oxide semiconductor material, the material composition, content of oxygen deficiency, the device structure, the gate insulator, the environment for stressing, the electrical biasing conditions, and the intensity of the illuminated light. AOS TFTs are generally sensitive to the oxygen or moisture during the bias stability test, resulting in threshold voltage shift. Post-deposition annealing process can be implemented to enhance the resistance of AOS channel to electrical stress and oxygen/moisture uptake from environment (Fuh *et al.*, 2014). The annealing process provides the energy for structure relaxation of AOS thin film, which reduces the content of oxygen deficiencies in AOS channel. Electrical trap states are originated from the oxygen deficiencies. A better transfer characteristic and electrical stability of AOS TFT can be resultantly achieved since both trap states in the AOS channel and at the interface of AOS/gate insulator are reduced effectively by post-deposition annealing. On the other hand, the channel passivation methods can be used to isolate the absorption oxygen on the channel surface and enhance electrical stability (Park *et al.*, 2008a,b; Pei *et al.*, 2009; Liu *et al.*, 2016b). Novel AOS materials and manufacture processes have been proposed for overcoming many critical challenges. It is a promising technology for next-generation FPD displays.

Organic Thin Film and OTFT

Organic semiconductors (OSCs) are another TFT device technology actively studied over the past two decades. Organic electronics are lighter, more flexible, and potentially lower manufacture cost than their inorganic counterparts. Steady improvement in the electrical performance and the stability of OSCs led the technology development to utilize various flexible substrates, such as glass, plastic, and even paper (Kim *et al.*, 2004; Moore, 2002; Klauk *et al.*, 2003; Tsumura *et al.*, 1986; Horowitz, 1998; Lee *et al.*, 2006; Schon and Batlogg, 2001; Lee and Subramanian, 2003). The organic semiconductors that are used are pentacene, poly (3-hexylthiophene) (P3HT), poly (3-alkylthiophene) (P3AT), and poly (3-octylthiophene) (P3OT), which can be deposited by vacuum evaporation and solution processing techniques (Shekar *et al.*, 2004). Their performance critically depends on the efficiency with which charge carriers (electrons and/or holes) move within the π -conjugated semiconductor materials. Most of the reported organic TFTs (OTFTs) were built as the bottom gate structure for the concern in deposition of active material on the insulator instead of at the bottom. Two categories of gate dielectrics are commonly used in organic transistors: inorganic and organic. Examples of inorganic dielectrics include silicon dioxide and silicon nitride, typically used in traditional silicon integrated circuits and a-Si:H TFT display backplanes. The solution processable organic dielectrics, Poly-vinylphenyle (PVP) and polymethylmethacrylate (PMMA) are typical candidates for organic gate dielectrics (Shekar *et al.*, 2004). The reported field-effect mobility of vacuum evaporation deposited pentacene-based TFTs have been quoted in the range of 0.1–1.5 cm²/Vs which is comparable to devices using a-Si:H as the semiconductor material. However, OTFTs typically suffer from several issues including high-voltage operation, and thermal stability. The proceeding development in the enhancement of field-effect mobility and thermal stability are still in progress.

Summary

TFT is a critical device to switch a pixel on and off in FPD technology. Nowadays, a-Si:H TFT array is mainly adopted as a backplane for large-area display applications. To cope with the increasing demand of image resolution with high frame rate, poly-Si TFT with high-field-effect mobility and driving current is gradually and extensively required in FPD industry. Additionally, AOS TFTs and organic TFTs have been studied as possible alternatives for Si-based TFTs to lower production costs and processing temperature for flexible display applications. At present, leading manufacturers have already demonstrated many prototypes of AMLCD and AMOLED panels with AOS TFT backplanes. Also, some of them have announced the mass production for innovative applications.

References

- Ban, A., Nishioka, Y., Shimada, T., Okamoto, M., Katayama, M., 1996. A simplified process for SVGA TFT-LCDs with single-layered ITO source bus-lines. SID 96 Digest, pp. 93–96.
- Cao, M., Zhao, T., Saraswat, K.C., Plummer, J.D., 1994. A simple EEPROM cell using twin polysilicon thin film transistor. IEEE Electron Device Lett. 15, 304–306.
- Fortunato, E., Barquinha, P., Martins, R., 2012. Oxide semiconductor thin-film transistors: A review of recent advances. Adv. Mater. 24, 2945–2986.
- Fuh, C.S., Liu, P.T., Huang, W.H., Sze, S.M., 2014. Effect of annealing on defect elimination for high mobility amorphous indium-zinc-tin-oxide thin-film transistor. IEEE Electron Device Lett. 35 (11), 1103–1105.
- Hatalis, M.K., Greve, D.W., 1988. Large grain polycrystalline silicon by low-temperature annealing of low-pressure chemical vapor deposited amorphous silicon films. J. Appl. Phys. 63, 2260–2266.
- Hayashi, F., Ohkubo, H., Takahashi, T., *et al.*, 1996. A highly stable SRAM memory cell with top-gated P–N drain poly-Si TFT's for 1.5 V operation. In: IEDM Tech. Dig., pp. 283–286.

- Horowitz, G., 1998. Organic field-effect transistors. *Adv. Mater.* 10 (5), 365–377.
- Jin, Z., Bhat, G.A., Yeung, M., Kwok, H.S., Wong, M., 1998. Nickel induced crystallization of amorphous silicon thin films. *J. Appl. Phys.* 84 (1), 194–200.
- Kim, Y.H., Moon, D.G., Han, J.I., 2004. Organic TFT array on a paper substrate. *IEEE Electron Device Lett.* 25 (10), 702–704.
- Klauk, H., Halik, M., Zschieschang, U., *et al.*, 2003. Pentacene organic transistors and ring oscillators on glass and on flexible polymeric substrates. *Appl. Phys. Lett.* 82 (23), 4175–4177.
- Kuriyama, H., Ishigaki, Y., Fujii, Y., *et al.*, 1998. A C-switch cell for low-voltage and high-density SRAM's. *IEEE Trans. Electron Devices* 45, 2483–2488.
- Lee, J.B., Subramanian, V., 2003. Organic transistors on fiber: A first step toward electronic textiles. In: *IEDM Tech. Dig.*, pp. 8.3.1–8.3.4.
- Lee, K.S., Smith, T.J., Dickey, K.C., *et al.*, 2006. High resolution characterization of pentacene/polyaniline interfaces in thin-film transistors. *Adv. Funct. Mater.* 16 (18), 2409–2414.
- Lee, S., Jeon, Y., Joo, S., 1995. Pd induced lateral crystallization of amorphous Si thin film. *Appl. Phys. Lett.* 66, 1671–1673.
- Lee, S.W., Joo, S.K., 1996. Low temperature poly-Si thin-film transistor fabrication by metal-induced lateral crystallization. *IEEE Electron Device Lett.* 17, 160–162.
- Levinson, J., Shepherd, F.R., Scanlon, P.J., *et al.*, 1982. Conductivity behavior in polycrystalline semiconductor thin film transistors. *J. Appl. Phys.* 53, 1193–1202.
- Lewis, A.G., Huang, T.Y., Wu, I.W., Bruce, R.H., Chiang, A., 1989. Physical mechanisms for short channel effect in polysilicon thin film transistors. In: *IEDM Tech. Dig.*, pp. 349–352.
- Lewis, A.G., Wu, I.-W., Huang, T.Y., Chiang, A., Bruce, R.H., 1990. Active matrix liquid crystal display design using low and high temperature processed polysilicon TFTs. In: *IEDM Tech. Dig.*, San Francisco, CA, pp. 843–846.
- Liu, P.T., Chang, C.H., Chang, C.J., 2016a. Suppression of photo-bias induced instability for amorphous indium tungsten oxide thin film transistors with bi-layer structure. *Appl. Phys. Lett.* 108 (26), 261603.
- Liu, P.T., Chang, C.H., Fuh, C.S., Liao, Y.T., Sze, S.M., 2016b. Effects of nitrogen on amorphous nitrogenated InGaZnO (a-IGZO: n) thin film transistors. *IEEE/OSA Journal of Display Technology* 12 (10), 1070–1077.
- Liu, P.T., Chou, Y.T., Teng, L.F., 2009. Environment-dependent metastability of passivation-free InZnO TFT after gate bias stress. *Appl. Phys. Lett.* 95, 233504.
- Maeda, H., Fujii, K., Yamagishi, N., *et al.*, 1992. A 15-in.-diagonal full-color high-resolution TFT-LCD. *SID 92 Digest*, pp. 47–50.
- Matare, H.F., 1984. Carrier transport at grain boundaries in semiconductors. *J. Appl. Phys.* 56, 2605–2631.
- Moore, S.K., 2002. Just one word: Plastics. *IEEE Spectrum* 39 (9), 55–57.
- Nomura, K., Ohta, H., Takaji, A., *et al.*, 2004. Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors. *Nature* 432, 488–492.
- Oh, J.H., Chung, H.J., Lee, N.I., Han, C.H., 2000. A high-endurance low-temperature polysilicon thin-film transistor EEPROM cell. *IEEE Electron Device Lett.* 21, 304–306.
- Oshima, H., Morozumi S., 1989. Feature trends for TFT integrated circuits on glass substrates. In: *IEDM Tech. Dig.*, Washington, DC, p. 157.
- Park, J., Kim, S., Kim, C., *et al.*, 2008b. High-performance amorphous gallium indium zinc oxide thin-film transistors through N₂O plasma passivation. *Appl. Phys. Lett.* 93, 053505.
- Park, W.J., Shin, H.S., Du Ahn, B., *et al.*, 2008a. Investigation on doping dependency of solution-processed Ga-doped ZnO thin film transistor. *Appl. Phys. Lett.* 93, 083508.
- Pei, Z.L., Pereira, L., Goncalves, G., *et al.*, 2009. Room-temperature cosputtered HfO₂ – Al₂O₃ multicomponent gate dielectrics. *Electrochem. Solid State Lett.* 12, G65.
- Schon, J.H., Batlogg, B., 2001. Trapping in organic field-effect transistors. *J. Appl. Phys.* 89 (1), 336–341.
- Shekar, B.C., Lee, J., Rhee, S., 2004. Organic thin film transistors: Materials, processes, and devices. *Korean J. Chem. Eng.* 21 (1), 267–285.
- Shimoda, T., Ohshima, H., Miyashita, S., *et al.*, 1998. High resolution light emitting polymer display driven by low temperature polysilicon thin film transistor with integrated driver. In: *Proceedings ASID*, Seoul, Korea, pp. 217–220.
- Smith, P.M., Carey, P.G., Sigmon, T.W., 1997. Excimer laser crystallization and doping of silicon films on plastics substrates. *Appl. Phys. Lett.* 70, 342–344.
- Street, R.A., 1991. Hydrogen diffusion and electronic metastability in amorphous silicon. *Physica B* 170 (69),
- Teng, L.F., Liu, P.T., Wang, W.Y., 2013. Electrical performance enhancement of Al-Zn-Sn-O thin film transistor by supercritical fluid treatment. *IEEE Electron Device Lett.* 34 (9), 1154–1156.
- Tsumura, A., Koezuka, H., Ando, T., 1986. Macromolecular electronic device: Field effect transistor with a polythiophene thin film. *Appl. Phys. Lett.* 49 (18), 1210–1212.
- Walker, A.J., Nallamothu, S., Chen, E.H., *et al.*, 2003. 3D TFT-SONOS memory cell for ultra-high density file storage applications. In: *VLSI Symp. Tech. Dig.*, pp. 29–30.
- Wang, H., Chan, M., Jagar, S., Wang, Y., Ko, P.K., 2000. Submicron super TFTs for 3-D VLSI applications. *IEEE Electron Device Lett.* 21 (9), 439–441.
- Young, N.D., Harkin, G., Bunn, R.M., McCulloch, D.J., French, I.D., 1996. The fabrication and characterization of EEPROM arrays on glass using a low-temperature poly-si TFT Process. *IEEE Trans. Electron Devices* 43, 1930–1936.
- Yue Kuo, , 2004. *Thin Film Transistors-Materials and Processes: Amorphous Silicon Thin Film Transistors*. 1. Boston: Kluwer Academic Publishers.

Further Reading

- Fortunato, G., Mariucci, L., Carluccio, R., Pecora, A., Foglietti, V., 2000. Excimer laser crystallization techniques for polysilicon TFTs. *Applied Surface Science* 154, 95–104.

Color Formation of LCD – Spatial and Temporal

Fang-Cheng Lin, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Structure of the Human Eye

The human eye is an adaptive lens. Fig. 1 shows the cross-section of the human eye (Webvision, 2017). Light goes into the pupil of the eye and continues through the lens which acts the same as a camera focusing light onto the retina. The retina contains two types of photoreceptors, rods and cones, named for their shapes. The rods are more numerous, roughly 125 million, and are more sensitive than the cones to detect small amounts of light, such as starlight. However, they are unable to distinguish color. Six to seven million cones perform in bright light, giving detailed colored views, but they fail at low light levels (Hecht, 1998). Cones are much more concentrated in the central yellow spot known as the “macula.” In the center of the macula is the “fovea centralis,” a 0.2 mm diameter rod-free area with very thin, densely packed cones. Also, the fovea is the area of the eye responsible for sharpest vision.

The distribution of rods and cones in the retina is not uniform, as illustrated in Fig. 2 (Osterberg, 1935). The density of cones falls rapidly to a constant level at about 10–15 degrees from the fovea. At about 15–20 degrees from the fovea, the density of the rods reaches a maximum. Because there is only one pigment type for rods, humans only see objects as shades of gray. In contrast, there are three types of cones which refer to as the short-wavelength (S-cone), middle-wavelength (M-cone), and long-wavelength (L-cone) sensitive cones, as shown in Fig. 3 (CIE 170-1:2006, 2017). This is why humans perceive vivid colors. While receiving light, the retina converts it into nerve signals, and then the nerve signals are sent through the optic nerve, at the back of the eye, to the brain. Hence, in the optical nerve region which located about 13–18 degrees of the retina, there is no any receptor to sense light stimuli and is so-called “blind spot.”

Color Matching Functions and Tristimulus

A color matching experiment was conducted in 1931 to describe the color that the human eye perceives, as illustrated in Fig. 4. Based on this study, a set of color matching functions (CMFs): $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ was derived as shown in Fig. 5(a) (UPENN, 2017). The CMFs were transformed to $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ using mathematics to avoid the negative values appearing on the CMFs, as shown in Fig. 5(b). This is the well-known CIEXYZ color space established by the Commission Internationale de l’Éclairage (CIE) in 1931. Consequently, the color perceived can be quantified by the CIE as three tristimulus: X, Y, and Z. The color (X, Y, Z) is affected by three components: illuminant (P), an object’s reflectance factor (R), and the CMFs of the human eye, ($\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$), and is given in Eq. (1).

$$X = \int P(\lambda)R(\lambda)\bar{x}(\lambda)d\lambda; Y = \int P(\lambda)R(\lambda)\bar{y}(\lambda)d\lambda; Z = \int P(\lambda)R(\lambda)\bar{z}(\lambda)d\lambda \quad (1)$$

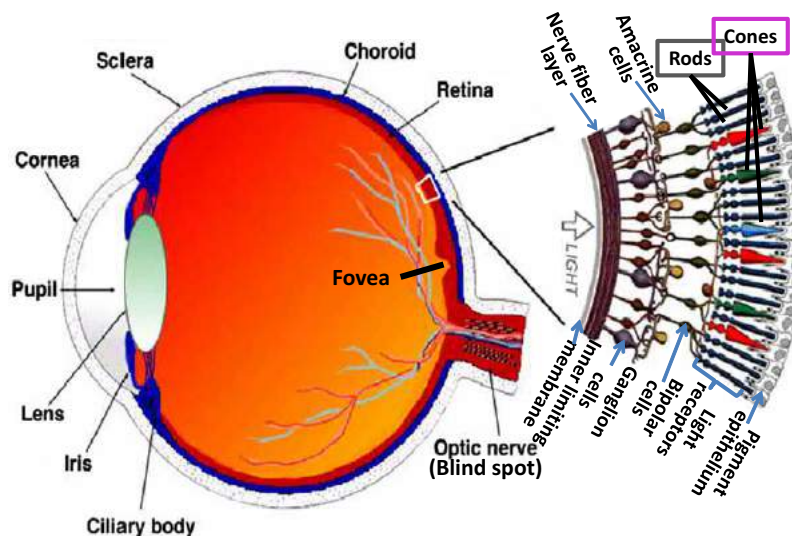


Fig. 1 A drawing of a cross-section through the human eye with a schematic enlargement of the retina including rod and cone light receptors (Webvision, 2017). The photoreceptors, cones and rods, receive the incident light and then send the response signals to ganglion cell, and finally optic nerves collect the signals to the brain.

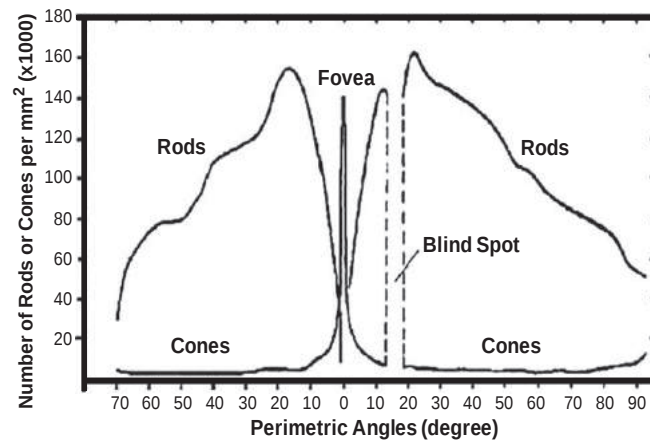


Fig. 2 Distribution of rods and cones along a line in the retina (Osterberg, 1935). The cones are much more concentrated in the macula, with an extremely high density in the center of the macula-fovea centralis. Oppositely, rods widely spread at the retina except for the center of the fovea, with a sharp decline in the fovea. An area, located about 13–18 degrees of the retina, existing no photoreceptor cells is the blind spot, where are full of the optic nerves and called optic disc and blind spot.

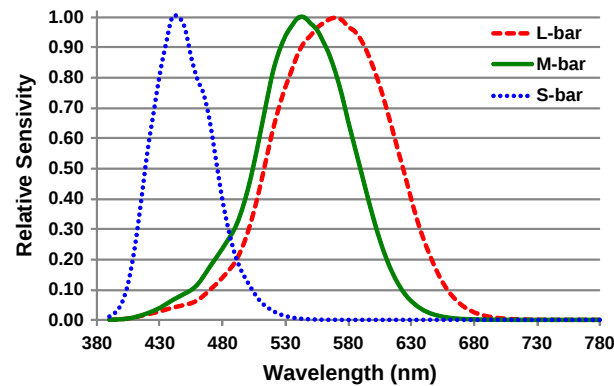


Fig. 3 Normalized response sensitivity of three cone-types – L-, M-, and S-cones, named after their responses at long, medium, and short wavelengths (CIE 170-1:2006, 2017).

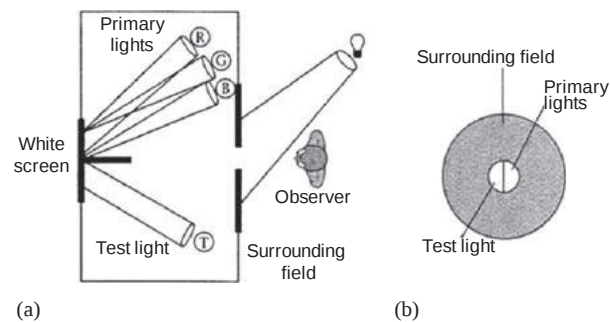


Fig. 4 Color matching experiment. (a) The observer sees a white screen divided into two halves, one-half illuminated by a test light, the other half illuminated by a mixture of one or more primary lights (here three are shown). (b) The observer can adjust the intensity of the primary light(s) to try to make the two halves of the screen match (UPENN, 2017).

Colors in Liquid Crystal Displays

Liquid crystal displays (LCDs) have become the dominant display technology for applications ranging from small mobile devices, notebooks, car navigation systems up to large-size television sets. This high penetration level can be attributed to a host of advantageous features, including high resolution, high brightness, light-weight, and a thin form factor. In an LCD, a vivid color

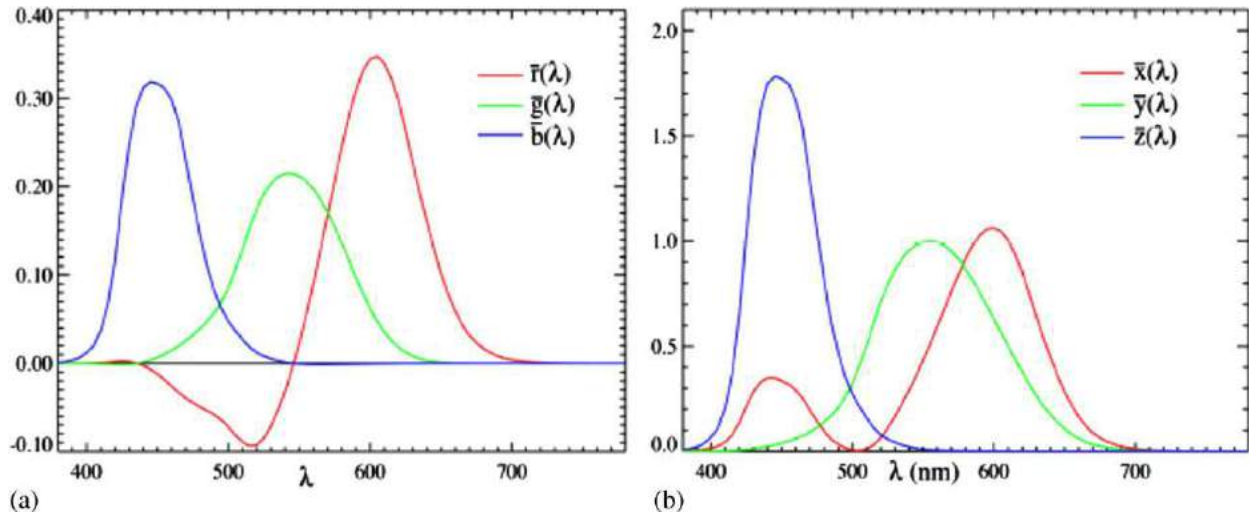


Fig. 5 Color matching functions (CMFs) from (a) human experiments, $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ and (b) transformation using mathematics, $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, to prevent from appearing negative values (Wikipedia, 2017).

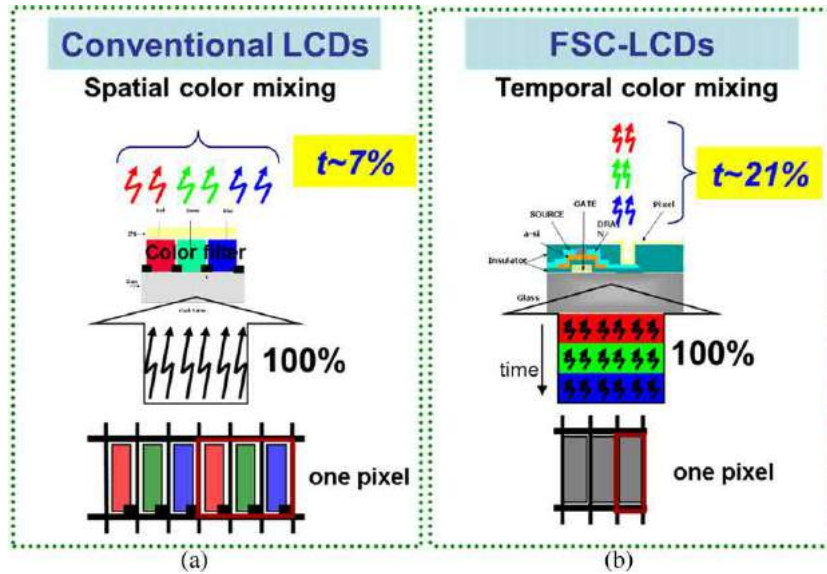


Fig. 6 Display mechanisms of (a) a conventional color filter type LCD by spatial color synthesis, with a poor optical throughput of 7% and (b) a field sequential color LCD by temporal color synthesis, with a triple optical throughput of 21%. Getting rid of color filters, the field sequential color LCD has a triple horizontal panel resolution compared with the color filter type LCD.

performed can be classified into two types: spatial color synthesis by micro color filters (typically red (R), green (G), and blue (B)) and temporal color synthesis by rapidly flash the primaries (RGB) time-sequentially, as shown in Fig. 6.

Spatial Color Synthesis

A traditional TFT-LCD is commonly illuminated using a constant full-on backlight which consists of conventional cold cathode fluorescence lamps (CCFLs) or white light emitting diodes (W-LEDs). The backlight polarization state changes to the linear state after propagating through a polarizer on the TFT-array glass. Different voltages between these two glasses control LC orientations and create different phase retardations due to the LC birefringence property. When the linear polarized lights propagate through the LC cell, the polarization states are modulated by differing LC orientations. Finally, lights from different sub-pixels continue to propagate into red, green, and blue color filters in various intensities and become colorized. After passing through the analyzer, a color image is finally created through human eye spatial synthesis. In a commercial 10-bit LCD, the intensity of each subpixel (i.e., R, G, and B subpixels) can range over 1024 shades of gray. Thus, this 10-bit LCD combining

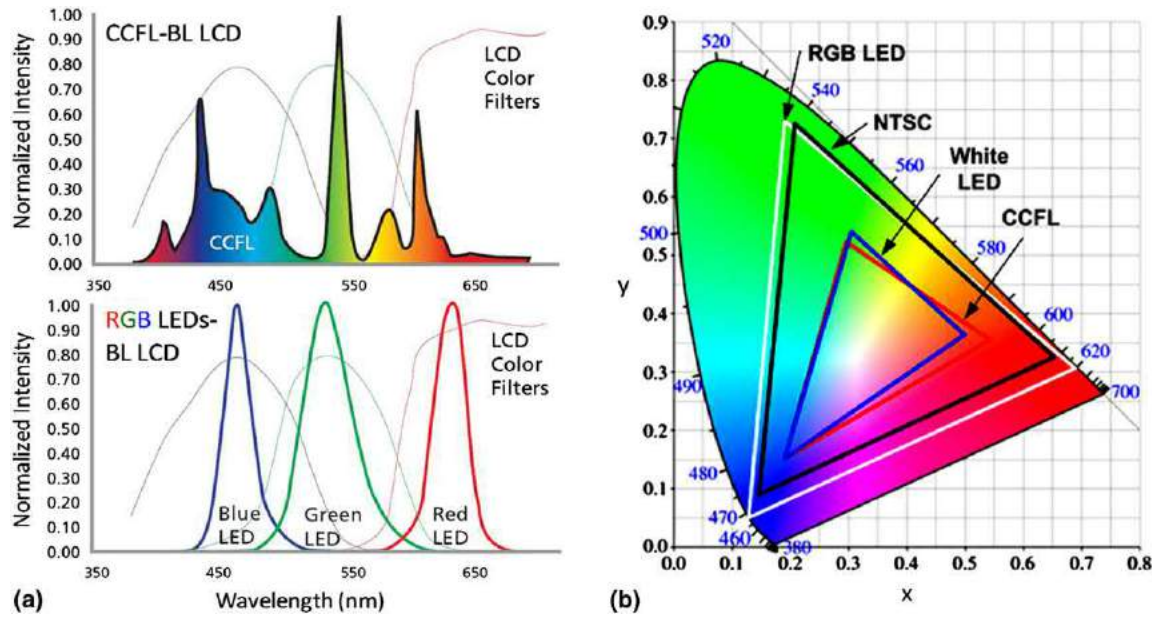


Fig. 7 (a) Typical color filter transmittance spectra with illuminant spectra of a CCFL backlight (top) and an RGB-LED backlight (bottom). (b) LCD color gamut with backlights employing the RGB LEDs, CCFLs and white LEDs (LED, 2017). An RGB-LED backlight has narrow emission spectra to fit the spectra of color filters of an LCD, resulting in less color cross-talk among different color filters. Thus, an LCD using the RGB-LED backlight reaches larger color gamut than that of using CCFL or W-LED backlight.

different intensities of R, G, and B subpixels can present a palette with 1.07 billion colors (1024 shades of red $\times$ 1024 shades of green $\times$ 1024 shades of blue).

LEDs are praised for several advantages, such as being mercury-free, having the small form factor, a long lifetime, high power efficiency, wide color gamut, and fast switching operation. Therefore, LEDs are being applied to the LCD backlight system and not only save power but also enable the panel lighter and thinner. Due to the fast response of LEDs, it is straightforward to switch the LEDs backlight on/off, thus the advanced technologies of scanning backlight driving (ClearLCD, 2017), high dynamic range (Seetzen *et al.*, 2003), and field sequential color (Hasebe and Kobayashi, 1985) are swiftly developed resulting from this new option for backlight systems.

Although the LCD has been applied in wide applications, some drawbacks, such as low optical throughput and limited color gamut, are still required to be overcome. The generated light propagates through an optical stack comprised of a set of polarizers, color filters, and diffusers. Substantial proportions of light are absorbed and scattered by these components. Overall, approximately only 7% of backlight reaches the front-of-screen image, more than 90% of the backlight is consumed. Additionally, the spectra of CCFL and W-LED widely covers all visible wavelengths (e.g., 400–700 nm), and the transmittance spectra of color filters are not pure peaks at red, green, and blue, as shown in Fig. 7(a). Consequently, the primary colors perceived by human eyes in a conventional LCD are often made up of several different frequencies resulting in color gamut shrinkage (Power Electronics, 2017). Typically, the color gamut of a traditional TFT-LCD using the CCFL or W-LED backlight is 72% NTSC only. Compared with a CCFL or a W-LED backlight module, an RGB-LED backlight has proper and narrow emission spectra to fit the spectra of color filters of an LCD. Thus, an LCD using the RGB-LED backlight reaches larger color gamut, typically as high as 120% NTSC, as shown in Fig. 7(b).

The issues of limited color gamut and low light efficiency are attributed to use wide-band CCFL/W-LED backlight and use color filters absorbing roughly 70% backlight intensity. As the world is increasingly aware of its ecological footprint and the harm that is causing to our planet, there is a great incentive to reduce the power dissipation of state-of-the-art LCDs. Hence, Hasebe and Kobayashi (1985) proposed a temporal color synthesis, field sequential color (FSC) technology, to present a vivid color image without using color filters. Using R, G, and B LEDs which are a narrower spectrum and near point light source to replace a conventional CCFL and W-LED backlight modules. The color gamut of an FSC-LCD using RGB-LEDs as a backlight easily reaches more than 120% NTSC.

Temporal Color Synthesis

Field sequential color LCDs (FSC-LCDs) without color filters was proposed to enhance light efficiency and color gamut. Differing from the spatial color mixing of the conventional color-filter type LCD, the R, G, and B fields for each frame in an FSC-LCD are sequentially and rapidly displayed. The three-primary system requires a minimum refresh rate of 180 Hz to prevent luminance flicker. The human eye can perceive full-color images by the temporal color mixing mechanism. Because the FSC-LCD gets rid of

color filterless, the advantages are low material cost, three times possible screen resolution, large color gamut, and high light efficiency (low power consumption) (Huang *et al.*, 2011).

However, FSC-LCDs notoriously suffer from color breakup (Jarvenpaa *et al.*, 2005), which occurs when relative velocities between displayed objects and the eyes of the viewer exist. In that case, the constituent color components do not overlap entirely at the retina, and the individual color fields may become visible in the front-of-screen image (Fig. 3(a)). This color breakup phenomenon may occur both during smooth motion pursuit, but particularly during saccadic eye movements which may result in large retinal color displacements. Color breakup artifacts are highly annoying, causing some viewers severe viewing discomfort and in some cases even nausea.

Static Color Breakup

Static color breakup phenomenon happens when observing static images. While humans watch a static image, eyes will move around the image to gather details. During or after a saccade, the static color breakup is perceived. The mechanism of static color breakup is illustrated in Fig. 8 (Jarvenpaa *et al.*, 2005; Yohso and Ukai, 2006). Fig. 8(a) shows three static white bars with a black background showing on an FSC-display. When perceiving the image, human eyes sweep and gaze between white bars with saccadic movements, like the white line in Fig. 8(a). In a conventional FSC-LCD using an RGB color sequence, the display shows red, green, and blue fields sequentially. When eyes move in saccade, the field images of different colors are projected onto the retina separately, as illustrated in Fig. 8(b). The effective factors of static color breakup are saccade velocity, image difference between objects and background, and display filed rate.

Dynamic Color Breakup

The other type of color breakup, dynamic color breakup, happens while perceiving dynamic images (Kurita and Kondo, 2000). Human eyes will pursue a moving object at the same velocity to focus the object on the fovea, as shown in Fig. 9. When a white bar

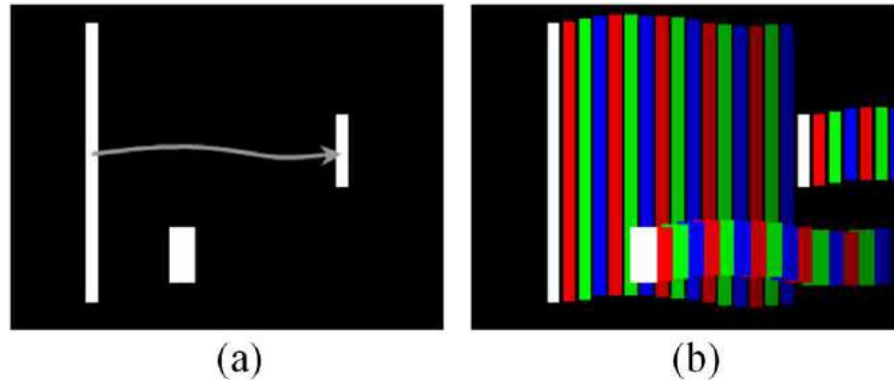


Fig. 8 Mechanism of the static color breakup. (a) Images on an FSC display and a saccade path, and (b) observed color breakup during or just after saccade (Jarvenpaa *et al.*, 2005).

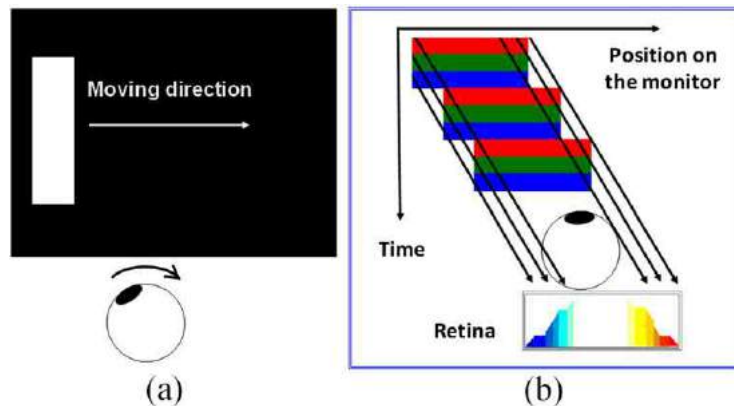


Fig. 9 Mechanism of the dynamic color breakup. (a) Human eyes synchronized with the moving image and (b) the eye-trace integration on the retina. The extra color field will be observed at the bar edge.

is moving from left to right on an FSC-LCD using an RGB color sequence, an observer perceives it through pursuit movement. Fig. 9(b) shows the spatial-temporal relation. The horizontal axis is the position of the FSC-LCD, the vertical axis is time, and oblique black lines indicate the eye-trace line. After eye integration, the three primary colors are projected onto the retina separately and causes the colorful band appeared on the edge of the white bar. From the observing mechanisms of these two types of color breakup, static color breakup is more severe and troublesome than dynamic color breakup because of faster eye velocity.

Color breakup suppression has been implemented on digital light processing (DLP) projectors by inserting additional mono color fields (Langendijk *et al.*, 2006) or increasing the field rate to 540 Hz or more (Mori *et al.*, 1999). Hence, most traditional color breakup solutions for FSC-LCDs involved either increasing the field rate or inserting additional mono-color fields (RGBY, RGBCY or RGBD). The black fields insertion method (360 Hz RGBKKK) has been reported to narrow the color breakup width further to be less perceptible (Koma and Uchida, 2003). Among these three methods, 360 Hz RGBKKK was concluded as the best color breakup suppression method. Although RGB-LED backlights can be switched very rapidly, the response time of liquid crystals prevents the implementation of many of the above methods (e.g., field rate > 300 Hz) for large-screen FSC-LCDs. In spite of the emergence of new fast panel technologies, it is unlikely that affordable panels with frame rates higher than 240 Hz will be widely available in the foreseeable future. In recent years, Lin *et al.* proposed Stencil-FSC (Lin *et al.*, 2009, 2015) and local-primary-desaturation (LPD) (Lin *et al.*, 2011; Zhang *et al.*, 2011) methods to effectively suppress this color breakup issue and make the FSC-technology potential to be practical.

Stencil-FSC:

The “stencil” is a unique technique of painting, and its main concept is to paint the base color first and paint the detail parts in the following. Sometimes multiple stencil layers are utilized for a single image to add colors or create the depth illusion. Additionally, in the human visual system, the luminance contrast sensitivity function is significantly higher than the chromatic contrast sensitivity functions in both temporal and spatial frequencies. It indicates that the visual system is more sensitive to small image changes in luminance contrast compared to chromatic contrast (Fairchild, 2005). Based on stencil concept and the human visual properties, the Stencil-FSC methods incorporate the local color backlight dimming technology. The LC panel performs as a monochrome stencil mask to produce the outline of an image, then the spatially modulated color LED backlight “paints” colors on the mask to create a multi-color field, as shown in Fig. 3(b)–(d). By doing so, the most luminance is gathered in the multi-color field and reduces the luminance of the residual field image to suppress color breakup successfully. For practical applications, the Stencil-FSC methods drive at 240 Hz (four field images), 180 Hz (three field images) and 120 Hz (two field images) field rates to effectively suppress color breakup for large-sized TFT-LCD applications (Fig. 10). Currently, 180 Hz Stencil-FSC has been implemented on a 65-in. color filter-less MVA LCD with 32×24 RGB-LED backlight resolution. Using the IEC 62087 video according to the average of APLs (average picture levels) curve to evaluate display power consumption, 180 Hz Stencil-FSC (52-Watt in average) averagely reduced power consumption to 20% compared with a 65-in. conventional color filter LCD (250-Watt).

However, 180 Hz Stencil-FSC induces green desaturation because redundant red and blue colors are mixed in the green-based color field (Huang *et al.*, 2009). This problem is especially apparent in an image that contains abundant green information, as illustrated by the magenta circle of Fig. 11. For an image, *Lotus*, the first field LC signals are taken from green color; in this case, the blue and red color LC signals are lower than green LC signals. Therefore, even though the red and blue backlight intensities under the purple circles are low, the blue and red lights also propagate through the green-based field and contribute redundant luminance resulting in a reduction of green color saturation. Consequently, a modified algorithm was proposed, 180 Hz min-based Stencil-FSC, the average color difference of CIEDE2000 of the *Lotus* is decreased from 2.9 to 0.3 (Huang *et al.*, 2009).

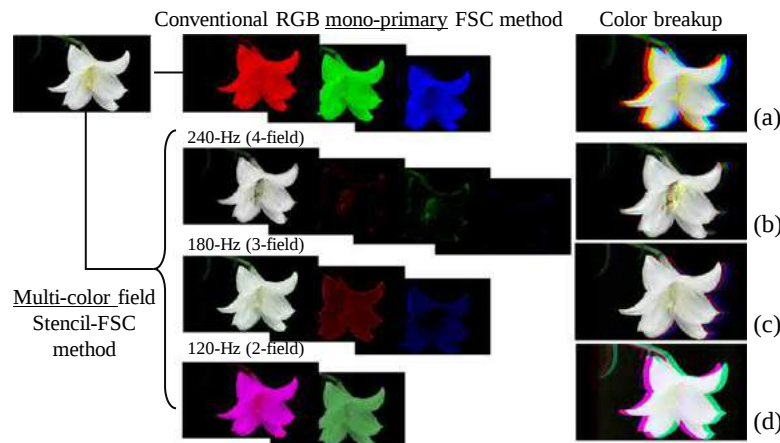


Fig. 10 Schematic plot and corresponding simulated color breakup images of (a) current RGB-FSC method and proposed Stencil-FSC methods in (b) 240 Hz, (c) 180 Hz, and (d) 120 Hz. Stencil-FSC produces a multi-color field image to accumulate most image information, resulting in less intensity of the remaining primary color fields. Thus, color breakup is reduced.

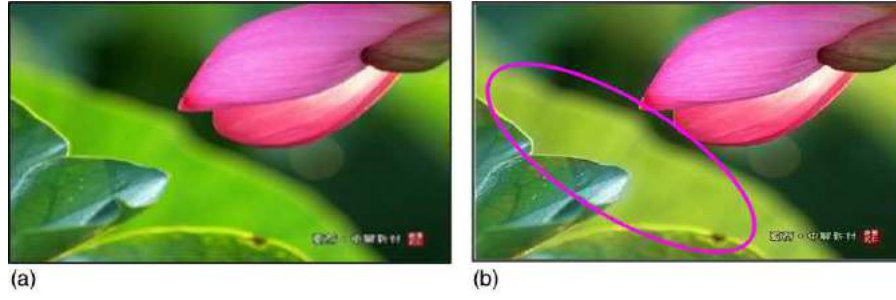


Fig. 11 Reduction of green color saturation using 180 Hz Stencil-FSC. (a) Target image, *Lotus* (provided by Jacky Lee). (b) Simulated image with an average color difference of 2.9 in CIEDE2000 (Huang et al., 2009).

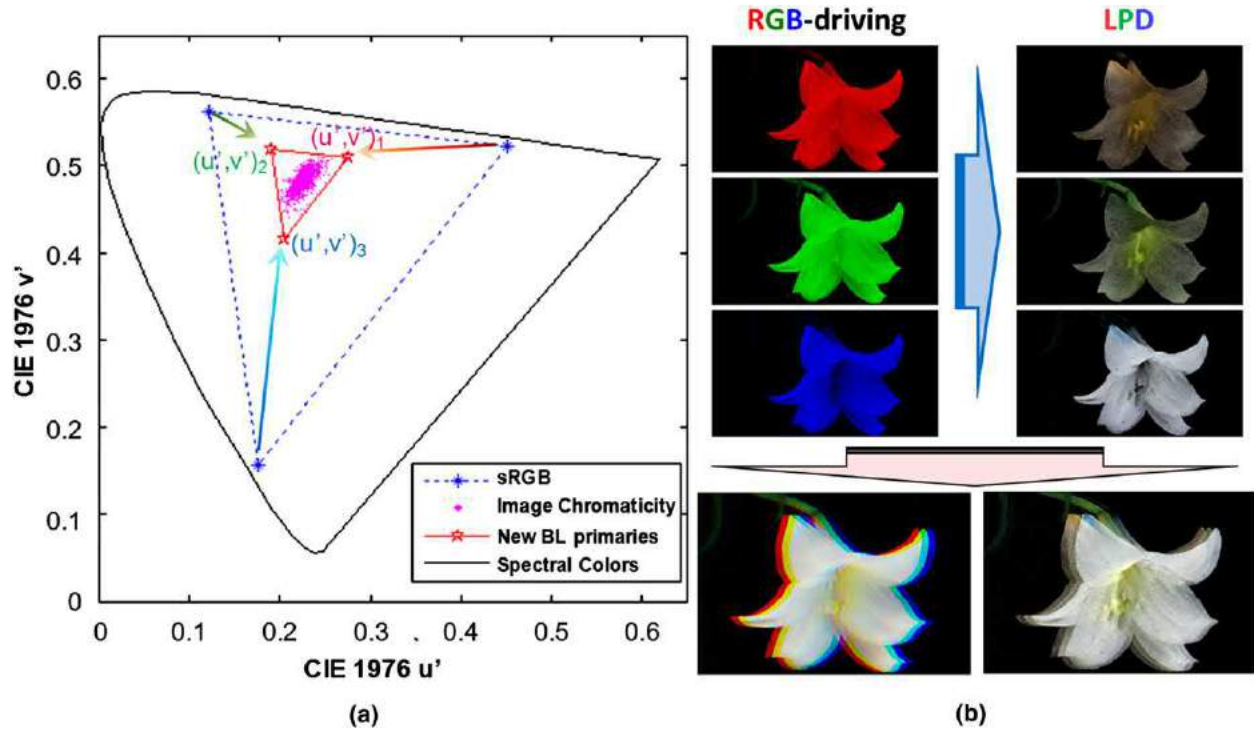


Fig. 12 (a) The color triangles of sRGB and LPD backlight primary colors in the CIE 1976 $u'v'$ uniform chromaticity diagram. (b) The three field images using the conventional RGB-driving (left) and the LPD method (right), and their color breakup images.

However, the intensity of the first field image is not as high as that of green-based one. Thus, the color breakup suppression would be slightly decreased.

To overcome the shortcomings in the previously addressed methods, a novel algorithm, the LPD process, was further proposed to redistribute the original three highly saturated primary color backlight fields, e.g., R, G, and B, into less saturated ones. Color breakup is suppressed with the reduced color differences between each field.

LPD:

Traditional FSC-LCDs utilize the maximum display gamut to display fully saturated R, G, and B field images sequentially at an 180 Hz field rate to create a full-color image. Without considering color contents of a target image, the color breakup phenomenon is very severe, as shown in Fig. 12(b)_left. In contrast, using the LPD method with a field rate of 180 Hz, the RGB-LED backlight module was segmented and independently controlled to provide a small but sufficient color gamut for each input image. The pixel chromaticity distributions of each backlight segment are first analyzed in the CIE 1976 $u'v'$ uniform color space. Next, the backlight primary colors enclosing all pixels chromaticities are dynamically desaturated, as illustrated in Fig. 12(a). By doing so, the color of the three backlight fields are desaturated by mixing another color(s) in the same image field (Fig. 12(b) right). The color differences between each field image are reduced. Thus color breakup is significantly reduced, as shown in the right of Fig. 12(b).

The color breakup reduction concept of Stencil-FSC is to display a multi-color image accumulating most image information in the first single field, thus lowering the intensities of the residual field images. Even human eyes have a fast relative velocity with an object on an FSC screen, the separated color fields are dark, thus suppressing color breakup. The algorithm is simple with only one convolution needed for backlight intensity evaluation. However, the 180 Hz Stencil-FSC method faces a trade-off between image fidelity and color breakup suppression. In contrast, the LPD method utilizes a small but sufficient color gamut to display three new desaturating field images depending on a target image and then reducing color differences between each field images. This is how LPD maintains high image fidelity and suppresses color breakup. However, three convolutions are required to evaluate backlight intensities for three field images when using LPD. This computation complexity causes a high hardware cost to execute LPD. Despite these minor issues, the FSC technique is still highly promising to apply on future display scenarios, such as ultra-low power LCD, light-weighted augmented reality and virtual reality displays, and low heat and high brightness vehicle head-up display.

See also: Displays – Introduction

References

- CIE 170-1:2006, 2017. Cone fundamentals for various field sizes and observer ages. Available at: <http://www.cis.rit.edu/mcsl/online/cie.php> (accessed 01.07.17).
- ClearLCD featuring Aptura™, 2017. Dynamic Scanning Backlight makes ClearLCD™ TV come alive. Available at: <https://wenku.baidu.com/view/d63879126edb6f1aff001f0f.html> (accessed 01.07.17).
- Fairchild, M.D., 2005. Color Appearance Models, second ed. John Wiley & Sons.
- Hasebe, H., Kobayashi, S., 1985. A full-color field sequential LCD using modulated backlight. In: Proceedings of SID Symposium Digest, vol. 16, pp. 81–83.
- Hecht, E., 1998. Optics, third ed. Addison Wesley Longman, Inc.
- Huang, Y.-P., Lin, F.-C., Shieh, H.-P.D., 2011. Eco-displays: The color LCD's without color filters and polarizers. J. Display Technol. 7 (12), 630–632.
- Huang, Y.P., Lin, F.C., Tsai, C.C., Shieh, H.P.D., 2009. Stencil field-sequential color method for color breakup suppression on 180 Hz LCD-TV. In: Eurodisplay Symp. Digest Tech Papers, vol. 29.
- Jarvenpaa, T., 2005. Measuring color breakup of stationary image in field-sequential-color. J. Soc. Info. Display 13, 139–144.
- Koma, N., Uchida, T., 2003. A new field-sequential-color LCD without moving-object color break-up. J. Soc. Info. Display 11, 413–417.
- Kurita, T., Kondo, T., 2000. Evaluation and improvement of picture quality for moving images on field-sequential color displays. In: International Display Workshop, pp. 69–72.
- Langendijk, E.H.A., Swinkels, S., Eliav, D., Chorin, M.B., 2006. Suppression of color breakup in color-sequential multi-primary projection displays. J. Soc. Info. Display 14, 325–329.
- Lin, F.C., Chang, C.W., Huang, Y.P., Shieh, H.P.D., 2015. High-image reproduction by the low-field-rate stencil-LPD method for field-sequential-color LCDs. J. Display Technol. 11 (12), 1069–1075.
- Lin, F.C., Huang, Y.P., Wei, C.M., Shieh, H.P.D., 2009. Color breakup suppression and low power consumption by stencil-FSC method in field-sequential LCDs. J. Soc. Info. Display 17 (3), 221–228.
- Lin, F.C., Zhang, Y., Langendijk, E.H.A., 2011. Color breakup suppression by local primary desaturation in field-sequential color LCDs. J. Display Technol. 7 (2), 55–61.
- Mori, M., Hatada, T., Ishikawa, K., *et al.*, 1999. Mechanism of color breakup in field-sequential-color projectors. J. Soc. Info. Display 7, 257–259.
- Osterberg, G.A., 1935. Topography of the layer of rods and cones in the human retina. Acta Ophthalmol Suppl 13, 1–103.
- Power Electronics, 2017. LED backlighting for LCDs requires unique drivers. In: Power electronics technology. Available at: http://powerelectronics.com/power_management/led_drivers/led-backlighting-lcd-power-efficiency-0512/ (accessed 01.07.17).
- Seetzen, H., Whitehead, L.A., Greg, W., 2003. A high dynamic range display using low and high resolution modulators. In: Proceedings of SID Symposium Digest, vol. 34, pp. 1450–1453.
- UPENN, 2017. Some characteristics of early color vision. Available at: http://www.ling.upenn.edu/courses/ling525/color_vision.html (accessed 01.07.17).
- Webvision, 2017. Simple anatomy of the retina. Available at: <http://webvision.med.utah.edu/sretina.html> (accessed 01.07.17).
- Wikipedia, 2017. CIE 1931 color space. Available at: http://en.wikipedia.org/wiki/CIE_1931_color_space (accessed 01.07.17).
- Yohso, A., Ukai, K., 2006. How color break-up occurs in the human-visual system: the mechanism of the color break-up phenomenon. J. Soc. Info. Display 14 (12), 1127–1133.
- Zhang, Y., Langendijk, E.H.A., Hammer, M., Lin, F.C., 2011. A field sequential color display with local primary desaturation backlight scheme. J. Soc. Inf. Display 17 (3), 242–248.

LCD Components

Fang-Cheng Lin, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

A thin film transistor liquid crystal display (TFT-LCD) is composed of two main structures: an LC panel and a backlight module (Schadt and Helfrich, 1971), as shown in Fig. 1. As liquid crystals (LCs) do not emit light directly, TFT-LCDs must rely on additional light sources (e.g., cold cathode fluorescence lamps, CCFLs or light emitting diodes, LEDs) which are generally behind the LC cell and are called "backlight." The components of an LC panel, a backlight module, and periphery driving are introduced, respectively.

LC Panel

The function of an LC panel is like a light valve to modulate the light intensity from a backlight. The LC panel is basically a sandwich structure with LC molecules filled between two glass substrates: a color filter glass and a TFT-array glass. The top glass, color filter glass, is composed of a polarizer (analyzer) and a color filter layer. In the color filter, each pixel segment is made up of three primary colors (i.e., red, green, and blue) sub-pixels approximate 318 micrometers in size for a 4K-UHD 55-in. LCD. The LCs control the light transmission using different molecular orientations that vary by applied voltages. The bottom glass, TFT-array glass, consists of a polarizer and TFTs. The number of TFTs is three times more than the number of pixels due to the requirement for red, green, and blue sub-pixels. The specific components of an LCD are described as below.

Glass

Two glasses (top color filter-glass and bottom TFT-glass) are used as substrates to maintain the stability of liquid crystals. The requirements of an LCD glass are much higher than house-decoration one. Being capable of withstanding a critical manufacturing environment (i.e., high temperature, strong acid chemicals usage), the LCD glass requires the features of low thermal expansion coefficient, chemically resistant, high surface uniformity. Also, alkali metal oxide cannot exist in the glass to prevent the TFT short circuit. To make a light weight and thin panel, the thickness of each glass would be less than 0.5 mm which is also the mainstream in 2016. Finally, the generation of TFT-LCDs manufacturing what we are familiar with is the size of a mother glass (without fabricating and cutting). The 10th-generation means a 2.85 m × 3.05 m mother glass, as shown in Fig. 2.

Polarizer

Each glass has a linear polarizer film. The one in the TFT glass is used to create a polarized light from a non-polarized backlight; the other one in the color filter glass is used to analyze the polarization state of the light passing through the LC cell to control the light intensity for each sub-pixel. Hence, the LCs can modulate the light polarization and control how much light can propagate through the LC panel. A current polarizer form is made from polyvinyl alcohol (PVA) plastic with an iodine doping (Shurcliff,

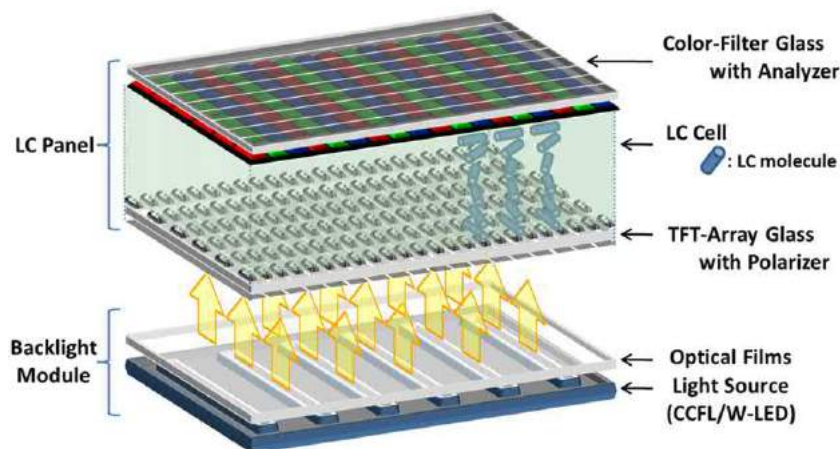


Fig. 1 Traditional LCD structure: an LC panel and a backlight module. The LC panel acts as a light valve to control the transmittance of the light from a backlight source; the backlight module provides the non-self-emitting LC panel with the light source, which is usually CCFLs or white-LEDs.

CORNING

Artwork is showing six 65-inch panels on a Gen 10 glass substrate

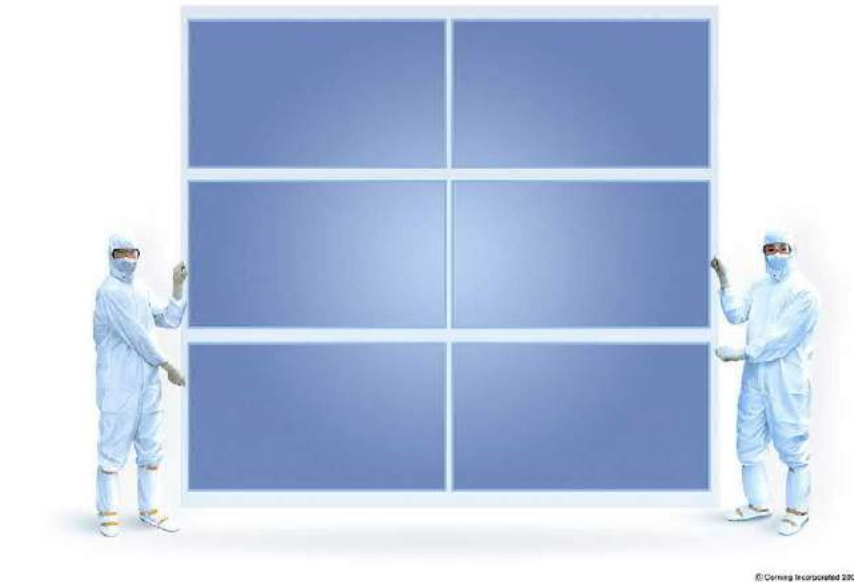


Fig. 2 A 10th-generation mother glass (2.85 m × 3.05 m) presented by Corning which co-located facility with Sharp in Japan in 2007. Available at: <http://mms.businesswire.com/bwapps/mediaserver/ViewMedia?mgid=118723&vid=5&download=1>.

1962; Gunning and Foschaar, 1983). Each side of PVA has a triacetyl cellulose (TAC) layer to increase the mechanical property of PVA. The transmission of a polarizer for LCDs is about 45% or less for visible wavelength.

Transparent Electrode

Transparent electrodes are necessary for the display. The liquid crystal is between the top and bottom electrode glass layers. The backlight needs to propagate through the bottom electrode, color filters, and the top electrode to be displayed. If the electrode is not transparent, it will reduce the aperture ratio, thus reducing the luminance of a display. Indium tin oxide (ITO) is commonly coated on the glass to make a thin and transparent conductive layer in an LCD. The ratio of indium and tin is around 90–10%, and the thickness of ITO layer is typically 100–300 nm. ITO film in the visible light range, the transparency is usually greater than 80%.

Color Filter

Typically, the display is the use of three primary colors (i.e., red, green, and blue) of the additive color mixture to achieve a full-color representation. To precisely show a vivid color image, each pixel has a set of RGB-color filter, names as R, G, and B sub-pixels, as shown in Fig. 3. There is a black matrix in between each sub-pixel to prevent the light crosstalk from neighbor sub-pixels causing a desaturation color. The transmission of the traditional RGB color filter for LCDs is about 30% only. Recently, some companies released 4-primary (RGBY by Sharp) (Yoshida, 2013) and 6-primary (RGBCYM by Panasonic) (Chan and Uei, 2011) color filters to enlarge the color gamut of an LCD. To increase the brightness, RGB with a white sub-pixel, RGBW, was also proposed, especially for a portable display application. Finally, sub-pixel rendering algorithm considering human visual system has also been developed for increasing the apparent resolution of a portable display, such as PenTile technology (Elliott, 2007).

Liquid Crystal (LC)

LCs possess a state of matter that has properties in between isotropic liquid and solid crystal phases. LC phases can be induced by three factors: the temperature variation, the level of solvent concentration, and the ratios of the organic-inorganic composition. In LCD technologies, the temperature variation is the most common used. The finding of LCs is that in 1888, Friedrich Reinitzer, an Austrian botanist, discovered the compound of cholesteryl benzoate has an intermediate state of matter in between lattice structure crystal and isotropic liquid at different temperatures. Next year, Reinitzer sent a letter to Otto Lehmann, a Germany physicist, about his findings. Lehmann used a polarized microscopy to observe this “new phase,” and found its property of birefringence (Δn). Then, they confirmed the existence of this new phase and named it as “Liquid Crystal.” Depending on the geometric

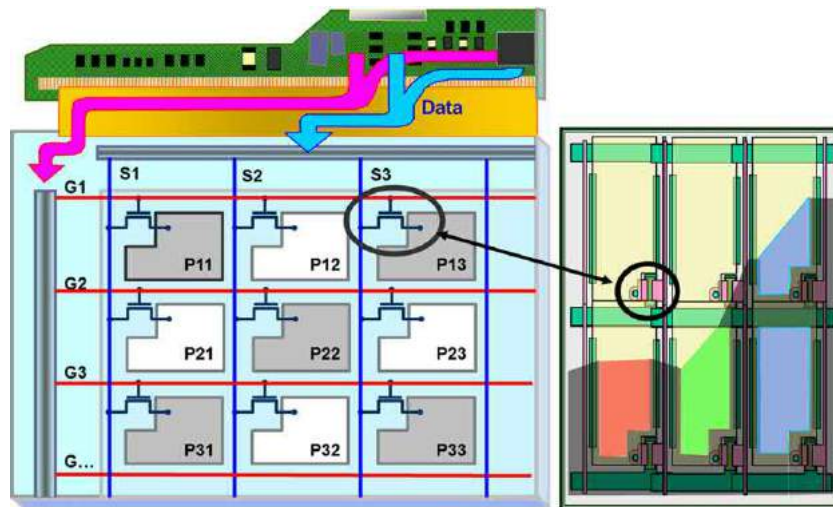


Fig. 3 Schematic diagram of an active-matrix TFT-LCD. The black circles represent a TFT switch in each sub-pixel. The gate lines control the TFT switch on to allow the designed signals in each column sent by the data lines to pass into liquid crystals to control the transmittance of each sub-pixel.

structure of LCs, they are divided into three types of phases– smectic phase, nematic phase, and cholesteric phase (Reinitzer and Chemie, 1888; Lehmann, 1908).

LCs are filled in-between two glass substrates with a cell gap (d). Because LC has optical birefringent property and is varied by different applied voltages, the LC cell can modulate the light polarization status and control the intensity of a backlight. In an LCD, the cell gap is in $\sim \mu\text{m}$ level and uses photo spacers to maintain the cell gap uniformity. The thinner cell gap with a high birefringence index makes the faster LC response, and then a higher frame rate can be achieved to increase the image motion clarity. A fast response LC is required to reach a high frame rate LCD for suppressing motion blur or reaching a field sequential color display. Several research groups have proposed fast response LC modes, such as ferroelectric LC, π cell, optically compensated bend (OCB), and blue phase LC (Meyer *et al.*, 1975; Bos *et al.*, 1984; Yamaguchi *et al.*, 1993; Castles *et al.*, 2012; Chen and Wu, 2014). These modes all can reach a sub-millisecond level but still need more efforts to make them commercialized.

Thin Film Transistor (TFT)

The TFT element is like a switch, and the LC element is like a capacitor. Controlling the TFT switch in on or off state can determine the voltage stored in the capacitor to update or keep the same status. An active-matrix TFT-LCD generally composes of a 2D driving circuit of addressing electrodes. Each pixel has three sub-pixels, and each sub-pixel has a TFT to be an electrical switch and maintain a particular voltage for LC molecules, as shown in Fig. 3. There are three main technologies for TFT manufacturing, including amorphous silicon (*a*-Si) (Powell, 1989), low temperature poly-silicon (LTPS) (Serikawa *et al.*, 1989), and oxide-TFT (Kamiya and Hosono, 2010).

Currently, most TFT-LCD utilizes *a*-Si technology because of high uniformity and low manufacturing cost by its mature process. Because the leakage current is small, *a*-Si is sufficient to be a switch. However, the mobility of *a*-Si TFT is too low ($\sim 1 \text{ cm}^2/\text{V s}$) to be applied on a high-resolution panel. In contrast, using LTPS process, the mobility is the highest ($\sim 100 \text{ cm}^2/\text{V s}$) among these three kinds of TFTs. It is appropriate to integrate circuit on a glass substrate and achieve system-on-panel (SOP), thus making a narrow bezel, high-resolution panel. However, the uniformity is insufficient, and processing cost is greater than others. Thus, LTPS is currently applied on small- and medium-sized panels. For an oxide-TFT, the mobility is in-between *a*-Si and LTPS with high uniformity. Therefore, oxide-TFT is suitable for a high resolution and large-sized LCD panel application. In 2004, additionally, Prof. Hosono proposed InGaZnO (IGZO) to be a TFT material (Kamiya and Hosono, 2010). The process of IGZO is compatible with ITO electrodes and becomes a transparent oxide-TFT or called transparent amorphous oxide semiconductor, TAOS.

Backlight Module

Light Sources

The light sources of a backlight are cold cathode fluorescence lamps (CCFLs) or white light emitting diodes (W-LEDs). Because LEDs are praised for several advantages, such as being mercury-free, having the small form factor, a long lifetime, high power efficiency, wide color gamut, and fast switching operation. Therefore, CCFLs are phased out, and point-like W-LEDs become the mainstream for backlight applications. Using LEDs not only saves power but also enables the panel lighter and thinner

(Sakai, 2004). Additionally, the R-, G-, B-LEDs packed in a single chip have also been presented as a backlight source to increase the color gamut of an LCD. This RGB-LED backlight makes an LCD is more compatible with an active matrix organic light emitting diodes display in image color performance. The advanced technologies of scanning backlight driving, high dynamic range (Seetzen *et al.*, 2003), and field sequential color are swiftly developed resulting from this new option for backlight systems.

The LED backlight of an LCD has two types depending on its placements– bottom-lit and edge-lit or side-lit. In a bottom-lit backlight, the LEDs are directly placed under an LC cell. By doing so, the high dynamic range or local dimming technologies can be applied to enhance the dynamic range of an image and simultaneously reducing power consumption. Additionally, the color saturation is also increased while dynamically and independently controlled RGB-LED signals depending on an input image content. However, this bottom-lit backlight requires an enough space (thickness, from 25 to 40 mm) to make a uniform backlight intensity distribution from a point source (from LED chips) to a uniform plane light distribution. Consequently, this makes an LCD bulky. The bottom-lit backlight used to be applied to a large-sized LCD TV. In contrast, a side-lit LCD places its light sources in one or two edges of an LCD bezel. With an enhanced specular reflector or a light guide, light from the edge is redirected to go into an LC cell normally. Therefore, edge-lit backlight module does not need a thick space to mix the light, thus making whole LCD panel thin (from 1 to 10 mm) and promising for portable display applications, such as smartphones, tablets, and laptops.

Optical Films

To increase the light efficiency and make a uniform light propagate through the LC panel, some optical films needs to be utilized, including the backlight reflector film, enhanced specular reflector (ESR), diffuser, brightness enhancement film (BEF), double brightness enhancement film (DBEF), as shown in Fig. 4 (Yi and Emmons, 2007).

Both bottom-lit and edge-lit types backlight, putting an ESR at the bottom of backlight module is necessary to reflect backlight. While the light is coupled from a light guide, a diffuser can scatter and make the light more uniform. However, the light spreads out and is wasted. To overcome this issue, a BEF, attached to the front surface of a backlight unit, is added to collect the light. Only light with a small incident angle can pass through to LC panel; light with a large incident angle will be totally reflected (total internal reflection) to ESR and redirect back to the BEF. Therefore, the uniform scattering light from a diffuser is collected by the BEF.

The non-polarized light is absorbed about 50% to become polarized light by a polarizer in the bottom-glass. To prevent this 50% waste, a DBEF can be utilized behind the bottom-glass. The DBEF can select the proportion of light which polarization direction is parallel to the polarizer on the bottom-glass (i.e., P1 in Fig. 4) and bounce the perpendicular proportion light without being absorbed and depolarized by the diffusion and scattering. (P2 in Fig. 4). This P2 polarized light becomes P1 after re-bouncing by the ESR and passes through the bottom polarizer. Using the BEF and DBEF can increase the brightness, reduce power usage or length the battery usage.

Quantum-Dot (QD) Film

In addition to the brightness enhancement films, a QD film is proposed to increase color saturation. In general, quantum dots are tiny particles, only several nanometers in size. If the material is made into the scale of the quantum dots, the electrons are easily excited to change the energy level, and then combine with the holes and finally emit light. The wavelength of light emitting energy is proportional to the size of quantum dots. Larger quantum dots emit longer wavelengths resulting in emission

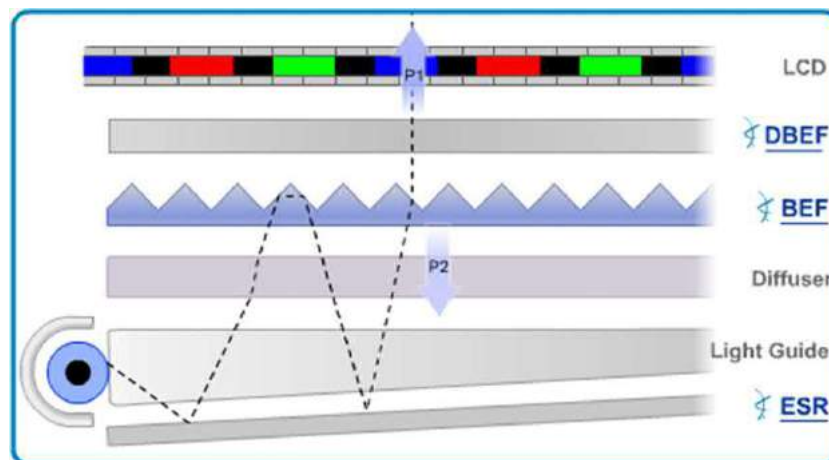


Fig. 4 Optical components of a backlight module from 3M Vikuiti™. When the light coupled from a light guide in an edge-lit backlight module, the optical components are utilized to increase light efficiency and uniformity.

colors such as orange or red. Smaller quantum dots emit shorter wavelengths resulting in colors like green or blue. Therefore, we can manipulate the quantum dot size to stimulate the red, green or blue light. Quantum-dot film (Luo *et al.*, 2014) is utilized to increase image color saturation by narrowing the spectral linewidths of red, green, and blue lights, as shown in Fig. 5. By modulating the size of CdSe, CdS, and CdTe molecules in nano-meter scale, the light spectra are narrowed, thus enlarging the color gamut of an LCD.

Periphery Component

TFT Driver Integrated Circuit

Basic functions of a driver IC (Pappas *et al.*, 2009) are used to drive TFT (gate driver) and supply the corresponding voltage to LCs (data driver). The main circuit structure of the LCD is shown in Fig. 6, including (a) gate driver IC to control TFTs on and off, (b) data driver IC to supply the corresponding voltage to LCs, (c) pulse width modulation to control backlight brightness, (d) DC

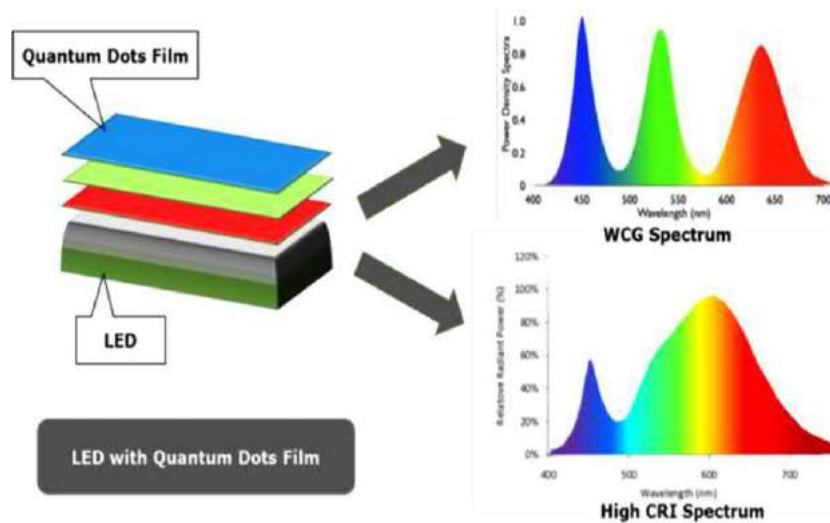


Fig. 5 Quantum-dot (QD) films are utilized in a W-LED backlight system. QD-backlight can reach a wider color gamut compared with W-LED backlight. Data are from 2016 HIS.

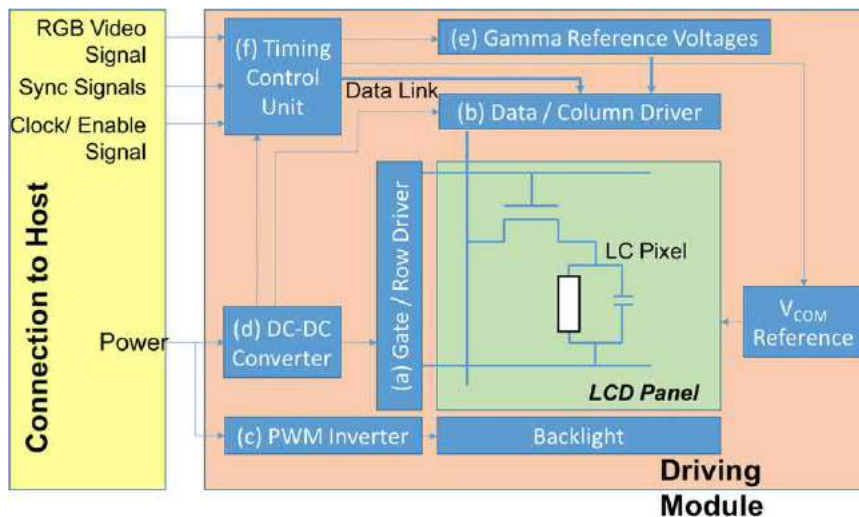


Fig. 6 Driving structure of an LCD, including (a) gate driver IC to switch TFTs on and off, (b) data driver IC to supply the corresponding voltage to LCs, (c) pulse width modulation to control backlight brightness, (d) DC transformer circuit to provide the LCD with the required voltage, (e) gamma reference circuit to correct gamma effect, and (f) timing controller (T-Con) to control the sync signals in horizontal and vertical, and data transmission timing signal. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

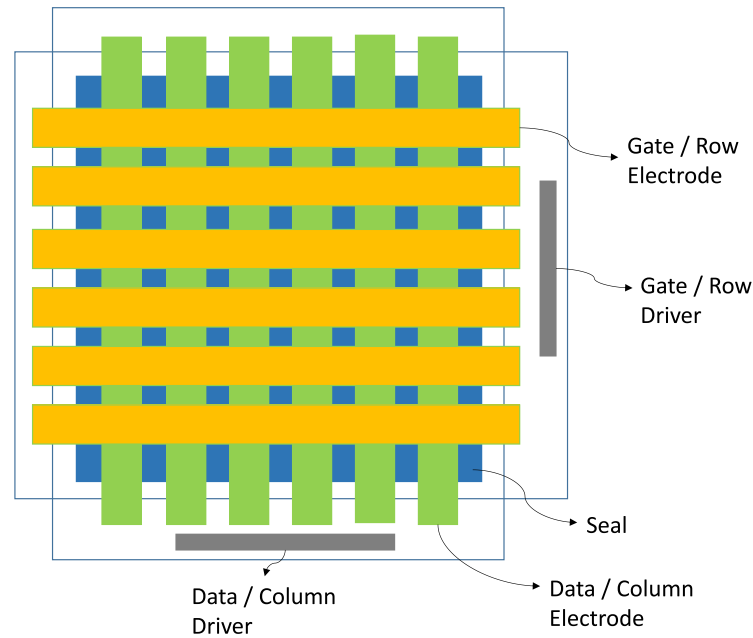


Fig. 7 Passive matrix driving method, which the Y-direction electrode is scanned time sequentially, and the brightness of a pixel is determined by the difference in driving voltage between the X-electrode and the Y-electrode. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

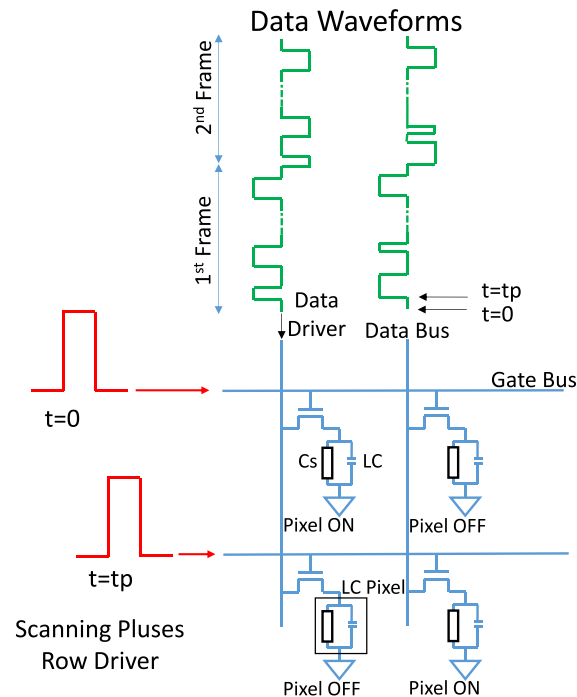


Fig. 8 Active matrix driving method. The data driving circuit sends a signal to each pixel, and a gate driver controls the TFT to open or close. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

transformer circuit to provide the LCD with the required voltage, (e) gamma reference circuit to correct gamma effect because the human eye's sense of light intensity is nonlinear with a gamma 2.2–2.5, and (f) timing controller (T-Con) is used to control the horizontal sync signals and vertical sync signals, and data transmission timing signal.

The TFT driving method can be divided into two kinds of architectures: (1) Passive matrix driving method (PMLCD): A pixel is arranged at the cross of an X- and Y- electrodes, as shown in Fig. 7. The Y-direction electrode is scanned time sequentially, and the

brightness of a pixel is determined by the difference in driving voltage between the X-electrode and the Y-electrode. The disadvantage of a passive matrix LCD is that each pixel has a relatively refreshing rate. Therefore, it is generally applied to a lower resolution display, such as a wearable device.

(2) Active matrix driving method (AMLCD): It usually has two components at each pixel: a switch TFT to control on and off and a capacitor to store the voltage, as shown in Fig. 8. The data driving circuit sends a signal to each pixel, and a gate driver controls the TFT to open or close. When a pixel of the TFT is turned on, then the pixel immediately have a voltage applied to charge the storage capacitor (Cs); when the TFT of a pixel is turned off, then this pixel has no applied voltage, and the LC can maintain its status by the voltage provided by its storage capacitor. This method provides higher response time (refreshing rate) than a PMLCD. Therefore, it is usually used in the high-level products, such as TVs or smartphones.

Trends of LC Displays

Although LCDs are still dominating to industrial, medical, and consumer electronics markets, the AMOLED is occupying the market of small-sized displays, especially in smartphones and AR/VR applications. To make LCDs continue to dominate in these markets, LCDs need to have features of a thinner, lighter, curvable, and flexible profile, lower power consumption/higher efficiency. The display industries also have to enhance LCD performances, such as brighter luminance, higher refreshing rate, higher image resolutions, and a larger color gamut. Therefore, the materials of thinner glasses, fast response LCs, higher TFT mobility and transparency, higher efficiency and color saturation of light sources are worth studying. After having a breakthrough in materials, scientists can implement existing technologies, such as local dimming and field sequential color techniques, on a new product to satisfy the consumers. The display industries also can propose advanced functions to make LCDs more useful, such as dynamically controlling refreshing rate to different operation scenarios. In a touch scenario, the refreshing rate is high to reduce the touching latency and make it more practical; in a reading mode, almost showing a static image, the refreshing rate can be adapted to lower frequency and reduce power consumption, thus lengthening the battery life. Integrating all efforts, the mature manufacturing process of LCDs still provides high quality and cost-effective to produce. Even though AMOLEDs have unique properties, but with featuring these advantages for varieties of display products, LCDs are still promising to hold the share of display markets not only in large-sized but also in small-sized display applications.

See also: Displays – Introduction. Liquid Crystal Physics and Materials

References

- Bos, P.J., Koehler, K.R., Beran, 1984. The pi-cell: A fast liquid-crystal optical-switching device. *Molecular Crystals and Liquid Crystals* 113 (1), 329–339.
- Castles, F., Day, F.V., Morris, S.M., *et al.*, 2012. Blue-phase templated fabrication of three-dimensional nanostructures for photonic applications. *Nature Materials* 11, 599–603.
- Chan, C.-C., Uei, G.-F., 2011. Multi-primary color display and color filter. US Patent 8040027.
- Chen, Y., Wu, S.T., 2014. Recent advances on polymer-stabilized blue phase liquid crystal materials and devices. *Journal of Applied Polymer Science* 131 (13), 40556.
- Elliott, C.H.B., 2007. Color display pixel arrangements and addressing means. US Patent 7307646.
- Gunning, W.J., Foschaar, J., 1983. Improvement in the transmission of iodine-polyvinyl alcohol polarizers. *Applied Optics* 22 (20), 3229–3231.
- Kamiya, T., Hosono, H., 2010. Material characteristics and applications of transparent amorphous oxide semiconductors. *NPG Asia Materials* 2, 15–22.
- Lehmann, O., 1908. Liquids crystals. *Berichte der Deutschen Chemischen Gesellschaft* 41, 3774.
- Luo, Z., Xu, D., Wu, S.T., 2014. Emerging quantum-dots-enhanced LCDs. *Journal of Display Technology* 10 (7), 526–539.
- Meyer, R.B., Liebert, L., Strzelecki, L., Keller, P., 1975. Ferroelectric liquid crystals. *Journal de Physique Lettres* 36, 69–71.
- Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.
- Powell, M.J., 1989. The physics of amorphous-silicon thin-film transistors. *IEEE Transactions on Electron Devices* 36 (12), 2753–2763.
- Reinitzer, F., 1888. Beiträge zur Kenntniss des Cholesterins. *Monatshette Chemie* 9 (1), 421–441.
- Sakai S., 2004. A thin LED backlight system with high efficiency for backlighting 22-in. TFT-LCDs. In: *Proceedings of SID Symposium Digest*, vol. 35, pp. 1218–1221.
- Schadt, M., Helfrich, W., 1971. Voltage dependent optical activity of a twisted nematic liquid crystal (TN-LCD). *Physical Review Letters* 27, 561.
- Seetzen, H., Whitehead, Lorne A., Ward, Greg, 2003. A high dynamic range display using low and high resolution modulators. In: *Proceedings of SID Symposium Digest*, vol. 34, pp. 1450–1453.
- Serikawa, T., *et al.*, 1989. Low-temperature fabrication of high-mobility poly-Si TFTs for large-area LCDs. *IEEE Transactions on Electron Devices* 36 (9), 1929–1933.
- Shurcliff, W.A., 1962. Polarized light production and use. Cambridge, Mass.: Harvard U.P.
- Yamaguchi, Y., Miyashita, T., Uchida, T., 1993. Wide-viewing-angle display mode for the active-matrix LCD Using bend-alignment liquid-crystal cell. In: *Proceedings of SID Symposium Digest*, vol. 24, pp. 277–280.
- Yi, J.R., Emmons, R.M., 2007. Thin direct-lit backlight for LCD display. US Patent 7220036.
- Yoshida, Y., Tomizawa, K., Miyazaki, A., 2013. Multi primary color display device. US Patent 8451302.

Relevant Website

<http://mms.businesswire.com/bwapps/mediaserver/ViewMedia?mgid=118723&vid=5&download=1>
Corning.

Projection Displays – LCD/MEMS/LCoS Based

Fleming Chuang, Coretronic Corporation, Taiwan, Republic of China

© 2018 Elsevier Ltd. All rights reserved.

Projection Display

A projection display system also called digital projector, projects an enlarged image on a screen. Projectors are commonly used in presentations, especially for the large size display, say, more than 100" diagonal. As shown in Fig. 1, a basic projection display includes light source which supplies the light, illumination optics which delivers the light and transforms the light beam into rectangular shape and homogenizes light onto microdisplay, and projection optics to project the image of the illuminated microdisplay and enlarge onto the screen.

The illumination optics might comprises the lens array or light pipe (or rod) and several mirrors, relay lens depends on the type of illumination method is used. In some cases, polarization of the light prior to image generation may be required for LC related microdisplay, say LCD and LCoS.

Light Source

The majority of today's projectors use HID lamp (High-intensity discharge lamps (HID lamps) are a type of electrical gas-discharge lamp which produces light by means of an electric arc between tungsten electrodes housed inside a translucent or transparent fused quartz or fused alumina arc tube) as the most common types of projector light source. In the past few years, the light source of projector move from the traditional projector lamp to solid state light source, such as LED and lasers. These new projector light sources are more efficient, environmentally-friendly and longer lasting than their predecessors.

The HID uses combination of rare earth metal salts and mercury vapor to deliver light.

The critical improvement of light lamp is to minimize the Arc gap of the lamp and increase the power to deliver more light in such a small Arc gap. As the smaller the Arc gap is, as higher the light collection efficiency of illumination optics can be.

Used in smaller projectors (so called, pico-projector or pocket-projector), LEDs are essentially applied. LED light sources are mercury-free and can power off and on instantly. The output lumens for the top line LED projectors are targeting for 3K lumens. The life time of LED projectors can last up to 20,000 h at least. It reaches its maximum brightness in several seconds and requires no cooling down period as shut-down it. In comparison old mercury lamps requires several minutes to boost-up to maximum brightness and it can be damaged if unplugged before the cooling down period ends.

In addition to the light source itself, it normally comprises the light collection optics and prisms as well as optical component needed for homogenization, and in some cases, polarization of the light prior to image generation. The imager module is the active component that generates the image from the incoming light.

In early 2010, Sony announced the development of a highly-efficient RGB laser light source with 10,000 h life-span. Since then, Laser projector became the primary developing direction for projection players. Also respectable is the lasers' considerably higher brightness, better contrast and wider color gamut compared to traditional projector lamp.

Microdisplay

Currently there are three types of microdisplay that are widely used in projection display, namely digital mirror devices (DMDs), transmissive liquid crystal displays (LCDs) panels and liquid crystal on silicon (LCOS). The Microdisplay is a key part of the system since most of the optical limitations are influenced by its properties (dimensions, polarization, acceptance angle, etc.), hence those will affect its design structures.

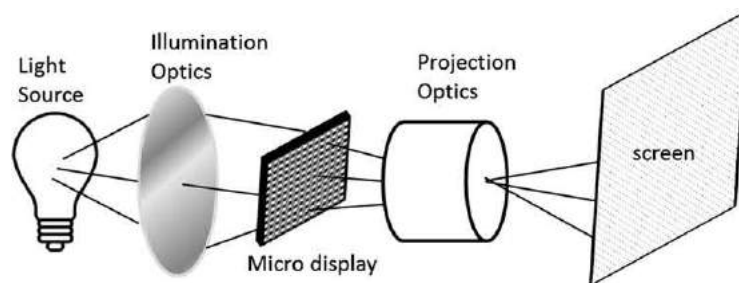


Fig. 1 Basic structure of a projection display.

Projection display – LCD

In 1989, Epson announced the first LCD projector. This projector was developed using three HTPS LCD (HTPS is an abbreviation of High Temperature Poly-Silicon, an active matrix transmissive LCD) panels. The classical structure of 3-LCD projector is shown in Fig. 3 which uses UHP lamp as a light source. The lamp projects white light onto a combination of mirrors that split the light into its three basic video colors. The three color images are combined using a cross dichroic prism (x-prism) (Robinson *et al.*, 2005a) (Cross dichroic prism (X-prism) is combination of 4 triangular prisms. It has function of combining R, G and B three color beams to form the color image) to form a full-color image consisting of millions of colors. The image passes through a projection lens and is projected onto a screen.

The LCD panel applied in projection is HTPS LCD. It has a thin-film transistor (TFT) generated by poly-silicon in each pixel. These pixel transistors act as a conduction switch by changing the scan line's voltage. They are produced in the same way as semiconductors. They are small and highly reliable because they can easily be miniaturized (pixilation or high open area ratio) and drivers can be generated on substrates by processing at a high temperature (Fig. 2).

HTPS LCD panel only transmits longitudinal waves of light. The most popular and effective polarization converter system is a strip half-wave plate. Polarization converter is equipped behind the second lens array (Lens B in Fig. 3). The lens array plate, comprising an array of lenslet, is used to divide light source of individual lenslet into small portion. By passing these focused second light sources through the half-wave plate (Fig. 4), the unified polarization state is formed and the divided light source is superimposed on to the LCD panel by condenser lens and field lens.

Below the projection lens (in Fig. 3), the Cross Dichroic Prism is combination of four triangular prisms. It has function of combining R, G, and B three color beams from respective LCD panels to form the color image. The R/G/B panel should be well aligned and lateral color of projection lens is also required with well control to achieve the crispy pixel imaging.

The 3-LCD projector has three panels to serve for R/G/B separately. Therefore, it does not have possible artifact problems-rainbow effect (Brennesholtz *et al.*, 2008) (The rainbow effect is a visual artifact possible on many single-chip DLP projectors. Because there is only one image chip, color is created sequentially with a rotating color wheel. The sequential color action of slower color wheels can be seen by some people). That principally exists in DLP projector if the frame fresh rate is not high enough. The distinguished advantage of LCD over DLP is the light collection efficiency. LCD can deliver higher lumen output ratio (amount of lamp brightness) in same level color performance and power consumption.

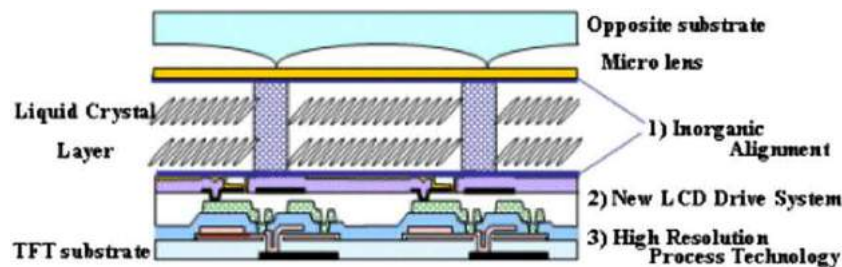


Fig. 2 First HTPS LCD panel using inorganic alignment layer. Available at: <https://phys.org/news/2005-02-world-https-lcd-panel-inorganic.html>.

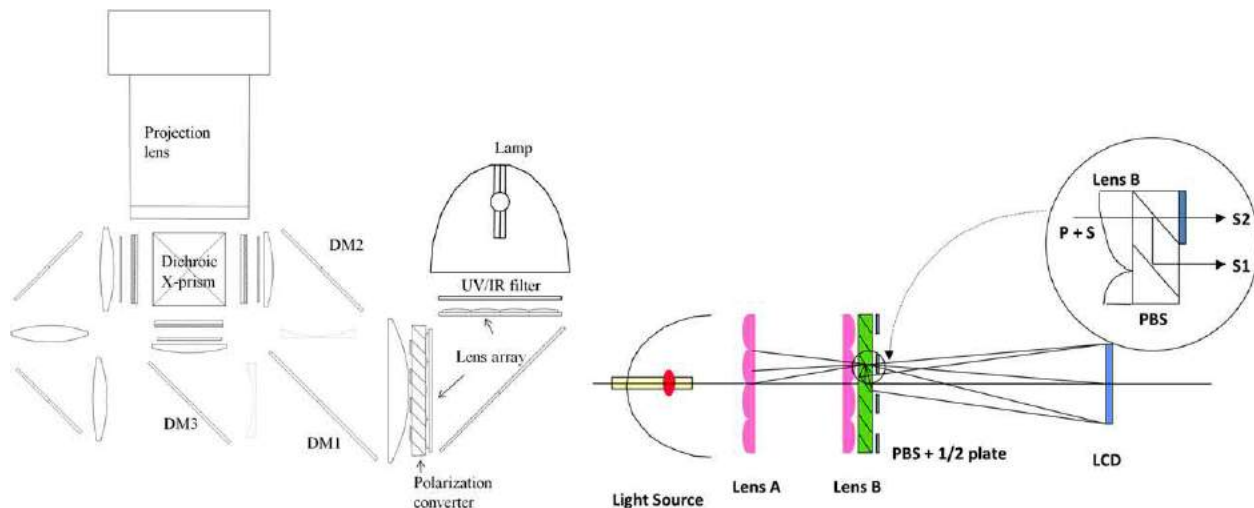


Fig. 3 Structure of LCD projector and Polarization converter (see “Relevant Websites” section).

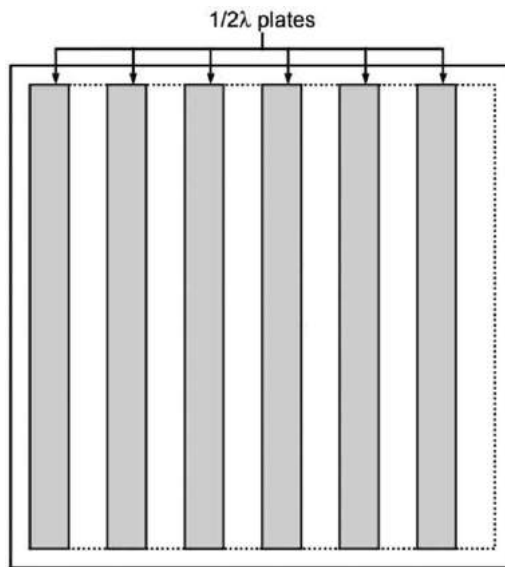


Fig. 4 Stripe Half-Wave-Plate.

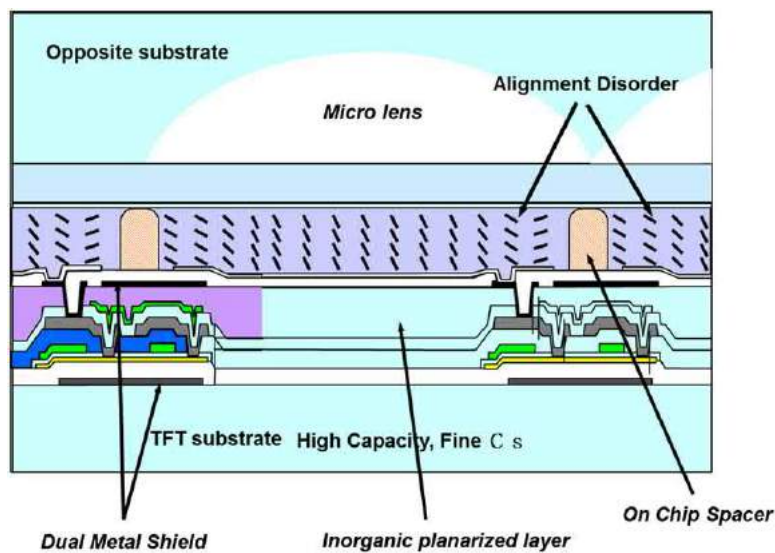


Fig. 5 Typical LCoS structure.

It is known as the screen door effect and resulted in such a sharp image that it was possible to see the individual pixels. With today's advanced resolution and micro-lens equipped in front of LCD panel, the pixels are so small that it becomes harder to notice the screen door effect even though it still exists.

An LCD projector uses three panels that light passes through to yield the full-colored image. Because of this, it is possible for any panel to fall out of alignment which will cause one of the three stacked images to be shifted slightly; this is called convergence. If it happens, then the on-screen image will be slightly fuzzy.

Projection Display – LCoS

Liquid crystal on silicon (LCoS or LCOS) is a miniaturized reflective active-matrix liquid-crystal microdisplay using a liquid crystal layer on top of a silicon backplane. LCoS was initially developed for projection display, but it is also used for wavelength selective switching, structured illumination, near-eye displays and optical pulse shaping, etc. By way of comparison, HTPS LCD projectors are using transmissive LCD, allowing light to pass through the liquid crystal; however, LCoS is reflective LCD. A typical LCoS structure is shown in [Fig. 5](#).

In an LCoS panel, a CMOS chip controls the voltage on square reflective aluminum electrodes buried just below the chip surface, each controlling one pixel with an independently addressable voltage. Typical LC cells are about 1–3 micrometers thick and pixel pitch range from sub-ten microns, as small as $2.79\ \mu\text{m}$ (see “Relevant Websites” section). A common voltage for all the pixels is supplied by a transparent conductive layer made of indium tin oxide (ITO) on the cover glass. The pixels are generally coated with a reflective aluminum layer, and a polyimide alignment layer. The liquid crystal industry can take advantage of the existing silicon technology to produce high resolution microdisplays that is easy and inexpensive to fabricate.

In LCOS, liquid crystals are applied to a reflective mirror substrate. The input light is always modulated into longitudinal polarization light (*s*-wave). The PBS (Robinson *et al.*, 2005a) (Polarizing Beam splitters (PBS) are used to split unpolarized light into two polarized parts. PBS is Beam splitters designed to split light by polarization state) is the most common light switching optical element to control the light in or out. (Fig. 6).

There are methods to increase the contrast of LCoS and LCD projection system. If a PBS is used as the exit analyzer, the isolating QWP (Quarter-wave plate (QWP) consists of a carefully adjusted thickness of a birefringent material such that the light associated with the larger index of refraction is retarded by 90° in phase (a quarter wavelength) with respect to that associated with the smaller index) can work with a retarder stack filter to enhance the ANSI contrast (Robinson *et al.*, 2005a). By applying the Wire grid polarizer into system, it could increase the NA and keeping high contrast at the same time (Brennesholtz and Stupp, 2008; Robinson *et al.*, 2005b).

There are a few typical optical architectures which have been using in the three-panel LCoS projection system. In the system, each panel separately modulates red, green, and blue light. such as, Philips color prism (US Patent 6,330,113), three PBS plus x-prism (Armitage *et al.*, 2006a), three wire-grid polarizer plus x-prism (Armitage *et al.*, 2006b), three PBS plus three lenses (US Patent 7,198,373), off Axis Illumination (Bone, 2001) and ColorQuad (US Patent 6,183,091), etc. You may see the details in the respective references for above diversified architectures. The most popular type is three PBS plus x-prism, shown in Fig. 7. They used dichroic mirrors to separate the three colors, each of which is directed to the corresponding LCoS panel with a separate PBS.

Skew rays are corrected by QWPs between the PBS and the panel. The reflected images are combined into a full color image in an x-prism prior to the projection lens. This design is conceptually simple and allows the optimization of each color channel separately.

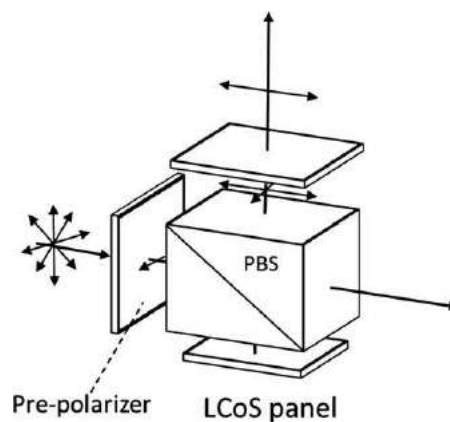


Fig. 6 On and off for PBS and LCoS.

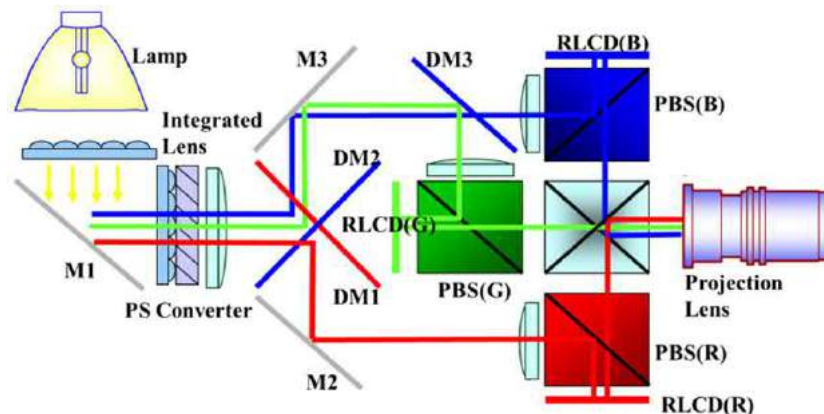


Fig. 7 3-PBS/x-prism LCoS projection system.

Projection Display – MEMS

The DMD™ chip is a micro-electro-mechanical systems (MEMS) array of fast digital micromirrors, monolithically integrated onto and controlled by an underlying silicon memory chip (Hutchison, 2007). (Left of Fig. 8).

The operation principal of DLP projector is shown in Fig. 8 (right). Grayscale patterns are created by programming on/off duty cycle of each mirror, and multiple light sources can be multiplexed to create full RGB color images. Each mirror represents one pixel. The light, rather than passing through the panel, is reflected from it. The mirrors move back and forth, varying the amount of light that reaches the projection lens from each pixel of the primary colors in a rapid rotating sequence to for the true color image.

Optical systems for DLP could be divided into single-panel and three-panel projection applications. (Fig. 9).

Optical systems for single-panel projection applications can be grouped into two main architectures, Telecentric and Non-Telecentric system. Telecentric architectures are defined by locating the exit pupil of the illumination system (entrance pupil of the projection lens) at or near infinity from the device surface. The chief rays of every bundle incident on every mirror then are essentially parallel to each other. For the illumination system, this provides uniform angles of incidence across the entire field, creating uniform black levels for the dark field. Non-Telecentric architectures differ from Telecentric in that the exit pupil of the illumination path is located a short, finite distance from the device, and the entrance pupil of the projection lens must be coincident with it. Some designs use a field lens (or lenses) instead of a prism directly in front of the device to perform the angle separation. The field lens must be on axis with the remainder of the projection lenses, but is shared by the illumination path. This presents some challenges in illumination design since these field lenses also are part of the illumination path, but are off-axis to the DMD and tilted to the illumination path. By designing the illumination pupil near the stop of the projection lens and folding the illumination path there, a very compact system can be designed.

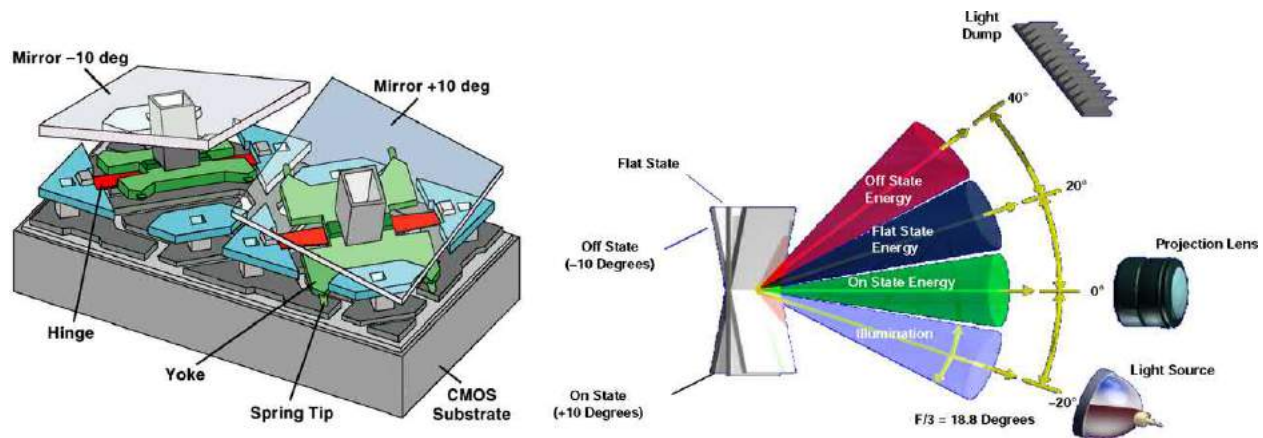


Fig. 8 Structure of DMD chip and simplified optical function. (Courtesy: Texas Instruments Inc.).

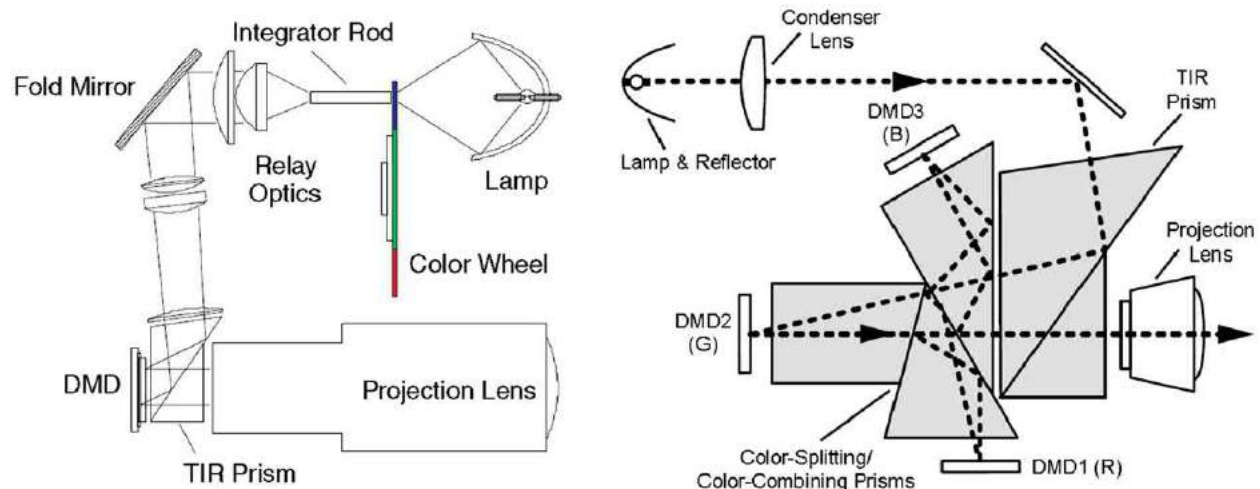


Fig. 9 Generic Optical System: single panel and 3 panels. (Courtesy: Texas Instruments Inc. (Im, H., Kim, W., Chung, J., 2010. A new transmissive type tilting actuator for a smooth projected television image. Proceedings IMechE 224 Part C: J. Mechanical Engineering Science 1327–1333.)).

A three-chip DLP projector uses a prism to split light from the light source, and each primary color of light is then routed to its own DLP chip, then recombined and routed out through the lens. It completely eliminates the “Rainbow” effect of single-chip DLP. 3-Chip DLP Projectors are used for high-performance, high brightness applications in large rooms and large audience venues. 3-Chip DLP technology is currently considered the top of the line technology for digital projection (Im *et al.*, 2010).

New Technologies, Challenges, and Applications in Projection Display Industry

There are many important improvements of the projection display undergoing, no matter what kind of microdisplay is applied. For example, color by applying new solid state light source (LED or Laser), product to meet new HDR standardization, component to improve the dynamic range, image resolution enhancement, UST lens, etc.

4K × 2K: True 4K and 4k by Smooth Picture

As the most advanced digital projection technology for cinema, with 4k digital cinema solutions deployed, DLP and LCoS have their own products. Both are based on 3-chip solutions. By utilizing three 1.4" 4K DMD chips, the manufactures could deliver the Cinema projector's brightness as high as 20K lumens. JVC and Sony have revealed its new native 4K HDR projector, which features a newly-developed laser light source with a claimed brightness of 3000 lumens. However, the price of true 4K 3 LCoS projector is very high no matter DLP or LCoS.

A pixel shifting technique which focuses on producing a smooth picture on the lower resolution panel, DLP and LCoS have their own approaches. So called e-shift by JVC, the operation principle is explained how they work (in Fig. 10). In this diagram at the left-corner of figure, the refractive index of the e-Shift device is changed through use of an electric image shift-switching signal. Following, the path of light emitted by the light sources (such as the D-ILA display device) changes, following the changed refractive index. By changing the path, one pixel is shifted diagonally (vertically and horizontally). By shifting pixels, the vertical and horizontal resolution is substantially doubled. It shows “quadruple” the amount of effective pixels, but it cannot be addressed that small portion of the pixel to display unique information. It has to be the entire pixel and when you put together the two sub frames the entire pixel from the non-shifted and shifted image have to be color coordinated to look like the original image.

JVC has 4K e-shift D-ILA front projector by three full HD resolution D-ILA devices. For the DLP camp, so called “Smooth Picture” technique is developed for the similar pixel shift purpose as well (Hutchison, 2007). For example, the transmissive type of Smooth picture module is operated by vibration to tilt the light from the DMD of optical system (equipped between projection lens and DMD). A voice coil motor (VCM) type actuator for smooth picture has to satisfy performance requirements such as higher driving force, lower power consumption, and lower cost. In updated development (Im *et al.*, 2010), TI developed the 0.67" 2.5K × 1.6K DMD for single chip projector. By applying smooth picture module, the high quality 4K picture is achieved. Several companies have introduced the 4K DLP projectors into market.

The e-shift device would also be available for reflective type (Hutchison and Neal, 2008) instead of transmissive type mentioned above. Compared to transmissive type, it requires a very good quality mirror and more precisely tilt angle control.

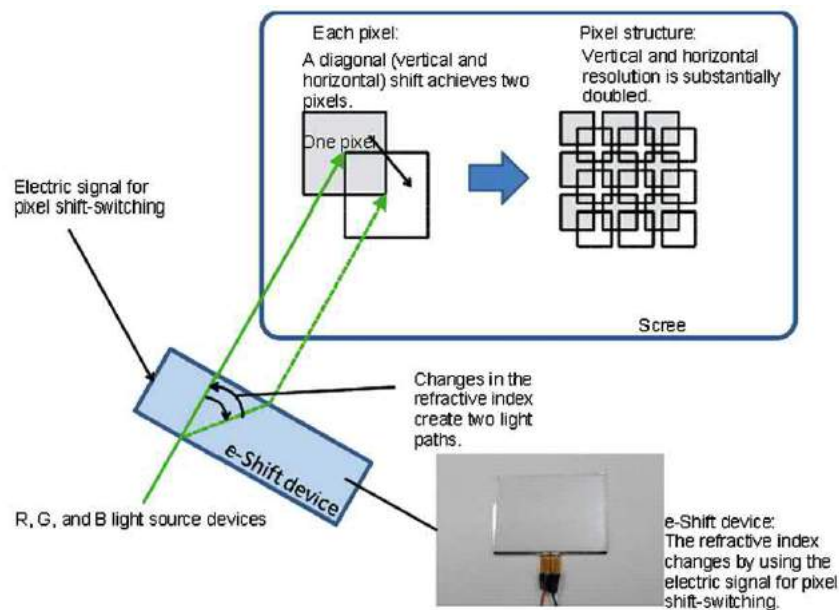


Fig. 10 Pixel shift-switching using the e-Shift device.

Ultra-Short-Throw

The terminology, Throw Ratio, refers to the ratio of the distance to the screen (throw) to the screen width (Fig. 11). A larger throw ratio corresponds to longer distance of a projector from the screen. For an Ultra-Short-Throw (UST), it is very short distance the image is thrown onto the screen, and it has a large effect on screen size. Often in home projector and individuals lacking the space in the room, projector to equip with an ultra-short-throw projection lens is quite popular in the market. As the throw ratio is as below as 0.35, we could call it UST. The critical issue for the projection lens design is that it requires a very large field of view; therefore, design and manufacturing become challenging.

Some special design techniques were developed in UST projector. In popular lens structure of projection system, the stop is positioned to make the system Telecentric or non-Telecentric to match the light beam of illumination system. It is always located close to focal point of the rear part (the lens elements at right of stop) of projection lens. Generally, multiple glass and plastic lenses are combined and large-diameter lenses are used to achieve the target performance. Thus, the wider the viewing angles, the larger the lenses and the more the number of the lenses. The requirements to widen the viewing angles of a projector are inversely proportional to reducing the projector's size and weight causing the manufacturing difficulties, especially; the aspherical surface is mostly applied for reducing the aberrations of lens system.

Some optical structures applied the reflecting mirror at most front position of projection lens was developed. There are two types of mirror: convex mirror and concave mirror. To reduce the size of lens element, forming the intermediate stop or intermediate image between the mirror and front lens element is also designed by many developers. Some distinguished design examples from the published patent are described below.

US Patent 7,009,765 (Fig. 12): One intermediate image and one intermediate stop to reduce the size of front lens group. No mirror is applied.

US Patent 20130162957 (Fig. 13): No intermediate image or stop applied, using convex mirror. The mirror and first lens will be large.

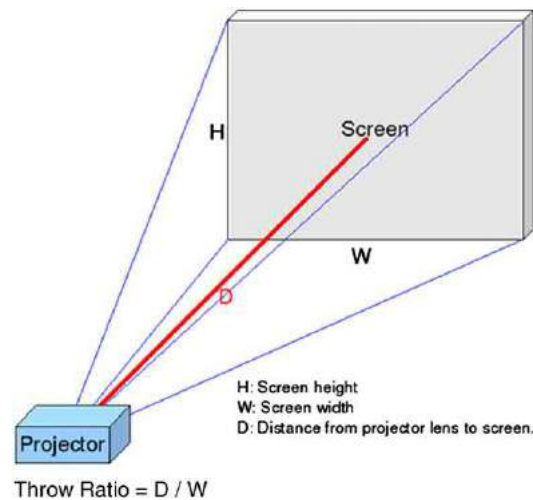


Fig. 11 Definition of Throw Ratio.

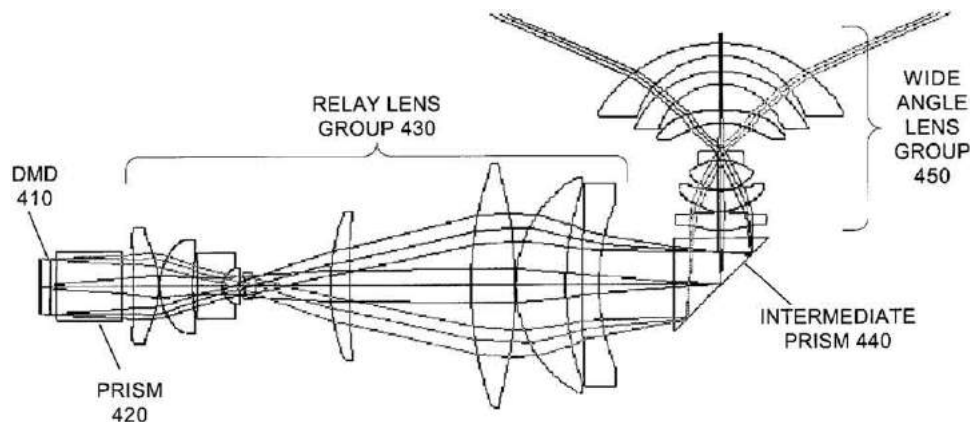


Fig. 12 US Patent 7,009,765B2, Assignee: Infocus.

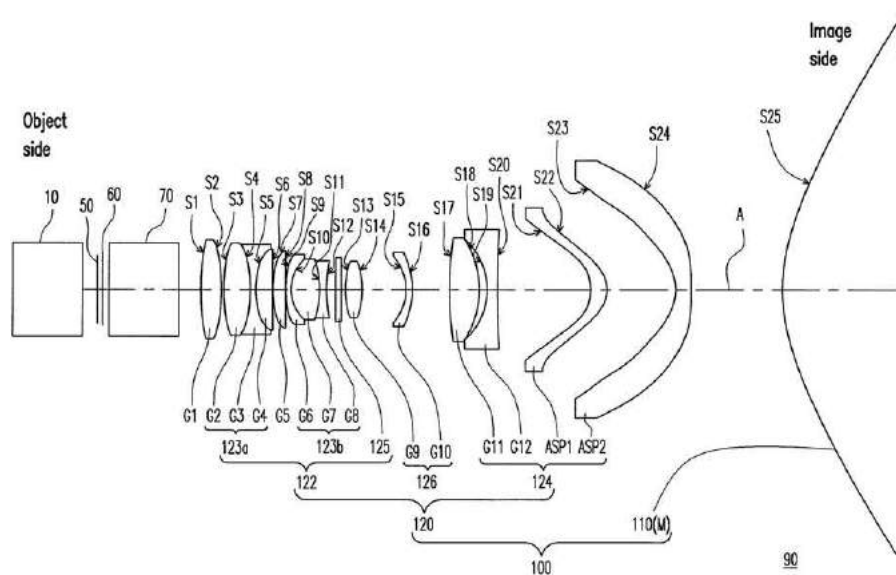


Fig. 13 US Patent 20130162957 A1, Assignee: Young Optics.

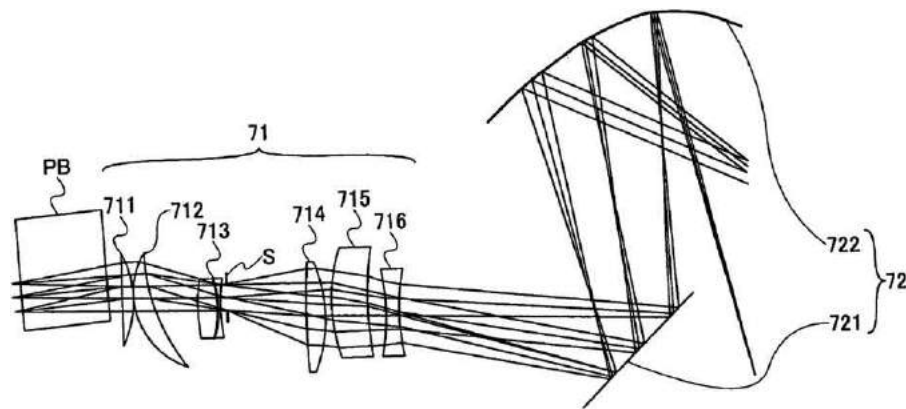


Fig. 14 US Patent 7441908, Assignee: Ricoh.

US Patent 7441908 (**Fig. 14**): One intermediate image is formed in front of concave free-form mirror.

US Patent 9,372,388 (**Fig. 15**): two intermediate image, one intermediate stop, and using concave mirror. This structure has the most compact reflection mirror and lens element.

Laser Phosphor Illumination Projector

Pure laser, also known as RGB laser, generates light directly from three individual red, green and blue lasers. The primary benefit of a RGB laser system is light output while also achieving higher performance in other standard image quality parameters such as color gamut, contrast ratio and dynamic range when compared to standard lamp-based systems. As such, RGB laser is ideal for largescale applications and giant screen cinema.

Laser phosphor illumination projector (**Fig. 16**) uses blue laser diodes as the light source instead of a high intensity discharge lamp to generate the three primary colors in a single DMD laser phosphor projector, the blue laser diode focuses laser light onto a phosphor wheel (Laser phosphor is a lampless projection illumination platform that uses blue laser diodes as the primary light source. To generate the three primary colors – red, blue, green – the blue light from the laser diodes shines onto a spinning wheel that is coated in a phosphor compound) to create red and green light, while blue laser light passes through an opening in the phosphor wheel. These red, green, and blue colors are then directed onto SLM chip, which directs the light through a lens and onto the projection screen. The basic concept is shown as **Fig. 16**. In a laser-phosphor light source, only uses the less expensive blue lasers, that can be made more cost-effective compared to pure laser projectors.

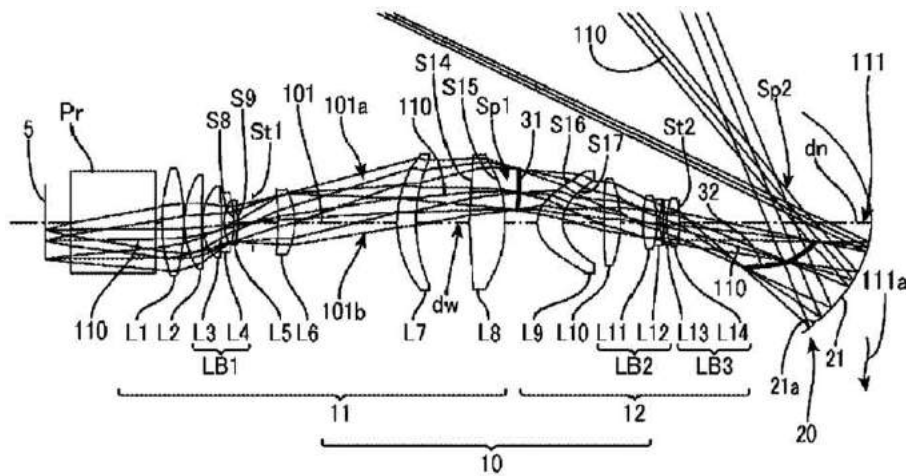


Fig. 15 US Patent 9,372,388 B2, Assignee: Nittoh K.K.

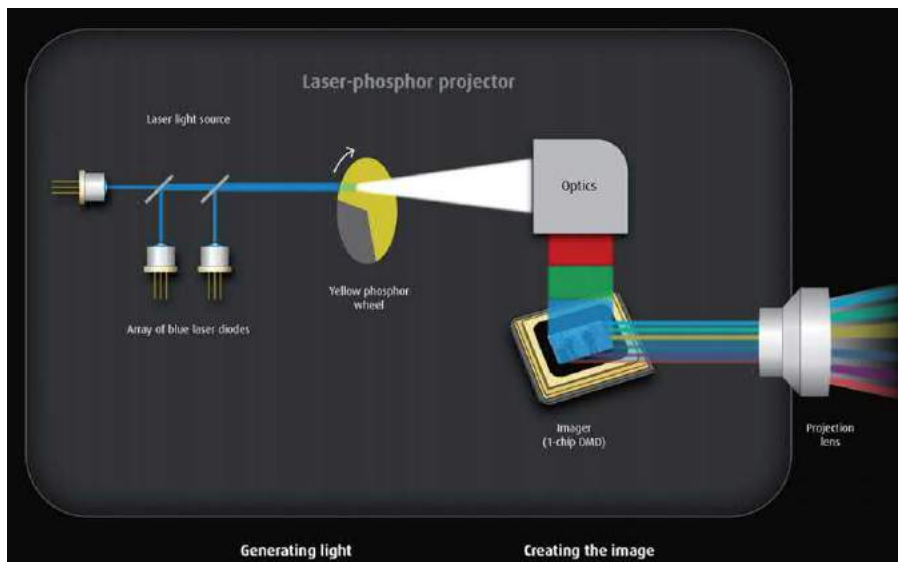


Fig. 16 Laser-phosphor projector (a single-chip projector) (Barco).

The primary advantage of a laser phosphor projector is the long life of the illumination system. As a lampless system, laser phosphor also eliminates the need for lamp and, in many designs, filter replacements, reducing the down-time, maintenance and costs associated with lamp-based projectors.

There are diversified optical structures in designing the laser phosphor projector. The phosphor wheel could be reflecting type (Fig. 17) or combined with color filter to make structure compact, like below patent application (Fig. 18).

Color Gamut and HDR

Displays cannot reproduce all the colors occurring in nature. HDR content preserves details in the darkest and brightest areas of a picture that are lost using current standards. It also allows for more natural, true-to-life colors that are closer to how we see them in real world. High Dynamic Range display provides viewers with an enhanced visual experience by providing images that have been produced to look correct on brighter display, that provide much brighter highlights, and provide improved detail in dark areas. Extension of the color gamut, necessary to realize more realistic color reproduction, is possible achieved by changing the chromaticity of the three primary colors. DCI-P3 describes a wider color gamut than Rec. 709 with three primary colors and is implemented for digital cinema based on xenon arc discharge lamps. Rec. 2020 reproduces images more realistically and includes the gamut of the major colorimetric standards but requires quasi-monochromatic light sources such as RGB lasers. The chromaticity diagram (CIE 1931) of Rec. 709, DCI-P3, and Rec. 2020 are shown in Fig. 19. The HDR regulation for HDR10

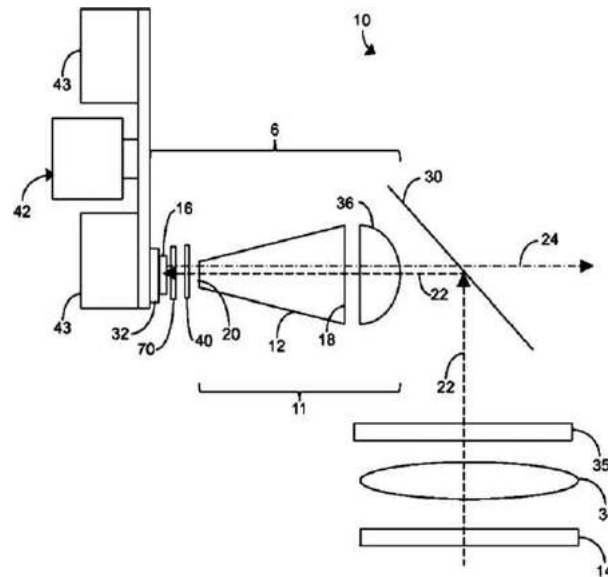


Fig. 17 US Patent 9,170,475: Light valve projector with a laser-phosphor light converter.

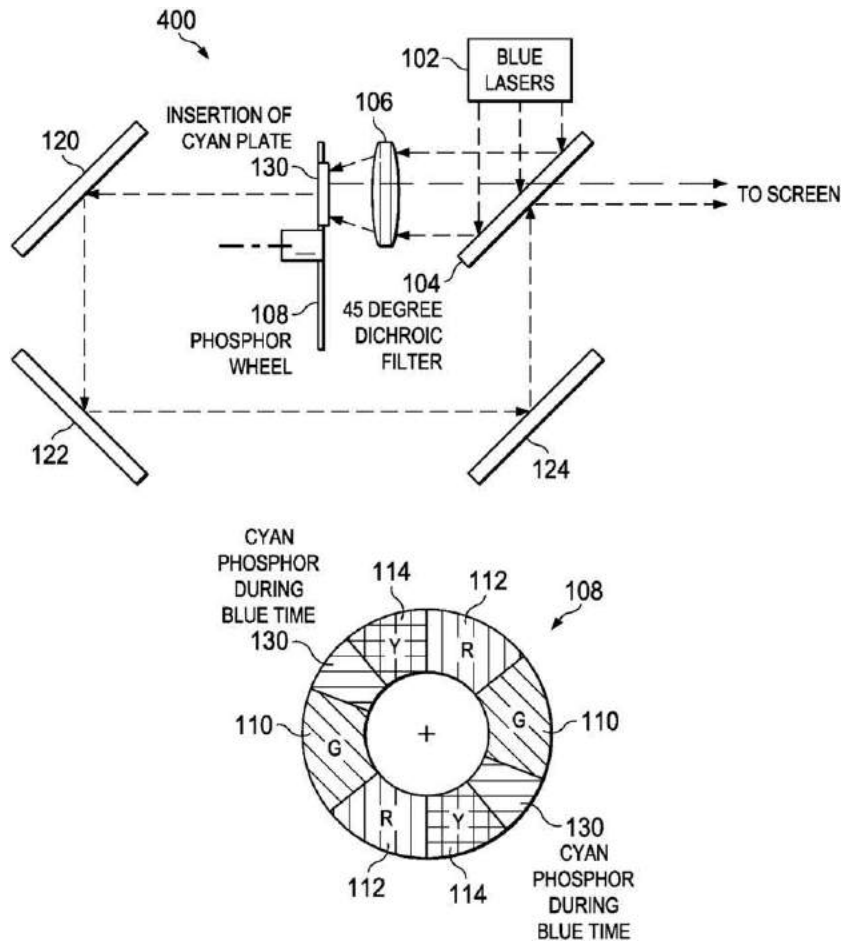


Fig. 18 US Patent 2014/0253882 A1: Hybrid laser excited phosphor illumination apparatus and method.

(see “Relevant Websites” section) and Dolby Vision (see “Relevant Websites” section) could be found in reference. Solid state light sources are gradually becoming the preferred light sources for projection displays instead of the traditional (broadband) short-arc xenon or ultra-high performance (UHP) discharge lamps.

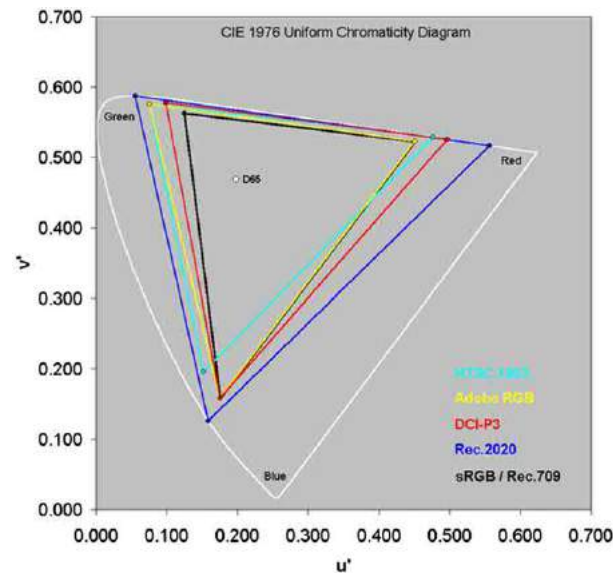


Fig. 19 Standard Color Gamut. Plotted on a CIE 1976 Uniform Chromaticity Diagram.

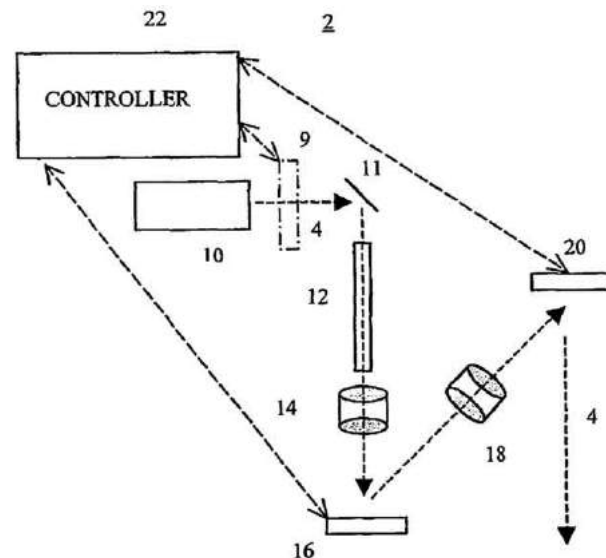


Fig. 20 US Patent 7224335 Assignee: IMAX DMD-based image display systems.

Dynamic irises (Laser phosphor is a lampless projection illumination platform that uses blue laser diodes as the primary light source. To generate the three primary colors – red, blue, green – the blue light from the laser diodes shines onto a spinning wheel that is coated in a phosphor compound) first came into general use with LCD projectors as a means for improving on this projector's relatively low native on/off contrast ratio and elevated black levels. Dynamic irises can now be found in many home theater projectors using DLP, LCoS and 3LCD digital display technologies including those with a relatively good native contrast ratio and low black levels.

When SLM based projection systems are to be used to meet the HDR regulation, it is desirable that the system has a high spatial resolution and that the system possesses a wide dynamic range image quality. SLM based systems such as those based on DMDs are typically limited to contrast ranges of 1000: 1 or less. IMAX proposed invention dual SLM structure is directed towards a projection image display systems that employ so-called reflective spatial light modulators (SLMs) such as those in the form of digital micro-mirror devices (DMDs). It relates to techniques for using multiple SLM devices arranged serially. The invented system (**Fig. 20**) comprises a controller, a first reflective SLM and a second reflective SLM. The first reflective SLM is operatively coupled to the controller to receive first image data from the controller and is aligned to receive light from a light source and to reflect imaging light correlated to the first image the second reflective SLM is operatively coupled to the controller to receive second image data from the controller and is aligned to receive imaging light from the first reflective SLM and to reflect modified imaging light.

Assume DMD has 100% optical efficiency and supports n bit depth, if the first DMD maintains 2^k equally spaced brightness levels, a serial DMD device should have a dynamic range performance equivalent to $k + n$ bit depth, at least within the lowest level (US Patent 7224335). Hence, by applying dual-DMD, the contrast ratio could meet the HDR regulation by well design.

See also: Displays – Introduction

References

- Armitage, D., Underwood, I., Wu, S.-T., 2006a. Introduction to Microdisplays, p. 327.
 Armitage, D., Underwood, I., Wu, S.-T., 2006b. Introduction to Microdisplays, p. 328.
 Barco. Laser-phosphor illumination in projectors, White paper. Available at: <http://www.barco.com/en/Page/2016/Entertainment/Downloads/Laser-phosphor-illumination/Downloadpage>.
 Bone, M., 2001. 46.2: Method to improve contrast in an off-axis projection engine. SID Symposium Digest of Technical Papers, 32: 1180–1183. doi:10.1889/1.1831770.
 Brennesholtz, M.S., Stupp, E.H., 2008. Projection Displays, p119.
 Hutchison, D.C., Neal, H.W., 2008. The design and implementation of a stereoscopic microdisplay television. IEEE Trans. Consumer Electron. 54, 254–261.
 Im, H., Kim, W., Chung, J., 2010. A new transmissive type tilting actuator for a smooth projected television image. Proceedings IMechE 224 Part C: J. Mechanical Engineering Science pp. 1327–1333.
 Robinson, M.D., Sharp, G., Chen, J., 2005a. Polarization Engineering for LCD Projection, p. 187.
 Robinson, M.D., Sharp, G., Chen, J., 2005b. Polarization Engineering for LCD Projection, p. 192.
 Hutchison, D.A., 2007. The SmoothPicture™ Algorithm: An Overview, Texas Instruments, DLP™ TV.
 US Patent 6,183,091 B1, Color imaging systems and methods.
 US Patent 6,330,113, Color separation prism with adjustable path lengths.
 US Patent 7,198,373, B2, Display apparatus using LCD panel.
 US Patent 7,224,335, DMD-based image display systems.

Relevant Websites

- http://www.3lcd.com/explore/polarization_charger.aspx
 3LCD.
<http://www.businesswire.com/news/home/20160510005412/en/World%E2%80%99s-Smallest-Native-4K-Imaging-Device-Introduced>
 Business Wire.
<https://www.dolby.com/us/en/technologies/home/dolby-vision.html>
 Dolby.
<http://www.uhdalliance.org/uhd-alliance-press-releasejanuary-4-2016/#more-1227>
 UHD Alliance.

3D Displays, Stereoscopic/Autostereoscopic 3D

Yi-Pai Huang and Chun-Ho Chen, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction of 3D Display

In recent years, the requirement is emerging for more vivid and realistic 3D experience. The development of 3D display technology comes from the desire of human's nature vision, and results from the maturity of manufactory and display technology. The public usually record and display images for memorizing and sharing activities with others. Therefore, display is the closest electronics product to human. The progress of display development represents the efforts of pursuing more realistic visual experience. When we review the display developing history, human's color perception ability is a strong driving force to replace black-and-white TVs with full color ones. More than one trillion vision cells also encourage us to replace standard definition displays with full high definition ones. In addition to the progress of color and definition, a depth perception is a disruptive innovation of display technology. 3D displays utilize this depth perception to demonstrate the real 3D world (Holliman, 2005). Watching the primary 3D displays needs to wear eye glasses, which is called the stereoscopic type (Hill and Jacobs, 2006). The glasses usually block human's line of sight and do not be assimilated into our life. The scenarios could be limited in movie theaters (Srivastava *et al.*, 2010). Therefore, the glasses-free technology, or called the auto-stereoscopic type, obtains more attention to be applied on billboards, TVs, and mobile phones.

Principle

Our world is a three-dimension space but there is only two-dimension intensity distribution in our retina. However, the human's vision system can transfer this planar intensity distribution to the stereo concept by multi depth cues. Depth cues are classified by psychological and physiological depth cues (Pastoor and Wopking, 1997).

Psychological Depth Cue

Even the objects in the same distance, viewers perceive the closer objects when the objects are more colorful and brighter. Besides, human considers the relative small object is far than the big object in the same image. This is due to the perspective cue as shown in Fig. 1(a). The spacing of two parallel lines seems to decrease when the distance of lines away from the viewer increases. The stereo perception also comes from the variation of light and shadow as shown in Fig. 1(b). In addition, the occlusion cue between objects help determining near and far objects. The near object blocks the viewing of far objects as shown in Fig. 1(c). Moreover, the texture gradient in near objects contains more detail and surface feature on objects infer 3D shape and distance as shown in Fig. 1(d).

Physiological Depth Cue

The crystalline lens in the eye structure, exactly as the focus lens in the camera, can accommodate itself to image the object on the retina. And the viewing lines of two eyes converge on the object. Based on the different depth, this converging angle varies. Near objects have wide converging angles and far objects have narrow converging angles. The average distance between two eyes is 6.5 cm. This distance induces a little difference between the viewing angles of two eyes. Therefore, the image disparity of object on

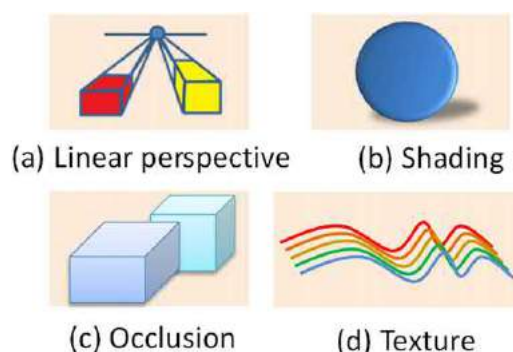


Fig. 1 Illustration of psychological depth cues such as (a) linear perspective: parallel lines converge at distance point on horizon, (b) shading: cast by an object gives a strong depth cue, (c) occlusion: invisible portion of objects behind an opaque object, and (d) texture: surface feature on objects infer 3D shape and distance.

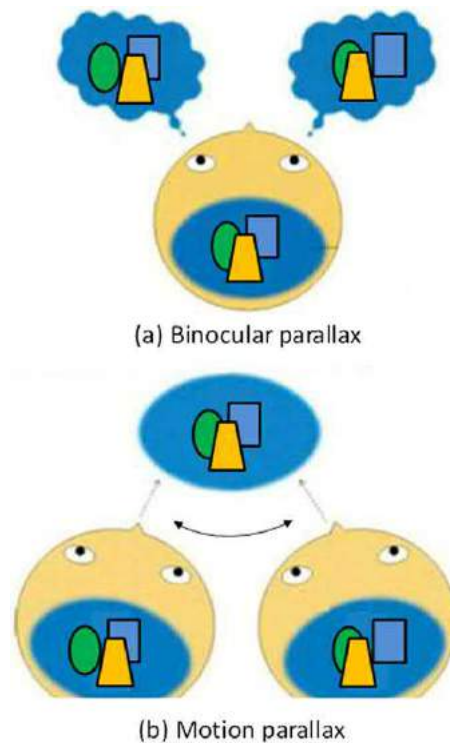


Fig. 2 (a) Binocular and (b) motion parallax provide the image disparity that our brain can merge to create 3D perception.

two retinas exists, called binocular parallax as shown in [Fig. 2\(a\)](#). According to this parallax, human can perceive the difference of object depth. Furthermore, when the viewer's eye moves, the image disparity of object also varies, called motion parallax as shown in [Fig. 2\(b\)](#). 3D displays apply the depth cues of parallax to realize stereo images. If the viewer wants to perceive a clear 3D image, two eyes can only receive individual images with the disparity.

Health Issue– Visual Fatigue

3D techniques use 2D left/right images and depth cue to yield 3D illusion in our brain. In the physiology, how the left/right images are received by the left/right eyes not only affects the 3D effectiveness but also the comfort level in the psychology. For examples, the interference between left and right image can degrade the 3D perception and may induce the visual fatigue ([Pastoor and Wopking, 1997](#)). One issue of 3D display is the accommodation–convergence conflict (A.C. conflict). While the human eye's convergence capability perceives the correct depth of 3D objects, the accommodation function makes the eyes focus on the display screen as shown in [Fig. 3\(b\)](#). Since there is a close interaction between convergence and accommodation, this A.C. conflict often causes visual fatigue in 3D viewers ([Pastoor and Wopking, 1997](#)). The relation between accumulated fatigue and eye disease is worth further exploring in the future. The 3D display revolution is still waiting for more research on physiological and psychological fields.

Stereoscopic and Auto-Stereoscopic Display

From the 19th to 20th century, 3D displays mainly belong to the stereoscopic type in need of special eye glasses. The glasses and other technical issues limit the popularization of 3D displays. By the end of the 20th century, more and more efforts have been made on the auto-stereoscopic display. These 3D displays are introduced as follows.

Polarizing Glass

The pixels of display are divided into odd rows and even rows. Micro-retarder film, arranged with corresponding rows, polarizes light into orthogonal polarization states ([Wu et al., 2008](#)). The viewer wears the glasses with polarization films for individual states. The binocular parallax can be achieved because another polarization will be blocked. The image interference could happen when view's head rotates with the linear polarization. Circular polarization is preferred to avoid this interference ([Jung et al., 2009](#)). The major issue is resolution decreasing, only a half of 2D images.

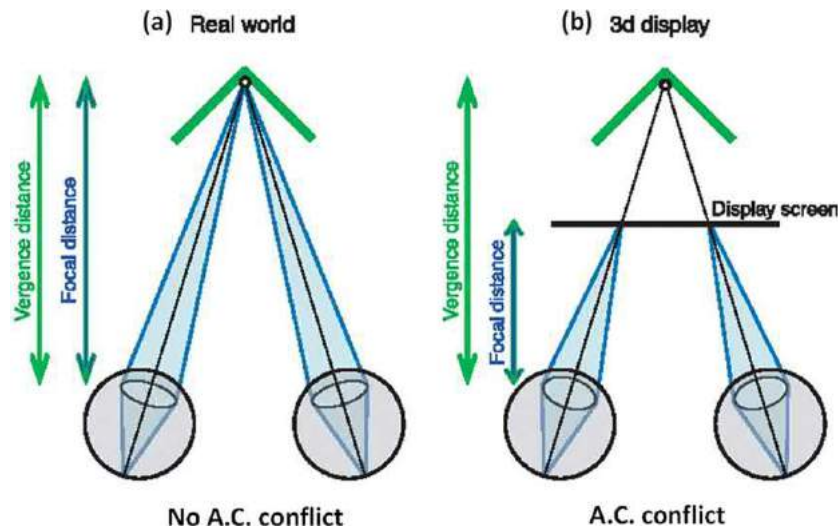


Fig. 3 In (a) real world, convergence and focal distance (accommodation) is the same; however, accommodation–convergence conflict (A.C. conflict) would happen on (b) 3D display. Reproduced from Hoffman, D.M., Girshick, A.R., Akeley, K., Banks, M.S., 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* 8, 33. doi: 10.1167/8.3.33

Anaglyph

Anaglyph is also called red-green glasses as shown in Fig. 4. Complementary colors for individual eyes are presented in one image. When viewer wears the red-green glasses, the glasses can filter out another color light. The binocular parallax can be achieved. In anaglyph 3D, partial color is filtered out by color glasses, resulting in poor color performance. Anaglyph 3D has also the resolution issue compared to 2D images.

Shutter Glasses

The display alternatively plays right and left images with a double frame rate. The viewer wears the glasses with liquid crystal shutters, which is synchronized by the display. The shutter switches on and off sequentially and each eye can receive its corresponding image. Because of vision persistence, these left and right image sequences are fused together in our brain, achieving the stereo perception. This technology can overcome the resolution decreasing issue, but the drawback is higher cost than other passive glasses. The battery is also needed for the active glasses.

Head Mounted Display

This kind of 3D display provides left and right two displays just in front of our eyes. The head mounted display can solve the interference between left and right images due to independent display for each eye (Tomilin, 1999). Only one viewer can watch 3D at once. Usually, the device is expensive, heavy, and uncomfortable for long-time wearing (Keller *et al.*, 2008).

These stereoscopic displays effectively yield a pair of disparity images to produce our stereo perception. However, viewers must wear eyeglasses when they watch the stereoscopic display. Therefore, the auto-stereoscopic display, which is eyeglasses-free, was developed.

E-Holography

This auto-stereoscopic display consists of red/green/blue lasers as light sources and acoustic optical modulator (AOM) as a phase grating (Javidi and Tajahuerce, 2000). Through the vertical scanning mirror and polygonal mirror, the lasers scan in vertical and horizontal direction to form a phase hologram (Son *et al.*, 2010). The 3D image size is limited by the AOM size and the scan speed of polygonal mirror must be synchronized with lasers (Takaki and Okada, 2010). This system will meet great challenge in data processing and mirror scan speed if the system scales up (Takaki *et al.*, 2010).

Volumetric

This volumetric display is composed by the spiral surface and bottom laser light source (Soltan *et al.*, 1995). When the light is projected to the rotated spiral surface, the light is scattered as a point of image. The spiral surface will scan all points in the working space to form the whole image as shown in Fig. 4. This kind of 3D display presents a transparent image without depth cue of occlusion (Refai, 2009). Meanwhile, in the region close to rotating axial, the slower speed induces a partial blurring image (Gately *et al.*, 2011). These effects degrade the stereo perception.

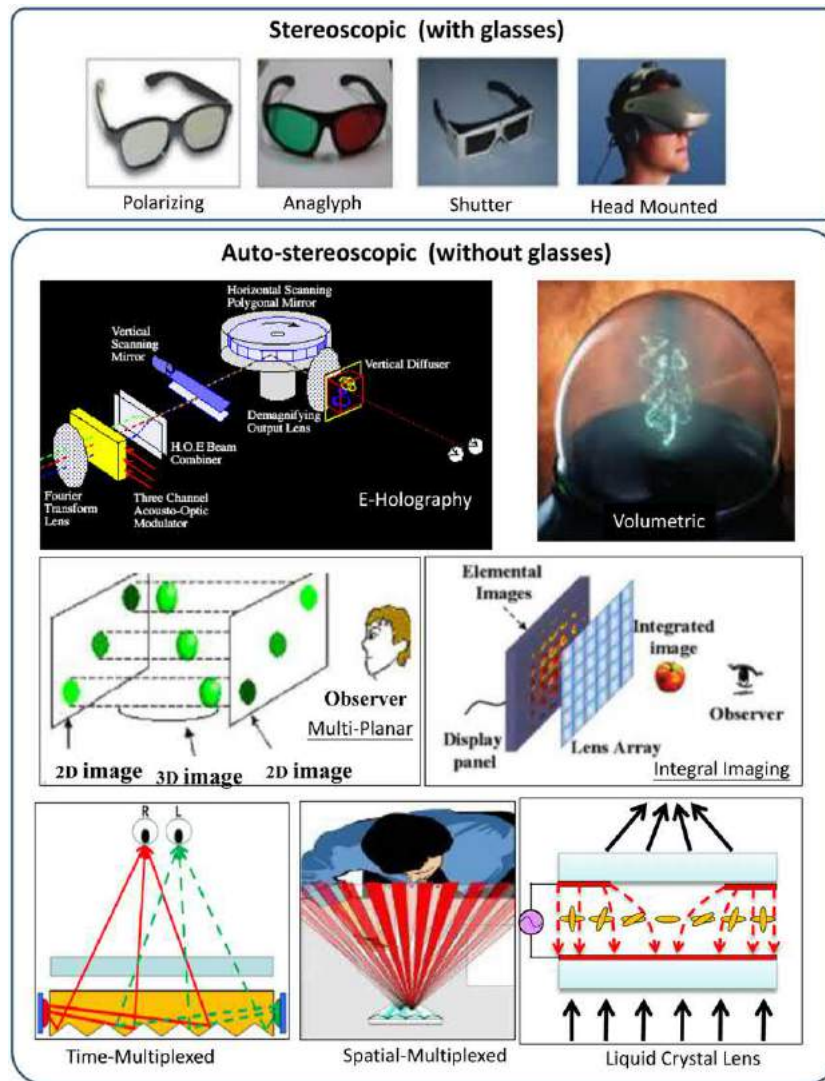


Fig. 4 3D displays classification based on with or without eyeglasses is divided into two groups, stereoscopic and auto-stereoscopic including various kinds of technologies (Jung *et al.*, 2009; Tomilin, 1999; Keller *et al.*, 2008; Javidi and Tajahuerce, 2000; Son *et al.*, 2010; Takaki and Okada, 2010; Takaki *et al.*, 2010, 2011; Soltan *et al.*, 1995; Refai, 2009; Gately *et al.*, 2011; Suyama *et al.*, 2004; Kuribayashi *et al.*, 2006; Hsu *et al.*, 2008; Berkel and Clarke, 1997; Young and Javidi, 2005; Chen *et al.*, 2009; Maphepo *et al.*, 2010; Tsuboi *et al.*, 2010; Takaki, 2010; Holliman *et al.*, 2011; Nakamura *et al.*, 2012; Hamagishi, 2009; Im *et al.*, 2008; Javidi and Okano, 2002; Son and Javidi, 2005; Sato, 1979; Woodgate, 2003, 2005; Willemsen and De Zwart, 2006; Krijn and de Zwart, 2008; Li and Wu, 2011; Ren and Wu, 2004; Naumov *et al.*, 1998; Ye and Sato, 2010; Chang *et al.*, 2014; Uehara *et al.*, 2013).

Multi-Planar

Two LCD panels are used in this multi-planar 3D display (Suyama *et al.*, 2004). These two LCD panels present the front and back image individually. The depth cues of various light and shadow in different viewing distance are applied to integrate front and back images. Precise alignment between the front and back panel could be a manufacturing issue (Kuribayashi *et al.*, 2006). In addition, only in the direct view direction, viewers can obtain the stereo perception. For other oblique viewing angle, the mismatch of two images cannot provide a 3D concept (Hsu *et al.*, 2008).

Time-Multiplexed

The LCD makers commonly apply the multiplexed type for 3D displays. One panel provides at least two views in time or spatial domain (Berkel and Clarke, 1997; Young and Javidi, 2005). In the time-multiplexed type, the left and right images are sequentially sent to left and right eye. Different from eyeglasses of stereoscopic display, the light guiding devices are implemented inside the LCD panel (Holliman *et al.*, 2011). When the switching speed is higher than that of our brain can distinguish,

the fusion of stereo images occurs. This 3D display preserves the resolution but requires a high frame rate LCD panel (Hamagishi, 2009).

Spatial-Multiplexed

Lenticular lens is attached on the LCD panel for guiding light from odd/even column pixels to left/right eyes (Im *et al.*, 2008). The cylindrical lens can focus and refract light from one image into more than two sub-images (Chen *et al.*, 2009; Takaki, 2010). The number of sub-images determines the number of views in stereo perception. More views can provide more continuous 3D viewing concept when viewers are moving their heads. In addition, parallax barrier is also popular as a light splitting device for disparity images (Maphepo *et al.*, 2010). The barrier is a black and transparent strip lines to restrict the light path from LCD panel to eyes. One eye of viewers only receives light from a specific LCD region (e.g., the odd or even column pixels); the other one eye receives light from the other region. Compared to the lenticular lens, the light intensity of barrier 3D display is lower due to the barrier's blocking light (Tsuboi *et al.*, 2010).

Integral Imaging Display

An integral imaging display uses the micro-lens array instead of the cylindrical lenses used by lenticular lens display (Javidi and Okano, 2002). Each micro lens directs light from the pixels and covers in different horizontal and vertical directions, thus producing a full parallax image. Therefore, the floating 3D image can be achieved by integral photography theory (Son and Javidi, 2005). The integral imaging display has an issue of image overlapping in oblique viewing angles. By modifying the pixel emitting into correct micro lens, this issue can be overcome (Javidi and Okano, 2002).

LC Lens for a 2D/3D Switchable Display

Spatial-multiplexed display degrades the image resolution, even just presents 2D contents. The advantage of LC lens-based switchable display is to preserve the high image resolution in 2D display mode (Sato, 1979). By switch-on the LC lens, display will show the 3D images (Ren and Wu, 2004; Naumov *et al.*, 1998; Ye and Sato, 2010). Three types of LC lens, polarization active micro-lens (Li and Wu, 2011), active LC lenticular lens (Krijn and de Zwart, 2008), and electric-field-driven (Willemsen and De Zwart, 2006), had been reported and successfully demonstrated. LC alignment direction (LCAD) (Woodgate, 2005) and electrode array direction (EAD) (Woodgate, 2003) was then proposed to have 3D rotational function for mobile display (Chang *et al.*, 2014; Uehara *et al.*, 2013). Additionally, the temporal scanning LC-lens (Huang and Chen, 2012), such as multi-electrode driven (MeD) Fresnel LC lens was proposed to improve the resolution of 3D images (Huang *et al.*, 2010).

Super-Multi-View Display

Due to the limitation of the number of multi-view, there is always discontinuity of view switching with respect to the viewing direction. This effect degrades the 3D perception. Super-multi-view (SMV) techniques use an extremely large number of views (e.g., 256 or 512 views) to generate more natural 3D display visual effects (Takaki *et al.*, 2011). The horizontal interval between views is smaller than the diameter of the human pupil. Light from slightly different viewpoints enters the pupil simultaneously. This is called the SMV display condition (Nakamura *et al.*, 2012). SMV provides smooth motion parallax and reduces the A.C. conflict since the brain unconsciously predicts the image change due to motion.

Augmented and Virtual Reality Display

Although wearing a head mounted display (HMD) is not very convenient for users, but Augmented Reality (AR) and Virtual Reality (VR) technologies can produce the virtual images to make people's brain perceived as a real 3D object. In addition, the hardware devices have been improved to be much lighter than that in the 20th century. AR/VR HMD becomes a popular system for many applications. The major difference between AR and VR is that VR make people thoroughly immersed with an ocular head mounted device, interactivity with sensory feedback. While AR only produces the virtual images covering part of observer's field of view and integrate the images with the environmental background.

There are several optical arrangements to render the images for AR and VR applications. For an occlusion type, mainly for VR application, the display can be placed closed to the eyes, then use optical lens to generate a virtual image with large field of view, as shown in Fig. 5(a). While in the case of see-through type, mainly for AR applications, the optical setup includes off-axis and partially reflective mode, light guided mode, or total internal reflection mode, to combine the virtual image with environmental scenes, as shown in Fig. 5(b).

The applications of AR/VR displays are not only limited on consumer electronics, such as video games, video entertainment, or live events. This technology can be further used on healthcare/medical, engineering, education, retail, and military field. Such as in the case of medical application, the VR display combined with a brain-wave detector can be used for detecting glaucoma damage

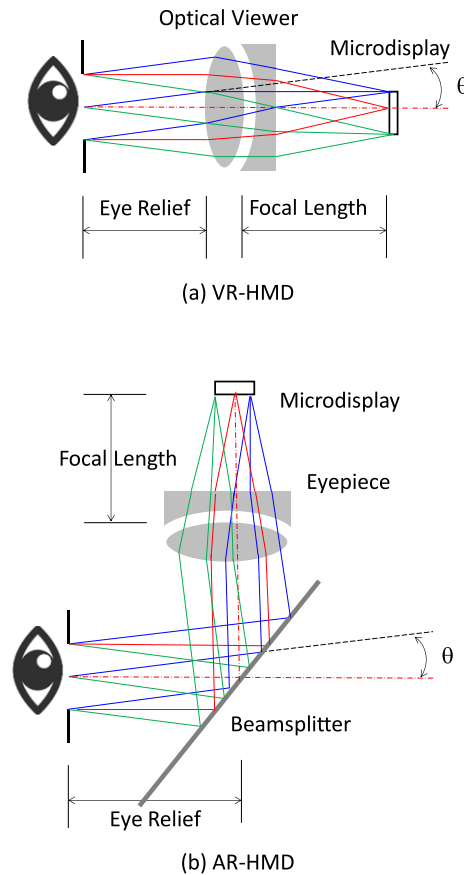


Fig. 5 Optical structure of (a) virtual reality head mounted display. The lens is placed in front of the eye and micro display located near eye is floated to the desired depth. (b) Augment reality head mounted display. The micro display is located at the side part of the observer and the displayed virtual information is bent to the eye by the optical combiner.

(Chien *et al.*, 2013; Lin *et al.*, 2015). It could even potentially allow users to routinely and continuously monitor the electrical brain activity associated with visual field in their living environments, representing a transformative way of diagnosing and monitoring disease progression in not only glaucoma but other functional visual losses.

See also: Displays – Introduction. Liquid Crystal Physics and Materials

References

- Berkel, C.V., Clarke, J.A., 1997. Characterization and optimization of 3D-LCD module design. *Proc. SPIE* 3012, 179–187.
- Chang, Y.C., Jen, T.H., Huang, Y.P., 2014. High-resistance liquid crystal lens array for rotatable 2D/3D auto-stereoscopic display. *Opt. Express* 22, 2714–2724.
- Chen, C.H., Huang, Y.P., Chuang, S.C., *et al.*, 2009. Liquid crystal panel for high efficiency barrier type autostereoscopic three-dimensional displays. *Appl. Opt.* 48, 3446–3454.
- Chien, Y.Y., Lin, F.C., Chou, C.C., *et al.*, 2013. Polychromatic high-frequency steady-state visual evoked potentials for brain-display interaction. In: *SID Symposium Digest of Technical Papers*, vol. 44, pp. 146–149.
- Gately, M., Zhai, Y., Yeary, M., Petrich, E., Sawalha, L., 2011. A three-dimensional swept volume display based on LED arrays. *J. Display Technol.* 7, 503–514.
- Hamagishi, G., 2009. Analysis and improvement of viewing conditions for two-view and multi-view display. In: *Proceeding of the SID Symposium Digest*, vol. 40, pp. 340–343.
- Hill, L., Jacobs, A., 2006. 3-D liquid crystal displays and their applications. *Proc. IEEE* 94, 575–590.
- Holliman, N., 2005. 3D display system. In: *Handbook of Optoelectronics*. Institute of Physics Press. ISBN: 0 7503 0646.
- Holliman, N.S., Dodgson, N.A., Favalora, G.E., Pockett, L., 2011. Three-dimensional displays: A review and applications analysis. *IEEE Trans. Broadcasting* 57, 362–371.
- Hsu, C.Y., Huang, Y.P., Chang, Y.C., 2008. Wide viewing angle 3D depth-fused display (DFD) system. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 1655–1658.
- Huang, Y.P., Chen, C.W., 2012. Superzone Fresnel liquid crystal lens for temporal scanning auto-stereoscopic display. *IEEE/OSA J. Display Tech.* 8 (11), 650–655.
- Huang, Y.P., Liao, L.Y., Chen, C.W., 2010. 2D/3D switchable autostereoscopic display with multi-electrically driven liquid crystal (MeD-LC) lenses. *J. Soc. Inf. Display* 18 (9), 642–646.
- Im, H.J., Jung, S.M., Lee, B.J., Hong, H.K., Shin, H.H., 2008. Mobile 3-D display based on a LTPS 2.4-in. VGA LCD panel attached with lenticular lens sheets. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 256–259.

- Javidi, B., Okano, F., 2002. Three-dimensional television, video, and display technologies. New York: Springer-Verlag.
- Javidi, B., Tajahuerce, E., 2000. Three dimensional object recognition using digital holography. *Opt. Lett.* 25, 610–612.
- Jung, S.M., Park, J.U., Lee, S.C., *et al.*, 2009. A novel polarizer glasses-type 3D displays with an active retarder. In: *Proceeding of the SID Symposium Digest*, vol. 40, pp. 348–351.
- Keller, K., State, A., Fuchs, H., 2008. Head mounted displays for medical use. *J. Display Technol.* 4, 468–472.
- Krijn, M.P.C.M., de Zwart, S.T., 2008. 2-D/3D displays based on switchable lenticulars. *J. Soc. Inf. Display* 16/8, 847–855.
- Kuribayashi, H., Date, M., Suyama, S., Tatada, H., 2006. A method for reproducing apparent continuous depth in a stereoscopic display. *J. Soc. Info. Display* 14, 493–498.
- Lin, F.C., Chien, Y.Y., Zao, J.K., *et al.*, 2015. High-frequency polychromatic visual stimuli for new interactive display systems. *SPIE Newsroom*. doi:10.1117/2.1201504.005851.
- Li, Y., Wu, S.T., 2011. Polarization independent adaptive microlens with a blue-phase liquid crystal. *Opt. Express* 19, 8045–8050.
- Maphepo, W., Huang, Y.P., Shieh, H.P.D., 2010. Enhancing the brightness of parallax barrier based 3D flat panel mobile displays without compromising power consumption. *J. Display Technol.* 6, 60–62.
- Nakamura, J., Takahashi, T., Chen, C.W., Huang, Y.P., Takaki, Y., 2012. Analysis of longitudinal viewing freedom of reduced-view super multi-view display and increased longitudinal viewing freedom using eye-tracking technique. *J. Soc. Info. Display* 20, 228–234.
- Naumov, A.F., *et al.*, 1998. Liquid-crystal adaptive lenses with modal control. *Opt. Letters* 23 (13), 992–994.
- Pastoor, S., Wopking, M., 1997. 3-D displays: A review of current technologies. *Displays* 17, 100–110.
- Refai, H.H., 2009. Static volumetric three-dimensional display. *J. Display Technol.* 5, 391–397.
- Ren, H.W., Wu, S.T., 2004. Getting in tune: Electronically controlled liquid crystal yields tunable-focal-length lenses. *SPIE's Optics Express Magazine*. 24–27.
- Sato, S., 1979. Liquid-crystal lens—cells with variable focal length. *Jpn. J. Appl. Phys.* 18, 1679–1684.
- Soltan, P., Trias, J., Dajlke, W., Lasher, M., McDonald, M., 1995. Laser based 3-D volumetric display system. *Naval Engineers J.* 107, 233–243.
- Son, J.Y., Javidi, B., 2005. Three-dimensional image methods based on multi-view images. *J. Display Technol.* 1, 125–140.
- Son, J.Y., Javidi, B., Yano, S., Choi, K.H., 2010. Recent developments in 3D image technologies. *J. Display Technol.* 6, 394–403.
- Srivastava, A.K., de, J.L., de la Tonnaye, B., Dupont, L., 2010. Liquid crystal active glasses for 3D cinema. *J. Display Technol.* 6, 522–530.
- Suyama, S., Ohtsuka, S., Takada, H., Uehira, K., Sakai, S., 2004. Apparent 3-D image perceived from luminance-modulated two 2-D images displayed at different depths. *Vis. Res.* 44, 785–793.
- Takaki, Y., 2010. Multi-view 3-D display employing a flat-panel display with slanted pixel arrangement. *J. Soc. Info. Display* 18, 476–482.
- Takaki, Y., Okada, N., 2010. Reduction of image blurring of horizontally scanning holographic display. *Opt. Express* 18, 11327–11334.
- Takaki, Y., Tanaka, K., Nakamura, J., 2011. Super multi-view display with a lower resolution flat-panel display. *Opt. Express* 19, 4129–4139.
- Takaki, Y., Yokouchi, M., Okada, N., 2010. Improvement of grayscale representation of the horizontally scanning holographic display. *Opt. Express* 18, 24926–24936.
- Tomilin, M.G., 1999. Head-mounted displays. *J. Optical Technol.* 66, 528–533.
- Tsuboi, M., Kimura, S., Takaki, Y., Horikoshi, T., 2010. Design conditions for attractive reality in mobile-type 3-D display. *J. Soc. Info. Display* 18, 698–703.
- Uehara, M.K.S., Takagi, A., Baba, M., 2013. LC GRIN lens mode with wide viewing angle for rotatable 2D/3D tablet. In: *Proceeding of the SID Symposium Digest*, pp. 154–157.
- Willemsen, O.H., De Zwart, S.T., 2006. 2-D/3D switchable displays. *J. Soc. Inf. Display* 14/8, 715–722.
- Woodgate, G.J., 2003. High efficiency reconfigurable 2D/3D autostereoscopic display. In: *2003 SID International Symposium Digest of Technical Papers*, pp. 394–397.
- Woodgate, G.J., 2005. A new architecture for high resolution autostereoscopic 2D/3D displays using free-standing liquid crystal microlenses. In: *2005 SID International Symposium Digest of Technical Papers*, pp. 378–381.
- Wu, Y.J., Jeng, Y.S., Yeh, P.C., Hu, C.J., Huang, W.M., 2008. Stereoscopic 3D display using pattern retarder. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 260–263.
- Ye, M., Sato, S., 2010. Low-voltage-driving liquid crystal lens. *Jpn. J. Appl. Phys.* 49, 100204-1–100204-3.
- Young, S.J., Javidi, B., 2005. Three-dimensional image methods based on multi-view images. *J. Display Technol.* 1, 125–140.

Wearable Displays

Paul Yang, Sun Yat-Sen University, Guangzhou, China

© 2018 Elsevier Ltd. All rights reserved.

Spectacle-Type

Spectacle-type of wearable display, also known as head mounted display (HMD), includes two major categories, augmented reality (AR) HMD and virtual reality (VR) HMD. Both AR and VR technologies produce the virtual images for human eyes and make people's brain perceived as something really exists. The major difference between AR and VR is that VR make people thoroughly immersed with ocular device, while AR only produce the virtual images covering part of observer's field of view and integrate the images with the environmental facilities.

There are several optical arrangements to render the images for AR and VR applications. For occlusion type, mainly for VR application, the display can be placed very closed to the eyes, or use mirror to direct the light into the human pupil to render images, and maximize the field of view (FOV), as shown in Fig. 1(a).

While in the case of see-through type, mainly for AR applications, the optical setup includes off-axis and partially reflective mode, light guided mode, and TIR mode, to maximize the observation of environmental scene, which thus decreased the FOV, as shown in Fig. 1(b). For different application segments, the optical designs may differ. For example, in the smart glasses application which can function like handless cellphone, the aesthetic appearance is more essential, while in the case of medical application, the resolution and contrast become much more important. As shown in Fig. 2(a) and (b), it is very helpful for a doctor to use

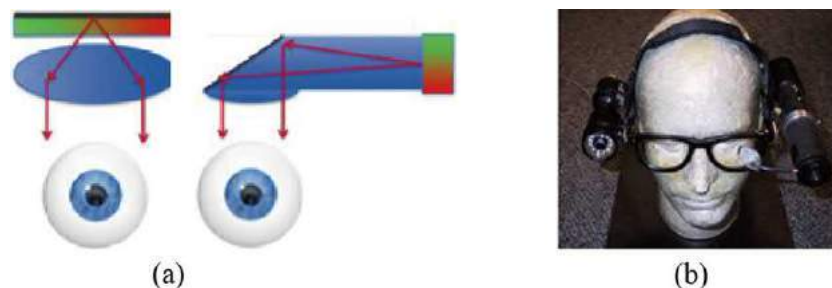


Fig. 1 The optical setup of AR and VR includes two basic arrangements: (a) totally occlusion with near-eye display and (b) optical combiner with light-guiding system. Reproduced from Kress, B., Saeedi, E., Brac-de-la-Perriere, V., 2014. The segmentation of the HMD market: Optics for smart glasses, smart eyewear, AR and VR headsets. In: Kazemi, A.A., Kress, B.C., Mendoza, E.A., *et al.* (Eds.), *Photonics Applications for Aviation, Aerospace, Commercial, and Harsh Environments V*, vol. 9202. Bellingham: Spie-Int Soc Optical Engineering; Bryant, R.C., Lee, C.M., Burstein, R. A., Seibel, E.J., 2004. 59.2: Engineering a low-cost wearable low vision aid based on retinal light scanning. *SID Digest* 35 (1), 1540–1543.

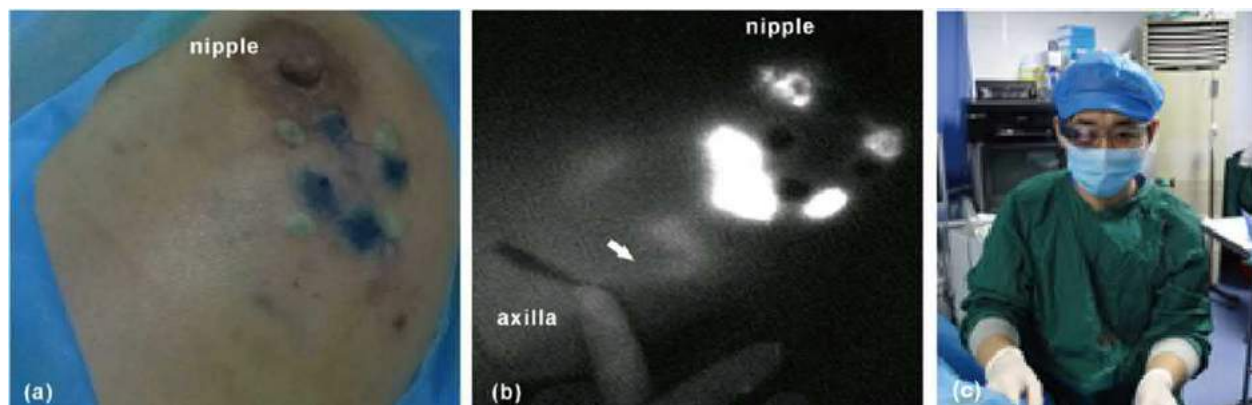


Fig. 2 The demonstration of using AR wearable glasses for medical operation: (a) the photograph of patient's breast, (b) the corresponding fluorescence image, and (c) the appearance of doctor wearing the AR glasses. Reproduced from Zhang, Z.S., Pei, J., Wang, D., *et al.*, 2016. A google glass navigation system for ultrasound and fluorescence dual-mode image-guided surgery. In: VoDinh, T., Mahadevan-Jansen, A., Grundfest, W.S. (Eds.), *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XIV*, vol. 9698. Bellingham: Spie-Int Soc Optical Engineering.

fluorescent image to identify the ordinary cell and the tumor cell. In this application, the doctor can conveniently observe by wearing the smart glasses.

Clothes-Integratable-Type

With the upcoming trend of "internet of things, IoT," people expect every device can be integrated with clothes or environments, which can make the users can "hands off." To realize this fantasy, the devices need to be flexible and stretchable, and so do the displays. Making display flexible is of one route, while making it stretchable is the other. Here, we introduce essential technologies for achieving flexible and stretchable displays, and therefore, these technologies can be combined for realizing clothes integrated type of wearable display, as shown in Fig. 3 (Lee *et al.*, 2016).

Flexible Displays Integrating With Clothes

Some wearable displays integrated with clothes were demonstrated, which are expected to provide more functions for users beyond the current smart phone applications. However, the appearance is still not aesthetic enough for fashion design, because the existing technology cannot be stretchable and conformable for fashion designers to realize their design (Fig. 4).

Flexible Displays Integrating With Human Body

Along with the development of flexible electronics, more and more sensors were prepared that contributes to human's convenience. For the further application, the integration with human body makes these devices more attractive.

Therefore, sensors are expected to be attached to the human body and people will need to know what the sensors detected. For example, the detected temperature sends the signal of your body condition that is needed to be known timely and a human-sensor interaction is therefore essential for practicability. In this regard, as an information carrier, a wearable display was also needed in these applications. The integration with human body makes this kind of flexible display different from the conventional. There are several design considerations should be included. Because the direct contact with human skin, the attached devices should not affect human's health as well as the ability of other activities. On one hand, the flexibility should be fully realized that does not

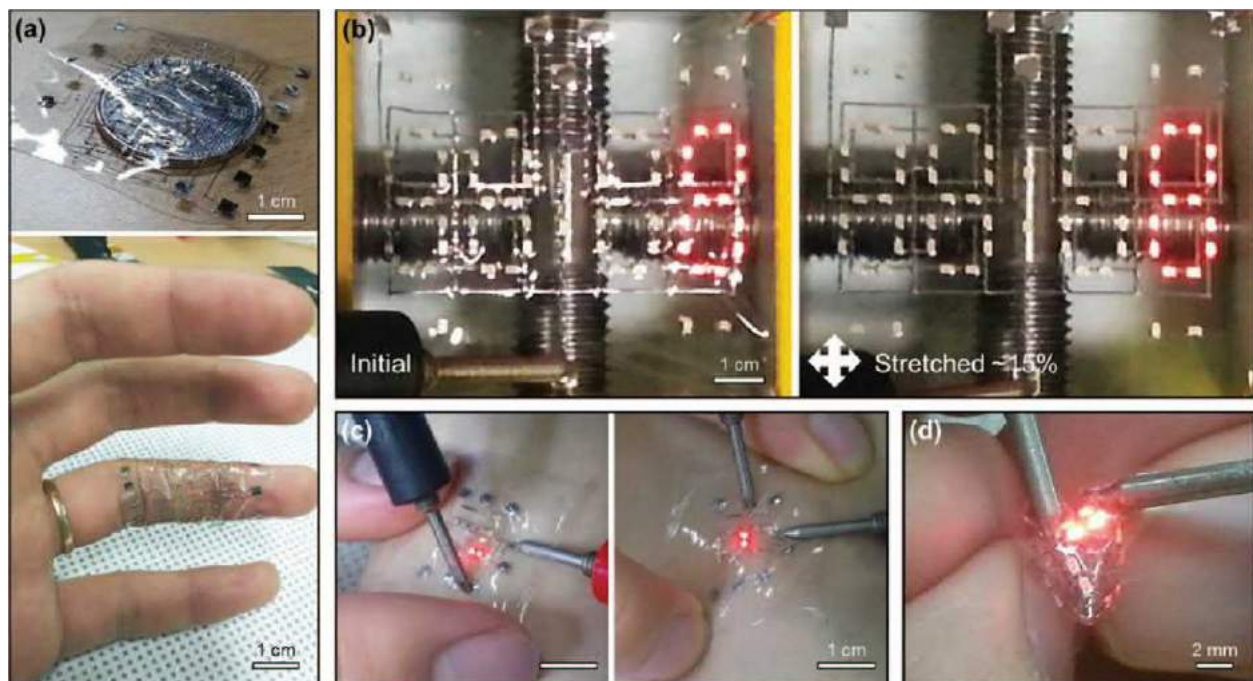


Fig. 3 (a) Photographs of freestanding stretchable 4-digit 7-segment on coin (upper) and skin (lower). (b–d) Photographs of operation of stretchable all-ink-jet-printed 4-digit 7-segment display with initial and stretched state (b), when attached to skin (c) and crumpled (d). Reproduced from Lee, B., Byun, J., Oh, E., *et al.*, 2016. All-ink-jet-printed wearable information display directly fabricated onto an elastomeric substrate. *SID Digest 47* (1), 672–675.



Fig. 4 The prototypes of the wearable-display wrist unit. Reproduced from Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56; Tajima, R., Miwa, T., Oguni, T., *et al.*, 2014. Truly wearable display comprised of a flexible battery, flexible display panel, and flexible printed circuit. *Journal of the Society for Information Display* 22, 237–244.

obstruct human's activities. To satisfy this requirements, our display should be flexible and stretchable to adapt muscle movement. So an integration with several rigid substrates cannot be accepted. On the other hand, the materials of good biological compatibility and breathability should be chosen as the flexible substrates. And the complex operating environment of human body raise new demands of endurance. Taking these considerations into account, the wearable display can not only succeed the advantages of light weight and facile fabrication of flexible devices and also be applied on human body that makes human's life more convenient.

Watch-Type

Display makers tried very hard to make the watch-type of display looks like conventional, especially the appearance needs to be round. The pixel array used in display is usually in rectangular arrangement. By applying special algorithm to dim the lightness a bit around the boundary, it is possible to render round watch appearance with conventional rectangular pixel arrangement (**Fig. 5 (a)**). In addition, by folding the COF underneath the watch, the form factor can be as slim as the actual watch.

Flexible Substrates

Flexible displays are made with flexible components including substrates, thin film transistors (TFTs), and display medium. Usually, the devices need to be annealed to stabilize and optimize the performances. However, flexible for ordinary TFT device processing. Polyethylene naphthalate (PEN) has better thermal durability, however, ordinary TFT needs 350°C of post-annealing, it is still too low in this case. Polyimide (PI) can endure 350°C and its thermal expansion coefficient matches with substrate made of plastics are not endurable for high temperature processing, while flexible substrate made of steel are too heavy to be made for mobile devices. As shown in **Table 1**, polyethylene terephthalate (PET) is the cheapest one, however, its glass transition temperature (T_g) is too low for other layers of a flexible device, however the cost of PI is still too high for consumer products. As a result, scientists are paying much attention develop a substrate with simultaneously high T_g , high transparency (for aesthetic purpose), high flexibility, and low cost.

Flexible TFT

Flexible TFT is also a critical component of flexible display, taking charge of controlling the grey scale of each pixel. The desired features of a good flexible TFT array are high switching speed, low power consumption, high aperture ratio, good flexibility, and reliable electrical performances. The material characteristics for these requirements are high mobility of semiconducting layer, low resistance of connection metal line, small capacitance coupling between layers, smaller area of TFT blocking the transmissive light, high power efficiency by eliminating the unwanted polarizers or color filters, capability of active layers to be deformed, and the stable material characteristics after photo or electrical stressing.

Here is the summary of flexible TFTs, as shown in **Table 2**. Organic TFT has the best capability of being bent, however, its mobility is still too low for high end applications, such as smart phone. Amorphous Si TFT (*a*-TFT) is the most mature TFT, well adopted in existing displays, however, its flexibility is not good enough for future wearable displays. While oxide TFT seems to have better overall performances in terms of flexibility, mobility, cost, and compatibility with existing manufacturing facilities. However, the oxide TFT still needs to be improved for its long-operation reliability, which is induced by the stressing and resulting the electrical performance degradation.

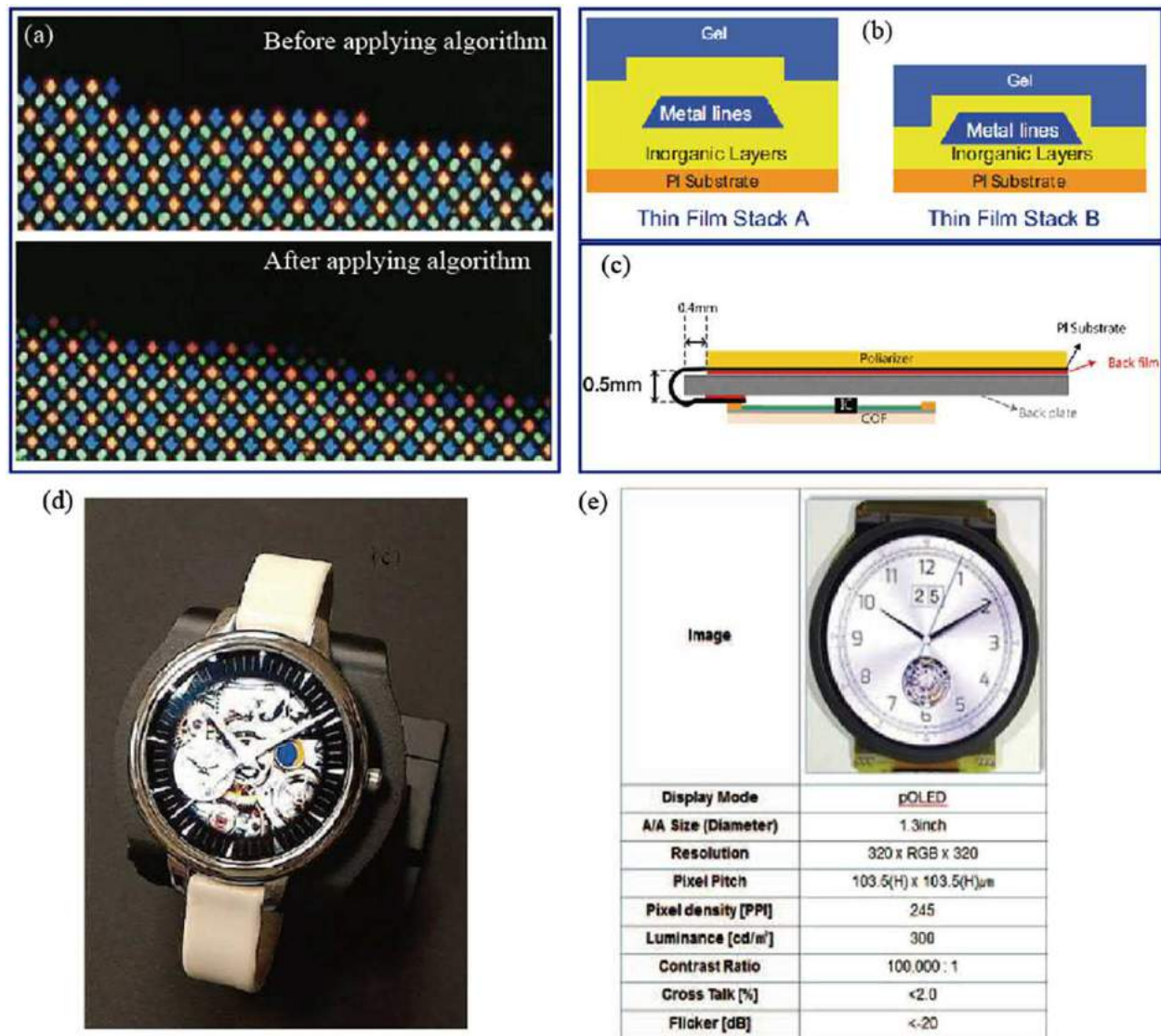


Fig. 5 (a) Image of before and after applying algorithm, (b) two thin film stacks with different total thickness of inorganic layers, (c) bending structure of the panel, (d) 1.26" circular plastic AMOLED panel, and (e) specification of circular plastic OLED display. Reproduced from Jung, D.S., Song, S.M., Kang, H.K., *et al.*, 2016. Panel design technology for circular OLED display. *SID Digest 47* (1), 844–846; Li Fong, L., Chia Hsum, T., Ko Ruey, J., *et al.*, 2016. A circular flexible AMOLED display with a 1-mm slim border. *SID Digest 47* (1), 847–850

Table 1 Basic properties of film used for base substrates

	PET	PEN	PC	PES	PAR	PCO	PI
CTE (−55 to 85°C) ppm/°C	1.5	13	60–70	54	53	74	17
Transmission at 400–700 nm (%)	>85	>85	>90	90	90	91.6	yellow
Water absorption	0.14	0.14	0.2–0.4	1.4	0.4	0.03	1.8
Young's modulus (GPa)	5.3	6.1	1.7	2.2	2.9	1.9	2.5
Tensile strength (MPa)	22.5	275	NA	83	100	50	231

To accommodate the exiting processes for flexible TFT array, the plastic substrate is usually bonded onto the rigid substrate with releasable adhesive. As shown in Fig. 6(a), the TFTs were formed onto a pre-laminated substrate with the structure of plastic/adhesive/carrier-plate by conventional photo-lithography, and be peeled off after the TFT array finished. Other modified approaches include using sacrificing layer as shown in Fig. 6(b) and using transfer process as shown in Fig. 6(c). They are all based on the concept of adopting existing TFT process, while forming the TFT arrays on the flexible substrates, as shown in Fig. 7.

Table 2 The comparison of flexible TFTs

		Poly-Si	Amorphous-Si	Organic TFT (Pentacene)
Performance	Mobility	Very Good	OK for PHOLEDs	OK
	Leakage	OK	Very good	OK
	Stability	Good	Issue	Issue
	Uniformity	Issue	OK	Issue
Manufacturability		Maturing	Excellent	N/A
Cost		> Medium	Medium	Low??
Plastic-substrate compatibility		Under development	OK	Excellent
Metal-foil-substrate compatibility		Good	Good	Good

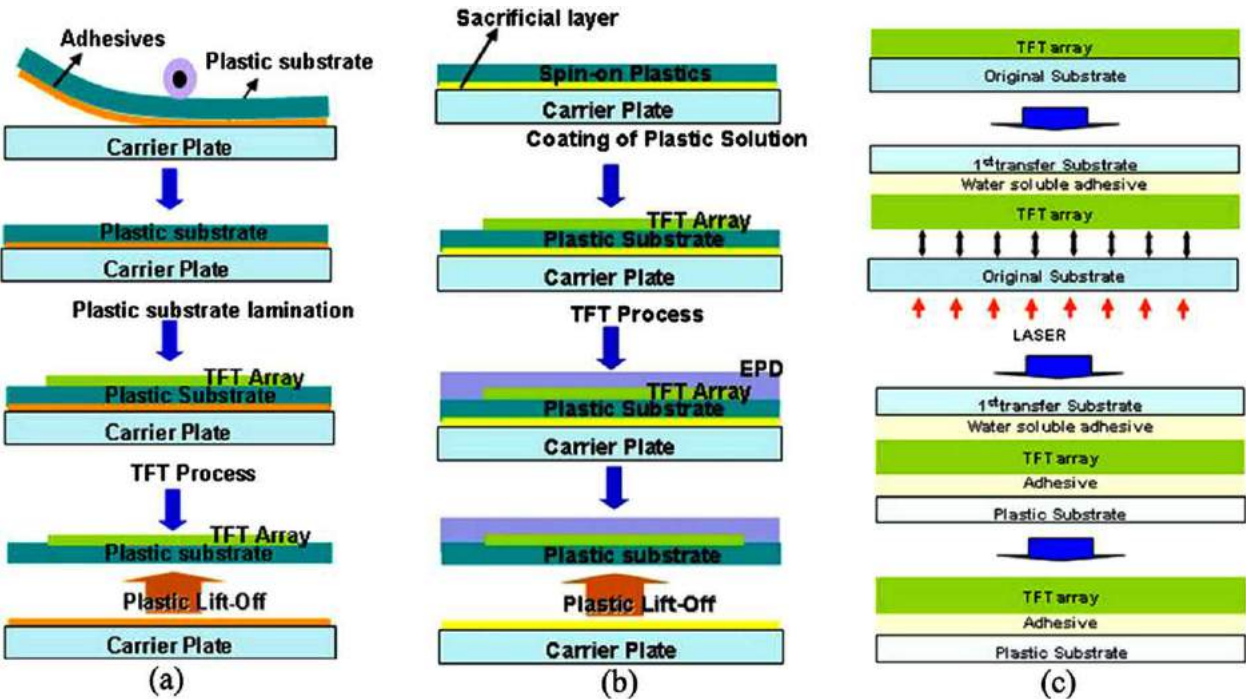


Fig. 6 The handling process for making flexible TFT array. Reproduced from Li, Z.M., Lv, Q.M., Huang, Y.J., 2010. Versatile soft backplane Technology. Journal of Industrial Materials.

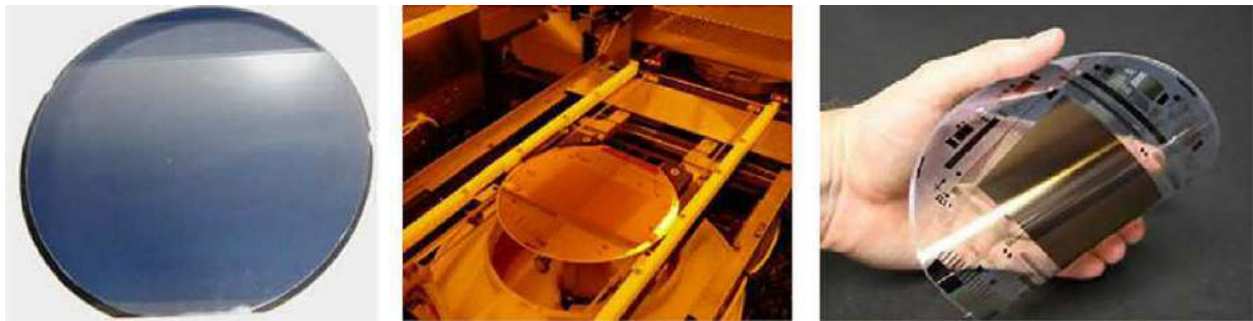


Fig. 7 A handling process for making flexible TFT array. Reproduced from Li, Z.M., Lv, Q.M., Huang, Y.J., 2010. Versatile soft backplane Technology. Journal of Industrial Materials.

Flexible Organic Light Emitting Diode

There are two major kinds of flexible display made of organics: small molecules, and polymer materials (usually called polymer light emitting diode, PLED, to differentiate with small molecular organic light emitting diode, OLED). The former one was firstly invented by Dr. Ching W. Tang in Kodak, and the latter was invented by Prof. Richard Friend's team at University of Cambridge. The lighting mechanism of OLED is to inject both electrons and holes from cathode and anode, respectively, where both carriers would transport in the organic layers to form excitons, and later on illuminate the light and go back to the energy ground state. To optimize the lighting efficiency, the rate of injection, transportation, and recombination of electrons and holes should be delicately designed.

Flexible OLED can perform the best image quality, however, it is very vulnerable to moisture and oxygen, which thus needs to combine with flexible passivation technology before it can be commercialized (the structure is shown in Fig. 8; Ma *et al.*, 2010). The mature technology used for passivation is the multi-stacking of organics and inorganics, where inorganics are used for blocking the oxygen or water molecules's diffusion, and the organics are used for balancing the inter-layer stress while being bent. The other popular method is called atomic layer deposition (ALD), which is made with certain reactive precursors, mixing with reacting gas and precisely forming an atomic layer for passivating the OLED. As the layer is of atomic level, the device is ultimately flexible.

Flexible E-Paper

Different from the lighting mechanism of flexible OLEDs, flexible electrophoretic displays (EPD) can reflect the ambient light to show the image. Because the image appearance is similar to actual paper, it is usually called E-paper. Not only EPD, but also many other kinds of reflective display technologies are promising for future E-paper applications, such as liquid powder, electro-wetting, electro-chromic, cholesteric liquid crystal, polymer dispersed liquid crystal displays, and so on. Among them, EPD is the first flexible display commercialized for e-reader applications, however its color image performance is not as good as flexible OLED, and it is not suitable for video application because of its slow updating speed. Even though, EPD has two unique advantages superior to flexible OLED, one is the good sunlight readability due to its particle scattering, and the other is the power-saving due to its bistability.

The fabrication processes of making E-paper includes two major categories: microcapsule type and micro-cup type, and both of them can be leveraged for future flexible display fabrication. Both type of E-paper includes the synthesis of particles, formulation with surfactant and additives, encapsulated with a small chamber (either capsule or cup), and then laminated onto the flexible substrate with roll-to-roll (R2R) processes.

To realize a full-color E-paper, many approaches have been proposed. The easiest way is to add on color-filter (Lai *et al.*, 2012), however, considering the low reflective efficiency, the image was not good enough. Usually users do not like this kind of vague color image. Until recently, E-ink demonstrated a structure utilizing multiple color particles within one microcup to render color images by selectively control certain particle to be on top of the surface (Telfer and McCreary, 2016). This approach has the best performance in terms of contrast, color saturation, and resolution. However, such method needs to mix many kinds of color particle in one small chamber, and thus the fabrication repeatability is hard to control and the driving scheme is quite complicated. To overcome such difficulty, transfer method to spatially deploy the particle color particles to sub-pixels, which only need two kinds of color particles within one chamber was demonstrated, and thus the process is much easier to control and the driving scheme is comparably simple.

Flexible Liquid Crystal Display

LCD is the most mature display technology in display thus far. It would be imperative if LCD can be flexible. However, the high birefringence of the plastic substrate is an issue since LCD is polarization dependent. As shown in Fig. 9(a) (Oka *et al.*, 2016), phase retardation caused by the high birefringence is relevant to thickness of the plastic substrate. So several approaches have been

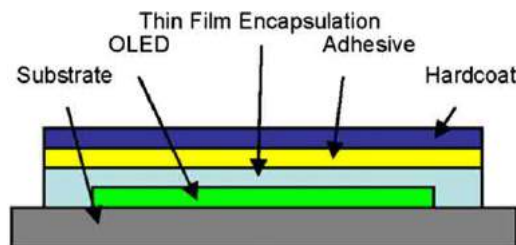


Fig. 8 A structure of thin film encapsulated OLED. Reproduced from Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56.

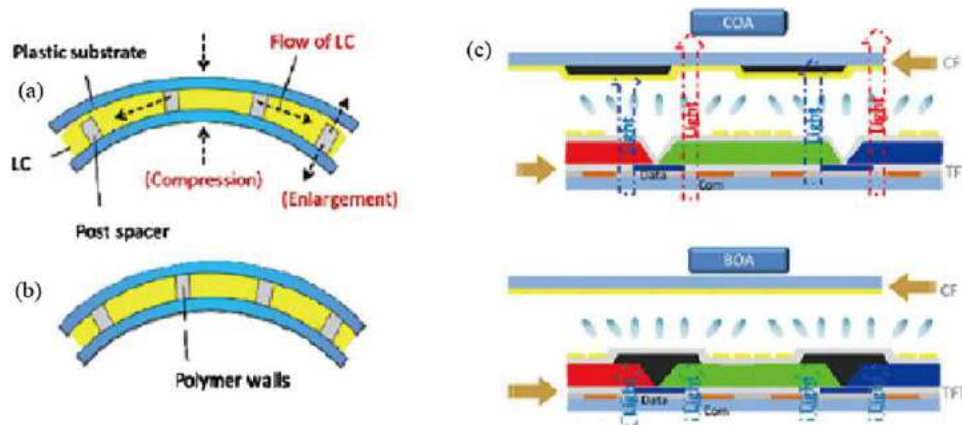


Fig. 9 (a) Deformation of plastic substrate during bending, (b) flexible LC device with polymer wall spacers, and (c) the principle of BOA to solve the BM shift issue in curved LCD panel. Reproduced from Ishinabe, T., Obonai, Y., Fujikake, H., 2016. A foldable ultra-thin LCD using a coat-debond polyimide substrate and polymer walls. *SID Digest 47* (1), 83–86; Ye, C.L., Fu, R.H., Qiu, J., *et al.*, 2016. The application of BOA on curved panel. *SID Digest 47* (1), 25–27.

taken: phase compensation film (Pin Hsiang *et al.*, 2016), in-cell polarizer (Fujiwara *et al.*, 2016), ultrathin substrate (Oka *et al.*, 2016; Pin Hsiang *et al.*, 2016) and low birefringence plastic substrate (Ishinabe *et al.*, 2016). Besides, thermal durability of plastic substrate is also important because of the high temperature during the LCD device fabrication process. Polyimide (PI) is widely applied as thermal durable plastic substrate, whose glass transition temperature ($> 200^{\circ}\text{C}$) is compatible with LCD fabrication. However, PI needs further material design to be colorless since it is a little yellowish.

In addition to plastic substrate, curvature mismatching between different layers is another critical issue, which needs flexible spacers, black matrix alignment, etc. (as shown in Fig. 9(b)). Recently, Prof. T. Ishinabe from Tohoku University demonstrated with a so-called coat-debond method to fabricate the low birefringence polyimide (PI) and polymer wall spacer to make a flexible LCD demo, which has testified that flexible LCD is possible to be a candidate for future flexible display applications. Polymer walls used as the flexible spacer which could stick both substrate together are able to maintain the cell gap when compressing the device (Ishinabe *et al.*, 2016).

Flexible Transparent Conductive Film

The desired feature of conductors for wearable displays are: flexibility, stretchability, transparency, and conductivity. Conventional conductors like Indium Tin oxide (ITO) seem too fragile in flexible applications. Currently, the most popular substitute materials are graphene and its derivatives, metal nanowire, carbon nanotubes, and two-dimensional materials, which are all compatible with flexible substrate. However, carbon based materials synthesize still needs vacuum and even toxic process, which increases the cost. Two-dimensional materials are also challenging for large size preparation process. Metal nanowire such as silver nanowire forms conducting network. According to percolation theory, the ratio of conducting component which is non-transparent could be decreased to increase the transmittance with high conductivity, which is considered as one of the promising material.

Since metal nanowire is not transparent, it's essential to trade off the transmittance and conductivity. Several methods have been proposed including metal mesh, alignment coating, stacking structure, etc. (Yang *et al.*, 2016).

Patterning of flex transparent conducting film (Flex-TCF) is critical for device applications. The conventional way is lithography, but the etching recipe and conditions are totally different from ones in glass-based process. The plastic substrate is softer and less chemical resistive than glass, which means the Flex-TCF needs milder treatment to avoid negative effects to the substrate.

In order to be compatible with solution-based process and roll-to-roll fabrication, hydrophile/hydrophobic treatment is a more appropriate patterning method. Common hydrophile treatment includes UV/Ozone, plasma and chemical decoration. Defining the hydrophile/hydrophobic region first, then coat the dispersion of conducting material on the substrate and tends to assemble in the hydrophile region.

Practical Processes for Flexible Display

To make a mass-producible flexible device, the key processes of thin-film deposition, registration, patterning, surface treatment, and lamination should be developed. As Prof. Jun Souk analyzed, the ultimate approach is to use roll-to-roll (R2R) process (Souk and Roh, 2012) without vacuum ambience and photo-lithography. These patterning methods were mature in the newspaper or magazine printing industry.

See also: Displays – Introduction

References

- Fujiwara, D., Ishinabe, T., Koma, N., Fujikake, H., 2016. Thin flexible liquid crystal display using dye-type in-cell polarizer and pet substrates. *SID Digest* 47 (1), 18–22.
- Ishinabe, T., Obonai, Y., Fujikake, H., 2016. A foldable ultra-thin LCD using a coat-debond polyimide substrate and polymer walls. *SID Digest* 47 (1), 83–86.
- Lai, Y.H., Chan, C.C., Hwu, K.L., Huang, W.M., 2012. Photolithographic color filter for 14.1-inch flexible color electrophoretic displays. *SID Digest* 43 (1), 1365–1367.
- Lee, B., Byun, J., Oh, E., *et al.*, 2016. All-ink-jet-printed wearable information display directly fabricated onto an elastomeric substrate. *SID Digest* 47 (1), 672–675.
- Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56.
- Oka, S., Sasaki, T., Tamaru, T., *et al.*, 2016. Optical compensation method for wide viewing angle IPS LCD using a plastic substrate. *SID Digest* 47 (1), 87–90.
- Pin Hsiang, C., Wen yuan, L., Zheng Han, C., *et al.*, 2016. Roll TFT-LCD with 20R curvature using optically compensated colorless-polyimide substrate. *SID Digest* 47 (1), 15–17.
- Souk, J.H., Roh, N.S., 2012. Processing flexible displays. *SID Digest* 40 (2), 622–624.
- Telfer, S.J., McCreary, M.D., 2016. A full-color electrophoretic display. *SID Digest* 43, 574–577.
- Yang, B.R., Liu, G.S., Han, S.J., *et al.*, 2016. Coating, patterning, and transferring processes of silver nanowire for flexible display and sensing applications. *Journal of the Society for Information Display* 24, 234–240.

Nonuniqueness in Imaging

Greg Gbur, University of North Carolina at Charlotte, Charlotte, NC, United States

© 2018 Elsevier Ltd. All rights reserved.

Glossary

Computed tomography Also known as computed axial tomography, a classic technique by which multiple X-ray absorption images of an object are taken and digitally reconstructed into a two- or three-dimensional image of the object's internal structure.

Condition number The ratio of the largest singular value of a matrix to its smallest. A large condition number (much greater than unity) implies that inversion techniques involving the matrix are highly susceptible to noise.

Ewald limiting sphere Sphere in Fourier space that characterizes the complete information that can be determined in a scattering experiment if data are taken for all directions of incidence and illumination.

Ewald sphere of reflection Sphere in Fourier space that characterizes the information that can be determined in a scattering experiment for a single direction of illumination.

Ill-conditioning Refers to a linear inversion problem with a high condition number, which is therefore highly unstable in the presence of noise.

Ill-posedness Refers to a linear inversion problem in which at least one of the following conditions is not satisfied: existence, uniqueness, and continuity.

Inverse problem A physical problem in which a "cause" is deduced from measurements of an "effect," in essence inverting the usual cause-effect type of problems usually studied in physics.

Inverse scattering problem Problem in which the structure of a scattering object is determined by measurements of the field scattered by the object.

Inverse source problem Problem in which the structure of a primary radiation source is determined by measurements of the radiated field. Known to be nonunique due to the existence of nonradiating sources.

Noncontinuity Refers to a linear inversion problem in which the reconstructed object is highly sensitive to even small changes in the image data.

Nonexistence Refers to a linear inversion problem in which the measured data may not correspond to any known object, due to the presence of noise.

Nonradiating source A primary radiation source which does not, in fact, produce any radiation.

Nonuniqueness Refers to a linear inversion problem in which multiple objects produce the same image.

Phase problem Imaging problem in which a source distribution is determined from measurements of the intensity of the image alone.

Prior knowledge Additional knowledge, usually dictated by the physics of an inverse problem, in order to make the solution unique.

Singular value decomposition Mathematical technique by which any matrix may be decomposed in such a way that the singular values (essentially diagonal elements of the decomposition) can be used to assess the stability of any inversion technique.

Overview

The techniques of optical imaging may be considered a subset of the broad class of so-called inverse problems, in which a general "cause" of a phenomenon is deduced from measurements of the "effect," namely, the phenomenon itself. This process of cause determination reverses the usual study of physics, in which the "effect" is deduced from knowledge of the "cause." In this context, the latter case is usually referred to as the forward problem. A familiar example of an inverse problem is computed tomography, the determination of the internal structure of a patient from measurements of the intensity of X-rays transmitted through the body. The cause is the human body itself, while the effect is the absorption of X-rays.

Because of this reversal of the familiar causal process of physics, inverse problems are often beset with difficulties. Chief among these is nonuniqueness, in which the measured data is insufficient, even in principle, to uniquely determine the properties of the object. In this article, we discuss the manifestation of nonuniqueness in a general imaging system, and the typical means by which it is overcome, if possible. Examples of nonuniqueness are given in the form of the inverse source problem and the phase retrieval problem.

Matrix Formulation of Imaging and Inverse Problems

We introduce the mathematics of nonuniqueness by considering an abstract imaging system, as illustrated in Fig. 1. We assume that the object data is characterized by the vector $|o\rangle$ and that the image data are characterized by the vector $|i\rangle$; the action of the imaging system is described by the matrix A . Assuming a linear relation between object and image, we may express the imaging process as

$$|i\rangle = A|o\rangle \quad (1)$$

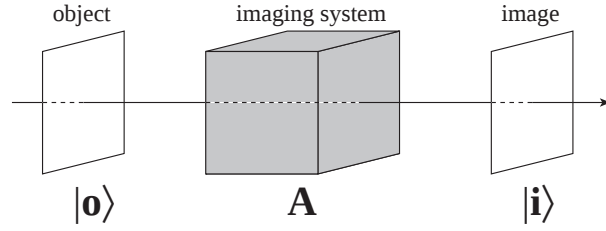


Fig. 1 Illustration of the notation for a general imaging system, in which an object vector $|o\rangle$ is converted to an image vector $|i\rangle$ through the action of an imaging system characterized by a matrix A .

It is to be noted that we have drawn the geometry in [Fig. 1](#) as a traditional imaging system, in which a planar object is mapped to a planar image through the action of a “black box” imager, such as a system of lenses. However, the formalism to be used here applies to much more general systems. For example, in computed tomography, the object may represent a three-dimensional (3D) map of an object’s absorption coefficient, and the image data may represent X-ray projection data acquired through multiple measurements with different orientations of the tomographic device. In general, we are considering systems in which the total measured data of the image is $|i\rangle$, and the desired information about the object is represented by $|o\rangle$.

It is also to be noted that the physics of such an imaging problem is usually described by continuous functions and operators; that is, the object $o(\mathbf{r}')$ is mapped to an image $i(\mathbf{r})$ through an integration with an integral kernel $A(\mathbf{r}, \mathbf{r}')$ over the domain D of the object, such that

$$i(\mathbf{r}) = \int_D A(\mathbf{r}, \mathbf{r}') o(\mathbf{r}') d^2 r' \quad (2)$$

Since practical imaging problems inevitably require the discretization of such integrals, however, [Eq. \(1\)](#) is sufficient for our description.

In the presence of noise, we may modify [Eq. \(1\)](#) by adding a noise vector $|n\rangle$ to the image, with the result

$$|i\rangle = A|o\rangle + |n\rangle \quad (3)$$

It is easiest to analyze the case for which A is a square $N \times N$ matrix, but more general $M \times N$ matrices may also be considered. When $M > N$, the image data collected is less than the amount of information available about the object, and it is expected that only a limited reconstruction of the object is possible. When $M < N$, more image data are collected than are in principle useful for reconstruction.

For a square Hermitian matrix, the properties of the imaging system may be characterized by the eigenvalues and eigenvectors of the matrix. However, as an imaging matrix is not necessarily Hermitian nor square, it is more appropriate to instead represent it using the singular value decomposition (SVD), given by

$$A = \sum_{i=1}^r s_i |y_i\rangle \langle x_i| \quad (4)$$

where $r = \min(M, N)$, the vectors $|y_i\rangle$ are the eigenvectors of the Hermitian $m \times m$ matrix $Y = AA^\dagger$, the vectors $|x_i\rangle$ are the eigenvectors of the Hermitian $n \times n$ matrix $X = A^\dagger A$, and $s_i = (x_i)^{1/2}$, with x_i the eigenvalues of the matrix X (or Y , if $N < M$). The vectors $|y_i\rangle$ are orthonormal, so that $\langle y_i | y_j \rangle = \delta_{ij}$, and also $\langle x_i | x_j \rangle = \delta_{ij}$. By [Eq. \(4\)](#), it can be readily seen then that

$$A|x_i\rangle = s_i|y_i\rangle, \quad A|y_i\rangle = s_i|x_i\rangle \quad (5)$$

The singular values are taken to be nonnegative by definition, which suggests that the signs of the eigenvectors must be taken appropriately. In matrix form, the SVD may be written as

$$A = USV^\dagger \quad (6)$$

where U is an $m \times m$ unitary matrix made up of the eigenvectors $|y_i\rangle$ of Y , V is an $n \times n$ unitary matrix made up of the eigenvectors $|x_i\rangle$ of X , and S is an $m \times n$ “diagonal” matrix such that $S_{ij} = \delta_{ij}s_i$. More information about the SVD can be found in [Trefethen and Bau \(1997\)](#).

We may apply this decomposition to formally find a solution to the imaging equation ([Eq. \(1\)](#)). Assuming for the moment that all the singular values $s_i \neq 0$, a solution is given by

$$|o\rangle = V\hat{S}U^\dagger|i\rangle \quad (7)$$

where $\hat{S}$ is an $n \times m$ diagonal matrix with diagonal elements equal to $1/s_i$. The meaning of this solution is elaborated upon in the next section.

Nonuniqueness and Prior Knowledge

Let us now restrict ourselves to the ideal case in which $m=n$; the matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$ are all square $m \times m$ matrices. The ability to solve our imaging problem then depends directly on the properties of the singular values s_i , and not on an overabundance or dearth of data.

An imaging problem is said to be well-posed if the solution exists, is unique, and is continuous. Any problem which fails to meet all of these conditions is said to be ill-posed, in a sense that was first introduced by Hadamard (1902). We next consider each of these conditions related to ill-posedness and how they are related, as well as basic strategies to ameliorate them.

The imaging system $\mathbf{A}$ may be viewed as a general mapping of the object data to the image data. The object data lies within an m -dimensional space called the object space, while the image data lies within another m -dimensional space called the image space. The set of all possible objects is referred to as the domain, while the set of all possible images is known as the range. The domain may be a subset of the object space, depending on how it is defined: for example, the object space might be considered as the set of all two variable functions, while the object domain might be defined as the set of all finite-valued two variable functions. The dimensionality of either the domain or range may be lower than that of the space in which it is embedded – or both. These regions, and the problems of nonuniqueness, noncontinuity, and nonexistence, are illustrated in Fig. 2.

Nonuniqueness is the case where two or more objects have exactly the same image. This arises when one or more of the singular values $s_i=0$. An object $|\mathbf{o}\rangle$ is said to lie in the null space of the imaging system if

$$\mathbf{A}|\mathbf{o}\rangle = 0 \quad (8)$$

Any object vector in the null space will contribute nothing to the image; therefore, if $\mathbf{o}_1$ represents an object, and $\mathbf{o}_2$ represents a null space vector, then $\mathbf{o}_1$ and $\mathbf{o}_1 + \mathbf{o}_2$ will have exactly the same image. Without additional information (to be discussed momentarily), there is no way to determine the contribution of the null space vector, or vectors, to the object. The values $s_i=0$ clearly make the inversion matrix $\hat{\mathbf{S}}$ singular, as it would contain terms of the form $1/s_i$.

As will be made explicit by example in the next section, the presence of nonuniqueness in an imaging problem may therefore be tied to the existence of invisible objects. An object which lies entirely in the null space will have no detectable image; it is invisible. Nonuniqueness in an imaging problem therefore implies the existence of invisible objects, and vice versa.

Nonexistence is closely related to nonuniqueness. The existence of null space vectors results in the range of the image having a smaller dimensionality than the domain of the object. There are, therefore, vectors in the image space that simply cannot be produced by an object. In a perfect imaging system, this would not be a problem; however, in the presence of additive noise $|\mathbf{n}\rangle$, it is possible for the image vector to be moved out of the range, in which case there exists no corresponding object.

Noncontinuity is a situation in which two very different objects produce very similar images. Strictly speaking, noncontinuity refers only to an integral imaging model, such as given by Eq. (2), in which the reconstructed object function $o(\mathbf{r})$ does not depend continuously on variations of the image function $i(\mathbf{r})$. In a discrete matrix model, the analogous problem is known as ill-conditioning, and arises when at least one of the singular values is extremely small compared to the others. To characterize this, we introduce the condition number κ , defined as the ratio of the largest singular value to the smallest,

$$\kappa \equiv \frac{s_{\max}}{s_{\min}} \quad (9)$$

A condition number which is not too much greater than unity (within two orders of magnitude) indicates a well-conditioned system, while an extremely large condition number represents ill-conditioning. Because the matrix $\hat{\mathbf{S}}$ depends on $1/s_i$, any vector corresponding to a very small singular value will be strongly amplified. In the presence of noise, this means that noise will also be highly amplified in an ill-conditioned system and will tend to drown out object information.

To resolve the issue of nonuniqueness, one must typically introduce prior knowledge based on known physics of the imaging system, as illustrated schematically in Fig. 2. Often, only a small class of mathematically achievable objects are physically reasonable, and restricting to this class will in some cases make the imaging problem unique; in other cases, additional prior knowledge may be needed. In phase retrieval problems, for instance, the size of the object is generally known and any

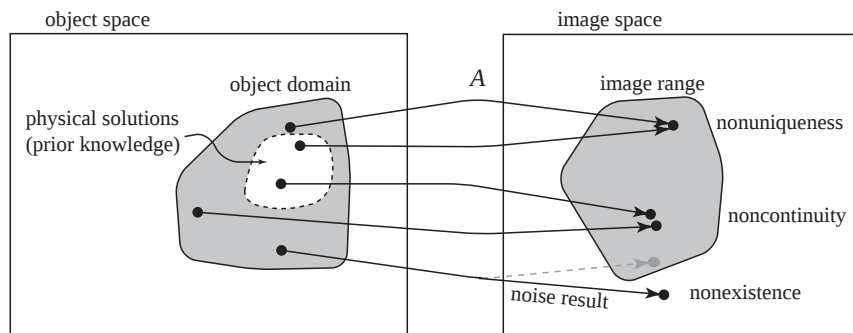


Fig. 2 Illustration of the different types of ill-posedness that can appear in an imaging system.

reconstruction which is much larger than this, or infinite, is rejected. Another commonly used condition used as prior knowledge is that the “physical” solution $|\mathbf{o}_t\rangle$ have minimum “energy,” i.e.,

$$\langle \mathbf{o}_t | \mathbf{o}_t \rangle \leq \langle \mathbf{o} | \mathbf{o} \rangle \text{ for any acceptable } \mathbf{o} \quad (10)$$

Prior knowledge needs to be chosen carefully based on the imaging problem being solved; it is possible to find a unique solution, which is nevertheless not useful or even closely related to the desired object.

In the context of the SVD, the best possible solution is often defined as the solution $|\hat{\mathbf{o}}\rangle$ which minimizes the quantity L , defined as

$$L \equiv |A|\hat{\mathbf{o}}\rangle - |\mathbf{i}\rangle| \quad (11)$$

It can be shown that this minimum solution is given by

$$|\mathbf{o}\rangle = V\hat{S}U^\dagger|\mathbf{i}\rangle \quad (12)$$

where $\hat{S}$ is the $n \times m$ diagonal matrix with diagonal elements equal to $1/s_i$ for $s_i \neq 0$ and 0 for $s_i = 0$; this is a generalization of Eq. (7). In essence, this solution calculates only those parts of the object that do not lie in the null space and ignores the rest. The solution can be further conditioned by setting the diagonal elements equal to zero for $s_i < S$, where S is a chosen critical value depending on practical properties of the system. In principle, throwing away those components of the solution which have small singular values results in a loss of information; however, as we have noted, those components would typically be corrupted by noise and overall be detrimental to the solution in any case.

The minimum solution of Eq. (12) in practice acts as an application of prior knowledge, though it may be said to be philosophically different. The minimum solution is unique because it simply ignores the null space of the operator A ; ideally, prior knowledge is used to deduce the part of the solution which lies in the null space.

The Inverse Source Problem

One of the simplest, and perhaps most infamous, examples of nonuniqueness in an imaging problem is found in the inverse source problem. The geometry and related notation is illustrated in Fig. 3. We consider a 3D monochromatic primary source of light, confined to a domain D , defined by the function $q(\mathbf{r})$ and which produces a monochromatic field $u(\mathbf{r})$. The field and source are related by the 3D Helmholtz equation,

$$[\nabla^2 + k^2]u(\mathbf{r}) = -4\pi q(\mathbf{r}) \quad (13)$$

where $k = \omega/c$, ω is the frequency and c is the vacuum speed of light. We consider the scalar wave equation for simplicity, but similar arguments can be applied to a full electromagnetic radiation problem.

The solution for the field everywhere inside and outside the source domain is given by

$$u(\mathbf{r}) = \int_D \frac{e^{ikR}}{R} q(\mathbf{r}') d^3 r' \quad (14)$$

with $R = |\mathbf{r} - \mathbf{r}'|$. At a distance far away from the source, we may make the far zone approximation,

$$\frac{e^{ikR}}{R} \approx \frac{e^{ikr}}{r} e^{-iks \cdot \mathbf{r}'} \quad (15)$$

where $\mathbf{r} = r\mathbf{s}$, with $\mathbf{s}$ a unit vector pointing in the direction observation. In the far zone, then, the field may be written as

$$u(r\mathbf{s}) \approx \frac{e^{ikr}}{r} \int_D q(\mathbf{r}') e^{-iks \cdot \mathbf{r}'} d^3 r' \quad (16)$$

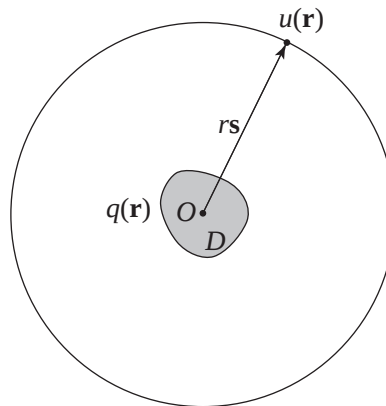


Fig. 3 Illustration of the notation and geometry of the inverse source problem.

The integral is now in the form of a 3D spatial Fourier transform, i.e.,

$$\tilde{q}(\mathbf{K}) = \frac{1}{(2\pi)^3} \int q(\mathbf{r}') e^{-i\mathbf{K} \cdot \mathbf{r}'} d^3 r' \quad (17)$$

In terms of this Fourier transform, we may write the far zone field as

$$u(r\mathbf{s}) \approx (2\pi)^3 \frac{e^{ikr}}{r} \tilde{q}(k\mathbf{s}) \quad (18)$$

It should be clear from this expression that any information about the source is contained in the Fourier transform, which depends only on the angular direction $\mathbf{s}$.

The inverse source problem may then be posed as an imaging problem as follows: given measurements of the field $u(r\mathbf{s})$ for all directions $\mathbf{s}$, determine the 3D structure $q(\mathbf{r})$ of the source. This problem has been shown to be severely nonunique due to the theoretical existence of invisible objects, known in this context as nonradiating sources. From Eq. (18), it can be seen that a nonradiating source is one for which

$$\tilde{q}(k\mathbf{s}) = 0 \text{ for all } \mathbf{s} \quad (19)$$

An example of such a nonradiating source is not difficult to come by (Kim and Wolf, 1986). Let us imagine a spherical source of radius a and uniform charge density q_0 , such that

$$q(\mathbf{r}) = \begin{cases} q_0, & |\mathbf{r}| \leq a \\ 0, & |\mathbf{r}| > a \end{cases} \quad (20)$$

The far zone field radiated by this source can be readily shown to be of the form,

$$u(r\mathbf{s}) = 2\pi \frac{e^{ikr}}{r} j_1(ka) \quad (21)$$

where $j_1(x)$ is the spherical Bessel function of order 1. The far zone field will be identically zero, i.e., the source will be nonradiating, if ka is equal to any of the zeros of $j_1(ka)$. The first zero of this Bessel function is at $ka = 4.49$, implying that the smallest nonradiating source of this structure has a radius $a = 4.49/k$.

Though radiationless sources have a long history in physics (Gbur, 2013), the influence of such sources on the inverse source problem was not clear until the 1980s, when a heated scientific argument clarified the matter (see Devaney and Sherman (1982) and the comments which follow).

It may be readily seen that the inverse source problem is what we might call fatally nonunique: the information present in the radiated field provides almost no information about the object structure. This can be seen by introducing a new nonradiating condition equivalent to Eq. (19). We multiply Eq. (16) by $\exp[i\mathbf{k} \cdot \mathbf{r}]$, and then integrate overall values of the unit vector $\mathbf{s}$; the result is a necessary and sufficient condition for a source to be nonradiating, of the form

$$\int_D q(\mathbf{r}') j_0(k|\mathbf{r} - \mathbf{r}'|) d^3 r' = 0 \text{ for all } \mathbf{r} \quad (22)$$

The spherical Bessel function $j_0(k|\mathbf{r} - \mathbf{r}'|)$ has a well-known modal expansion, namely,

$$j_0(k|\mathbf{r} - \mathbf{r}'|) = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^n j_n(kr) j_n(kr') Y_n^{m*}(\theta, \phi) Y_n^m(\theta', \phi') \quad (23)$$

where $Y_n^m(\theta, \phi)$ are the spherical harmonics.

Let us now consider a general spherical source $q(\mathbf{r})$ of radius a . Integrated over such a spherical domain, the spherical harmonics are orthogonal. The condition given by Eq. (22) is therefore equivalent to the requirement that the nonradiating source $q(\mathbf{r}')$ be orthogonal to the infinite set of modes given by $j_n(kr') Y_n^m(\theta', \phi')$. However, for a fixed value of n and m , i.e., a fixed angular dependence, it is clear that there are an infinite number of orthogonal modes with the same angular dependence but different radial dependence (see examples given in Gbur and Wolf (1997)). It may be loosely stated that, for each mode that produces a measurable radiation pattern, there are an infinite number of modes with the same angular dependence that do not radiate.

Because of the fatal nonuniqueness of the inverse source problem, it is not generally possible to deduce anything significant about the source structure from the radiation pattern. The problem can be made unique by assuming (Porter and Devaney, 1982) that the source has minimum energy E , i.e., by minimizing the functional

$$E = \int_D |q(\mathbf{r}')|^2 d^3 r' \quad (24)$$

but this condition gives the same result as the SVD problem of Eq. (11), in that all the nonradiating source components are identically zero. This minimum energy condition does not seem to hold in realistic radiation problems.

Often, though not always, the uniqueness or nonuniqueness of an imaging problem can be determined by a simple dimensional analysis. From Eq. (18), it is to be noted that a measurement of the complete radiation pattern of a 3D source $q(\mathbf{r})$ determines the Fourier components $\tilde{q}(k\mathbf{s})$ on a spherical shell in 3D Fourier space. Because the function $\tilde{q}(\mathbf{K})$ is the Fourier transform of a function of finite extent, it follows that it is the boundary value of an analytic function in three complex variables,

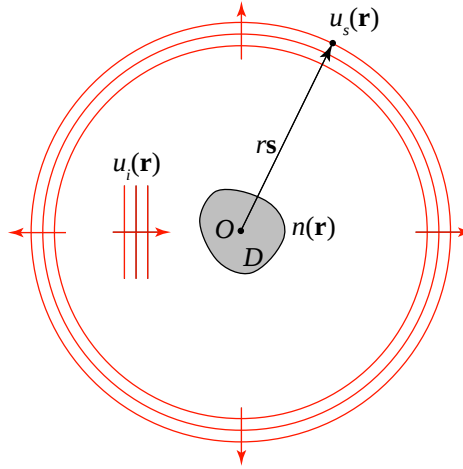


Fig. 4 Illustration of the notation and geometry of the inverse scattering problem.

K_x , K_y , and K_z . However, such a function can only be analytically continued if information is known in a 3D volume of space: the information on the two-dimensional (2D) shell is not sufficient to analytically continue the existing Fourier data to a volume.

Nonuniqueness can sometimes be the result of an insufficiently diverse dataset. This can be illustrated by the inverse scattering problem under weak scattering conditions, as illustrated in [Fig. 4](#). An incident field $u_i(\mathbf{r})$ impinges on a localized scattering object of refractive index $n(\mathbf{r})$, producing a scattered field $u_s(\mathbf{r})$. The total field $u(\mathbf{r}) = u_i(\mathbf{r}) + u_s(\mathbf{r})$, and the incident field is defined as the field that would exist in the complete absence of the scatterer.

The total field satisfies the Helmholtz equation with inhomogeneous wave number,

$$[\nabla^2 + n^2(\mathbf{r})k^2]u(\mathbf{r}) = 0 \quad (25)$$

As is shown in [Born and Wolf \(1999, Chapter 13\)](#), this equation can be written in the form

$$[\nabla^2 + k^2]u_s(\mathbf{r}) = -4\pi F(\mathbf{r})u(\mathbf{r}) \quad (26)$$

where $F(\mathbf{r})$ is the scattering potential, given by

$$F(\mathbf{r}) = \frac{k^2}{4\pi} [n^2(\mathbf{r}) - 1] \quad (27)$$

[Eq. \(26\)](#) is formally the same as [Eq. \(13\)](#) for the source problem, and the far zone solution may be found in the same way to be

$$u_s(r\mathbf{s}) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u(\mathbf{r}') e^{-ik\mathbf{s} \cdot \mathbf{r}'} d^3r' \quad (28)$$

The difficulty with this solution is that the scattered field appears on both sides of the equation, implicitly in $u(\mathbf{r})$, making [Eq. \(28\)](#) an integral equation that cannot in general be solved. If one assumes that the scattered field is extremely weak, however, then $u(\mathbf{r}') \approx u_i(\mathbf{r}')$, and we have

$$u_s(r\mathbf{s}) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u_i(\mathbf{r}') e^{-ik\mathbf{s} \cdot \mathbf{r}'} d^3r' \quad (29)$$

Let us now further assume that the incident field is a plane wave propagating in a direction specified by the unit vector $\mathbf{s}_0$, of the form $u_i(\mathbf{r}) = \exp[ik\mathbf{s}_0 \cdot \mathbf{r}]$. Then the scattered field is of the form,

$$u_s(r\mathbf{s}) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u_i(\mathbf{r}') e^{-ik(\mathbf{s}-\mathbf{s}_0) \cdot \mathbf{r}'} d^3r' = (2\pi)^3 \frac{e^{ikr}}{r} \tilde{F}[k(\mathbf{s} - \mathbf{s}_0)] \quad (30)$$

The scattered field measured in a direction $\mathbf{s}$ therefore gives information about the Fourier components of the field at position $k(\mathbf{s} - \mathbf{s}_0)$; if measurements are made for all directions $\mathbf{s}$, one gets Fourier information on a sphere of radius k centered on $k\mathbf{s}_0$, known as an Ewald sphere of reflection. However, as in the inverse source case, the information on this Ewald sphere is not sufficient to reconstruct the scattering potential $F(\mathbf{r})$. If the scattered field is measured for all directions of incidence $\mathbf{s}_0$ and scattering $\mathbf{s}$, essentially over a continuum of Ewald spheres, one acquires information within a sphere of radius $2k$, known as the Ewald limiting sphere Σ_L . This geometry is illustrated in [Fig. 5](#).

With the Fourier components determined within a 3D volume, it is possible to perform an inverse Fourier transform to determine a bandlimited reconstruction of the scattering object.

As the information on any finite number of Ewald spheres of reflection is insufficient to uniquely reconstruct the scattering object, one would expect that it is possible to make objects which are invisible, or nonscattering, for any finite number of directions of illumination, and this was in fact shown to be the case ([Devaney, 1978](#)). Evidently, as more data is taken, these nonscattering solutions play a correspondingly smaller role in the reconstruction process.

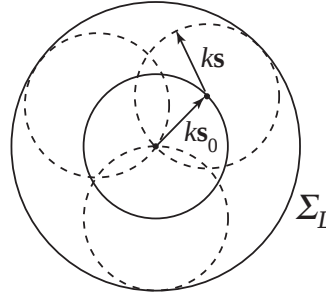


Fig. 5 Illustration of several Ewald spheres of reflection and the Ewald limiting sphere Σ_L .

Phase Retrieval

The discussion so far has been concerned with linear imaging systems, in which the image data has a linear relation to the object data. However, ill-posedness, in particular nonuniqueness, can also appear in nonlinear imaging, and the analysis of such systems is much less straightforward and must be dealt with on a case-by-case basis.

We consider here the so-called phase problem, or the phase retrieval problem, in which an object must be reconstructed from measurements of the image intensity alone. We assume that the object $f(x, y)$ and image $F(u, v)$ are related by a simple Fourier relation, i.e.,

$$F(u, v) = \int_S f(x, y) e^{-i2\pi(ux+vy)} dx dy \quad (31)$$

where S represents the object plane. The intensity of the image is given by $I(u, v) = |F(u, v)|^2$, and we are concerned with determining $f(x, y)$ from this data alone, if possible.

The nonlinear nature of this problem, due to the implicit reconstruction of the phase, results in some unexpected results. One of the most notable and famous is the observation that the problem is nonunique for one-dimensional (1D) systems but typically unique in 2D systems; we now examine how this is possible.

In considering the uniqueness of this imaging problem, we follow the discussion and notation of [Bruck and Sodin \(1979\)](#). The problem is almost certainly nonunique in general, and certain prior knowledge needs to be applied immediately. It is typically assumed that the source distribution is of finite extent and takes on only nonnegative and finite values.

The source is taken to consist of discrete points, represented by Dirac delta functions,

$$f(x, y) = \sum_{j=1}^m \sum_{k=1}^n f_{jk} \delta(x - x_j) \delta(y - y_k) \quad (32)$$

where $x_j = \Delta j$ and $y_k = \Delta k$, with Δ the spacing between points. The image data are then in the form of a discrete Fourier transform,

$$F(u, v) = \sum_{j,k} f_{jk} e^{-i2\pi\Delta(ju+kv)} \quad (33)$$

where for convenience we have used the Fourier transform in the form,

$$F(u, v) = \int f(x, y) e^{-i2\pi(ux+vy)} dx dy \quad (34)$$

The expression (33) may be simplified by writing the complex exponentials as complex variables, namely,

$$z \equiv e^{-i2\pi\Delta u}, \quad w \equiv e^{-i2\pi\Delta v} \quad (35)$$

where z and w lie on the unit circle in the complex plane. In this way, the image data may be represented as a polynomial in z and w ,

$$F(u, v) \equiv p(z, w) = \sum_{j,k} f_{jk} z^j w^k \quad (36)$$

The measured intensity of the image is given by the squared modulus of $p(z, w)$; since $|z| = |w| = 1$, $z^* = z^{-1}$ and $w^* = w^{-1}$, which means that we may write

$$p(z, w) p^*(z, w) = p(z, w) p(z^{-1}, w^{-1}) \quad (37)$$

From basic Fourier theory, we know that any spatial shift in the overall object position by $s\Delta$ along x and $t\Delta$ along y will result in the polynomial $p(z, w) \rightarrow p(z, w) z^s w^t$. The same argument applies to the measured intensity. We define the autocorrelation polynomial of the image as

$$Q(z, w) = p(z, w) p(z^{-1}, w^{-1}) z^m w^n \quad (38)$$

The additional powers $z^m w^n$ make this a true polynomial, unlike the product of Eq. (37), which will possess positive and negative powers of z and w . As noted, however, the additional powers only modify the object by a trivial spatial shift.

Let us now consider the 1D phase problem, in which the object is $f(x)$ and the image polynomial is of the form

$$p(z) = \sum_{j=0}^m f_j z^j \quad (39)$$

with $f_j \geq 0$. By the fundamental theorem of algebra, any such polynomial may be factored into a product of m terms,

$$p(z) = \prod_{j=1}^m (z - z_j) \quad (40)$$

With the constraint that the $f_j \geq 0$, the zeros z_j of this polynomial must either satisfy $\text{Re}\{z_j\} \leq 0$ and $\text{Im}\{z_j\} = 0$, or must come in a complex conjugate pair, i.e., if z_j is a zero, then so is z_j^* . This must be the case because the coefficients f_j will feature every combination of products of the z_j 's.

With this in mind, the autocorrelation function may be written as

$$Q(z) = \prod_{j=1}^m (z - z_j)(z - z_j^{-1}) \quad (41)$$

and we note that the phase problem may be stated as determining the coefficients f_j from the values of this autocorrelation function, or equivalently determining the z_j 's. The function $Q(z)$ is found by direct measurement, but it features both z_j and z_j^{-1} as its zeros. If all $|z_j| = 1$, then both z_j and z_j^* must be zeros of the polynomial $p(z)$; the solution is uniquely determined. If all $|z_j| \neq 1$, then it is not possible from the autocorrelation function to determine if a measured z_j or its inverse is the true zero of the polynomial $p(z)$. Since there are two choices for each of the zeros, there are 2^m possible solutions for $p(z)$ for the given $Q(z)$ —the solution is nonunique. If there are r zeros with $|z_j| = 1$, then there will be 2^{m-r} possible solutions, and the solution is still nonunique. In general it can be seen that the phase problem is nonunique in 1D.

For the 2D case, however, the situation is quite different. This arises from the fact that the polynomial $p(z, w)$ cannot be represented in a factorized form,

$$p(z, w) \neq \prod_{j,k} (z - z_j)(w - w_k) \quad (42)$$

If it could, we could make arguments similar to the 1D case to prove nonuniqueness. It is very easy, though, to find examples which do not satisfy Eq. (42). For instance, the polynomial

$$p(z, w) = z^2 + w^3 \quad (43)$$

cannot be written in the form of Eq. (42). Without factorization, the earlier nonuniqueness argument breaks down. This alone does not demonstrate that the problem becomes unique, but experimental and computational studies have shown that it is generally so. Of course, in the rare case that the image polynomials can be factorized, uniqueness will fail.

Reconstruction of the phase can be done in many cases by the venerable algorithm of [Gerchberg and Saxton \(1972\)](#), though this method also runs into issues of nonuniqueness when the data possesses certain symmetries.

References

- Born, M., Wolf, E., 1999. *Principles of Optics*, seventh ed. Cambridge: Cambridge University Press.
- Bruck, Y., Sodin, L., 1979. On the ambiguity of the image reconstruction problem. *Optics Communications* 30, 304–308.
- Devaney, A., 1978. Nonuniqueness in the inverse scattering problem. *Journal of Mathematical Physics* 19, 1526–1531.
- Devaney, A., Sherman, G., 1982. Nonuniqueness in inverse source and scattering problems. *IEEE Transactions on Antennas and Propagation* 30, 1034–1037.
- Gbur, G., 2013. Invisibility physics: Past, present, and future. In: Wolf, E. (Ed.), *Progress in Optics*, vol. 58. Amsterdam: Elsevier, pp. 65–114.
- Gbur, G., Wolf, E., 1997. Sources of arbitrary states of coherence that generate completely coherent fields outside the source. *Optics Letters* 22, 943–945.
- Gerchberg, R., Saxton, W., 1972. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35, 1–6.
- Hadamard, J., 1902. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin* 13, 49.
- Kim, K., Wolf, E., 1986. Non-radiating monochromatic sources and their fields. *Optics Communication* 59, 1–6.
- Porter, R., Devaney, A., 1982. Holography and the inverse source problem. *Journal of the Optical Society of America* 72, 327–330.
- Trefethen, L., Bau, D., 1997. *Numerical Linear Algebra*. Philadelphia, PA: SIAM.

View Angles of Liquid Crystal Displays

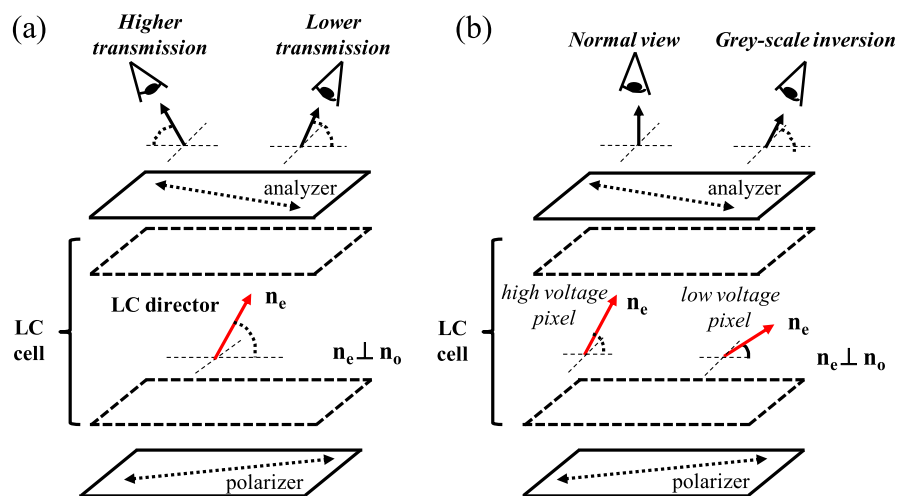
Huang-Ming Philip Chen, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

The viewing angle of liquid crystal displays is defined as the maximum angle at which a display panel can be viewed within acceptable visual performance when a contrast ratio exceeds the minimum ratio of 10:1. The viewing angle is measured from one position to the opposite direction, which gives a maximum of 180° in one axis. Four viewing angles need to be taken into account: the horizontal (0°), vertical (90°), and $\pm 45^\circ$ axes. The detail measurements can be performed either by hand or by a conoscope at smaller increment degrees. The contour data mapping exhibits the viewing angle and viewing cone in displays. The narrow viewing angle of twist nematic liquid crystal displays (TN LCDs) is an intrinsic issue of LCDs due to the various retardation changes that are associated with LC molecular motion under electric field, also known as birefringent mode. The contrast ratio is reduced by the light leakage at oblique angle caused by elliptical polarization states. Two major approaches have been applied side by side to solve the issue. One approach is working on optical compensation films to compensate polarization loss in LC cells (Mori *et al.*, 1997). The other is dedicating in LC driving modes and pixel designs to achieve spatial effect in different viewing angles, such as In-Plane-Switching (IPS), or Multi-Domain Vertical Alignment (MVA) modes. Due to the manufacture limitation, both approaches need to work closely to achieve desired wide viewing angle LCDs. In addition to the enhancement of viewing angle in LCDs, many new optical designing concepts have been proposed and realized in the latest researches beyond the aforementioned framework to control the viewing angles that fit the specific needs of display applications.

Optical Compensation Film

The key of designing optical compensation film is based on the LC device in a particular driven state, which is associated with the birefringent mode in LC optics. When the linear polarized light passes through the LC medium, it also loses its linear polarization state and turns into elliptically polarized light. The long axis of elliptically polarized light does not closely follow the local LC director. The light leakage occurs when elliptically polarized light passes through the second polarizer (or called analyzer) at oblique angle. The polarization states of oblique angle can be calculated and mapped on the surface of a Poincaré Sphere. The calculation of correction leading to the linear polarization point can be applied as an optical compensation (Zhu *et al.*, 2006). The directors' profile is taken into consideration when designing the optical compensation film accordingly. It is important to note that this compensation method cannot resolve all viewing angle issues, especially in the gray-scale inversion, which requires divided pixels in domains with different orientations. A pixel divided into two domains may provide a significant reduction in the gray-scale inversion. A pixel that splits into four domains can give relatively symmetrical appearances to improve both gray level and contrast in a panel.



An optical compensation film (or so called Wide View film) is characterized by its refractive indices n_x , n_y , and n_z , in which n_z is the refractive index in the direction perpendicular to the film; n_x and n_y are the in-plane at X and Y axis. In order to compensate an on-state of TN LCD, an axial symmetric compensator with negative birefringence ($n_x = n_y > n_z$) can be applied. In general, the compensation films can be classified by its optical characteristic, such as A, C, O-plate, and biaxial films. Based on the specific manufacture processes, these films can also be described, as stretching films, or LC coated films.

	<i>A plate</i>	<i>C plate</i>
Positive (+)	$n_x > n_y = n_z$	$n_z > n_x = n_y$
Negative (-)	$n_x = n_z > n_y$	$n_x = n_y > n_z$

Many references have indicated the viewing angle issues can be solved by various optical compensation films for different LCD modes. In general, the viewing angle of TN LCD mode can be compensated by bend-splay hybrid orientation discotic material coated films. The viewing angle of MVA mode can be compensated by either biaxial film itself or positive A (A^+) plate in conjunction with negative C (C^-) plate. The IPS mode, on the other hand, requires either negative A (A^-) plate or positive C (C^+) plate to enhance its viewing angle. The complex structures of compensation films (i.e., multiple optical composite layers) were proposed in some reports to solve the finite polarization deviation in the LC cells. These calculated solutions may be unrealistic when facing limited manufacturing ability or rising costs. The Fuji film Corporation (Ito *et al.*, 2013) has revealed their significant Wide View film (WV film) development in 2013 (Ito *et al.*, 2013). The hybrid alignment of polymerized discotic materials is the main component of compensation film for TN LCDs. The grayscale inversion can be significantly minimized because the LCs' director in the middle of the cell is parallel to the axis of the polarizer. The retardation of the LC has least transmittance effect when the darker gray level moving to the brighter gray level. As a result, controlling the orientation of polymerized discotic materials is the key to formulate new WV films in improving the TN-LCDs' gray scale image qualities. In addition, they also developed a precise patterning chemical method by using one shot of non-polarized UV light to control and create patterned discotic alignments. The resulting film is called film patterned retarder (FPR) with out-of-plane retardation (R_{th}) close to 0 nm. The advantage of FRP is to reduce the oblique cross-talk in 3D display applications. Furthermore, the achromatic quarter wave films can be prepared by coating the rod-like LC layer on top of the polymerized discotic materials. Their achromatic quarter wave film measurement outcome suggests that the retardation is similar to the ideal curve below 550 nm. The deviation is much smaller than one-film type achromatic quarter wave film in the range of 550–650 nm. This new achromatic quarter wave films are great assets to LCD as well as to OLED for the true black appearance.

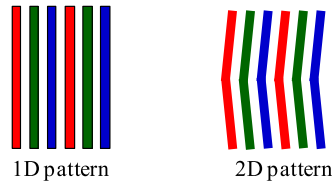
The optical isotropic Blue Phase LCD is assumed to have a good dark state and high contrast ratio in theory. The undesired light leakage occurs in the voltage off-state. Wu *et al.* explains that the causes of the dark-state light leakage are coming from the edge of protrusive electrodes, polarization rotation effect of BPLC material, and scattering from polymer network (Liu *et al.*, 2014). In this article, it states that the major source of light leakage comes from the polarization rotation effect in the dark state. Fortunately, this light leakage can be compensated by the Poincaré Sphere method. There are three major solutions using compensation films to suppress light leakage for the polarization rotation effect: (1) Small angle rotation of analyzer with biaxial film compensates the light leakage. This method can only reach the contrast ratio as high as 3700:1 in the normal direction. This simple approach faces more challenges when one wishes to further improve the contrast ratio through compensation. (2) Utilizing broadband and wide view circular polarizers to compensate the light leakage is proposed by Hong. The contrast ratio at normal direction can be as high as 12,000:1; and the viewing angle can be as wide as the entire viewing zone (CR 100:1). This approach requires that the two layers of biaxial films to form one circular polarizers for one side. Consequently, there are four layers of biaxial films that need to be aligned in order to obtain good results. The complicated film alignments may hinder one to apply this method for compensating the light leakage. (3) Using dispersive positive A (A^+) film is another method to compensate the light leakage. The birefringence of dispersive positive A film is designed to vary from 450 to 650 nm, for example: 0.008 for 450 nm, 0.005 for 550 nm, and 0.004 for 650 nm. The last method which requires not only dispersive positive A film, but also biaxial films is the same method as (1) and (2). The result of BPLCD can reach contrast ratio at normal direction as high as 13,500:1 with relatively wide viewing zone.

The flexible LCDs research becomes much more active recently in competing with the flexible OLED devices. The capability of flexible LCDs remains challenged. Many research groups have turned their attentions toward the enhancement of the optical performance that may exclusively occur in the flexible LCD. Ishinabe *et al.* (2014) published the article indicating that the contrast ratio of 650:1 and a wide viewing angle of 150° based on polycarbonate substrates in 2014. Jeong *et al.* (2016) attempted to eliminate the bending-induced retardation of a plastic film by coating polystyrene (PS) or poly (methyl methacrylate) (PMMA) on the polycarbonate substrates. The negative birefringence of the coated PS or PMMA layer is formed when bending the polycarbonate substrates. As a result, the bending-induced retardation from polycarbonate substrate can be compensated. Oka *et al.* (2016) (Japan Display) revealed their design using the polyimide (PI) and acrylic resin layers as C^+ -plate to compensate their flexible FFS LCD in 2016. The retardation can be controlled by the thickness of PI layer. The positive C (C^+)-plate ($R_{th} = -450$ nm) can be reached when PI is at 3.4 μm in thickness. A 2.95-in. prototype sheet LCD with contrast ratio of 860:1 in 400 μm panel thickness was demonstrated.

Improvement From LC Modes

Various nematic LC modes have been proposed and realized for the wide view liquid crystal displays. Hitachi manufactured the first super-TFT liquid crystal display (was later called In-Plane-Switching (IPS) mode) to enhance the viewing angle back in 1996. It was evolved from IPS mode to super IPS (S-IPS) mode to solve the color shifting issue in 1998, followed by the advance super IPS (AS-IPS) to enhance the transmittance in 2002. In order to solve the aspect ratio, transmittance, and contrast issues, the concept of Fringe Field Switching (FFS) was proposed by Lee in 1998 and was put into mass production of the IPS-Pro LCD TV

(as the abbreviation of IPS-Provectus) at the end of 2004. The FFS mode has over 3 times of higher contrast and 1.5 times of higher transmission than original IPS mode. Besides the electrode design, the optical compensation films still require for IPS or FFS modes to further enhance their viewing angle at $\pm 45^\circ$ and $\pm 135^\circ$. Due to the high demand of portable LC devices, p-FFS and n-FFS were proposed by using positive (p) and negative (n) dielectric LC materials. The pros and cons between p-FFS and n-FFS were mentioned by Wu (Chen *et al.*, 2015). In short, p-FFS gains the response time, yet n-FFS achieves the higher transmittance. Both have excellent viewing angle and minor color shift. The n-FFS has smaller image flicker, but it needs a 2D electrode patterned design to suppress its grayscale inversion.



Multi-Domain Vertical Alignment (MVA) liquid crystal display using negative dielectric LC to achieve wide view display was first developed by Fujitsu in 1997. The protrusion electrode divided one pixel into 4 domains. Single sided protrusion and the open slit on the counter side MVA structure was announced to achieve similar effect in 1999. AU Optronics revealed their AMVA with fast pixel response times yet small brightness changes. Their second generation of AMVA (AMVA II) was presented with the improved contrast ratio to 2500:1 in 2006. In 2009, the third generation Advanced MVA III (AMVA III) was developed to fulfill the high quality image and fast response time. The AMVA III adopted the new cell process based on polymer sustained alignment (PSA) process and new pixel electrode pattern design without protrusion and ITO main slit. Moreover, 8-domain pixel electrode pattern with new ART (additional refreshing transistor) architecture was implemented to gain the precise control in the lower levels of gray scale. As a result, a 46-in. AMVA III LCD can obtain contrast ratio greater than 8000:1 with 180° viewing angle (Chen *et al.*, 2009). It is notable that Samsung Electronics Co. Ltd. received a Gold Award for their 40-in. High Contrast, Wide Color Gamut, LED-Backlit LCD TV at the 45th annual Society for Display (SID) 2007 exhibition. This prototype LED TV featured a dynamic contrast ratio of 10,000:1 and claimed its viewing angle at 180° in all direction. MVA has superior viewing angle, good dark state, color reproduction, and fast response time. Comparing to IPS mode, the optical compensation film of MVA mode is much easier to reach its high contrast ratio and wide viewing angles. As a result, MVA mode dominates the TV market lately.

Another multi-domain VA LCD approach is called Axially Symmetric Vertical Alignment (ASVA) mode with circularly symmetric at 360° viewing angle. It was first proposed as the Axis Symmetric Micro-Cell (ASM) method by Yamada *et al.* (1995). Axis symmetry micro-cell (ASM) is fabricated by phase separation of liquid crystal and polymer wall, in which the LC molecules are surrounded by polymer walls under UV irradiation. The chiral compound is added to ensure nematic liquid crystal with a 90° twisted orientation. The ASM method has the advantage not only wide viewing angle due to the symmetry orientation axis, but also disclination-free due to the mono-domain orientation. They continue to develop a negative dielectric liquid crystal in the ASM (N-type ASM) mode in 1998. The N-type ASM mode combined with compensation film is able to reach 300:1 contrast ratio and over $\pm 80^\circ$ viewing angle. The difference between P-type and N-type is that axially symmetrical direction and tilted angle are set prior to carrying out the final curing of UV irradiation. The only drawback is the response time falling around 30 ms. They proposed the ASVA mode with multiple protrusion structures to define the plurality of liquid crystal regions in 2001. Liquid crystal molecules in the plurality of liquid crystal regions are aligned axially and symmetrically around an axis perpendicular to the surface. Each region is approximately a square of 70 μm .

In addition to IPS and MVA modes, researchers have continually improved the image quality by all means. Optically compensated bend (OCB) mode is known to possess the advantages of wide viewing angles and fast response time over nematic liquid crystal displays. It was modified from the original pi-cell (proposed by Bos and Koehler/Beran, 1984) to a much thinner cell gap with optical compensation films by Yamaguchi *et al.* (1993). Bos *et al.* proposed using a negative birefringence film with the optic axis in the normal direction (negative C-plate) to compensate the pi cell in 1993. Yamaguchi *et al.* (1993) suggested the use of a biaxial film for optical compensation of the OCB cell. The viewing angle performance, however, was not satisfied by these compensation methods. Mori offered the OCB-WV film consists of a hybrid polymerized discotic material layer alignment (as in TN WV film) with biaxial film to suppress the light leakage in the on-state of bend mode (Mori, 2005). Chen *et al.* proposed using reactive monomer to suppress the splay-to-bend transition (i.e., bias-free OCB mode) in 2009. The thin splay orientation of LC polymer layer not only provides the surface pre-tilted angle to suppress the splay-to-bend transition, but simultaneously becomes an inner-retarder to minimize the light leakage. The contrast ratio is improved up to a factor of 11 without using compensation films (Chen *et al.*, 2009).

The Latest Approaches of Wide View Angle LCDs

The latest approaches are focused in the optical structure design to enhance LCD's viewing angle after 2010. The design implements a diffuse optics on top of LCD. Kälantär reported the concept of shaping the luminance cone and making spatially uniform

luminance by a light scattering film in SID Symposium 2011. A directional backlight with a narrow angular luminance distribution was applied in the backlight unit. The result suggested that the normal and oblique contrasts were enhanced by a factor of 1.1 and 6 on the VA panel; and 2.5 and 2.7 on the TN panel, respectively. The imaging blur and low intensity were the main drawback of using conventional diffusing film. Yamamoto *et al.* (2014) recommended microstructure film to control passing light on top of LCD in 2014. The microstructure film consists of plate-like transparent polymer and the black dots were randomly arranged on the base film. The cone shape air cavities were placed beneath the black dots to control the passing light. Their result suggested that the microstructure film has over five times higher contrast ratio than the conventional diffusing film. In a similar way, Yang and Chen proposed photo-luminescent LCD with R, G, B-quantum dots (QDs) on top of the conventional TN LC cell to improve the LCD viewing angle in 2016. In particular, the single wavelength excitation at 410 nm gives the great advantage for LCDs without considering the retardation dispersion. Thus, the higher contrast ratio at normally black mode achieved $\pm 70^\circ$ viewing angle (CR>10:1) and 3 ms under 4 V driving voltage in the QD array over TN-LC (Yang *et al.*, 2016).

Controlled Viewing Angle of LCDs

The criticism of the viewing angle has been raged over the years when TN LCD first appeared in the market. The issues of viewing angle are mostly attributed to the unique birefringent mode controlling the light intensity between crossed-polarizers. Wide viewing angle (WVA) LCDs dominate the display market. However, the narrow viewing angle (NVA) or dual-view LCD is getting its popularity for security, personal privacy, or safety reasons. The birefringent mode is more prominent for designing the directional viewing control system. A number of approaches have been realized in the controllable viewing angle LCDs. The double-cell architecture is the first proposed method, in which one additional liquid crystal panel is placed on top of the main panel to control the viewing angle (Jeong *et al.*, 2007). It is the simplest approach, but suffers from the panel thickness and greater costs. The second design is subpixel divided method, in which each pixel is divided into two domains: the primary pixel domain is for displaying images whereas the sub-pixel domain is to control the viewing angle (Lim *et al.*, 2010). The third method is to adopt dual light sources in conjunction with two light-guided films in the backlight system without changing the design of LC panel. The resulting controllable LCD is able to obtain the viewing angular ranges from 140° to 60° (Chien *et al.*, 2006). The fourth is based on a HAN (hybrid aligned nematic) cell, driven by fringe- and vertical-electric fields (Jeong *et al.*, 2008). The viewing angle in the horizontal direction can be controlled from 120° to about 20° . The last method, but not the least, is utilizing a patterned vertical alignment LCD (PVA-LCD) with the circular polarizers. The WVA mode is turned on when no bias voltage on the common electrode. A bias voltage (2.42 V) is applied on common electrode to turn on the NVA mode. The contrast ratio of WVA mode is more than 100:1 in all viewing angles. The viewing angle can be reduced to 20° when the NVA mode is turned on (Yu *et al.*, 2014).

Dual-view liquid crystal display is able to simultaneously deliver different information in the right and left viewing directions, i.e., two distinct viewing cones for right and left directions. After SHARP launched "Dual View LCD" technology in 2005, the dual view display technology was soon adopted by the automobile makers, in their concept cars. There are several approaches have been realized in the dual-view LCD. One of the approaches is to create spatial-multiplexing type, in which the main pixel of the DV LCD contains both a right and left sub-pixel. Parallax barriers and lenticular lenses are commonly applied to direct light from different pixels to achieve the intended field of view. Controlling the viewing angles of the right sub-pixel (RSP) and the left sub-pixel (LSP) can be fulfilled by two different tilted surfaces in one pixel, or by patterned electrodes, in which it provides the inclined electric field to control the LC direction. The other approach is based on the time-multiplexed type, which uses the directional backlight module with complex reflectors and fast response LC cell, such as OCB, Blue Phase modes. The negative aspect is that the aperture ratio is decreased as a result of the dead zones formed nearby the electrodes. In 2017, Chen *et al.* proposed a two-domain twisted nematic liquid crystal with a patterned E-type polarizer which is capable of suppressing the crosstalk between right and left views. The crosstalk is successfully lowered to less than 0.07% for all gray levels, while keeping the viewing angle around 50° for both left and right views (Chen *et al.*, 2017).

Summary

Viewing angle is an intrinsic matter of LCD associated with its birefringent mode optics. It has been intensively studied to achieve wide viewing angle for monitor and TV applications. The methods such as optical compensation films, pixel designs, and driving modes are available to resolve the negative viewing angle issues. Since Samsung succeeded in delivering the widest viewing angle in 2007, the viewing angle is no longer a critical issue in LCDs. It is possible to extend the viewing angle in all directions based on fulfilling the minimum contrast ratio 10:1 requirement. On the contrary, the small viewing angle of LCDs becomes a benefit when one needs to control and manipulate as needed, especially, for privacy purposes. Thus, the latest developments of alternating maximum and minimum viewing angles are valuable traits in LCDs.

References

- Bos, P.J., Koehler/Beran, K.R., 1984. The pi-cell: A fast liquid-crystal optical-switching device. *Molecular Crystals and Liquid Crystals* 113, 329–339.
- Chen, C.P., Wu, Y., Zhou, L., *et al.*, 2017. Crosstalk-free dual-view liquid crystal display using patterned E-type polarizer. *Applied Optics* 56, 380–384.
- Chen, H., Gao, Y., Wu, S.-T., 2015. n-FFS vs. p-FFS: Who wins? *SID Symposium Digest* 46, 735–738.
- Chen, S.-F.F., Chen, H.-M.P., Chow, L., Chang, Y.-Y., Shieh, H.-P.D., 2009. Splay-to-bend transition-free reactive monomer modified Pi-cell. *IEEE Photonics Technology Letters* 21, 636–638.
- Chen, T.-S., Huang, C.-W., Hsieh, C.-C., *et al.*, 2009. Advanced MVA III technology for high-quality LCD TVs. *SID Symposium Digest* 41, 776–779.
- Chien, K.-W., Hsu, Y.-J., Chen, H.-M., 2006. Dual light source for backlight systems for smart viewing-adjustable LCDs. *SID Symposium Digest* 37, 1425–1427.
- Ishinabe, T., Sato, A., Fujikake, H., 2014. Wide-viewing-angle flexible liquid crystal displays with optical compensation of polycarbonate substrates. *Applied Physics Express* 7, 111701.
- Ito, Y., Watanabe, J., Saitoh, Y., *et al.*, 2013. Innovation of optical films using polymerized discotic materials: Past, present and future. *SID Symposium Digest* 44, 526–529.
- Jeong, E., Lim, Y.J., Chin, M.H., *et al.*, 2008. Viewing-angle controllable liquid crystal display using a fringe- and vertical-field driven hybrid aligned nematic liquid crystal. *Applied Physics Letters* 92, 261102.
- Jeong, E., Lim, Y.J., Rhee, J.M., Lee, S.H., 2007. Viewing angle switching of vertical alignment liquid crystal displays by controlling birefringence of homogeneously aligned liquid crystal layer. *Applied Physics Letters* 90, 051116.
- Jeong, J., Park, D., Lee, J.-H., 2016. Simultaneous retardation compensation during bending of plastic film coated with a polymer layer of opposite birefringence. *Applied Optics* 55, 8738–8742.
- Lim, Y.J., Kim, J.H., Her, J.H., *et al.*, 2010. Viewing angle controllable liquid crystal display with high transmittance. *Optics Express* 18, 6824–6830.
- Liu, Y., Lan, Y.-f., Hong, Q., Wu, S.-T., 2014. Compensation film designs for high contrast wide-view blue phase liquid crystal displays. *Journal of Display Technology* 10, 3–4.
- Mori, H., 2005. The wide view (WV) film for enhancing the field of view of LCDs. *Journal of Display Technology* 1, 179–186.
- Mori, H., Itoh, Y., Nishiura, Y., Nakamura, T., Shinagawa, Y., 1997. Performance of a novel optical compensation film based on negative birefringence of discotic compound for wide-viewing-angle twisted-nematic liquid-crystal displays. *Japanese Journal of Applied Physics* 36, 143–147.
- Oka, S., Sasaki, T., Tamaru, T., *et al.*, 2016. Optical compensation method for wide viewing angle IPS LCD using a plastic substrate. *SID Symposium Digest* 47, 87–90.
- Yamada, N., Kohzaki, S., Funada, F., Awane, K., 1995. Axially symmetric aligned microcell (ASM) mode: Electro-optic characteristics of new display mode with excellent wide viewing angle. *SID Symposium Digest* 26, 575–578.
- Yamaguchi, Y., Miyashita, T., Uchida, T., 1993. Wide-viewing-angle display mode for the active-matrix LCD using bend-alignment liquid-crystal cell. *SID Symposium Digest* 24, 277–280.
- Yamamoto, E., Yui, H., Katsuta, S., *et al.*, 2014. Wide viewing LCDs using novel microstructure film. *SID Symposium Digest* 45, 385–388.
- Yang, J.-P., Hsiang, E.-L., Chen, H.-M.P., 2016. Wide viewing angle TN LCD enhanced by printed quantum-dots film. *SID Symposium Digest* 47, 21–24.
- Yu, Y., Dou, H., Ma, H., Sun, Y., 2014. Continuous viewing angle controllable patterned vertical alignment liquid crystal display. *Liquid Crystals* 41, 1595–1599.
- Zhu, X., Ge, Z., Wu, S.T., 2006. Analytical solutions for niaxial-film-compensated wide-view liquid crystal displays. *Journal of Display Technology* 2, 2–20.

Organic Light-Emitting Diode (OLED)

Chin H (Fred) Chen and Wen-Shi Wen, Shine Materials Technology Co., Ltd., Kaohsiung, Taiwan, R.O.C.
Chin-Ti Chen, Chemistry Institute, Academia Sinica, Taipei, Taiwan, R.O.C.

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic light-emitting diode (OLED) with its superb attributes in thinness, light-weight, brightness, color, dynamic contrast, viewing angle, response time, power efficiency and form factor has been declared as the “ultimate display” of our time. It is an emissive device with a very simple structure which comprises a stack of extremely thin layers ($<0.2\ \mu\text{m}$) of organic materials sandwiched between two electrodes. The anode which is usually made of a transparent semi-conductor of high work function (e.g., ITO) is capable of injecting holes into the organics while the cathode often selected from a highly reflective metal of low work function (e.g., Li, Ca, Mg, LiF/Al) is capable of injecting electrons. If the corresponding adjacent organic layers are chosen properly with a set of matching highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) and applied a positive bias voltage (DC driven), the injected electron and hole are transported to recombine near the heterojunction of the corresponding electron/hole transporting layers (ETL/HTL) to produce exciton, from which light emission will occur and emit through the ITO/glass. The color of the emitted photon will roughly correspond to the bandgap energy of the organic luminescent material with an internal quantum efficiency (η_{INT}) governed by the equation of ($\eta_{\text{INT}} = \gamma \cdot \eta_r \cdot \psi_{\text{PI}}$), where γ is the probability of carrier recombination (e/h ratio), η_r is the efficiency of exciton production and ψ_{PI} is the photoluminescence quantum efficiency. Electro-generated excitons statistically have a ratio of singlet/triplet ratio about 1/3. Abided by *Pauli* exclusion principle, only singlet is spin-allowed to efficiently transition radiatively to its ground state for most fluorescent materials while the spin-forbidden decay of triplet (phosphorescence) is usually too slow nor efficient to be useful. It follows that the theoretical limit of η_{INT} is limited to only about 25% without the capability of harvesting the triplet exciton. Moreover, as a Lambertian source, not all the light generated within OLED can escape out of the planar cell through various layers of media with different refractive index (n), for there is a further loss which can be estimated by the Fresnel equation of $\alpha = 1/(2n^2)$. Therefore the external quantum efficiency (η_{EXT}) can be further calculated by the equation of ($\eta_{\text{EXT}} = \alpha \cdot \eta_{\text{INT}}$) where α is designated as the out-coupling efficiency which without means of any light enhancement is about 20% for most organic materials with an average $n \approx 1.5 - 1.9$. Lastly, the power efficiency which is inversely proportional to the driving voltage must also be considered in order to meet the ever increasing demand of high resolution mobile displays. To drive down power consumption by minimizing carrier injection barriers and increasing the probability of recombination near the interface, a typical bottom-emission OLED architecture of today is shown in [Fig. 1](#) (top) with the addition of several key functional organic layers such as electron injection layer (EIL), hole injection layer (HIL) and a R,G,B doped emitter to control color of emission.

Over the last 15 years, tremendous progress has been made in OLED materials overcoming the limitation of η_{INT} of fluorescence which has been enhanced to nearly 100% by exploiting phosphorescence as well as delayed fluorescence. Power efficiency can also be boosted by the discovery of bipolar host material and the introduction of extremely electron deficient (*p*-) and donating (*n*-) dopants into the respective HTL and ETL to increase its conductivity as illustrated in [Fig. 1](#) (bottom) with a typical *p-i-n* OLED energy diagram with its LUMO and HOMO bending after applying a positive bias voltage. Tandem OLED to increase luminance without increasing drive current density nor jeopardizing OLED lifetime and light enhancement technology have also been developed to extract more light out of OLED with the understanding of its optics. As a result of these discoveries, the most eco-friendly, sun-like area luminaire of indoor OLED lighting has now become possible.

Materials

Early OLED materials development has all been focused on the design and synthesis of fluorescent molecules which can only harvest singlet excitons. In the first decade of 21st century, unusually high EL efficiency and lifetime have been reported for a number of blue fluorescence OLEDs, many of which are anthracene-derived fluorophores, such as non-dopant deep blue emitter TAT with $\lambda_{\text{max}}^{\text{EL}}$ 444 nm, CIE_{x,y} (0.16, 0.09), (η_{EXT} 7.18%) and doped sky-blue emitter of arylamine-substituted distyrylarylene DSA-Ph with $\lambda_{\text{max}}^{\text{EL}}$ 465 nm, CIE_{x,y} (0.15, 0.29) in MADN host (η_{EXT} 8.7%). In depth study of TAT has revealed that both triplet-triplet annihilation (TTA) and the outcoupling enhancement by dipole orientation of emitting molecules contribute to the production of singlet exciton from which exceedingly high η_{EXT} is thus achievable.

Major breakthrough of EL efficiency of OLED started to emerge before the end of last century. One of the early examples was a green phosphorescence material known as $\text{Ir}(\text{ppy})_3$ which is an organometallic-based coordination complex containing heavy metal Iridium. Utilizing the heavy atom effect that promotes spin-orbital coupling, these 2nd generation luminescent materials all possess a hybrid emissive excited state of singlet and triplet excitons with a rather short excited-state lifetime (in μs) thus easing up the forbidden phosphorescence at room temperature. Based on this emissive mechanism, numerous heavy metal containing

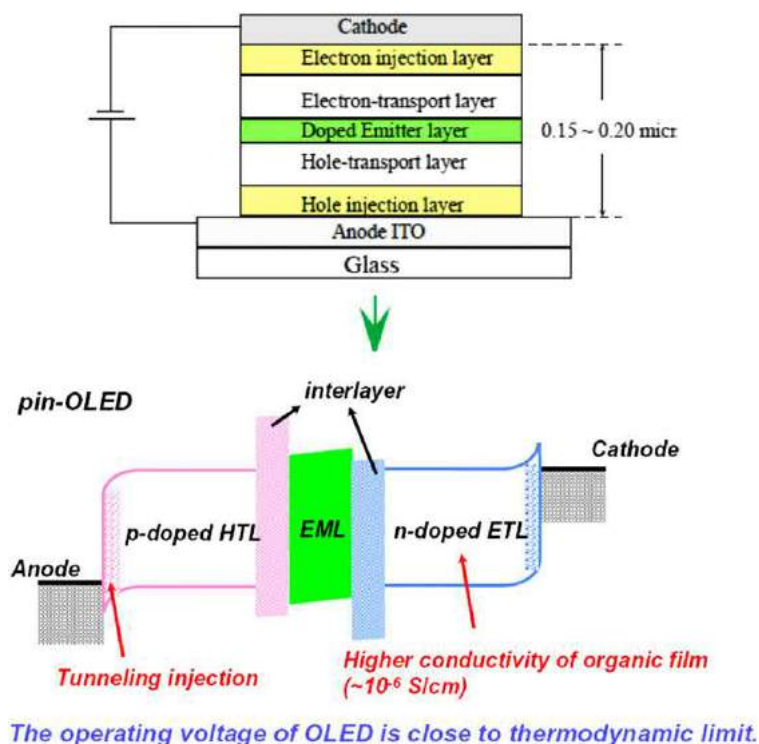


Fig. 1 A typical bottom-emission OLED architecture (top) and energy profile of *p-i-n* OLED upon positive bias (bottom).

(mostly Ir) red, green or even some light blue phosphorescence EL materials have been reported. Meanwhile, hosts derived from carbazole with high triplet energy as well as many triplet exciton blocking materials to prevent luminescence quenching have also been developed. In dopant/host emitter device with appropriate blocking layer, EL power efficiency of >100 lm/W for green $\text{Ir}(\text{ppy})_3$, >25 lm/W for red $\text{Ir}(\text{mphpq})_2(\text{acac})$ equivalent to $\eta_{\text{EXT}} \sim 25\%$ and near 50 lm/W for sky-blue FIrpic was reported at brightness of 1000 cd/m^2 . Mainly due to the high EL efficiency (hence low power consumption), today's AMOLED display products in the marketplace all have adopted phosphorescence EL materials for their green and red emitters. The blue phosphorescent material owing to its intrinsic instability and lack of color saturation issues remains to be the *Holy Grail* that has been pursued relentlessly. As a result, display industry presently favors the use of the deep blue anthracene-based fluorescent materials especially the deep blue with $\text{CIE}_y (<0.10)$ as dopant for its AMOLED products.

The 3rd generation of luminescence was discovered in 2011 which is based on the up-conversion of triplet exciton into its singlet excited state leading to efficient thermally activated delayed fluorescence (TADF). This type of materials which have the advantage of containing no heavy metal like Ir or Pt (which is expensive and rare) and the potential of harvesting triplet exciton as well as the possibility of breaking the Ir-materials patent gridlock have quickly caught the fancy of a great many researchers. In just a few years thereafter, green 4CZIPN ($\lambda_{\text{max}}^{\text{PL}}$ 507 nm) and DACT-II ($\lambda_{\text{max}}^{\text{PL}}$ 516 nm), greenish blue ($\lambda_{\text{max}}^{\text{PL}}$ 480 nm) *SpiroAC-TRZ* and a near blue ($\lambda_{\text{max}}^{\text{PL}}$ 460 nm) *DMAC-DPS* with $\text{CIE}_{x,y}$ (0.16, 0.20) have been reported to achieve η_{EXT} as high as 19.3%, 29.6%, 36.7% and 19.5%, respectively. All these TADF materials have a common structural feature containing a twisted non-coplanar configuration formed by steric crowding between an electron donor (usually arylamine) and an acceptor (e.g., CN, SO_2 , $\text{C}=\text{O}$, $\text{P}=\text{O}$) through certain π -conjugation (such as benzene ring) which has been shown to possess unusually narrow singlet-triplet state bandgap (<0.1 eV) conducive for room temperature intersystem crossing.

In 2015, a series of TADF molecules have been adopted as an assisted dopant (such as ACRSA) in OLEDs to transfer its singlet excitons to the singlet state of a more fluorescent dopant emitter (such as *TBPe*) via Förster energy transfer process to produce what is called "hyperfluorescence". To differentiate from TADF emitter, a 3.5th generation of OLED materials has been declared after its discovery which opens up a golden opportunity for further research in the pursuit of highly efficient yet still stable emitters particularly for the blue color. This is because the twisted molecular configuration associated with TADF molecules usually have a rather large FWHM which is adverse to the color purity required for display applications. Utilization of conventional fluorescence molecule as the end emitting dopant can alleviate the problem so as to regain its color purity yet still maintain high EL efficiency as well as long device lifetime. Although TADF materials are non-heavy-metal containing, recent study showed that most TADF OLEDs still suffered from fast degradation and "roll-off" of efficiency at high brightness particularly acute for the blue. This is mainly attributed to the synergetic oxidation of the TADF emitters in the devices, including at least an electrochemical oxidation and a photo-oxidation processes. Evolution of OLED materials development and a few examples of key molecular structures are illustrated in Fig. 2.

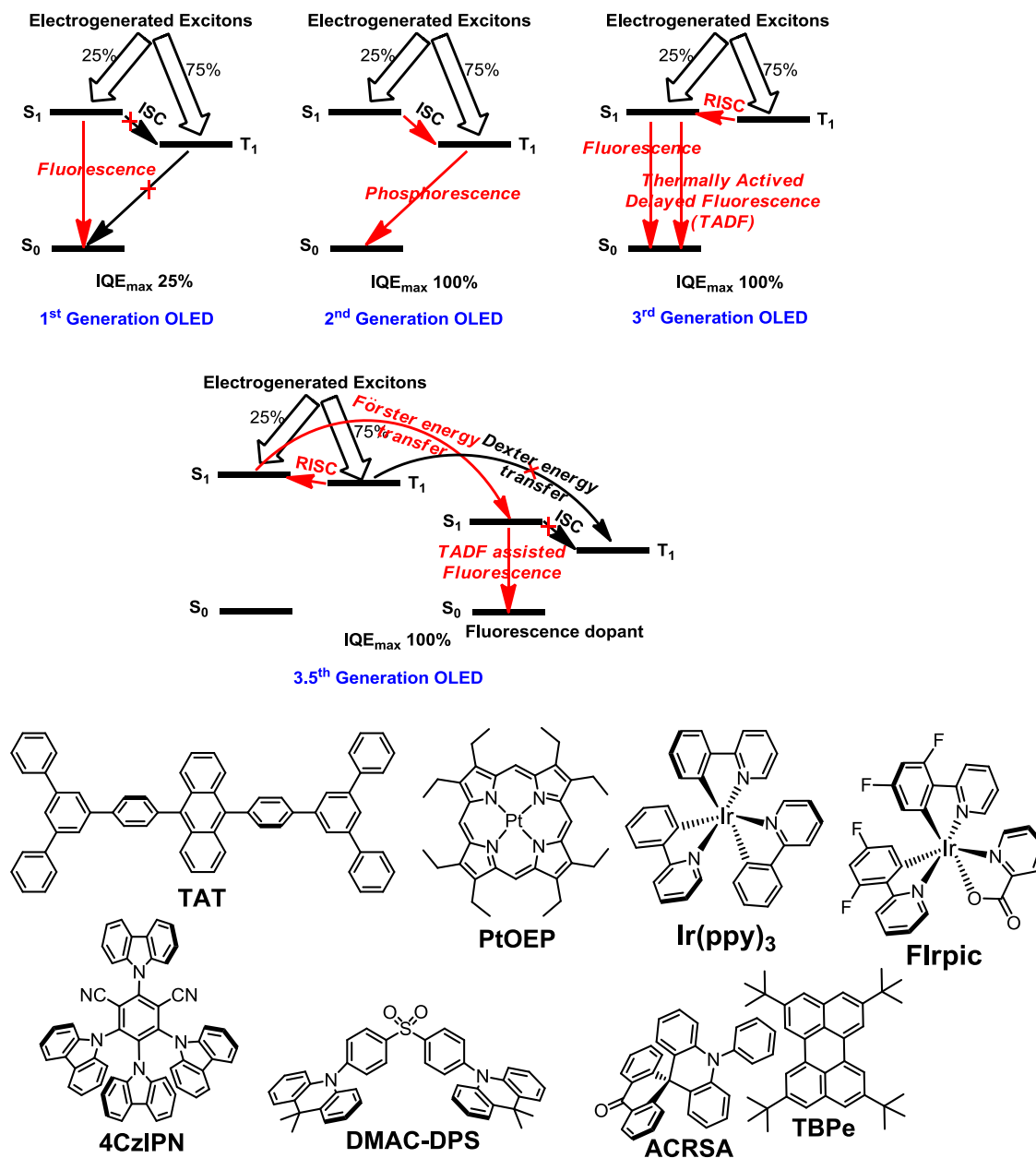


Fig. 2 Evolution of OLED materials development and examples of key molecular structures.

Devices

All good OLED materials invariably need good device architecture to maximize their electroluminescence (EL). This is particularly true in regard to overall performance involving key measurement parameters like stability, power consumption, luminance efficiency and color. OLED device structures can be roughly classified into bottom-emission and top-emission (TOLED) depending upon whether the bottom anode (ITO) or the top cathode is made transparent (with its anode reflective), respectively. If both electrodes are transparent, a *transparent* OLED can also be made. For certain *n*-type TFT driving needs, there have also been developed the corresponding bottom-emission and top-emission *inverted* OLED where the anode and cathode are reversed instead.

For driver electronics, there are passive-matrix (PM-OLED) used mostly for low-information content, small size alphanumeric displays (because of its line-by-line scan requiring very high peak-current density and limitation for high resolution) and the active-matrix TFT driving (AM-OLED) which is the most popular. Low temperature poly-silicon (LTPS) TFT with its high *p*-channel mobility ($\sim 100 \text{ cm}^2/\text{V s}$) is best suited for small to medium displays like the smartphones and tablets while Oxide (IGZO) TFT is particularly suitable for large size displays like OLED TV. To increase the aperture ratio for high-resolution, it is most advantageous to couple LTPS-TFT backplane with a TOLED structure in which its cathode is made semi-transparent (e.g., Ag:Mg 15 nm) and its

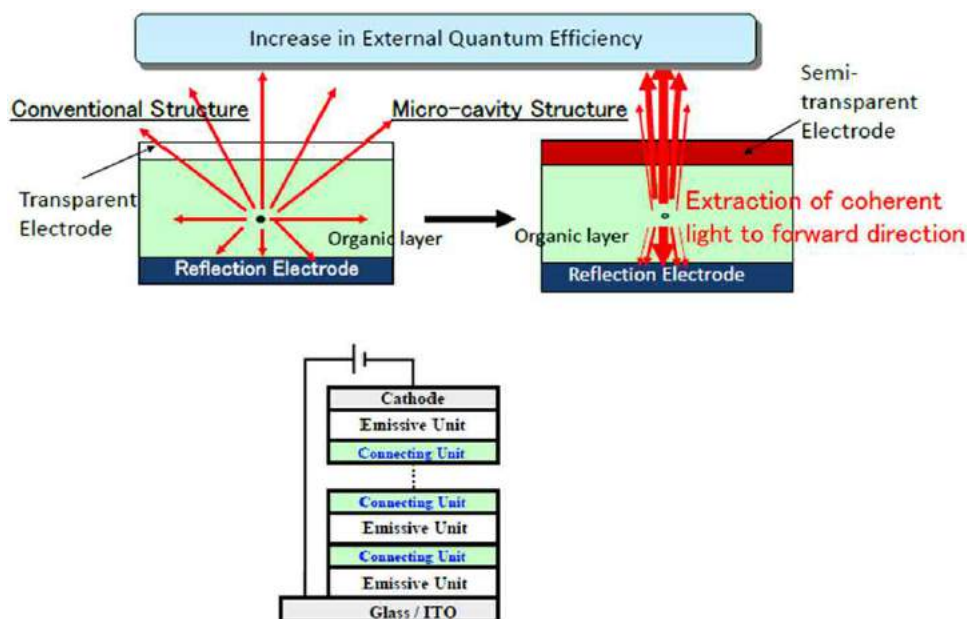


Fig. 3 Microcavity effect of TOLED (top) and a typical tandem OLED structure (bottom).

anode reflective (Ag) as exemplified in Fig. 3 (top). There's also an added advantage in its capability of creating a microcavity within the organic media between the two metallic electrodes one of which is semi-transparent and the other reflective. Its optical effect in terms of constructive (*antinode*) and destructive (*node*) interferences which can be modeled as a Fabry-Perot cavity embedded light source within 2 mirrors, is the dramatic spectral narrowing and intensity enhancement commonly observed in a TOLED. By clever adjustment of the optical length L of the μ -cavity (which is usually done by adjusting the HTL to minimize impact on drive voltage) for each of the color pixel according to the first order approximation of $[2L/\lambda_{\max} + \phi/2\pi = m]$ (m =integer) and [full width at half maximum ($FWHM$) = $\lambda^2/2L \times 1 - (R_1R_2)^{1/2}/\pi(R_1R_2)^{1/4}$] where L is the optical length and R_1R_2 are the reflectance of the 2 mirrors, greatly enhanced color gamut by reducing FWHM of each of the R, G, B emission wavelength (λ) and EQE (η_{EXT}) of AM-OLED can be readily achieved.

Another key development in OLED device is the discovery of tandem OLED where multiple OLEDs can be coupled together by transparent connecting charge-generation layer (CGL) which functions as a reversed p/n degenerate contact to create and inject electron and hole back to the ETL and HTL, respectfully for recombination as depicted in Fig. 3 (bottom). As a consequence, luminance of tandem OLED (L_n) can be enhanced at the same driving current density according to the equation of $[L_n = n \times L]$ where n is the number of stacked units and L is the luminance of each single unit. This in turn means that the lifetime of tandem OLED can also be greatly enhanced because the intrinsic stability of organic luminescent materials is singularly dependent upon the driving condition and inversely proportional to the total amount of electric charge passed through the device. Thus, brightness of 3000 cd/m² required for most general indoor lighting is readily attainable with relatively small driving current density.

Recently, due to the high interest of developing OLED as a green and healthy *area* indoor solid state lighting alternative to LED, a great deal of effort has been aimed to extract more light trapped inside OLED. Detailed analysis by rigorous electromagnetic modeling of optics have revealed 4 types of light entrapment, namely radiation mode, substrate mode, waveguide mode and Plasmon mode as depicted in Fig. 4 (top). It also shows that coupling into different modes is highly dependent on the position of emission center which oscillates as illustrated in Fig. 4 (bottom). The second *antinode* appears to give the highest outcoupling efficiency while the surface Plasmon mode can be minimized by fabricating OLED emitter away from its reflective metallic electrode. To date, the most noteworthy light extraction technology is the so-called internal enhancement layer (IES) which is a high-index coupling layer embedded with light scattering particles incorporated in-between the ITO and the glass substrate where η_{EXT} as high as 45% has been achieved in white OLED.

Applications

As display technology, OLED has considerable attributes that are unsurpassed by others. There have been a great many super AMOLED products (e.g., smartphones, tablets, monitors, VR/AR microdisplays, wearables, TVs, etc.) that have successfully entered the consumer electronics market and indeed, many of which had won Best Display Awards (including most recently the 2016 "Best of Innovation" for a 77-in 4K OLED TV). It has also been declared "the best panel ever" as tested by the independent Display Mate Labs in major measurements categories like winning the "highest color accuracy, infinite contrast ratio, lowest screen reflectance, and smallest brightness variation with viewing angle."

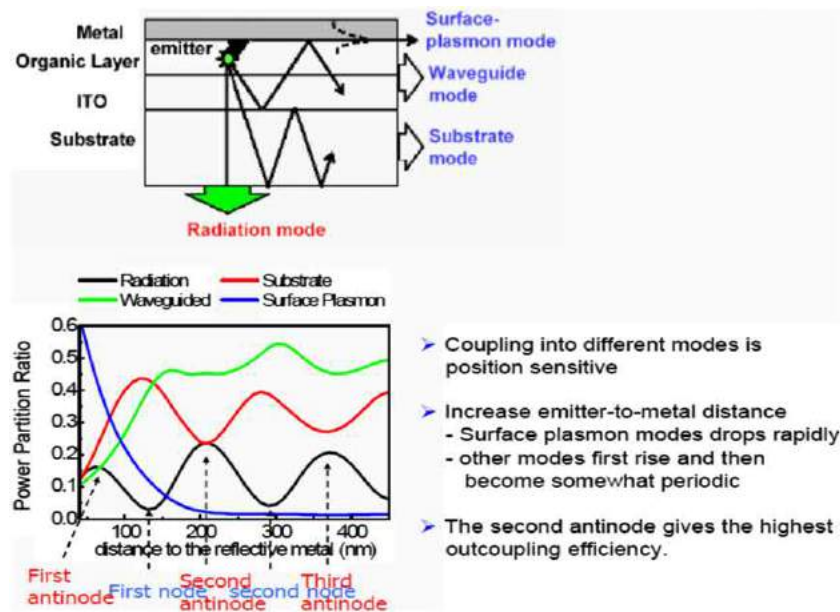


Fig. 4 Electromagnetic modeling of light generated inside OLED (top) and their couplings with respect to the position of emission (bottom).



Fig. 5 Examples of OLED displays and lighting.

As predicted by DisplaySearch and IHS, OLED display revenues will grow to about \$44B in 2020 in which LTPS-TFT AMOLED smartphones will be the main driver overtaking TFT-LCD to become the leading display technology in mobile sector while OLED TV business will begin to grow significantly after 2016. In addition, owing to the tremendous improvement of power efficiency ($> 100 \text{ lm/W}$) as well as lifetime recently, WOLED lighting which has the potential of emitting warm light across large surfaces

and bringing to bear novel *form factor* into the lighting sector is also forecasted to gain market share slowly and eventually to become competitive with the existing LED solid state lighting particularly in indoor illumination. A glossary of several OLED display and lighting products is reproduced in Fig. 5.

Perspectives

After more than 30 years of R&D, OLED technology has finally begun to hit the big time in the display market. It is expected that flexible OLED (FOLED) will be a “killer application” for which OLED industry has long been searching. Indeed, we have witnessed the successful launching of many unbreakable and bendable FOLEDs which are fabricated on thin plastics and continue looking forward to the ultimate realization of foldable and rollable OLEDs in years to come. But in order to become a viable and sustainable business to compete effectively with LCD as well as a host of other new emerging displays like those of Quantum Dot, microLED and laser, many challenges must be overcome. These include issues of deep blue emitter efficiency and lifetime, large area fabrication, mass production, cost down, thin-film encapsulation, robust RGB pixelation (to achieve high yield and resolution), improvement of luminance and power efficiency, etc. Control of optics to maximize out-coupling efficiency of OLED (without impacting on image quality for display application) will remain to be an interesting topic of research and further study in the future.

It suffices to state that the future of OLED as display and lighting is extremely bright and colorful, which presents great opportunities for numerous invention and innovation in R&D as the technology matures and continues to improve over time. We will witness significant growth of OLED investment and business as well as proliferation of AMOLED, FOLED, wearable and VR/AR products like smart phones, PMP, tablets, ultrabooks, monitors and TVs in the marketplace in the foreseeable future. By then, it is predicted that many area OLED lighting luminaire products will also be with us as one of the major green, energy-saving, most healthy (no glare nor blue-light hazard) solid state lighting fixtures in our home. Perhaps, the last question to be pondered is: Will Roll-to-Roll printing of OLED not be too far?

It is impossible to describe OLED science and technology in any breath or depth in these few pages. Interested readers are encouraged to read further with the list of books and references cited in the following.

Acknowledgement

We gratefully thank M.-T. Lee of AU Optronics (Taiwan), Y.-S. Tyan of First-O-Lite (China) and C. C. Wu of National Taiwan University for permission of using excerpts of some of their presentation and lecture notes for graphic illustration in this review.

See also: OLED/Introduction

Further Reading

- Baranoff, E., Yum, J.-H., Graetzel, M., Nazeeruddin, Md. K., 2009. *J. Organometallic Chem.* 694 (17), 2661–2670.
- Han, T.-H., Jeong, S.-H., Lee, Y., *et al.*, 2015. *J. Information Display* 16 (2), 71–84.
- Kaji, H., Suzuki, H., Fukushima, T., *et al.*, 2015. *Nat. Commun.* 6, 8476.
- Kim, S.T., 2006. Keynote in IMID 2006, August 23, 2006, Daegu, Korea.
- Ko, S.H. (Ed.), 2011. *Organic light emitting diode- material, process and devices* (Book, 2011).
- Lee, J.-H., Kim, J.-J., Lee, S., *et al.*, 2013. *J. Information Display* 14 (1), 39–48.
- Liu, C.-C., Liu, S.-H., Tien, K.-C., *et al.*, 2009. *Appl. Phys. Lett.* 94, 103302.
- Liu, J., Chen, C.-T., Chen, C.H., 2015. Introduction to organic light-emitting diodes (OLED). In: Kriss, M. (Ed.), *The Handbook of Digital Imaging*. John Wiley & Sons, Ltd., pp. 577–626.
- Mazzeo, M. (ed), 2010. *Organic Light Emitting Diode* (Book, 2010).
- Nakanotani, H., Adachi, C., 2014. *Nat. Commun.* 5, 4016.
- Raychaudhuri, P.K., Madathil, J.K., Shore, J.D., Van Slyke, S.A., 2004. Performance enhancement of top- and bottom- emitting organic light-emitting devices using microcavity structures. *J. Soc. Inf. Display* 12, 315–321.
- So, F., Adachi, C. (Eds.), 2013. *Organic light emitting materials and devices*. In: XVII SPIE Proceedings, vol. 8829, San Diego, CA.
- Tyan, Y.-S., Rao, Y.-Q., Ren, X.-F., *et al.*, 2009. *Proceedings of SID* 2009, p. 895.
- Uoyama, H., Goushi, K., Shizu, K., Nomura, H., Adachi, C., 2012. Highly efficient light-emitting diodes from delayed fluorescence. *Nature* 492, 234–239.
- Zhang, Q., Li, B., Huang, S., *et al.*, 2014. *Nat. Photonics* 8, 326.

Optical Characteristics of Display Devices

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Definition of a Display

A display can be defined as an object that transfers information to the viewer and/or an output unit that gives a visual representation of data (Lee *et al.*, 2008; Yeh and Gu, 2010; Tsujimura, 2017; Kobayashi *et al.*, 2009).

Optical Characteristics of a Display

Considering that the information provided by a display is delivered via light, the optical characteristics of a display are of great importance. A display has, in general, the optical characteristics of luminance, reflectivity, contrast ratio, color, surface gloss, resolution, and image sharpness (sometimes called small area contrast). Measurement methods can be obtained from the Electronics Industries Alliance (EIA) JT-31 (Optical Characteristics of Display Devices), Society of Automotive Engineers (SAE), International Organization for Standardization (ISO), and Video Electronics Standards Association (VESA) Standards, to name a few (see Relevant Website section).

Radiometric and Photometric Units

To measure light, a certain system of units is needed. Generally, the radiometric or photometric units are adopted for different purposes. The radiometric units refer to a measurement of any portion of the electromagnetic spectrum, such as visible, infrared, and ultraviolet light. In this system, the radiant power is in Watt (W), which is the total radiant energy per unit time (Joule per second, J/s). Different from the radiometric units, the photometric units only deal with visible light and consider the human perception. As the sensitivity of the human vision system varies with wavelengths, the weights of different wavelengths are also different. In the photometric system, the most basic quantity is luminous intensity, which is a measure of the wavelength-weighted power emitted by a light source in a particular direction per unit solid angle. The International System of Units (SI unit) of luminous intensity is candela (cd) (in the English system, cp), which is also an SI base unit. Table 1 lists some common radiometric quantities with their units and corresponding photometric quantities with units.

Luminance

The luminance of a display is the output luminous intensity per unit area in a unit solid angle, which is given in units of candela/square meter or cd/m² (in the English system fL), as shown in Eq. (1) and Fig. 1 (Warren, 2007; Wyszecki and Stiles, 1982). Luminance is one of the most important specifications of a display device, which is closely associated with a display's perceptive brightness (Booker, 1981; Ronchi *et al.*, 1980). Generally, a computer monitor may have a luminance of several hundred nits while an outdoor display may have several thousands (Gong *et al.*, 2012; Reinhard *et al.*, 2010). In addition, the luminance may vary with the region on a screen (spatial luminance profile) and viewing angle (angular luminance profile). To measure luminance of a display, a photometer, photometer/colorimeter, or spectroradiometer can be adopted.

$$L = \frac{d\Phi}{d\Omega \cdot dS \cdot \cos\theta} \quad (1)$$

where L is luminance, Φ is luminous flux, Ω is solid angle, S is area, and θ is viewing angle.

Table 1 Common quantities and units in the radiometric and photometric systems

Radiometric system		Photometric system	
Quantity	Unit	Quantity	Unit
Radiant energy	J	Luminous energy	lm · s
Radiant intensity	W · sr ⁻¹	Luminous intensity	cd (lumen · sr ⁻¹)
Radiant flux	W (J · s ⁻¹)	Luminous flux	lumen (cd · sr)
Radiance	W · sr ⁻¹ · m ⁻²	Luminance	nit (cd · m ⁻²)
Irradiance	W · m ⁻²	Illuminance	lux (lm · m ⁻²)

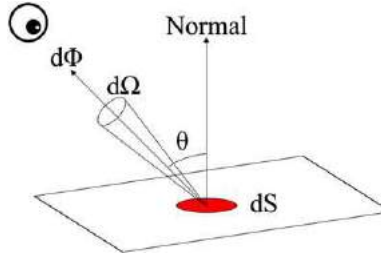


Fig. 1 Calculation of luminance perceived from a direction θ relative to the normal. $d\Phi$, $d\Omega$, and dS are infinitesimal luminous flux, solid angle, and area, respectively.

Illuminance

The illuminance is the total luminous flux incident on a surface, per unit area, as given in Eq. (2) (Warren, 2007; Wyszecki and Stiles, 1982). Its units are lumen per m^2 or lux (and in the English system Ft.cd). This is measured using illuminance meters. The luminance of a display may also vary with the region on the screen.

$$E = \frac{d\Phi}{dS} \quad (2)$$

where E is illuminance, Φ is luminous flux and S is area.

Reflectivity

The reflectivity of a display is the ratio of the reflected light to the incident light. For CRTs, the reflectivity is measured with incident light at 45 degrees, while a smaller angle of 30 degree is used in some cases for LCDs (Keller, 1997). The reflected light from the surface is measured with a detector normal to (perpendicular to) the surface.

Color

A normal human vision system has three kinds of cone cells, which respectively have sensitivity peaks in short (S, 420 nm–440 nm), middle (M, 530 nm–540 nm), and long (L, 560 nm–580 nm) wavelengths. According to the additive color-mixing principle, various color systems have been developed from the working mechanism of the human color vision. With the aid of a color system, although color perception is different from different individuals, the standard responses to a color stimuli can be evaluated and we can measure colors of different objects and display images. Color can be measured with a colorimeter or spectroradiometer (Smith, 1978; Ohno, 2000). The output data are given in various color systems. Color systems, such as the 1931 CIE (Commission International de l'Eclairage) Color System (x, y), the 1960 CIE (u, v) or 1976 CIE (u', v') color coordinates are all used (Wikipedia, 1931, 1960; see Relevant Website section). All these color systems can be converted to each other by simple mathematical relations and these conversions are usually available to the user in the data output. The choice of the color space/system is dependent on application but the system in most general use for displays is the 1976 CIE (u', v'), where the values of u' and v' can be converted from tristimulus values XYZ in 1931 CIE (x, y) by Eq. (3) (Ford and Roberts, 1998). This color system has become more widely used because equally perceived color differences in this color system are at almost equal distances in the color space (McGuire, 1992). In reality, the quest for a system with equal distances equating to equal color differences can be achieved through a complicated matrix approach in which the constants used in the matrix vary from point to point in the color space.

$$\begin{aligned} u' &= \frac{4X}{X + 15Y + 3Z} = \frac{4x}{-2x + 12y + 3} \\ v' &= \frac{9Y}{X + 15Y + 3Z} = \frac{9y}{-2x + 12y + 3} \end{aligned} \quad (3)$$

Gloss

Gloss is a measure of surface reflectivity under Snell's Law criteria (equal incidence and reflection angles). The measurement angles of 20, 45, 60, and 85 degrees are used to measure the gloss of common surfaces (Keller, 1997). Display glass surface measurements tend to use 60 degree gloss angle. The surface of a display can be made rougher to scatter incident light away from the viewer or it can be optically coated to reduce the reflected light to achieve improved viewing. Some times both methods are used (Dearborn Glass Co, 1971). Reduced gloss surfaces are also involved with the reduction in surface glare. The level of increased roughness of the surface affects the user's ability to see details.

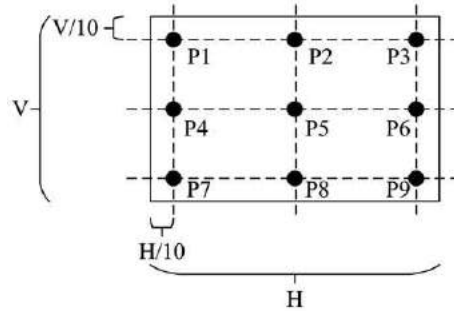


Fig. 2 9-point testing of the uniformity of a display, where the black dots denote the testing points.

Uniformity

The uniformity of a display is how consistent an image appears on the screen in terms of luminance, color, etc. For instance, luminance uniformity is defined as the ratio of the smallest luminance on the screen to the largest, which is usually required to be higher than 0.75 for a qualified display (Keller, 1997). To measure the luminance uniformity (or uniformity in terms of other metrics), a 9-point testing is often implemented, which measures nine testing points uniformly distributed on the screen with the aid of certain test charts, as shown in Fig. 2. Sometimes, a similar 16-point testing is also adopted. Also based on the 9-point testing, the chromatic uniformity in 1976 CIE (u' , v') color space can be measured, as given by Eq. (4).

$$\Delta u'v' = \frac{\sum \sqrt{(u'_i - u'_0)^2 + (v'_i - v'_0)^2}}{N} \quad (4)$$

where (u'_0 , v'_0) is the chromaticity coordinate of a reference, such as the central point of the screen or a designated color.

Contrast Ratio

When people talk about contrast they generally mean “contrast ratio.” The contrast ratio of an area on a display is a unitless parameter that is defined as the ratio of the luminance of the brightest color to that of the darkest color (Lee *et al.*, 2008; Yeh and Gu, 2010). Contrast ratio is a general concept that has not been officially standardized across different types of displays and measurement methods. For LCDs or other modern displays, the most widely adopted conceptions are static contrast ratio and dynamic contrast ratio.

Static contrast ratio is the fundamental and inherent characteristic of a display that is the luminosity ratio of the brightest color to the darkest that the display is able to produce simultaneously at any instant of time. Two methods are usually used to measure the static contrast ratio, as (1) native on/off method, which sequentially adopts a full black signal and then a full white signal, and (2) ANSI method, which adopts a checkerboard pattern containing fully black and white colors simultaneously (Pavlovych and Stuerzlinger, 2005). The native on/off contrast ratio is always a little higher than the ANSI one, so commercial products are always characterized by ANSI contrast ratio. In LCDs, certain backlight intrinsically leaks through LC cells even if a black signal is displayed; i.e., LC cells cannot be completely closed, which restricts the contrast ratio of an LCD. Generally, modern LCDs can achieve a native on/off contrast ratio of approximately a thousand to one. Next-generation displays, such as OLED and micro-LED displays (Pavlovych and Stuerzlinger, 2005; Putman, 2016), directly emit light from the screen without backlight; thus the pixels actually produce “no light” for a black signal, and then, the contrast ratio can be way much higher.

Dynamic contrast ratio is the luminosity ratio of the brightest to the darkest color over time, which is usually measured by dynamic on/off method. What is different from the native on/off method is that a mechanical iris or electronic signal processing is allowed to help increase the contrast ratio. Local dimming or some brightness boost technologies that can be found in modern displays can also be utilized (Huang *et al.*, 2011). Obviously, dynamic contrast ratio measured in such a manner is (much) higher than static contrast ratio. Some LCD manufacturers announce that their products can even reach a dynamic contrast ratio of higher than 10,000,000:1. However, since software or circuit based contrast ratio enhancement technologies and measurement conditions have not been standardized, dynamic contrast ratio is usually not compared across different display devices.

Additionally, the effect of ambience on contrast ratio measurement should also be considered sometimes. If the measurement is implemented in a room with reflective walls, light threw out from the display will be reflected back by the walls and then degrade the contrast ratio.

Finally, the variation of contrast ratio with viewing angle also needs to be paid attention. Modern display technologies, such as emissive OLED displays or in-plane switching (IPS) LCDs, can realize quite homogenous contrast ratio over a wide range of viewing angles, while twisted nematic (TN) LCDs have a poorer performance, like that shown in Fig. 3.

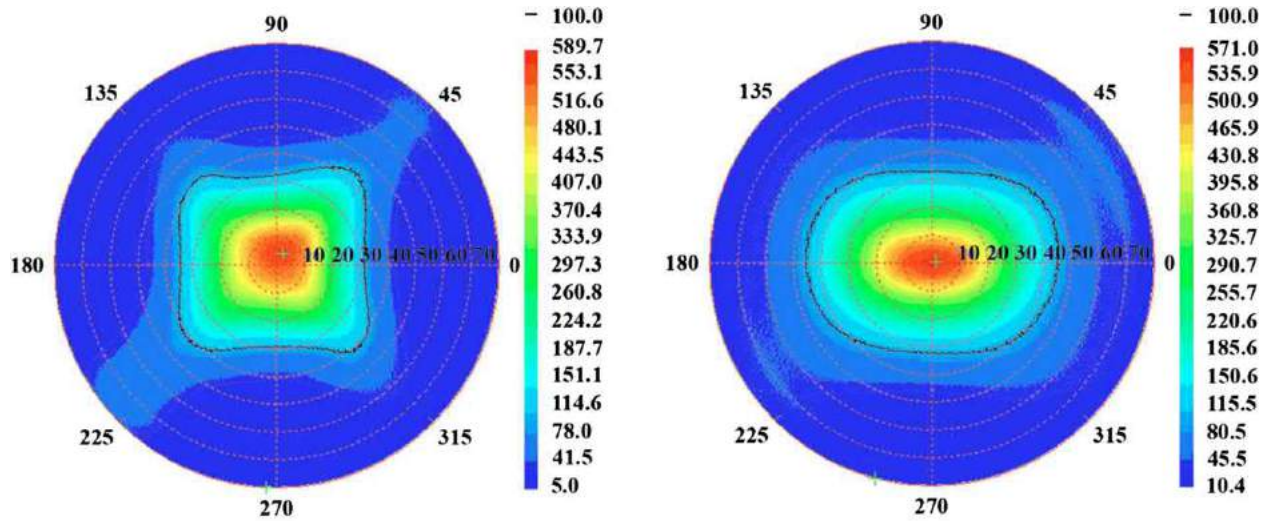


Fig. 3 Iso-contrast contours of angular-dependent contrast ratios of a conventional TN LCD (left) and a viewing-angle-enhanced LCD (Hwang *et al.*, 2009). Reproduced from Hwang, S., Min, J., Lee, M.G., *et al.*, 2009. Novel wide viewing liquid crystal display with improved off-axis image quality in a twisted nematic configuration. *Optical Engineering* 48(11). doi: 10.1117/1.3265524.

Ambient Illumination and Reflection

A display for users to see the information is exposed in an ambient illumination, like bright daylight; hence ambient contrast ratio (ACR) that takes reflected ambient light into account is usually used. ACR of a display with an overall reflectivity R can be calculated with Eq. (5) (Singh *et al.*, 2012; Qin *et al.*, 2015). It can be seen that higher ambient luminance contributing both to the numerator and the denominator leads to a lower ACR.

$$ACR = \frac{L_{on} + R \cdot L_{amb}}{L_{off} + R \cdot L_{amb}} \quad (5)$$

where L_{on} , L_{off} , and L_{amb} is luminance of the brightest color, the darkest color, and the ambience.

With the development of flat panel displays, displays are widely used in portable electronic devices. It is a common scenario that a user watches a display in the bright daylight; thus the "specular reflection" (Snell's Law type of reflection) on the front glass of the display will degrade the contrast ratio severely. This reflection is reduced by methods such as anti-glare (AG) and anti-reflecting (AR) surfaces applied to the front surface of a display. Surface roughness is used in AG surfaces to scatter a portion of the specular light away from the viewer. Surface roughness can be achieved through surface abrading, surface plastic coating, surface silicate coating, and surface acid etching to name a few methods. AR surfaces use optical interference techniques to reduce the reflected light to the viewer. In some cases, both methods are applied to a display surface. If an AG coating is used, the image sharpness will be affected by the reduction in specular reflection through the use of surface roughness; thus ambient light reflection and image sharpness should be traded off to acquire the optimum image quality. On the other hand, AR coatings improve glare reduction without scattering and thus give higher resolution capability to the displays but this is usually obtained at a higher price than the AG coating (Chen, 2001).

Moreover, like OLED displays, the ambient light reflection is caused by the metallic electrodes inside the display, hence AG or AR coating on the front surface does not take effect. The most common approach for such an internal reflection is to use a linear polarizer along with a quarter wavelength retarder, so that the ambient light can be absorbed before it exits the display, as shown in Fig. 4 (Singh *et al.*, 2012; Qin *et al.*, 2015). Nevertheless, the use of a polarizer cuts off at least half of the emissive light, and a considerable thickness of the polarizer plus the retarder is not expected in flexible displays. Therefore, some new techniques are under development, such as "black cathode (Xie and Hung, 2004)," microlens array with black matrices (Qin *et al.*, 2015), etc.

Bidirectional Reflectance Distribution Function

An effective tool is bidirectional reflectance distribution function (BRDF), which can describe reflection properties at different viewing angles versus light with different incident angles (Marschner *et al.*, 2000). Fig. 5 and Eq. (6) explains how BRDF works with four variables. Its value equals to the ratio of the reflected radiance in the outgoing direction to the irradiance in the incoming direction. BRDF can comprehensively depict the reflective property of a surface, so that it can be widely utilized in optical

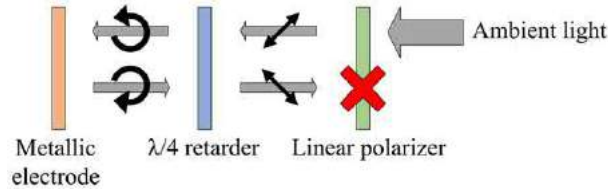


Fig. 4 Elimination of ambient light reflection in a display containing metallic electrodes by using a linear polarizer with a quarter wavelength retarder. The straight and circular arrows demote linearly and circularly polarized light, respectively. The directions of the arrows denote the polarization direction.

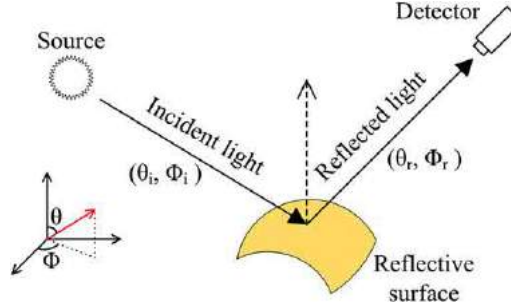


Fig. 5 Illustration of the definition of BRDF. θ and Φ denote the direction. The subscripts “i” and “r” denote incoming and reflected light, respectively.

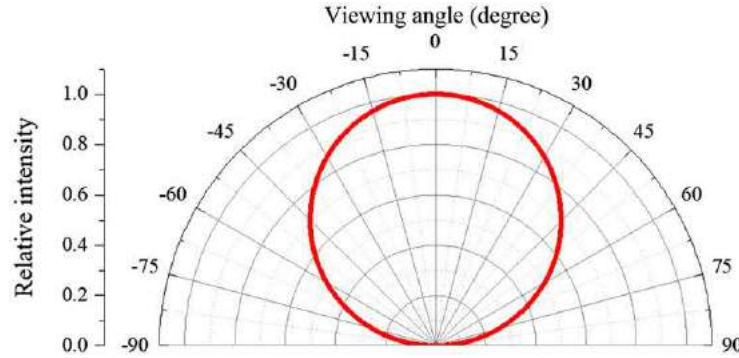


Fig. 6 Angular luminous intensity distribution of a Lambertian surface in a polar coordinate system.

design, computer graphic, computer vision, etc. Common BRDF models include Lambertian, Lommel–Seeliger, Phong etc. (Langovoy and Wübbeler, 2013)

$$BRDF(\theta_i, \Phi_i, \theta_r, \Phi_r) = \frac{L(\theta_r, \Phi_r)}{E(\theta_i, \Phi_i)} \quad (6)$$

where L donates radiance and E donates irradiance.

Lambertian Surface

An ideal diffusive surface that is important in displays obeys Lambert's cosine law, so called Lambertian surface (Warren, 2007). A Lambertian surface reflects light with luminous intensity proportional to cosine of the viewing angle (the angle between the direction of reflected light and the surface normal), whose luminous intensity profile is shown in Fig. 6. More important, its luminance is isotropic regardless the viewing angle, resulting in appearing equally bright to observers in all directions. BRDF of Lambertian surface is a special case that BRDF is a constant of ρ/π where ρ is the albedo.

Viewing Angle and Viewing Cone

Viewing angle is defined as the maximum angle, within which a display can provide acceptable visual performances (luminance, chromaticity, etc.) to the observers (Lee *et al.*, 2008; Yeh and Gu, 2010; Teunissen *et al.*, 2008). A display may perform differently in different directions, hence horizontal and vertical viewing angles are usually used to gauge a display. In fact, the most frequently used viewing angle is defined by the luminance. According to different definitions, the viewing angle in terms of luminance can be defined as the angle where the luminance descends to one half or one third of the on-axis value. To be more comprehensive, a more complicated concept, as viewing zone, is sometimes adopted for a display. Viewing cone is defined as “the range of viewing directions that can satisfactorily be used for the intended task without reduced visual performance” in ISO standards, which is actually the multitude of viewing angles in all viewing directions, leading it to appear in an inverted “cone” with its apex on the display. According to different usage scenarios, different aspects of performances, such as luminance, chromaticity, contrast, can be adopted to define the viewing cone. To conveniently represent a viewing cone, a contour or heat map in a polar coordinate system is usually recommended.

Resolution and Image Sharpness

The resolution of a television, computer monitor, cellphone display etc. is a specification that denotes the distinct pixel amount in each dimension that can be displayed. Usually, resolution with the unit of pixel is expressed in width $\times$ height. For example, 1920×1080 means 1920 pixels in the horizontal direction and 1080 pixels in the vertical direction. For modern displays with a fixed pixel array, such as LCD, digital light processing (DLP) projector, OLED display, etc., the display resolution is directly the number of physical pixels. Some common display resolutions are shown in Table 2. The resolution of a display also refers to the image sharpness of it; that is, the higher resolution, the finer image detail that can be displayed.

Pixels Per Inch

Sometimes, to give not only the total pixel number, but also the pixel density, another parameter associated with display resolution, as pixels per inch (PPI), is also used. By giving PPI, the viewing distance, over which the human eyes cannot distinguish between pixels of a display, can be determined. For instance, a pixel density around 400 PPI is satisfactory for high-end smartphones; virtual-reality (VR) and augmented-reality (AR) devices usually call for a PPI larger than 600; and light field displays even requires a PPI of several thousands. Table 3 shows PPIs of some typical consumer electronic products released these years.

Color Temperature

Color temperature originally comes from ideal black-body radiators that emit light with different hues at different temperatures, and is adopted to evaluate the hue of light sources in Kelvin (K) (Wyszecki and Stiles, 1982). The Planckian locus in Fig. 7 shows the chromaticities of a black-body radiator of different temperatures on CIE 1931 (x, y) chromaticity diagram. Generally speaking, the lower the color temperature is, a white light appears more reddish; the higher the color temperature is, the white is more bluish. For stimuli that do not exactly follow the form of black-body radiators, the concept of correlated color temperature (CCT) is commonly used, which is the temperature of a Planckian radiator whose perceived color most closely resembles that of the given stimulus at the same brightness and under specified viewing conditions. In Fig. 8, chromaticity coordinates having the same CCT are marked with black lines, revealing that different colors may have the same CCT.

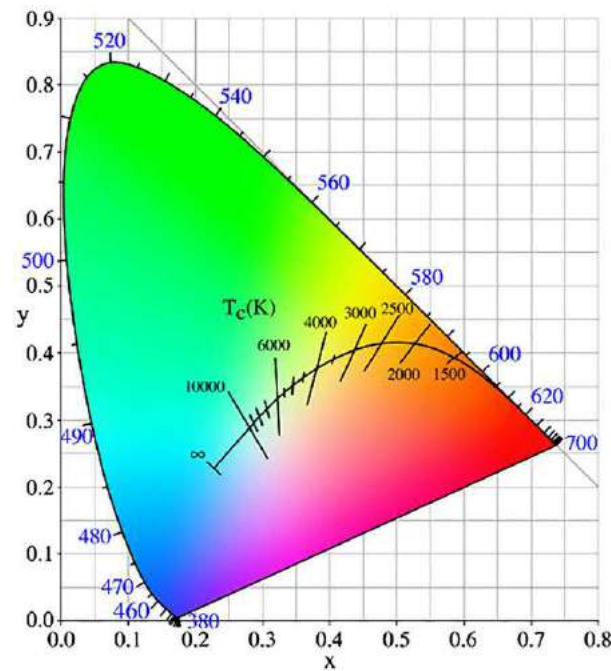
Under NTSC, PAL TV norms, a display should reproduce electrically black and white signals with the minimal color saturation at a color temperature of 6500K (The value may differ in some other norms). By considering that it is high-cost to precisely preprogram or calibrate a display's color temperature to be 6500K or other designated values, many consumer products have

Table 2 Some common display resolutions

Standard name	Aspect ratio	Width (pixel)	Height (pixel)
SVGA	4:3	800	600
XGA	4:3	1024	768
WXGA	16:9	1280	720
HD	16:9	1366	768
HD +	16:9	1600	900
UXGA	4:3	1600	1200
FHD	16:9	1920	1080
WQHD	16:9	2560	1440
4K UHD	16:9	3840	2160
8K UHD	16:9	7680	4320

Table 3 PPIs of some typical consumer electronic products released these years

Product	Release year	Diagonal size (inch)	PPI
Smartphones			
Apple iPhone 4	2010	3.5	330
Samsung Galaxy S II	2011	4.3	216
Apple iPhone 6 Plus	2014	5.5	401
Google Nexus 6	2014	6	490
Sony Xperia XZ Premium	2017	5.8	801
Tablets and laptops			
Apple MacBook Air 13"	2010	13.3	128
Microsoft Surface RT	2012	10.6	148
Amazon Kindle Fire HD	2013	7	216
Apple iPad Air	2013	9.7	264
Apple MacBook Pro Retina 15"	2015	15.4	220

**Fig. 7** Planckian locus on the CIE 1931 (x, y) chromaticity diagram that shows chromaticities of black-body radiators corresponding to different temperatures.

deviations from this requirement, and consequently produce some chromatic differences. To meet different personal preferences of color reproduction, some displays also provide a function of color temperature adjustment.

Color Triangle

In an additive color system, the gamut of all the possible combinations of all the primary colors can be plotted in a polygon with vertexes defined by these primaries (Wyszecki and Stiles, 1982; Smith, 1978). For three primaries, it is a triangle, which is simultaneously the chromaticity gamut when the amounts of the primaries are constrained to be nonnegative, as the Maxwell's triangle shown in Fig. 8. In a Maxwell's triangle, corners of the triangle are three primary colors – red (R), green (G), and blue (B), and other colors are represented by distances from each of the three sides of the triangular to a point inside. After the CIE system was developed, color triangles are used along with chromaticity diagrams. Generally, the sum of the three chromaticity values in modern chromaticity diagrams has a fixed value and only two of the three values are depicted, hence color triangles are plotted in a Cartesian coordinate system. Fig. 9 shows color triangles of some typical color gamuts on CIE 1931 (x, y) chromaticity diagram.

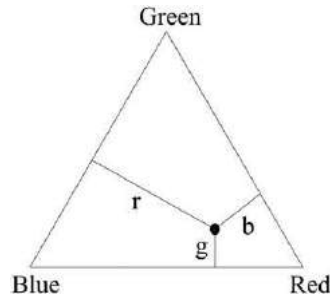


Fig. 8 Maxwell's color triangle, in which a color point (the black dot) can be represented by the distances from each of the three sides, as r , g , and b .

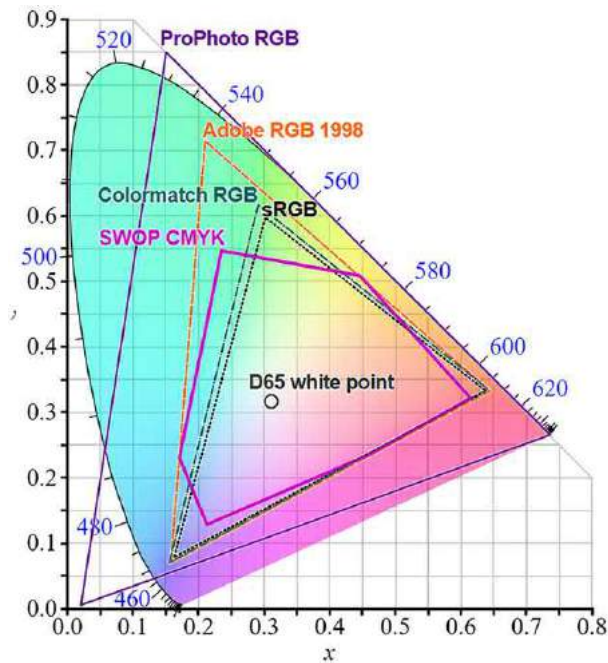


Fig. 9 Color triangles of some typical RGB and CMYK (cyan, magenta, yellow, black) color gamuts, where the name of the color gamuts are marked in the figure.

Response Time

Response time of a display is the time for a pixel takes to change status and is usually with the unit of millisecond ([Lee et al., 2008](#); [Tsujiyura, 2017](#)). Generally, the “change” is defined as that a pixel goes from black to white and back to black again, hence response time is the sum of T_r (time rising) and T_f (time falling). Sometimes, grey-to-grey (GTG) response time that represents the time for a pixel takes to change from a grey value to another is also adopted to acquire a better-looking specification. A long response time may create comet-like tails while objects are quickly moving on a display, so called trailer effect or motion blur; therefore, response time is important to fast moving contents such as sports or video games. LCDs, as the most widely used display technology today, have an inherent issue of low response time, because the liquid crystals need some time to twist. Modern LCDs can achieve a response time of several milliseconds. Blue phase mode LCDs can even decrease the response time to about one millisecond ([Ge et al., 2009](#)).

Refresh Rate and Frame Rate

For displays, refresh rate is the number of times in a second that a display hardware updates its buffer, as known as temporal resolution, while frame rate is the number of times in a second that the image source provides an entire frame of new data to a display. Sometimes, the refresh rate may not equal to the frame rate. For example, an LCD operated under a refresh rate of 60 Hz is

provided with 30 frames per second, so that its refresh rate (60 Hz) is unequal to the frame rate (30 Hz). Since pixels of an LCD do not flash from a frame to another, a minimal refresh rate is not required to eliminate the refresh-induced flicker, whereas this is important for CRTs, whose refresh rate usually needs to be higher than 85 Hz to eliminate the flicker. On the other hand, a high refresh rate of LCDs can smooth the motion objects in the image. Generally, a monitor or TV LCD for regular applications has a refresh rate of 60 Hz. Moreover, LCDs dedicated to video games usually have a refresh rate of 120 Hz, or even 180 Hz to match the high frame rate provided by a graphics card. For some special applications, such as field sequential color displays (Huang *et al.*, 2011), the refresh rate needs to be 240 Hz or even higher. For LCDs with a high refresh rate, the pixel response time should be short enough to realize a rapid grey-to-grey conversion; otherwise, the images will be distorted.

See also: Color Formation of LCD – Spatial and Temporal

References

- Booker, R.L., 1981. Luminance–brightness comparisons of separated circular stimuli. *JOSA* 71 (2), 139–144.
- Chen, D., 2001. Anti-reflection (AR) coatings made by sol–gel processes: A review. *Solar Energy Materials and Solar Cells* 68 (3), 313–336.
- Dearborn Glass Co., 1971. Method of producing glare-reducing glass surface, U.S. Patent 3,616,098, issued October 26.
- Ford, A., Roberts, A., 1998. *Colour Space Conversions*, 1998. London: Westminster University, pp. 1–31.
- Ge, Z., Gauza, S., Jiao, M., Xianyu, H., Wu, S.-T., 2009. Electro-optics of polymer-stabilized blue phase liquid crystal displays. *Applied Physics Letters* 94 (10), 101104.
- Gong, R., Haisong, X., Wang, B., Ronnier Luo, M., 2012. Image quality evaluation for smart-phone displays at lighting levels of indoor and outdoor conditions. *Optical Engineering* 51 (8), doi:10.1117/1.OE.51.8.084001.
- Huang, Y.P., Lin, F.C., Shieh, H.P.D., 2011. Eco-displays: The color LCD's without color filters and polarizers. *Journal of Display Technology* 7 (12), 630–632.
- Hwang, S., Min, J., Lee, M.G., *et al.*, 2009. Novel wide viewing liquid crystal display with improved off-axis image quality in a twisted nematic configuration. *Optical Engineering* 48 (11), doi:10.1117/1.3265524.
- Keller, P.A., 1997. *Electronic Display Measurement: Concepts, Techniques, and Instrumentation*. Wiley-Interscience.
- Kobayashi, S., Mikoshiba, S., Lim, S. (Eds.), 2009. *LCD Backlights* 21. John Wiley & Sons.
- Langovoy, M., Wübbeler, G., 2013. Empirical BRDF models for standard reference materials, EURAMET JRP-i21 Deliverable 4.1.
- Lee, J.H., Liu, D.N., Wu, S.T., 2008. *Introduction to Flat Panel Displays*, 20. John Wiley & Sons.
- Marschner, S.R., Stephen, H.W., Lafortune, E.P.F., Torrance, K.E., 2000. Image-based bidirectional reflectance distribution function measurement. *Applied Optics* 39 (16), 2592–2600.
- McGuire, R.G., 1992. Reporting of objective color measurements. *HortScience* 27 (12), 1254–1255.
- Ohno, Yoshi, 2000. CIE fundamentals for color measurements. In: *NIP & Digital Fabrication Conference*, vol. 2000, no. 2, Society for Imaging Science and Technology, pp. 540–545.
- Pavlovych, A., Stuerzlinger, W., 2005. A high-dynamic range projection system. In: *Photonics North 2005 International Society for Optics and Photonics*, 596920.
- Putman, P.H., 2016. Display Technology: The Next Chapter. *SMPTE Motion Imaging Journal* 125 (3), 30–34.
- Qin, Zong, Yeh, Yen-Wei, Huang, Yi-Pai, David Shieh, Han-Ping, 2015. High ambient contrast ratio OLED display with microlens array and cruciform black matrices. In: *Photonics Conference (IPC)*, 2015, pp. 32–33. IEEE.
- Reinhard, E., Heidrich, W., Debevec, P., *et al.*, 2010. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann.
- Ronchi, L.R., Macii, R., Stefanacci, S.R., Bassan, M., 1980. Brightness and luminance of light-emitting diodes: Influence of size and defocus. *Color Research & Application* 5 (4), 207–211.
- Singh, R., Narayanan Unni, K.N., Solanki, A., 2012. Improving the contrast ratio of OLED displays: An analysis of various techniques. *Optical Materials* 34 (4), 716–723.
- Smith, A.R., 1978. Color gamut transform pairs. *ACM Siggraph Computer Graphics* 12 (3), 12–19.
- Teunissen, K., Qin, S., Heynderickx, I., 2008. A perceptually based metric to characterize the viewing-angle range of matrix displays. *Journal of the Society for Information Display* 16 (1), 27–36.
- Tsujimura, T., 2017. *OLED Display Fundamentals and Applications*. John Wiley & Sons.
- Warren, J.S., 2007. *Modern optical engineering. The Design of Optical Systems*. 205–216.
- Wikipedia, 1931. CIE color space, <https://en.wikipedia.org/wiki/CIE-1931-color-space>.
- Wikipedia, 1960. CIE color space, <https://en.wikipedia.org/wiki/CIE-1960-color-space>.
- Wyszecki, G., Stiles, W.S., 1982. *Color Science*, 8. New York: Wiley.
- Xie, Z.Y., Hung, L.S., 2004. High-contrast organic light-emitting diodes. *Applied Physics Letters* 84 (7), 1207–1209.
- Yeh, P., Gu, C., 2010. *Optics of Liquid Crystal Displays*, 67. John Wiley & Sons.

Relevant Websites

- <https://en.wikipedia.org/wiki/CIELUV>
CIELUV.
- <https://en.wikipedia.org/wiki/Electronic-Industries-Alliance>
Electronic Industries Alliance.
- <https://www.nist.gov/director/pao/nist-general-information>
NIST.
- <http://standards.sae.org/>
SAE.
- <https://www.vesa.org/about-vesa/>
VESA.

Reflective Display Technologies

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Inc. All rights reserved.

In addition to transmissive and emissive displays, reflective displays are also widely used in display applications that call for ultra-low power consumption, sunlight readability, flexibility, low cost etc., such as electronic paper (e-paper), digital signages, electronic shelf labels, wearable displays. For instance, e-paper is a typical application, which is predicted to share more than 10% of the traditional printing and publishing market (Wu and Yang, 2001; Luxresearch, 2015). From 2015 to 2020, the overall market of reflective displays will grow by 30 million dollars reportedly (DisplaySearch, 2016). Several mainstream reflective display technologies including electrophoretic, electro-wetting etc., will be introduced in this chapter.

Electrophoretic Display (EPD)

Electrophoretic is an electro-kinetic phenomenon that refers to the motion of dispersed particles in a fluid when an electric field is applied, which was observed in as early as 1807 (Dukhin and Derjaguin, 1974). Based on this phenomenon, electrophoretic displays (EPDs) are developed (Dukhin and Derjaguin, 1974; Comiskey *et al.*, 1998; Kim *et al.*, 2004; Amundson *et al.*, 2001; Inoue *et al.*, 2002; Oh *et al.*, 2009; Lee and Liang, 2003; Liang *et al.*, 2003; Huang *et al.*, 2013; Li *et al.*, 2016). Generally, an EPD contains two electrode layers sandwiching an ink layer where white particles (usually positively charged titanium dioxide) and black particles (usually negatively charged carbon black) are suspended, as shown in Fig. 1(a). By applying an appropriate voltage across the electrode layers, the particles will move under the electrostatic force. When the white particles gather at the top electrode, the ambient light will be diffusely reflected and a white state will be observed. In the opposite case, the black particles are pushed to the top electrode, hence the ambient light is absorbed and a black state is given. To form an image, active matrix driving technology is usually utilized to separately drive a number of regions of the display. In addition to white and black particles with opposite charges, another alternative principle of EPD is to use particles with opposite charges and colors respectively at both sides, as shown in Fig. 1(b). When an electric field is applied, the particles will be rotated and moved by the field stress; hence, the color appearing to the observers changes. For an actual EPD, the driving waveform, which forces the charged particles to move towards or away from the observers, plays an important role in sequentially erasing the original image, activating the particles, and displaying a new image. Therefore, the driving waveform algorithm should be carefully optimized for driving time, grayscale, flicker and ghost.

The most significant advantage of EPD is its power saving feature introduced by the “bistability,” which denotes that a system has two stable equilibrium states. For displays, “bistability” means a display can retain a still image without power consumption in two states (usually black and white). In an EPD, positions of the charged particles are retained when the applied voltage is removed. Consequently, the displayed image can be kept when the power source is removed and the display consumes little power when it shows still images. In contrast, in traditional LCDs, the backlight needs to be lightened and the driving circuit needs to be rapidly refreshed all the time. Next, EPDs use basically the same pigments as the regular ink for printed paper to reflect the ambient light, hence it has the same satisfactory contrast ratio and viewing angle as printed paper, especially in the bright sunlight. Thirdly, the displayed images of EPDs are essentially generated by reflected ambient light, hence the healthily harmful light possibly in transmissive and emissive displays can be prevented. Fourthly, the resolution of EPDs can directly benefit from active matrix driving technology that is fast growing (Amundson *et al.*, 2001). Additionally, all components of an EPD can be made flexible to realize a bendable or even rollable display. Therefore, EPDs are now widely used for e-readers and gain increasing attention for architecture displays, wearable displays, electronic signages etc. However, the response time of EPDs of hundreds of milliseconds somewhat restricts their application for animation or video contents. In addition, due to the difficulty of accurately controlling the charged particles, the bit number of gray-level of an EPD is a little poor in comparison with conventional LCDs. For example, 6-bit gray-level is quite prominent for an EPD while an entry-level LCD can easily achieve 8- or even 10-bit gray-level. This issue of bit number further affects the color rendering capacity. Currently, an active research area of EPD is to enhance the color gamut to expand the application scenarios.

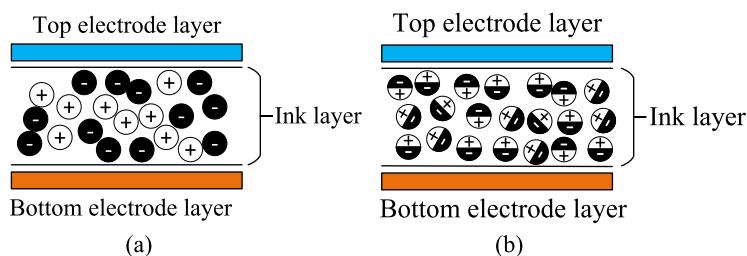


Fig. 1 Basic principle of EPD: (a) using black and white particles with opposite charges; (b) using opposite charges and colors at both sides of a single particle.

Microencapsulated Electrophoretic Display

In an EPD, to solve the issues of lateral drift caused by the gravitation and particle agglomeration caused by the direct contact between the particles and electrodes, the particles suspended in a fluid are enclosed in tiny microcapsules with a scale of dozens of microns, so called microencapsulated EPD (Oh *et al.*, 2009). Microencapsulated EPDs can achieve improved reliability and realize direct coating of microencapsulated electrophoretic materials on the substrate, which leads to potential flexibility. Fig. 2 shows the principle of a monochrome microencapsulated EPD using white and black pigments with opposite charges. The applied voltage pushes the charged white or black pigments to the top electrode to make the display appear white or black to the observers. The voltage applied to each microcapsule can also be bifurcated, then the mixture of white and black particles can realize different gray levels. If color filters are added onto these microcapsules, reflected ambient light passing through the color filters can realize a chromatic microencapsulated EPD (Oh *et al.*, 2009), as shown in Fig. 3.

Microcup[®] Electrophoretic Display

Another practical EPD technology uses microcups to enclose the particles and fluid, so called microcup[®] EPD (Lee and Liang, 2003; Liang *et al.*, 2003; Huang *et al.*, 2013). An example of monochrome microcup[®] EPD with two types of pigment is shown in Fig. 4. The size of a microcup, which can be prepared by a lithographic or embossing process, may vary from 50 to 180 μm in diameter, 12 to 40 μm in height, and 8 to 25 μm in partition width, depending on the application. A unique feature of microcup[®] EPD is that the microcups can be easily imprinted onto a flexible substrate and driving backplane with a mature roll-to-roll (R2R) processing, thus flexible microcup[®] EPDs can be produced at relatively low cost rather than microencapsulated EPDs. Additionally, microcup[®] EPDs can also be made chromatic by using chromatic fluid or a three-pigment-ink system (Oh *et al.*, 2009; Lee and Liang, 2003; Liang *et al.*, 2003; Huang *et al.*, 2013; Li *et al.*, 2016), as shown in Fig. 5. Benefiting from no use of color filters, the reflectivity can be made quite high for a chromatic microcup[®] EPD.

Electro-Wetting Display (EWD)

The electro-wetting phenomenon refers to a modification of the wetting properties of a surface, which is typically hydrophobic, with an applied electric field (Lippmann, 1875). In electro-wetting, two terms including surface tensions and electrostatics are involved to acquire a force equilibrium, as given by Eq. (1).

$$\gamma_{LV}\cos\theta = \gamma_{SV} - \gamma_{SL} + \frac{\epsilon_0\epsilon_r}{2d}V^2 \quad (1)$$

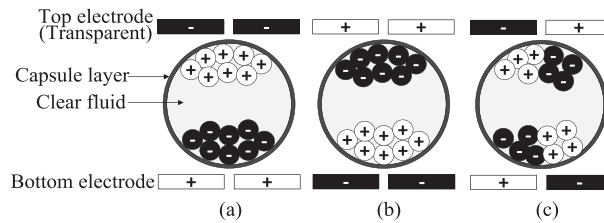


Fig. 2 Principle of a monochrome microencapsulated EPD where (a), (b) and (c) show the cases of white, black and gray visible to the observer, respectively.

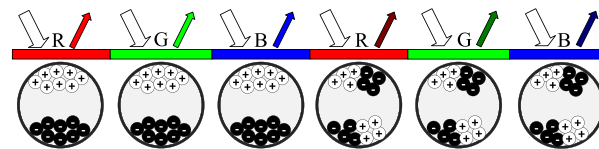


Fig. 3 Principle of a chromatic microencapsulated EPD where red (R), green (G) and blue (B) color filters are added.

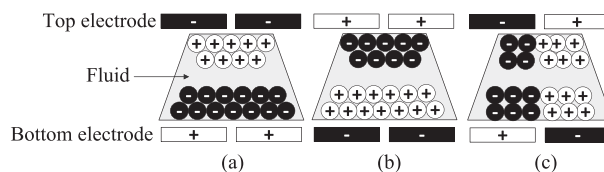


Fig. 4 Principle of a monochrome microcup[®] EPD where (a), (b) and (c) show the cases of white, black and gray, respectively.

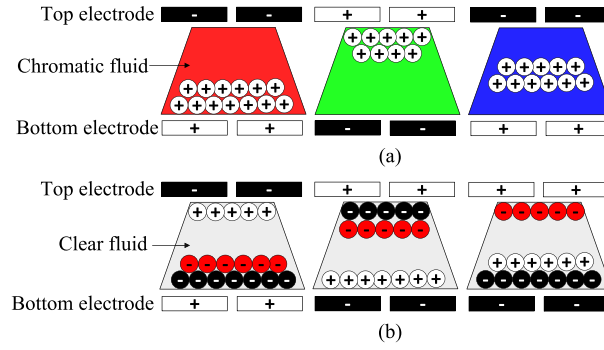


Fig. 5 Principle of a chromatic microcup® EPD: (a) using chromatic fluid and control the position of particles in a microcup to generate colors with different brightness; (b) using a three-pigment-ink system to generate white, black and one more color, which is usually used to enable electronic signages.

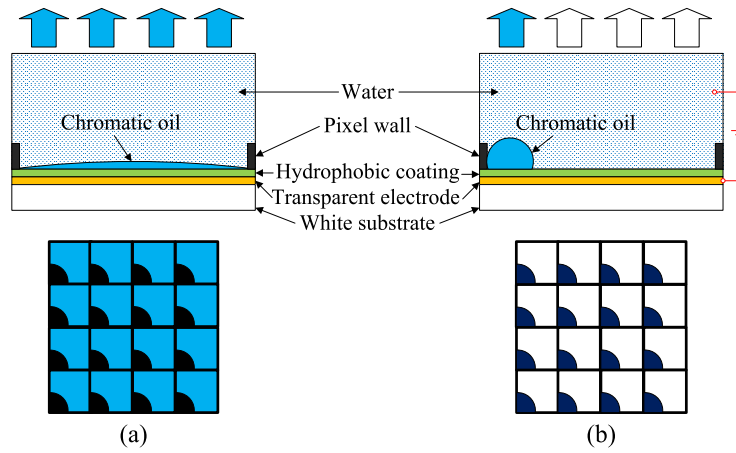


Fig. 6 Principle of a typical EWD using chromatic oil and water: (a) colored (blue) pixel; (b) white pixel where part of the electrode is omitted in the lower left corner of each pixel to control the oil motion.

where γ_{LV} , γ_{SV} , γ_{SL} is respectively surface tension of the liquid/vapor, solid/vapor, and solid/liquid interface; θ is the contact angle; ϵ_r is dielectric constant; d is the thickness of the hydrophobic insulator.

Electro-wetting has been utilized for a variety of applications, such as displays, optical filters, fiber optics, adaptive lenses, lab-on-a-chip. Among these applications, electro-wetting display (EWD) is one of the most important, which was firstly proposed in 2003 (Hayes and Feenstra, 2003). Fig. 6 shows the principle of a typical EWD (Hayes and Feenstra, 2003; Feenstra and Hayes, 2009; Hayes et al., 2004; Blankenbach, 2008; Burns et al., 2005; Charipar et al., 2015; You and Steckl, 2010). In a pixel, the non-polarized oil containing chromatic dye is dropped on a hydrophobic electrode with an insulating coating, and enclosed by hydrophilic ribs. The oil and electrode are immersed in a polarized water solution. When no voltage is applied, the oil forms a homogeneous film between the water and the insulating coating to acquire the minimum Gibbs free energy, resulting in a colored pixel, as shown in Fig. 6(a). When a voltage is applied across the hydrophobic insulating coating, an electrostatic force is added to the original force equilibrium. To acquire the minimum Gibbs free energy again, the water is moved into contact with the insulating coating; thus, the oil is pushed aside and the white substrate is exposed to form a white pixel, as shown in Fig. 6(b). For a typical pixel size of about 200 μm , the interfacial tension force is more than 1000 times stronger than the gravitational force, thus the displayed images are stable in all orientations. Moreover, the state of a pixel can be continuously tuned between on and off states by setting a certain balance between the electrostatic and surface tension forces. Consequently, different gray levels can be analogously acquired for an EWD. If pulse width modulation (PWM) is adopted, different gray levels can also be digitally realized. Certainly, black or transparent oils can also be utilized in combination with a substrate of different colors, and the working principle is similar except the pixel color.

The switching time between on and off states of an EWD is typically several milliseconds; hence, an EWD is sufficiently fast to show video contents. Current EWDs can go through over a billion times of switches without a degradation of electro-optical performances, revealing a long working life of EWD for video contents. Moreover, under a low operation voltage of around 20 V, the power consumption of an EWD is just about one third of that of conventional transmissive LCDs, leading to an attractive feature of power saving. Due to the reflection of ambient light via the oil or substrate, reflectivity, contrast ratio, viewing angle, sunlight readability etc. of EWDs is comparable to those of regular printed paper. Finally, since the electro-optical performances of

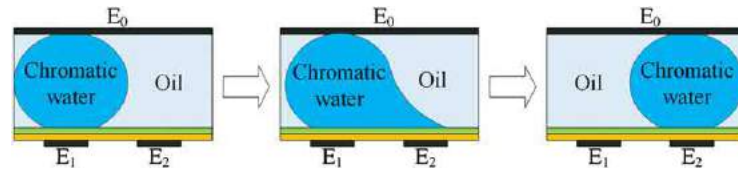


Fig. 7 Principle of a bistable EWD utilizing droplet movement. E_0 is common electrode, and E_1 , E_2 are structured control electrodes, which are applied opposite waveforms in phase while moving the droplet.

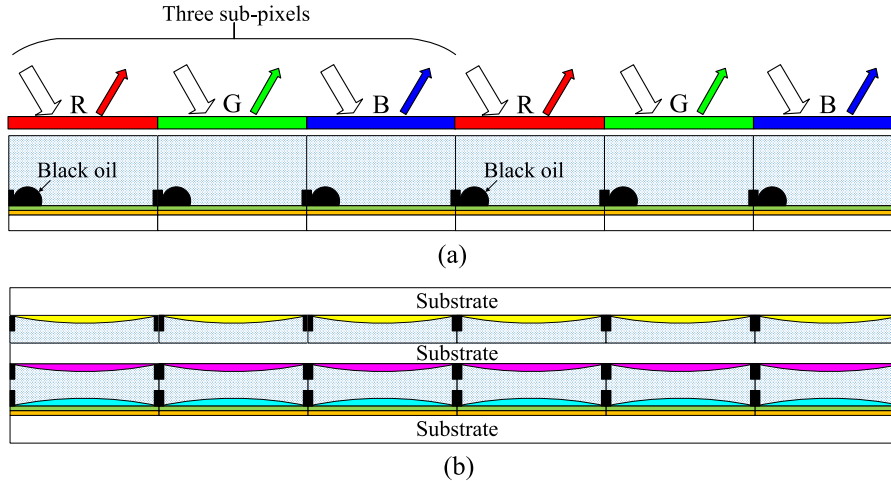


Fig. 8 Principle of a chromatic EWD: (a) adding a color filter onto each pixel containing black oil; (b) stacking three monochrome layers (cyan, magenta and yellow).

an EWD pixel is insensitive to the cell-gap, which is hard to be homogeneous across the screen for a flexible display, the fabrication of flexible EWDs is relatively easy and low-cost (Burns *et al.*, 2005). Furthermore, EWDs and conventional LCDs share around 90% of the production process, leading to a promising prospect of commercialization. However, an EWD utilizing droplet (oil) contracting and relaxing is not bistable, hence it consumes power when still images are shown.

To make EWD adapted for applications that need a bistability, such as e-readers and digital signages, some researchers have realized bistable EWDs (Charipar *et al.*, 2015) by utilizing droplets moving from one position to another, as shown in Fig. 7. Nevertheless, the response time of a bistable EWD is lower than that of a non-bistable one.

EWDs can also be adopted to realize low-cost and power-saving color displays (You and Steckl, 2010). The first approach is simply adding color filters onto each pixel containing black oil to form a chromatic sub-pixel, as shown in Fig. 8(a). As the manufacturing process for traditional LCDs using color filters can be directly utilized here, the cost is considerably low while a higher optical efficiency than that of LCDs can be offered because of no polarizers. The second approach is stacking three monochrome layers that can be separately controlled, as shown in Fig. 8(b). By adjusting the driving voltage of each layer, a desired color can be generated via subtractive color-mixing principle. This triple-layer architecture can directly adopt the processes for single-layer to suppress the cost, in addition to a much higher optical efficiency. Benefiting from the triple-layer structure, the color rendering capacity of such an EWD is much higher than that of other reflective displays. Furthermore, due to the subtractive color-mixing principle, the color gamut and brightness have a win-win relationship for a triple-layer EWD, while a trade-off needs to be endured for a conventional LCD adopting additive colors.

Other Reflective Display Technologies

In addition to EPD and EWD, there are several other reflective display technologies, such as quick response liquid powder display (QR-LPD) (Hattori *et al.*, 2004), cholesteric liquid crystal display (ChLCD) (Tamaoki, 2001), interferometric modulator display (IMOD, trademarked mirasol[®] by Qualcomm) (Bita *et al.*, 2012), memory LCD (Tamaki *et al.*, 2014). Although they have not been widely commercialized, several attractive features can be found in these technologies. Here their principles and features are briefly summarized in Tables 1 and 2.

- (1) QR-LPD[®]: A QR-LPD[®] is based on electrophoresis, where high-fluidity black and white “liquid powders” at nanometer scale are charged and suspended in the air to be driven by two electrodes sandwiching the powders. When a negative voltage is applied to the upper transparent electrode, the black powders with positive charges aggregate on the top, and then bring

Table 1 Principles and features of several reflective display technologies (1)

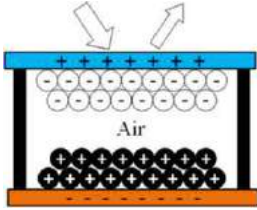
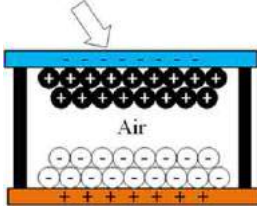
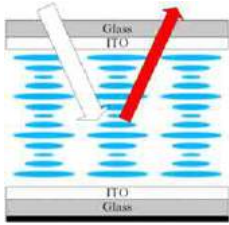
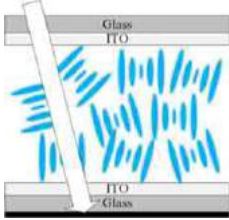
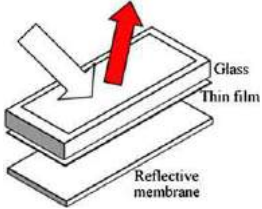
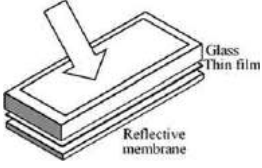
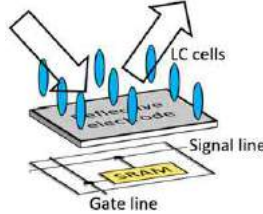
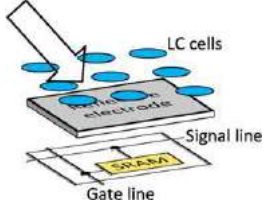
Technology	QR-LPD	ChLCD
Principle	White state  Black state 	 
Advantages	Fast response (< 1 ms), low temperature driving, wide viewing angle, power saving due to bistability	No use of color filter or polarizer, wide color gamut, power saving due to bistability
Disadvantages	High driving voltage, low contrast ratio	Slow response (~ 300 ms), low resolution, high cost, high driving voltage
Leading companies	Bridgestone	Fujitsu, Fujitsu Xerox, Kent Displays, ITRI

Table 2 Principles and features of several reflective display technologies (2)

Technology	IMOD	Memory LCD
Principle	White state  Black state 	 
Advantages	Fast response (~ 10 μ s), power saving due to bistability, high contrast ratio, high reflectivity ($> 60\%$), high robustness	Low power consumption due to bistability
Disadvantages	Low resolution of gray scale, high cost	Limited resolution and color rendering capacity due to the complication of SRAM circuit integration
Leading companies	Qualcomm	Sharp

about a black appearance, while a positive voltage applied to the upper electrode produces a white appearance in the opposite case. Reflection by the white powders makes a QR-LPD to have a paper-like appearance and wide viewing angle. Due to the high fluidity of the liquid powders induced by their surface morphology and electronic property, response time of

- a QR-LPD is ultra-fast, as shorter than 1 ms. Additionally, the electric field forces between the electrodes and liquid powders give a QR-LPD a bistability feature, resulting in ultra-low power consumption. By optimizing the charge amounts for liquid powders of different colors, a colorful QR-LPD can be realized.
- (2) ChLCD: A cholesteric liquid crystal with chirality caused by a helical structure is contained in a ChLCD. The liquid crystal is arranged in layers with no positional ordering within layers, while director axes of the layers vary periodically, leading to Bragg reflection to the incoming light. When the incident wavelength is between $n_o p$ and $n_e p$, where n_o is the ordinary refractive index, n_e is the extraordinary refractive index, and p is the cholesteric pitch over which the cholesteric liquid crystal molecules fully rotate 360° along their helical, a certain light wavelength is selectively reflected. In this way, it is possible to realize vivid color images without color filters and polarizers. On the other hand, a focal conic domain configuration makes the light pass through and absorbed by an absorption layer beneath to make a dark state. A ChLCD is also bistable (planar state and focal conic state), thus little power is consumed to maintain a still image. To achieve a full-color ChLCD, a stacked-panel technology can be adopted.
 - (3) IMOD: Vivid colors on a butterfly's wings or a peacock's feathers, which are generated by light interference, are imitated. In an IMOD, each pixel is actually a micro-electromechanical system (MEMS) device that contains an air gap sandwiched by a thin-film-coated glass substrate on the top and a reflective conductive membrane at the bottom. The thickness of the air gap at micrometer scale can be tuned by applying a voltage on the top glass substrate and bottom membrane. In this way, a constructive or destructive interference occurs; and then the incoming light with a certain wavelength can be selectively reflected or collapsed. To build a full-color pixel, grids of red, green and blue color subpixels are combined and gray levels are produced by using spatial dithering. An IMOD calls for very little power to maintain either one of its two states to realize a bistable display. Moreover, the interference light from the ambience leads to a great sunlight readability.
 - (4) Memory LCD: Memory LCD, as known as memory-in-pixel (MIP) reflective LCD, adopts reflective electrodes under LC cells and integrates a static random access memory (SRAM) circuit and LC driver into each subpixel (under the reflective electrodes) using conventional TFT technologies. When a still image is displayed, benefiting from the SRAM circuit, only the pixel circuit is driven, leading to no charging current to the data line and a very small amount of charging current to the pixel capacitor. Therefore, a memory LCD is bistable-like and the power consumption of it is ultra-low. Nevertheless, due to the considerable complication of integrating a memory circuit into each subpixel, the resolution of a memory LCD is highly limited. In addition, the bit number of the in-pixel memory is also limited, as only one or two bit; thus, the color rendering capacity is also low in comparison with other reflective display technologies.

References

- Amundson, K., Ewing, J., Kazlas, P., *et al.*, 2001. Flexible, active-matrix display constructed using a microencapsulated electrophoretic material and an organic-semiconductor-based backplane. *SID Symposium Digest of Technical Papers*, vol. 32(1), pp. 160–63.
- Bita, I., Govil, A., Gusev, E., 2012. Mirasol[®]—MEMS-based direct view reflective display technology. In *Handbook of Visual Display Technology*. Berlin/Heidelberg: Springer.
- Blankenbach, K., Schmoll, A., Bitman, A., Bartels, F., Jerosch, D., 2008. Novel highly reflective and bistable electrowetting displays. *Journal of the Society for Information Display* 16 (2), 237–244.
- Burns S.E., Reynolds, K., Reeves, W., *et al.*, 2005. Flexible active-matrix displays. *SID Symposium Digest of Technical Papers*, vol. 36 (1), 19–21.
- Charipar, K.M., Charipar, N.A., Bellemare, J.V., Peak, J.E., Piqué, A., 2015. Electrowetting displays utilizing bistable, multi-color pixels via laser processing. *Journal of Display Technology* 11 (2), 175–182.
- Comiskey, B., Albert, J.D., Yoshizawa, H., Jacobson, J., 1998. An electrophoretic ink for all-printed reflective electronic displays. *Nature* 394 (6690), 253–255.
- DisplaySearch, 2016. E-paper Display Revenues to Reach US\$9.6B by 2018. [https://www.thefreelibrary.com/E-paper+display+revenues+to+reach+US\\$249.6B+by+2018%3a+DisplaySearch.-a0208133110](https://www.thefreelibrary.com/E-paper+display+revenues+to+reach+US%249.6B+by+2018%3a+DisplaySearch.-a0208133110).
- Dukhin, S.S., Derjaguin, B.V., 1974. *Electrokinetic Phenomena*. New York: J. Wiley and Sons.
- Feenstra J. and Hayes R., 2009. Electro-wetting Displays, Liquavista White Paper. <http://www.liquavista.com/media/772/LQV0905291LL5-15.pdf>.
- Hattori, R., Yamada, S., Masuda, Y., Nihei, N., 2004. A novel bistable reflective display using quick-response liquid powder. *Journal of the Society for Information Display* 12 (1), 75–80.
- Hayes R.A., Feenstra, B.J., Camps, I.G.J., *et al.*, 2004. A high brightness colour 160 PPI reflective display technology based on electrowetting. *SID Symposium Digest of Technical Papers*, vol. 35 (1), pp. 1412–1415.
- Hayes, R.A., Feenstra, B.J., 2003. Video-speed electronic paper based on electro-wetting. *Nature* 425 (6956), 383–385.
- Huang, Y.-P., Lin, F.-C., Wu, S.-I., Yang, P.-R., 2013. Mechanism and improvement of charged-particles transition in microcup electrophoretic displays. *Journal of Display Technology* 9 (8), 619–625.
- Inoue, S., Kawai, H., Kanbe, S., Saeki, T., Shimoda, T., 2002. High-resolution microencapsulated electrophoretic display (EPD) driven by poly-Si TFTs with four-level grayscale. *IEEE Transactions on Electron Devices* 49 (9), 1532–1539.
- Kim, M.K., Am Kim, C., Ahn, S.D., Kang, S.R., Suh, K.S., 2004. Density compatibility of encapsulation of white inorganic TiO₂ particles using dispersion polymerization technique for electrophoretic display. *Synthetic Metals* 146 (2), 197–199.
- Lee, H., Liang, R.C., 2003. SiPix Microcup[®] electronic paper – an introduction. *Advanced Display* 3, 4–9.
- Li, G.X., Qin, S.C., Feng, Y.Q., Fang, S., Meng, S.X., 2016. Preparation and characterization of microcapsule-encapsulated colored electrophoretic fluid in trifluorotoluene system for electrophoretic display. *Journal of Display Technology* 12 (10), 1145–1151.
- Liang, R.C., Hou, J., Zang, H., Chung, J., Tseng, S., 2003. Microcup[®] displays: electronic paper by roll-to-roll manufacturing processes. *Journal of the Society for Information Display* 11 (4), 621–628.
- Lippmann, M.G., 1875. Relation entre les phénomènes électriques et capillaires. *Annales de Chimie et de Physique* 5, 494.
- Luxresearch, 2015. Picking the win, place and show in the emerging display technologies horserace. <http://blog.luxresearchinc.com/blog/tag/reflective-displays/>.
- Oh, S.W., Kim, C.W., Cha, H.J., Pal, U., Kang, Y.S., 2009. Encapsulated-dye all-organic charged colored ink nanoparticles for electrophoretic image display. *Advanced Materials* 21 (48), 4987–4991.

- Tamaki, M., Fukunaga, Y., Mitsui, M., *et al.*, 2014. A memory-in-pixel reflective-type LCD using newly designed system and pixel structure. *Journal of the Society for Information Display* 22 (5), 251–259.
- Tamaoki, N., 2001. Cholesteric liquid crystals for color information technology. *Advanced Materials* 13 (15), 1135–1147.
- Wu, S.-T., Yang, D.-K., 2001. *Reflective Liquid Crystal Displays*. New York: Wiley.
- You, H., Steckl, A.J., 2010. Three-color electro-wetting display device for electronic paper. *Applied Physics Letters* 97 (2), 023514.

In Situ Optical Tissue Diagnostics/Miniaturized Optoelectronic Sensors for Tissue Diagnostics

Seung Yup Lee, Georgia Institute of Technology/Emory University, Atlanta, GA, United States

Yooree Grace Chung and Mary-Ann Mycek, University of Michigan, Ann Arbor, MI, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Optical sensing technology has been widely employed in a variety of biomedical applications due to its noninvasive and nonionizing nature (Mycek and Pogue, 2003). The core components comprising the optical sensor are a light source and a detector. The light source excites a sample with different intensities, wavelengths, time-scales, and coherences of light suited to the biophysical characteristics being investigated. The detector receives the re-emitted light that has interacted with tissues and typically converts a photon signal into an electrical signal to be read by an electronic recording system. Conventional light sources include various types of lasers and lamps to generate photons. Light detecting devices commonly used in tissue optical sensing include photomultiplier tubes (PMTs), silicon photodiodes (PDs), and image sensors such as complementary metal-oxide-semiconductor (CMOS) and charged-couple device (CCD).

Typically, these optical components are bulky, resulting in a large footprint of the final instrument that integrates all the optical and electronic hardware units (Pitts and Mycek, 2001; Lloyd *et al.*, 2010). The large size of these photonic medical devices can limit their mobility, requiring patients to visit the clinical location of the device. Further, if an optical fiber is employed to deliver light to tissue, disadvantages could include inefficient coupling, limited bend radius, the possibility of breakage and low spatial information.

During the past couple of decades, advances in techniques for handling new semiconductive materials, micro-/nano-fabrication technologies, and tiny electronics with high computing power have revolutionized the engineering of miniaturized optoelectronic sensors. Compared to fiber-based optical sensors, miniaturized optoelectronic sensors offer substantial benefits. Dramatic reductions in size, cost and power consumption have great potential to make light-based sensing more ubiquitous than ever. Beyond implementing point-of-care diagnostics, these sensors can be wearable, enabling continuous, noninvasive and object monitoring anytime and anywhere. The end terminal of this “fiber-less” sensor is an electrical signal that can be wirelessly transmitted, thereby providing subject’s mobility with tremendous flexibility. Miniaturized sensors also enable optical inter-rogation in human organs where traditional fiber-optic-based approaches fulfill very limited tasks.

In this article, we will review the basic construction of miniaturized sensors, major optoelectronic components and assembly techniques, data processing algorithms and their tissue diagnostics applications.

Optoelectronic Sensing System

Fig. 1 illustrates a block diagram of the miniaturized optoelectronic sensing system. The system can be divided into two parts. One is a sensing unit comprised of optoelectronic modules that send and receive photons to and from biological tissues, in which the conversion process occurs between photonic and electrical signal. The other is a control electronics unit that controls the excitation and reading sequence via an electrical signal sent or received by the sensing unit. Depending on the sensor design and application, only the sensing unit can be constructed in a miniaturized form (under a few millimeter scale), communicating with relatively large control electronics located on a remote site via physical wiring. The detailed technique for sensor unit fabrication is discussed in a later section. Alternatively, the two parts can be integrated in one platform utilizing the CMOS microfabrication technique of electrical circuitry or direct use of minimal electronic bare chips. This integrated optoelectronic sensor can independently work as a separate sensor node.

The sensing unit consists of an optoelectronics source, an optoelectronics detector and micro-optic components such as micro-lenses or color filters. The optoelectronic source is powered to emit light at a predetermined intensity given voltage and current supplied by the control electronics. The use of a micro-lens on the source’s emission aperture adjusts the divergence of the beam. The incident photon interacts with a variety of micro-organisms, and its properties are altered (e.g., intensity – attenuated by natural chromophores such as hemoglobin, wavelength – fluorescence molecules) as it propagates through the tissue. The optoelectronic detector captures remitted light exiting the tissues, which contains information about tissue morphology, physiology or metabolism. Before light enters the detector, a micro-lens can determine an acceptance angle, and a micro-fabricated filter selects the wavelength range. We will not discuss the types and technologies of the micro-lens and optical filters as they are beyond the scope of this article.

Optical contrasts used by optoelectronic sensors are classified into two groups: exogenous and endogenous. Exogenous agents such as fluorescence protein, oxygen-sensitive dye and gold nanoparticle can be inserted or labeled with a specific target molecule to enhance the signal. For label-free sensing, the optoelectronic sensor can employ endogenous contrasts. In reflectance mode, scattering from micro-organelles such as nuclei and mitochondria, and absorption by natural chromophores such oxy- and deoxy-

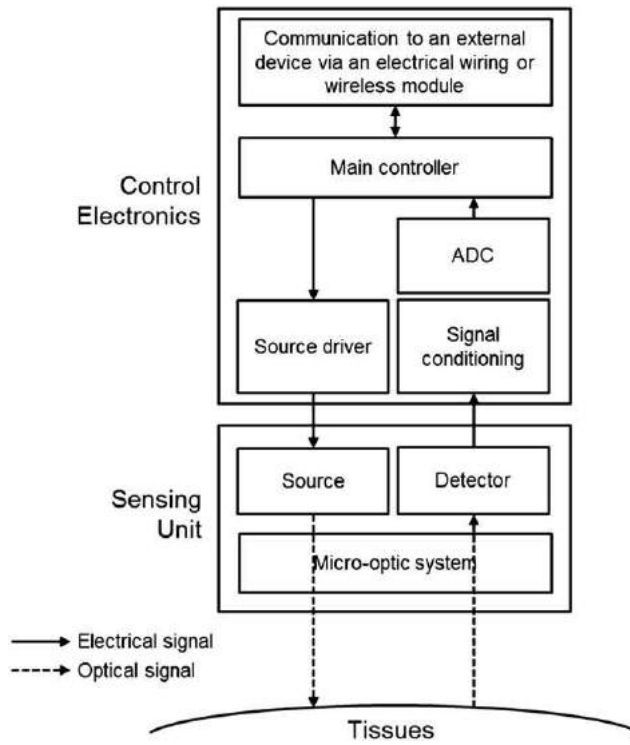


Fig. 1 Schematic diagram of an optoelectronic sensor for tissue diagnostics. ADC: analog-to-digital converter.

hemoglobin, bilirubin and beta-carotene can alter the detected reflectance. In fluorescence mode, metabolic cofactors such as nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) and flavin adenine dinucleotide (FAD) or extracellular matrix (e.g., collagen) can serve as intrinsic fluorescence sources (Lloyd *et al.*, 2013).

In general, there are three different types of techniques in optical sensing relating to domain: continuous, frequency domain and time domain. Most current miniaturized optoelectronic sensors are based on continuous sensing (only light intensity measurements) due to the technical complexity in miniaturization of necessary components for frequency- and time-domain sensing. Thus, we premise the continuous sensing technique when we discuss design, data analysis and applications of the optoelectronic sensors.

Designing optoelectronic sensors for tissue diagnostics considers numerous engineering factors including electrical, optical, thermal, and mechanical aspects. The signal-to-noise ratio (SNR) is one of the most important metrics to achieve the basic function of the designed sensor. SNR is defined as

$$\text{SNR(dB)} = 20 \log \frac{\mu}{\sigma}. \quad (1)$$

Here, μ is the mean and σ is the standard deviation of the repetitive measurements on a sample. Typically, tissue-simulating phantoms mimicking optical properties of targeted biological tissues are employed to obtain the SNR of the sensor. SNR is determined by various elements such as source intensity, detector sensitivity, source–detection separation and electronics design (source driver and signal conditioning circuit). The maximum of excitation intensity is ideally desirable for highest SNR, but it needs to be carefully selected to prevent thermal damage of interrogated tissues. In applications requiring sensor implantation or direct contact with human tissues, a biocompatible encapsulation of the sensor is indispensable to avoid tissue compromise via physical and chemical isolation. At the same time, this encapsulation protects the sensor against the watery environment of tissues.

Thermal stability is another critical factor to be considered in optoelectronic sensor design for two main reasons. Because the operation of the optoelectronic components is highly sensitive to temperature, the optical performance of the designed sensor could vary without careful design of heat sink or transfer. In a miniaturized form, a heat sink might not be a feasible option. Ideal design would have negligible or stabilized temperature increase from heat dissipation of source operation, even during long-term use. Optical and thermal simulation tools are useful for evaluating the initial feasibility of the design.

Optoelectronic Components

Laser diode (LD)

A laser diode (LD), also known as an injection diode laser, is a forward-biased semiconductor diode that emits coherent light when electrons and holes are stimulated by an electrical current into the p–n junction (Fig. 2(a)). LDs consist of two stacked layers of

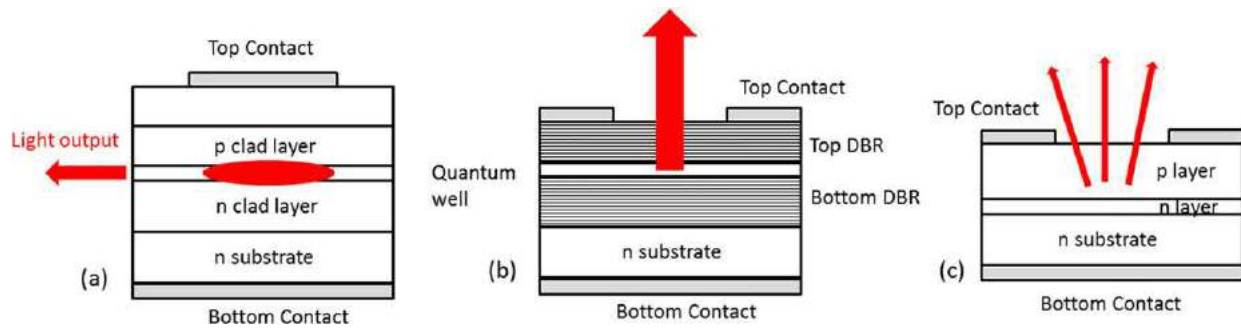


Fig. 2 Illustration of basic structure of (a) laser diode (LD), (b) vertical-cavity surface-emitting laser (VCSEL), and (c) light-emitting diode (LED).

semiconductors as the optical gain medium. The top layer is a p-doped semiconductor, and the bottom layer is an n-doped semiconductor. The layer in between the two semiconductors is the p–n junction, known as the intrinsic region. The intrinsic region is the resonant optical cavity of the laser. When a current is passed through the semiconductors, electrons from the n-type semiconductor and holes from the p-type semiconductor flow into the intrinsic region. Recombination of an electron and a hole releases energy in the form of a photon of light. This photon remains in the intrinsic region and stimulates further release of electrons and holes. The newly formed photons will then be in the same phase, direction and polarization, thereby producing coherent light. Light emerges from one end of the diode and diffracts but can be formed into a straight beam with the use of a collimating lens. Because laser beams are concentrated to a small point, LDs are useful for targeted medical applications that require greater penetration.

Vertical-cavity surface-emitting laser (VCSEL)

A vertical-cavity surface-emitting laser (VCSEL) is a forward-biased semiconductor diode that emits light perpendicular to the laser's top surface as compared to a conventional edge-emitting laser that emits light in the parallel direction (**Fig. 2(b)**). A VCSEL consists of three main layers: the upper p-type distributed Bragg reflector (DBR), the active region with quantum wells, and the lower n-type DBR. The two DBR layers are highly reflective mirrors with a reflectivity of 99.5–99.9%. Photons formed in the active region oscillate between the two DBR layers until they escape as coherent light perpendicularly from the top of the laser. The perpendicular light beam enables three times as many VCSELs to be produced from the same size wafer as compared to edge-emitting lasers. The perpendicular beam also allows VCSELs to be incorporated into two-dimensional arrays to maximize output power and device reliability. Other advantages of VCSELs over edge-emitting lasers include easy coupling with optical fibers due to circular beams, less power consumption, greater quality control during production, and lower production costs.

Light emitting diode (LED)

A light-emitting diode (LED) is a forward-biased semiconductor diode that produces incoherent light. An LED includes a p–n junction where electrons and holes generated from an electric current are recombined (**Fig. 2(c)**). Energy is released as light and heat. The mechanism of light generation is similar to LDs; however, LEDs do not have stimulated emission or coherent light emission. Because LEDs distribute their power incoherently, they are useful for treatment of larger areas of shallow depth. Low power consumption and heat generation associated with LEDs result in improved device efficiency. LEDs are composed of inorganic semiconductors, commonly gallium arsenide phosphide, with long life times. Failure occurs as a gradual, rather than abrupt, decrease in brightness. The durability, compact size, and low production costs of LEDs make them well suited for medical applications. Unlike LDs that emit a coherent light with a narrow bandwidth (~ 1 nm) in a small divergence angle, LEDs emit an incoherent light with a broader bandwidth (typically 25–50 nm) in a larger divergence angle. LEDs come in smaller packages and have more flexibility in available wavelengths than LDs.

Silicone photodiode

A photodiode (PD) is an optical signal receiver. PDs function as semiconductor diodes that detect light with the use of a p–n junction. Incident light is focused onto the junction of the PD. As light enters the charge-depleted region between the p-type and n-type semiconductors and strikes an atom within the crystal structure of the PD, electron–hole pairs are generated. The built-in electric field of the charge-depleted region causes the electrons to drift toward the cathode and the holes to drift toward the anode, creating a photocurrent. The greater the intensity of incident light, the more electron–hole pairs are created. Thus, the generated photocurrent is dependent on light intensity. The total current of a PD includes dark currents, which must be accounted for in calibration. There are different modes of PD operation: the photovoltaic mode, the photoconductive mode, and avalanche PDs. The photoconductive mode creates voltage by restricting the photocurrent flow. In the photoconductive mode, the PD is reverse biased, which increases the width of the charge-depleted region. This width increase decreases the capacitance of the p–n junction, reduces the response time of the PD and increases the linearity of light intensity measurements. However, the photoconductive mode also increases the amount of dark current and noise. Avalanche PDs are highly reverse biased, multiplying the current, and increasing the response of the PD.

Silicone photomultiplier (SiPM)

A silicon photomultiplier (SiPM) is an avalanche PD array joined together on a silicon substrate. Silicon can detect light in the wavelength range 350–800 nm. The array is composed of PD and quenching resistor microcells arranged in parallel. SiPMs have single-photon sensitivity because each individual microcell can detect an incident photon. When a photon is detected, the microcell avalanches as the silicon becomes conductive and amplifies the current (Geiger discharge process). The quenching resistors then stop the avalanche, and the microcells are reset to detect another photon. The released current from all of the microcells can be measured to detect the intensity of light. Therefore, the number of microcells activated is proportional to the amount of incident light. SiPMs are advantageous due to their sensitivity, low operating voltage, and high timing performance. Because of their single photon sensitivity, SiPMs are useful in low-light applications. Currently, SiPMs are being considered for use in positron emission tomography and single-photon emission computed tomography.

Organic LED and PD (OLED and OPD)

Organic LEDs and PDs are composed of organic semiconductors and polymers. Organic polymers allow for greater flexibility and transparent displays. OLEDs and OPDs have several advantages over inorganic LEDs and PDs. Inorganic-based optoelectronics are difficult to scale for large-area production. Organic optoelectronics, however, can be manufactured with flexible substrates and with solution processing, resulting in cheaper and larger volume production. Inorganic components also tend to be more rigid and bulkier. Flexible organic polymers allow organic parts to be used in a wider variety of medical applications. The flexibility of OLEDs and OPDs and their large scalability also introduce the potential for low-cost, disposable medical sensors (Lochner *et al.*, 2014).

Assembly Techniques

Small footprint components soldered on conventional print circuit board (PCB)

The most basic method of constructing an optoelectronic sensor is assembling commercially available packaged components onto a conventional printed circuit board (PCB). Most optoelectronic-based medical devices that have become a standard of care in clinics fall into this category in a hand-held form, allowing for point-of-care diagnostics. The pulse oximeter is the most successful device. The bilirubinometer, which screens for neonatal jaundice by measuring subcutaneous serum bilirubin concentration, is another good example. Their operating principles will be discussed in detail later in the application section. The smallest size of packaged LEDs and PDs currently available is in the range of ~ 1 mm. These components can be manually or automatically soldered onto the PCB, but are not feasible for applications requiring further miniaturization.

Die-level chip assembled on fabricated board with die-attached material

An alternative to using packaged optoelectronic components is assembling unpackaged die-level chips onto micro-fabricated circuit boards for the applications smaller than 1 mm (Fig. 3(a)). The direct soldering of chips is not feasible due to densely patterned interconnections (a few tens of micrometers between patterned lines) or no solder contact metal on the chips. For chips without solder metals on their anode/cathode electrode, a special metal alloy needs to be patterned in the process of fabricating the board. One bonding option utilizes a eutectic alloy like indium–gold (In–Au), which means the melting temperature of the alloy is lower than that of each metal. One advantage of eutectic bonding, in terms of assembly, is that the eutectic-bonded components are not re-melted by another heat process when even higher temperatures are applied (Fig. 4(a)). For the top electrode, wire-bonding is required to make a connection to the bottom interconnections. Some LED manufacturers supply directly attached LED chips of which cathode and anode are on the same side as solder metals. A simple reflow process bonds both electrodes, thus requiring no wire-bonding of the top electrode. Temperature profiles in each heating process need to be carefully controlled because non-optimized processing of temperatures can deteriorate the performance of the optoelectronic components as shown in Fig. 4(b). Even a bare die of an electronic chip can be directly attached to the flexible substrate board (Kim *et al.*, 2017).

Monolithic fabrication of integrated sensor

Maximum miniaturization of a sensor is achieved by monolithically fabricating an integrated sensor (O'Sullivan *et al.*, 2010). The source and detector are fabricated together using fabrication process flow and patterned by photolithography techniques from a single wafer (Fig. 3(b)). The VCSEL is a suitable semiconductor light source for this type of the integrated sensors because it emits light upward with higher intensity and requires simpler fabrication design compared to LED. The fabrication order requires careful consideration since the performance of each source and detector can be critically affected. For high-quality VCSEL lasing, the VCSEL layer needs to be grown before the PD. One advantage of monolithic fabrication is that source–detector separation can be sophisticatedly determined on the order of 100 μm . For a very short distance, the additional structure blocking a direct light pipelining from source to detector is necessary. An interference filter or absorbing-based filter can be used as a blocker and deposition techniques are compatible with the general semiconductor fabrication process of source/detector.

Data Processing Algorithm for Optoelectronic Sensor

Using physical principles on light–tissue interaction, there are various analysis algorithms depending on specific applications. This section will discuss two methods for analyzing diffuse reflectance: (1) the ratiometric method to measure a simple reflectance ratio

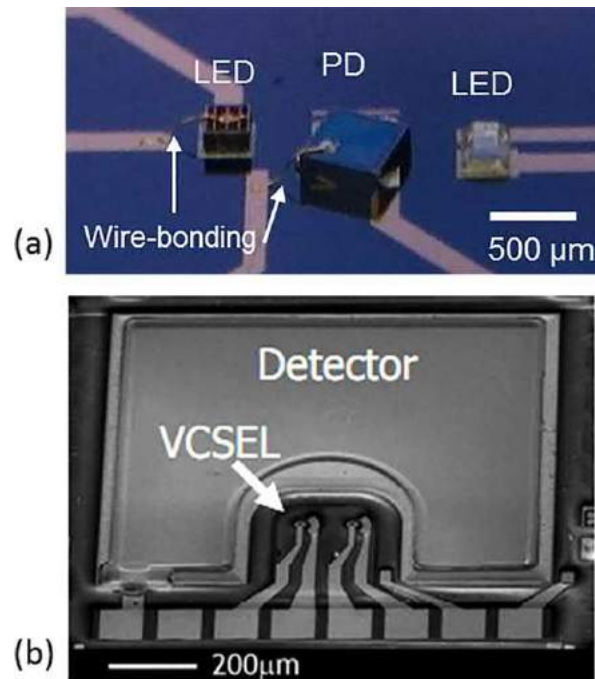


Fig. 3 (a) Die-level chips assembled on the fabricated micro-printed circuit board patterned with solder materials for a needle-compatible optoelectronic sensor assessing reflectance ratios. (b) Monolithically integrated fluorescence sensor. From O'Sullivan, T., Munro, E.A., Parashurama, N., *et al.*, 2010. Implantable semiconductor biosensor for continuous in vivo sensing of far-red fluorescent molecules. *Optics Express* 18, 12513–12525.

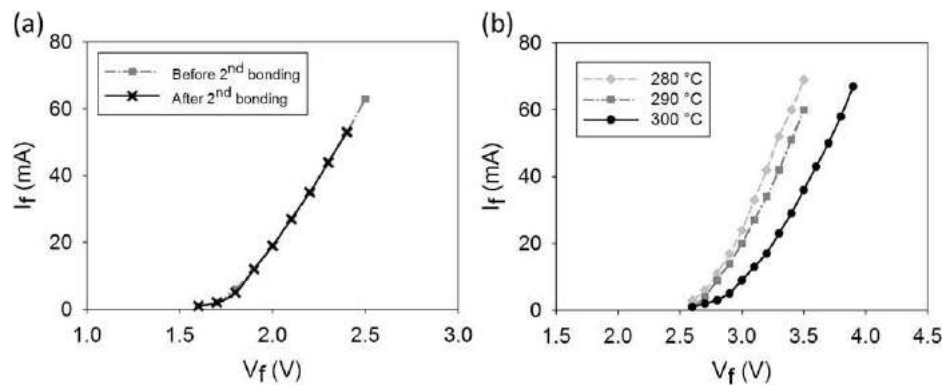


Fig. 4 (a) Forward voltage–forward current characteristics of the eutectic-bonded LED before and after the following heat process for 2nd LED bonding. (b) Forward voltage–forward current characteristics of the bonded die-level LEDs with different temperatures.

between two discrete wavelengths (one is a contrast and the other is a reference), and (2) the quasi-spectral method that uses multiple wavelengths to mimic a continuous spectrum.

Ratiometric analysis

Intensity-based sensing at single wavelength would be the simplest way to analyze diffuse reflectance but requires a rigorous calibration process to address intensity-induced distortions of such varying source intensity. The ratiometric method avoids this process by using one measurement as reference reflectance intensity. Calculation of the ratio in regard to the reference accounts for measurement errors. For example, in human pancreatic tissue assessment, the reflectance ratio, $R_{470 \text{ nm}}/R_{650 \text{ nm}}$ is suitable to distinguish malignant tissues (adenocarcinoma) from non-malignant tissues, including normal tissues and chronic pancreatitis. In a previous optical study on freshly excised human pancreatic tissues, the reflectance intensities at 450–470 nm differed significantly among these tissue types. This was attributed to different scattering properties depending on nuclei size, refractive index, and collagen contents. On the other hand, the reflectance intensities at 630–650 nm in three tissue types were very close, which served as a reference (Chandra *et al.*, 2007).

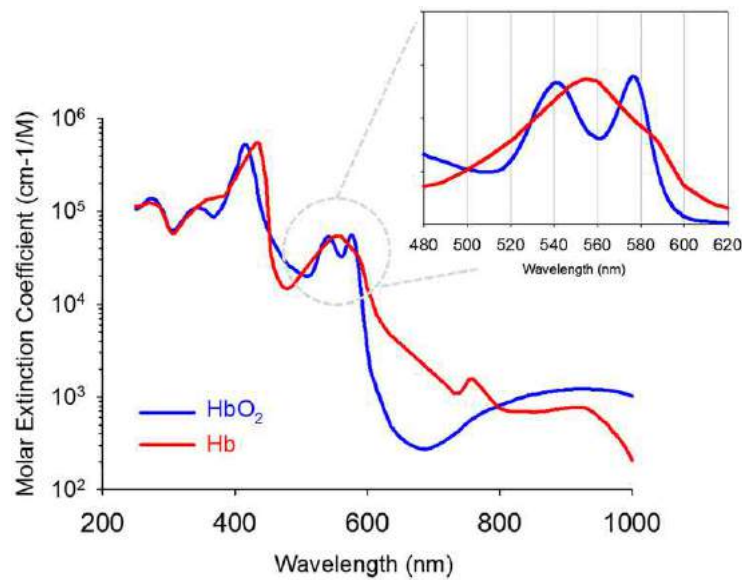


Fig. 5 Wavelength-dependent molar extinction coefficients of oxy- (HbO_2) and deoxy- (Hb) hemoglobin. Inset shows isosbestic points.

More generally, employing different wavelengths related to hemoglobin absorption has potential to assess total hemoglobin (Hb) concentration and tissue oxygenation (Hu *et al.*, 2013). Previous studies demonstrated that the ratiometric analysis based on isosbestic and non-isosbestic points has the capability to estimate tissue hemoglobin concentration and oxygenation. Isosbestic points have the same molar extinction for oxyhemoglobin and deoxyhemoglobin wavelengths, for example, 500, 529, 545, 570 and 584 nm (Fig. 5). Non-isosbestic points are the wavelengths showing the biggest difference in absorbance between oxy- and deoxy-hemoglobin, for example, 516, 539, 560, 577 and 593 nm. However, this study employed a full reflectance spectrum acquired by a fiber-based diffuse reflectance spectroscopy. For the optoelectronic sensors, two LEDs with wavelengths as close to isosbestic and non-isosbestic wavelengths as possible can be selected to enable rapid estimation of blood parameters. Rapid and non-invasive assessment of Hb concentration and oxygenation will have many clinical applications, including cancer detection, anemia detection, perfusion monitoring, etc. The main advantage of the ratiometric analysis is the rapid processing time that potentially enables real-time monitoring.

Quasi-spectral analysis

Unlike conventional optical spectroscopy that takes advantage of a continuous spectrum for evaluating tissues, the number of wavelengths available to optoelectronic sensors is limited due to technical difficulties in manufacturing a miniaturized light-dispersing component. However, multiple monochromatic sources with different wavelengths associated with one detector or, alternatively, a broadband source coupled with multispectral bandpass filter installed onto the detector, make the quasi-spectral analysis feasible. Even with a limited number of wavelengths, traditionally employed analysis algorithms utilizing a full spectrum can be applied to extract biophysically relevant parameters or optical properties directly from sensor measurements.

Analytical models have been employed to analyze diffuse reflectance and fluorescence spectrum obtained from human biological tissues (Wilson and Mycek, 2011). The mathematical equation is based on fundamental principles for light scattering and absorption at a given fiber optics probe geometry (Wilson *et al.*, 2010). A reflectance spectrum is calculated from many input variables including nuclei size and refractive index (Lee *et al.*, 2014). By fitting this mathematical model to the obtained reflectance spectrum, the best parameters generating the least error between the model and the experiment are extracted. The original model uses a wide range of wavelengths in the reflectance spectrum from 400 to 750 nm with 2 nm step for a total of 176 wavelengths. However, using only four discrete wavelengths (460, 520, 560, and 630 nm) can also produce results comparable to the model with 176 wavelengths. The reconstructed model fits from two models applied to the same spectrum obtained from human pancreatic tissues are comparable (Fig. 6(a)). The extracted parameters (scattering amplitude) are also similar in each pancreatic tissue type (Fig. 6(b)).

This preliminary comparison demonstrates the possibility of quasi-spectral sensing using the optoelectronic microprobe for noninvasive assessment of biophysically relevant parameters.

Applications

Clinical standard of care

The pulse oximeter is the most representative optoelectronic sensing device and has become a standard of care in clinics. Arterial oxygen saturation (SpO_2) and pulse rate are noninvasively measured (Chan *et al.*, 2013). Consisting of a small LED and PD, the

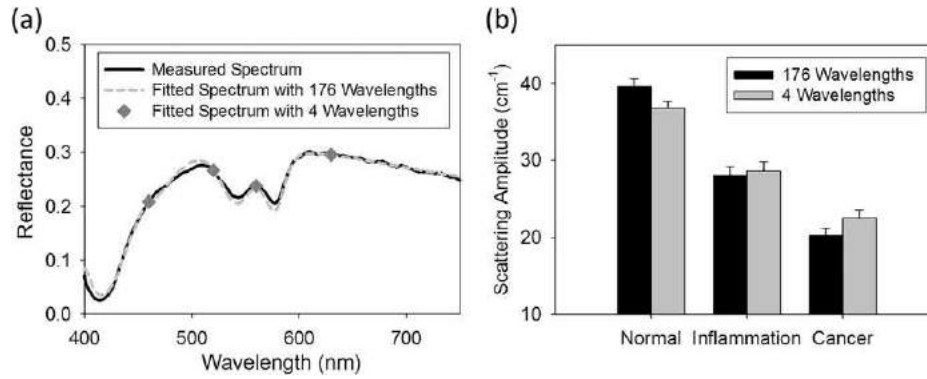


Fig. 6 (a) Reconstructed spectra by fitting procedure using 176 and 4 wavelengths on the measured spectrum. (b) Comparison of the extracted parameters (scattering amplitude) per different tissue types by inversion algorithm using 176 vs. 4 wavelengths.

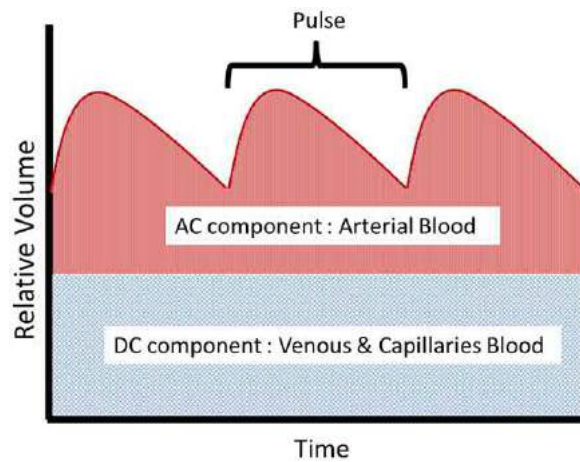


Fig. 7 Relative volume of the non-pulsatile (DC) and pulsatile (AC) tissue compartment during a cardiac cycle.

device can be placed on the tip of a patient's finger or ear lobe, allowing for continuous monitoring. Pulse oximetry takes advantage of red and near-infrared (IR) light that can penetrate tissue and be absorbed by hemoglobin. A red LED-emitting light at 660 nm and a near-IR LED with a wavelength of 940 nm are placed on one side of the device. As the light is transmitted through tissue, oxygenated hemoglobin absorbs IR light more than red light, and deoxygenated hemoglobin absorbs red light more than IR light. A PD located on the opposite side of the tissue detects transmitted light. The ratio of transmitted red and IR light is used to determine oxygen saturation of hemoglobin. Pulse rate can also be measured by the transmitted light. A consistent amount of light is absorbed by the tissue and steady venous flow, resulting in a direct current (DC) component of the signal. Pulsatile arterial blood flow creates an alternating current (AC) signal (Fig. 7). The modulation ratio (R) is calculated by

$$\text{Modulation Ratio } (R) = \frac{A_{\text{red-AC}}/A_{\text{red-DC}}}{A_{\text{IR-AC}}/A_{\text{IR-DC}}} \quad (2)$$

where A is the absorbance. This R is typically inversely proportional to SpO_2 and determines absolute SpO_2 values based on a calibration curve empirically obtained from controlled measurements.

The bilirubinometer is another exemplary device that is routinely used to screen neonatal jaundice in preterm infants. Jaundice is monitored by serum bilirubin levels, typically requiring neonates to endure multiple painful blood drawings. However, because cutaneous bilirubin levels are highly correlated with serum levels, transcutaneous bilirubinometers provide a noninvasive method of determining serum bilirubin levels by measuring the yellowness of the skin. Measurements are made using diffuse reflectance spectroscopy. Placing the device on the neonate's forehead or upper sternum activates a pressure-sensitive probe. A light-generating tube transmits light through subcutaneous tissue. Fiber-optic filaments send resultant scattered light to a spectrophotometric module. Light is then split into different spectra between 460 and 540 nm (different devices vary on the number of wavelengths). The absorbance at different wavelengths is compared to known absorption coefficients for all skin chromophores. The amount of yellowness in the skin is calculated as an indirect measurement of bilirubin values.

Cancer diagnostics

Not many optoelectronic sensors have been introduced to investigate human cancerous tissues because there is not a significant need for miniaturization in clinical diagnostics of human cancers. Over the last three decades, optical spectroscopy has demonstrated its potential for noninvasive detection of cancer in various human organs with a sub-cellular level sensitivity evaluating morphological, physiological biochemical changes upon tumor progression. However, only a few attempts to build optical sensing systems with optoelectronic components for either source or detector have been reported. Most of them were to implement sensor arrays for breast cancer imaging. Recently, the needs for miniaturized sensors have emerged to be compatible with an endoscopic channel (2–5 mm) or even with a hollow needle (< 1 mm) targeting some solid tumors such as pancreas, liver, and lung that are accessible only via a hollow needle. Some studies have delivered a fiber optics through this hollow-needle channel to optically interrogate lung and pancreas. However, the miniaturized optoelectronic sensor fitting in this needle could allow for arrangement of multiple source–detector units through a side-view window, which would increase an optical sampling volume (Lee *et al.*, 2017).

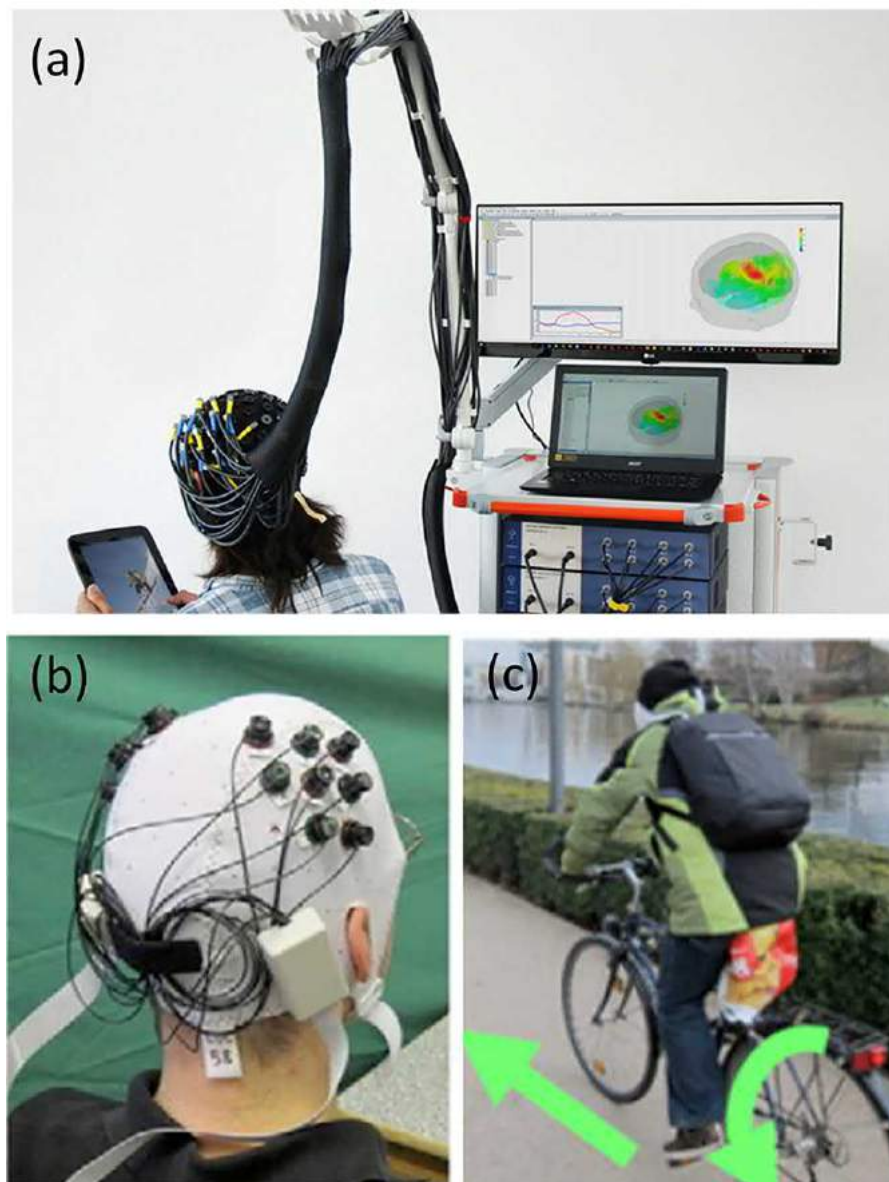


Fig. 8 (a) Commercialized fNIRS system tethered to a control system via fiber bundles, limiting subjects mobility. (b) Miniaturized fNIRS system only with electrical connections and compact electronic box attached to a head gear. (c) Subject wearing the miniaturized fNIRS system can move freely, demonstrating how this type of sensor can expand a kind of possible experiments. (b) and (c) from Piper, S.K., Krueger, A., Koch, S.P., *et al.*, 2014. A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage* 85, Part 1, 64–71.

Brain research

Since first demonstrated in 1993, functional near-infrared spectroscopy (fNIRS) has been increasingly employed to noninvasively investigate human brain activity by assessing changes in oxygen consumption. Using a deep penetration depth (a few millimeters) of near infrared light, fNIRS is capable of quantifying changes of oxy- and deoxy-hemoglobin concentration in specific brain regions upon external stimulations. The design with multiple pairs of source–detector covers a large area of human brain implementing 2D topological imaging or 3D tomographic imaging that can localize brain activity (Wilson *et al.*, 2016). However, this multichannel system inevitably involves heavy and cumbersome fiber bundles tethered to a remote control system. Most commercial systems use fiber bundles (Fig. 8(a)). The recent wireless fNIRS system based on a miniaturized optoelectronic source–detector pair and signal processing unit has been developed to enable evaluation of brain activity on subjects in a more natural environment (Fig. 8(b) and (c)) (Piper *et al.*, 2014).

Technical advances in optoelectronics/electronics have also opened a new era in functional neuroimaging of animals. Most current optical imaging systems are built with bulky optics and electronic hardware, requiring animals to be restrained and anesthetized, limiting types of experiments and affecting the baseline measurements. In contrast, a miniaturized head-mount optical imaging system enables awake and un-anesthetized animals to be freely moving. Current systems are capable of one photon and two photon imaging using external contrast agents and cerebral blood flow and optical intrinsic imaging via endogenous contrast (Yu *et al.*, 2015).

Summary

Recent technical advances in micro/nano-fabrication techniques, new materials and optoelectronic technologies have enabled fabrication of novel miniaturized optical sensors for in situ tissue diagnostics. In this article, we have discussed the basic construction of the sensors, core optoelectronic source and detector components, sensor fabrication techniques, data processing algorithms, and their clinical applications. The constant innovation in relevant technologies will accelerate development of these miniaturized optical sensors, which will ultimately lead to a paradigm shift in tissue optical diagnostics through noninvasive, objective, wearable, and wireless sensing.

References

- Chan, E.D., Chan, M.M., Chan, M.M., 2013. Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. *Respiratory Medicine* 107, 789–799.
- Chandra, M., Scheiman, J., Heidt, D., *et al.*, 2007. Probing pancreatic disease using tissue optical spectroscopy. *Journal of Biomedical Optics* 12, 060501.
- Hu, F., Vishwanath, K., Lo, J., *et al.*, 2013. Rapid determination of oxygen saturation and vascularity for cancer detection. *PLoS One* 8, e82977.
- Kim, J., Gutruf, P., Chiarelli, A.M., *et al.*, 2017. Miniaturized battery-free wireless systems for wearable pulse oximetry. *Advanced Functional Materials* 27, 1604373.
- Lee, S.Y., Lloyd, W.R., Wilson, R.H., *et al.*, 2014. Photon-tissue interaction model for quantitative assessment of biological tissues. *Proceeding of SPIE* 2014, 893528.
- Lee, S.Y., Na, K., Pakela, J.M., *et al.*, 2017. Compact system with handheld microfabricated optoelectronic probe for needle-based tissue sensing applications. *Proceeding of SPIE* 2017.100540Y-100540Y-7.
- Lloyd, W.R., Chen, L.-C., Wilson, R.H., Mycek, M.-A., 2013. Biophotonics: Clinical fluorescence spectroscopy and imaging. In: Maitland, D.J., Moore JR., J.E. (Eds.), *Biomedical Technology and Devices Handbook*, second ed. Boca Raton, FL: CRC Press-Taylor&Francis Group.
- Lloyd, W., Wilson, R.H., Chang, C.-W., Gillispie, G.D., Mycek, M.-A., 2010. Instrumentation to rapidly acquire fluorescence wavelength-time matrices of biological tissues. *Biomedical Optics Express* 1, 574–586.
- Lochner, C.M., Khan, Y., Pierre, A., Arias, A.C., 2014. All-organic optoelectronic sensor for pulse oximetry. *Nature Communications* 5, 5745.
- Mycek, M.-A., Pogue, B.W. (Eds.), 2003. *Handbook of Biomedical Fluorescence*. New York, NY: Marcel Dekker, Inc.
- O'Sullivan, T., Munro, E.A., Parashurama, N., *et al.*, 2010. Implantable semiconductor biosensor for continuous in vivo sensing of far-red fluorescent molecules. *Optics Express* 18, 12513–12525.
- Piper, S.K., Krueger, A., Koch, S.P., *et al.*, 2014. A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage* 85 (Part 1), 64–71.
- Pitts, J.D., Mycek, M.-A., 2001. Design and development of a rapid acquisition laser-based fluorometer with simultaneous spectral and temporal resolution. *Review of Scientific Instruments* 72, 3061–3072.
- Wilson, R.H., Chandra, M., Chen, L.C., *et al.*, 2010. Photon-tissue interaction model enables quantitative optical analysis of human pancreatic tissues. *Optics Express* 18, 21612–21621.
- Wilson, R.H., Mycek, M.-A., 2011. Models of light propagation in human tissue applied to cancer diagnostics. *Technology in Cancer Research and Treatment* 10, 121–134.
- Wilson, R.H., Vishwanath, K., Mycek, M.-A., 2016. Optical methods for quantitative and label-free sensing in living human tissues: Principles, techniques, and applications. *Advances in Physics: X* 1, 523–543.
- Yu, H., Senarathna, J., Tyler, B.M., Thakor, N.V., Pathak, A.P., 2015. Miniaturized optical neuroimaging in unrestrained animals. *NeuroImage* 113, 397–406.

In Situ Optical Tissue Diagnostics/Laser Speckle-Based Spectroscopy Techniques in Biomedicine

Karthik Vishwanath and Sara Zanfardino, Miami University, Oxford, OH, United States

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

$\vec{r}$ Vector location of a point of interest in space
 τ Correlation time
 $\Phi(\vec{r})$ Photon fluence rate
 μ_a Absorption coefficient of medium

μ_s Scattering coefficient of medium
 D Optical diffusion coefficient of medium
 $S(\vec{r})$ Input source term
 g Anisotropy coefficient for single scattering of medium

Glossary

Absorption coefficient A wavelength-dependent medium property characterizing the probability of photon absorption in the medium.

Anisotropy coefficient A wavelength-dependent medium property characterizing the mean cosine of the angle between the directions of photon incidence and scattering in the medium.

Coherence The correlation between two (electromagnetic) waves either at different points in space (spatial coherence) or at two different points in time (temporal coherence).

Photon fluence rate The number of photons passing through a specified area per unit of time.

Photon radiance The net photon power passing through a specified area, along a specified direction, per unit of time.

Scattering coefficient A wavelength-dependent medium property characterizing the probability of photon scattering in the medium.

Speckle Small “spots” of bright and dark spots formed when a coherent source is incident on any surface due to constructive and destructive interference at the detector plane.

Temporal autocorrelation function The correlation of a time-dependent function, with itself across the full time-period measured of the original function.

Turbid medium Any medium appears “cloudy” or “murky” due multiple light scattering.

Introduction

Optical spectroscopic methods and techniques have been used for sensing of turbid media such as biological tissues and is an active area of research in the biomedical sciences. The basic prescription in most applications of these methods is to non-invasively derive optical properties of media using experimental measurements in combination with mathematical modeling and analysis to yield quantitative analogues of physiological markers related to human health and disease. Clinical applications over the last two decades have reported using optical spectroscopy for several applications including detection (including surgical margin identifications) of cancer, detection of atherosclerotic plaques, monitoring of burns, wounds and their healing, measuring cerebral functions, and imaging of retinal and dental health.

Optical techniques used for biosensing may be classified as two broad categories – static or dynamic – depending on whether the experimental techniques and methods used are inherently time-integrated, or time-resolved. A large fraction of optical methods that are used to characterize tissue optical properties including diffuse optical spectroscopy and tomography fall under the realm of static techniques – wherein one or more light sources are used to illuminate a medium of interest and a set of measurements (usually at several spatial locations) which are translated to produce a static spatial map of optical properties. It is to be noted that repeated measurements of such static optical techniques as defined here may still be used to study tissue (or biological) dynamics, at time-scales of the acquisition rates of the measurements. Dynamic optical spectroscopic techniques on the other hand make a set of measurements (again at one or more spatial locations) by measuring changes or fluctuations in light intensities (or a shift in the frequency spectrum) across several decades in time and seek to relate these fluctuations to a dynamical varying physical quantity such as motion of particles in the medium.

This article describes a dynamic optical technique – diffuse correlation spectroscopy (DCS) – which combines the theoretical formalism of incoherent photon propagation in a strongly scattering medium, to model experimental measurements of dynamic speckle intensity fluctuations (correlations) measured from the same medium. In DCS, measured speckle intensity fluctuations from an incident monochromatic source is used to compute temporal intensity autocorrelations which in turn are used to extract information about movement of scatterers within the volume of tissue probed by the diffuse optical waves transporting the incident light field to the detector. A closely related optical technique to DCS is Laser Doppler flowmetry which detects the scattered, frequency-shifted spectrum of a monochromatic (coherent) light source after passing through a scattering medium containing moving scatterers. Although these two techniques are closely connected analytically and theoretically as Fourier

transforms of each other, their experimental implementations and analysis are quite distinct. This article will only focus on the description, analysis and applications of DCS for tissue sensing.

Historical Overview and Background

When coherent light is scattered by a material, temporal measurements of the scattered field intensity exhibits fluctuations. Theoretical calculations of such fluctuations of light intensity when scattered by molecules (or other particles) have been studied since beginning of 20th century and include contributions from Einstein, Raleigh, Smoluchowski, and Debye. With the invention of the laser and fast detector-electronics, dynamic light scattering (DLS) became an accessible experimental technique to measure and study a wide variety of dynamic phenomena in complex media such as molecular rotation or diffusion in complex fluids, colloidal particle compositions and biological systems. Historically, DLS theory was developed first for weakly scattering systems (single scattering limit). However, experimentally, the methods of DLS are as applicable to multiply scattering systems as they were to weakly scattering media. DLS was soon applied to study of particle dynamics in various multiply scattering materials and the theoretical modeling of such experimental measurements came to be called diffusing wave spectroscopy (DWS). Experimentally, DCS uses the same experimental theme that is employed in DLS (or DWS) – but instead provides an elegant theoretical encapsulation to analyze the experimental measurements using the now well-established framework of diffusive optical transport in turbid media as sketched in Fig. 1.

Experimentally, all of these techniques employ a long-coherence length laser to illuminate the sample of interest and collect the scattered light intensity from a small volume. Fluctuations in the measured intensity incident on a photodetector and photon-counting system, $I(\vec{r}, t)$ can be obtained. The experimentally calculated intensity unnormalized and normalized correlation functions of these fluctuations are denoted respectively using their standard notations as:

$$G_2(\vec{r}, \tau) = \langle I(\vec{r}, 0) \cdot I(\vec{r}, \tau) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T I(\vec{r}, t) I(\vec{r}, t + \tau) dt \quad (1)$$

$$g_2(\vec{r}, \tau) = \frac{\langle I(\vec{r}, 0) \cdot I(\vec{r}, \tau) \rangle}{\langle I(\vec{r}, 0) \rangle^2} \quad (2)$$

The integral in Eq. (1) defines the temporal correlation function of $G_2(r, \tau)$ for an infinitely long time-signal $I(\vec{r}, t)$ while the angular brackets indicate ensemble averages. Although it is usually assumed that the system under observation is ergodic and thus ensemble averages are time averages this may not always be true – we discuss this in the section under Experimental Considerations below. In practice, the temporal correlation is usually calculated using a correlator board (or computer) to numerically evaluate the integral in Eq. (1). The smaller dt is, the more accurately measurements can be expected to match theoretical predictions, with the longest possible correlation time being the total duration of continuous data acquisition. Photodetectors with sampling rates in the 10–100 MHz range are typically used to collect DCS data for 1–10 s for in vivo tissue sensing.

Theoretically, the (detected) intensity is related to the scattered complex electric field from a monochromatic coherent source as $I(\vec{r}, t) = E(\vec{r}, t)E^*(\vec{r}, t) = |E(\vec{r}, t)|^2$ and the corresponding correlations of the detected unnormalized and normalized electric fields are defined as per convention by:

$$G_1(\vec{r}, \tau) = \langle E(\vec{r}, 0)E^*(\vec{r}, \tau) \rangle \quad (3)$$

$$g_1(\vec{r}, \tau) = \frac{\langle E(\vec{r}, 0)E^*(\vec{r}, \tau) \rangle}{\langle E(\vec{r}, 0) \rangle} \quad (4)$$

From the above definitions some properties of the field correlation functions can be computed. $G_1(\vec{r}, 0) = \langle |E(\vec{r}, t)|^2 \rangle = \langle I(\vec{r}, 0) \rangle$, giving $g_1(\vec{r}, 0) = 1$. Also, for long correlation times $G_1(\vec{r}, \tau) = \langle E(\vec{r}, t) \rangle^2 = 0$, giving $g_1(\vec{r}, \tau) = 0$. Given that it is the field fluctuations that give rise to the measured intensity fluctuations, theoretical methods are

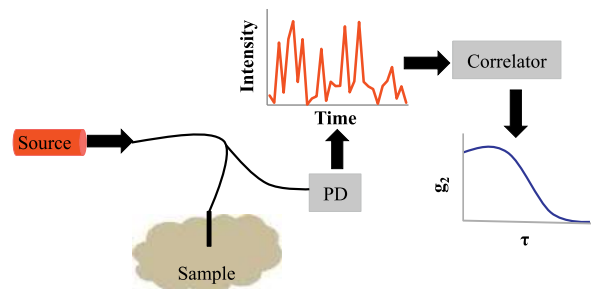


Fig. 1 Schematic representation of the DCS setup and its components. The illumination is provided by a laser source coupled to the sample using fiber optical probes. The detected signal is measured using a photodetector (PD) that can provide continuous (temporal) measurements. The monitored intensity fluctuations are recorded using a correlator device to compute the normalized intensity correlation g_2 .

used to generate models that predict behavior of $G_1(\vec{r}, \tau)$ and $g_1(\vec{r}, \tau)$. If the temporal fluctuations in $E(\vec{r}, t)$ are random and assumed to arise due to uncorrelated motions of several randomly (diffusing or moving) particles, then these fluctuations in $E(\vec{r}, t)$ will be normally distributed. Under this condition, $G_1(\vec{r}, \tau)$ and $G_2(\vec{r}, \tau)$ (and their normalized counterparts) are theoretically related through the Siegert relations:

$$G_2(\vec{r}, \tau) = \langle I(\vec{r}, 0) \rangle^2 + |G_1(\vec{r}, \tau)|^2 \quad (5)$$

$$g_2(\vec{r}, \tau) = 1 + |g_1(\vec{r}, \tau)|^2 \quad (6)$$

In DCS the functional form for $G_1(\vec{r}, \tau)$ is hypothesized to follow the same diffusion transport equation governing photon propagation in a multiply scattering medium, while $g_1(\vec{r}, \tau)$ is modeled using the random, independent single scattering model using the formalism of DLS.

Diffusive Photon Correlation Transport

Photon propagation in a turbid medium can be accurately modeled as a random walk of neutral particles in the medium using optical transport coefficients of the medium – typically specified as absorption, scattering and anisotropy. The radiative transport equation provides a theoretical analog for describing propagation of photon radiance within such media. The mathematical form of the transport equation prevents its use practice since it remains intractable for appropriately modeling experimental (boundary) conditions. The transport equation however reduces to a diffusion type equation if two conditions are met (a) if the medium's absorption is small in comparison to its scattering and (b) if the detected light has been scattered sufficiently and can be thus be considered emanating from an isotropic source. Both these conditions can largely experimentally satisfied – the first by selecting light with appropriate wavelengths, since the absorption and scattering properties of tissue are strongly wavelength dependent (for e.g. the near-infrared (NIR) wavelengths for tissue spectroscopy). The second condition may be satisfied by keeping the source-detector separations large in comparison to the mean scattering transport length of light in the medium of interest.

Under steady-state conditions (i.e. the source is CW or when the detected signal is time-integrated for longer than a few ns, in general) the diffusion equation for a semi-infinite homogenous medium with uniform optical properties the diffusion equation is written as:

$$[D\nabla^2 - \mu_a]\Phi(\vec{r}) = S(\vec{r}) \quad (7)$$

Here, $\Phi(\vec{r})$ represents the photon fluence rate (Wm^{-2}) as the net light energy at all points of interest in the medium, $S(\vec{r})$ describes the input source term, and D the medium's diffusion coefficient given by $D = 3[\mu_a + \mu_s(1 - g)]$, where μ_a is the medium's absorption coefficient, μ_s its scattering coefficient and g the single-scattering anisotropy coefficient. Under the formalism of DCS, the unnormalized field correlation $G_1(\vec{r}, \tau)$ is propagated using the diffusion equation as:

$$\left[D\nabla^2 - \mu_a - \frac{1}{3}\mu_s(1 - g)k_0^2\langle \Delta r^2(\tau) \rangle \right] G_1(\vec{r}, \tau) = S(\vec{r}) \quad (8)$$

The $\langle \Delta r^2(\tau) \rangle$ in Eq. (8) is the only additional term for the transport of the field correlation in a homogenous turbid medium and represents the mean-square displacement of the diffusing (or moving) particles, which add to the dynamics of the decay in the field correlation. Since closed form expressions of $\Phi(\vec{r})$ for Eq. (7) are readily available for several geometries, they can be used to calculate $G_1(\vec{r}, \tau)$ for measurements made in those geometries by replacing μ_a with $\mu_a - \frac{1}{3}\mu_s(1 - g)k_0^2\langle \Delta r^2(\tau) \rangle$.

Experimental Considerations

Experimentally, one has to make choices about the illuminating source, the delivery and collection optics to couple to the medium, and the detection system (see Fig. 1). Fig. 2 shows representative examples of experimental measurements obtained using a fiber-based DCS system while for different light sources, different detection fiber sizes and different types of scattering media.

Light Sources

Given that the theoretical formulation for modeling measurements assumes a monochromatic plane wave, DCS sources are usually single mode, stabilized, long-coherence length lasers operating in between 600–800 nm. In order for the experimental data to contain reliable dynamic information about the motion of scatterers, the temporal coherence length must be much longer the average path length of photons diffusing from the source and the detector placed on the tissue. Most applications reported with DCS use sources with coherence lengths longer than several meters.

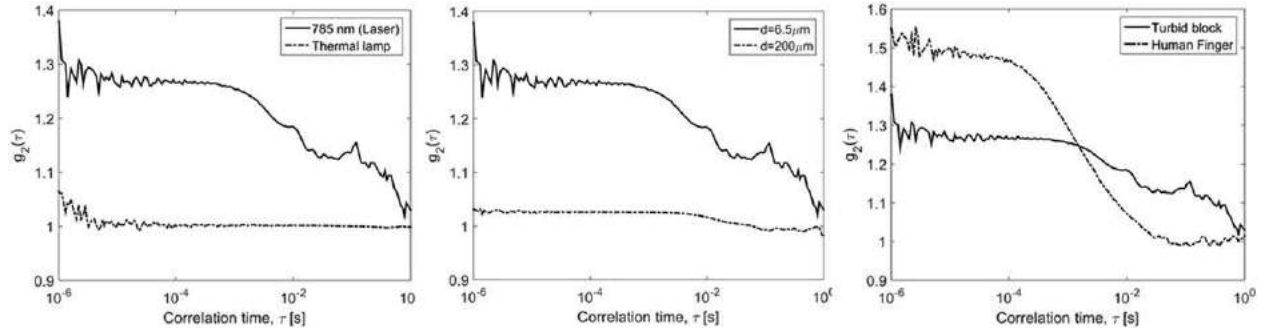


Fig. 2 Measured representative intensity correlation curves (g_2) depicting data obtained for changing incident light source (left), changing detection fiber radius (middle) and different media (right). In all the plots, the solid line represents the autocorrelation detected using a fiber probe (source fiber radius = 200 μm , detector fiber radius = 6.5 μm and center-center distance of 1.5 mm) in reflection geometry on a turbid solid block. The dashed line in the left shows the correlation when the source was changed to a thermal halogen lamp. In the middle figure, the second fiber probe was used that had identical properties as the first probe but with detector radius of 200 μm while using the laser light for illumination. The dashed line in the right figure shows data from the first probe in contact with a human finger under laser illumination. These data indicate how both the intercept on the y-axis as well as the decay curves change across different conditions.

Coupling Optics

Although it is possible to make DCS devices capable of operating in a non-contact mode (i.e. coupling light to and from the medium without physical contact), the most practical method of using DCS for biomedical applications has been to couple light from the source to a medium (tissue site) using a source-fiber and obtaining diffusely scattered measurements using one or more detection fibers. The detection fiber is typically a low-mode fiber (ideally capable of monitoring fluctuations of a single speckle) while the source-fiber is a multi-mode fiber. The common end of this probe sets the source-detector geometry used to probe the sample volume and usually is larger than 5–10 mm for most DCS systems to stay within validity limits of diffusion theory.

Detection Systems

The ability to extract temporal correlations relies which not only requires sampling a fluctuating intensity signal at sampling rates of 10–100 MHz but also to store and process these signals to obtain the required intensity autocorrelation function. Traditional photon detection systems used an amplified photomultiplier tube (PMT) that with high enough response rates, while most recent systems employ detectors are small-area single photon avalanche photodiodes (APD). The digitized output sampled off the photodetector is to a digital correlator – usually a dedicated hardware board that computes required time correlation functions.

Modeling Measured Intensity Correlations

As described above, the experimental observations yield $g_2(\vec{r}, \tau)$ that is the intensity correlation for a measured at the detector situated at location $\vec{r}$ relative to the source and the medium. Thus, the analytical expressions for $G_1(\vec{r}, \tau)$ are converted to calculate $g_1(\vec{r}, \tau)$ which in turn is related to the experimental measurement of $g_2(\vec{r}, \tau)$ as:

$$g_2(\vec{r}, \tau) = 1 + \beta |g_1(\vec{r}, \tau)|^2 \quad (9)$$

Here, β is an experimental factor that is lesser than unity and depends on specific experimental configuration details such as the coherence length, coupled modes, dark noise and detection aperture size and thus is usually obtained as a free parameter during the fitting process. Besides β , experimentally measured data is fit to a specified functional formulation to extract parameters contributing to $\langle \Delta r^2(\tau) \rangle$. For example two commonly used models in tissue spectroscopy include diffusive flow where $\langle \Delta r^2(\tau) \rangle = 6D_B\tau$ where D_B is a diffusion coefficient, or random flow where $\langle \Delta r^2(\tau) \rangle = V^2\tau^2$ and V^2 is the second moment of the moving particle's velocity distribution. It is important to note that analytical fitting process is possible only by having estimates for the optical absorption and scattering coefficients of the medium, at the wavelength of the incident source.

Applications to In Vivo Tissue Sensing

Fiber-based DCS measurements have been obtained across a variety of clinical settings to address problems in monitoring progress of treatments, for monitoring cerebral blood flow during ischemia and for assessment of wound healing. Ultimately, the largest factor of contrast in biological tissues for the decorrelation of the measured intensity autocorrelations is attributed to the motion

of red blood cells in the vascular system. Thus, nearly any health application that requires monitoring of blood flow and its relative changes across perturbations could potentially use DCS a simple, non-invasive clinical adjunct.

DCS has been explored to investigate tissue perfusion since angiogenesis (i.e. the growth of new blood vessels) is now believed to be a marker of tumor growth and progression. Further, since treatments of tumors including chemotherapy and photodynamic therapy would be expected to impact blood flow and its dynamics during such treatments, DCS measurements have been applied to investigate these changes longitudinally in cancers of the prostate, breast and of the head and neck. In many cases DCS has been shown to provide an early indicator of treatment response (and failure). Given that traditional therapy monitoring techniques that include methods such as PET, MRI or CT, which many times are not accessible or amenable for use, DCS may provide a convenient, clinically-adoptable alternative.

DCS has also been used in applications to monitor cerebral blood flow – such monitoring studies were focused particularly on investigating perfusion rates and changes in the brain during ischemic strokes. An ischemic stroke the most common type of hemorrhagic stroke where a blood clot in the body constricts or depletes blood supply to certain regions of the brain. Since such a blockage would result in markedly different perfusion rates, DCS was an appropriate candidate for continuous and non-invasive blood in patients.

Other clinical applications of DCS have explored monitoring patient recovery from a traumatic wound injury, monitoring the blood supply to maxillofacial autographs during early stages of transplant, and monitoring the blood supply to skeletal tissues and muscles during certain high-intensity activities such as exercise. Several of these studies that employ DCS for measurement of tissue perfusion have been validated against clinically used “gold-standards” for flow-monitoring such as CT and MRI.

Preliminary and current data in clinical trials is promising for the field of non-invasive biomedical optical sensing. With an increasing number of studies showing successful applications of dynamic light scattering spectroscopy for in vivo and real-time monitoring of blood flow in tissues, it is likely that DCS-based devices will become more widespread. These application areas will benefit if DCS devices become commercially available as out-of-the box instruments and could become a routine and important tool used for sensing in vascular function or dysfunction in humans.

Further Reading

- Ackerson, B.J., Dougherty, R.L., Reguigui, N.M., Nobbmann, U., 1992. Correlation transfer – Application of radiative-transfer solution methods to photon-correlation problems. *Journal of Thermophysics and Heat Transfer* 6, 577–588.
- Berne, B.J., Pecora, R., 1990. In dynamic light scattering with applications to chemistry. New York: Wiley-Interscience.
- Boas, D.A., Yodh, A.G., 1997. Spatially varying dynamical properties of turbid media probed with diffusing temporal light correlation. *Journal of the Optical Society of America* 14, 192–215.
- Bonner, R., Nossal, R., 1981. Model for laser Doppler measurements of blood flow in tissue. *Applied Optics* 20, 2097–2107.
- Briers, J.D., 2001. Laser Doppler, speckle and related techniques for blood perfusion mapping and imaging. *Physiological Measurement* 22, R35–R66.
- Durduran, T., Choe, R., Baker, W.B., Yodh, A.G., 2010. Diffuse optics for tissue monitoring and tomography. *Reports on Progress in Physics* 73 (7), doi:10.1088/0034-4885/73/7/076701.
- Farzam, P., Zirak, P., Binzoni, T., Durduran, T., 2013. Pulsatile and steady-state hemodynamics of the human patella bone by diffuse optical spectroscopy. *Physiological Measurement* 34, 839–857.
- Kim, M.N., Durduran, T., Frangos, S., *et al.*, 2010. Noninvasive measurement of cerebral blood flow and blood oxygenation using near-infrared and diffuse correlation spectroscopies in critically brain-injured adults. *Neurocritical Care* 12, 173–180.
- Pine, D.J., Weitz, D.A., Chaikin, P.M., Herbolzheimer, E., 1988. Diffusing wave spectroscopy. *Physical Review Letters* 60, 1134–1137.
- Yu, G., 2012. Near-infrared diffuse correlation spectroscopy in cancer diagnosis and therapy monitoring. *Journal of Biomedical Optics* 17, 010901.
- Wilson, R.H., Vishwanath, K., Mycek, M.-A., 2016. Optical methods for quantitative and label-free sensing in living human tissues: Principles, techniques, and applications. *Advances in Physics*. 1–21.

Photodynamic Therapy and Photobiomodulation: Can All Diseases be Treated with Light?

Michael R Hamblin, Massachusetts General Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States; and Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction. The Janus Face of Photomedicine

Janus (or Ianus) was the Roman god of beginnings, gates, transitions, time, doorways, passages, and endings. He is usually depicted as having two faces looking in opposite directions, since he looks to the future and also to the past ([Fig. 1](#)). It is thought that the month of January was named for Janus (Ianuarius) because he presided over the ending of one year and the beginning of the next. Janus also presided over the beginning and ending of conflict, and hence over war and peace. The doors of his temple were open in time of war, and closed to mark the onset of peace. The most famous temple to Janus in Rome is called the Ianus Geminus, or “Twin Janus.” When its doors were open, neighboring cities knew that Rome was at war. As a god of transitions, he had functions pertaining to birth and to journeys and exchange.

The two-faced nature of the Roman God, Janus has become inextricably linked with many aspects of biology and medicine. The concept of a process, mechanism, treatment, or drug having both good and bad effects is so widespread and pervasive, that the analogy of Janus has been seized upon for a variety of reasons. Perhaps the best-known application of the Janus name in biology, is Janus kinase (JAK), which represents a family of intracellular, non-receptor tyrosine kinases that transduce cytokine-mediated signals via the JAK-STAT pathway. They were initially named “just another kinase” 1 and 2 (since they were just two of a large number of discoveries in a PCR-based screen of kinases ([Wilks, 1989](#))), but were ultimately published as “Janus kinase”. The name was taken from the two-faced Roman god of beginnings and endings, Janus, because the JAKs possess two near-identical phosphate-transfer domains. One domain mediates the kinase activity, while the other negatively regulates the kinase activity of the first domain.

We intend to show in this chapter that light therapy has many characteristics that can accurately be described as a “two-faced” or “Janus effect”. Depending on whether or not light is combined with a non-toxic photosensitizing dye (photodynamic therapy), and on the overall dose of light (energy density and power density), light is able to kill just about anything that is living such as cancers, microorganisms, blood vessels, parasites, pests, unwanted tissues etc, etc. On the opposite side of the coin (or perhaps on the other side of Janus’ face), light of the correct wavelength (photobiomodulation) and at the right dose (generally accepted to be “low” as in “low level laser therapy”) can have exactly the opposite effect, being able to heal, regenerate, protect, revitalize and restore any kind of dead, damaged, stressed, dying, degenerating cells, tissue, or organ system. Between the bad destroying face of photodynamic therapy, and the good healing face of photobiomodulation, it may well be concluded that indeed it can be said “all diseases can be treated with light”.

Photodynamic Therapy – The Killer

Photodynamic therapy (PDT) is the combination of non-toxic dyes and harmless visible light that combine together in the presence of oxygen to produce highly toxic reactive oxygen species (ROS) that can kill unwanted cells like cancer cells and pathogenic microorganisms ([Hamblin and Mroz, 2008](#)). PDT was discovered over one hundred years ago in 1900 when Oscar Raab (a medical student working with Herman von Tappeiner) on the toxicity of acridine dyes to protozoa. Raab realized that the



Fig. 1 Different depictions of the two-faced Roman god Janus.

variable results he obtained in his experiments using acridine orange and paramecia, could be attributed to variations in the level of ambient sunlight on the days when the experiments were performed. von Tappeiner and a dermatologist named Jesionek published the first report (just 3 years after Raab's initial paper) describing the use of 1% eosin solution painted on the skin, with white light illumination to successfully treat patients with basal cell carcinoma of the skin. In the decade that followed, von Tappeiner's group showed that this "photodynamic" effect required the simultaneous presence of a photosensitizer, light, and oxygen to mediate cellular toxicity. Freiderich Meyer-Betz then performed the first human trial of hematoporphyrin-mediated PDT by injecting himself with 200 mg of the drug and standing in sunlight.

It was however not until 1978 that Thomas Dougherty and his colleagues at Roswell Park Cancer Institute (RPCI) reported the first large-scale trial of PDT in human patients with cancer (Cengel *et al.*, 2016). In total, 113 primary and metastatic skin tumors were treated in 25 patients with diagnoses ranging from malignant melanoma to cutaneous T-cell lymphoma to metastatic breast and endometrial cancers. Many of those tumors were recurrent or had progressed after multiple cytotoxic chemotherapy agents and had failed local ionizing radiotherapy.

Since those early days PDT has continued to progress and has significantly widened the range of diseases and disorders that can be treated (Hamblin and Mroz, 2008), in addition to cancer (Agostinis *et al.*, 2011). The ROS produced during PDT can rapidly and completely kill any type of pathogenic microorganism, whether it be different classes of bacteria, viruses, yeasts, filamentous fungi, parasites (such as *Leishmania*), and also remove unwanted tissue such as atherosclerotic plaque, undesirable blood vessels, and scars.

Mechanisms of PDT

Fig. 2 shows the mechanisms that operate during PDT in a so-called Jablonski diagram. The ground state PS has two electrons with opposite spins (singlet state) in the lowest energy molecular orbital. Following the absorption of light (photons), one of these electrons is boosted into a higher-energy orbital but retains its spin (first excited singlet state). This is a short-lived (nanoseconds) species and can lose its energy by emitting light (fluorescence) or by internal conversion into heat. The fact that most PS are fluorescent has led to the development of sensitive assays to quantify the amount of PS in cells or tissues, and allows in vivo fluorescence imaging in living animals or patients to measure the pharmacokinetics and distribution of the PS. The excited singlet state PS may also undergo the process known as intersystem crossing whereby the spin of the excited electron inverts to form the relatively long-lived (microseconds to milliseconds) excited triplet-state that has both electron spins parallel. The long life of the triplet state is attributed to the relative rarity of other triplet molecules with which it can interact.

The PS excited triplet can undergo three broad kinds of reactions that are usually known as type I and type II (see Jablonski diagram in Fig. 2). Firstly, in a type I reaction, the triplet PS can gain an electron from a neighboring reducing agent. In cells these reducing agents are commonly either nicotinamide adenine dinucleotide (NADH) or NAD-phosphate, reduced form (NADPH). The PS is now a radical anion bearing an additional unpaired electron ($PS^{\bullet-}$). Alternatively two triplet PS molecules can react together involving intermolecular electron transfer to produce a pair consisting of a radical cation and a radical anion. The PS radical anions may further react with oxygen to carry out an electron transfer to produce reactive oxygen species, in particular superoxide anion. The damaging effects of superoxide are relatively mild, however it can cause much more oxidative damage when it reacts with itself to produce hydrogen peroxide and oxygen, in the process called "dismutation". Hydrogen peroxide, can add to an organic (carbon-containing) substrate and result in an oxidizing "chain reaction". This is common in the oxidative damage of fatty acids and other lipids. Hydrogen peroxide is important in the production of the highly reactive hydroxyl radical ($HO^{\bullet}$), by a second one-electron reduction mediated by $PS^{\bullet-}$. Superoxide will additionally react with nitric oxide ($NO^{\bullet}$) (also a radical) to produce peroxynitrite ($OONO^-$), another highly reactive molecule that can oxidize many functional groups and can also nitrate proteins on tyrosine residues.

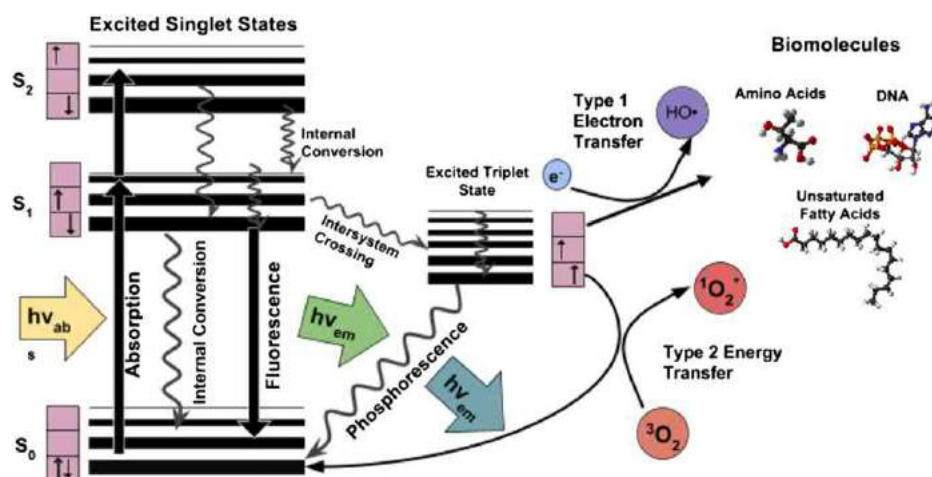


Fig. 2 Jablonski diagram showing photochemical reactions that occur during PDT.

In a type II reaction, the triplet PS can transfer its energy directly to molecular oxygen (itself a triplet in the ground state), to form excited-state singlet oxygen. Type II processes are thought to best preserve the PS molecular structure in a repeatedly photoactivatable state, and in some circumstances a single PS molecule can generate 10,000 molecules of singlet oxygen. The PS can in some circumstances also react with the singlet oxygen it produces in a process known as oxygen-dependent photobleaching.

Singlet oxygen generated during type II photochemical reactions, is believed to be the most important molecule responsible for PDT-induced cellular damage (Henderson and Miller, 1986; Weishaupt *et al.*, 1976). However, because of the high reactivity and short half-life of singlet oxygen only molecules and structures that are proximal to the area of its production (areas of PS localization) are directly affected by PDT. The half-life of singlet oxygen in biological systems is < 40 ns, and, therefore, the radius of the action of singlet oxygen is of the order of 20 nm (Moan and Berg, 1991). In singlet oxygen the spin restriction that reduces reaction rates of triplet oxygen with non-radicals is excluded. Consequently, singlet oxygen has an empty low-energy valence orbital that makes it a potent oxidizing/oxygenating agent compared to its ground-state counterpart. Singlet oxygen interacts with molecules via two essential mechanisms: physical interactions with quenchers or chemical reactions with biomolecules and scavengers. In the first phenomenon singlet oxygen transfers its excitation energy to an acceptor molecule (quencher) promoting it to an excited state while oxygen returns to ground-state. The excited quencher can be cleaved forming new products or react with biomolecules and cause damages. Antioxidant molecules, such as carotenoids, can alternatively dissipate the energy quenched from singlet oxygen in the form of heat, preventing any sort of oxidative damage. Even though, most carotenoids are very hydrophobic substances and exclusively accumulate in cell membranes (Kearns, 1971).

Singlet oxygen mainly reacts with molecules containing thiol groups or double bonds. Hence, it presents more restricted reaction patterns in relation to free radicals produced by type I reactions. Carbon and nitrogen atoms form double bonds sharing two electrons with opposite spin orientation. Note that ground-state oxygen cannot receive those two paired electrons because both low-energy orbitals are already occupied by one unpaired electron. Since singlet oxygen has an empty low-energy orbital, it has no restrictions to react with double bonds (Bland, 1976; Kearns, 1971).

Type I photodynamic reactions start off via donation of one electron from the excited PS ($^3\text{PS}^*$) to O_2 producing superoxide radical ($\text{O}_2^{\bullet-}$) by a 1-electron reduction, and then going on to form hydrogen peroxide by a 2-electron reduction of O_2 that can yet produce more radicals. Radical species are typically paramagnetic because they carry one or more unpaired electrons in valence shell orbitals. This feature makes radicals particularly reactive. They are susceptible to react with several possible biomolecules passing the unpaired electron to another molecule also turning it into radical species or releasing another free radical. Therefore, a single radical formed can initiate chain reactions that lead to damage of several biomolecules. Radical reactions are rather unspecific and can attack lipids, sugars, proteins and nucleic acids forming radical moieties within these biomolecules. Such radical moieties then allow biomolecules to react with each other forming cross-links between them (e.g. lipid-lipid, lipid-protein, protein-DNA, etc.). Either radical addition to biomolecules through hydroxylation and peroxidation or cross-link formation will potentially inhibit biomolecular function. As direct consequences, proteins are denatured, polysaccharides cleaved, nucleic acids are damaged, membranes become less fluid, more porous and may even be disrupted. Proteins can yet be targeted for proteasome degradation.

Lipid peroxidation represents a critical source of cellular damage that often leads to necrosis. Because hydroxyl radicals are extremely reactive non-polar species they can rapidly abstract a hydrogen atom from saturated lipids forming lipid radicals ($\text{L}^\bullet$). Secondly, ground-state oxygen molecules can add to the lipid radical forming peroxy radicals that can subsequently react with other lipids that will propagate chain reactions until recombination or dismutation reactions take place. Since cell membrane consists of double layer of phospholipids, closely interacting with each other and with membrane proteins, radical chain reactions can spread through the membrane causing serious damages that ultimately result in necrosis via membrane rupture (Halliwell and Chirico, 1993). To minimize the extent of this type of damage, nearly all cells in nature accumulate antioxidants molecules and enzymes in membranes preventing the propagation of radical reactions (Buettner, 1993; Halliwell and Gutteridge, 2015). All amino acids, including when modified, are susceptible to damage caused by radical reactions. Protein function is highly associated to its structural conformation, that is resultant of the chemical characteristics of each amino acid in the polymeric chain. Of course, some amino acids are more important than others for protein structure and active site, however, the oxidation of any amino acid in the chain can potentially alter protein conformation and consequently its function. In the case of membrane proteins, note that transmembrane domains are constructed with chains of hydrophobic amino acids. If these amino acids are oxidized they become more hydrophilic and may detach from cell membrane to be solvated by water either in the cytosol or the extracellular media (Davies, 2016).

There are an astonishing array of known compounds that can act as photosensitizers in PDT. In the interests of space we can only give a few examples of their chemical structures here. Fig. 3 shows the structure of Photofrin (the first PS to be clinically used for various cancers, and still one of the commonest). Foscan or Temoporfin (*m*-tetrahydroxyphenylchlorin) is another PS that is commonly used in the clinic to treat different types of cancers. Methylene blue is a phenothiazinium dye, which is the most commonly used PS to treat infections.

Antimicrobial Photodynamic Inactivation and PDT for infectious Disease

Rise of antibiotic resistance

Together with vaccines and public health measures to control transmission of communicable diseases, antibiotics have helped to dramatically reduce mortality from infectious disease during the 20th century (Guyer *et al.*, 2000). However, the continuing global challenge of pathogens such as influenza, tuberculosis, and malaria, as well as the emergence of human

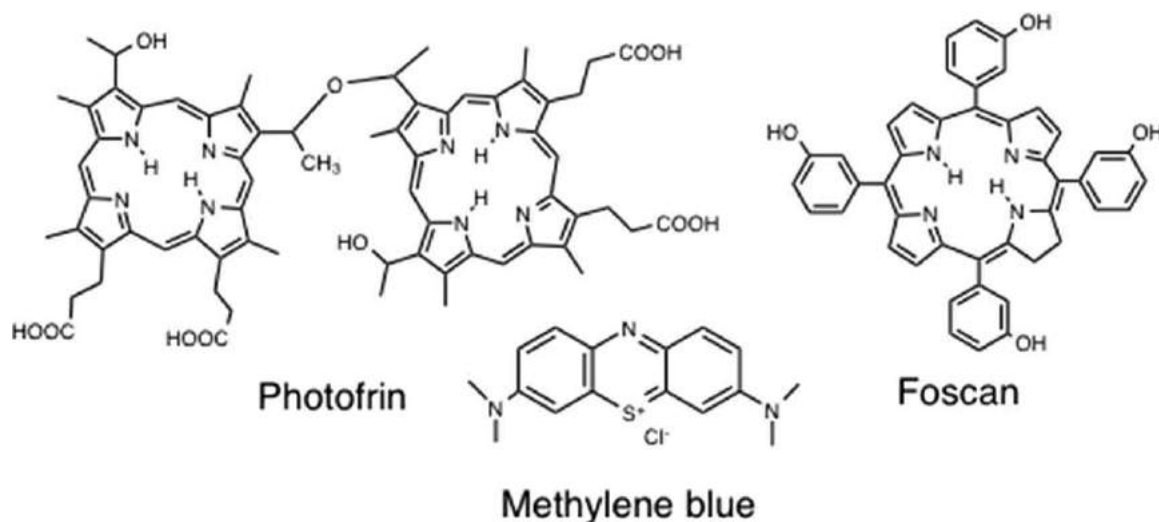


Fig. 3 Chemical structures of three clinically used PS: Photofrin, Foscan, Methylene blue.

immunodeficiency virus, means that infectious disease remains an important problem (Fauci and Morens, 2012). The real value of antibiotics goes beyond simply preventing death and illness due to infection, considering that antibiotics also permit the serious iatrogenic assault on the immune system (immunosuppression) that may occur during cancer treatment or in organ transplantation, where antibiotics have helped to keep surgical complication rates down (Youngblood *et al.*, 2002). Therefore, antibiotics are an extremely valuable resource across the whole spectrum of modern medicine (Smith and Coast, 2013). However, multidrug-resistant (MDR) and pandrug-resistant (PDR) bacterial strains and the related infections they cause, have become emerging threats to public health throughout the world (Kraus, 2008). These infections are associated with an approximately two-fold higher death rate, and with considerably prolonged hospital stays (Munoz-Price *et al.*, 2013). The infections caused by antibiotic resistant microbial strains are often exceptionally hard to treat due to the limited range of therapeutic options (Yoneyama and Katsumata, 2006). Therefore, there is a need for an immediate and all-out search for alternative antimicrobial methods against MDR strains, towards which no resistance likely to develop (Hamblin and Hasan, 2004; Maisch, 2007; Maisch *et al.*, 2011). Recently, Karen Bush *et al.* pointed out that novel non-antibiotic approaches for the prevention and treatment of infectious disease should be considered high-priority international research and development goals (Bush *et al.*, 2011).

In 2015 the O'Neill report garnered much international attention when it delivered the alarming forecast that by 2050 (if nothing were done to stem the growth of MDR bacteria) there would have been 300 million premature deaths that would have cost the world economy \$100 trillion (O'Neill, 2015). One of the most promising and innovative approaches to kill drug-resistant bacteria and treat resistant infections is antimicrobial photodynamic inactivation (aPDI).

History of aPDI

The initial discovery of the PDT/PDI effect was made by in 1900 by a PhD student called Oscar Raab working in the laboratory of Prof Hermann von Tappeiner in Munich, Germany. Raab was studying the toxicity of acridine dyes towards paramecia (a single-celled ciliated protozoan microorganism living in ponds). Raab found that his results were quite variable, but recognized that the apparent toxicity depended on the time of day, and the amount of light coming through the laboratory window, with no toxicity when the experiment was done in the dark. There were sporadic reports of killing of different kinds of microorganisms (bacteria, fungi and viruses) for the next 90 years, but it wasn't until 1990 that major progress was made.

Pathogen classification and PS structure

In the 1990s it was observed that there was a fundamental difference in susceptibility to PDT between Gram (+) and Gram (−) bacteria. It was found that in general neutral or anionic PS molecules are efficiently bound to and photodynamically inactivate Gram (+) bacteria whereas they are bound to a greater or lesser extent only to the outer membrane of Gram (−) bacterial cells but do not inactivate them after illumination (Malik *et al.*, 1992). The high susceptibility of Gram (+) species was explained by their physiology as their cytoplasmic membrane is surrounded by a relatively porous layer of peptidoglycan and lipoteichoic acid that allowed PS to cross. The cell envelope of Gram (−) bacteria consists of an inner cytoplasmic membrane and an outer membrane, which are separated by the peptidoglycan-containing periplasm (Fig. 4). The outer membrane forms a physical and functional barrier between the cell and its environment. In the outer membrane several different proteins are present, some of them function as pores to allow passage of nutrients, whereas others have an enzymatic function or are involved in maintaining the structural integrity of the outer membrane and shape of the bacteria.

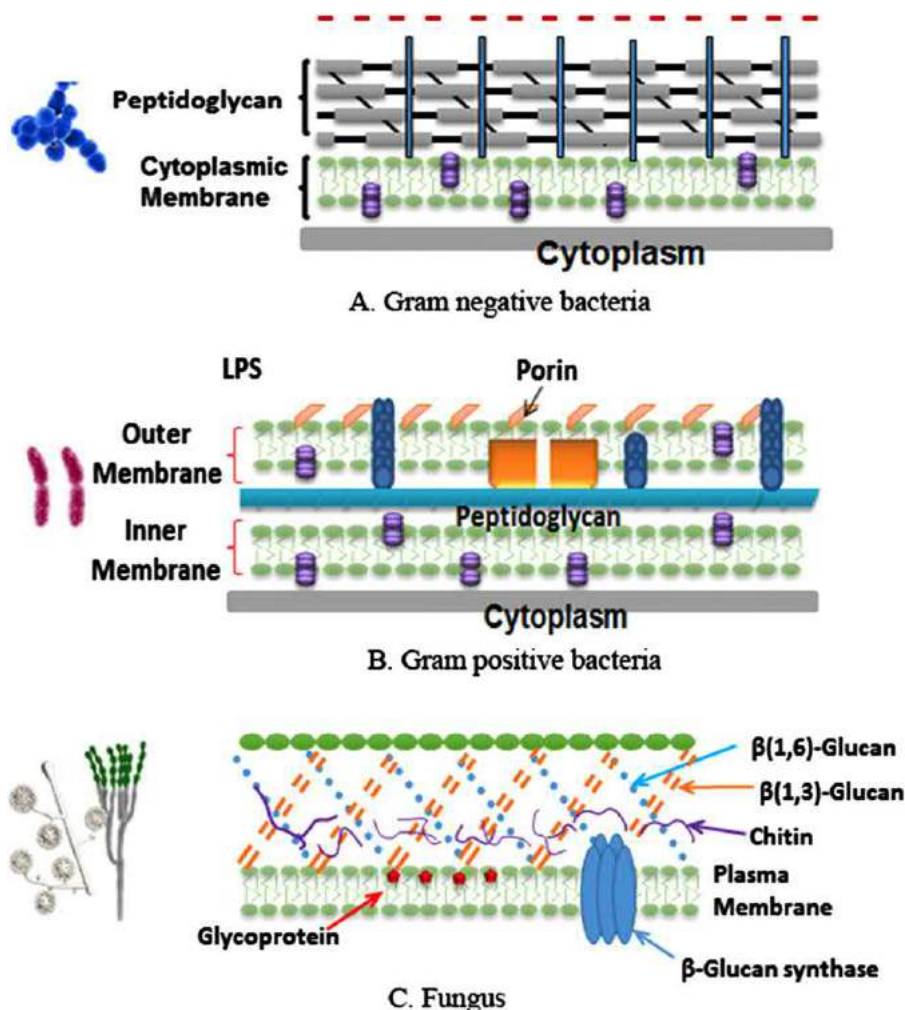


Fig. 4 Cell wall structures of (A) Gram-positive bacteria; (B) Gram-negative bacteria; (C) Fungi.

Then several groups of workers devised approaches that would allow PDI of Gram (–) species. The group in Israel consisting of Malik, Nitzan and co-workers (Nitzan *et al.*, 1992) used the polycationic peptide polymyxin B nonapeptide (PMBN) which increased the permeability of the Gram (–) outer membrane and allowed PS that are normally excluded from the cell to penetrate to a location where the reactive oxygen species generated upon illumination to execute fatal damage. PMBN does not release LPS from the cells but “expands” the outer leaflet of the membrane allowing PS such as deuteroporphyrin (DP) to penetrate and allow PDI of *E. coli* and *P. aeruginosa*. They demonstrated an interaction between PMBN and DP in solution and speculated that this binding assisted the penetration. They used this method to kill a multi-antibiotic resistant strain of *Acinetobacter baumannii* and found that DP seemed to work much better in concert with PMBN than many other PS including porphyrins, phthalocyanines and merocyanine 540 (Nitzan *et al.*, 1998). They also found that the growth medium of the bacteria made a large difference in their susceptibility to PDI, with the high protein brain-heart infusion leading to less killing than the low protein nutrient broth. They found that the type of protein present in the medium made a difference in addition to the concentration. A similar approach was taken by Bertoloni *et al.* (1990) who found that the use of tris-ethylenediamine tetra-acetic acid (EDTA) to release LPS or the induction of competence with calcium chloride sensitized *E. coli* and *Klebsiella pneumoniae* to PDI by haematoporphyrin or zinc phthalocyanine.

A second approach adopted by several groups is to use a PS molecule with an intrinsic positive charge. The groups of Wilson (2004) and Wainwright (1998) have used the phenothiazinium salts such as MB and toluidine blue O to carry out PDI of a large range of bacteria both Gram (+) and Gram (–). The group in Italy led by Jori have used cationic porphyrins (meso-tetra (N-methyl)-4-pyridyl)-porphine tetraiodide and tetra-(4N,N,N-trimethyl-anilinium)-porphine to photoinactivate Gram (–) species such as *Vibrio anguillarum* and *E. coli* (Merchat *et al.*, 1996). They found that washing the loosely bound PS from the cells before illumination decreased the killing and explained this finding by supposing that the first dose of light on PS bound to the outside of the outer membrane causes an initial limited photodamage that then allows further penetration of the PS. The group in Leeds, UK led by Brown has used cationic phthalocyanines for PDI of Gram (–) bacteria (Minnock *et al.*, 1996). They investigated

E. coli DH5a and in particular the mechanism of uptake. They found that incubation with PPC in the dark led to increased sensitivity of the bacteria to hydrophobic but not hydrophilic antibiotics. Incubation with PPC also led to increased uptake of radiolabeled protoporphyrin that was reversed in the presence of up to 50 mM Mg^{++} ions. These observations were consistent with the uptake of PPC proceeding through the self-promoted uptake pathway (see next section).

However there are other reports of PDI of Gram (–) bacteria in which it is clear that the PS does not have to penetrate the bacterium to be effective, or indeed even come into contact with the cells. According to these papers, if singlet oxygen can be generated in sufficient quantities near to the bacterial outer membrane it will be able to diffuse into the cell to inflict damage on vital structures. In one set of studies, the bacteria were separated from the PS by a layer of moist air and singlet oxygen in the gas phase was generated and allowed to diffuse across the gap before contacting the bacteria (Valduga *et al.*, 1993). Gram (–) species were harder to kill than Gram (+), and the intracellular content of carotenoids was found to protect the bacteria from photo-inactivation. In another study, the PS Rose Bengal was covalently bound to small polystyrene beads that were allowed to mix with the bacteria in suspension (Dahl *et al.*, 1989).

How can these two conflicting sets of findings; one pointing out the necessity for PS penetration into Gram (–) bacteria, and the other pointing out that extracellular generated singlet oxygen is effective, be reconciled? One possibility is that indeed singlet oxygen can diffuse into bacteria producing fatal damage if it is produced at the outer surface or in solution in close proximity. The diffusion distance of singlet oxygen in solution has been estimated to be approximately 20–100 nm (Klaper *et al.*, 2016). The failure of some PS that bind to Gram (–) species to produce any killing however, must mean that the reactive species produced on illumination were unable to diffuse inward to sensitive sites. The fact that these PS can lead to efficient PDI of Gram (+) species shows that the molecules have not undergone any conformational change that has rendered the PS no longer photoactive. Hence we hypothesize that PS that operate chiefly via Type I mechanisms need to penetrate the Gram (–) outer membrane, while those that act mainly by Type II PS may not, and that the former may be more efficient.

Virulence

Photodynamic therapy is one of the few antimicrobial therapies, that can not only kill microbial cells, but can also destroy many virulence factors that microbial cells produce in order to facilitate their adherence, invasion, survival, and toxicity in host cells and tissues. These virulence factors are usually biomolecules such as proteins, lipids and carbohydrates, that are just as susceptible to oxidative damage by the ROS produced during PDT, as the microbial cells themselves.

PDT for infectious disease

In recent years the application of PDT for the treatment of localized infections called aPDI has steadily increased in popularity (Fig. 5). This application can superficially appear much simpler than the more well-recognized application of PDT for cancer treatment (see below). The complexity of dealing with the vascular effects of PDT is absent, because the PS is almost always topically applied into/onto the infected area of tissue. Moreover, there have not as yet been many reports about activation of the immune system being triggered after carrying out aPDI for an infectious disease (see later).

So it is generally assumed that a PS with cationic charges, which is more or less selective for microbial cells, is injected or otherwise topically applied into the infected area, where it binds to the microbial cells after a short-drug light interval, in order to

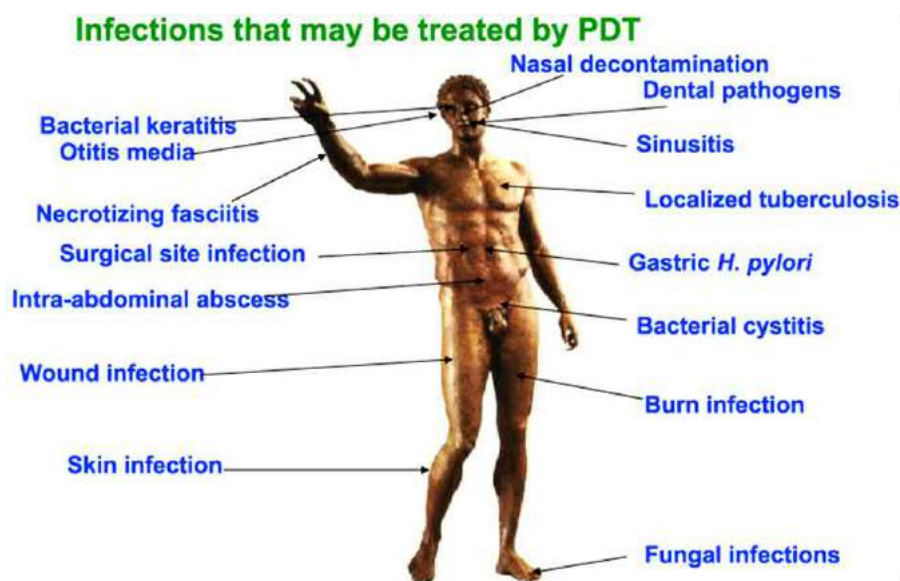


Fig. 5 Localized infections that could possibly be treated by PDT.

reduce the uptake of the PS by the surrounding host cells, and increase selectivity. However what is not often realized is the big difference between tumors and infections. In tumors all the surrounding tissue is considered to be undesirable, diseased and in need of removal. This is completely different to the case of an infection. An infection is defined as greater than 10(5) CFU of microbial cells per gram of tissue. The mass of 10(5) CFU of bacterial cells is approximately 1 µg. So the ratio between the mass of the bacterial tissue and the mass of the host tissue is in the region of 1 to 1,000,000 (Hamblin and Dai, 2010). So in sharp contrast to the case, where in tumors the goal is to destroy 100% of the surrounding tissue, in infections the goal is to destroy only 0.0001% of the surrounding tissue (just the bacterial cells). For this to be anyway near feasible the selectivity of the PS for bacteria has got to be exquisite so that 999,999 PS molecules out of every million bind to their target bacteria and not to the host cells. As can be imagined this is rather unlikely. However the situation is not as dire as the foregoing has made it out to be. The reason for this is that a short drug light interval is used for aPDI and even if the selectivity for bacteria over host cells is somewhat less than 1 million to 1, then the poly-cationic nature of the PS means that it is fairly slow to be taken up by host mammalian cells. So if light is delivered after say 30 s from topical delivery of the PS, it can be reasonably save to assume that not many of the surrounding host cells would be killed.

There have not been many reports of damage to host tissue being caused by aPDI in animal models. One study (Gad *et al.*, 2004) did use a model of a subcutaneous abscess caused by Gram-positive bioluminescent *Staphylococcus aureus* injected into the mouse thigh muscle. A small volume of PS solution was injected into the infected area and light was shone as a spot onto the surface. Two different PS were compared, the polycationic conjugate, poly-L-lysine chlorin(e6) (pL-ce6), and the anionic PS, free chlorin(e6). Both PS were equally effective in destroying the bacteria as judged by non-invasive bioluminescence imaging. However the non-bacterial targeted free chlorin(e6) caused more damage to the surrounding mouse thigh muscle as measured by a leg function test. The bacterial-targeted pL-ce6 was able to kill the bacteria without damaging the thigh muscle.

There has only been one report of activation of the immune system after PDT, when used against an infection in an animal model (Tanaka *et al.*, 2012). This involved injection of bioluminescent *S. aureus* into the mouse knee to produce a form of septic arthritis. When methylene blue was injected into the knee and 660 nm light delivered to the surface, the bacteria were killed in a similar manner to the experiment described above. However when the MB PDT was carried out before the bacteria were injected into the knee, rather than afterwards, an interesting observation was made. PDT carried out 24 h (but not 2 h) before infection protected the mice from developing a severe knee infection when the bacteria were subsequently injected, as shown by the fact that the infection resolved earlier (Tanaka *et al.*, 2012). This effect was dependent on neutrophils in the mouse knee as demonstrated by use of several blocking antibodies and inhibitors. Although there was no immediate influx of neutrophils into the mouse knee immediately after PDT, the neutrophils did appear to be “primed” to react quicker to attack the bacteria, once these were injected into the knee joint. Therefore, PDT of the mouse knee could be envisaged as a ‘non-specific vaccination procedure’, which could in principle be used in cases of orthopedic surgery, which might be considered high-risk for infection.

As mentioned above, for many years PDT was studied as a cancer therapy by designing PS that could be administered either systemically (by intravenous injection) or applied topically (e.g. aminolevulinic acid), and some time later the tumor would be irradiated with light (either as a surface spot or by insertion of interstitial fiber optics) (Agostinis *et al.*, 2011). However starting in the 1990s it was realized that PDT could also exert a powerful antimicrobial effect, if PS could be designed that could selectively bind to microbial cells, while not binding to host mammalian cells (Malik *et al.*, 1992). The best way of achieving this goal of antimicrobial photodynamic inactivation (aPDI) was to ensure that the PS had a pronounced cationic charge, as it was realized that microbial cells in general have a more pronounced negative charge compared to mammalian cells and cationic PS will bind selectively. Moreover the binding of the PS to the microbial cells is relatively rapid, while uptake of the cationic PS by mammalian cells is slow, thus giving good selectivity when a short drug-light interval (few minutes) is employed (Dai *et al.*, 2009). The advantages of aPDI as a potential clinical antimicrobial therapy were bolstered when it was realized that aPDI works equally well regardless of the antibiotic resistance status of the microbial cells (Vera *et al.*, 2012) and moreover, that aPDI has (so far) not been shown to produce resistance in bacteria (Maisch, 2015) even after 20 successive cycles of partial killing followed by regrowth (Giuliani *et al.*, 2010). Another advantage of aPDI is that the PS is applied topically or locally into the infected area. Many chronic infections involve a build up of microbial biofilms, into which it is now well-recognized, that systemically administered antibiotics fail to penetrate. However, aPDI has been shown to kill biofilm-grown cells both in vitro and in vivo (de Melo *et al.*, 2013). This anti-biofilm application has found particular application in dental infections such as periodontitis (Kikuchi *et al.*, 2015) and peri-implantitis (Vohra *et al.*, 2014). Moreover, infections in burns or damaged tissue suffer from a compromised blood supply, so systemically administered antibiotics fail to reach the site of infection in sufficient concentrations. The killing of microbial cells with aPDI is rapid (seconds) while the action of antibiotics can take hours or days, giving a potential advantage against fast-spreading infections such as necrotizing fasciitis. Moreover the broad-spectrum nature of aPDI means that treatment can be instituted before the infectious agents have been identified (Ferreyra *et al.*, 2016). Although many infections can occur deep inside the body, it is now possible to deliver both PS and light to almost any anatomical region, via endoscopes and narrow-diameter interstitially-inserted needles and fiber optics (Gad *et al.*, 2004).

PDT for Cancer

Cancer is still the most often-discussed indication for PDT and many papers have been published discussing modes of cell death in vitro (Fig. 6), and investigating pharmacokinetics and tumor response in vivo.

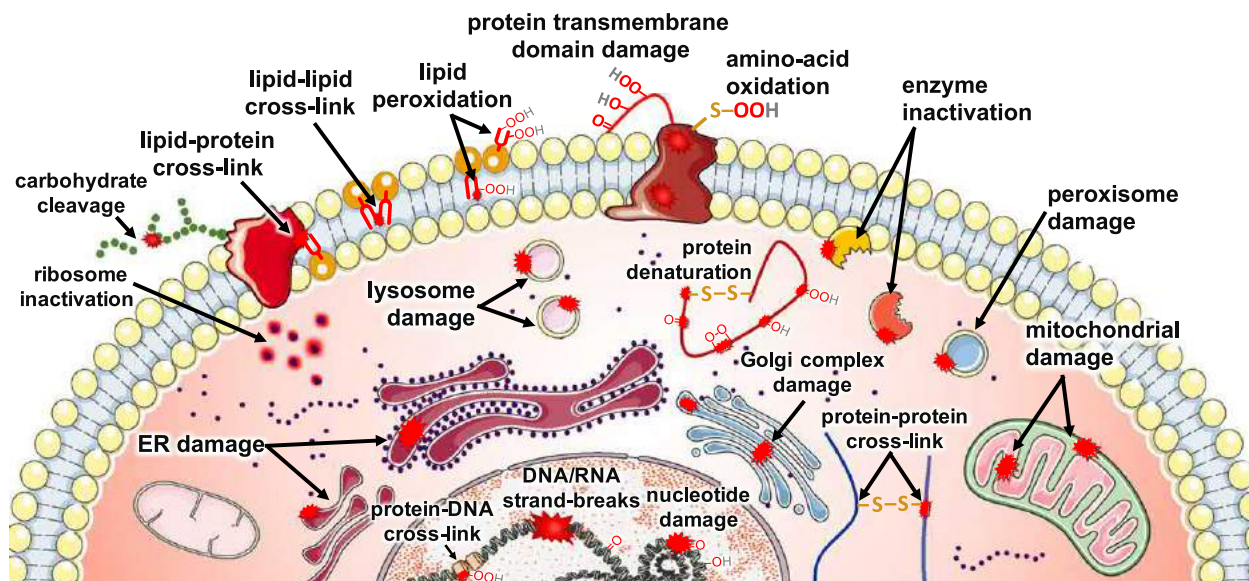


Fig. 6 PDT causes damage to biomolecules and intracellular organelles.

Cellular mechanisms of PDT

PDT and apoptosis

Apoptosis is a very complex, multi-step, multi-pathway cell-death program that is genetically encoded in every cell of the body (Pettigrew and Cotter, 2009). It can be initiated either through the activation of death receptors or the mitochondrial release of cytochrome c (Rustin, 2002). Both events eventually lead to activation of caspase cascades. Caspases are cysteine aspartic acid proteases and some members of the family are 'executioner caspases' such as caspase-3, -6 and -7 (Gougeon and Kroemer, 2003; Perfettini and Kroemer, 2003). The active executioner caspases cleave cellular substrates, which leads to characteristic biochemical and morphological changes observed in dying cells (Rathmell and Thompson, 1999). Chromatin condensation and nuclear shrinkage; cleavage of the inhibitor of the DNase CAD (caspase activated deoxyribonuclease) causes DNA fragmentation. Cleavage of cytoskeletal proteins leads to cell fragmentation and formation of apoptotic bodies (Savill and Fadok, 2000). The apoptotic process is tightly controlled by various proteins (Cotter, 2009; Okada and Mak, 2004). It is well known that resistance of tumor cells to apoptosis might be an essential feature of cancer development thus, modulation of the key elements of apoptosis signaling may directly influence tumor-cell death induced by various forms of cancer therapy (Igney and Krammer, 2002a,b).

PDT is thought to induce photodamage to Bcl-2 and related antiapoptotic proteins and activate the proapoptotic members of the family (Oleinick *et al.*, 2002). Alteration in expression of members of Bcl-2 family proteins following PDT has been reported in various cell lines and tumors. During apoptosis, the pro-apoptotic proteins Bax and Bak oligomerize in the mitochondrial outer membrane and probably disturb its integrity, leading to release of pro-apoptotic signaling molecules like cytochrome c, which then allows the activation of caspase 9 (Cory and Adams, 2002).

Accumulating evidence suggests that PDT is an efficient inducer of apoptosis in many cancer cell lines, whereas cells that lack the ability to efficiently undergo apoptosis are not protected from the lethal effects of PDT. Bax was not found to be essential for Photofrin-PDT induced apoptosis in human lung adenocarcinoma cells (ASTC-a-1) (Wu *et al.*, 2011). In the absence of Bax/ Bak, autophagy appears to be predominantly responsible for cell death following PDT with different photosensitizers (Buytaert *et al.*, 2006a,b; Kessel, 2006).

PDT produces a level of oxidative stress that can result in activation and translocation of the transcription factor NF- κ B to the nucleus, as originally shown for treatment of L1210 murine leukemia cells with Photofrin and light (Ryter and Gomer, 1993). NF- κ B provides an anti-apoptotic signal or cell survival pathway for cells exposed to PDT (Coupienne *et al.*, 2010).

PDT and necrosis

Necrotic cell death has been described as a violent and rapid form of degeneration affecting large fractions of cell populations, characterized by cytoplasmic swelling, destruction of organelles and disruption of the plasma membrane, leading to the release of intracellular contents and consequent inflammation (Danial and Korsmeyer, 2004).

Necrosis has been referred to as accidental cell death, caused by physical or chemical damage and has generally been considered an unprogrammed process (Kitsis and Molkentin, 2010). It is characterized by formation of a pyknotic nucleus, cytoplasmic swelling, and progressive disintegration of cytoplasmic membranes, all of which lead to cellular fragmentation and release of material into the extracellular compartment. In necrosis, decomposition is principally mediated by proteolytic activity, but the precise identities of proteases and their substrates are poorly defined (Proskuryakov and Gabai, 2010).

Studying the factors and parameters that cause cellular necrosis after PDT is not as easy as studying those factors which lead to apoptosis. The crucial factors in determining the type of cell death, e.g. apoptosis or necrosis following PDT are: the cell type, the presence of an intact set of apoptosis machinery, the subcellular localization of the PS, the light dose applied to activate it locally, and the oxygen partial pressure (Buytaert *et al.*, 2007). One factor that can be agreed upon by all commentators is that high dose PDT (either a high photosensitizer concentration or a high light fluence or both) tends to cause cell death by necrosis, while PDT administered at lower doses tend to predispose cells towards apoptotic cell death.

It is possible that high doses of PDT can photochemically inactivate essential enzymes and other components of the apoptotic cascade such as caspases. For instance, Lavie and coworkers (Lavie *et al.*, 1999) used the perylenequinones (hypericin and dimethyl tetrahydroxyhelianthrene) and found high dose PDT inhibited apoptosis by interfering with lamin phosphorylation, or by photodynamic cross-linking of lamins.

Xue *et al.* (2001) compared Pc4-mediated PDT of MCF7 cells that lack caspase 3 with the same cell line with caspase 3 transfected back in. They found apoptotic indicators only in the caspase expressing cells which also showed more loss of viability by an assay involving reduction of a tetrazolium dye; however both cell lines showed an equal degree of cytotoxicity by a clonogenic assay.

Dahle, Steele and Moan reported (Dahle *et al.*, 1999) that the mode of cell death induced by PDT depended on cell density. They seeded Madison Darby canine kidney II cells (MDCK II) at two different densities and incubated them with meso-tetra(4-sulfonatophenyl)porphine (TPPS4) for 18 h, washed and irradiated with blue light. Four hours later the cells were studied by fluorescence microscopy. With <55% total cell death the apoptotic fraction was significantly higher for cells in confluent monolayers than for cells growing in microcolonies at equitoxic doses. Confluent cells were 2.9 times more sensitive than cells in microcolonies partly due to a 1.5 times higher uptake of TPPS4 in monolayer cells. The difference in mode of cell death for the different cell densities was not related to any observable difference in subcellular localization pattern of TPPS4 at equitoxic doses of PDT.

PDT and autophagy

Autophagy is a catabolic cellular mechanism that allows the cell to maintain a balance between the synthesis, degradation, and recycling of cellular products (Mathew *et al.*, 2007). A variety of autophagic processes exist, all of which involve the lysosomal degradation of the cellular organelles and proteins. The most well-known mechanism proceeds in the following manner: A double membrane structure called autophagosome surrounds the target region, creating a vesicle that separates its contents from the rest of the cytoplasm. This vesicle is then transported and fused to the lysosome, forming a structure called the autophagolysosome, the contents of which are subsequently degraded by lysosomal hydrolases (Kroemer and Jaattela, 2005). Besides facilitating the disposal of unwanted proteins, organelles, and invading microorganisms, autophagy also allows a cell to reallocate its nutrients from unnecessary processes to life-essential ones in times of starvation or stress. The term “autophagy” was coined in 1963 at the Ciba Foundation Symposium on Lysosomes, where it was used to describe the presence of membrane vesicles that contain disintegrating organelles and cytoplasm components. Evidence soon confirmed that autophagy is an adaptive, energy-generating process. Molecular control of autophagy came into focus in the late 1990s. Signaling pathways governing this process became better understood after the identification of the target of rapamycin kinase (TOR), which controls cell growth and protein synthesis (Bjornsti and Houghton, 2004).

It is still unclear exactly how autophagy affects the outcome of PDT (Kessel, 2006; Kessel *et al.*, 2006; Reiners *et al.*, 2010). In general, mammalian cells use autophagy as a defense against ROS mediated damage by clearing the cell of damaged organelles (Scherz-Shouval *et al.*, 2007). Depending on the type of ROS and degree of oxidative injury, PDT may stimulate autophagy (Sasnauskiene *et al.*, 2009) that either acts in a cytoprotective manner or induces autophagic cell death (Scherz-Shouval and Elazar, 2007). Autophagy may play a role in PDT induced apoptosis, but the two processes can also occur independently of one another (Kessel and Oleinick, 2009). A study done on murine leukemia L1210 cells found that a wave of autophagy occurs right before apoptosis. (Kessel and Arroyo, 2007). It was also found that prevention of autophagy by silencing the Atp7 gene allowed photo-killing to occur at lower light doses. This observation is consistent with the theory that autophagy is a defense mechanism against PDT induced ROS (Kessel and Reiners, 2007). The situation is different for tumor cells that do not have the ability to undergo apoptosis through a deficiency in Bax or Bac, which regulate the apoptotic pathway. In these cells, PDT induced autophagy stimulates a necrotic, caspase-independent cell death (Buytaert *et al.*, 2006a). Suppression of autophagy in apoptosis-deficient cells resulted in the inhibition of cell death during PDT. In general, the induction of autophagy in PDT treated cells occurs independently of an apoptotic outcome. While autophagy seems to play a pro-survival role in tumor cells that are capable of apoptosis, it has been shown to promote death in cells that are apoptosis-deficient.

PS pharmacokinetics

When PS preparations are injected into the bloodstream they tend to bind to various serum proteins and this binding can dramatically affect their pharmacokinetics and biodistribution (Boyle and Dolphin, 1996). Different PS can have very different pharmacokinetics and this can directly affect the illumination parameters and particularly the pharmacokinetics will determine the drug-light interval (Agostinis *et al.*, 2011). Intravenously injected PS undergo a gradual transition from being bound to serum proteins, then becoming bound to endothelial cells, then they are bound to the adventitia of the vessels, then bound either to the extracellular matrix or to the cells within the tumor, and eventually they will be cleared from the tumor tissue by removal by lymphatic vessels or by blood vessels, and they will be excreted either by the kidneys or the liver. The effect of PDT on the tumor largely depends at which

stage of this continuous process light is delivered. The anti-tumor effects of PDT are divided into three main mechanisms. Direct tumor cell death by apoptosis or necrosis can occur, if the PS has been allowed to be taken up by tumor cells. Powerful anti-vascular effects can lead to thrombosis and hemorrhage in tumor blood vessels that subsequently leads to shut-down of tumor blood vessels, which kills the tumor by deprivation of oxygen and nutrients. Finally the release of damage associated molecular patterns (DAMPs) leads to acute inflammation and the release of cytokines (Garg *et al.*, 2011). The DAMPs, cytokines and stress response proteins induced in the tumor by PDT work together and lead to an influx of leukocytes that can both contribute to tumor destruction as well as to stimulate the immune system to recognize and destroy tumor cells even at distant locations.

When PS are injected into the bloodstream a certain sequence of events commences that can take very different lengths of time to reach completion, for PS with different chemical structures. Firstly, depending on the delivery solvent or vehicle that is used for the PS injection, the PS must come to equilibrium with the components of the circulating blood. This can involve the PS disaggregating from itself being released from its delivery vehicle and binding instead to various protein components of serum (see later). In addition the blood has many circulating cells (erythrocytes and leukocytes) that would in principle be available to bind injected PS (Maugain *et al.*, 2004). Secondly, the circulating PS must bind to the walls of the blood vessels, and it is thought that the nature of the various blood vessels in the tumor and in surrounding normal tissues that vary significantly in size, speed of blood flow, and physiological characteristics such as leakiness and tortuosity in tumors and various normal organs, governs to a great extent where the PS localizes or accumulates. Thirdly, the PS will penetrate through the wall of the blood vessel at a rate that probably depends on how strong the initial binding of the PS to the intimal surface was. PS that bind strongly to the blood vessel wall will take a longer time to cross the entire wall of the blood vessel, and those that have an initial weaker binding will pass through more quickly. Fourthly, after extravasation, the PS will diffuse throughout the parenchyma of the organ or tumor to which it has been delivered. If the organ happens to be the liver or other metabolically active organ, the PS may be subject to chemical changes by metabolic enzymes, but this is thought to be rather unlikely for most tetrapyrrole-based PS in common clinical use. Fifthly, the PS will eventually be eliminated from the tissue. Although this elimination has not been much studied, if the tissue containing the PS is a tumor or other non-metabolic organ, elimination will probably be by lymphatic drainage and the PS will then return into the general blood circulation via the thoracic duct. Sixthly, the PS will be excreted from the body. For the majority of PS in clinical use, excretion is from the liver into the bile and thence to the intestine where it is lost via fecal elimination. However entero-hepatic re-circulation, which is the term used for the re-absorption of the PS from the small intestine, back into the blood stream, has been reported for some PS.

Measurement of pharmacokinetics requires sequential measurements of PS concentration in tissue or other biological substance or fluid in order to construct temporal profiles that can be compared for different PS. This is frequently done for blood where samples can be readily removed in small amounts at frequent intervals and analysis then yields values of serum half-lives (first or second order) in minutes, hours or days. Alternatively, analysis of PS concentrations in urine or fecal samples can give terminal elimination half-lives. All these methods require a method of quantitatively analyzing PS concentrations in samples of biological material. The fact that the majority of PS used for PDT are also fluorescent, has led to the development of assays involving homogenization or dissolving the tissue or biological fluid and measurement of fluorescence in solution after centrifugation or extraction steps (Bellnier *et al.*, 1997). Calibration curves can be prepared with mixtures of known amounts of PS and weighed amounts of biological material. Care must be taken when these experiments are carried out on small rodents such as mice and rats because these animals accumulate a naturally red-fluorescent compound derived from chlorophyll in their diet that ends up in the skin and the digestive tract amongst other organs (Gudgin Dickson *et al.*, 1995). This can be avoided by feeding the animals a chlorophyll-free diet for enough time before the experiment to allow the interfering fluorescent compound to be eliminated (Allison *et al.*, 1994). In some cases the PS is quantified by a traditional analytical chemistry technique such as high performance liquid chromatography (Almeida-Lopes *et al.*, 2001). This method has the advantage of detecting possible metabolism products that have been formed from the originally pure PS after injection. Finally in some cases of experimental animal models, the PS has been presynthesized in the laboratory to include a radioisotope label (for instance carbon-14 or tritium) or a chelator can be attached that can bind a metal radioisotope and biological samples can then be quantified using a scintillation counter (Bellnier *et al.*, 1989; Little *et al.*, 1984; Schuitemaker *et al.*, 1993). An alternative approach to carrying out analysis of samples of biological material removed sequentially, is to use a continuous optical monitoring method to determine PS concentration in tissue based on fluorescence. In this approach a non-invasive or minimally invasive approach with fiber-based fluorescence detection systems can be used to measure the variation over time of a fluorescence signal linked to the PS of interest (Frisoli *et al.*, 1993). In some small animal models the skin can be removed from a subcutaneous tumor so the fiber that collects fluorescence can be in contact with the underlying tumor as well as the adjacent skin (Sheng *et al.*, 2004).

PS biodistribution

In experimental animals it has sometimes been possible to establish the overall pattern of distribution in the different organs after intravenous injection of the PS. A study (Bellnier *et al.*, 1989) using radioisotope labeled Photofrin in mice with subcutaneous fibrosarcoma tumors that were sacrificed 24h later found highest concentrations in liver, adrenal glands, and urinary bladder. Next was pancreas, kidney, spleen, and these organs were higher than stomach, bone, lung, and heart. Muscle concentration was low and brain was lowest. Only skeletal muscle, brain, and skin located distant to the tumor had peak concentrations lower than tumor tissue; the skin overlying the tumors showed concentrations not significantly different from tumor. The higher molecular weight components of Photofrin were partially retained in the liver and spleen at 75 days after injection.

In most cases where organ biodistribution data have been reported for small animals such as mice and rats, the highest concentration of PS has been found in the liver (Bertoloni *et al.*, 1990; Bhargava *et al.*, 2012). Chan *et al.* studied a set of sulfonated phthalocyanines found that the accumulation in liver was inversely proportional to the degree of sulfonation and hence to the overall lipophilicity of the PS molecules (Bhat *et al.*, 2005). The liver is known to have a highly permeable blood supply with fenestrated endothelium that contains pores that allow molecules to pass easily out of the blood vessels. This porous vasculature is required by the role of the liver in detoxifying the body of extraneous molecules (particularly organic molecules such as PS). These molecules are then excreted into the bile via the gall bladder in a similar fashion to the production of bile acids from cholesterol. The bile is routed into the duodenum where it is possible for unchanged PS to be absorbed into the circulation a second time from the small intestine (a process known as entero-hepatic recycling) as well as being excreted via the feces. It is known that soon after injection, large amounts of PS can accumulate in the lungs. Lung accumulation has been reported (Brun *et al.*, 2004) as the mode of dose-limiting dark toxicity when the PS injected doses were increased eventually leading to lung hemorrhage and acute interstitial pneumonia after injection of a high dose of 80 mg/kg of Pc4 into mice. The explanation is probably that large injected doses of PS can aggregate in the blood vessels and the particles so formed can easily collect in the fine capillary network of the lungs.

The amount of PS that accumulates in the spleen seems to vary widely with the structure of the PS (overall charge and hydrophobicity). Woodburn *et al.* (1992) studied a range of porphyrins with varying octanol/water partition coefficients and found the accumulation in the spleen varied the most compared to other organs. Egorin *et al.* (1999) compared organ distribution of Pc4 administered by IV injection of the same dose (10 mg/kg) dissolved in several solvents and found that the accumulation in the spleen was the most variable.

The kidney and bladder frequently accumulate high amounts of PS, despite the clearance route being almost exclusively via liver and bile (Richter *et al.*, 1990) rather than the kidneys and urine. The digestive organs (stomach, large and small intestines) seem to take up middling amounts of PS (i.e. less than liver but more than muscle) (Schuitmaker *et al.*, 1993). The skin has not been found to accumulate high amounts of PS in experimental animals, which might be considered somewhat surprising in view of the fact, that skin photosensitivity due to PS accumulation in patients' skin is the main cause of PDT-related side-effects. The lowest concentrations of PS are generally found in organs such as heart, skeletal muscle, bone, eye and brain. These are organs that are known to have a relatively impermeable blood supply. The blood-brain barrier serves to exclude PS from the parenchyma of the brain, while the fact that the blood-brain barrier is frequently breached by the growth of brain tumors is probably responsible for the very large (over 100) tumor-to-normal brain ratios reported for some PS (Boyle and Dolphin, 1996).

Since the original observation of the tumor localizing ability of hematoporphyrin derivative (HPD), many workers have investigated the underlying mechanism of this effect whereby PS preferentially localize in tumors, and in other specific organs and anatomical locations (Hamblin and Newman, 1994; Jori, 1989; Larroque *et al.*, 1996). The precise definition of "tumor to normal tissue ratio" has also been subject to debate. Some investigators working with subcutaneous tumors in experimental animals use the ratio between the tumor and peritumoral muscle or skin, while others use the ratio between the tumor to distant muscle and skin. Furthermore, there has been much effort made to determine which molecular features in the chemical structures of the PS are optimal for maximizing the selectivity for the tumor over normal tissue and organs. This has proved quite complicated because the pharmacokinetics of different PS can vary dramatically. For instance, one PS can have its best tumor to normal tissue ratio at a relatively early time point after administration such as 3 h, while for another PS this time point can be as long as 7 days after injection. One of the notable properties of these tetrapyrrole molecules that are commonly employed as PS, that may be relevant to their tumor localizing ability, is their tendency to bind strongly to serum proteins and also to bind to each other (aggregation). This high tendency to bind means that most PS when injected into the bloodstream behave as macromolecules rather than as small molecules, either because they are more or less firmly bound to large protein molecules or because they have formed intermolecular aggregates of similar size. Many workers have reported (Kessel and Poretz, 2000; Kongshaug *et al.*, 1989; Maziere *et al.*, 1990) on the distribution of PS between the various classes of serum proteins when mixed with serum (either human or animal) *in vitro*. These serum proteins are usually divided into four classes: (a) albumin and other heavy proteins; (b) high density lipoprotein (HDL); (c) low density lipoprotein (LDL); and (d) very low density lipoprotein (VLDL). However even this study has been complicated by the fact that the most lipophilic PS are insoluble in aqueous media and need to be delivered in a solvent mixture which may alter the serum protein distribution of the PS. It has been argued that PS that preferentially bind to LDL are better tumor localizers (see below) (Jori and Reddi, 1993) but this is by no means always the case (Korbelik, 1992).

It is useful to make a distinction between selective accumulation and selective retention. The tumor localizing ability of the PS with the faster pharmacokinetics is probably due to selective accumulation in the tumor, while the localization of PS with slower acting pharmacokinetics is more likely due to selective retention. In the selective accumulation model it is thought that the increased vascular permeability to macromolecules typical of tumor neovasculature is chiefly responsible for the preferential extravasation of the PS. These quick-acting PS frequently bind to albumin, which is of ideal size and Stokes radius to pass through the "pores" in the endothelium of the tumor micro-vessels (Yuan *et al.*, 1993). The selective retention model where PS can be retained in tumors, while they are eliminated from surrounding normal tissue and organs, has been the subject of much speculation. As mentioned above, a popular theory maintains that the binding of the PS to LDL is of major importance (Jori and Reddi, 1993). According to this theory, it is proposed that cancer cells over-express the LDL (apoB/E) receptor. Up-regulation of the expression of LDL receptors is one strategy that rapidly growing malignant cells have developed to obtain sufficient cholesterol that the tumor cells need for the biosynthesis of lipids required due to the rapid turnover of their cellular membranes. There is experimental evidence both for and against this LDL receptor theory (Allison *et al.*, 1994; Korbelik, 1992). Other theories have been proposed to account for the selective retention of PS in tumor tissue. One is that tumors are characterized by poorly

developed lymphatic drainage, and that macromolecules, which extravasate from the hyperpermeable tumor neovasculature, are retained in the extravascular space (Buytaert *et al.*, 2006b). This theory has been termed the well known “enhanced permeability and retention” (EPR) effect (Maeda *et al.*, 2016). Another theory is attributed to macrophages, which are phagocytic cells which infiltrate solid tumors to varying extents (Blanco *et al.*, 2015). These tumor-associated macrophages (TAMs) have been shown to accumulate up to thirteen times the amount of some PS compared to the cancer cells themselves (Buytaert *et al.*, 2007). The explanation for this preferential uptake by TAMs has been proposed to be due to either the phagocytosis of aggregates of PS (Canti *et al.*, 1994), or to the preferential uptake by TAMs of lipoproteins which have been subtly altered by the binding of porphyrins (Blanco *et al.*, 2015). Another theory proposes that the low pH commonly found in tumors (due to the Warburg effect in which the metabolism changes from oxidative phosphorylation to glycolysis, and produces large amounts of lactic acid) has the effect of trapping some of the anionic PS, which are normally ionized at normal physiological pH. These PS then become neutrally charged at low pH and hence more lipophilic, when they encounter the acidic pH in the tumor environment (Pottier and Kennedy, 1990). Yet another proposed mechanism is the selective accumulation of PS in tumors is related to the propensity of tumors to accumulate lipid droplets which may contain lipophilic molecules (Freitas, 1990). A final theory concerns the tendency of some tumors to have very high populations of tumor-associated macrophages, that may engulf particles of aggregated PS (Korbelik and Kros, 1995), or else take up PS that have bound to LDL, and then altered its behavior to be recognized by scavenger receptors (Hamblin and Newman, 1994).

Direct destruction of tumors by PDT

Many of the studies relating to this topic have been concerned with attempts to distinguish between direct tumor cell killing *in vivo* as described in the previous chapter, and indirect tumor cell killing caused by the vascular effects of PDT (see later).

Some studies have investigated an experimental model in which tumors in mice that had been treated with a potentially curative dose of PDT *in vivo* were removed from the mice, and then dissociated into a suspension of individual tumor cells. These tumor cells were still able to form colonies in a clonogenic assay if the tumors were removed immediately from the host after PDT (Henderson and Fingar, 1989; Henderson *et al.*, 1985). However when these tumors were left in the host animal for increasing lengths of time, a progressive loss in clonogenicity of the dissociated tumor cells was observed. These times corresponded to the emergence of PDT-induced tumor hypoxia as determined radiologically. These studies suggested a central role for vascular damage in governing the tumor response to PDT in mouse models.

Anti-vascular effect of PDT

Photodynamic alteration of tissue microcirculation was first reported in 1963 (Castellani *et al.*, 1963). A study by Star *et al.* (1986) utilized a window chamber in rats bearing an implanted mammary tumor, to make direct observations of the microcirculation of the tumor and the adjacent normal tissue before, during, and at various times after PDT mediated by HPD. An initial slowing down of blood flow and vasoconstriction of the tumor vessels was followed by heterogeneous responses including eventual complete cessation of blood flow, hemorrhage, and in some larger vessels, the formation of platelet aggregates. Other workers have histologically examined tissues removed from animal models that had been treated with PDT such as subcutaneous urothelial tumors and normal rat jejunum and reported a range of vascular responses including disruption of blood flow, breakdown of the blood brain barrier in the normal brain of mice, and endothelial cell damage in subcutaneous tumors and normal tissue (Bhuvaneswari *et al.*, 2009; Tseng *et al.*, 1988). Many reports cited above directly implicate the endothelium as a primary target for PDT *in vivo*; this stimulated research into the relative sensitivity of endothelial cells to PDT and the responses of endothelial cells that could initiate the various phenomena at the vessel level. Gomer *et al.* (1988) showed that bovine endothelial cells were significantly more sensitive to Photofrin-mediated PDT than smooth muscle cells or fibroblasts from the same species. This increased sensitivity to PDT, measured by clonogenic assay, was not a result of increased Photofrin accumulation. Sensitivity to HPD-mediated PDT of bovine aorta endothelial cells and human colon adenocarcinoma cells was investigated by West *et al.* (1990). Exponentially growing endothelial cells were significantly more sensitive than tumor cells with a similar rate of proliferation, and the difference in sensitivity was accompanied by greater PS accumulation in the endothelial cells.

Intravital microscopy of the rat cremaster muscle showed that Photofrin-PDT caused vessel constriction with an increase in vessel permeability and leukocyte adhesion was observed (Fingar *et al.*, 1992). The cyclooxygenase inhibitor indomethacin was able to inhibit the vessel constriction and reduce the increase in vascular permeability, implicating the involvement of the metabolites of cyclooxygenase called eicosanoids (such as thromboxane or prostacyclin). An increase in vascular permeability, thrombus formation and blood flow reduction were observed in animals treated with verteporfin-PDT in a dose- and time-dependent manner (Debefe *et al.*, 2010; Fingar *et al.*, 1999; He *et al.*, 2008). Blood vessels with lower flow rates were more sensitive to verteporfin-PDT-induced vascular shutdown than vessels with a higher flow rate. In agreement with light and electron microscopic studies, intravital fluorescence microscopic imaging demonstrated in live animals that verteporfin-PDT decreased endothelial cell viability and caused platelet aggregation (Khurana *et al.*, 2008). Rapid shutdown of tumor blood flow by formation of blood clots was also found after PDT with Tookad and MV6401, which was also confirmed by histological assessment (Dolmans *et al.*, 2002; Madar-Balakirski *et al.*, 2010). Many optical imaging studies have led to the conclusion that PDT is essentially a vascular disrupting therapy.

Immune stimulation effect of PDT: Innate immunity

PDT frequently provokes a strong acute inflammatory reaction in the targeted site, and this can often be observed as a fairly rapid induction of localized edema (Dougherty *et al.*, 1998). This reaction is a consequence of PDT-induced oxidative stress. Thus, PDT can be ranked among cancer therapies (including cryotherapy, hyperthermia, and focused ultrasound ablation) that all produce a chemical or physical destruction of tumor tissue that is perceived by the host as localized acute trauma. This tissue damage prompts the host to launch protective actions designed by evolution to react against threats to tissue integrity and to try to restore homeostasis at the affected site (Korbelik, 2006). The acute inflammatory response is the principal process engaged in this context. Its main task is containing the disruption of homeostasis, ensuring removal of damaged cells, and to protect against possible invasion of pathogenic microorganisms that may gain entry to the body as a result of a traumatic breach in external barriers. The inflammatory response also contributes to the local healing process with restoration of normal tissue function. The inflammation elicited by PDT is orchestrated by the innate immune system (Korbelik, 2006). The principal cells involved in this process consist of neutrophils, monocyte/macrophages, dendritic cells and mast cells. The recognition systems of the innate immune system are based upon a wide array of pattern-recognition receptors (PRRs), which are responsible for detecting the presence of tissue damage revealed to its sensors as the appearance of “altered-self” (Korbelik, 2006). PDT appears particularly effective in generating rapidly an abundance of alarm/danger signals, also called damage-associated molecular patterns (DAMPs) at the treated site that can be detected by the innate immune cells (Korbelik, 2006). Many DAMPs are nuclear or cytosolic proteins, or even small molecules that are normally safely contained inside healthy cells. When released outside from damaged or necrotic cells, or exposed on the surface of the cell following damage, they can encounter an oxidizing environment, which results in their changes in their conformation or chemical structure (Garg *et al.*, 2015). DNA can be released outside the nucleus from necrotic cells, and becomes a DAMP (Magna and Pisetsky, 2016). Other DAMPs (Garg *et al.*, 2015) include heat-shock proteins, high mobility group box protein 1 (HMGB1), hyaluronan fragments, extracellular ATP, uric acid, heparin sulfate and S100 molecules (members of a family of calcium modulated proteins).

Dendritic cells (DCs) are professional antigen presenting cells that have been described as the “orchestrators” of the immune response (Bhargava *et al.*, 2012). DCs express many receptors that specifically recognize DAMPs. Since PDT of tumor cells causes both cell death and cell stress (Dougherty *et al.*, 1998; Henderson and Gollnick, 2003; Oleinick and Evans, 1998) it is hypothesized that the activation of DCs by PDT-treated cells is the result of recognition of DAMPs released/secreted/exposed by PDT from dying cells (Gollnick *et al.*, 2006; Korbelik *et al.*, 2005, 2007). HSP70 is a well-characterized DAMP that interacts with the danger signal receptors, TLRs (Toll-like receptors) 2 and 4 (Vabulas *et al.*, 2002) and is induced by PDT (Gomer *et al.*, 1996). The level of expression of HSP70 in PDT-treated tumor cells appears to correlate with an ability to stimulate DC maturation (Gollnick *et al.*, 2004) and initiation of inflammation (Korbelik *et al.*, 2005; Stott and Korbelik, 2007).

The onset of PDT-induced inflammation is marked by dramatic changes in the tumor vasculature that becomes permeable to serum proteins and expresses adhesion molecules that circulating inflammatory cells can recognize and to which they can adhere (Korbelik, 2006). This process is responsible for the well-known “rolling” behavior of neutrophils and other leukocytes. The inflammatory process is predominantly initiated by signals originating from photooxidative damage produced in perivascular regions of the tumor with chemotactic gradients reaching out from the damaged endothelium. The inflammatory cells, initially consisting of neutrophils, which are then followed by mast cells and monocytes/macrophages, rapidly and massively invade tumors after PDT (Fig. 7) (Dougherty *et al.*, 1998; Krosi *et al.*, 1995). The primary task of these infiltrating cells is to neutralize the source of DAMPs by removing debris containing injured and dead cells and oxidatively damaged tissue components. Damage and dysfunction of PDT-treated tumor vasculature frequently results in vascular occlusion that serves to “seal off” the damaged tumor tissue until it is removed by invading phagocytic cells (Korbelik, 2006). Depletion of neutrophils or inhibition of their activity after PDT was shown to diminish the therapeutic effect (de Vree *et al.*, 1997; Korbelik and Cecic, 1999, 2003; Kousis *et al.*, 2007). Among the cytokines involved in the regulation of the inflammatory process, the most important ones for the PDT response are IL-1 β and IL-6 (Gollnick *et al.*, 2003; Sun *et al.*, 2002). Blocking the function of various adhesion molecules was also shown to inhibit the PDT response (Gollnick *et al.*, 2003; Sun *et al.*, 2002). On the other hand, blocking anti-inflammatory cytokines such as IL-10 and TGF- β can significantly improve the cure rates after PDT (Korbelik, 2006).

Fig. 7 illustrates the stimulation of immune response after PDT.

Immune stimulation effect of PDT: Adaptive immunity

A wide range of pre-clinical animal studies and even some clinical studies have demonstrated that PDT can influence the adaptive immune response in disparate ways. Some PDT regimens result in potentiation of adaptive immunity, while others lead to immunosuppression. The precise reasons why PDT can lead to potentiation vs. suppression are unclear; however it appears as though the effect of PDT on the immune system is dependent upon the treatment regimen, as well as the area treated and the precise PS type (Hunt and Levy, 1998; Kousis *et al.*, 2007). PDT induced immune suppression is largely confined to cutaneous and transdermal PDT regimens involving large surface areas (Hunt and Levy, 1998; Yusuf *et al.*, 2008).

Broadly speaking PDT efficacy appears to be dependent upon the induction of adaptive anti-tumor immunity. Long-term tumor response and cures are diminished or absent in immunocompromised mice such as nude mice or SCID mice (Korbelik and Cecic, 1999; Korbelik *et al.*, 1996). Reconstitution of these animals by injecting bone marrow or T cells isolated from immunocompetent (normal) mice can result in increased PDT efficacy. Canti *et al.* (1994) were the first to show PDT-induced immune stimulation, demonstrating that cells isolated from tumor-draining lymph nodes of PDT-treated mice were able to confer tumor

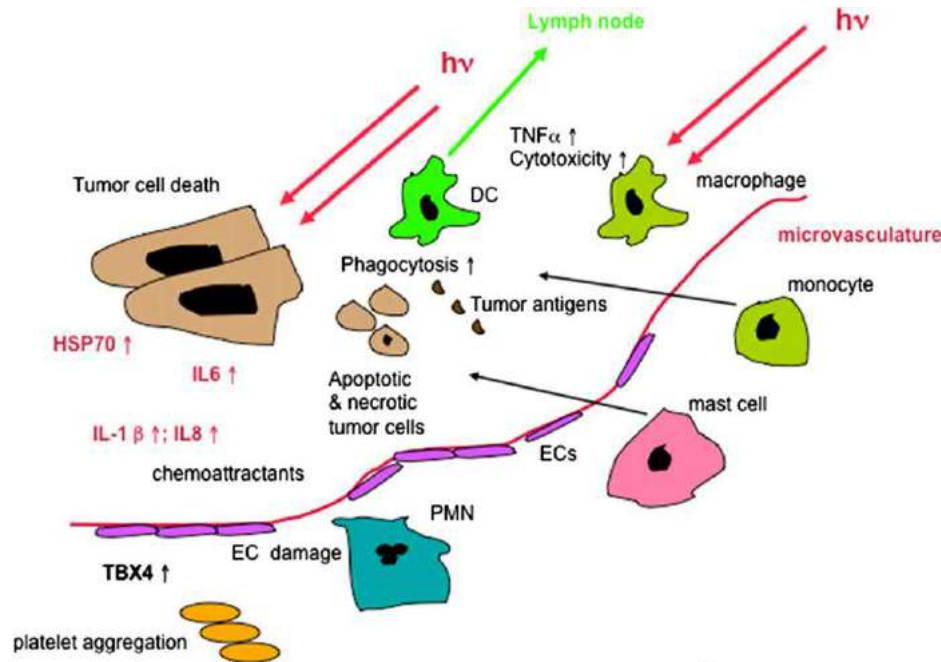


Fig. 7 Stimulation of immune response after PDT of tumors.

resistance when injected into naïve mice. Subsequent studies demonstrated that PDT directed against murine tumors resulted in the generation of immune memory such as rejection of a tumor rechallenge in mice that had obtained a long-term cure after PDT (Korbelik and Dougherty, 1999).

In order for the adaptive immune response produced by PDT to be highly effective, it appears that the tumor cell line employed should express what has been called a “tumor rejection antigen”. This has been demonstrated by testing PDT of syngeneic mouse tumors that have been engineered to express a model tumor antigen in immunocompetent mice. One of these model antigens was the bacterial protein beta-galactosidase (b-gal) (Mroz *et al.*, 2010). Although the b-gal tumors grew as rapidly as their wild-type counterparts, there was 100% long-term cures after BPD-mediated PDT with the b-gal tumors but no long-term cures with the wild-type tumors. The strength of the anti-tumor immune response was demonstrated by the fact that if mice had two b-gal tumors (one in each leg) and only one was treated with PDT, the contralateral untreated tumor underwent spontaneous regression in 70% of the mice. A similar result was seen in mice bearing the mastocytoma P815 that is well known to express the naturally-occurring tumor rejection antigen called P1A (Mroz *et al.*, 2013). The CD8 cytotoxic T-cells can recognize specific peptide epitopes derived from the amino acid sequence of the tumor antigen.

MHC-I is critical for allowing the CD8⁺ T cells to recognize the tumor antigen, and tumors that lack MHC-I are resistant to cell-mediated anti-tumor immune reactions (Maeurer *et al.*, 1996). Since MHC1 expression and expression of tumor associated antigens (such as P1A) are both susceptible to down-regulation by epigenetic silencing, an epigenetic reversal agent (the demethylating compound, 5-aza-2'-deoxycytidine) could potentiate the induction of anti-tumor immunity and increase cures in several mouse tumor models (Wachowska *et al.*, 2014).

The mechanism whereby PDT enhances anti-tumor immunity has been examined for the past several decades. PDT activates both humoral and cell-mediated anti-tumor immunity, although the importance of the humoral response (antibodies) is at present unclear. PDT efficacy in mice and humans is reduced in the absence of CD8⁺ T cell activation and/or tumor infiltration (Abdel-Hady *et al.*, 2001; Kabingu *et al.*, 2007; Korbelik and Cecic, 1999). Therefore most mechanistic studies have focused on the means by which PDT potentiates CD8⁺ T cell activation. It is clear that induction of anti-tumor immunity following PDT is dependent upon induction of inflammation (Henderson *et al.*, 2004). PDT-induced acute local and systemic inflammation is postulated to culminate in the maturation and activation of dendritic cells (DCs). Mature DCs are critical for activation of tumor specific CD8⁺ T cells and induction of anti-tumor immunity (Reis e Sousa, 2004). DCs are activated in response to PDT (Gollnick *et al.*, 2003) and migrate to tumor draining lymph nodes where they present the antigens they have taken up from damaged tumor cells to naïve T cells in the so-called “immune synapse” (Benvenuti, 2016). The T-cells in the lymph node become activated and proliferate dramatically and a “flood” of tumor-specific CTLs is released into the blood stream where they can search for other tumors cells to attack and destroy (Gollnick *et al.*, 2003; Sur *et al.*, 2008). Generation of CD8⁺ effector and memory T cells is frequently, but not always dependent upon the presence and activation of CD4⁺ T cells (Castellino and Germain, 2006). PDT induced anti-tumor immunity may (Korbelik and Cecic, 1999) or may not depend on CD4⁺ T cells (Kabingu *et al.*, 2007) and may be augmented by natural killer (NK) cells (Kabingu *et al.*, 2007).

PDT-mediated enhancement of anti-tumor immunity is believed to be due, at least in part, to stimulation of DCs by dead and dying tumor cells, suggesting that in vitro PDT-treated tumor cells may act as effective anti-tumor vaccines (Gollnick *et al.*, 2002).

This hypothesis has been proven by several studies using a wide variety of different PS and tumor models in both preventative and therapeutic settings (Gollnick *et al.*, 2002; Korbelik and Cecic, 2003; Korbelik *et al.*, 2009; Korbelik and Sun, 2006).

Mechanistic studies showed that incubation of immature DCs with PDT-treated tumor cells leads to enhanced DC maturation, activation and increased ability to stimulate T cells (Gollnick *et al.*, 2002; Jalili *et al.*, 2004). Furthermore, opsonization of PDT-treated tumor cells by complement proteins increases the efficacy of PDT-generated vaccines (Korbelik and Sun, 2006).

Induction of anti-tumor immune response after PDT

How does an effective anti-tumor immune response triggered after PDT manifest itself? This depends on whether pre-clinical animal tumor models (most often mice or rats) are being studied, or whether clinical studies are being considered. In an ideal scenario the immune stimulation would be strong enough for an animal or a patient with metastatic cancer who had the primary tumor treated with PDT, and then to have the distant untreated metastases regress over time as the host T-cells recognize specific tumor antigens and then attack and destroy their cognate target cells.

The implications of PDT-induced anti-tumor immunity and efficacious PDT-generated vaccines are significant and provide an exciting possibility for using PDT in the treatment of metastatic disease and as an adjuvant in combination with other cancer modalities. Several pre-clinical studies demonstrated that PDT is able to control the growth of tumors present outside the treatment field (Gomer *et al.*, 1987; Kabingu *et al.*, 2007) although others have failed to demonstrate control of distant disease following PDT (van Duijnhoven *et al.*, 2003).

This ideal scenario was demonstrated (Mroz *et al.*, 2010) in a mouse model where a BALB/c syngeneic colon carcinoma called CT26 was implanted subcutaneously in the leg and treated by a vascular PDT regimen mediated by liposomal verteporfin and 690 nm light. Although an apparently complete local response was achieved by application of PDT, the mice all suffered a local recurrence of the tumor in the leg and were sacrificed within a few weeks. By contrast, when the CT26 tumor was engineered to express a model tumor antigen (β -galactosidase) to form a tumor cell line called CT26. CL25, the tumors grew in the mice the same as CT26 wild-type. However CT26. CL25 tumors gave 100% permanent cures after PDT, and the cured mice were resistant to a rechallenge with CT26. CL25 tumor cells (but not to CT26 cells). When mice with two bilateral CT26. CL25 tumors (1 in each leg) had only the right-leg tumor treated with PDT, the distant untreated left-leg tumor spontaneously regressed over a period of several weeks. In about 70% the regression of both tumors was permanent, while in the remaining 30% of mice where the untreated CT26. CL25 tumor was observed to recur, it was found that the recurrent tumor has lost expression of the β -galactosidase antigen. Loss of antigen expression is a recognized route by which highly plastic tumors can evade the host immune response.

The efficacy of clinical PDT also appears to depend on induction of anti-tumor immunity. Patients with vulval intraepithelial neoplasia (VIN) who did not respond to ALA-PDT were more likely to have tumors that lacked major histocompatibility complex class I molecules (MHC-I) than patients who responded to ALA-PDT (Abdel-Hady *et al.*, 2001). VIN patients who responded to PDT had increased CD8⁺ T cell infiltration into the treated tumors as compared to non-responders. In other clinical studies patients who were treated with PDT for actinic keratoses and Bowen's disease were divided into immunocompetent and immunosuppressed sub-groups. Both sub-groups had similar initial response rates to PDT, but immunosuppressed patients had less long-term responses and were more likely to suffer appearance of new lesions (Dragieva *et al.*, 2004). In a recent clinical study, PDT of multifocal angiosarcoma of the head and neck resulted in increased immune cell infiltration into distant untreated tumors that was accompanied by tumor regression (Thong *et al.*, 2007, 2008). PDT was also shown to be an effective surgical adjuvant in non-small-cell lung cancer patients with pleural spread (Friedberg *et al.*, 2004). Recent reports have shown that clinical anti-tumor PDT also increases anti-tumor immunity. PDT of basal cell carcinoma (BCC) increased immune cell reactivity against a BCC-associated antigen (Kabingu *et al.*, 2009).

Photobiomodulation – The Healer

Light alone when it is delivered at a relatively low irradiance and fluence is called low level light (laser) therapy (LLLT) and has almost the complete opposite effect to PDT. The internationally accepted terminology for LLLT has recently been changed to "photobiomodulation" (PBM) therapy, due to widespread uncertainties about just exactly what parameters constituted "low level" (Anders *et al.*, 2015). LLLT (or PBM) was discovered in a rather serendipitous manner by Endre Mester in the 1960s. The discovery of the laser in 1960 by Theodore Maiman, had been recognized as a landmark in modern physics, and garnered an enormous amount of international attention (Maiman, 1960). In 1961 Leon Goldman started to experiment with the effect that these newly discovered laser beams had on the skin (Goldman *et al.*, 1963), and asked whether they could be used to perform "bloodless surgery". In the early 1960s Paul McGuff in Boston made medical history by using a laser beam to vaporize human cancer cells that had been transplanted into a hamster (McGuff *et al.*, 1965). In 1965 Mester started laser research in Hungary and tried to repeat McGuff's experiments by implanting tumor cells beneath the skin of laboratory rats and exposing them to the beam from a customized ruby laser. However the tumor cells were not destroyed by doses of what was presumed to be high-power laser energy, but instead, the skin incisions that had been made to implant the cancer cells appeared to heal faster in laser-treated animals, compared to incisions of control animals that were not treated with light (Mester *et al.*, 1968b). Moreover the regrowth of hair on the depilated rat skin, was observed to be faster after exposure to his ruby laser (Mester *et al.*, 1968c). After being initially puzzled by these contradictory findings, he realized that the output power of his custom-designed ruby laser was much weaker than he originally thought it to be, and instead of being photo-ablative against the tumor tissue, the low-power light stimulated the skin to

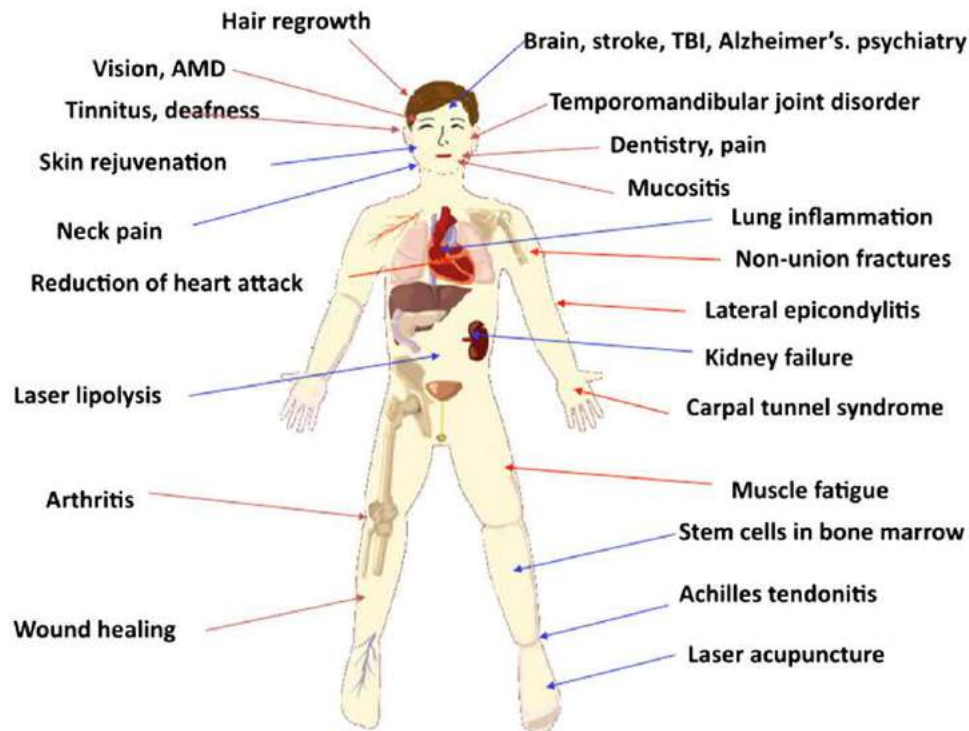


Fig. 8 Wide range of medical indications susceptible to treatment by PBM.

heal faster and caused the hair to regrow. Since those early days the range of medical indications that can be treated by LLLT or PBM has continued to expand almost exponentially as shown in [Fig. 8](#).

Mechanisms of Action

The first law of photobiology lays down that in order for a photon to have a biological effect it has to be absorbed by a specific photoacceptor molecule also known as a "chromophore". The leading hypothesis is that light is absorbed by chromophores in the mitochondria such as cytochrome c oxidase leading to production of more ATP and activation of a host of cell signaling pathways leading to gene transcription (Chen *et al.*, 2009). A secondary hypothesis is that longer wavelength light (in the infrared region) is absorbed by water in sensitive cellular locations such as ion channels (see [Fig. 9](#) for a graphical illustration). Instead of killing cells (as is the case with PDT) PBM stimulates cellular proliferation and prevents cells from dying (Huang *et al.*, 2009). The medical applications of PBM range from serious diseases such as stroke, heart attack, spinal cord injury, traumatic brain injury to less serious indications such as wound healing, reduction of pain and inflammation and depression (Hashmi *et al.*, 2010).

Cytochrome C oxidase

Cytochrome C oxidase (Cox) is the terminal enzyme of the electron transport chain, mediating the electron transfer from cytochrome c to molecular oxygen. Several lines of evidence show that Cox acts as a photoacceptor and transducer of photsignals in the red and near-infrared regions of the light spectrum (Anneroth *et al.*, 1988). It seems that PBM increases the availability of electrons for the reduction of molecular oxygen in the catalytic center of Cox, increasing the mitochondrial membrane potential (MMP) and the levels of ATP, cAMP and ROS as well (Barcia, 2012).

PBM increases the activity of complexes I, II, III, IV and succinate dehydrogenase in the electron transfer chain. Cox is known as complex IV and, as mentioned before, appears to be the primary photoacceptor. This assumption is supported by the increased oxygen consumption during low-level light irradiation (the majority of the oxygen consumption of a cell occurs at complex IV in the mitochondria), and by the fact that sodium azide, a Cox inhibitor, prevents the beneficial effect of PBM. Besides ATP and cAMP, nitric oxide (NO) level is increased, either by release from metal complexes in Cox (Cox has two heme and two copper centers) or by up-regulation of the activity of Cox as a nitrite reductase (Barolet *et al.*, 2009).

Retrograde mitochondrial signaling

One of the most accepted mechanisms for light-cell interaction was proposed by Karu (Barrett and Gonzalez-Lima, 2013), referring to the retrograde mitochondrial signaling that occurs with light activation in the visible and infrared range. According to Karu, the first step is the absorption of a photon with energy $h(\nu)$ by the chromophore Cox. This interaction increases

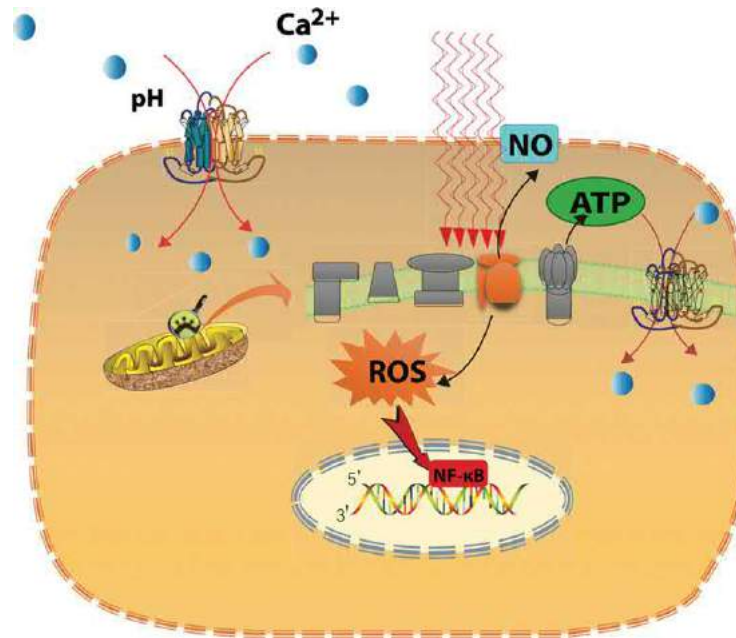


Fig. 9 Main mechanisms of action of PBM at a cellular level.

mitochondrial membrane potential ($\Delta\Psi_m$), causing an increase in the synthesis of ATP and changes in the concentrations of reactive oxygen species (ROS), Ca^{2+} and NO. Furthermore, there is a communication between mitochondria and the nucleus, driven by changes in the mitochondria ultrastructure, i.e. changes in the fission-fusion homeostasis in a dynamic mitochondrial network. The alteration in the mitochondrial ultrastructure induces changes in ATP synthesis, in the intracellular redox potential, in the pH and in cyclic adenosine monophosphate (cAMP) levels. Activator protein-1 (AP1) and NF- κ B have their activities altered by changes in membrane permeability and ion flux at the cell membrane. Some complementary routes were also suggested by Karu, such as the direct upregulation of some genes (Bellnier *et al.*, 1997).

Light/heat sensitive ion channels

The most well-known ion channels that can be directly gated by light are the channelrhodopsins (ChRs), which are seven-transmembrane-domain proteins that can be naturally found in algae providing them with light perception. Once activated by light, these cation channels open and depolarize the membrane. They are currently being applied in neuroscientific research in the new discipline of optogenetics (Bellnier *et al.*, 1989).

However members of another broad group of ion-channels are now known to be more or less light sensitive (Benvenuti, 2016). These channels are called “transient receptor potential” (TRP) channels as they were first discovered in a *Drosophila* mutant (Benvenuti, 2016). TRP channels are responsible for vision in insects. There are now known to be at least 50 different TRP isoforms distributed amongst seven subfamilies (Bernstein, 2005). These are called the TRPC (‘Canonical’) subfamily, the TRPV (‘Vanilloid’), the TRPM (‘Melastatin’), the TRPP (‘Polycystin’), the TRPML (‘Mucolipin’), the TRPA (‘Ankyrin’) and the TRPN (‘NOMPC’) subfamilies. A wide range of stimuli modulate the activity of different TRP such as light, heat, cold, sound, noxious chemicals, mechanical forces, hormones, neurotransmitters, spices, and voltage. TRP are calcium channels modulated by phosphoinositides (Bertoloni *et al.*, 1990).

The evidence that light mediated activation of TRP is responsible for some of the mechanisms of action of PBM is somewhat sparse at present, but is slowly mounting.

Mast cells are known to accumulate at the site of skin wounds, and there is some degree of evidence suggesting that these cells play a role in the biological effects of laser irradiation on promoting wound healing. Yang and co-workers demonstrated that after laser irradiation (532 nm), the intracellular $[\text{Ca}^{2+}]$ was increased and, as a consequence, there was a release of histamine. If the TRPV4 inhibitor, ruthenium red, was used, the histamine release was blocked, indicating the central role of these channels in promoting histamine-dependent wound healing after laser irradiation (Bhargava *et al.*, 2012).

It seems that TRPV1 ion channels are involved in the degranulation of mast cells and laser-induced mast cell activation. It was demonstrated that capsaicin, temperatures above 42°C and acidic pH could induce the expression of TRPV1 in oocytes, and these ion channels can be activated by green light (532 nm) in a power-dependent manner, although blue and red light were not able to activate them (Bhat *et al.*, 2005). Infrared light (2780 nm) attenuates TRPV1 activation by capsaicin in cultured neurons, decreasing the generation of pain stimuli. TRPV4 is also attenuated by laser light, but the antinociceptive effect was less intense, therefore the antinociception in this model is mainly dependent on TRPV1 inhibition (Bhuvanawari

et al., 2009) The stimulation of neurons with pulsed infrared light (1875 nm) is able to generate laser-evoked neuronal voltage variations and, in this case, TRPV4 channels were demonstrated to be the primary effectors of the chain reaction activated by the laser (Bisht *et al.*, 1994).

Flavins

Rachel Lubart showed that flavins were able to absorb light in the 400–500 nm wavelength range (Eichler *et al.*, 2005). This led to the production of ROS inside cells. Since ROS appear to be produced by different wavelengths of light that target different chromophores, it is probably a general mechanism involved in PBM effects (Lavi *et al.*, 2010).

Direct cell-free light-mediated effects on molecules

There have been some scattered reports that light can have effects on some important molecules in cell free systems (in addition to the established effect on Cox). The latent form of transforming growth factor beta has been reported to be activated by light exposure (Bisht *et al.*, 1999). Superoxide dismutase (Cu,Zn) from bovine erythrocytes that had been inactivated by exposure to pH 5.9 was reactivated by exposure to He-Ne laser light (632.7 nm) (Bjordan *et al.*, 2003). The same treatment also reactivated the heme-containing catalase. Amat *et al.* showed that irradiation of ATP in solution by 655 nm or 830 nm light appeared to produce changes in its enzyme reactivity, fluorescence and Mg²⁺ binding capacity (Bjornsti and Houghton, 2004). However other workers were unable to repeat this somewhat surprising result (Blanco *et al.*, 2015).

PBM for Healing of Wounds and Tissue Damage

The first demonstration the clinical effectiveness of PBM was in non-healing wounds in patients in the early 1970s (Mester *et al.*, 1972). Since then many authors have published studies in both animal models and clinical trials showing positive effects. Our study in excisional wounds in mice showed that certain wavelengths (635 nm and 810 nm) were better than others and moreover there was a biphasic dose response with an optimum fluence of 2 J/cm² (Demidova-Rice *et al.*, 2007).

Wound healing was one of the first applications of PBM, when HeNe lasers were used by Mester *et al.* to treat skin ulcers (Mester *et al.*, 1971, 1972, 1976, 1968b). PBM is believed to affect all three phases of wound healing (Thomas *et al.*, 1995): the inflammatory phase, in which immune cells migrate to the wound, the proliferative phase, which results in increased production of fibroblasts and macrophages, and the remodeling phase, in which collagen deposition occurs at the wound site and the extracellular matrix is rebuilt.

PBM is believed to promote wound healing by inducing the local release of cytokines, chemokines, and other biological response modifiers that reduce the time required for wound closure, and increase the mean breaking strength of the wound (Bisht *et al.*, 1999; Hawkins *et al.*, 2005; Meyers, 1990). Proponents of PBM speculate that this result is achieved by increasing the production and activity of fibroblasts and macrophages, improving the mobility of leukocytes, promoting collagen formation, and inducing neovascularization (Hawkins and Abrahamse, 2005; Loevschall and Arenholt-Bindslev, 1994; Medrado *et al.*, 2003; Noble *et al.*, 1992; Reddy *et al.*, 2001; Skinner *et al.*, 1996).

However, there is a lack of convincing clinical studies that either prove or disprove the efficacy of PBM in wound healing. The results that are currently available are conflicting and do not lead to any clear conclusions. For example, Abergel *et al.* found that the 632.8 nm HeNe laser did not have any effect on the cellular proliferation of fibroblasts, while the 904 nm GaAs laser actually lowered fibroblasts proliferation (Abergel *et al.*, 1987a). In contrast, other studies noted an increase in proliferation of human fibroblasts exposed to 904 nm GaAs lasers (Pereira *et al.*, 2002), rat myofibroblasts exposed to 670 nm GaAs lasers (Medrado *et al.*, 2003), and gingival fibroblasts exposed to diode lasers (670, 692, 780, and 786 nm) (Almeida-Lopes *et al.*, 2001). *In vivo* studies in both animal and human models show similar discrepancies. A study by Kana *et al.* claimed that treatment of open wounds in rats with HeNe and argon lasers resulted in faster wound closure (Kana *et al.*, 1981). Bisht *et al.* found a similar increase in granulation tissue and collagen expression in rats using the same treatment as Kana (Bisht *et al.*, 1994). However, Anneroth *et al.* failed to observe any beneficial effects after laser treatment in a comparable rat model (Anneroth *et al.*, 1988). In human studies, Schindl *et al.* reported that application of a HeNe laser was beneficial in promoting wound healing in 3 patients (Schindl *et al.*, 2000), whereas Lundeberg *et al.* found no statistically significant difference between leg ulcer patients treated with an HeNe laser and those treated with a placebo (Lundeberg and Malm, 1991).

The scarcity of well-designed clinical trials makes it difficult to assess the impact of PBM on wound healing. Our task is further complicated by the difficulty in comparing studies, due to the large number of factors involved. In addition to the multiple parameters that must be adjusted in order to apply PBM, such as the wavelength and power of the light, the effectiveness of the treatment also depends on many factors such as the location and nature of the wound, and the physiologic state of the patient. For example, impaired wound healing is one of the major chronic complications of diabetes (Goodson and Hunt, 1979; Raskin *et al.*, 1975), and is thought to result from various factors, including decreased collagen production and impaired functionality of fibroblasts, leukocytes, and endothelial cells (Goodson and Hunt, 1979; Spanheimer *et al.*, 1988). It has therefore been hypothesized that PBM could have beneficial effects in stimulating wound healing in diabetic patients (Schindl *et al.*, 1998, 2002; Yu *et al.*, 1997). Thus, in order to obtain a convincing verdict on the impact of PBM on wound healing, we will require several large, randomized, placebo controlled, and double blind trials that compare the effects of PBM on wounds that are as similar as possible. A greater understanding of the cellular and biochemical mechanisms of PBM would

also be useful in assessing these studies, as it would enable us to pinpoint exactly what criteria to use in determining the effectiveness of the therapy.

PBM for Pain and Inflammation

PBM can be used with great benefit in numerous inflammatory disorders such as arthritis, carpal tunnel syndrome, and tendinitis. We showed that 810-nm laser was very effective in a rat model of inflammatory arthritis caused by intra-articular injection of zymosan (Castano *et al.*, 2007). A sufficiently long illumination time was more important than total fluence or the irradiance in determining the benefit of the therapy.

There appears to be more firm evidence to support the success of PBM in alleviating pain and treating chronic joint disorders, than in healing wounds. A review of 16 randomized clinical trials including a total of 820 patients found that PBM reduces acute neck pain immediately after treatment, and up to 22 weeks after completion of treatment in patients with chronic neck pain (Chow *et al.*, 2009). PBM has also been shown to relieve pain due to cervical dentinal hypersensitivity (Sandford and Walsh, 1994), or from periodontal pain during orthodontic tooth movement (Wahl and Bastanier, 1991). A study of 88 randomized controlled trials indicated that PBM can significantly reduce pain and improve health in chronic joint disorders such as osteoarthritis, patellofemoral pain syndrome, and mechanical spine disorders (Bjordan *et al.*, 2003). However, the authors of the study urge caution in interpreting the results due to the wide range of patients, treatments, and trial designs involved.

PBM for Cosmetic and Esthetic Applications

Skin rejuvenation

PBM is a novel treatment option for non-thermal and non-ablative skin rejuvenation which has been shown to be effective for improving skin conditions such as wrinkles and skin laxity (Barolet *et al.*, 2009; Bhat *et al.*, 2005; Dierickx and Anderson, 2005; Russell *et al.*, 2005; Weiss *et al.*, 2004a,b). A wide range of different light sources have been used to deliver light for these treatments, particularly to the face, and some are shown in Fig. 10. PBM is a treatment modality that has been shown to provide increased rates of skin rejuvenation and wound healing with great efficacy, while also reducing post-operative pain, edema and several types of inflammation, making it highly desirable modality (Kim and Calderhead, 2011; R.G. *et al.*, 2008). Early studies by Abergel *et al.* and Yu *et al.* reported an increase in production of pro-collagen, collagen, basic fibroblast growth factors (bFGF) and proliferation of fibroblasts after exposure to low-energy laser irradiation in vitro and in vivo animal models (Abergel *et al.*, 1987b; Yu *et al.*, 1994). Use of light sources with wavelengths of 633 nm/830 nm is most common in cases of clinical application involving wound healing and skin rejuvenation. PBM is now used for the treatment of even chronic non-healing wounds through restoration of imbalances in collagenesis/collagenase and allows for rapid and enhanced wound healing in general (Kim and Calderhead, 2011). Lee *et al.* conducted a study to investigate the histological and ultrastructural alterations that followed a series of light treatments utilizing light emitting diodes (LEDs) with parameters of 830 nm, 55 mW/cm², 66 J/cm² and 633 nm, 105 mW/cm², 126 J/cm². They observed alteration in the levels of MMPs and tissue inhibitors of metalloproteinases (TIMPs)



Fig. 10 Selection of light sources used for PBMT of skin conditions.

(Lee *et al.*, 2007). The study also showed increased mRNA levels of interleukin-1 beta (IL-1 β), tumor necrosis factor alpha (TNF- α), intercellular adhesion molecule 1 (ICAM-1), and connexin 43 (C $\times$ 43) following LED phototherapy whereas IL-6 levels were decreased (Lee *et al.*, 2007). Subsequently the study also demonstrated a well-marked increase in the amount of collagen in the post-treatment specimens (Lee *et al.*, 2007). In fractional laser resurfacing it is thought that the deliberate development of microscopic photothermally-mediated wounds is responsible for the recruitment of pro-inflammatory cytokines IL-1 β and TNF- α *in order cause wound repair*. The generation of such a wound healing cascade thus contributes to new collagen synthesis (Lee *et al.*, 2007). PBM may also induce this wound healing process through athermal and atraumatic induction of a subclinical 'quasi-wound', even without any actual wounding created by thermal damage which can possibly cause complications as in some other laser treatments (Lee *et al.*, 2007). MMP activities are known to be inhibited by TIMPs, suggesting the possibility of other mechanisms for increased collagen synthesis through induction of TIMPs.

Collectively viewing these findings, they are suggestive of the idea that an increased production of IL-1 β and TNF- α might be responsible for induction of MMP activity as an early response to light treatment, which might possibly contribute to the removal of photodamaged collagen fragments in order to facilitate new collagen biosynthesis. Furthermore, as a consequence of the therapy there may be increased concentrations of TIMPs that most likely play a role in the protection of the newly synthesized collagen, from proteolytic degradation by MMPs (Lee *et al.*, 2007). Subsequently, heightened expression of C $\times$ 43 may possibly enhance cell-cell communication between dermal components, especially between fibroblasts, allowing for greater synchrony between cellular responses, following the effects of PBM in order to promote synthesis of new collagen in a greater area including even the regions that did not receive light irradiation (Lee *et al.*, 2007).

A clinical study conducted by Weiss *et al.* demonstrated the benefits of PBM over traditional thermal-based rejuvenation modalities. A group of 300 patients were administered PBM (590 nm, 0.10 J/cm²) alone, and another group of 600 patients received a combination of PBM with a thermal-based photorejuvenation procedure. Of the patients who received just the light treatment, 90% reported an observed softening of skin textures as well as a reduction in skin coarseness and fine lines that ranged from small alterations to significant changes (Weiss *et al.*, 2005b). It was observed that patients who received a form of PBM (n=152) reported a noticeable reduction in post-treatment erythema and an overall impression of increased efficacy versus patients that received treatment through a thermal photorejuvenation laser or light source lacking any sort of PBM photomodulation (Kucuk *et al.*, 2010; Weiss *et al.*, 2005b). Reduction in post-treatment erythema can most likely be attributed to the anti-inflammatory effects of PBM (Barolet *et al.*, 2009). Utilizing different pulse sequence parameters, a multicenter clinical trial was conducted, wherein 90 patients received 8 PBM treatments over 4 weeks (Geronemus *et al.*, 2003; McDaniel *et al.*, 2003; Weiss *et al.*, 2004a,b, 2005a). The study found good overall results with more than 90% of patients improving by at least one Fitzpatrick photoaging category and 65% of the patients displaying global improvement in facial texture, fine lines, background erythema and pigmentation with results peaking at 4 to 6 months following completion of the 8 treatments. A noticeable increase in papillary dermal collagen and reductions in MMP-1 were generally observed. A study conducted by Barolet *et al.* also proved to be consistent with the aforementioned studies. The study used a 3-D model of tissue-engineered Human Reconstructed Skin (HRS) to investigate the potential of PBM (660 nm, 50 mW/cm², 4 J/cm²) in collagen and MMP-1 modulation. The results showed up-regulation of collagen and down-regulation MMP-1 *in vitro* (Barolet *et al.*, 2009). A split-face, single-blinded clinical study was then carried out to assess the results of this light treatment on skin texture and appearance of individuals with aged/photoaged skin (Barolet *et al.*, 2009). Profilometry quantification (to measure wrinkles) demonstrated that more than 90% of individuals had a reduction in rhytid depth and surface roughness, and, 87% of the individuals reported that they have experienced a reduction in the Fitzpatrick wrinkling severity score following 12 PBM treatments (Barolet *et al.*, 2009).

PBM for hair regrowth

As mentioned above in the late 1960s, Endre Mester, a Hungarian physician, conducted a series of experiments to investigate the anti-cancer effect of lasers using a low-powered ruby laser (694 nm) on mice and much to his surprise, not only did laser exposure not cure cancer instead it seemed to enhance hair growth in the shaved area of the mice that had been set up for the original experiment (Mester *et al.*, 1968a). Quite recently, lasers have garnered much attention due to their remarkable ability to cause selective hair removal, however in some cases it was observed that lasers in fact produced undesired hair growth effects such as increase in hair density, color or coarseness or a combination of any of these, in areas of the skin that were treated for hair removal (Moreno-Arias *et al.*, 2002a,b; Vlachos and Kontoes, 2002; Wikramanayake *et al.*, 2012a). This phenomenon is known as 'paradoxical hypertrichosis' and its incidence varies from 0.6% to 10% (Wikramanayake *et al.*, 2012a). A group of researchers also observed that low-powered laser irradiation of small vellus hairs cause them to transform into larger terminal hairs, and thus termed this process as "terminalization" of vellus hair follicles (Bernstein, 2005; Bouzari and Firooz, 2006). Different mechanisms have been proposed by dermatologists (who did not not really believe in PBM) in an attempt to explain these effects of lasers on hair removal. In one particular study this ability was attributed to the fact that polycystic ovarian syndrome was present in 5 out of 49 females undergoing intense pulsed laser (IPL) laser treatment for facial hirsutism (Moreno-Arias *et al.*, 2002a). Another study suggested that despite the heat generating effects of laser treatments, the heat produced was not sufficient enough to induce total hair follicle thermolysis but, it may be sufficient to stimulate follicular stem cell proliferation and differentiation by increasing levels of heat shock proteins (HSPs) such as HSP27 which influence regulation of cell growth and differentiation (Wikramanayake *et al.*, 2012a).

In 2007, the FDA approved PBM as a possible treatment modality for hair loss (Wikramanayake *et al.*, 2012a). It is believed that non-thermal PBM can stimulate reentry of anagen hair follicles into the telogen stage, bringing about greater rates of proliferation

in active anagen follicles, can prevent development of premature catagen stage follicles and extend the duration of the anagen phase (Leavitt *et al.*, 2009; Wikramanayake *et al.*, 2012a).

In one study conducted by Yamazaki and colleagues, irradiated the backs of Sprague Dawley rats using a linearly polarized IR source, where an up-regulation of hepatocyte growth factor (HGF) and HGF activator expression was discovered (Miura *et al.*, 1999). Another study reported increases in temperature of the skin and improved blood flow around the stellate ganglion area, following treatment with PBM (Wajima *et al.*, 1996). The exact mechanism for the action of minoxidil in treating hair loss is not entirely understood, but it is known that minoxidil does contain nitric oxide (NO) which is an important cellular signaling molecule and vasodilator (Proctor, 1989) that is influential to a variety of physiological and pathological processes (Hou *et al.*, 1999). Furthermore, NO is a regulator of the opening of ATP-dependent potassium (K^+) channels and is thus responsible for the hyperpolarization of cell membranes (Rossi *et al.*, 2012). Also, it has been suggested that ATP sensitive K^+ channels of the mitochondria and elevated levels of NO might be involved in the mechanism of action of PBM (Karu, 2008; Karu *et al.*, 2005; Tuby *et al.*, 2006) in areas of the brain and heart (Karu, 2008; Karu *et al.*, 2004; Y.D. Ignatov, 2005). Thus given the dependency of both minoxidil and PBM on the aforementioned factors there is possibly some mechanistic overlap between the two modalities. A study conducted by Weiss *et al.* demonstrated that PBM was able to modulate 5- α reductase, the enzyme responsible for the conversion of testosterone to DHT, as well as alter the genetic expression of vascular endothelial growth factor (VEGF), which plays an influential role in hair follicle growth and thus PBM is able to stimulate hair growth (Castex-Rizzi *et al.*, 2002; Weiss *et al.*, 2005a,b; Yano and Detmar, 2001). Furthermore, it has been demonstrated that PBM may stimulate hair growth through modulation of inflammatory processes and immunological responses (Meneguzzo *et al.*, 2012). A study conducted by Wikramanayake *et al.* on C3H/HeJ AA mouse models supported this assumption, wherein the mice were exposed to a laser-comb and it was observed that the treatment led to an increase in the quantity of hair follicles where the majority of the follicles in anagen phase were seen to have decreased inflammatory infiltrates (Wikramanayake *et al.*, 2012a). Taking into account the disruptive effect that inflammatory infiltrates have on hair follicles along with the notion that several cytokines such as interferon gamma (IFN- γ), IL-1 α and β , TNF- α , MHC and Fas-antigen and macrophage migration inhibitory factor are all involved in cyclic hair growth as well as the pathogenesis of AA, PBM may be able to play significant role in the treatment of AA due to its modulating effects on inflammation (Wikramanayake *et al.*, 2012a).

A clinical study was carried out to investigate the effect of PBM on treatment of alopecia areata consisting of a sample size of 15 patients (6 men, 9 women) utilizing Super Lizer TM, a medical instrument operating on polarized linear light with a high output (1.8 W) of NIR radiation (600–1600 nm) possessing sufficient penetration depth to reach deep subcutaneous tissue (Yamazaki *et al.*, 2003). The patients received a 3 min laser treatment on the scalp either once a week or once every 2 weeks and were administered additional carpronium chloride 5% twice daily to all lesions (Yamazaki *et al.*, 2003). Supplemental oral antihistamines, cepharanthin and glycyrrhizin (extracts of medicinal Chinese herbs) were prescribed as well (Yamazaki *et al.*, 2003). The results of the study showed that 47% of the patients experienced hair growth 1.6 months earlier on areas irradiated with laser when compared to the areas that were not irradiated (Yamazaki *et al.*, 2003).

In another study 24 male androgenetic alopecia (AGA) patients were evaluated via global photography and phototrichogram using 655 nm red light and 780 nm IR light once a day for a period of 10 min (Kim *et al.*, 2007). Following 14 weeks of treatment significant increases in hair density and anagen/telogen ratio were observed at both the vertex and occiput, with 83% patients reporting that the treatment resulted in satisfactory results (Kim *et al.*, 2007).

Satino *et al.* conducted a study to investigate the efficacy of PBM on hair growth and tensile strength involving 28 male and 7 female AGA patients (Satino, 2003). Each patient was given a 655 nm HairMax LaserComb® to use at home for a period of 6 months, applying it for five to ten minutes per day on alternate days (Satino, 2003). Regarding the tensile strength of the hair, the results showed improvements in hair growth in all treated areas for both male and female sexes, however in the case of males the greatest improvements were observed in the vertex area whereas for females, the best improvements were seen in the temporal area (Satino, 2003). With regards to hair count, again all treated areas of both sexes showed improvement, but the vertex area showed the greatest improvement for the male patients (Satino, 2003). Leavitt *et al.* conducted a double-blind, sham device-controlled, multi-center, randomized 26 week trial where they tested the same device on 110 male AGA patients (Leavitt *et al.*, 2009). The patients were made to use the device three times a week for fifteen minutes for a total period of 26 weeks (Leavitt *et al.*, 2009). Noticeable increases in mean terminal density of hair were reported in the treatment group when compared to the sham treatment group (Leavitt *et al.*, 2009). Also, subjective assessment of the patients over the 26 week period reported prominent improvements in overall hair regrowth, decreased rate of hair loss, thicker feeling hair, scalp health and hair shine (Leavitt *et al.*, 2009).

Around 65% of the patients receiving chemotherapy for cancer develop alopecia which can have detrimental effects on psychological health of the patient (Trueb, 2009). It has been proposed that PBM could serve as treatment modality to stimulate and promote hair growth in cases of chemotherapy induced alopecia. In one study, a rat model was given varying regimens of chemotherapy in conjunction with PBM administered with a device possessing components (laser unit and switch, lacking comb or handle) of the Hair Max LaserComb (Wikramanayake *et al.*, 2012b). In all rats that were given laser treatment, hair regrowth occurred at a faster rate when compared to the sham treatment group. Additionally, PBM did not hinder the efficacy of the chemotherapeutic procedures (Wikramanayake *et al.*, 2012b).

PBM for fat reduction and cellulite treatment

In 2000 Neira *et al.* demonstrated the use of low level laser as a new alternative to liposuction, and successfully utilized it with doses that did not produce any detectable increases in tissue temperature or cause any noticeable macroscopic alterations in the tissue

structure (Neira *et al.*, 2000, 2002). Prior investigations concerned with the effects of LLLT on wound healing, pain relief and edema prevention paved the way for this therapeutic application (GD *et al.*, 1991; PR, 1989). The development of LLLT as a therapeutic modality to augment liposuction while avoiding macroscopic tissue alterations were based on determination of optimal parameters such as wavelength and power output for use (J.L., 2000). Evidence suggests that the best wavelengths suitable for biomodulation range between 630 and 640 nm (FAH. and XY., 1996; Frohlich, 1970, 1975; Fàà, 1968; Sroka *et al.*, 1997; van Breugel and Bar, 1992). Nierra *et al.* made several intriguing observations regarding the effects of LLLT on adipocytes. They utilized low level diode laser (635 nm) and a maximal power of 10 mW with energy values ranging from 1.2 to 3.6 J/cm² (Neira *et al.*, 2002). Using scanning electron microscopy (SEM) and transmission electron microscopy (TEM) it was demonstrated that adipocyte plasma membranes exhibited transitory pore formation as result of irradiation. It was formulated that this enabled the release of intracellular lipids from the adipocytes and thus supplemented the liposuction as it was expected to reduce the time taken for the procedure, allowed for extraction of greater volumes of fat and overall, reduced the energy expenditure of the surgeon.

Although the findings associated with LLLT enjoyed much praise and enthusiasm, an extensively study conducted put these findings surrounding LLLT into question (Brown *et al.*, 2004). In their study, cultured human preadipocytes did not show any differences when compared to non-irradiated cells after 60 min of irradiation using an LLLT source (635 nm and fluence 1 J/cm²) (Brown *et al.*, 2004). Furthermore, histological examination of lipoaspirates, in a porcine model exposed to LLLT for 30 min and human lipoaspirates, failed to demonstrate transitory pores when analyzed using SEM (Brown *et al.*, 2004). Additional data, raised questions regarding the ability of red light (635 nm) to effectively penetrate below the skin, into the sub-dermal tissues (Kolari and Airaksinen, 1993). Peter Foddor supportively stated: "One could postulate that the presence of the black dots on scanning electron microscopy images on the surface of fat cells reported by Neira *et al.* could represent an artifact." (Brown *et al.*, 2004). Since the data reported by Brown *et al.* from 2004, there have been several publications reporting the efficacy of LLLT.

In the original paper published by Neira *et al.* the fat liberating effects of LLLT on adipocytes were attributed to its ability to induce transitory micropores which were visualized with the help of SEM (Neira *et al.*, 2002). Furthermore, it was postulated that this stimulated the release of intracellular lipids from the adipocytes. Based on this, it was formulated that up to 99% of the fat stored within the adipocytes could be released and subsequently removed with the help of LLLT (635nm, 10 mW intensity, 6 min irradiation time) (Neira *et al.*, 2002). Re-cultured adipocytes exhibited a tendency to attain their native cellular conformation which was further confirmed by Caruso-Davis *et al.* utilizing a live-dead assay to assess the viability of these adipocytes following irradiation (Caruso-Davis *et al.*, 2011). Increase in ROS following LLLT has been proposed to bring about lipid peroxidation within the cell membrane which may cause damage and may present as the transitory pores. This may cause temporary damage that presents as transitory micropores. (Chen *et al.*, 2011; Geiger *et al.*, 1995; Karu, 2008; Tafur and Mills, 2008). However, when Brown *et al.* attempted to replicate Neira *et al.*'s findings (Neira *et al.*, 2002), they failed to visualize any transitory micropores via SEM (Brown *et al.*, 2004). No further SEM studies have documented these pores, but many publications have reported findings that indirectly support the transitory micropore formation theory. Another proposed mechanism that explains the release of intracellular lipids from adipocytes, suggests the activation of the complement cascade which is responsible for induction of adipocyte apoptosis and subsequent release of intracellular lipid components (Caruso-Davis *et al.*, 2011). To test the feasibility of this theory Caruso-Davis *et al.* exposed differentiated human adipocytes to plasma and exposed one group of cells to laser, while the control group received no laser intervention (Caruso-Davis *et al.*, 2011). No differences were noticed with regard to complement system components in either group and it was concluded that the LLLT did not act through activation of the complement cascade (Caruso-Davis *et al.*, 2011).

PBM for Vision and Hearing Problems

PBM has been used to treat disorders of sensory systems such as vision and hearing. Light can easily be shone into the eyes, and its known ability to heal and preserve cells and restore function means that PBM works well in the retina (Eells *et al.*, 2016; Geneva, 2016). It should be remembered that the retina is one of the most metabolically active tissue in the body, with a high demand for oxygen and glucose (Sanchez-Chavez *et al.*, 2012). Most studies of PBM for ophthalmology appear to have used red light (more specifically 670 nm from LEDs). PBM has been reported to be beneficial (in patients as well as in animal models) for age-related macular degeneration (Ivandic and Ivandic, 2008), diabetic retinopathy (Tang *et al.*, 2013, 2014), retinopathy of prematurity (Natoli *et al.*, 2013), light-induced photoreceptor degeneration (Albarracin *et al.*, 2011; Qu *et al.*, 2010), amblyopia (Ivandic and Ivandic, 2012), retinitis pigmentosa (Ivandic and Ivandic, 2014), and blindness caused by methanol toxicity (Eells *et al.*, 2003).

In a similar approach to shining light in the eyes for problems with vision, investigators have studied the effect of shining light in the ears to treat disorders of hearing. In this case however there is no ideal clear optical path for the light to get to the intended target cells in the inner ear. PBM has been studied for noise-induced hearing loss (Lee *et al.*, 2016; Tamura *et al.*, 2015), chronic tinnitus and sensorineural hearing loss (Tauber *et al.*, 2003), antibiotic-induced ototoxicity (Rhee *et al.*, 2013) and Meniere's disease (Teggi *et al.*, 2008).

PBM for Stroke

There are a plethora of plausible mechanisms (shown in Fig. 11) that suggest the beneficial effects of PBM on the brain will be pronounced, and this prediction has largely been proved to be true. Perhaps the most well-investigated application of PBM to the

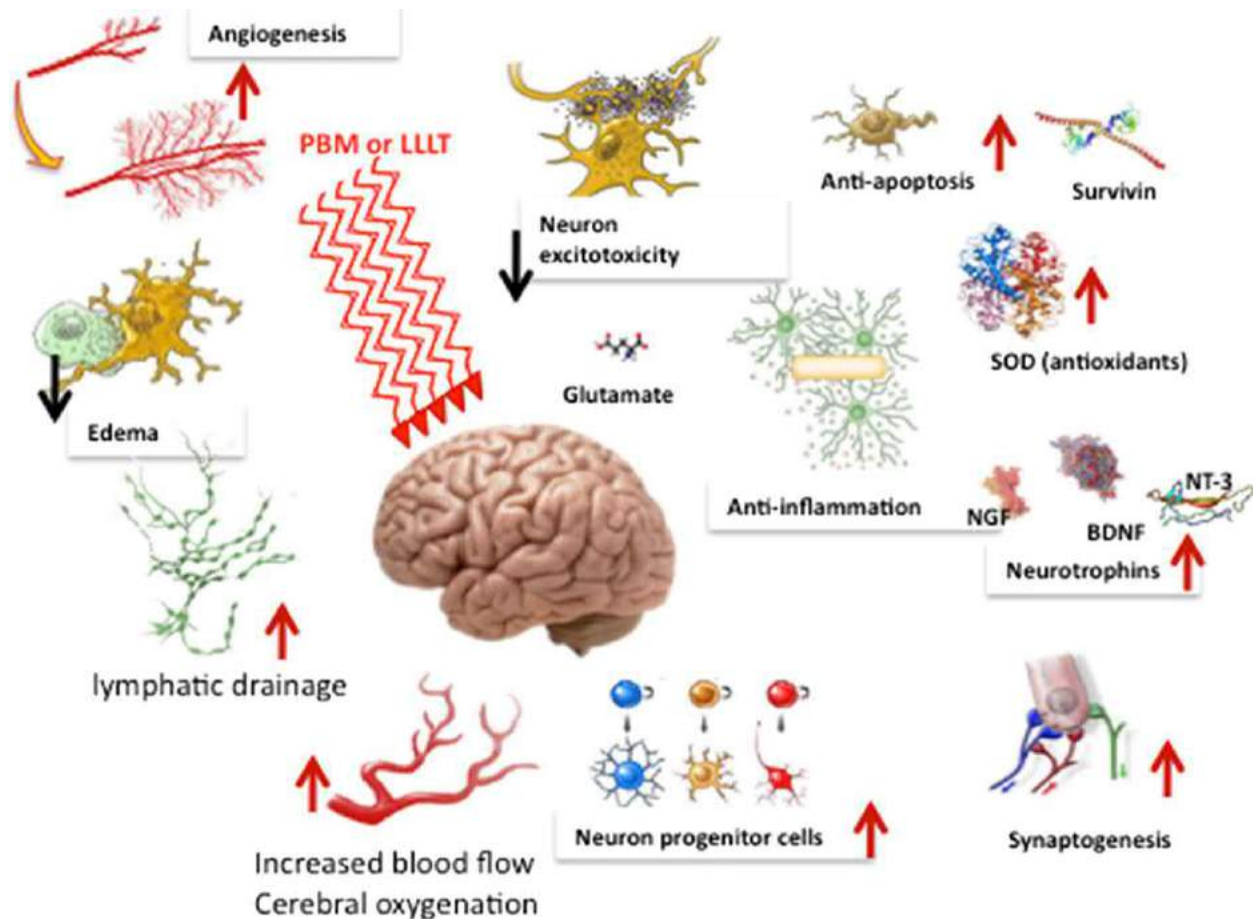


Fig. 11 Many different mechanisms may contribute to the beneficial effects of PBMT on brain disorders.

brain, lies in its possible use as a treatment for acute stroke (Zhao *et al.*, 2015). Animal models such as rats and rabbits, were first used as laboratory models. These animals had experimental strokes induced by a variety of methods, and were then treated with light (usually 810 nm laser) within 24h of stroke onset (Peplow, 2015). In these studies intervention by tPBM within 24 h had meaningful beneficial effects. For the rat models, stroke was induced by middle cerebral artery occlusion (MCAO) via an insertion of a filament into the carotid artery or via craniotomy (Oron *et al.*, 2006). Stroke induction in the “rabbit small clot embolic model” (RSCM) was by injection of a preparation of small blood clots (made from blood taken from a second donor rabbit) into a catheter placed in the right internal carotid artery.

The promising results from pre-clinical animal studies led to a series of three clinical trials called being conducted for treatment of acute stroke in humans. These were called the “Neurothera Effectiveness and Safety Trials” (NEST-1 (Lampl *et al.*, 2007), NEST-2 (Huisa *et al.*, 2013), and NEST-3 (Zivin *et al.*, 2014)) using an 810 nm laser applied to the shaved head within 24 h of patients suffering an ischemic stroke. The first study, NEST-1, enrolled 120 patients between the ages of 40 to 85 years of age with a diagnosis of ischemic stroke involving a neurological deficit that could be measured. The purpose of this first clinical trial was to demonstrate the safety and effectiveness of laser therapy for stroke within 24h (Lampl *et al.*). tPBM significantly improved outcome in human stroke patients, when applied at ~18 h post-stroke, over the entire surface of the head (20 points in the 10/20 EEG system) regardless of stroke (Lampl *et al.*, 2007). Only one laser treatment was administered, and 5 days later, there was significantly greater improvement in the Real- but not in the Sham-treated group ($p < .05$, NIH Stroke Severity Scale). This significantly greater improvement was still present at 90 days post-stroke, where 70% of the patients treated with Real-PBM had a successful outcome, while only 51% of Sham-controls did. The second clinical trial, NEST-2, enrolled 660 patients, aged 40 to 90, who were randomly assigned to one of two groups (331 to PBM, 327 to sham) (Zivin *et al.*). Beneficial results ($p < .04$) were found for the moderate and moderate-severe (but not for the severe) stroke patients, who received the Real laser protocol (Zivin *et al.*, 2009). These results suggested that the overall severity of the individual stroke should be taken into consideration in future studies, and very severe patients are unlikely to recover with any kind of treatment. The last clinical trial, NEST-3, was planned for 1000 patients enrolled. Patients in this study were not to receive tissue plasminogen activator, but the study was prematurely terminated by the DSMB for futility (an expected lack of statistical significance) (Zivin *et al.*). NEST1 was considered successful, even though as a phase 1 trial, it was not designed to show efficacy. NEST2 was partially successful when the patients were stratified, to exclude

very severe strokes or strokes deep within the brain (Huisa *et al.*, 2013). There has been considerable discussion in the scientific literature on precisely why the NEST-3 trial failed (Lapchak and Boitano, 2016). Many commentators have wondered how could tPBM work so well in the first trial, in a sub-group in the second trial, and fail in the third trial. Lapchak's opinion is that the much thicker skull of humans compared to that of the other animals discussed above (mouse, rat and rabbit), meant that therapeutically effective amounts of light were unlikely to reach the brain (Lapchak and Boitano, 2016). Moreover the time between the occurrence of a stroke and initiation of the PBMT may be an important factor. There are reports in the literature that neuroprotection must be administered as soon as possible after a stroke (Lapchak, 2013a,b). Furthermore, stroke trials in particular should adhere to the RIGOR (rigorous research) guidelines and STAIR (stroke therapy academic industry roundtable) criteria (Lapchak *et al.*, 2013). Other contributory causes to the failure of NEST3 may have been included the decision to use only one single tPBM treatment, instead of a series of treatments. Moreover, the optimum brain areas to be treated in acute stroke remain to be determined. It is possible that certain areas of the brain that have sustained ischemic damage should be preferentially illuminated and not others.

Somewhat surprisingly, there have not as yet been many trials of PBM for rehabilitation of chronic stroke patients with only the occasional report to date. Naeser reported in an abstract the use of tPBM to treat chronic aphasia in post-stroke patients (Naeser *et al.*, 2012). Boonswang *et al.* (2012) reported a single patient case in which PBM was used in conjunction with physical therapy to rehabilitate chronic stroke damage. However the findings that PBM can stimulate synaptogenesis in mice with TBI, does suggest that tPBM may have particular benefits in rehabilitation of stroke patients. Norman Doidge, in Toronto, Canada has described the use of PBM as a component of a neuroplasticity approach to rehabilitate chronic stroke patients (Doidge, 2015).

PBM for Traumatic Brain Injury (TBI)

There have been a number of studies looking at the effects of PBM in animal models of TBI. Oron's group was the first (Oron *et al.*, 2007) to demonstrate that a single exposure of the mouse head to a NIR laser (808 nm) a few hours after creation of a TBI lesion could improve neurological performance and reduce the size of the brain lesion. A weight-drop device was used to induce a closed-head injury in the mice. An 808 nm diode laser with two energy densities (1.2–2.4 J/cm² over 2 min of irradiation with 10 and 20 mW/cm²) was delivered to the head 4 h after TBI was induced. Neurobehavioral function was assessed by the neurological severity score (NSS). There were no significant difference in NSS between the power densities (10 vs 20 mW/cm²) or significant differentiation between the control and laser treated group at early time points (24 and 48 h) post TBI. However, there was a significant improvement (27% lower NSS score) in the PBM group at times of 5 days to 4 weeks. The laser treated group also showed a smaller loss of cortical tissue than the sham group (Oron *et al.*).

Hamblin's laboratory then went on (in a series of papers (Oron *et al.*, 2007)) to show that 810nm laser (and 660nm laser) could benefit experimental TBI both in a closed head weight drop model (Wu *et al.*, 2012a), and also in controlled cortical impact model in mice (Ando *et al.*, 2011). Wu *et al.* (Wu *et al.*) explored the effect that varying the laser wavelengths of PBM had on closed-head TBI in mice. Mice were randomly assigned to PBM treated group or to sham group as a control. Closed-head injury (CHI) was induced via a weight drop apparatus. To analyze the severity of the TBI, the neurological severity score (NSS) was measured and recorded. The injured mice were then treated with varying wavelengths of laser (665, 730, 810 or 980 nm) at an energy level of 36 J/cm² at 4 h directed onto the scalp. The 665 nm and 810 nm groups showed significant improvement in NSS when compared to the control group at day 5 to day 28. Conversely, the 730 and 980 nm groups did not show a significant improvement in NSS and these wavelengths did not produce similar beneficial effects as in the 665 nm and 810 nm PBM groups (Wu *et al.*). The tissue chromophore cytochrome c oxidase (CCO) is proposed to be responsible for the underlying mechanism that produces the many PBM effects that are the byproduct of PBM. CCO has absorption bands around 665 nm and 810 nm while it has low absorption bands at the wavelength of 730 nm (Karu *et al.*). It should be noted that this particular study found that the 980 nm did not produce the same positive effects as the 665 nm and 810 nm wavelengths did; nevertheless previous studies did find that the 980 nm wavelength was an active one for PBM. Wu *et al.* proposed these dissimilar results may be due to the variance in the energy level, irradiance etc. between the other studies and this particular study (Wu *et al.*).

Ando *et al.* (2011) used the 810 nm wavelength laser parameters from the previous study and varied the pulse modes of the laser in a mouse model of TBI. These modes consisted of either pulsed wave at 10 Hz or at 100 (50% duty cycle) or continuous wave laser. For the mice, TBI was induced with a controlled cortical impact device via open craniotomy. A single treatment with an 810 nm Ga-Al-As diode laser with a power density of 50 mW/m² and an energy density of 36 J/cm² was given via tPBM to the closed head in mice for a duration of 12 min at 4 h post CCI. At 48 h to 28 days post TBI, all laser treated groups had significant decreases in the measured neurological severity score (NSS) when compared to the control. Although all laser treated groups had similar NSS improvement rates up to day 7, the PW 10 Hz group began to show greater improvement beyond this point. At day 28, the forced swim test for depression and anxiety was used and showed a significant decrease in the immobility time for the PW 10 Hz group. In the tail suspension test, which measures depression and anxiety, there was also a significant decrease in the immobility time at day 28, and this time also at day 1, in the PW 10 Hz group.

Studies using immunofluorescence of mouse brains showed that tPBM increased neuroprogenitor cells in the dentate gyrus (DG) and subventricular zone at 7 days after the treatment (Xuan *et al.*, 2013). The neurotrophin called brain derived neurotrophic factor (BDNF) was also increased in the DG and SVZ at 7 days, while the marker (synapsin-1) for synaptogenesis and neuroplasticity was increased in the cortex at 28 days but not in the DG, SVZ or at 7 days (Xuan *et al.*, 2015). Learning and memory as measured by the

Morris water maze was also improved by tPBM (Xuan *et al.*, 2014). Whalen's laboratory (Khuman *et al.*, 2012) and Whelan's laboratory (Quirk *et al.*, 2012) also successfully demonstrated therapeutic benefits of tPBM for TBI in mice and rats respectively.

Zhang *et al.* (2014) showed that secondary brain injury occurred to a worse degree in mice that had been genetically engineered to lack "Immediate Early Response" gene X-1 (IEX-1) when exposed to a gentle head impact (this injury is thought to closely resemble mild TBI in humans). Exposing IEX-1 knockout mice to PBM 4 h post injury, suppressed proinflammatory cytokine expression of interleukin (IL)-1 β and IL-6, but upregulated TNF- α . The lack of IEX-1 decreased ATP production, but exposing the injured brain to PBM elevated ATP production back to near normal levels.

Dong *et al.* (2015) even further improved the beneficial effects of PBM on TBI in mice, by combining the treatment with metabolic substrates such as pyruvate and/or lactate. The goal was to even further improve mitochondrial function. This combinatorial treatment was able to reverse memory and learning deficits in TBI mice back to normal levels, as well as leaving the hippocampal region completely protected from tissue loss; a stark contrast to that found in control TBI mice that exhibited severe tissue loss from secondary brain injury.

Margaret Naeser and collaborators have tested PBM in human subjects who had suffered TBI in the past (Naeser and Hamblin, 2015). Many sufferers from severe or even moderate TBI, have very long lasting and even life-changing sequelae (headaches, cognitive impairment, and difficulty sleeping) that prevent them working or living any kind of normal life. These individuals may have been high achievers before the accident that caused damage to their brain (McClure, 2011). Initially Naeser published a report (Naeser *et al.*, 2011) describing two cases she treated with PBM applied to the forehead twice a week. A 500 mW continuous wave LED source (mixture of 660 nm red and 830 nm NIR LEDs) with a power density of 22.2 mW/cm² (area of 22.48 cm²), was applied to the forehead for a typical duration of 10 min (13.3 J/cm²). In the first case study the patient reported that she could concentrate on tasks for a longer period of time (the time able to work at a computer increased from 30 min to 3 h). She had a better ability to remember what she read, decreased sensitivity when receiving haircuts in the spots where PBM was applied, and improved mathematical skills after undergoing PBM. The second patient had statistically significant improvements compared to prior neuropsychological tests after 9 months of treatment. The patient had a 2 standard deviation (SD) increase on tests of inhibition and inhibition accuracy (9th percentile to 63rd percentile) on the Stroop test for executive function and a 1 SD increase on the Wechsler Memory scale test for the logical memory test (83rd percentile to 99th percentile) (Naeser *et al.*, 2010).

Naeser *et al.* then went on to report a case series of a further eleven patients (Naeser *et al.*, 2010). This was an open protocol study that examined whether scalp application of red and near infrared (NIR) light could improve cognition in patients with chronic, mild traumatic brain injury (mTBI). This study had 11 participants ranging in age from 26 to 62 (6 males, 5 females) who suffered from persistent cognitive dysfunction after mTBI. The participants' injuries were caused by motor vehicle accidents, sports related events and for one participant, an improvised explosive device (IED) blast. tPBM consisted of 18 sessions (Monday, Wednesday, and Friday for 6 weeks) and commenced anywhere from 10 months to 8 years post-TBI. A total of 11 LED clusters (5.25 cm in diameter, 500 mW, 22.2 mW/cm², 13 J/cm²) were applied for about 10 min per session (5 or 6 LED placements per set, Set A and then Set B, in each session). Neuropsychological testing was performed pre-LED application and 1 week, 1 month and 2 months after the final treatment. Naeser and colleagues found that there was a significant positive linear trend observed for the Stroop Test for executive function, in trial 2 inhibition ($p=0.004$); Stroop, trial 4 inhibition switching ($p=0.003$); California Verbal Learning Test (CVLT)-II, total trials 1–5 ($p=0.003$); CVLT-II, long delay free recall ($p=0.006$). Improved sleep and fewer post-traumatic stress disorder (PTSD) symptoms, if present beforehand, were observed after treatment. Participants and family members also reported better social function and a better ability to perform interpersonal and occupational activities. Although these results were significant, further placebo-controlled studies will be needed to ensure the reliability of this these data (Naeser *et al.*, 2014).

Henderson and Morris (2015) used a high-power NIR laser (10–15 W at 810 and 980 nm) applied to the head to treat a patient with moderate TBI. The patient received 20 NIR applications over a 2 month period. They carried out anatomical magnetic resonance imaging (MRI) and perfusion single-photon emission computed tomography (SPECT). The patient showed decreased depression, anxiety, headache, and insomnia, whereas cognition and quality of life improved, accompanied by changes in the SPECT imaging.

PBM for Neurodegenerative Diseases

PBM has also been considered as a candidate for treating degenerative brain disorders such as Alzheimer's disease (AD), and Parkinson's disease (PD) (Moges *et al.*, 2009; Zhang *et al.*, 2008). Although only preliminary studies have so far been carried out, these are encouraging indications that merit further investigation.

Alzheimer's disease

There was a convincing study (De Taboada *et al.*, 2011) carried out in an A β PP transgenic mouse of AD. tPBM (810 nm laser) was administered at different doses 3 times/week for 6 months starting at 3 months of age. The numbers of A β plaques were significantly reduced in the brain with administration of tPBM in a dose-dependent fashion. tPBM mitigated the behavioral effects seen with advanced amyloid deposition and reduced the expression of inflammatory markers in the transgenic mice. In addition, TLT showed an increase in ATP levels, mitochondrial function, and c-fos expression suggesting there was an overall improvement in neurological function.

There has been a group of investigators in Northern England who have used a helmet containing 1072 nm LEDs to treat AD, but somewhat surprisingly no peer-reviewed publications have described this approach (see Relevant Website section). However a small pilot study (19 patients) that took the form of a randomized placebo-controlled trial investigated the effect of the Vielicht Neuro system (a combination of tPBM and intranasal PBM) on patients with dementia and mild cognitive impairment (Saltmarche *et al.*, 2016). This was a controlled single blind pilot study in humans to investigate the effects of PBM on memory and cognition. 19 participants with impaired memory/cognition were randomized into active and sham treatments over 12 weeks with a 4-week no-treatment follow-up period. They were assessed with MMSE and ADAS-cog scales. The protocol involved in-clinic use of a combined transcranial-intranasal PBM device; and at-home use of an intranasal-only PBM device and participants/caregivers noted daily experiences in a journal. Active participants with moderate to severe impairment (MMSE scores 5–24) showed significant improvements (5-points MMSE score) after 12 weeks. They also reported better sleep, fewer angry outbursts and decreased anxiety and wandering. Declines were noted during the 4 week no-treatment follow-up period. Participants with mild impairment to normal (MMSE scores of 25 to 30) in both the active and sham sub-groups showed improvements. No related adverse events were reported.

An interesting paper from Russia (Maksimovich, 2015) described the use of intravascular PBM to treat 89 patients with AD who received PBM (46 patients) or standard treatment with memantine and rivastigmine (43 patients). The PBM consisted of threading a fiber-optic through a catheter in the femoral artery and advancing it to the distal site of the anterior and middle cerebral arteries and delivering 20 mW of red laser for 20–40 min. The PBM group had improvement in cerebral microcirculation leading to permanent (from 1–7 years) reduction in dementia and cognitive recovery.

Parkinson's disease

The majority of studies on PBM for Parkinson's disease have been in animal models and have come from the laboratory of John Mitrofanis in Australia (Johnstone *et al.*, 2015). Two basic models of Parkinson's disease were used. The first employed administration of the small molecule (MPTP or 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) to mice (Shaw *et al.*, 2010). MPTP was discovered as an impurity in an illegal recreational drug to cause Parkinson's like symptoms (loss of substantia nigra cells) in young people who had taken this drug (Barcia, 2012). Mice were treated with tPBM (670-nm LED, 40 mW/cm², 3.6 J/cm²) 15 min after each MPTP injection repeated 4 times over 30 h. There were significantly more (35%–45%) dopaminergic cells in the brains of the tPBM treated mice (Shaw *et al.*, 2010). A subsequent study showed similar results in a chronic mouse model of MPTP-induced Parkinson's disease (Peoples *et al.*, 2012). They repeated their studies in another mouse model of Parkinson's disease, the tau transgenic mouse strain (K3) that has a progressive degeneration of dopaminergic cells in the substantia nigra pars compacta (SNc) (Purushothuman *et al.*, 2013). They went on to test a surgically implanted intracranial fiber designed to deliver either 670 nm LED (0.16 mW) or 670 nm laser (67 mW) into the lateral ventricle of the brain in MPTP-treated mice (Moro *et al.*, 2014). Both low power LED and high power laser were effective in preserving SNc cells, but the laser was considered to be unsuitable for long term use (6 days) due to excessive heat production. As mentioned above, these authors also reported a protective effect of abscopal light exposure (head shielded) in this mouse model (Johnstone *et al.*, 2014). Recently this group has tested their implanted fiber approach in a model of Parkinson's disease in adult Macaque monkeys treated with MPTP (El Massri *et al.*, 2016). Clinical evaluation of Parkinson's symptoms (posture, general activity, bradykinesia, and facial expression) in the monkeys were improved at low doses of light (24 J or 35 J) compared to high doses (125 J) (Moro *et al.*, 2016).

The only clinical report of PBM for Parkinson's disease in humans was an abstract presented in 2010 (Maloney *et al.*, 2010). Eight patients between 18 to 80 years with late stage PD participated in a non-controlled, non-randomized study. Participants received tPBM treatments of the head designed to deliver light to the brain stem, bilateral occipital, parietal, temporal and frontal lobes, and treatment along the sagittal suture. A Visual Analog Scale (VAS), was used to record the severity of their symptoms of balance, gait, freezing, cognitive function, rolling in bed, and difficulties with speech pre-procedure and at study endpoint with 10 being most severe and 0 as no symptom. Compared with baseline, all participants demonstrated a numerical improvement in the VAS from baseline to study endpoint. A statistically significant reduction in VAS rating for gait and cognitive function was observed with average mean change of -1.87 ($p < 0.05$) for gait and a mean reduction of -2.22 ($p < 0.05$) for cognitive function. Further, freezing and difficulty with speech ratings were significantly lower (mean reduction of 1.28 ($p < 0.05$) for freezing and 2.22 ($p < 0.05$) for difficulty with speech).

PBM for Psychiatric Disorders

A common and well-accepted animal model of depression is called "chronic mild stress" (Willner, 1997). After exposure to a series of chronic unpredictable mild stressors, animals develop symptoms seen in human depression, such as anhedonia (loss of the capacity to experience pleasure, a core symptom of major depressive disorder), weight loss or slower weight gain, decrease in locomotor activity, and sleep disorders (Anisman and Matheson, 2005). Wu *et al.* used Wistar rats to show that after 5 weeks of chronic stress, application of tPBM 3 times a week for 3 weeks (810 nm laser, 100 Hz with 20% duty cycle, 120 J/cm²) gave significant improvement in the forced swimming test (FST) (Wu *et al.*, 2012b). In a similar study Salehpour *et al.* (2016) compared the effects of two different lasers (630 nm at 89 mW/cm², and 810 nm at 562 mW/cm², both pulsed at 10 Hz, 50% duty cycle). The 810 nm laser proved better than the 630 nm laser in the FST, in the elevated plus maze and also reduced blood cortisol levels.

The first clinical study in depression and anxiety was published by Schiffer *et al.* in 2009 (Schiffer *et al.*, 2009). They used a fairly small area 1 W 810 nm LED array applied to the forehead in patients with major depression and anxiety. They found improvements in the Hamilton depression rating scale (HAM-D), and the Hamilton anxiety rating scale (HAM-A), 2 weeks after a single treatment. They also found increases in frontal pole regional cerebral blood flow (rCBF) during the light delivery using a commercial NIR spectroscopy device. Cassano and co-workers (Cassano *et al.*, 2015) used tPBM with an 810 nm laser (700 mW/cm² and a fluence of 84 J/cm² delivered per session for 6 sessions) in patients with major depression. Baseline mean HAM-D17 scores decreased from 19.8 ± 4.4 (SD) to 13 ± 5.35 (SD) after treatment ($P=0.004$).

Enhancement of Performance

PBM has been studied in two broad areas of performance enhancement, namely athletic performance and cognitive performance.

PBM for muscle and athletic performance

PBM has been used for prevention of muscle damage after exercise, including delayed onset muscle soreness, and improvement in muscle performance, demonstrated by increases in muscle torque, power or work; workload capacity; fatigue resistance; functional or sport activity, and exercise recovery.

The use of PBM to prevent muscle damage has been investigated in animal models, irradiating skeletal muscles before a bout of intense exercise (called in this chapter “muscular pre-conditioning”) and by following the muscle damage by measuring the creatine kinase (CK) levels in the bloodstream. The first study that used photobiomodulation by PBM to prevent muscle damage was carried out by Lopes-Martins *et al.* (2006) in rats. These authors investigated the effects of different doses of 655 nm laser (0.5 J/cm²; 1.0 J/cm² and 2.5 J/cm²) to prevent muscle fatigue and muscle damage (CK) induced by neuromuscular electrical stimulation. This study reported a PBM dose response to decrease CK activity.

Another experimental study used exercise-trained rats on an inclined treadmill and measured inhibition of inflammation, reduction of CK activity and reduction of oxidative stress. They also found increases in defense against oxidative stress (increased activity of superoxide dismutase) 24 h and 48 h after exercise (Gladwin and Shiva, 2009). These previous studies and other similar reports (de Almeida *et al.*, 2011; Leal *et al.*, 2010a,b; Liu *et al.*, 2009; Lopes-Martins *et al.*, 2006; Santos *et al.*, 2014; Sussai *et al.*, 2010) were important to establish the scientific basis to conduct clinical trials for the prevention of muscle damage by PBM.

Regarding clinical trials of improved muscle performance and exercise recovery, the first studies published were on delayed onset muscle soreness (Douris *et al.*, 2006; Vinck *et al.*, 2006) and resistance to muscle fatigue (Gorgey *et al.*, 2008; Leal *et al.*, 2008) during one or a few bouts of exercise; and improved muscle strength (Ferraresi *et al.*, 2011) and resistance to muscle fatigue (Vieira *et al.*, 2012) after exercise training programs combined with PBM. PBM has been tested in real life athletic competition such as in a team of volleyball players (Ferraresi *et al.*, 2015). A randomized double-blind placebo-controlled trial delivered PBM using LEDs (25 clusters of 4 infrared LEDs 850 nm, 130 mW and 25 clusters of 4 red LEDs 630 nm; 80 mW). PBM was applied over the *quadriceps femoris* muscles, hamstrings, and *triceps surae* of volleyball players before official matches. A professional male volleyball team (12 athletes) was randomized to receive one of four different total doses over each muscle group in a double-blind protocol. PBM at a dose of 210 J or higher avoided significant increases in creatine kinase (a marker of muscle damage) that were found in the placebo group.

PBM for enhancement of cognitive performance

The second area of performance enhancement by PBM (that is not yet as advanced as athletic performance) is the possibility of improving cognitive function in normal people or healthy animals using PBM. The first report was in middle aged (12 months) CD1 female mice (Michalikova *et al.*, 2008). Exposure of the mice to 1072 nm LED arrays led to improved performance in a 3D maze compared to sham treated age-matched controls. Francisco Gonzalez-Lima at the University of Texas Austin, has worked in this area for some time (Gonzalez-Lima and Barrett, 2014). Working in rats they showed that transcranial PBM (9 mW/cm² with 660 nm LED array) induced a dose-dependent increase in oxygen consumption of 5% after 1 J/cm² and 16% after 5 J/cm² (Rojas *et al.*, 2012). They also found that tPBM reduced fear renewal and prevented the reemergence of extinguished conditioned fear responses (Rojas *et al.*, 2012).

In normal human volunteers Gonzalez-Lima used transcranial PBM (1064 nm laser, 60 J/cm² at 250 mW/cm²) delivered to the forehead in a placebo-controlled, randomized study, to influence cognitive tasks related to the prefrontal cortex, including a psychomotor vigilance task (PVT), a delayed match-to-sample (DMS) memory task, and the positive and negative affect schedule (PANAS-X) to show improved mood (Barrett and Gonzalez-Lima, 2013). Subsequent studies in normal humans showed that tPBM with 1064 nm laser could improve performance in the Wisconsin Card Sorting Task (considered the gold standard test for executive function) (Blanco *et al.*, 2015). They also showed that tPBM to the right forehead (but not the left forehead) had better effects on improving attention bias modification (ABM) in humans with depression (Disner *et al.*, 2016).

A study by Salgado *et al.* used transcranial LED PBM on cerebral blood flow in healthy elderly women analyzed by transcranial Doppler ultrasound (TCD) of the right and left middle cerebral artery and basilar artery. Twenty-five non-institutionalized elderly women (mean age 72 years old), with cognitive status > 24, were assessed using TCD before and after transcranial LED therapy. tPBM (627 nm, 70 mW/cm², 10 J/cm²) was performed at four points of the frontal and parietal region for 30 s each twice a week for 4 weeks. There was a significant increase in the systolic and diastolic velocity of the left middle cerebral artery (25 and 30%,

respectively) and the basilar artery (up to 17 and 25%), as well as a decrease in the pulsatility index and resistance index values of the three cerebral arteries analyzed (Salgado *et al.*, 2015).

Conclusions

The goal of this chapter has been to argue the case that different aspects of photomedicine can exert two complementary, but completely opposite therapeutic effects. PDT can kill undesirable cells such as cancer cells and pathogenic microorganisms and remove unwanted tissue, while PBM can heal, restore and protect tissues that have been damaged or are at risk. Ironically both of these therapeutic approaches often use remarkably similar light parameters, namely wavelengths of red light (between 600 and 700 nm) at power densities and fluences that are also surprisingly comparable. However PBM is more versatile than PDT, which has an upper wavelength limit of about 800 nm. PBM on the other hand works very well with NIR light above 800 nm. In fact 810 nm is probably the single most popular wavelength for PBM. Several commentators have pointed out that the mechanisms of action of PDT and PBM have aspects in common; specifically the production of ROS (Lubart and Friedmann, 2011). However the amount of ROS produced is very different. In PDT the type of ROS (predominantly singlet oxygen and hydroxyl radicals produced in large quantities) produced by exogenously added PS are highly damaging to cells. These ROS are produced by photochemical reactions occurring between the administered PS and the light. By contrast the much lower amounts of ROS produced in PBM, arise from the action of light on the mitochondria inside the cells containing intrinsic chromophores, and are much less damaging, their duration is short, and they tend to stimulate protective responses and trigger signaling pathways.

We can illustrate the dual simultaneous use of PDT and PBM by describing an application of photomedicine in dentistry. Periodontitis is caused by pathogenic bacteria that accumulate in the dental pocket (between the tooth and the gums). An inflammatory response to these bacteria leads to production of matrix metalloproteinase enzymes that degrade the periodontal ligament eventually leading to tooth loss. PDT is used to treat periodontitis by injecting a small amount of a solution of methylene blue into the dental pocket and inserting a narrow fiber optic to deliver red light to excite the dye and kill the bacteria. However the red light spreads throughout the gingival tissue (containing no MB) where it acts to reduce inflammation and stimulate healing of the periodontal ligament. This dual action mechanism may explain why PDT is so effective for this condition.

References

- Abdel-Hady, E.S., Martin-Hirsch, P., Duggan-Keen, M., *et al.*, 2001. Immunological and viral factors associated with the response of vulval intraepithelial neoplasia to photodynamic therapy. *Cancer Res.* 61 (1), 192–196.
- Abergel, R.P., Lyons, R.F., Castel, J.C., Dwyer, R.M., Uitto, J., 1987a. Biostimulation of wound healing by lasers: Experimental approaches in animal models and in fibroblast cultures. *J. Dermatol. Surg. Oncol.* 13, 127–133.
- Abergel, R.P., Lyons, R.F., Castel, J.C., Dwyer, R.M., Uitto, J., 1987b. Biostimulation of wound healing by lasers: Experimental approaches in animal models and in fibroblast cultures. *J. Dermatol. Surg. Oncol.* 13 (2), 127–133.
- Agostinis, P., Berg, K., Cengel, K.A., *et al.*, 2011. Photodynamic therapy of cancer: An update. *CA Cancer J. Clin.* 61 (4), 250–281.
- Albarracin, R., Eells, J., Valter, K., 2011. Photobiomodulation protects the retina from light-induced photoreceptor degeneration. *Invest. Ophthalmol. Vis. Sci.* 52 (6), 3582–3592.
- Allison, B.A., Pritchard, P.H., Levy, J.G., 1994. Evidence for low-density lipoprotein receptor-mediated uptake of benzoporphyrin derivative. *Br. J. Cancer* 69 (5), 833–839.
- Almeida-Lopes, L., Rigau, J., Zangaro, R.A., *et al.*, 2001. Comparison of the low level laser therapy effects on cultured human gingival fibroblasts proliferation using different irradiance and same fluence. *Lasers Surg. Med.* 29, 179–184.
- Anders, J.J., Lanzafame, R.J., Arany, P.R., 2015. Low-level light/laser therapy versus photobiomodulation therapy. *Photomed. Laser Surg.* 33 (4), 183–184.
- Ando, T., Xuan, W., Xu, T., *et al.*, 2011. Comparison of therapeutic effects between pulsed and continuous wave 810-nm wavelength laser irradiation for traumatic brain injury in mice. *PLoS One* 6 (10), e26212.
- Anisman, H., Matheson, K., 2005. Stress, depression, and anhedonia: Caveats concerning animal models. *Neurosci. Biobehav. Rev.* 29 (4-5), 525–546.
- Anneroth, G., Hall, G., Ryden, H., Zetterqvist, L., 1988. The effect of low-energy infra-red laser radiation on wound healing in rats. *Br. J. Oral Maxillofac. Surg.* 26, 12–17.
- Barcia, C., 2012. Who else was intoxicated with MPTP in Santa Clara? *Parkinsonism Relat. Disord.* 18 (9), 1005–1006.
- Barolet, D., Roberge, C.J., Auger, F.A., Boucher, A., Germain, L., 2009. Regulation of skin collagen metabolism in vitro using a pulsed 660 nm LED light source: Clinical correlation with a single-blinded study. *J. Invest. Dermatol.* 129 (12), 2751–2759.
- Barrett, D.W., Gonzalez-Lima, F., 2013. Transcranial infrared laser stimulation produces beneficial cognitive and emotional effects in humans. *Neuroscience* 230, 13–23.
- Baxter, G.D., Bell, A.J., Allen, J.M., Ravey, J., 1991. Low level laser therapy: Current clinical practice in Northern Ireland. *Physiotherapy* 77 (3), 171–178.
- Bellnier, D.A., Greco, W.R., Parsons, J.C., *et al.*, 1997. An assay for the quantitation of Photofrin in tissues and fluids. *Photochem. Photobiol.* 66 (2), 237–244.
- Bellnier, D.A., Ho, Y.K., Pandey, R.K., Missert, J.R., Dougherty, T.J., 1989. Distribution and elimination of Photofrin II in mice. *Photochem. Photobiol.* 50 (2), 221–228.
- Benvenuti, F., 2016. The dendritic cell synapse: A life dedicated to T cell activation. *Front. Immunol.* 7, 70.
- Bernstein, E.F., 2005. Hair growth induced by diode laser treatment. *Dermatol. Surg.* 31 (5), 584–586.
- Bertoloni, G., Rossi, F., Valduga, G., Jori, G., van Lier, J., 1990. Photosensitizing activity of water- and lipid-soluble phthalocyanines on *Escherichia coli*. *FEMS Microbiol. Lett.* 59 (1-2), 149–155.
- Bhargava, A., Mishra, D., Banerjee, S., Mishra, P.K., 2012. Dendritic cell engineering for tumor immunotherapy: From biology to clinical translation. *Immunotherapy* 4 (7), 703–718.
- Bhat, J., Birch, J., Whitehurst, C., Lanigan, S.W., 2005. A single-blinded randomised controlled study to determine the efficacy of Omnilux Revive facial treatment in skin rejuvenation. *Lasers Med. Sci.* 20 (1), 6–10.
- Bhuvaneswari, R., Gan, Y.Y., Soo, K.C., Olivo, M., 2009. The effect of photodynamic therapy on tumor angiogenesis. *Cell. Mol. Life Sci.* 66 (14), 2275–2283.
- Bisht, D., Gupta, S.C., Mishra, V., *et al.*, 1994. Effect of low intensity laser radiation on healing of open skin wounds in rats. *Indian J. Med. Res.* 100, 43–46.
- Bisht, D., Mehrortra, R., Singh, P.A., Atri, S.C., Kumar, A., 1999. Effect of helium-neon laser on wound healing. *Indian J. Exp. Biol.* 37, 187–189.

- Bjrdal, J.M., Couppe, C., Chow, R.T., Tuner, J., Ljunggren, E.A., 2003. A systematic review of low level laser therapy with location-specific doses for pain from chronic joint disorders. *Aust. J. Physiother.* 49, 107–116.
- Bjornsti, M.A., Houghton, P.J., 2004. The TOR pathway: A target for cancer therapy. *Nat. Rev. Cancer* 4 (5), 335–348.
- Blanco, N.J., Maddox, W.T., Gonzalez-Lima, F., 2015. Improving executive function using transcranial infrared laser stimulation. *J. Neuropsychol.*
- Bland, J., 1976. Biochemical effects of excited state molecular oxygen. *J. Chem. Educ.* 53 (5), 5.
- Boonswang, N.A., Chicchi, M., Lukachek, A., Curtiss, D., 2012. A new treatment protocol using photobiomodulation and muscle/bone/joint recovery techniques having a dramatic effect on a stroke patient's recovery: a new weapon for clinicians. *BMJ Case Rep.* 2012.
- Bouzar, N., Firooz, A.R., 2006. Lasers may induce terminal hair growth. *Dermatol. Surg.* 32 (3), 460.
- Boyle, R.W., Dolphin, D., 1996. Structure and biodistribution relationships of photodynamic sensitizers. *Photochem. Photobiol.* 64 (3), 469–485.
- Brown, S.A., Rohrich, R.J., Kenkel, J., *et al.*, 2004. Effect of low-level laser therapy on abdominal adipocytes before lipoplasty procedures. *Plast Reconstr Surg* 113 (6), 1796–1804. (discussion 1805–6).
- Brun, P.H., DeGroot, J.L., Dickson, E.F., Farahani, M., Pottier, R.H., 2004. Determination of the in vivo pharmacokinetics of palladium-bacteriopheophorbide (WST09) in EMT6 tumour-bearing Balb/c mice using graphite furnace atomic absorption spectroscopy. *Photochem. Photobiol. Sci.* 3 (11–12), 1006–1010.
- Buettner, G.R., 1993. The pecking order of free radicals and antioxidants: lipid peroxidation, alpha-tocopherol, and ascorbate. *Arch. Biochem. Biophys.* 300 (2), 535–543.
- Bush, K., Courvalin, P., Dantas, G., *et al.*, 2011. Tackling antibiotic resistance. *Nat. Rev. Microbiol.* 9 (12), 894–896.
- Buytaert, E., Callewaert, G., Hendrickx, N., *et al.*, 2006a. Role of endoplasmic reticulum depletion and multidomain proapoptotic BAX and BAK proteins in shaping cell death after hypericin-mediated photodynamic therapy. *FASEB J* 20 (6), 756–758.
- Buytaert, E., Callewaert, G., Vandenheede, J.R., Agostinis, P., 2006b. Deficiency in apoptotic effectors Bax and Bak reveals an autophagic cell death pathway initiated by photodamage to the endoplasmic reticulum. *Autophagy* 2 (3), 238–240.
- Buytaert, E., Dewaele, M., Agostinis, P., 2007. Molecular effectors of multiple cell death pathways initiated by photodynamic therapy. *Biochim. Biophys. Acta* 1776 (1), 86–107.
- Canti, G.L., Lattuada, D., Nicolini, A., *et al.*, 1994. Immunopharmacology studies on photosensitizers used in photodynamic therapy. *Proc. SPIE* 2078, 268–275.
- Caruso-Davis, M.K., Guillot, T.S., Podichetty, V.K., *et al.*, 2011. Efficacy of low-level laser therapy for body contouring and spot fat reduction. *Obes. Surg.* 21 (6), 722–729.
- Cassano, P., Cusin, C., Mischoulon, D., *et al.*, 2015. Near-infrared transcranial radiation for major depressive disorder: Proof of concept study. *Psychiatry J.* 2015, 352979.
- Castano, A.P., Dai, T., Yaroslavsky, I., *et al.*, 2007. Low-level laser therapy for zymosan-induced arthritis in rats: Importance of illumination time. *Lasers Surg. Med.* 39 (6), 543–550.
- Castellani, A., Pace, G.P., Concioli, M., 1963. Photodynamic effect of haematoporphyrin on blood microcirculation. *J. Pathol. Bacteriol.* 86, 99–102.
- Castellino, F., Germain, R.N., 2006. Cooperation between CD4 + and CD8 + T cells: When, where, and how. *Annu. Rev. Immunol.* 24, 519–540.
- Castex-Rizzi, N., Lachgar, S., Charveron, M., Gall, Y., 2002. Implication of VEGF, steroid hormones and neuropeptides in hair follicle cell responses. *Ann. Dermatol. Venerol.* 129 (5 Pt 2), 783–786.
- Cengel, K.A., Simone 2nd, C.B., Glatstein, E., 2016. PDT: What's past is prologue. *Cancer Res* 76 (9), 2497–2499.
- Chen, A.C., Arany, P.R., Huang, Y.Y., *et al.*, 2011. Low-level laser therapy activates NF- κ B via generation of reactive oxygen species in mouse embryonic fibroblasts. *PLoS One* 6 (7), e22453.
- Chow, R.T., Johnson, M.I., Lopes-Martins, R.A., Bjrdal, J.M., 2009. Efficacy of low-level laser therapy in the management of neck pain: A systematic review and meta-analysis of randomised placebo or active-treatment controlled trials. *Lancet* 374, 1897–1908.
- Cory, S., Adams, J.M., 2002. The Bcl2 family: regulators of the cellular life-or-death switch. *Nat. Rev. Cancer* 2 (9), 647–656.
- Cotter, T.G., 2009. Apoptosis and cancer: The genesis of a research field. *Nat. Rev. Cancer* 9 (7), 501–507.
- Coupienne, I., Piette, J., Bontems, S., 2010. How to monitor NF- κ B activation after photodynamic therapy. *Methods Mol. Biol.* 635, 79–95.
- Dahle, J., Steen, H.B., Moan, J., 1999. The mode of cell death induced by photodynamic treatment depends on cell density. *Photochem. Photobiol.* 70 (3), 363–367.
- Dahl, T.A., Midden, W.R., Hartman, P.E., 1989. Comparison of killing of gram-negative and gram-positive bacteria by pure singlet oxygen. *J. Bacteriol.* 171 (4), 2188–2194.
- Dai, T., Huang, Y.Y., Hamblin, M.R., 2009. Photodynamic therapy for localized infections-state of the art. *Photodiagnosis Photodyn. Ther.* 6 (3–4), 170–188.
- Daniel, N.N., Korsmeyer, S.J., 2004. Cell death: critical control points. *Cell* 116 (2), 205–219.
- Davies, M.J., 2016. Protein oxidation and peroxidation. *Biochem. J.* 473 (7), 805–825.
- de Almeida, P., Lopes-Martins, R.A., Tomazoni, S.S., *et al.*, 2011. Low-level laser therapy improves skeletal muscle performance, decreases skeletal muscle damage and modulates mRNA expression of COX-1 and COX-2 in a dose-dependent manner. *Photochem. Photobiol.* 87 (5), 1159–1163.
- Debelve, E., Cheng, C., Schaefer, S.C., *et al.*, 2010. Photodynamic therapy induces selective extravasation of macromolecules: insights using intravital microscopy. *J. Photochem. Photobiol. B* 98 (1), 69–76.
- de Melo, W.C., Avci, P., de Oliveira, M.N., *et al.*, 2013. Photodynamic inactivation of biofilm: taking a lightly colored approach to stubborn infection. *Expert Rev. Anti Infect. Ther.* 11 (7), 669–693.
- Demidova-Rice, T.N., Salomatina, E.V., Yaroslavsky, A.N., Herman, I.M., Hamblin, M.R., 2007. Low-level light stimulates excisional wound healing in mice. *Lasers Surg. Med.* 39 (9), 706–715.
- De Taboada, L., Yu, J., El-Amouri, S., *et al.*, 2011. Transcranial laser therapy attenuates amyloid-beta peptide neuropathology in amyloid-beta protein precursor transgenic mice. *J. Alzheimers Dis.* 23 (3), 521–535.
- de Vree, W.J., Essers, M.C., Koster, J.F., Sluiter, W., 1997. Role of interleukin 1 and granulocyte colony-stimulating factor in photofrin-based photodynamic therapy of rat rhabdomyosarcoma tumors. *Cancer Res.* 57 (13), 2555–2558.
- Dierickx, C.C., Anderson, R.R., 2005. Visible light treatment of photoaging. *Dermatol. Ther.* 18 (3), 191–208.
- Disner, S.G., Beevers, C.G., Gonzalez-Lima, F., 2016. Transcranial laser stimulation as neuroenhancement for attention bias modification in adults with elevated depression symptoms. *Brain Stimul.*
- Doidge, N., 2015. *The Brain's Way of Healing: remarkable Discoveries and Recoveries from the Frontiers of Neuroplasticity*. New York, NY: Viking Press.
- Dolmans, D.E., Kadambi, A., Hill, J.S., *et al.*, 2002. Vascular accumulation of a novel photosensitizer, MV6401, causes selective thrombosis in tumor vessels after photodynamic therapy. *Cancer Res.* 62 (7), 2151–2156.
- Dong, T., Zhang, Q., Hamblin, M.R., Wu, M.X., 2015. Low-level light in combination with metabolic modulators for effective therapy of injured brain. *J. Cereb. Blood Flow Metabolism: Off. J. Int. Soc. Cereb. Blood Flow Metabolism*.
- Dougherty, T.J., Gomer, C.J., Henderson, B.W., *et al.*, 1998. Photodynamic therapy. *J. Natl. Cancer Inst.* 90 (12), 889–905.
- Douris, P., Southard, V., Ferrigi, R., *et al.*, 2006. Effect of phototherapy on delayed onset muscle soreness. *Photomed. Laser Surg.* 24 (3), 377–382.
- Dragieva, G., Hafner, J., Dummer, R., *et al.*, 2004. Topical photodynamic therapy in the treatment of actinic keratoses and Bowen's disease in transplant recipients. *Transplantation* 77 (1), 115–121.
- Eells, J.T., Gopalakrishnan, S., Valters, K., 2016. Near-infrared photobiomodulation in retinal injury and disease. *Adv. Exp. Med. Biol.* 854, 437–441.
- Eells, J.T., Henry, M.M., Summerfelt, P., *et al.*, 2003. Therapeutic photobiomodulation for methanol-induced retinal toxicity. *Proc. Natl. Acad. Sci. USA* 100 (6), 3439–3444.
- Egorin, M.J., Zuhowski, E.G., Sentz, D.L., *et al.*, 1999. Plasma pharmacokinetics and tissue distribution in CD2F1 mice of Pc4 (NSC 676418), a silicone phthalocyanine photodynamic sensitizing agent. *Cancer Chemother. Pharmacol.* 44 (4), 283–294.
- Eichler, M., Lavi, R., Shainberg, A., Lubart, R., 2005. Flavins are source of visible-light-induced free radical formation in cells. *Lasers Surg. Med.* 37 (4), 314–319.
- El Massri, N., Moro, C., Torres, N., *et al.*, 2016. Near-infrared light treatment reduces astrogliosis in MPTP-treated monkeys. *Exp. Brain Res.*
- Fàà, H., 1968. Long-range coherence and energy storage in biological systems. *Int. J. Quantum Chem.* 2 (5), 641–649.

- FAH, A.-W., XY, Z., 1996. Comparison of the effects of laser therapy on wound healing using different laser wavelengths. *Laser Ther* 1996 (8), 127–135.
- Fauci, A.S., Morens, D.M., 2012. The perpetual challenge of infectious diseases. *New England J. Med.* 366 (5), 454–461.
- Ferraresi, C., de Brito Oliveira, T., de Oliveira Zafalon, L., *et al.*, 2011. Effects of low level laser therapy (808 nm) on physical strength training in humans. *Lasers Med. Sci.* 26 (3), 349–358.
- Ferraresi, C., Dos Santos, R.V., Marques, G., *et al.*, 2015. Light-emitting diode therapy (LEDT) before matches prevents increase in creatine kinase with a light dose response in volleyball players. *Lasers Med. Sci.* 30 (4), 1281–1287.
- Ferreira, D.D., Reynoso, E., Cordero, P., *et al.*, 2016. Synthesis and properties of 5,10,15,20-tetrakis[4-(3-N,N-dimethylaminopropoxy)phenyl] chlorin as potential broad-spectrum antimicrobial photosensitizers. *J. Photochem. Photobiol. B* 158, 243–251.
- Fingar, V.H., Kik, P.K., Haydon, P.S., *et al.*, 1999. Analysis of acute vascular damage after photodynamic therapy using benzoporphyrin derivative (BPD). *Br. J. Cancer* 79 (11–12), 1702–1708.
- Fingar, V.H., Wieman, T.J., Wiehle, S.A., Cerrito, P.B., 1992. The role of microvascular damage in photodynamic therapy: the effect of treatment on vessel constriction, permeability, and leukocyte adhesion. *Cancer Res.* 52 (18), 4914–4921.
- Freitas, I., 1990. Lipid accumulation: the common feature to photosensitizer-retaining normal and malignant tissues [news]. *J. Photochem. Photobiol. B* 7 (2–4), 359–361.
- Friedberg, J.S., Mick, R., Stevenson, J.P., *et al.*, 2004. Phase II trial of pleural photodynamic therapy and surgery for patients with non-small-cell lung cancer with pleural spread. *J. Clin. Oncol.* 22 (11), 2192–2201.
- Frisoli, J.K., Tudor, E.G., Flotte, T.J., *et al.*, 1993. Pharmacokinetics of a fluorescent drug using laser-induced fluorescence. *Cancer Res.* 53 (24), 5954–5961.
- Frohlich, H., 1970. Long range coherence and the action of enzymes. *Nature* 228 (5276), 1093.
- Frohlich, H., 1975. The extraordinary dielectric properties of biological materials and the action of enzymes. *Proc. Natl Acad. Sci. USA* 72 (11), 4211–4215.
- Gad, F., Zahra, T., Francis, K.P., Hasan, T., Hamblin, M.R., 2004. Targeted photodynamic therapy of established soft-tissue infections in mice. *Photochem. Photobiol. Sci.* 3 (5), 451–458.
- Garg, A.D., Galluzzi, L., Apetoh, L., *et al.*, 2015. Molecular and translational classifications of DAMPs in immunogenic cell death. *Front. Immunol.* 6, 588.
- Garg, A.D., Krysko, D.V., Vandenabeele, P., Agostinis, P., 2011. DAMPs and PDT-mediated photo-oxidative stress: exploring the unknown. *Photochem. Photobiol. Sci.* 10 (5), 670–680.
- Geiger, P.G., Korytowski, W., Girotti, A.W., 1995. Photodynamically generated 3-beta-hydroxy-5 alpha-cholest-6-ene-5- hydroperoxide: toxic reactivity in membranes and susceptibility to enzymatic detoxification. *Photochem. Photobiol.* 62 (3), 580–587.
- Geneva, I.I., 2016. Photobiomodulation for the treatment of retinal diseases: a review. *Int. J. Ophthalmol.* 9 (1), 145–152.
- Geronemus, R.G., Weiss, R.A., Weiss, M.A., *et al.*, 2003. Non-ablative LED photomodulation light activated fibroblast stimulation clinical trial. *Lasers Surg. Med.* 25, 22.
- Giuliani, F., Martinelli, M., Cocchi, A., *et al.*, 2010. In vitro resistance selection studies of RLP068/Cl, a new Zn(II) phthalocyanine suitable for antimicrobial photodynamic therapy. *Antimicrob Agents Chemother.* 54 (2), 637–642.
- Gladwin, M.T., Shiva, S., 2009. The ligand binding battle at cytochrome c oxidase: how no regulates oxygen gradients in tissue. *Circ. Res.* 104 (10), 1136–1138.
- Goldman, L., Blaney, D.J., Kindel Jr., D.J., Franke, E.K., 1963. Effect of the laser beam on the skin. [Preliminary report] *J Invest Dermatol* 40, 121–122.
- Gollnick, S.O., Evans, S.S., Baumann, H., *et al.*, 2003. Role of cytokines in photodynamic therapy-induced local and systemic inflammation. *Br J Cancer* 88 (11), 1772–1779.
- Gollnick, S.O., Kabingu, E., Kousis, P.C., Henderson, B.W., 2004. Stimulation of the host immune response by photodynamic therapy (PDT). *Proc SPIE* 5319, 60–70.
- Gollnick, S.O., Owczarczak, B., Maier, P., 2006. Photodynamic therapy and anti-tumor immunity. *Lasers Surg Med* 38 (5), 509–515.
- Gollnick, S.O., Vaughan, L., Henderson, B.W., 2002. Generation of effective antitumor vaccines using photodynamic therapy. *Cancer Res.* 62 (6), 1604–1608.
- Gomer, C.J., Ferrario, A., Murphree, A.L., 1987. The effect of localized porphyrin photodynamic therapy on the induction of tumour metastasis. *Br J Cancer* 56 (1), 27–32.
- Gomer, C.J., Rucker, N., Murphree, A.L., 1988. Differential cell photosensitivity following porphyrin photodynamic therapy. *Cancer Res.* 48 (16), 4539–4542.
- Gomer, C.J., Ryter, S.W., Ferrario, A., *et al.*, 1996. Photodynamic therapy-mediated oxidative stress can induce expression of heat shock proteins. *Cancer Res.* 56 (10), 2355–2360.
- Gonzalez-Lima, F., Barrett, D.W., 2014. Augmentation of cognitive brain functions with transcranial lasers. *Front. Syst. Neurosci.* 8, 36.
- Goodson, W.H., Hunt, T.K., 1979. Wound healing and the diabetic patient. *Surg. Gynecol. Obstet.* 149, 600–608.
- Gorgey, A.S., Wade, A.N., Sobhi, N.N., 2008. The effect of low-level laser therapy on electrically induced muscle fatigue: A pilot study. *Photomed. Laser Surg* 26 (5), 501–506.
- Gougeon, M.L., Kroemer, G., 2003. Charming to death: caspase-dependent or -independent? *Cell Death Differ* 10 (3), 390–392.
- Gudgin Dickson, E.F., Holmes, H., Jori, G., *et al.*, 1995. On the source of the oscillations observed during in vivo zinc phthalocyanine fluorescence pharmacokinetic measurements in mice. *Photochem. Photobiol.* 61 (5), 506–509.
- Guyer, B., Freedman, M.A., Strobino, D.M., Sondik, E.J., 2000. Annual summary of vital statistics: Trends in the health of Americans during the 20th century. *Pediatrics* 106 (6), 1307–1317.
- Halliwell, B., Chirico, S., 1993. Lipid peroxidation: its mechanism, measurement, and significance. *Am. J. Clin. Nutr.* 57 (5 Suppl), 715S–724S. (discussion 724S–725S).
- Halliwell, B., Gutteridge, J.M.C., 2015. *Free Radicals in Biology and Medicine*, 5 edn, 1. Oxford: Oxford Press.
- Hamblin, M.R., Dai, T., 2010. Can surgical site infections be treated by photodynamic therapy? *Photodiagnosis Photodyn. Ther.* 7 (2), 134–136.
- Hamblin, M.R., Hasan, T., 2004. Photodynamic therapy: A new antimicrobial approach to infectious disease? *Photochem. Photobiol. Sci.* 3 (5), 436–450.
- Hamblin, M.R., Mroz, P., 2008. *Advances in Photodynamic Therapy: Basic, Translational and Clinical*. Norwood, MA: Artech House.
- Hamblin, M.R., Newman, E.L., 1994. On the mechanism of the tumour-localising effect in photodynamic therapy. *J Photochem. Photobiol. B* 23 (1), 3–8.
- Hashmi, J.T., Huang, Y.Y., Sharma, S.K., *et al.*, 2010. Effect of pulsing in low-level light therapy. *Lasers Surg. Med.* 42 (6), 450–466.
- Hawkins, D., Abrahamse, H., 2005. Biological effects of helium-neon laser irradiation on normal and wounded human skin fibroblasts. *Photomed. Laser Surg.* 23, 251–259.
- Hawkins, D., Houreld, N., Abrahamse, H., 2005. Low level laser therapy (LLL) as an effective therapeutic modality for delayed wound healing. *Ann. NY Acad. Sci.* 1056, 486–493.
- He, C., Agharkar, P., Chen, B., 2008. Intravital microscopic analysis of vascular perfusion and macromolecule extravasation after photodynamic vascular targeting therapy. *Pharmaceutical Res.* 25 (8), 1873–1880.
- Henderson, B.W., Fingar, V.H., 1989. Oxygen limitation of direct tumor cell kill during photodynamic treatment of a murine tumor model. *Photochem. Photobiol.* 49 (3), 299–304.
- Henderson, B.W., Gollnick, S.O., 2003. Mechanistic Principles of Photodynamic Therapy. In: Vo-Dinh, T. (Ed.), *Biomedical Photonics Handbook*. Boca Raton, FL: CRC Press, pp. 36.1–36.27.
- Henderson, B.W., Gollnick, S.O., Snyder, J.W., *et al.*, 2004. Choice of oxygen-conserving treatment regimen determines the inflammatory response and outcome of photodynamic therapy of tumors. *Cancer Res.* 64 (6), 2120–2126.
- Henderson, B.W., Miller, A.C., 1986. Effects of scavengers of reactive oxygen and radical species on cell survival following photodynamic treatment in vitro: comparison to ionizing radiation. *Radiat. Res.* 108 (2), 196–205.
- Henderson, B.W., Waldow, S.M., Mang, T.S., *et al.*, 1985. Tumor destruction and kinetics of tumor cell death in two experimental mouse tumors following photodynamic therapy. *Cancer Res.* 45 (2), 572–576.
- Henderson, T.A., Morris, L.D., 2015. SPECT perfusion imaging demonstrates improvement of traumatic brain injury with transcranial near-infrared laser phototherapy. *Adv. Mind Body Med.* 29 (4), 27–33.
- Hou, Y.C., Janczuk, A., Wang, P.G., 1999. Current trends in the development of nitric oxide donors. *Curr. Pharm. Des.* 5 (6), 417–441.

- Huang, Y.Y., Chen, A.C., Carroll, J.D., Hamblin, M.R., 2009. Biphasic dose response in low level light therapy. *Dose Response* 7 (4), 358–383.
- Huisa, B.N., Stemer, A.B., Walker, M.G., *et al.*, 2013. Transcranial laser therapy for acute ischemic stroke: a pooled analysis of NEST-1 and NEST-2. *Int. J. Stroke* 8 (5), 315–320.
- Hunt, D.W., Levy, J.G., 1998. Immunomodulatory aspects of photodynamic therapy. *Expert Opin. Investig. Drugs* 7 (1), 57–64.
- Y.D. Ignatov, A.I.V., Vlasov, T.D., *et al.*, 2005. Effects of helium-neon laser irradiation and local anesthetics on potassium channels in pond snail neurons. *Neurosci. Behav. Physiol.* 35, 871–875.
- Igney, F.H., Krammer, P.H., 2002a. Death and anti-death: Tumour resistance to apoptosis. *Nat. Rev. Cancer* 2 (4), 277–288.
- Igney, F.H., Krammer, P.H., 2002b. Immune escape of tumors: apoptosis resistance and tumor counterattack. *J. Leukoc. Biol.* 71 (6), 907–920.
- Ivancic, B.T., Ivancic, T., 2008. Low-level laser therapy improves vision in patients with age-related macular degeneration. *Photomed. Laser Surg.* 26 (3), 241–245.
- Ivancic, B.T., Ivancic, T., 2012. Low-level laser therapy improves visual acuity in adolescent and adult patients with amblyopia. *Photomed. Laser Surg.* 30 (3), 167–171.
- Ivancic, B.T., Ivancic, T., 2014. Low-level laser therapy improves vision in a patient with retinitis pigmentosa. *Photomed. Laser Surg.* 32 (3), 181–184.
- Jalili, A., Makowski, M., Switaj, T., *et al.*, 2004. Effective photoimmunotherapy of murine colon carcinoma induced by the combination of photodynamic therapy and dendritic cells. *Clin. Cancer Res.* 10 (13), 4498–4508.
- J.I., O., 2000. *Energy Medicine: The Scientific Basis*, 1st edn. Edinburgh: Churchill Livingstone.
- Johnstone, D.M., el Massri, N., Moro, C., *et al.*, 2014. Indirect application of near infrared light induces neuroprotection in a mouse model of parkinsonism – an abscopal neuroprotective effect. *Neuroscience* 274, 93–101.
- Johnstone, D.M., Moro, C., Stone, J., Benabid, A.L., Mitrofanis, J., 2015. Turning on lights to stop neurodegeneration: the potential of near infrared light therapy in Alzheimer's and Parkinson's disease. *Front. Neurosci.* 9, 500.
- Jori, G., 1989. In vivo transport and pharmacokinetic behavior of tumour photosensitizers. *Ciba Found. Symp.* 146, 78–86.
- Jori, G., Reddi, E., 1993. The role of lipoproteins in the delivery of tumour-targeting photosensitizers. *Int. J. Biochem.* 25 (10), 1369–1375.
- Kabingu, E., Oseroff, A.R., Wilding, G.E., Gollnick, S.O., 2009. Enhanced systemic immune reactivity to a Basal cell carcinoma associated antigen following photodynamic therapy. *Clin. Cancer Res.* 15 (13), 4460–4466.
- Kabingu, E., Vaughan, L., Owczarczak, B., Ramsey, K.D., Gollnick, S.O., 2007. CD8+ T cell-mediated control of distant tumours following local photodynamic therapy is independent of CD4+ T cells and dependent on natural killer cells. *Br. J. Cancer* 96 (12), 1839–1848.
- Kana, J.S., Hutschenreiter, G., Haina, D., Waidelich, W., 1981. Effect of low-power density laser radiation on healing of open skin wounds in rats. *Arch. Surg.* 116, 293–296.
- Karu, T.I., 2008. Mitochondrial signaling in mammalian cells activated by red and near-IR radiation. *Photochem. Photobiol.* 84 (5), 1091–1099.
- Karu, T.I., Pyatibrat, L.V., Afanasyeva, N.I., 2004. A novel mitochondrial signaling pathway activated by visible-to-near infrared radiation. *Photochem. Photobiol.* 80 (2), 366–372.
- Karu, T.I., Pyatibrat, L.V., Afanasyeva, N.I., 2005. Cellular effects of low power laser therapy can be mediated by nitric oxide. *Lasers Surg. Med.* 36 (4), 307–314.
- Kearns, D.R., 1971. Physical and chemical properties of singlet molecular oxygen. *Chem. Rev.* 71 (4), 32.
- Kessel, D., 2006. Death pathways associated with photodynamic therapy. *Med Laser Appl.* 21 (4), 219–224.
- Kessel, D., Arroyo, A.S., 2007. Apoptotic and autophagic responses to Bcl-2 inhibition and photodamage. *Photochem. Photobiol. Sci.* 6 (12), 1290–1295.
- Kessel, D., Oleinick, N.L., 2009. Initiation of autophagy by photodynamic therapy. *Methods Enzymol.* 453, 1–16.
- Kessel, D., Poretz, R.D., 2000. Sites of photodamage induced by photodynamic therapy with a chlorin e6 triacetoxymethyl ester (CAME). *Photochem. Photobiol.* 71 (1), 94–96.
- Kessel, D., Reiners Jr., J.J., 2007. Apoptosis and autophagy after mitochondrial or endoplasmic reticulum photodamage. *Photochem. Photobiol.* 83 (5), 1024–1028.
- Kessel, D., Vicente, M.G., Reiners Jr., J.J., 2006. Initiation of apoptosis and autophagy by photodynamic therapy. *Lasers Surg. Med.* 38 (5), 482–488.
- Khuman, J., Zhang, J., Park, J., *et al.*, 2012. Low-level laser light therapy improves cognitive deficits and inhibits microglial activation after controlled cortical impact in mice. *J. Neurotrauma* 29 (2), 408–417.
- Khurana, M., Moriyama, E.H., Mariampillai, A., Wilson, B.C., 2008. Intravital high-resolution optical imaging of individual vessel response to photodynamic treatment. *J. Biomed. Opt.* 13 (4), 040502.
- Kikuchi, T., Mogi, M., Okabe, I., *et al.*, 2015. Adjunctive application of antimicrobial photodynamic therapy in nonsurgical periodontal treatment: A review of literature. *Int. J. Mol. Sci.* 16 (10), 24111–24126.
- Kim, S.S., Park, M.W., Lee, C.J., 2007. Phototherapy of androgenetic alopecia with low level narrow band 655-nm red light and 780-nm infrared light. *J. Am. Acad. Dermatol.* 56, AB112. American Academy of Dermatology Proceedings of the 65th Annual Meeting.
- Kim, W.S., Calderhead, R.G., 2011. Is light-emitting diode phototherapy (LED-LLLT) really effective? *Laser Ther.* 20 (3), 205–215.
- Kitsis, R.N., Molkentin, J.D., 2010. Apoptotic cell death "Nixed" by an ER-mitochondrial necrotic pathway. *Proc. Natl Acad. Sci. USA* 107 (20), 9031–9032.
- Klaper, M., Fudickar, W., Linker, T., 2016. Role of distance in singlet oxygen applications: A model system. *J. Am. Chem. Soc.* 138 (22), 7024–7029.
- Kolari, P.J., Airaksinen, O., 1993. Poor penetration of infra-red and helium neon low power laser light into the dermal tissue. *Acupunct. Electrother. Res.* 18 (1), 17–21.
- Kongshaug, M., Moan, J., Brown, S.B., 1989. The distribution of porphyrins with different tumour localising ability among human plasma proteins. *Br. J. Cancer* 59 (2), 184–188.
- Korbelik, M., 1992. Low density lipoprotein receptor pathway in the delivery of Photofrin: How much is it relevant for selective accumulation of the photosensitizer in tumors? *J. Photochem. Photobiol. B* 12 (1), 107–109.
- Korbelik, M., 2006. PDT-associated host response and its role in the therapy outcome. *Lasers Surg. Med.* 38 (5), 500–508.
- Korbelik, M., Cecic, I., 1999. Contribution of myeloid and lymphoid host cells to the curative outcome of mouse sarcoma treatment by photodynamic therapy. *Cancer Lett.* 137 (1), 91–98.
- Korbelik, M., Cecic, I., 2003. Mechanism of tumor destruction by photodynamic therapy. In: Nalwa, H.S. (Ed.), *Handbook of Photochemistry and Photobiology*. Stevenson Ranch, CA, USA: American Scientific Publishers, pp. 39–77.
- Korbelik, M., Dougherty, G.J., 1999. Photodynamic therapy-mediated immune response against subcutaneous mouse tumors. *Cancer Res.* 59 (8), 1941–1946.
- Korbelik, M., Krosi, G., 1995. Photofrin accumulation in malignant and host cell populations of a murine fibrosarcoma. *Photochem. Photobiol.* 62 (1), 162–168.
- Korbelik, M., Krosi, G., Krosi, J., Dougherty, G.J., 1996. The role of host lymphoid populations in the response of mouse EMT6 tumor to photodynamic therapy. *Cancer Res.* 56 (24), 5647–5652.
- Korbelik, M., Merchant, S., Huang, N., 2009. Exploitation of immune response-eliciting properties of hypocrellin photosensitizer SL052-based photodynamic therapy for eradication of malignant tumors. *Photochem. Photobiol.* 85 (6), 1418–1424.
- Korbelik, M., Stott, B., Sun, J., 2007. Photodynamic therapy-generated vaccines: Relevance of tumour cell death expression. *Br J Cancer* 97 (10), 1381–1387.
- Korbelik, M., Sun, J., 2006. Photodynamic therapy-generated vaccine for cancer therapy. *Cancer Immunol. Immunother.* 55 (8), 900–909.
- Korbelik, M., Sun, J., Cecic, I., 2005. Photodynamic therapy-induced cell surface expression and release of heat shock proteins: Relevance for tumor response. *Cancer Res.* 65 (3), 1018–1026.
- Kousis, P.C., Henderson, B.W., Maier, P.G., Gollnick, S.O., 2007. Photodynamic therapy6 enhancement of antitumor immunity is regulated by neutrophils. *Cancer Res.* 67 (21), 10501–10510.
- Kraus, C.N., 2008. Low hanging fruit in infectious disease drug development. *Curr. Opin. Microbiol.* 11 (5), 434–438.
- Kroemer, G., Jaattela, M., 2005. Lysosomes and autophagy in cell death control. *Nat Rev Cancer* 5 (11), 886–897.
- Krosi, G., Korbelik, M., Dougherty, G.J., 1995. Induction of immune cell infiltration into murine SCCVII tumour by photofrin-based photodynamic therapy. *Br. J. Cancer* 71 (3), 549–555.

- Kucuk, B.B., Oral, K., Selcuk, N.A., Toklu, T., Civi, O.G., 2010. The anti-inflammatory effect of low-level laser therapy on experimentally induced inflammation of rabbit temporomandibular joint retrodiscal tissues. *J. Orofac. Pain* 24 (3), 293–297.
- Lamp, Y., Zivin, J.A., Fisher, M., *et al.*, 2007. Infrared laser therapy for ischemic stroke: A new treatment strategy: Results of the NeuroThera Effectiveness and Safety Trial-1 (NEST-1). *Stroke* 38 (6), 1843–1849.
- Lapchak, P.A., 2013a. Fast neuroprotection (fast-NPRX) for acute ischemic stroke victims: The time for treatment is now. *Transl. Stroke Res.* 4 (6), 704–709.
- Lapchak, P.A., 2013b. Recommendations and practices to optimize stroke therapy: Developing effective translational research programs. *Stroke* 44 (3), 841–843.
- Lapchak, P.A., Boitano, P.D., 2016. Transcranial near-infrared laser therapy for stroke: How to recover from futility in the NEST-3 clinical trial. *Acta Neurochir. Suppl.* 121, 7–12.
- Lapchak, P.A., Zhang, J.H., Noble-Haeusslein, L.J., 2013. RIGOR guidelines: Escalating STAIR and STEPS for effective translational research. *Transl. Stroke Res.* 4 (3), 279–285.
- Larroque, C., Pelegrin, A., Van Lier, J.E., 1996. Serum albumin as a vehicle for zinc phthalocyanine: Photodynamic activities in solid tumour models. *Br J Cancer* 74 (12), 1886–1890.
- Lavie, G., Kaplinsky, C., Toren, A., *et al.*, 1999. A photodynamic pathway to apoptosis and necrosis induced by dimethyl tetrahydroxyhelianthone and hypericin in leukaemic cells: Possible relevance to photodynamic therapy. *Br. J. Cancer* 79 (3–4), 423–432.
- Lavi, R., Shainberg, A., Shneyvays, V., *et al.*, 2010. Detailed analysis of reactive oxygen species induced by visible light in various cell types. *Lasers Surg. Med.* 42 (6), 473–480.
- Leal Junior, E.C., Lopes-Martins, R.A., Dalan, F., *et al.*, 2008. Effect of 655-nm low-level laser therapy on exercise-induced skeletal muscle fatigue in humans. *Photomed. Laser Surg.* 26 (5), 419–424.
- Leal Junior, E.C., Lopes-Martins, R.A., de Almeida, P., *et al.*, 2010a. Effect of low-level laser therapy (GaAs 904 nm) in skeletal muscle fatigue and biochemical markers of muscle damage in rats. *Eur. J. Appl. Physiol.* 108 (6), 1083–1088.
- Leal Junior, E.C., Lopes-Martins, R.A., Frigo, L., *et al.*, 2010b. Effects of low-level laser therapy (LLLT) in the development of exercise-induced skeletal muscle fatigue and changes in biochemical markers related to postexercise recovery. *J. Orthop. Sports Phys. Ther.* 40 (8), 524–532.
- Leavitt, M., Charles, G., Heyman, E., Michaels, D., 2009. HairMax Laser comb laser phototherapy device in the treatment of male androgenetic alopecia: A randomized, double-blind, sham device-controlled, multicentre trial. *Clin. Drug Investig.* 29 (5), 283–292.
- Lee, J.H., Chang, S.Y., Moy, W.J., *et al.*, 2016. Simultaneous bilateral laser therapy accelerates recovery after noise-induced hearing loss in a rat model. *Peer J.* 4, e2252.
- Lee, S.Y., Park, K.H., Choi, J.W., *et al.*, 2007. A prospective, randomized, placebo-controlled, double-blinded, and split-face clinical study on LED phototherapy for skin rejuvenation: Clinical, profilometric, histologic, ultrastructural, and biochemical evaluations and comparison of three different treatment settings. *J. Photochem. Photobiol. B* 88 (1), 51–67.
- Little, F.M., Gomer, C.J., Hyman, S., Apuzzo, M.L., 1984. Observations in studies of quantitative kinetics of tritium labelled hematoporphyrin derivatives (HpDI and HpDII) in the normal and neoplastic rat brain model. *J. Neurooncol.* 2 (4), 361–370.
- Liu, X.G., Zhou, Y.J., Liu, T.C., Yuan, J.Q., 2009. Effects of low-level laser irradiation on rat skeletal muscle injury after eccentric exercise. *Photomed. Laser Surg.* 27 (6), 863–869.
- Loevschall, H., Arenholt-Bindslev, D., 1994. Effect of low level diode laser irradiation of human oral mucosa fibroblasts in vitro. *Lasers Surg. Med.* 14, 347–354.
- Lopes-Martins, R.A., Marcos, R.L., Leonardo, P.S., *et al.*, 2006. Effect of low-level laser (Ga-Al-As 655 nm) on skeletal muscle fatigue induced by electrical stimulation in rats. *J. Appl. Physiol.* (1985) 101 (1), 283–288.
- Lubart, R., Friedmann, H., 2011. LLLT and PDT. *Laser Ther.* 20 (3), 233.
- Lundeberg, T., Malm, M., 1991. Low-power HeNe laser treatment of venous leg ulcers. *Ann. Plast. Surg.* 27, 537–539.
- Madar-Balakirski, N., Tempel-Brami, C., Kalchenko, V., *et al.*, 2010. Permanent occlusion of feeding arteries and draining veins in solid mouse tumors by vascular targeted photodynamic therapy (VTP) with Tookad. *PLOS ONE* 5 (4), e10282.
- Maeda, H., Tsukigawa, K., Fang, J., 2016. A retrospective 30 years after discovery of the enhanced permeability and retention effect of solid tumors: Next-generation chemotherapeutics and photodynamic therapy – problems, solutions, and prospects. *Microcirculation* 23 (3), 173–182.
- Maeurer, M.J., Gollin, S.M., Storkus, W.J., *et al.*, 1996. Tumor escape from immune recognition: Loss of HLA-A2 melanoma cell surface expression is associated with a complex rearrangement of the short arm of chromosome 6. *Clin. Cancer Res.* 2 (4), 641–652.
- Magna, M., Pisetsky, D.S., 2016. The alarmin properties of DNA and DNA-associated nuclear proteins. *Clin. Ther.* 38 (5), 1029–1041.
- Maiman, T.H., 1960. Stimulated optical radiation in ruby. *Nature* 187 (4736), 493–494.
- Maisch, T., 2007. Anti-microbial photodynamic therapy: useful in the future? *Lasers Med. Sci.* 22 (2), 83–91.
- Maisch, T., 2015. Resistance in antimicrobial photodynamic inactivation of bacteria. *Photochem. Photobiol. Sci.* 14 (8), 1518–1526.
- Maisch, T., Hackbarth, S., Regensburger, J., *et al.*, 2011. Photodynamic inactivation of multi-resistant bacteria (PIB) – a new approach to treat superficial infections in the 21st century. *J. Deutschen Dermatologischen Gesellschaft = J. German Soc. Dermatol.: JDDG* 9 (5), 360–366.
- Maksimovich, I.V., 2015. Dementia and cognitive impairment reduction after laser transcatheter treatment of Alzheimer's disease. *World J. Neurosci.* 5, 189–203.
- Malik, Z., Ladani, H., Nitzan, Y., 1992. Photodynamic inactivation of Gram-negative bacteria: problems and possible solutions. *J. Photochem. Photobiol. B* 14 (3), 262–266.
- Maloney, R., Shanks, S., Maloney, J., 2010. The application of low-level laser therapy For the symptomatic care of late stage Parkinson's disease: A non-controlled, non-randomized study (abstract). *Lasers Surg. Med.* 185.
- Mathew, R., Karantz-Wadsworth, V., White, E., 2007. Role of autophagy in cancer. *Nat. Rev. Cancer.* 7 (12), 961–967.
- Maugain, E., Sasnouski, S., Zorin, V., *et al.*, 2004. Foscan-based photodynamic treatment in vivo: correlation between efficacy and Foscan accumulation in tumor, plasma and leukocytes. *Oncol. Rep* 12 (3), 639–645.
- Maziere, J.C., Santus, R., Morliere, P., *et al.*, 1990. Cellular uptake and photosensitizing properties of anticancer porphyrins in cell membranes and low and high density lipoproteins. *J. Photochem. Photobiol. B* 6 (1–2), 61–68.
- McClure, J., 2011. The role of causal attributions in public misconceptions about brain injury. *Rehab. Psychol.* 56 (2), 85–93.
- McDaniel, D.H., Newman, J., Geronemus, R., *et al.*, 2003. Non-ablative non-thermal LED photomodulation, a multicenter clinical photoaging trial. *Lasers Surg. Med.* 15, 22.
- McGuff, P.E., Deterling Jr., R.A., Gottlieb, L.S., 1965. Tumorcidal effect of laser energy on experimental and human malignant tumors. *New England J. Med.* 273 (9), 490–492.
- Medrado, A.R., Pugliese, L.S., Reis, S.R., Andrade, Z.A., 2003. Influence of low level laser therapy on wound healing and its biological action upon myofibroblasts. *Lasers Surg. Med* 32, 239–244.
- Meneguzzo, D.T., Lopes, L.A., Pallota, R., *et al.*, 2012. Prevention and treatment of mice paw edema by near-infrared low-level laser therapy on lymph nodes. *Lasers Med. Sci.*
- Merchat, M., Spikes, J.D., Bertoloni, G., Jori, G., 1996. Studies on the mechanism of bacteria photosensitization by meso-substituted cationic porphyrins. *J. Photochem. Photobiol. B* 35 (3), 149–157.
- Mester, E., Ludany, G., Sellyei, M., *et al.*, 1968a. Studies on the inhibiting and activating effects of laser beams. *Langenbecks Arch. Chir.* 322, 1022–1027.
- Mester, E., Ludany, G., Sellyei, M., Szende, B., Tota, J., 1968b. The stimulating effect of low power laser rays on biological systems. *Laser Rev.* 1, 3.
- Mester, E., Nagylucskay, S., Doklen, A., Tisza, S., 1976. Laser stimulation of wound healing. *Acta Chir. Acad. Sci. Hung.* 17, 49–55.
- Mester, E., Spiry, T., Szende, B., Tota, J.G., 1971. Effect of laser rays on wound healing. *Am. J. Surg.* 122, 532–535.
- Mester, E., Szende, B., Gartner, P., 1968c. The effect of laser beams on the growth of hair in mice. *Radiobiol. Radiother. (Berl)* 9 (5), 621–626.
- Mester, E., Szende, B., Spiry, T., Scher, A., 1972. Stimulation of wound healing by laser rays. *Acta Chir. Acad. Sci. Hung.* 13 (3), 315–324.

- Meyers, A.D., 1990. Lasers and wound healing. *Arch. Otolaryngol. Head Neck Surg.* 116, 1128.
- Michalikova, S., Ennaceur, A., van Rensburg, R., Chazot, P.L., 2008. Emotional responses and memory performance of middle-aged CD1 mice in a 3D maze: Effects of low infrared light. *Neurobiol. Learn Mem.* 89 (4), 480–488.
- Minnock, A., Vernon, D.I., Schofield, J., *et al.*, 1996. Photoinactivation of bacteria. Use of a cationic water-soluble zinc phthalocyanine to photoinactivate both gram-negative and gram-positive bacteria. *J. Photochem. Photobiol. B* 32 (3), 159–164.
- Miura, Y.Y.M., Tsuboi, R., Ogawa, H., 1999. Promotion of rat hair growth by irradiation using Super Lizer™. *Jpn. J. Dermatol.* 109 (13), 2149–2152.
- Moan, J., Berg, K., 1991. The photodegradation of porphyrins in cells can be used to estimate the lifetime of singlet oxygen. *Photochem. Photobiol.* 53 (4), 549–553.
- Moges, H., Vasconcelos, O.M., Campbell, W.W., *et al.*, 2009. Light therapy and supplementary Riboflavin in the SOD1 transgenic mouse model of familial amyotrophic lateral sclerosis (FALS). *Lasers Surg. Med.* 41, 52–59.
- Moreno-Arias, G.A., Castelo-Branco, C., Ferrando, J., 2002b. Side-effects after IPL photodepilation. *Dermatol. Surg.* 28 (12), 1131–1134.
- Moreno-Arias, G., Castelo-Branco, C., Ferrando, J., 2002a. Paradoxical effect after IPL photoepilation. *Dermatol. Surg.* 28 (11), 1013–1016. (discussion 1016).
- Moro, C., Massri, N.E., Darlot, F., *et al.*, 2016. Effects of a higher dose of near-infrared light on clinical signs and neuroprotection in a monkey model of Parkinson's disease. *Brain Res.*
- Moro, C., Massri, N.E., Torres, N., *et al.*, 2014. Photobiomodulation inside the brain: a novel method of applying near-infrared light intracranially and its impact on dopaminergic cell survival in MPTP-treated mice. *J. Neurosurg.* 120 (3), 670–683.
- Mroz, P., Szokalska, A., Wu, M.X., Hamblin, M.R., 2010. Photodynamic therapy of tumors can lead to development of systemic antigen-specific immune response. *PLoS One* 5 (12), e15194.
- Mroz, P., Vatansever, F., Muchowicz, A., Hamblin, M.R., 2013. Photodynamic therapy of murine mastocytoma induces specific immune responses against the cancer/testis antigen P1A. *Cancer Res.*
- Munoz-Price, L.S., Poirel, L., Bonomo, R.A., *et al.*, 2013. Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases. *Lancet Infectious Dis.* 13 (9), 785–796.
- Naeser, M.A., Hamblin, M.R., 2015. Traumatic brain injury: a major medical problem that could be treated using transcranial, red/near-infrared LED photobiomodulation. *Photomed. Laser Surg.*
- Naeser, M.A., Martin, P.I., Lundgren, K., *et al.*, 2010. Improved language in a chronic nonfluent aphasia patient after treatment with CPAP and TMS. *Cogn. Behav. Neurol.* 23 (1), 29–38.
- Naeser, M.A., Saltmarche, A., Krengel, M.H., Hamblin, M.R., Knight, J.A., 2011. Improved cognitive function after transcranial, light-emitting diode treatments in chronic, traumatic brain injury: Two case reports. *Photomed. Laser Surg.* 29 (5), 351–358.
- Naeser, M.A., Zafonte, R., Krengel, M.H., *et al.*, 2014. Significant improvements in cognitive performance post-transcranial, red/near-infrared light-emitting diode treatments in chronic, mild traumatic brain injury: Open-protocol study. *J. Neurotrauma* 31 (11), 1008–1017.
- Naeser, M., Ho, M., Martin, P.E., *et al.*, 2012. Improved language after scalp application of red/near-infrared light-emitting diodes: pilot study supporting a new, noninvasive treatment for chronic aphasia. *Procedia – Social Behav. Sci.* 61, 138–139.
- Natoli, R., Valtter, K., Barbosa, M., *et al.*, 2013. 670nm photobiomodulation as a novel protection against retinopathy of prematurity: evidence from oxygen induced retinopathy models. *PLOS ONE* 8 (8), e72135.
- Neira, R., *et al.* 2000. Low level assisted lipoplasty: A new technique. In: *Proceedings of the World Congress on Liposuction*, Dearborn, MI.
- Neira, R., Arroyave, J., Ramirez, H., *et al.*, 2002. Fat liquefaction: effect of low-level laser energy on adipose tissue. *Plast Reconstr Surg* 110 (3), 912–922. (discussion 923–5).
- Nitzan, Y., Balzam-Sudakevitz, A., Ashkenazi, H., 1998. Eradication of *Acinetobacter baumannii* by photosensitized agents in vitro. *J. Photochem. Photobiol. B* 42 (3), 211–218.
- Nitzan, Y., Gutterman, M., Malik, Z., Ehrenberg, B., 1992. Inactivation of gram-negative bacteria by photosensitized porphyrins. *Photochem. Photobiol.* 55 (1), 89–96.
- Noble, P.B., Shields, E.D., Blecher, P.D., Bentley, K.C., 1992. Locomotory characteristics of fibroblasts within a three-dimensional collagen lattice: modulation by a Helium/Neon soft laser. *Lasers Surg. Med.* 12, 669–674.
- Okada, H., Mak, T.W., 2004. Pathways of apoptotic and non-apoptotic death in tumour cells. *Nat. Rev. Cancer* 4 (8), 592–603.
- Oleinick, N.L., Evans, H.H., 1998. The photobiology of photodynamic therapy: Cellular targets and mechanisms. *Radiat. Res.* 150 (5 Suppl), S146–S156.
- Oleinick, N.L., Morris, R.L., Belichenko, I., 2002. The role of apoptosis in response to photodynamic therapy: What, where, why, and how. *Photochem Photobiol Sci* 1 (1), 1–21.
- O'Neill, J., 2015. Tackling a global health crisis: Initial steps The Review on Antimicrobial Resistance Chaired by Jim O'Neill 2015.
- Oron, A., Oron, U., Chen, J., *et al.*, 2006. Low-level laser therapy applied transcranially to rats after induction of stroke significantly reduces long-term neurological deficits. *Stroke* 37 (10), 2620–2624.
- Oron, A., Oron, U., Streeter, J., *et al.*, 2007. Low-level laser therapy applied transcranially to mice following traumatic brain injury significantly reduces long-term neurological deficits. *J. Neurotrauma* 24 (4), 651–656.
- Peoples, C., Spana, S., Ashkan, K., *et al.*, 2012. Photobiomodulation enhances nigral dopaminergic cell survival in a chronic MPTP mouse model of Parkinson's disease. *Parkinsonism Relat. Disord.* 18 (5), 469–476.
- Peplow, P.V., 2015. Neuroimmunomodulatory effects of transcranial laser therapy combined with intravenous tPA administration for acute cerebral ischemic injury. *Neural Regen. Res.* 10 (8), 1186–1190.
- Pereira, A.N., Eduardo Cde, P., Matson, E., Marques, M.M., 2002. Effect of low-power laser irradiation on cell growth and procollagen synthesis of cultured fibroblasts. *Lasers Surg. Med.* 31, 263–267.
- Perfettini, J.L., Kroemer, G., 2003. Caspase activation is not death. *Nat. Immunol.* 4 (4), 308–310.
- Pettigrew, C.A., Cotter, T.G., 2009. Deregulation of cell death (apoptosis): implications for tumor development. *Discov. Med.* 8 (41), 61–63.
- Pottier, R., Kennedy, J.C., 1990. The possible role of ionic species in selective biodistribution of photochemotherapeutic agents toward neoplastic tissue. *J. Photochem. Photobiol. B* 8 (1), 1–16.
- PR, K., 1989. Low level laser therapy: a review. *Laser Med. Sci.* 4 (3), 141–150.
- Proctor, P.H., 1989. Endothelium-derived relaxing factor and minoxidil: Active mechanisms in hair growth. *Arch. Dermatol.* 125 (8), 1146.
- Proskuryakov, S.Y., Gabai, V.L., 2010. Mechanisms of tumor cell necrosis. *Curr. Pharm. Des.* 16 (1), 56–68.
- Purushothuman, S., Nandasena, C., Johnstone, D.M., Stone, J., Mitrofanis, J., 2013. The impact of near-infrared light on dopaminergic cell survival in a transgenic mouse model of parkinsonism. *Brain Res.* 1535, 61–70.
- Qu, C., Cao, W., Fan, Y., Lin, Y., 2010. Near-infrared light protect the photoreceptor from light-induced damage in rats. *Adv. Exp. Med. Biol.* 664, 365–374.
- Quirk, B.J., Torbey, M., Buchmann, E., Verma, S., Whelan, H.T., 2012. Near-infrared photobiomodulation in an animal model of traumatic brain injury: improvements at the behavioral and biochemical levels. *Photomed. Laser Surg.* 30 (9), 523–529.
- Raskin, P., Marks, J.F., Burns, H., Plumer, M.D., Siperstein, M.D.L., 1975. Capillary basement membrane within diabetic children. *Am. J. Med.* 58, 365–375.
- Rathmell, J.C., Thompson, C.B., 1999. The central effectors of cell death in the immune system. *Annu. Rev. Immunol.* 17, 781–828.
- Reddy, G.K., Stehno-Bittel, L., Enwemeka, C.S., 2001. Laser photostimulation accelerates wound healing in diabetic rats. *Wound Repair Regen.* 9, 248–255.
- Reiners Jr., J.J., Agostinis, P., Berg, K., Oleinick, N.L., Kessel, D., 2010. Assessing autophagy in the context of photodynamic therapy. *Autophagy* 6 (1), 7–18.
- Reis e Sousa, C., 2004. Activation of dendritic cells: Translating innate into adaptive immunity. *Curr. Opin. Immunol.* 16 (1), 21–25.
- Glen Calderhead, R., Kubota, J., Trelles, M.A., Ohshiro, T., 2008. One mechanism behind LED phototherapy for wound healing and skin rejuvenation: Key role of the mast cell. *Laser Ther.* 17, 141–148.

- Rhee, C.K., He, P., Jung, J.Y., *et al.*, 2013. Effect of low-level laser treatment on cochlea hair-cell recovery after ototoxic hearing loss. *J. Biomed. Opt.* 18 (12), 128003.
- Richter, A.M., Cerruti-Sola, S., Sternberg, E.D., Dolphin, D., Levy, J.G., 1990. Biodistribution of tritiated benzoporphyrin derivative (3H-BPD-MA), a new potent photosensitizer, in normal and tumor-bearing mice. *J. Photochem. Photobiol. B* 5 (2), 231–244.
- Rojas, J.C., Bruchey, A.K., Gonzalez-Lima, F., 2012. Low-level light therapy improves cortical metabolic capacity and memory retention. *J. Alzheimers Dis.* 32 (3), 741–752.
- Rossi, A., Cantisani, C., Melis, L., *et al.*, 2012. Minoxidil use in dermatology, side effects and recent patents. *Recent Pat. Inflamm. Allergy Drug Discov.* 6 (2), 130–136.
- Russell, B.A., Kellett, N., Reilly, L.R., 2005. A study to determine the efficacy of combination LED light therapy (633 nm and 830 nm) in facial skin rejuvenation. *J. Cosmet. Laser Ther.* 7 (3–4), 196–200.
- Rustin, P., 2002. Mitochondria, from cell death to proliferation. *Nat. Genet.* 30 (4), 352–353.
- Ryter, S.W., Gomer, C.J., 1993. Nuclear factor kappa B binding activity in mouse L1210 cells following photofrin II-mediated photosensitisation. *Photochem. Photobiol.* 58, 753–756.
- Salehpour, F., Rasta, S.H., Mohaddes, G., Sadigh-Eteghad, S., Salarirad, S., 2016. Therapeutic effects of 10-Hz pulsed wave lasers in rat depression model: A comparison between near-infrared and red wavelengths. *Lasers Surg. Med.*
- Salgado, A.S., Zangaro, R.A., Parreira, R.B., Kerppers, I.I., 2015. The effects of transcranial LED therapy (TCLT) on cerebral blood flow in the elderly women. *Lasers Med. Sci.* 30 (1), 339–346.
- Saltmarche, A.E., Naeser, M.A., Ho, K.F., Hamblin, M.R., Lim, L., 2016. Significant improvement in cognition after transcranial and intranasal photobiomodulation: A controlled, single-blind pilot study in participants with dementia (Abstract). In: Alzheimer's Association International Conference, Toronto, Canada.
- Sanchez-Chavez, G., Pena-Rangel, M.T., Riesgo-Escovar, J.R., Martinez-Martinez, A., Salceda, R., 2012. Insulin stimulated-glucose transporter Glut 4 is expressed in the retina. *PLOS ONE* 7 (12), e52959.
- Sandford, M.A., Walsh, L.J., 1994. Thermal effects during desensitisation of teeth with gallium-aluminium-arsenide lasers. *Periodontology* 15, 25–30.
- Santos, L.A., Marcos, R.L., Tomazoni, S.S., *et al.*, 2014. Effects of pre-irradiation of low-level laser therapy with different doses and wavelengths in skeletal muscle performance, fatigue, and skeletal muscle damage induced by tetanic contractions in rats. *Lasers Med. Sci.* 29 (5), 1617–1626.
- Sasnauskienė, A., Kadziauskas, J., Vezelyte, N., Jonusiene, V., Kirveliėne, V., 2009. Apoptosis, autophagy and cell cycle arrest following photodamage to mitochondrial interior. *Apoptosis* 14 (3), 276–286.
- Satino, J.L.M.M., 2003. Hair regrowth and increased hair tensile strength using the hair max laser comb for low-level laser therapy. *Int. J. Cos. Surg. Aest. Dermatol.* 5, 113–117.
- Savill, J., Fadok, V., 2000. Corpse clearance defines the meaning of cell death. *Nature* 407, 784–788.
- Scherz-Shouval, R., Elazar, Z., 2007. ROS, mitochondria and the regulation of autophagy. *Trends Cell. Biol.* 17 (9), 422–427.
- Scherz-Shouval, R., Shvets, E., Fass, E., *et al.*, 2007. Reactive oxygen species are essential for autophagy and specifically regulate the activity of Atg4. *EMBO J.* 26 (7), 1749–1760.
- Schiffer, F., Johnston, A.L., Ravichandran, C., *et al.*, 2009. Psychological benefits 2 and 4 weeks after a single treatment with near infrared light to the forehead: A pilot study of 10 patients with major depression and anxiety. *Behav. Brain Funct.* 5, 46.
- Schindl, A., Heinze, G., Schindl, M., Pernerstorfer-Schon, H., Schindl, L., 2002. Systemic effects of low-intensity laser irradiation on skin microcirculation in patients with diabetic microangiopathy. *Microvasc. Res.* 64, 240–246.
- Schindl, A., Schindl, M., Pernerstorfer-Schon, H., 2000. Low intensity laser irradiation in the treatment of recalcitrant radiation ulcers in patients with breast cancer – long-term results of 3 cases. *Photodermatol. Photoimmunol. Photomed.* 16, 34–37.
- Schindl, A., Schindl, M., Schon, H., *et al.*, 1998. Low-intensity laser irradiation improves skin circulation in patients with diabetic microangiopathy. *Diabetes Care* 21, 580–584.
- Schuitmaker, J.J., Feitsma, R.I., Journee-De Korver, J.G., Dubbelman, T.M., Pauwels, E.K., 1993. Tissue distribution of bacteriochlorin a labelled with ^{99m}Tc-perthetate in hamster Greene melanoma. *Int. J. Radiat. Biol.* 64 (4), 451–458.
- Shaw, V.E., Spana, S., Ashkan, K., *et al.*, 2010. Neuroprotection of midbrain dopaminergic cells in MPTP-treated mice after near-infrared light treatment. *J. Comp. Neurol.* 518 (1), 25–40.
- Sheng, C., Pogue, B.W., Wang, E., Hutchins, J.E., Hoopes, P.J., 2004. Assessment of photosensitizer dosimetry and tissue damage assay for photodynamic therapy in advanced-stage tumors. *Photochem. Photobiol.* 79 (6), 520–525.
- Skinner, S.M., Gage, J.P., Wilce, P.A., Shaw, R.M., 1996. A preliminary study of the effects of laser radiation on collagen metabolism in cell culture. *Aust. Dent. J.* 41, 188–192.
- Smith, R., Coast, J., 2013. The true cost of antimicrobial resistance. *BMJ* 346, f1493.
- Spanheimer, R.G., Umpierrez, G.E., Stumpf, V., 1988. Decreased collagen production in diabetic rats. *Diabetes* 37, 371–376.
- Sroka, R., Fuchs, C., Schaffer, M., *et al.*, 1997. Biomodulation effects on cell mitosis after laser irradiation using different wavelengths. *Laser Surg. Med. Supplement* (9), 6.
- Star, W.M., Marijnissen, H.P., van den Berg-Blok, A.E., *et al.*, 1986. Destruction of rat mammary tumor and normal tissue microcirculation by hematoporphyrin derivative photoradiation observed in vivo in sandwich observation chambers. *Cancer Res.* 46 (5), 2532–2540.
- Stott, B., Korbelik, M., 2007. Activation of complement C3, C5, and C9 genes in tumors treated by photodynamic therapy. *Cancer Immunol. Immunother.* 56 (5), 649–658.
- Sun, J., Cecic, I., Parkins, C.S., Korbelik, M., 2002. Neutrophils as inflammatory and immune effectors in photodynamic therapy-treated mouse SCCVII tumours. *Photochem. Photobiol. Sci.* 1 (9), 690–695.
- Sur, B.W., Nguyen, P., Sun, C.H., Tromberg, B.J., Nelson, E.L., 2008. Immunophototherapy using PDT combined with rapid intratumoral dendritic cell injection. *Photochem. Photobiol.* 84 (5), 1257–1264.
- Sussai, D.A., Carvalho Pde, T., Dourado, D.M., *et al.*, 2010. Low-level laser therapy attenuates creatine kinase levels and apoptosis during forced swimming in rats. *Lasers Med. Sci.* 25 (1), 115–120.
- Tafur, J., Mills, P.J., 2008. Low-intensity light therapy: Exploring the role of redox mechanisms. *Photomed. Laser Surg.* 26 (4), 323–328.
- Tamura, A., Matsunobu, T., Mizutani, K., *et al.*, 2015. Low-level laser therapy for prevention of noise-induced hearing loss in rats. *Neurosci. Lett.* 595, 81–86.
- Tanaka, M., Mroz, P., Dai, T., *et al.*, 2012. Photodynamic therapy can induce a protective innate immune response against murine bacterial arthritis via neutrophil accumulation. *PLOS ONE* 7 (6), e39823.
- Tang, J., Du, Y., Lee, C.A., *et al.*, 2013. Low-intensity far-red light inhibits early lesions that contribute to diabetic retinopathy: In vivo and in vitro. *Invest. Ophthalmol. Vis. Sci.* 54 (5), 3681–3690.
- Tang, J., Herda, A.A., Kern, T.S., 2014. Photobiomodulation in the treatment of patients with non-center-involving diabetic macular oedema. *Br. J. Ophthalmol.* 98 (8), 1013–1015.
- Tauber, S., Schorn, K., Beyer, W., Baumgartner, R., 2003. Transmeatal cochlear laser (TCL) treatment of cochlear dysfunction: a feasibility study for chronic tinnitus. *Lasers Med. Sci.* 18 (3), 154–161.
- Teggi, R., Bellini, C., Fabiano, B., Bussi, M., 2008. Efficacy of low-level laser therapy in Meniere's disease: a pilot study of 10 patients. *Photomed. Laser Surg.* 26 (4), 349–353.
- Thomas, D.W., O'Neill, I.D., Harding, K.G., Shepherd, J.P., 1995. Cutaneous wound healing: A current perspective. *J. Oral Maxillofac. Surg.* 53, 442–447.
- Thong, P.S., Olivo, M., Kho, K.W., *et al.*, 2008. Immune response against angiosarcoma following lower fluence rate clinical photodynamic therapy. *J. Environ. Pathol. Toxicol. Oncol.* 27 (1), 35–42.
- Thong, P.S., Ong, K.W., Goh, N.S., *et al.*, 2007. Photodynamic-therapy-activated immune response against distant untreated tumours in recurrent angiosarcoma. *Lancet Oncol* 8 (10), 950–952.

- Trueb, R.M., 2009. Chemotherapy-induced alopecia. *Semin. Cutan Med. Surg.* 28 (1), 11–14.
- Tseng, M.T., Reed, M.W., Ackermann, D.M., *et al.*, 1988. Photodynamic therapy induced ultrastructural alterations in microvasculature of the rat cremaster muscle. *Photochem. Photobiol.* 48 (5), 675–681.
- Tuby, H., Maltz, L., Oron, U., 2006. Modulations of VEGF and iNOS in the rat heart by low level laser therapy are associated with cardioprotection and enhanced angiogenesis. *Lasers Surg. Med.* 38 (7), 682–688.
- Vabulas, R.M., Wagner, H., Schild, H., 2002. Heat shock proteins as ligands of toll-like receptors. *Curr. Top. Microbiol. Immunol.* 270, 169–184.
- Valduga, G., Bertoloni, G., Reddi, E., Jori, G., 1993. Effect of extracellularly generated singlet oxygen on gram-positive and gram-negative bacteria. *J. Photochem. Photobiol. B* 21 (1), 81–86.
- van Breugel, H.H., Bar, P.R., 1992. Power density and exposure time of He-Ne laser irradiation are more important than total energy dose in photo-biomodulation of human fibroblasts in vitro. *Lasers Surg. Med.* 12 (5), 528–537.
- van Duijnhoven, F.H., Aalbers, R.I., Rovers, J.P., Terpstra, O.T., Kuppen, P.J., 2003. Immunological aspects of photodynamic therapy of liver tumors in a rat model for colorectal cancer. *Photochem. Photobiol.* 78 (3), 235–240.
- Vera, D.M., Haynes, M.H., Ball, A.R., *et al.*, 2012. Strategies to potentiate antimicrobial photoinactivation by overcoming resistant phenotypes. *Photochem. Photobiol.* 88 (3), 499–511.
- Vieira, W.H., Ferraresi, C., Perez, S.E., Baldissera, V., Parizotto, N.A., 2012. Effects of low-level laser therapy (808 nm) on isokinetic muscle performance of young women submitted to endurance training: A randomized controlled clinical trial. *Lasers Med. Sci.* 27 (2), 497–504.
- Vinck, E., Cagnie, B., Coorevits, P., Vanderstraeten, G., Cambier, D., 2006. Pain reduction by infrared light-emitting diode irradiation: A pilot study on experimentally induced delayed-onset muscle soreness in humans. *Lasers Med. Sci.* 21 (1), 11–18.
- Vlachos, S.P., Kontoes, P.P., 2002. Development of terminal hair following skin lesion treatments with an intense pulsed light source. *Aesthetic Plast. Surg.* 26 (4), 303–307.
- Vohra, F., Al-Ritaiy, M.Q., Lillywhite, G., Abu Hassan, M.I., Javed, F., 2014. Efficacy of mechanical debridement with adjunct antimicrobial photodynamic therapy for the management of peri-implant diseases: a systematic review. *Photochem. Photobiol. Sci.* 13 (8), 1160–1168.
- Wachowska, M., Gabrysiak, M., Muchowicz, A., *et al.*, 2014. 5-Aza-2'-deoxycytidine potentiates antitumour immune response induced by photodynamic therapy. *Eur. J. Cancer.*
- Wahl, G., Bastanier, S., 1991. Soft laser in postoperative care in dentoalveolar treatment. *ZWR* 100, 512–515.
- Wainwright, M., 1998. Photodynamic antimicrobial chemotherapy (PACT). *J. Antimicrob. Chemother.* 42 (1), 13–28.
- Wajima, Z., Shitara, T., Inoue, T., Ogawa, R., 1996. Linear polarized light irradiation around the stellate ganglion area increases skin temperature and blood flow. *Masui* 45 (4), 433–438.
- Weishaupt, K.R., Gomer, C.J., Dougherty, T.J., 1976. Identification of singlet oxygen as the cytotoxic agent in photoinactivation of a murine tumor. *Cancer Res.* 36 (7 PT 1), 2326–2329.
- Weiss, R.A., McDaniel, D.H., Geronemus, R., *et al.*, 2004a. Non-ablative, non-thermal light emitting diode (LED) phototherapy of photoaged skin. *Laser Surg. Med.* 16, 31.
- Weiss, R.A., McDaniel, D.H., Geronemus, R.G., Weiss, M.A., 2005a. Clinical trial of a novel non-thermal LED array for reversal of photoaging: Clinical, histologic, and surface profilometric results. *Lasers Surg. Med.* 36 (2), 85–91.
- Weiss, R.A., McDaniel, D.H., Geronemus, R.G., *et al.*, 2005b. Clinical experience with light-emitting diode (LED) photomodulation. *Dermatol. Surg.* 31 (9 Pt 2), 1199–1205.
- Weiss, R.A., Weiss, M.A., Geronemus, R.G., McDaniel, D.H., 2004b. A novel non-thermal non-ablative full panel LED photomodulation device for reversal of photoaging: Digital microscopic and clinical results in various skin types. *J. Drugs Dermatol.* 3 (6), 605–610.
- West, C.M., West, D.C., Kumar, S., Moore, J.V., 1990. A comparison of the sensitivity to photodynamic treatment of endothelial and tumour cells in different proliferative states. *Int. J. Radiat. Biol.* 58 (1), 145–156.
- Wikramanayake, T.C., Rodriguez, R., Choudhary, S., *et al.*, 2012a. Effects of the Lexington LaserComb on hair regrowth in the C3H/HeJ mouse model of alopecia areata. *Lasers Med. Sci.* 27 (2), 431–436.
- Wikramanayake, T.C., Villasante, A.C., Mauro, L.M., *et al.*, 2012b. Low-level laser treatment accelerated hair regrowth in a rat model of chemotherapy-induced alopecia (CIA). *Lasers Med. Sci.* 28 (3), 701–706.
- Wilks, A.F., 1989. Two putative protein-tyrosine kinases identified by application of the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* 86 (5), 1603–1607.
- Willner, P., 1997. Validity, reliability and utility of the chronic mild stress model of depression: A 10-year review and evaluation. *Psychopharmacology (Berl)* 134 (4), 319–329.
- Wilson, M., 2004. Lethal photosensitisation of oral bacteria and its potential application in the photodynamic therapy of oral infections. *Photochem. Photobiol. Sci.* 3 (5), 412–418.
- Woodburn, K.W., Styli, S., Hill, J.S., *et al.*, 1992. Evaluation of tumour and tissue distribution of porphyrins for use in photodynamic therapy. *Br. J. Cancer* 65 (3), 321–328.
- Wu, Q., Xuan, W., Ando, T., *et al.*, 2012a. Low-level laser therapy for closed-head traumatic brain injury in mice: effect of different wavelengths. *Lasers Surg. Med.* 44 (3), 218–226.
- Wu, S., Zhou, F., Zhang, Z., Xing, D., 2011. Bax is essential for Drp1-mediated mitochondrial fission but not for mitochondrial outer membrane permeabilization caused by photodynamic therapy. *J. Cell Physiol.* 226 (2), 530–541.
- Wu, X., Alberico, S.L., Moges, H., *et al.*, 2012b. Pulsed light irradiation improves behavioral outcome in a rat model of chronic mild stress. *Lasers Surg. Med.* 44 (3), 227–232.
- Xuan, W., Agrawal, T., Huang, L., Gupta, G.K., Hamblin, M.R., 2015. Low-level laser therapy for traumatic brain injury in mice increases brain derived neurotrophic factor (BDNF) and synaptogenesis. *J. Biophotonics* 8 (6), 502–511.
- Xuan, W., Vatansever, F., Huang, L., Hamblin, M.R., 2014. Transcranial low-level laser therapy enhances learning, memory, and neuroprogenitor cells after traumatic brain injury in mice. *J. Biomed. Opt.* 19 (10), 108003.
- Xuan, W., Vatansever, F., Huang, L., *et al.*, 2013. Transcranial low-level laser therapy improves neurological performance in traumatic brain injury in mice: Effect of treatment repetition regimen. *PLOS ONE* 8 (1), e53454.
- Xue, L.Y., Chiu, S.M., Oleinick, N.L., 2001. Photodynamic therapy-induced death of MCF-7 human breast cancer cells: A role for caspase-3 in the late steps of apoptosis but not for the critical lethal event. *Exp. Cell Res.* 263 (1), 145–155.
- Yamazaki, M., Miura, Y., Tsuboi, R., Ogawa, H., 2003. Linear polarized infrared irradiation using Super Lizer is an effective treatment for multiple-type alopecia areata. *Int. J. Dermatol.* 42 (9), 738–740.
- Yano, K.B.L.F., Detmar, M., 2001. Control of hair growth and follicle size by VEGF-mediated angiogenesis. *J. Clin. Invest.* 107, 409–417.
- Yoneyama, H., Katsumata, R., 2006. Antibiotic resistance in bacteria and its future for novel antibiotic development. *Biosci. Biotechnol. Biochem.* 70 (5), 1060–1075.
- Youngblood, W.J., Gryko, D.T., Lammi, R.K., *et al.*, 2002. Glaser-mediated synthesis and photophysical characterization of diphenylbutadiene-linked porphyrin dyads. *J. Organic Chem.* 67 (7), 2111–2117.
- Yuan, F., Leunig, M., Berk, D.A., Jain, R.K., 1993. Microvascular permeability of albumin, vascular surface area, and vascular volume measured in human adenocarcinoma LS174T using dorsal chamber in SCID mice. *Microvasc. Res.* 45 (3), 269–289.
- Yusuf, N., Katiyar, S.K., Elmets, C.A., 2008. The immunosuppressive effects of phthalocyanine photodynamic therapy in mice are mediated by CD4+ and CD8+ T cells and can be adoptively transferred to naive recipients. *Photochem. Photobiol.* 84 (2), 366–370.
- Yu, W., Naim, J.O., Lanzafame, J., 1997. Effects of photostimulation on wound healing in diabetic mice. *Lasers Surg. Med.* 20, 56–63.
- Yu, W., Naim, J.O., Lanzafame, R.J., 1994. The effect of laser irradiation on the release of bFGF from 3T3 fibroblasts. *Photochem. Photobiol.* 59 (2), 167–170.
- Zhang, L., Xing, D., Zhu, D., Chen, Q., 2008. Low-power laser irradiation inhibiting Abeta25–35-induced PC12 cell apoptosis via PKC activation. *Cell Physiol. Biochem.* 22, 215–222.

- Zhang, Q., Zhou, C., Hamblin, M.R., Wu, M.X., 2014. Low-level laser therapy effectively prevents secondary brain injury induced by immediate early responsive gene X-1 deficiency. *J. Cereb. Blood Flow Metabolism: Off. J. Int. Soc. Cereb. Blood Flow Metabolism* 34 (8), 1391–1401.
- Zhao, Y., Vanhoutte, P.M., Leung, S.W., 2015. Vascular nitric oxide: Beyond eNOS. *J. Pharmacol. Sci.* 129 (2), 83–94.
- Zivin, J.A., Albers, G.W., Bornstein, N., *et al.*, 2009. Effectiveness and safety of transcranial laser therapy for acute ischemic stroke. *Stroke* 40 (4), 1359–1364.
- Zivin, J.A., Sehra, R., Shoshoo, A., *et al.*, 2014. Neuro thera(R) efficacy and safety trial-3 (NEST-3): a double-blind, randomized, sham-controlled, parallel group, multicenter, pivotal study to assess the safety and efficacy of transcranial laser therapy with the neuro thera(R) laser system for the treatment of acute ischemic stroke within 24h of stroke onset. *Int. J. Stroke* 9 (7), 950–955.

Relevant Website

<https://www.sciencebasedmedicine.org/science-by-press-release-a-helmet-to-fight-alzheimers-disease/>
Science-Based Medicine.

Resolution and Multiple Scattering in Imaging

Edwin A Marengo, Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

Zambrano-Nunez Maytee, Technological University of Panama, Panama City, Panama

Edson S Galagarza, Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Light carries information about the sources from which it comes and the objects with which it interacts. Imaging can be defined as the physical or computational processing of this information in order to create a real or virtual replica (an “image”) of the source or object or its associated optical wavefield. This function can be performed by a physical imaging system, such as a classical lens-based system, as shown in Fig. 1. Alternatively, it can be implemented computationally, by a two-layer system composed of a physical interface, which senses the field produced by the object and a computational block that implements inversion algorithms to estimate the radiating object’s properties, which can, in turn, render image representations of that object and its surrounding field (see Fig. 2). The most critical property of an imaging system, be it physical or computational, is the degree of accuracy with which objects under surveillance can be successfully imaged or reproduced by that system. This characterization of the imaging system, usually termed “imaging resolution,” is affected by both fundamental physical limits (the inherent finite resolving power of the system, also termed the “diffraction limit”) as well as environmental effects (propagation model deviations) and sensor imperfections (aberrations). This contribution outlines the key principles to define and compute the resolution of typical imaging systems including those in which the phenomenon of multiple scattering can play a major role. Particular attention is given to the

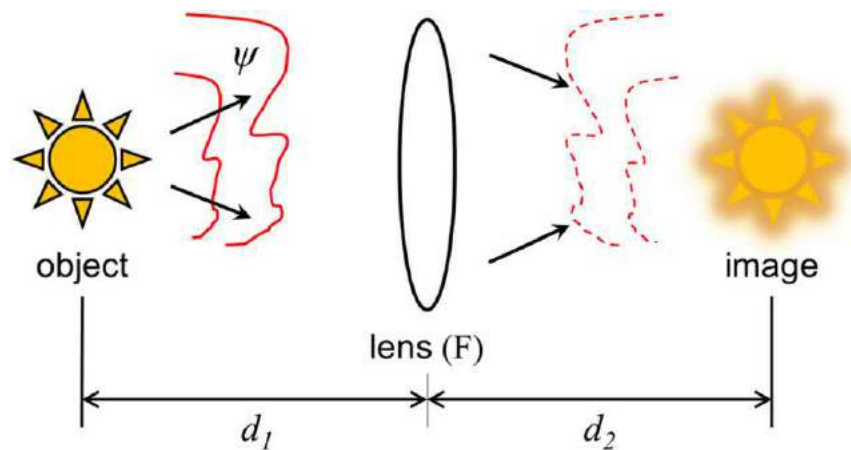


Fig. 1 Physical imaging. In this lens-based system, imaging occurs if $1/d_1 + 1/d_2 = 1$ where F is the lens focal length.

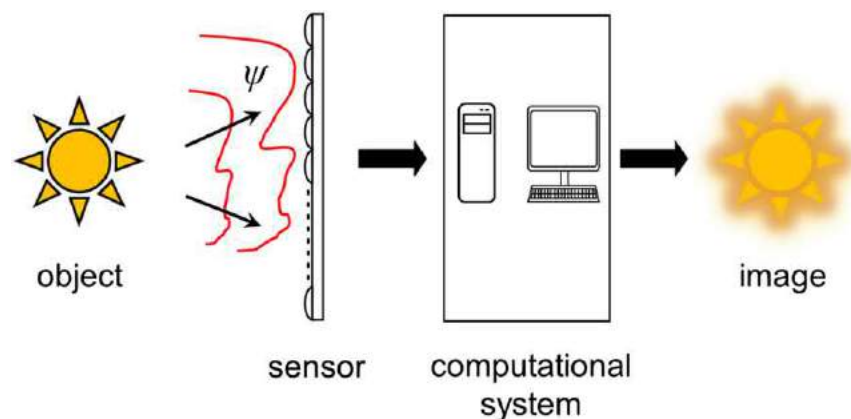


Fig. 2 Computational Imaging.

intrinsic, fundamental resolution limit of the systems considered, which can be thought of as a fundamental bound to the resolution. The latter is indeed attainable in state-of-the-art systems (termed “diffraction-limited”).

Inverse Source and Scattering Problems

Two different forms of imaging are of interest in this contribution. The most basic one, to be termed here “source-field imaging,” involves the imaging of sources and their fields, and can be addressed within the theoretical framework of the inverse source problem (reconstruction of an unknown source from knowledge of its field in a region located outside the source support, such as in the far zone). This includes as a special case the inverse diffraction problem of reconstructing the field in a boundary or region from knowledge of the field in another region. The sources in question can be self-emitting (primary) or induced (secondary), as in a scatterer that is illuminated by light from another source (see Fig. 3). The second kind of imaging problem under consideration, to be termed “scatterer imaging,” involves the determination of a scatterer’s constitutive properties (e.g., spatial distribution of its index of refraction) from scattering data corresponding to a number of transmit-and-receive experiments.

In source-field imaging the associated mapping from the sought-after source or field (the object) to the field data is linear. In the classical theoretical formulations of this problem one can talk about a well-defined resolution limit arising from bandlimitation considerations in Fourier space. In contrast, in the scatterer imaging or inverse scattering problem, the associated forward map from scatterer properties to the data is, in general, nonlinear, so that alternative methods, such as the statistical Cramer–Rao bound (CRB) (Kay, 1993, chapter 3) need to be adopted so as to quantify the resolution (Sentenac *et al.*, 2007). Scattering has, however, a weak scatterer regime, called “the Born approximation,” in which the map in question is linear. This specialized regime is of interest in many practical applications and within it there is an associated classical resolution limit which has characteristics similar to those of the related source-field imaging problem.

Background and Literature

Resolution is traditionally quantified as the minimum distance that two point sources or scatterers can have while being distinguishable in the corresponding image. The image of a point source or scatterer is usually called the point spread function (PSF) of the imaging system. According to the well-known Rayleigh criterion (Goodman, 2005, p. 158), the resolution corresponds to the separation for which the peak of the image of one source occurs at the first null or zero of the image of the other. This is, however, only an approximate measure of resolution which applies, in fact, to a quite specific type of object. As pointed out by Toraldo di Francia in his classical studies on resolving power and information (Toraldo di Francia, 1952, 1955), imaging resolution is more appropriately characterized by the entire PSF and its associated information content, for example, in the spectral domain, from which one can in turn determine the effective dimensionality or number of degrees of freedom (NDF) of the entire imaging system. In many of the imaging problems that arise in optics, this detailed resolution characterization can be done quite conveniently in spatial Fourier domain, as we explain next for both inverse source and inverse scattering scenarios.

Among the major contributions to our understanding of imaging resolution and the associated field data “information content” and degrees of freedom of classical imaging systems, it is worthwhile mentioning the pioneering work of Abbe who established the resolution limit of the microscope around 1873 and Lord Rayleigh who further expanded the relevant theory in 1879 (Lord Rayleigh, 1879). Other contributions include those of Sparrow (1916), Schuster (1924), Houston (1927), and Buxton (1937), who proposed alternatives to the Rayleigh resolution criterion. Important advances at the crossroads of antenna theory, information theory, and optics were done in the 1940s and 1950s by several authors including Schelkunoff, Bouwkamp and Bruijn (1946), Woodward and Lawson (1948), and Toraldo di Francia (1955, 1952), which addressed the theoretical and practical

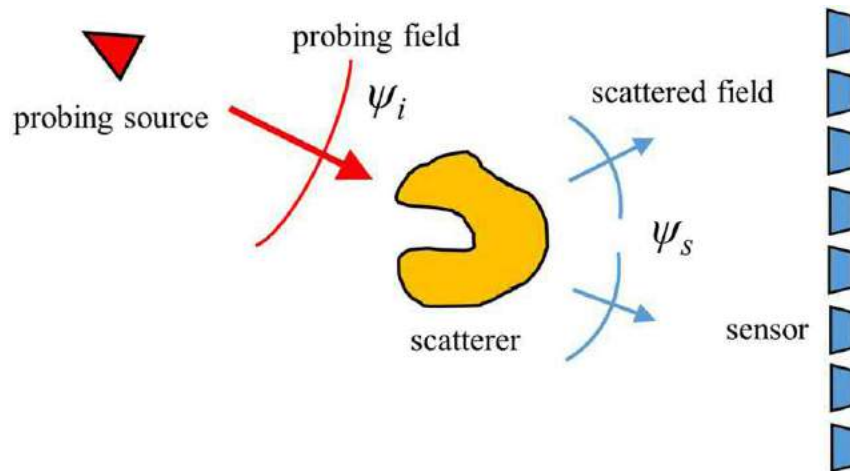


Fig. 3 Scattering experiment. The scatterer generates scattered field $\psi_s(\mathbf{r})$ in response to excitation or probing with incident field $\psi_i(\mathbf{r})$.

limits associated to the launching of and imaging with far fields. Among other results, these and other subsequent investigations (Ronchi, 1961; Rushforth and Harris, 1968) conclusively rendered the important insight that the resolution limits of far-field data are, for the most part, only practical (as opposed to theoretical), insofar as they are the result of noise and perturbations in realistic data and signal models. For example, if data were perfect and thus lacked noise, then it would be possible for many imaging systems (see Sections Propagating and Evanescent Spectra of the Radiated Field and Transverse Resolution of this article), to analytically extend the data corresponding to a region in Fourier space (called the visible region) into another one (called invisible) so as to increase the signal bandwidth and improve the resolution (Goodman, 2005; Section 6.6). This is one of the computational approaches to “superresolution” but it is useful only for perfect field data. In realistic systems, data are not perfect, and thus this approach is of limited practical applicability. But this illustrates that the conventional “fundamental” resolution limits are mostly practical, due to the almost inevitable presence of noise and perturbations, such as systematic errors. For instance, it is not impossible to physically process the far field, via a suitable device or instrument, so as to achieve superresolution. But (as in the computational case) such an analog or “physical processing” approach is viable only insofar as the physical field arriving at the instrument carries itself little noise or interference from spurious sources. Several promising approaches along these lines are mentioned in Section Superresolution, which are based on metamaterials. There we also mention other, more robust and already proven approaches to superresolution. Recent overviews on resolution and the related topic of superresolution can be found in den Dekker and van den Bos (1997), Heintzmann and Ficz (2006), Zheludev (2008), Testorf and Fiddy (2010).

Resolution in Source-Field Imaging

Scalar Wave Model

In the scalar approximation, the light emitted in free space by a monochromatic source oscillating at temporal frequency f (in Hertz) can be represented by a scalar wavefield whose spatial dependence is defined by a wavefunction $\psi(\mathbf{r})$ which obeys the inhomogeneous Helmholtz equation

$$(\nabla^2 + k_0^2)\psi(\mathbf{r}) = \rho(\mathbf{r}) \quad (1)$$

where $\mathbf{r} \in \mathbb{R}^3$, ρ represents the spatial dependence of the source (at said temporal frequency), ∇^2 is the Laplacian operator, and the wavenumber

$$k_0 = \omega/c = 2\pi/\lambda \quad (2)$$

where $\omega = 2\pi f$ is the angular oscillation frequency (in radians per second), c is the speed of light in free space, and λ is the wavelength. In this model of wave propagation, the wavefield $\psi(\mathbf{r})$ radiated by the source $\rho(\mathbf{r})$ is given by

$$\psi(\mathbf{r}) = \int_{\tau} d^3r' \rho(\mathbf{r}') G_0(\mathbf{r} - \mathbf{r}') \quad (3)$$

where d^3r' denotes differential volume, τ is the spatial support of the source, and G_0 is the free space Green function defined by

$$G_0(\mathbf{r} - \mathbf{r}') = -\frac{e^{ik_0|\mathbf{R}|}}{4\pi|\mathbf{R}|} \quad (4)$$

where $i = \sqrt{-1}$ and

$$\mathbf{R} = \mathbf{r} - \mathbf{r}' \quad (5)$$

Green's function G_0 admits a number of alternative representations, such as the Fourier-integral representation,

$$G_0(\mathbf{R}) = -\frac{1}{8\pi^3} \int_{-\infty}^{\infty} d^3K \frac{\exp(i\mathbf{K} \cdot \mathbf{R})}{k_0^2 - K^2 + i\varepsilon} \quad (6)$$

where $\mathbf{K} \in \mathbb{R}^3$, d^3K is differential volume in Fourier space, $K = |\mathbf{K}|$, and ε is an infinitesimal value added to remove the singularity at $K = k_0$.

Three-Dimensional Resolution of Far Fields

In many applications the fields are sensed in the far zone of the source region (τ), where $|\mathbf{r} - \mathbf{r}'| \simeq r - \hat{\mathbf{r}} \cdot \mathbf{r}'$, where $r = |\mathbf{r}|$ and $\hat{\mathbf{r}} = \mathbf{r}/r$, which when substituted in Eq. (4) gives the far zone behavior of G_0 ,

$$G_0(r\hat{\mathbf{r}} - \mathbf{r}') = -\frac{e^{ik_0r}}{4\pi r} e^{-ik_0\hat{\mathbf{r}} \cdot \mathbf{r}'} \quad (7)$$

Substituting this in Eq. (3) gives the far zone behavior of the field ψ ,

$$\psi(r\hat{\mathbf{r}}) \sim f(\hat{\mathbf{r}}) \frac{e^{ik_0r}}{r} \quad (8)$$

where the quantity $f(\hat{\mathbf{r}})$ is the far-field radiation pattern defined by

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \int_{\tau} d^3 r \rho(\mathbf{r}) e^{-ik_0 \hat{\mathbf{r}} \cdot \mathbf{r}'} \quad (9)$$

The latter is, apart from a multiplicative factor, equal to the spatial Fourier transform of the source

$$\tilde{\rho}(\mathbf{K}) = \int_{\tau} d^3 r \rho(\mathbf{r}) e^{-i\mathbf{K} \cdot \mathbf{r}'} \quad (10)$$

evaluated at $\mathbf{K} = k_0 \hat{\mathbf{r}}$, i.e.,

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \tilde{\rho}(\mathbf{K})|_{\mathbf{K} = k_0 \hat{\mathbf{r}}} \quad (11)$$

This means that the information about the source that is contained in the far field is the spatial Fourier transform of the source evaluated over the surface of a sphere, called the Ewald sphere (shown in Fig. 4), that has radius $K = k_0$ and is centered at the origin in Fourier space. The $K = k_0$ and $K \neq k_0$ components are sometimes referred to as the “on-the-shell” and “off-the-shell” components of the Fourier spectrum.

The fact that the far field carries only on-the-shell information about the spatial Fourier transform of ρ means that there is a loss of information about the source in the wave propagation process (from source to far field). This, in turn, implies that there is a fundamental limit to our capacity to reconstruct an unknown source from knowledge of the far field alone. It is, in fact, impossible to uniquely deduce the unknown source from the far field alone, since there are infinite nontrivial sources (called nonradiating sources) for which the spatial Fourier transform vanishes identically over the Ewald sphere (Devaney and Wolf, 1973; Bleistein and Cohen, 1977; Marengo and Ziolkowski, 2000; Gbur, 2003). One can, however, evaluate the unique, wavelike source which best fits the Fourier data over that sphere and has zero off-the-shell Fourier components. This is the classical approach to source imaging, and is the one that will lead us next to the fundamental resolution limit. Mathematically, this corresponds to the solution of minimum functional energy or 2 norm, or “minimum energy solution” (Devaney, 2012, chapter 5) to the inverse problem, without prior information about the source support. Minimum energy solutions consistent with a known source support can also be considered in certain applications (Marengo et al., 1999, 2000). In this case, the prior support information reliably enables superresolution (Schmidt-Weinmar, 1978).

The inverse problem of reconstructing an unknown source ρ , whose support τ is unknown, and that generates a given far-field radiation pattern $f(\hat{\mathbf{r}})$, with full-view data corresponding to 4π steradians, can thus be cast as computing the inverse Fourier transform (IFT) of the “on-the-shell” Fourier transform of the source. This is equivalent to Fourier inverting the quantity $\tilde{\rho}(\mathbf{K})H(\mathbf{K})$, where $H(\mathbf{K}) = \delta(K - k_0)/K^2$, which implies in view of the convolution theorem that the reconstructed source $\hat{\rho}$ is related to the source ρ by

$$\hat{\rho}(\mathbf{r}) = \int_{\tau} d^3 r' \rho(\mathbf{r}') h(\mathbf{r} - \mathbf{r}') \quad (12)$$

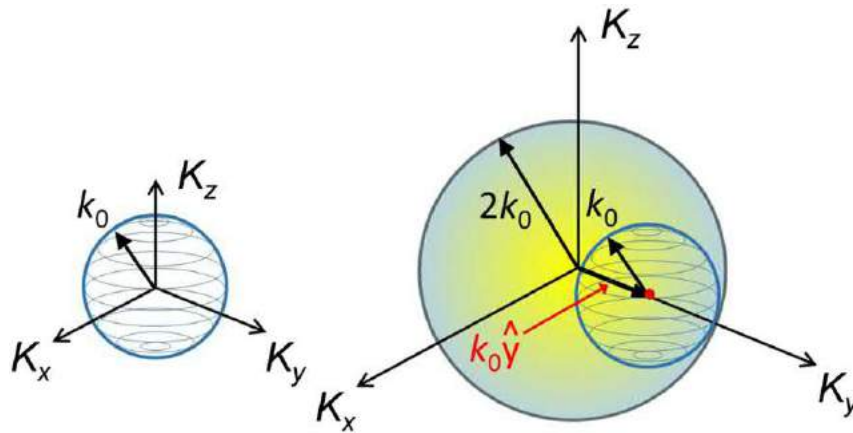


Fig. 4 Ewald sphere (left), and limiting Ewald sphere and ball (right) in 3D Fourier space. The Ewald sphere ($|\mathbf{K}| = k_0$) defines the Fourier domain region populated by far-field data of inverse source problems. The limiting Ewald ball ($|\mathbf{K}| \leq 2k_0$) defines the corresponding data region for full-view far-field inverse scattering within the weak scattering or first Born approximation. The figure also shows a shifted Ewald sphere ($|\mathbf{K} - k_0 \hat{\mathbf{y}}| = k_0$). It is contained inside the limiting Ewald ball, and corresponds to far-zone scattered field data under probing with a plane wave traveling in the direction of the unit vector $-\hat{\mathbf{y}}$. By probing the scatterer with plane waves traveling in all directions, the resulting scattered field data populates the entire limiting Ewald ball.

where the PSF h of this imaging system, equal to the IFT of H , can be shown to be given by

$$h(\mathbf{r} - \mathbf{r}') = \frac{1}{2\pi^2} j_0(k_0 |\mathbf{r} - \mathbf{r}'|) = -\frac{2}{\pi k_0} \text{Im} G_0(\mathbf{r} - \mathbf{r}') \quad (13)$$

where j_0 is the spherical Bessel function of the first kind and order 0 and Im denotes the imaginary part. A plot of this PSF is shown in Fig. 5. The first zero of $j_0(k_0 \Delta_{3D}^R)$ corresponding to $k_0 \Delta_{3D}^R = \pi$, gives a separation

$$\Delta_{3D}^R = \lambda/2 \quad (14)$$

which can be regarded, in view of the Rayleigh criterion, as the sought-after resolution that is attainable with this 3D, source-inversion-based computational imaging system. In particular, without the involvement of prior information about the source (such as support, parametric form, point-like nature, sparsity, or other), then two point sources are well resolved with far-field data only if they are separated by a distance equal to or greater than about half a wavelength.

Propagating and Evanescent Spectra of the Radiated Field

In many applications, field sensing is done over a plane, say $z = z_1$, which is itself contained in a half-space $z > z_0$ that is exterior to the source support τ , as shown in Fig. 6. In these applications, the field can be conveniently analyzed using a quite insightful classical representation called the angular-spectrum expansion, whose main features are outlined next.

Consider the Cartesian representation of $\mathbf{r}$ and $\mathbf{r}'$, in particular, $\mathbf{r} = (x, y, z)$ and $\mathbf{r}' = (x', y', z')$. Similarly, let $\mathbf{K} = (K_x, K_y, K_z)$. Let $\mathbf{R}_t = (x - x', y - y')$ and $\mathbf{K}_t = (K_x, K_y)$ denote the transverse components of $\mathbf{R}$ and $\mathbf{K}$, respectively, and introduce $Z = z - z'$. Then

$$\begin{aligned} \mathbf{R} &= \mathbf{R}_t + Z\hat{z} \\ \mathbf{K} &= \mathbf{K}_t + K_z\hat{z} \end{aligned} \quad (15)$$

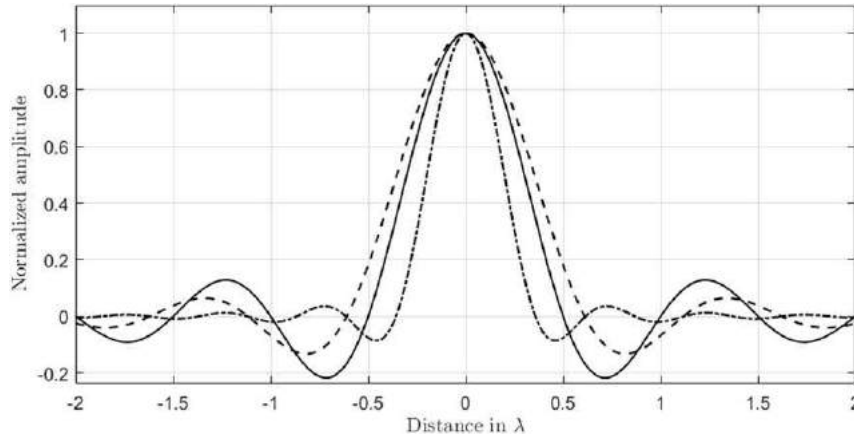


Fig. 5 Plots of the normalized point spread function (PSF) $h(d)/h(0)$ vs. d (in λ) relevant to three-dimensional (3D) source-field imaging (solid line), two-dimensional (2D) transverse field imaging (dashed line), and 3D Born-approximable inverse scattering (dashed-dotted line).

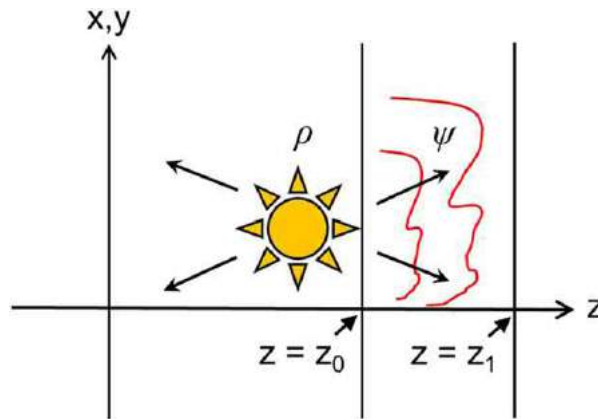


Fig. 6 Geometry used in the discussion of the plane wave expansion of the radiated field.

Integrating over K_z in Eq. (6) one arrives, via classical contour integration in the complex plane (see Devaney, 2012, Section 4.1), to the Weyl expansion of Green's function G_0 , in particular,

$$G_0(\mathbf{R}) = -\frac{i}{8\pi^2} \int_{-\infty}^{\infty} d^2 K_t \frac{e^{i(\mathbf{K}_t \cdot \mathbf{R}_t \pm \gamma Z)}}{\gamma(\mathbf{K}_t)} \quad (16)$$

where “+” is used if $Z > 0$ and “−” if $Z < 0$ and

$$\gamma(\mathbf{K}_t) = \begin{cases} \sqrt{k_0^2 - K_t^2} & \text{if } K_t \leq k_0 \\ i\sqrt{K_t^2 - k_0^2} & \text{if } K_t > k_0 \end{cases} \quad (17)$$

where $K_t = |\mathbf{K}_t|$.

In view of Eq. (17), the integral in Eq. (16) can be conveniently decomposed into two parts, one corresponding to integration over the ball in two-dimensional $\mathbf{K}_t$ space defined by $K_t \leq k_0$, and another corresponding to integration over the complementary region (the exterior of this ball). These two regions are illustrated in Fig. 7. The waves associated to the $K_t \leq k_0$ region, sometimes called “the visible region,” are propagating, and go to the far zone, while the waves corresponding to the $K_t > k_0$ region, sometimes referred to as “the invisible region,” decay exponentially with $|Z|$ and are called evanescent. The latter carry the high-frequency features of G_0 but cannot be sensed effectively for $|Z|$ greater than a few wavelengths. Moreover, since this field representation can be applied to any $\hat{\mathbf{z}} \in S^2$ over the unit sphere S^2 , this further implies that these high-frequency, evanescent features of G_0 cannot be sensed a few wavelengths away from the source point $\mathbf{r}'$. For $|Z|$ greater than a few wavelengths, $G_0(\mathbf{R})$ can thus be approximated as

$$G_0(\mathbf{R}) \simeq -\frac{i}{8\pi^2} \int_{K_t \leq k_0} d^2 K_t \frac{e^{i(\mathbf{K}_t \cdot \mathbf{R}_t \pm \gamma Z)}}{\gamma(\mathbf{K}_t)} \quad (18)$$

This is a free field, obeying the homogeneous Helmholtz equation in all space.

Physically, the Weyl expansion (Eq. (16)) thus characterizes Green's function G_0 as superpositions (an spectrum) of two kinds of plane waves, namely, propagating and evanescent waves. The expansion coefficient $1/\gamma(\mathbf{K}_t)$ appearing in Eq. (16) can be thought of as the plane wave spectrum of G_0 . Its value in the visible and invisible regions defines the respective propagating and evanescent spectra of G_0 . Interestingly, it can be shown that (Devaney, 2012, p. 122)

$$\text{Im} G_0(\mathbf{R}) = -\frac{i}{8\pi^2} \int_{K_t \leq k_0} \frac{d^2 K_t}{\gamma(\mathbf{K}_t)} e^{i\mathbf{K}_t \cdot \mathbf{R}_t} \cos[\gamma(\mathbf{K}_t)Z] \quad (19)$$

hence the imaginary part of G_0 is determined only by the propagating spectrum. This is consistent with the fact, demonstrated earlier, that the imaginary part of G_0 governs, under conventional far-field sensing, the 3D imaging resolution, giving rise to the fundamental Rayleigh-criterion-based $\lambda/2$ resolution limit.

Now, using Eqs. (3) and (16) one finds that the field $\psi(\mathbf{r})$ can be represented for $z > z_0$ as

$$\psi(\mathbf{r}) = -\frac{i}{8\pi^2} \int_{-\infty}^{\infty} \frac{d^2 K_t}{\gamma(\mathbf{K}_t)} \hat{\rho}[\mathbf{K}_t + \gamma(\mathbf{K}_t)\hat{\mathbf{z}}] e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(\mathbf{K}_t)Z]} \quad (20)$$

where $\mathbf{r}_t = (x, y)$. Thus the corresponding plane wave spectrum of the field is governed by the spatial Fourier transform $\hat{\rho}$ of the source ρ , and its analytic continuation for complex argument, as required in connection with the evanescent portion of the field spectrum. For z near the boundary $z = z_0$ this field has significant evanescent components, which carry high-frequency information about the source and can thus lead to high imaging resolution in the associated inversion. On the other hand, $z - z_0$ greater than a

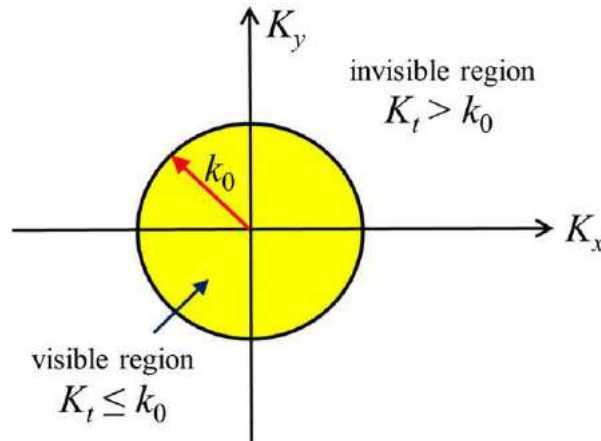


Fig. 7 Regions in Fourier domain corresponding to the visible (propagating) and invisible (evanescent) field components. The visible region is populated by far-field data.

few wavelengths the field is mostly propagating and can be approximated as

$$\psi(\mathbf{r}) = -\frac{i}{8\pi^2} \int_{K_t \leq k_0} \frac{d^2 K_t}{\gamma(\mathbf{K}_t)} \tilde{\rho}[\mathbf{K}_t + \gamma(\mathbf{K}_t)\hat{\mathbf{z}}] e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(\mathbf{K}_t)z]} \quad (21)$$

so that it carries only the low-frequency information about the source, which is in fact identical to the far-field one that lies in the Ewald sphere in the Fourier domain, as was explained earlier.

Transverse Resolution

In many optical imaging systems, such as in conventional microscopy, the object to be imaged lies over a planar surface. In this context it is convenient to characterize the resolution limits associated to the imaging of the field over a planar surface. Furthermore, this characterization is also relevant to the imaging of more arbitrary objects. In particular, it sheds insight on the resolution with which the near field in the vicinity of the object can be imaged, which is in turn linked to the resolution with which the object itself can be resolved. Thus, referring to the geometry in Fig. 6, we consider next the classical resolution associated to the imaging of the field over the boundary $z=z_0$ that bounds the support τ of radiating source ρ . It is assumed that the field data are gathered over the boundary $z=z_1$ where $z_1 - z_0$ is greater than several wavelengths so that the evanescent components cannot be reliably sensed there.

This can be addressed within the plane wave spectrum representation of the field by considering an equivalent Huygens' source linked to the field, for example, the double-layer singular surface source

$$\rho_s(\mathbf{r}) = -2\psi(\mathbf{r}_t, z_0)\delta'(z - z_0) \quad (22)$$

where $\delta'(z) = \frac{d}{dz}\delta(z)$, which leads to the first Rayleigh–Sommerfeld formula (Devaney, 2012, p. 75; Goodman, 2005, p. 48)

$$\psi(\mathbf{r}) = -2 \int_{-\infty}^{\infty} d^2 \mathbf{r}'_t \psi(\mathbf{r}'_t, z_0) \frac{\partial}{\partial z'} G_0(\mathbf{r} - \mathbf{r}'), \quad z > z_0 \quad (23)$$

Using this equivalent source in Eq. (21) one obtains

$$\psi(\mathbf{r}_t, z_1) = \frac{i}{4\pi^2} \int_{K_t \leq k_0} d^2 K_t \tilde{\psi}(\mathbf{K}_t, z_0) e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(\mathbf{K}_t)z_1]} \quad (24)$$

Therefore, the information about the near field at $z=z_0$ that is contained in the measured field at $z=z_1$ is its 2D spatial Fourier transform $\tilde{\psi}(\mathbf{K}_t, z_0)$ evaluated over the ball of radius k_0 centered at the origin in 2D Fourier space (the visible region shown in Fig. 7). This means that the largest spatial frequency magnitude contained in the data is $k_0 = 2\pi/\lambda$, so that the minimum spatial period that can be resolved is λ . This is a fundamental measure of the achievable resolution, in particular,

$$\Delta_{2D} = \lambda \quad (25)$$

based on standard bandlimitation considerations. In addition, one can also consider the Rayleigh resolution criterion. For this purpose, we note that the transfer function of the propagation map is $H(\mathbf{K}_t) = U(k_0 - K_t)$ where U is the unit-step function. The respective IFT corresponds to the PSF h of the imaging system which can be shown to be equal to

$$h(\mathbf{r}_t) = 2\pi k_0 J_1(k_0 \rho) / \rho \quad (26)$$

where $\rho = |\mathbf{r}_t|$. A plot of this PSF is shown in Fig. 5. The first zero of this PSF occurs at $k_0 \rho \approx 1.22\pi$, corresponding to a (Rayleigh-criterion-based) 2D or transverse field resolution of

$$\Delta_{2D}^R = 0.61\lambda \quad (27)$$

This corresponds to the resolution limit if the field is sensed over the entire plane $z=z_1$. In addition to the intrinsic loss of information (the evanescent wave part) due to wave propagation, there is also an additional loss due to the finite size of the sensing apparatus. In this realistic case if the aperture covers a polar angle 2θ then the detected transverse frequencies are such that $K_t \leq k_0 \sin \theta$ and therefore the respective PSF is

$$h(\mathbf{r}_t) = 2\pi k_0 \sin \theta J_1(k_0 \sin \theta \rho) / \rho \quad (28)$$

which gives the Rayleigh criterion transverse resolution

$$\Delta_{2D}^R = \frac{0.61\lambda}{\text{NA}} \quad (29)$$

where $\text{NA} = \sin \theta$ is the numerical aperture of the apparatus.

Resolution in Scatterer Imaging

In many application areas (such as radar, lidar, biomedical imaging, geophysical imaging, nondestructive testing, to mention a few) a medium is probed using known incident waves, and the medium's response, in the form of the resulting perturbation or scattered waves, is measured with sensors located in the far zone (see Fig. 3). By processing the measured scattered field data, it is possible to obtain images of the scattering medium or target, which give information about its localization, shape, and material

properties (e.g., index of refraction). If the surrounding background medium is free space, then the scattered field, to be denoted as $\psi_s(\mathbf{r})$, obeys, within the scalar formulation, a differential equation of the form

$$(\nabla^2 + k_0^2)\psi_s(\mathbf{r}) = V(\mathbf{r})\psi(\mathbf{r}) \quad (30)$$

where k_0 is the wavenumber in free space, $V(\mathbf{r})$ is the space-dependent scattering potential, which is related to the scatterer's space-dependent index of refraction $n(\mathbf{r})$ via

$$V(\mathbf{r}) = k_0^2[1 - n^2(\mathbf{r})] \quad (31)$$

and $\psi(\mathbf{r})$ is the total field, which is given by

$$\psi(\mathbf{r}) = \psi_i(\mathbf{r}) + \psi_s(\mathbf{r}) \quad (32)$$

where $\psi_i(\mathbf{r})$ is the probing or incident field and the scattered field $\psi_s(\mathbf{r})$ obeys the radiation condition at infinity, as required by causality considerations.

It follows from Eqs. (1), (3), and (30) (via the substitution $\rho \rightarrow V\psi$) that

$$\psi_s(\mathbf{r}) = \int_{\tau} d^3r' G_0(\mathbf{r} - \mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}') \quad (33)$$

where τ denotes the scatterer's support. Furthermore, using Eq. (7) in Eq. (33) we obtain (in analogy to Eqs. (8) and (9)) the far zone behavior of the scattered field:

$$\psi(r\mathbf{s}) \sim f_s(\mathbf{s}; \psi_i) \frac{e^{ik_0 r}}{r} \quad (34)$$

where the quantity $f_s(\mathbf{s}; \psi_i)$ is the far-zone scattering amplitude, which depends on the observation direction $\mathbf{s} \in S^2$ and the incident field ψ_i , and is given by

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \quad (35)$$

For the typical situation in which the incident field is a uniform plane wave traveling in the direction of the unit vector $\mathbf{s}_0$, i.e.,

$$\psi_i(\mathbf{r}) = e^{ik_0 \mathbf{s}_0 \cdot \mathbf{r}} \quad (36)$$

then it is convenient to denote $f_s(\mathbf{s}; \psi_i)$ as $f_s(\mathbf{s}; \mathbf{s}_0)$. For this case we obtain from Eqs. (32), (35), and (36)

$$f_s(\mathbf{s}; \mathbf{s}_0) = f_s^B(\mathbf{s}; \mathbf{s}_0) + f_s^{MS}(\mathbf{s}; \mathbf{s}_0) \quad (37)$$

where f_s^B is the scattering amplitude under the weak scattering or Born approximation, which is given by

$$f_s^B(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) e^{-ik_0(\mathbf{s} - \mathbf{s}_0) \cdot \mathbf{r}} = -\frac{1}{4\pi} \tilde{V}(\mathbf{K})|_{\mathbf{K} = k_0(\mathbf{s} - \mathbf{s}_0)} \quad (38)$$

where $\tilde{V}$ denotes the spatial Fourier transform of the scattering potential function V , and f_s^{MS} is the multiple scattering contribution to the total scattering amplitude, which is given by

$$f_s^{MS}(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi_s(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \quad (39)$$

If multiple scattering is weak, so that $|\psi_s| \ll |\psi_i|$ and consequently $|f_s^{MS}| \ll |f_s^B|$ then from Eq. (37) approximately

$$f_s(\mathbf{s}; \mathbf{s}_0) \simeq f_s^B(\mathbf{s}; \mathbf{s}_0) \quad (40)$$

and according to Eq. (38) the scattering amplitude is given by the spatial Fourier transform $\tilde{V}$ of the scattering potential V evaluated at $\mathbf{K} = k_0(\mathbf{s} - \mathbf{s}_0)$. If the scattering amplitude is measured for a fixed incidence direction $\mathbf{s}_0$ and for all the observation directions $\mathbf{s}$, the resulting scattering data populates the shifted Ewald sphere where $|\mathbf{K} - k_0 \mathbf{s}_0| = k_0$ (see Fig. 4). If probing is done using all the incidence directions $\mathbf{s}_0$ and sensing is done in all observation directions $\mathbf{s}$ then the resulting scattering data populates in Fourier space a ball of radius $2k_0$ that is centered at the origin. It is bounded by the so-called limiting Ewald sphere, and is illustrated and referred to as the limiting Ewald ball in Fig. 4. We focus next on this situation of full-view probing and sensing. In this weak scattering regime, the map from the scattering potential V to the scattering amplitude is linear, so that one can compute a well-defined PSF for the relevant computational imaging. The PSF in question is given by (Simonetti, 2006, Eq. (17))

$$h(\mathbf{r}) = \frac{8j_1(2k_0 r)}{\lambda^2 r} \quad (41)$$

A plot of this PSF is shown in Fig. 5. Based on the first zero of the spherical Bessel function j_1 (whose approximate value is 4.493), an approximate value of the Born approximation scattering resolution based on Rayleigh's criterion is

$$\Delta_{3D}^R \simeq 0.358\lambda \quad (42)$$

But this is only a rough resolution estimate. Moreover, it is pertinent only for the special case of two point-like scatterers. It is more general to simply state that the largest spatial frequency magnitude contained in the data is $2k_0 = 4\pi/\lambda$ which implies, in turn, that the minimum spatial period that can be resolved is $\lambda/2$, and this is the familiar resolution limit discussed in Wolf's classical treatment of the inverse scattering problem in the framework of the Born approximation (Wolf, 1969, Eq. (22)). In particular,

based on this minimum spatial period criterion which is relevant to arbitrary scatterers,

$$\Delta_{3D,s} = \lambda/2 \quad (43)$$

On the other hand, if multiple scattering is not negligible, then the scattering amplitude is given by Eq. (37) where the term f_s^{MS} is given by Eq. (39) and cannot be neglected. Substituting from Eq. (33) in Eq. (39) we find that

$$\begin{aligned} f_s^{\text{MS}}(\mathbf{s}; \mathbf{s}_0) &= -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi_s(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \\ &= -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \int_{\tau} d^3r' G_0(\mathbf{r} - \mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}') \end{aligned} \quad (44)$$

which clearly showcases the nonlinear nature of the map from V to f_s^{MS} (and, in turn, to f_s). This prohibits the use of bandlimitation considerations in Fourier domain to quantify the corresponding imaging resolution. Alternative criteria that apply also to nonlinear systems, such as a criterion based on the fundamental CRB have been proposed (Sentenac *et al.*, 2007), and applied to quantify the effects of multiple scattering in imaging (Sentenac *et al.*, 2007; Simonetti *et al.*, 2008; Marengo and Berestesky, 2013; Marengo *et al.*, 2012).

Simonetti (2006) has shown that multiple scattering can enhance imaging resolution, relative to the Born-approximation-based resolution limits, thanks to the conversion of evanescent waves in the scatterer into propagating waves that are accessible in the far zone. This mechanism implies the encoding of high-frequency information about the scatterer, beyond that associated (in the weak scattering approximation) to the limiting Ewald ball. Thus superresolution is possible thanks to multiple scattering. The key argument can be outlined with the help of Eqs. (6) and (44). Using Eq. (6) in Eq. (44) one obtains

$$f_s^{\text{MS}}(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int d^3K \frac{\tilde{V}(k_0 \mathbf{s}_0 - \mathbf{K}) T(\mathbf{K}, k_0 \mathbf{s}_0)}{k_0^2 - K^2 + i\epsilon} \quad (45)$$

where

$$T(\mathbf{K}, k_0 \mathbf{s}_0) = \int_{\tau} d^3r \rho(\mathbf{r}) e^{-i\mathbf{K} \cdot \mathbf{r}} V(\mathbf{r}) \psi(\mathbf{r}) \quad (46)$$

The integral in Eq. (45) is over the entire $\mathbf{K}$ space, so that it depends on the value of the spatial Fourier transform $\tilde{V}$ of V for all spatial frequencies, including those outside the limiting Ewald ball. As a result, it is possible, thanks to multiple scattering, to encode information in the far-field scattering amplitude about the high spatial frequencies of the object corresponding to the exterior of the limiting Ewald ball, leading to the possibility of superresolution.

There is a lot of evidence that multiple scattering can, indeed, lead to superresolution. This includes evidence based on experimental data (Chen and Chew, 1998), numerical inversions incorporating multiple scattering (Cui *et al.*, 2004; Belkebir *et al.*, 2006), and algorithm-independent, estimation-theoretic studies based on the fundamental CRB corresponding to a given noise level (Sentenac *et al.*, 2007; Simonetti *et al.*, 2008; Marengo and Berestesky, 2013). On the other hand, it has been found that while multiple scattering is typically beneficial and enables superresolution beyond what is possible for weak scatterers obeying the Born approximation, for certain sensing conditions its impact is minor and can even be detrimental (Sentenac *et al.*, 2007; Marengo and Berestesky, 2013). The quantification of the role of multiple scattering in superresolution is so far inconclusive and is an important area of ongoing research.

Superresolution

The resolution limit discussed so far applies to conventional imaging systems employing far-field probing (involving propagating waves only) and far-field sensing. If the interrogation of the object involves probing or sensing in the near zone, then imaging resolution is greatly improved (leading to “superresolution”) thanks to the sensing of the high-frequency information about the object that is contained in the corresponding near-field evanescent waves. This is the idea behind near-field optics techniques such as near-field scanning microscopy (Betzig and Trautman, 1992) and total internal reflection tomography (Carney and Schotland, 2001; Chaumet *et al.*, 2004).

Another approach to superresolution is based on the analyticity of the radiated fields, and consists of analytically continuing the field data associated to the propagating (or visible) region of the plane wave spectrum into the complementary evanescent (or invisible) region of the spectrum (Goodman, 2005, Section 6.6). However, the required analytic continuation is highly sensitive to noise in the data, and consequently this method is hard to implement in practice. A more reliable methodology which is well-known (Schmidt-Weinmar, 1978) to render superresolution in many applications is to incorporate into the inversion and imaging algorithms prior information about the object, such as support information (see Marengo *et al.*, 1999, 2000; Schmidt-Weinmar, 1978; Marengo, 2014; Goodman, 2005, Section 6.6.4). In the latter approach, the superresolution is due to the exploitation of prior information about the sought-after object. Here it is important to emphasize that the focus of the present contribution has been on the classical assessment of resolution, in which no prior information about the object is assumed. In modern imaging, several techniques have emerged which successfully exploit prior information, and can even achieve superresolution thanks to the relevant prior. In this connection, the incorporation of sparsity priors by means of the methods of compressive sensing has been particularly successful (Candes, 2006).

Recently, there has also been a lot of interest in achieving superresolution via direct physical imaging using innovative devices and sensors. One of these approaches is the so-called superlens proposed by Pendry (2000), in which left-handed metamaterials play a key role. In these approaches, the idea is to build a device that can amplify or sense in the far zone the evanescent waves emanating from a radiating object. The experimental realization has been pursued by many groups (Grbic and Eleftheriades, 2004), and several alternative technologies such as the hyperlens (Jacob *et al.*, 2006) have also been proposed.

References

- Belkebir, K., Chaumet, P.C., Sentenac, A., 2006. Influence of multiple scattering on three-dimensional imaging with optical diffraction tomography. *Journal of the Optical Society of America A* 23, 586–595.
- Belzig, E., Trautman, J.K., 1992. Near-field optics: Microscopy, spectroscopy, and surface modification beyond the diffraction limit. *Science* 257, 189–195.
- Bleistein, N., Cohen, J.K., 1977. Nonuniqueness in the inverse source problem in acoustics and electromagnetics. *Journal of Mathematical Physics* 18, 194–201.
- Bouwkamp, C.J., de Bruijn, N.G., 1946. The problem of optimum antenna current distribution. *Philips Research Reports* 1, 135–158.
- Buxton, A., 1937. Note on optical resolution. *Philosophical Magazine* 23, 440–442.
- Candes, E.J., 2006. Compressive sampling. In: *Proceedings of the International Congress of Mathematicians, Madrid*, pp. 1–20.
- Carney, P.S., Schotland, J.C., 2001. Three-dimensional total internal reflection tomography. *Optics Letters* 26, 1072–1074.
- Chaumet, P.C., Belkebir, K., Sentenac, A., 2004. Superresolution of three-dimensional optical imaging by use of evanescent waves. *Optics Letters* 29, 2740–2742.
- Chen, F.-C., Chew, W.C., 1998. Experimental verification of super resolution in nonlinear inverse scattering. *Applied Physics Letters* 72, 3080–3082.
- Cui, T.J., Chew, W.C., Yin, X.X., Hong, W., 2004. Study of resolution and super resolution in electromagnetic imaging for half-space problems. *IEEE Transactions on Antennas and Propagation* 52, 1398–1411.
- den Dekker, A.J., van den Bos, A., 1997. Resolution: A survey. *Journal of the Optical Society of America A* 14, 547–557.
- Devaney, A.J., 2012. *Mathematical Foundations of Imaging, Tomography and Wavefield Inversion*. Cambridge: Cambridge University Press.
- Devaney, A.J., Wolf, E., 1973. Radiating and nonradiating classical current distributions and the fields they generate. *Physical Review* 8, 1044–1047.
- Gbur, G., 2003. Nonradiating sources and other invisible objects. In: Wolf, E. (Ed.), *Progress in Optics*, vol. 45. Amsterdam: Elsevier.
- Grbic, A., Eleftheriades, G.V., 2004. Overcoming the diffraction limit with a planar left-handed transmission-line lens. *Physical Review Letters* 92, 117403.
- Goodman, J.W., 2005. *Introduction to Fourier Optics*, third ed. Englewood, CO: Roberts & Company.
- Heintzmann, R., Ficz, G., 2006. Breaking the resolution limit in light microscopy. *Briefings in Functional Genomics and Proteomics* 5, 289–301.
- Houston, W.V., 1927. A compound interferometer for fine structure work. *Physical Review* 29, 478–484.
- Jacob, Z., Alekseyev, L.V., Narimanov, E., 2006. Optical hyperlens: Far-field imaging beyond the diffraction limit. *Optics Express* 14, 8247–8256.
- Kay, S.M., 1993. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. New Jersey, NJ: Prentice Hall.
- Lord Rayleigh, F.R.S., 1879. Investigations in optics, with special reference to the spectroscope. London, Edinburgh, and Dublin *Philosophical Magazine and Journal of Science*, Series 5 8, 261–274.
- Marengo, E.A., 2014. Inverse diffraction theory and computation of minimum source regions of far fields. *Mathematical Problems in Engineering* 2014, 1–18. Article ID: 513953.
- Marengo, E.A., Berestesky, P., 2013. Cramer–Rao bound study of multiple scattering effects in target separation estimation. *International Journal of Antennas and Propagation* 2013, 1–10. Article ID: 572923.
- Marengo, E.A., Devaney, A.J., Ziolkowski, R.W., 1999. New aspects of the inverse source problem with far field data. *Journal of the Optical Society of America A* 16, 1612–1622.
- Marengo, E.A., Devaney, A.J., Ziolkowski, R.W., 2000. Inverse source problem and minimum energy sources. *Journal of the Optical Society of America A* 17, 34–45.
- Marengo, E.A., Zambrano-Nunez, M., Berestesky, P., 2012. Cramer–Rao study of multiple scattering effects in target localization. *International Journal of Antennas and Propagation* 2012, 1–13. Article ID: 390312.
- Marengo, E.A., Ziolkowski, R.W., 2000. Nonradiating and minimum energy sources and their fields: Generalized source inversion theory and applications. *IEEE Transactions on Antennas and Propagation* 48, 1553–1562.
- Pendry, J.B., 2000. Negative refraction makes a perfect lens. *Physical Review Letters* 85, 3966–3969.
- Ronchi, V., 1961. Resolving power of calculated and detected images. *Journal of the Optical Society of America* 51, 458–460.
- Rushforth, C.K., Harris, R.W., 1968. Restoration, resolution, and noise. *Journal of the Optical Society of America* 58, 539–545.
- Schuster, A., 1924. *An Introduction to the Theory of Optics*, third ed. London: Arnold.
- Schmidt-Weinmar, H.G., 1978. Spatial resolution of subwavelength sources from optical far-zone data. In: Baltes, H.P. (Ed.), *Inverse Source Problems in Optics*. Berlin: Springer-Verlag, pp. 83–118.
- Sentenac, A., Guerin, C.-A., Chaumet, P.C., *et al.*, 2007. Influence of multiple scattering on the resolution of an imaging system: A Cramer–Rao analysis. *Optics Express* 15, 1340–1347.
- Simonetti, F., 2006. Multiple scattering: The key to unravel the subwavelength world from the far-field pattern of a scattered wave. *Physical Review E* 73, 036619.
- Simonetti, F., Fleming, M., Marengo, E.A., 2008. Illustration of the role of multiple scattering in subwavelength imaging from far-field measurements. *Journal of the Optical Society of America A* 25, 292–303.
- Sparrow, C.M., 1916. On spectroscopic resolving power. *Astrophysics Journal* 44, 76–86.
- Testorf, M.E., Fiddy, M.A., 2010. Superresolution imaging: Revisited. In: Hawkes, P.W. (Ed.), *Advances in Imaging and Electron Physics*, vol. 163. San Diego, CA: Academic Press.
- Toraldo di Francia, G., 1952. Super-gain antennas and optical resolving power. *Il Nuovo Cimento* 9, 426–438.
- Toraldo di Francia, G., 1955. Resolving power and information. *Journal of the Optical Society of America* 45, 497–501.
- Wolf, E., 1969. Three-dimensional structure determination of semi-transparent objects from holographic data. *Optics Communications* 1, 153–156.
- Woodward, P.M., Lawson, J.D., 1948. The theoretical precision with which an arbitrary radiation-pattern may be obtained from a source of a given size. *Journal of Electrical Engineering* 95, 363–370.
- Zheludev, N.I., 2008. What diffraction limit? *Nature Materials* 7, 420–422.

Coherent Diffractive Imaging

Lei Tian, Boston University, Boston, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The infamous “phase problem” stems from that optical detection devices (e.g., charge-coupled devices, photodetectors, and human eyes) cannot directly record the phase of the electromagnetic field (at 10^{15} Hz and higher), but only the average power, which is proportional to the magnitude square of the field. As a result, obtaining phase involves additional complexity, both experimentally and computationally.

In general, the first step of solving the phase problem is to convert the originally “invisible” phase information to the observable intensity variations, i.e., phase contrast. In fact, any *complex* transfer function will produce some phase contrast. For example, the Zernike phase contrast is achieved by introducing a $\pi/2$ phase shift (and thus a pure imaginary transfer function component) to the direct illumination beam. The second step is to carry out a phase retrieval procedure to computationally recover the quantitative phase values.

Many techniques have been developed to recover phase information. Depending on the physical mechanism of realizing phase contrast, they can be broadly categorized into interferometric and “non-interferometric” techniques – the former relies on interferometry and the latter does not. Here, we focus on the non-interferometric techniques. A distinctive advantage of these techniques is that phase contrast can be produced simply by some diffraction process, which means that the experimental setup can be made simple (e.g., lensless). The non-interferometric techniques are often referred to as “coherent diffractive imaging (CDI),” emphasizing the role of diffraction in forming phase contrast. As compared to interferometry, CDI often requires more sophisticated computational algorithms to reconstruct the phase, since the diffraction measurements are often related to the phase by nonlinear equations.

From a theoretical point of view, phase retrieval may not have a unique solution; furthermore, even if a unique solution exists, it is not guaranteed that the algorithm will converge to the global solution because the underlying problem is nonlinear and nonconvex. The success of the phase retrieval algorithms depends on not only the numerical procedure itself, but also the physical measurement process that defines the nonlinear problem to solve. Generally speaking, successful phase retrieval requires both redundancy and diversity in the data. The optimal design involves trade-offs between experimental complexity, cost, accuracy, and computational requirement. The optimal implementation is highly application dependent, which requires joint designs of physical optical instrumentation and computation.

Fourier Measurement-Based Techniques

The Fourier measurement is the most widely used method in CDI. A typical experimental setup based on far-field diffraction is illustrated in Fig. 1. Under coherent illumination, the measured far-field intensity is simply the Fourier intensity of the object,

$$I(\mathbf{u}) = \left| \int g(\mathbf{r}) \exp\{-i2\pi\mathbf{r} \cdot \mathbf{u}\} d^2\mathbf{r} \right|^2 = |\mathcal{F}\{g\}|^2 \quad (1)$$

where $\mathcal{F}\{\cdot\}$ denotes the two-dimensional (2D) Fourier transform (FT), $g(\mathbf{r}) = |g(\mathbf{r})| \exp\{i\phi(\mathbf{r})\}$ is the object’s complex transmittance function having amplitude $|g(\mathbf{r})|$ and phase $\phi(\mathbf{r})$, and $\mathbf{r}$ and $\mathbf{u}$ denote the spatial and Fourier coordinates, respectively. The knowledge of the Fourier intensity alone does not allow unique recovery of the object, since there are infinitely many possible phase functions that satisfy the equation. The general strategy is to introduce additional prior knowledge and/or constraints to

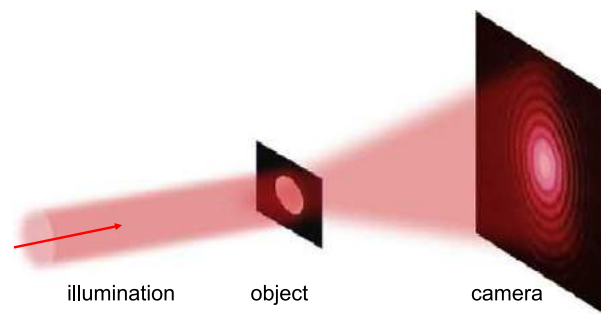


Fig. 1 Schematics of the Fourier measurement-based coherent diffractive imaging (CDI) technique.

regularize the problem. In the following, we will describe different algorithms that have been designed for solving problems that incorporate different priors.

Alternating Projection Algorithms

Alternating projection is the most popular class of phase retrieval algorithms. The general idea is to treat the constraint at the Fourier domain as one set, and the constraint at the real-space (at the object plane) as the other. The algorithm then iteratively updates the solution by projecting the estimated field between the two sets in an alternating order.

The pioneering work of Gerchberg and Saxton (GS) in 1972 assumes both the Fourier magnitude $|G(\mathbf{u})| = \sqrt{I(\mathbf{u})}$ and the real-space magnitude $|g(\mathbf{r})|$ are known. The algorithm starts with some phase initial guess at the real space, for example, all-zero phase or uniformly random-distributed values within certain range. It then iteratively imposes the real-space and Fourier magnitude constraints in the following four steps, as illustrated in Fig. 2. In the s th iteration, the algorithm first takes the FT of the estimated object $g_s(\mathbf{r})$ to get the estimated Fourier spectrum $G_s(\mathbf{u})$,

$$G_s(\mathbf{r}) = \mathcal{F}\{g_s(\mathbf{u})\} \quad (2)$$

Second, the computed Fourier magnitude is then replaced by the measurement, while keeping the estimated phase unchanged.

$$G'_s(\mathbf{u}) = G_s(\mathbf{u}) \cdot \frac{\sqrt{I(\mathbf{u})}}{|G_s(\mathbf{u})|} \quad (3)$$

Third, it takes the inverse FT, $\mathcal{F}^{-1}\{\cdot\}$, of the updated Fourier spectrum $G'_s(\mathbf{u})$ to update the object field $g'_s(\mathbf{r})$,

$$g'_s(\mathbf{r}) = \mathcal{F}^{-1}\{G'_s(\mathbf{u})\} \quad (4)$$

Fourth, the real-space magnitude constraint is imposed by replacing the estimate with the known values, while keeping the estimated phase unchanged.

$$g_{s+1}^{\text{GS}}(\mathbf{r}) = |g(\mathbf{r})| \cdot \frac{g'_s(\mathbf{r})}{|g'_s(\mathbf{r})|} \quad (5)$$

The iteration continues until the reconstruction error between the estimated and the measured Fourier magnitude, $\sum_{\mathbf{u}} ||G_s(\mathbf{u})| - \sqrt{I(\mathbf{u})}|$, below certain value.

The error-reduction (ER) algorithm proposed by Fienup in 1978 extends the GS algorithm to incorporate other types of priors, such as the object support and nonnegativity constraints. The ER algorithm follows the similar four steps in each iteration.

The only difference is in the fourth step, in which other real-space constraints are imposed by

$$g_{s+1}^{\text{ER}}(\mathbf{r}) = \begin{cases} g'_s(\mathbf{r}), & m \notin \gamma \\ 0, & m \in \gamma \end{cases} \quad (6)$$

where γ defines the set of indices for which $g'_s(\mathbf{r})$ violates the real-space constraints.

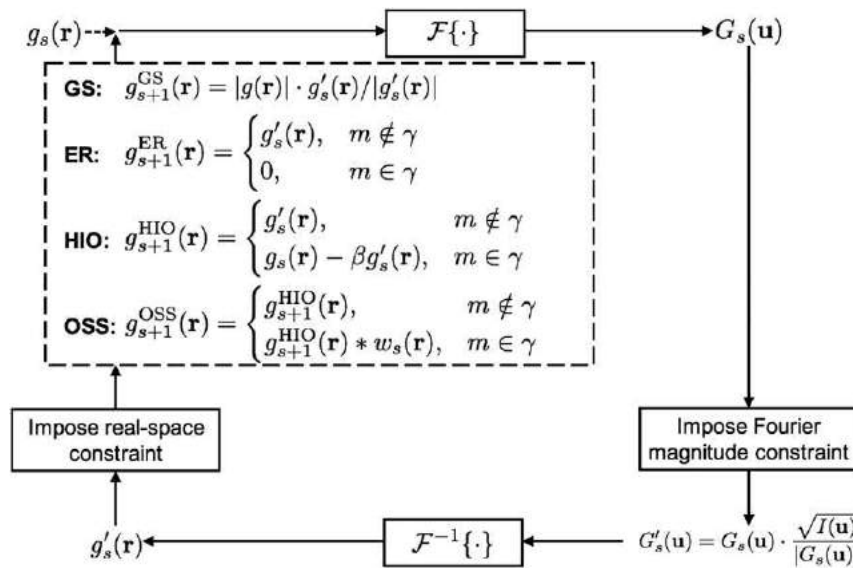


Fig. 2 Block diagram of the Gerchberg–Saxton (GS), error reduction (ER), hybrid input–output (HIO), and oversampling smoothness (OSS) algorithms.

The reconstruction errors of both GS and ER algorithms are shown to be monotonically decreasing with iterations. Despite this convergence property, it is not guaranteed that the true solution can be found, since the algorithm may converge to a local minimum. The underlying problem, which can be written in the form of a multivariate minimization of a cost function with respect to the unknown phase function, is both nonlinear and nonconvex in a high-dimensional space. Here, the (non)convexity characterizes the hyper-surface formed by the cost function. A nonconvex surface consists of multiple peaks and valleys, which implies the existence of multiple local minimal solutions.

The hybrid input–output (HIO) algorithm is another commonly used alternating projection algorithm. It follows the same iterative framework as GS and ER with a modification to the fourth step, as given by

$$g_{s+1}^{\text{HIO}}(\mathbf{r}) = \begin{cases} g_s'(\mathbf{r}), & m \notin \gamma \\ g_s(\mathbf{r}) - \beta g_s'(\mathbf{r}), & m \in \gamma \end{cases} \quad (7)$$

where β is a feedback constant, with a typical value between 0.5 and 1. Empirically, it is shown that HIO is able to avoid local minima and converges to the global solution in the noise-free case. In practice, the solution tends to oscillate as the iteration increases. The algorithm may stagnate and fails to converge to the global minimum.

Many algorithms have been developed to overcome these limitations, including the combination of HIO and ER, solvent flipping (SF), different map (DM), hybrid projection–reflection (HPR), average successive reflections (ASR), relaxed averaged alternating reflectors (RAAR), oversampling smoothness (OSS), and continuous HIO (CHOI). As an example, both CHIO and OSS attempt to smooth the updates by removing sharp transitions in the estimates. The OSS update is related to HIO by

$$g_{s+1}^{\text{OSS}}(\mathbf{r}) = \begin{cases} g_{s+1}^{\text{HIO}}(\mathbf{r}), & m \notin \gamma \\ g_{s+1}^{\text{HIO}}(\mathbf{r}) * w_s(\mathbf{r}), & m \in \gamma \end{cases} \quad (8)$$

where $*$ denotes convolution. For the set of indices that violates the real-space constraints, the HIO solution is low-pass filtered by a Gaussian function $w_s(\mathbf{r})$, whose variance increases with iterations.

The convergence property of HIO and its variants is greatly affected by the support constraint. In many applications, the support is not exactly known a priori. Estimation techniques have been developed based on the support of the object's auto-correlation function (i.e., FT of the measured Fourier intensity). The support can be further refined during iterations using the “shrink-wrap” algorithm.

The performance of the alternating projection phase retrieval algorithm is also dependent on the initial guess and it may converge to different solutions due to the nonconvexity of the problem. Initializations with different random phase distributions have been proposed. The final solution is then constructed by some numerical averaging scheme. For example, the phase retrieval transform function (PRTF) is proposed to quantify the computationally recovered spatial resolution, as defined by

$$\text{PRTF}(\mathbf{u}) = \langle |\mathcal{F}\{\hat{g}_t(\mathbf{r})\}| \rangle_t / \sqrt{I(\mathbf{u})} \quad (9)$$

where $\hat{g}_t(\mathbf{r})$ denotes the estimated object from the t th reconstruction, and $\langle \cdot \rangle$ takes the average over all the reconstructions using different initial guesses. In cases where a low-resolution object is known or can be first estimated and then used as the initial guess, the algorithm tends to converge to the global solution more reliably and reconstruct high resolution. Examples include the “expanding Fourier modulus” method and ptychographic algorithms.

Partial Coherence Algorithms

The FT model in Eq. (1) assumes that the illumination beam is spatially and temporally fully coherent, which does not hold for many applications. Directly applying the coherent phase retrieval algorithm suffers from significant reconstruction artifacts in partially coherent CDI problem.

The diffraction intensity $I^{\text{PC}}(\mathbf{u})$ under partially coherent illumination is described through the mutual intensity function $J(\mathbf{r}_1, \mathbf{r}_2)$, which is a measure of the statistical correlations between a pair of points at $\mathbf{r}_1$ and $\mathbf{r}_2$ in the partially coherent field,

$$I^{\text{PC}}(\mathbf{u}) = \int \int J(\mathbf{r}_1, \mathbf{r}_2) g(\mathbf{r}_1) g^*(\mathbf{r}_2) \exp\{-i2\pi\mathbf{u} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\} d^2\mathbf{r}_1 d^2\mathbf{r}_2 \quad (10)$$

The first approach to simplify Eq. (10) is to use the coherent mode decomposition model, which states that any partially coherent field can be decomposed into a set of coherent fields that are mutually incoherent,

$$J(\mathbf{r}_1, \mathbf{r}_2) = \sum_{n=1}^N \mu_n \psi_n(\mathbf{r}_1) \psi_n^*(\mathbf{r}_2) \quad (11)$$

The coherent modes are described by a set of orthonormal functions $\psi_n(\mathbf{r})$. The real, nonnegative number μ_n characterizes the amount of energy contained in each mode. Eq. (11) assumes that the majority of the energy in the illumination is contained in the first N modes. The diffraction intensity can be rewritten as

$$I^{\text{PC}}(\mathbf{u}) = \sum_{n=1}^N \mu_n \left| \int \psi_n(\mathbf{r}) g(\mathbf{r}) \exp\{-i2\pi\mathbf{u} \cdot \mathbf{r}\} d^2\mathbf{r} \right|^2 \quad (12)$$

which is the incoherent sum of the coherent diffraction intensities from the object illuminated with each coherent mode.

The second approach further assumes the illumination is statistically stationary and the mutual intensity function depends only on the spatial separations $\Delta \mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$. The diffraction intensity takes the following convolution form,

$$I^{\text{PC}}(\mathbf{u}) = \left| \int g(\mathbf{r}) \exp \{-i2\pi \mathbf{u} \cdot \mathbf{r}\} d^2 \mathbf{r} \right|^2 * S(\mathbf{u}) = I^{\text{C}}(\mathbf{u}) * S(\mathbf{u}) \quad (13)$$

where $I^{\text{C}}(\mathbf{u})$ is the intensity under coherent illumination; $S(\mathbf{u})$ is the FT of $J(\Delta \mathbf{r})$, which represents the intensity distribution of the primary source according to the Van Cittert–Zernike theorem.

When the coherence function of the illumination is known, the partially coherent phase retrieval can be achieved using a modified alternating projection algorithm that follows the similar four-step process. The only modification is to the second Fourier spectrum estimation step,

$$G'_s(\mathbf{u}) = G_s(\mathbf{u}) \cdot \sqrt{\frac{I(\mathbf{u})}{I_s^{\text{PC}}(\mathbf{u})}} \quad (14)$$

where $I_s^{\text{PC}}(\mathbf{u})$ denotes the estimated partially coherent diffraction intensity using either Eq. (12) or (13) in the s th iteration.

It has been shown heuristically that simultaneous recovery of the primary source (Eq. (13)) and the object function is possible from a single Fourier measurement and the object support constraint. The solution is found by solving the minimization problem,

$$\min_{\mathbf{g}, \mathbf{S}} \|\mathbf{I} - \mathbf{I}^{\text{C}} * \mathbf{S}\|_2^2 \quad (15)$$

where $\|\cdot\|_2^2$ represents the ℓ_2 norm of a vector, $\mathbf{g}$, $\mathbf{I}$, and $\mathbf{S}$ are vectors presenting the discretized object, intensity measurement, and source function, respectively. An alternating minimization strategy can be used that alternates between the two subproblems of coherent phase retrieval and source recovery in each iteration. The coherent phase retrieval can be implemented using any previous algorithm. The source is found by linear deconvolution assuming the coherent intensity is known. Since the full problem is nonlinear and nonconvex, the solution may still converge to a local minimum.

Three-Dimensional Recovery Problems

The 2D Fourier model in Eq. (1) relies on the assumption that the object is thin; the object function is determined by the projection approximation. When the object is thick and *weakly scattering*, the three-dimensional (3D) scattering process can be modeled using the first Born approximation, which accounts for only single scattering. The 3D CDI technique attempts to reconstruct the volumetric distribution of the object, which is characterized by the complex refractive index distribution $n(\mathbf{r}, z)$ as

$$g(\mathbf{r}, z) = \frac{\pi}{\lambda^2} (1 - n^2(\mathbf{r}, z)) \quad (16)$$

Denote the 3D FT of the object as $G(\mathbf{u}, w) = \mathcal{F}_{3D}\{g(\mathbf{r}, z)\}$, where w denotes the spatial frequency along the axis of illumination (z -axis). A single diffraction intensity $I(\mathbf{u})$ corresponds to a spherical cap on the Ewald sphere,

$$I(\mathbf{u}) = \left| G\left(\mathbf{u}, \sqrt{\frac{1}{\lambda^2} - |\mathbf{u}|^2} - \frac{1}{\lambda}\right) \right|^2 \quad (17)$$

To better cover the object's 3D Fourier space, diffraction data need to be collected at various orientations by either rotating the object or the illumination. In the reconstruction, alternating projection algorithms in Section Alternating Projection Algorithms have been extended to solve the 3D phase retrieval problem.

Defocus-Based Techniques

Defocus is another popular CDI method that is based on diffraction in free-space. The experimental procedure involves simply moving the camera (or the object) axially, while capturing multiple images through-focus, as illustrated in Fig. 3. The in-focus

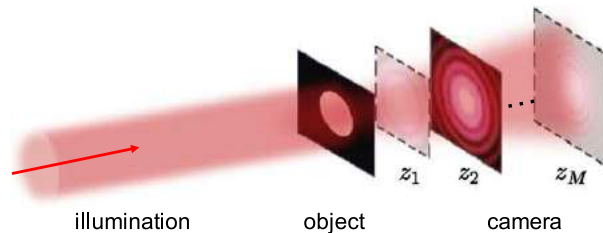


Fig. 3 Schematics of defocus-based coherent diffractive imaging (CDI) techniques.

intensity image contains no phase information. The defocused intensity is characterized by the depth (z) dependent free-space point-spread-function (PSF) $h(\mathbf{r}, z)$. The measured intensity at z under coherent illumination is

$$I(\mathbf{r}; z) = \left| \int g(\mathbf{r}') h(\mathbf{r} - \mathbf{r}', z) d^2 \mathbf{r}' \right|^2 = |\mathcal{P}_z\{g\}|^2 \quad (18)$$

where $\mathcal{P}_z\{\cdot\}$ denotes the propagator at defocus distance z .

Many phase retrieval algorithms have been demonstrated for defocus measurements. The validity of each method depends on many experimental parameters, such as the defocus range, the object property, and the illumination coherence.

Iterative Algorithms

To solve the nonlinear Eq. (18), one can employ iterative algorithms. When only a single defocused intensity is available, the technique is also known as Fresnel CDI, which is analogous to the Fourier technique. All the iterative algorithms discussed in Section Fourier Measurement-Based Techniques can be applied by simply replacing $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ by the (Fresnel) forward-propagator $\mathcal{P}_z\{\cdot\}$ and back-propagator $\mathcal{P}_z^{-1}\{\cdot\} = \mathcal{P}_{-z}\{\cdot\}$, respectively. All the limitations in the Fourier techniques still apply – the algorithm may stagnant and not converge due to the nonlinearity and nonconvexity.

In many practical applications, it is possible to take a series of defocused intensity measurement, $I(\mathbf{r}; z_m)$, $m=1, 2, \dots, M$. The most widely applied phase retrieval algorithm is the modified GS algorithm. In each iteration, one propagates the estimated field to each defocused planes, where the estimated amplitude is replaced by the measurement. To impose the amplitude constraints across multiple planes, two updating schemes have been implemented: sequential and global. The sequential GS algorithm imposes the amplitude constraint one plane at a time as the field propagates between the defocused and the object planes. The global GS algorithm instead updates the object field by averaging over estimates by imposing amplitude constraints from all defocused planes.

An alternative approach is based on solving the minimization of a cost function using the nonlinear optimization framework. For example, consider the amplitude-based least-square problem,

$$\min_{\mathbf{g}} E(\mathbf{g}) = \frac{1}{2} \sum_{m=1}^M \left\| \sqrt{\mathbf{I}_m} - |\mathbf{P}_{z_m} \mathbf{g}| \right\|_2^2 \quad (19)$$

where $\mathbf{g}$ and $\mathbf{I}_m$ are vectors presenting the discretized object function and intensity measurement at z_m , respectively. The discrete propagator $\mathbf{P}_{z_m}$ can be efficiently implemented using the fast Fourier transform (FFT), $\mathbf{P}_{z_m} = \mathbf{K}^H \mathbf{H}_{z_m} \mathbf{K}$, where $\mathbf{K}$ is the 2D discrete Fourier transform (DFT) matrix, $\mathbf{H}_{z_m} = \text{diag}(\exp[-i\pi\lambda z_m |\mathbf{u}|^2])$ is a diagonal matrix describing the Fresnel transfer function at the distance z_m , $\mathbf{u}$ is the discrete spatial frequency coordinates along the lateral dimensions, λ is the wavelength, $\text{diag}(\mathbf{a})$ denotes the diagonal matrix with its corresponding diagonal entries equal to the elements in the vector $\mathbf{a}$, and $\cdot^H$ denotes the matrix Hermitian. The back-propagation matrix $\mathbf{P}_{z_m}^{-1}$ satisfies $\mathbf{P}_{z_m}^{-1} = \mathbf{P}_{z_m}^H$, since $\mathbf{P}_{z_m}$ is unitary.

The reconstruction algorithm is then implemented following standard gradient-based optimization procedure. It starts with some initial guess of the object $\mathbf{g}_0$. At the s th iteration, the general equation for updating the object is

$$\mathbf{g}_{s+1} = \mathbf{g}_s + \alpha_s \mathbf{d}_s \quad (20)$$

where α_s is a step size, which can be a constant or found by a line-search algorithm. The search direction $\mathbf{d}_s$ is crucial for the algorithm convergence, which is determined by the gradient and/or second-order derivative (Hessian) of the cost function.

Both the gradient and Hessian can be conveniently calculated using $\mathbb{C}\mathbb{R}$ -calculus, in which the complex variable $\mathbf{g}$ is augmented as $[\mathbf{g}, \mathbf{g}^*]^T$ and $\bullet^*$ denotes the complex conjugate. The gradient of the cost function in Eq. (19) is

$$\nabla E(\mathbf{g}, \mathbf{g}^*) = \sum_{m=1}^M \begin{bmatrix} -\mathbf{P}_{z_m}^H \cdot \text{diag}(\mathbf{P}_{z_m} \mathbf{g} / |\mathbf{P}_{z_m} \mathbf{g}|) \\ -\mathbf{P}_{z_m}^T \cdot \text{diag}(\mathbf{P}_{z_m}^* \mathbf{g}^* / |\mathbf{P}_{z_m} \mathbf{g}|) \end{bmatrix} [\sqrt{\mathbf{I}_m} - |\mathbf{P}_{z_m} \mathbf{g}|] \quad (21)$$

where $\cdot^T$ denotes the matrix transpose.

In the first-order gradient descent (GD) methods, the search direction is simply the gradient of the cost function. In the second-order methods, the Hessian of the cost function $\nabla^2 E(\mathbf{g}, \mathbf{g}^*)$ is used to improve convergence speed. The search direction is then obtained by solving the linear equation

$$\nabla^2 E(\mathbf{g}, \mathbf{g}^*) \begin{bmatrix} \mathbf{d} \\ \mathbf{d}^* \end{bmatrix} = -\nabla E(\mathbf{g}, \mathbf{g}^*) \quad (22)$$

The Hessian matrix in practical phase retrieval problem is huge (size of N^2 , where N is the size of the object), so directly computing the inverse of Hessian is computationally prohibitive. Quasi-Newton's methods, such as the limited-memory Broyden-Fletcher-Goldfarb-Shanno method (L-BFGS) circumvents these problems by approximating the inverse of Hessian from the gradients of a few previous iterations. In modified Gauss-Newton methods, the inverse of Hessian is computed iteratively (e.g., using the conjugate gradient method) by solving the linear Eq. (22). Both the L-BFGS and modified Gauss-Newton methods have been applied to defocus phase retrieval problems and achieve better performance than the GS and GD algorithms.

Recursive Algorithms

The recursive algorithms, such as Kalman filter, attempts to progressively improve the phase estimate as each new intensity is included in the reconstruction. The intensity evolution upon propagation is treated with a complex augmented state space model. Denote the field at the m th plane and its FT as $\mathbf{a}_m$ as $\mathbf{b}_m$, respectively. Propagation of the intensity between consecutive defocus planes can be written in the Fourier domain as

$$\mathbf{b}_{m-1} = \mathbf{K}\mathbf{a}_{m-1} \quad (23)$$

$$\mathbf{b}_m = \mathbf{H}_m \mathbf{b}_{m-1} \quad (24)$$

where $\mathbf{H}_m$ is the matrix describing propagation by a distance $z_m - z_{m-1}$. The observation model assumes noisy intensity measurements,

$$\mathbf{I}_m = |\mathbf{a}_m|^2 + \mathbf{v}_m \quad (25)$$

where $\mathbf{v}_m$ assumes pixel-dependent gaussian noise with zero mean and a covariance matrix $\mathbf{R}_m \equiv \langle \mathbf{v}_m \mathbf{v}_m^T \rangle = \text{diag}(|\mathbf{a}_m|^2)$.

In the reconstruction, the Kalman filter phase retrieval (KF-PR) updates the estimates of both the field as well as its covariance matrix. The method is shown to be robust in the presence of significant noise corruption. The main computational requirement of KF-PR is the inversion of the covariance matrix. Sparsity priors in the covariance matrix have been exploited to significantly reduce the computational cost. Partial coherence KF-PR algorithm has also been demonstrated using the convolution model in Section Partial Coherence Algorithms.

Direct-Inversion Methods

When the object is weakly scattering in the sense of the first Born approximation being valid, and/or the defocus distance is small as compared to the Fresnel distance, the nonlinear measurement can be linearized, which allows implementations of direct-inversion algorithms. The transport of intensity equation (TIE) relates phase and amplitude to variations of intensity along the optical axis z ,

$$\frac{\partial I(\mathbf{r})}{\partial z} = -\frac{\lambda}{2\pi} \nabla_{\perp} \cdot [I(\mathbf{r}) \nabla_{\perp} \phi(\mathbf{r})] \quad (26)$$

where $I(\mathbf{r})$ is the intensity at focus, $\phi(\mathbf{r})$ is the phase, and $\nabla_{\perp}$ denotes the gradient operator in lateral directions $\mathbf{r}$ only. The problem is linear for small defocus distances.

Inversion algorithms for the TIE can be understood by considering the case of a pure-phase object, where $I(\mathbf{r})$ is constant. In this situation, the TIE simplifies to a Poisson equation, whose solution can be efficiently found by an FFT-based algorithm,

$$\phi(\mathbf{r}) = \mathcal{F}^{-1} \left\{ \mathcal{F} \left\{ \frac{2\pi}{\lambda I} \frac{\partial I(\mathbf{r})}{\partial z} \right\} / (4\pi^2 |\mathbf{u}|^2 + \delta) \right\} \quad (27)$$

The small regularization constant δ prevents instabilities around zero frequencies. When $I(\mathbf{r})$ is not constant, the inversion of the TIE requires solving two Poisson equations. The phase reconstruction algorithm does not require additional phase unwrapping step and is computationally efficient. In practice, one cannot measure intensity derivative directly and so it is usually estimated by finite difference methods. However, intensity is not linear upon propagation, which corrupts the derivative estimate. Higher-order TIE method has been demonstrated to mitigate this nonlinearity error by performing pixel-wise spatial domain polynomial fitting.

Contrast transfer function (CTF) is a linear model derived from the weak-object approximation,

$$g(\mathbf{r}) = \exp[-\alpha(\mathbf{r}) + i\phi(\mathbf{r})] \approx 1 - \alpha(\mathbf{r}) + i\phi(\mathbf{r}) \quad (28)$$

where $\alpha(\mathbf{r})$ characterizes the object's absorption. The FT of intensity at a distance z , $\mathcal{I}(\mathbf{u}; z)$ is linearly related to the FT of the phase $\Phi(\mathbf{u})$ and absorption $A(\mathbf{u})$ as

$$\mathcal{I}(\mathbf{u}; z) = \delta(\mathbf{u}) - 2A(\mathbf{u}) \cos(\pi\lambda|\mathbf{u}|^2 z) + 2\Phi(\mathbf{u}) \sin(\pi\lambda|\mathbf{u}|^2 z) \quad (29)$$

where $\delta(\mathbf{u})$ denotes the Dirac delta function.

The CTF suggests that the low-frequency phase information is only captured at large defocus distances. The intensity's FT follows a sinusoidal oscillation along the propagation direction with a frequency $\pi\lambda|\mathbf{u}|^2$, which increases as the spatial frequency becomes larger. The optimal design of defocus distances has shown to be exponentially spaced, which significantly reduces the number of images required to ensure high-quality phase recovery.

Phase reconstruction can be achieved by solving a linear deconvolution problem,

$$\min_{A, \Phi} \sum_m \sum_{\mathbf{u}} \left| \mathcal{I}(\mathbf{u}; z_m) - [\delta(\mathbf{u}) - 2A(\mathbf{u}) \cos(\pi\lambda|\mathbf{u}|^2 z_m) + 2\Phi(\mathbf{u}) \sin(\pi\lambda|\mathbf{u}|^2 z_m)] \right|^2 + \sum_{\mathbf{u}} [\alpha \|A(\mathbf{u})\|_2^2 + \beta \|\Phi(\mathbf{u})\|_2^2] \quad (30)$$

An alternative algorithm is based on pixel-wise spatial frequency domain curve-fitting to sinusoids. A Gaussian process regression-based method takes the CTF as the guideline to apply a thresholding operation in the curve-fitting process for suppressing the unwanted high frequency components in the fitted functions. The algorithm has been empirically shown to be robust even beyond the weak-object approximation, where the actual data deviates from the CTF model.

Transverse Translation-Based Techniques

The main limitation of the Fourier and defocus-based CDI techniques is the requirement of an isolated object. In the Fourier CDI technique, the size of the object cannot exceed the reciprocal of the detector's pixel size, due to the sampling requirement and Fourier transform relation between the object and image space. In the defocus based CDI technique, the object should be smaller than the detector. The transverse translation-based technique, commonly known as "ptychography," circumvents this problem by taking a series of Fourier diffraction images, while translating the illumination or the object, as illustrated in Fig. 4. A phase retrieval algorithm is then devised to *coherently* "stitch" all the intensity images from each sub field-of-view (FOV).

At the j th transverse position $\mathbf{r}_j$, the intensity $I_j(\mathbf{u})$ is simply the far-field diffraction of the exit wave from each sub-FOV, which can be modeled as the product of the object $g(\mathbf{r})$ with the shifted "probe" (illumination) function $p(\mathbf{r} - \mathbf{r}_j)$, assuming the object is thin,

$$I_j(\mathbf{u}) = \left| \int g(\mathbf{r}) p(\mathbf{r} - \mathbf{r}_j) \exp \{-i2\pi \mathbf{r} \cdot \mathbf{u}\} d^2 \mathbf{r} \right|^2 = |\mathcal{F}\{g(\mathbf{r}) p(\mathbf{r} - \mathbf{r}_j)\}|^2, \quad j = 1, 2, \dots, J \quad (31)$$

The phase retrieval algorithm reconstructs $g(\mathbf{r})$ from multiple intensity measurements at different probe positions. The success of this technique is highly dependent on the experimental parameters, such as the amount of overlap in the probe regions, the object property, and the illumination coherence. For example, a minimum of 60% of probe overlap has been reported both numerically and experimentally. It should be noted that even with the multiple intensity measurements, the underlying problem is still nonlinear and nonconvex, so the solution can get stuck in local minima. It has been shown empirically that transverse translation-based techniques significantly improve the convergence property, as compared to Fourier and defocused based techniques, because of the redundancy and diversity in the dataset.

Alternating Projection Algorithms

The most widely used ptychographic phase retrieval algorithm is based on the alternating projection framework. Similar to the modified GS algorithm in Section Iterative Algorithms, the multiple translational measurements can be processed using either the sequential or global scheme. The iterative algorithm allows simultaneous recovery of both the object and the probe functions.

The sequential algorithm is initialized with $g_0(\mathbf{r})$ and $p_0(\mathbf{r})$. In each iteration, it repeatedly applies two projection operations to each measurement one-by-one. In the j th subroutine within the s th iteration, it starts by calculating each exit field $\psi_{j,s}(\mathbf{r})$ from the j th sub-FOV,

$$\psi_{j,s}(\mathbf{r}) = g_{j,s}(\mathbf{r}) p_{j,s}(\mathbf{r} - \mathbf{r}_j) \quad (32)$$

The Fourier projection Π_F^S is applied to each exit field by imposing the amplitude constraint in two steps. It first replaces the Fourier magnitude by the measurement, while keeping the phase unchanged. An updated exit field, $\psi'_{j,s}(\mathbf{r})$ is then calculated by taking the inverse FT:

$$\Pi_F^S[\psi_{j,s}(\mathbf{r})] : \quad \Psi'_{j,s}(\mathbf{u}) = \mathcal{F}\{\psi_{j,s}(\mathbf{r})\} \cdot \frac{\sqrt{I_j(\mathbf{u})}}{|\mathcal{F}\{\psi_{j,s}(\mathbf{r})\}|} \quad (33)$$

$$\psi'_{j,s}(\mathbf{r}) = \mathcal{F}^{-1}\{\Psi'_{j,s}(\mathbf{u})\} \quad (34)$$

Next, the algorithm implements the overlap projection Π_O^S , which updates the object and probe functions:

$$\Pi_O^S[\psi_{j,s}(\mathbf{r})] : \quad g_{j+1,s}(\mathbf{r}) = g_{j,s}(\mathbf{r}) + \alpha \frac{p_{j,s}^*(\mathbf{r} - \mathbf{r}_j)}{|p_{j,s}(\mathbf{r})|_{\max}^2} [\psi'_{j,s}(\mathbf{r}) - \psi_{j,s}(\mathbf{r})] \quad (35)$$

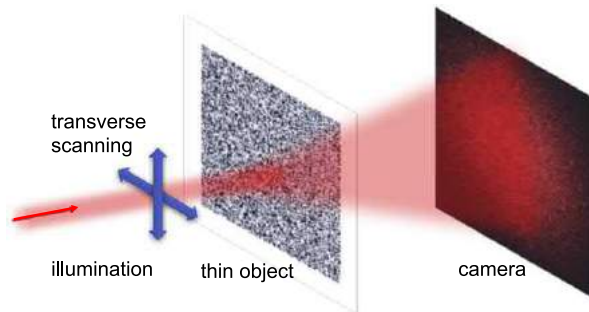


Fig. 4 Schematics of transverse translation-based coherent diffractive imaging (CDI) techniques.

$$p_{j+1,s}(\mathbf{r}) = p_{j,s}(\mathbf{r}) + \beta \frac{g_{j,s}^*(\mathbf{r} + \mathbf{r}_j)}{|g_{j,s}(\mathbf{r})|_{\max}^2} \left[\psi'_{j,s}(\mathbf{r} + \mathbf{r}_j) - \psi_{j,s}(\mathbf{r} + \mathbf{r}_j) \right] \quad (36)$$

where α and β are step sizes and are often set to 1. The results are then used as the input to the $(j+1)$ th subroutine until all the diffraction intensities have been used.

In the global alternating projection scheme, the Fourier projection Π_F^G is accomplished by imposing all the amplitude constraints simultaneously to all the exit wave $\psi_s(\mathbf{r}) = (\psi_{1,s}(\mathbf{r}), \psi_{2,s}(\mathbf{r}), \dots, \psi_{j,s}(\mathbf{r}))$:

$$\Pi_F^G[\psi_s(\mathbf{r})] = \left(\Pi_F^S[\psi_{1,s}(\mathbf{r})], \Pi_F^S[\psi_{2,s}(\mathbf{r})], \dots, \Pi_F^S[\psi_{j,s}(\mathbf{r})] \right) \quad (37)$$

The overlap projection, Π_O^G , updates the object and probe functions by taking weighted averages:

$$\Pi_O^G[\psi_s(\mathbf{r})] : g_{s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J p_s^*(\mathbf{r} - \mathbf{r}_j) \psi'_{j,s}(\mathbf{r})}{\sum_{j=1}^J |p_s(\mathbf{r} - \mathbf{r}_j)|^2} \quad (38)$$

$$p_{s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J g_s^*(\mathbf{r} + \mathbf{r}_j) \psi'_{j,s}(\mathbf{r} + \mathbf{r}_j)}{\sum_{j=1}^J |g_s(\mathbf{r} + \mathbf{r}_j)|^2} \quad (39)$$

Other alternating projection algorithms, as introduced in Section Alternating Projection Algorithms, can also be applied. For example, the DM algorithm has been widely used in ptychographic phase retrieval and takes the following procedure to update the exit field,

$$\psi_{s+1}(\mathbf{r}) = \psi_s(\mathbf{r}) + \Pi_F^G[2\Pi_O^G[\psi_s(\mathbf{r})] - \psi_s(\mathbf{r})] - \Pi_O^G[\psi_s(\mathbf{r})] \quad (40)$$

Mode Multiplexing Algorithms

The redundancy and diversity in the ptychographic dataset makes it possible to recover additional information besides the object and probe functions. Consider applying the mode decomposition model (Eq. (11)) to both the probe and the object, the diffraction pattern at the j th transverse position is,

$$I_j(\mathbf{u}) = \sum_{n=1}^N \sum_{m=1}^M \left| \int g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j) \exp \{-i2\pi \mathbf{r} \cdot \mathbf{u}\} d^2 \mathbf{r} \right|^2 = \sum_{n=1}^N \sum_{m=1}^M |\mathcal{F}\{g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j)\}|^2, \quad j = 1, 2, \dots, J \quad (41)$$

where $g_{(m)}(\mathbf{r})$ is the m th mode of the object and $p_{(n)}(\mathbf{r})$ the n th mode of the probe.

The alternating projection algorithms can be extended to the mode decomposition problem to simultaneously reconstruct the M mutually incoherent object modes and the N mutually incoherent probe modes. Each iteration starts by calculating the $M \times N$ coherent exit field from all possible pairs of the object and probe modes,

$$\psi_{(m,n)}(\mathbf{r}) = g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j) \quad (42)$$

The Fourier and overlap projections are then applied similar to the coherent case. In the sequential scheme, the Fourier update Eq. (33) is replaced by

$$\Psi'_{(m,n),j,s}(\mathbf{u}) = \mathcal{F}\{\psi_{(m,n),j,s}(\mathbf{r})\} \cdot \frac{\sqrt{I_j(\mathbf{u})}}{\sqrt{\sum_{n=1}^N \sum_{m=1}^M |\mathcal{F}\{\psi_{(m,n),j,s}(\mathbf{r})\}|^2}} \quad (43)$$

which is applied to all the exit field. The rest of the steps follow the same operations as described in Eqs. (34–36), which are applied simultaneously to each exit wave.

The global scheme makes the same modifications to the Fourier projection as the sequential scheme. The overlap projection updates all the object and probe modes by taking the weighted average over all the modes and measurements,

$$g_{(m)s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J \sum_{n=1}^N p_{(n)s}^*(\mathbf{r} - \mathbf{r}_j) \psi'_{(m,n),j,s}(\mathbf{r})}{\sum_{j=1}^J \sum_{n=1}^N |p_{(n)s}(\mathbf{r} - \mathbf{r}_j)|^2} \quad (44)$$

and

$$p_{(n)s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J \sum_{m=1}^M g_{(m)s}^*(\mathbf{r} + \mathbf{r}_j) \psi'_{(m,n),j,s}(\mathbf{r} + \mathbf{r}_j)}{\sum_{j=1}^J \sum_{m=1}^M |g_{(m)s}(\mathbf{r} + \mathbf{r}_j)|^2} \quad (45)$$

The general mode decomposition model above has been applied to various situations. The probe mode decomposition model is useful to account for the effects of partially coherent illumination. A major limitation of ptychography is the long acquisition time required to collect all the images. “Fly-scan” ptychography attempts to improve the data collection speed by continuously scanning the probe during the exposure time. It has been shown that the motion blurs in fly-scan measurements can be accounted for using the probe mode decomposition model. Multi-spectral and polarization multiplexed ptychographic measurements have been attempted. Spectrally and polarization resolved reconstructions have been demonstrated by adapting the mode multiplexing algorithms to the problem of decomposing both the object and probe by their respective spectral and polarization channels.

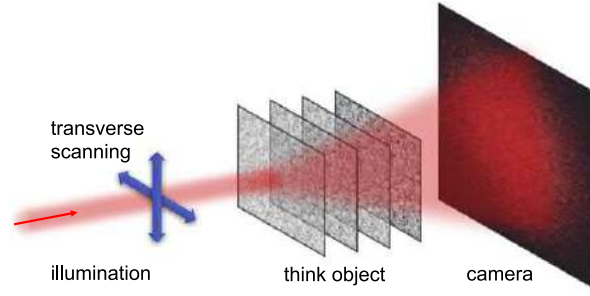


Fig. 5 Schematics of the multi-slice ptychography.

Multi-Slice Algorithms

The multiplicative model used in 2D ptychography breaks down when the object is thick and multiple scattering effects become nonnegligible. The multi-slice model is proposed to extend the depth of field and partially accounts for the forward multiple scattering effects.

A thick object is modeled as M thin slices, each described by a transmittance function $g^{(m)}(\mathbf{r})$ ($m = 1, 2, \dots, M$) (Fig. 5). The layer in-between is assumed uniform (e.g., air) of thickness Δz_m . The exit field from the object is modeled by a series of multiply-and-propagate operations,

$$\psi_j(\mathbf{r}) = \mathcal{P}_{\Delta z_{M-1}} \left\{ \dots \mathcal{P}_{\Delta z_2} \left\{ \mathcal{P}_{\Delta z_1} \left\{ p_j(\mathbf{r} - \mathbf{r}_j) g^{(1)}(\mathbf{r}) \right\} g^{(2)}(\mathbf{r}) \right\} \dots \right\} g^{(M)}(\mathbf{r}) \quad (46)$$

where the propagator can be computed using the angular spectrum method, $\mathcal{P}_{\Delta z}\{\cdot\} = \mathcal{F}^{-1} \left\{ \exp \left[i 2 \pi \Delta z \sqrt{1/\lambda^2 - |u|^2} \right] \mathcal{F}\{\cdot\} \right\}$. After passing through the object, the far-field diffraction intensity is measured,

$$I_j(\mathbf{u}) = |\mathcal{F}\{\psi_j(\mathbf{r})\}|^2 \quad (47)$$

The reconstruction algorithm can be implemented by extending the sequential 2D alternating projection algorithm to accounting for the multi-slice model. The j th subroutine within the s th iteration starts with the estimated j th exit field $\psi_{j,s}(\mathbf{r})$. Next, the Fourier projection is applied to update the exit field,

$$\psi'_{j,s}(\mathbf{r}) = \prod_{\mathbf{F}}^S [\psi_{j,s}(\mathbf{r})] \quad (48)$$

To update each object slice, the algorithm then recursively applies the “overlap-projection back-propagation” operation until the first slice is reached. The incident field $\phi_{j,s}^{(m)}(\mathbf{r})$ and the exit field $\psi_{j,s}^{(m)}(\mathbf{r})$ for the m th slice is related by

$$\psi_{j,s}^{(m)}(\mathbf{r}) = \phi_{j,s}^{(m)}(\mathbf{r}) g_{j,s}^{(m)}(\mathbf{r}) \quad (49)$$

Eq. (49) takes the same form as Eq. (32), except for the coordinate change. As a result, the overlap projection operation in 2D ptychography can be easily extended here. The modified overlap projection $\prod_{\mathbf{O}}^{S'}$ essentially separates the field scattered from the current slice from the field scattered from all previous slices,

$$\prod_{\mathbf{O}}^{S'} [\psi_{j,s}^{(m)}(\mathbf{r})] : g_{j+1,s}^{(m)}(\mathbf{r}) = g_{j,s}^{(m)}(\mathbf{r}) + \alpha \frac{\phi_{j,s}^{(m)*}(\mathbf{r})}{|\phi_{j,s}^{(m)}(\mathbf{r})|_{\max}^2} [\psi_{j,s}^{(m)'}(\mathbf{r}) - \psi_{j,s}^{(m)}(\mathbf{r})] \quad (50)$$

$$\phi_{j,s}^{(m)'}(\mathbf{r}) = \phi_{j,s}^{(m)}(\mathbf{r}) + \beta \frac{g_{j,s}^{(m)*}(\mathbf{r})}{|g_{j,s}^{(m)}(\mathbf{r})|_{\max}^2} [\psi_{j,s}^{(m)'}(\mathbf{r}) - \psi_{j,s}^{(m)}(\mathbf{r})] \quad (51)$$

The back-propagation operation is then applied to update the exit field of the $(m-1)$ th slice,

$$\psi_{j,s}^{(m-1)'}(\mathbf{r}) = \mathcal{P}_{\Delta z_{m-1}}^{-1} \left\{ \phi_{j,s}^{(m)'}(\mathbf{r}) \right\} \quad (52)$$

At the first slice, the overlap projection of 2D ptychography is applied to update the object slice and the probe functions: $\prod_{\mathbf{O}}^S [\psi_{j,s}^{(1)}(\mathbf{r})]$.

Illumination Angle Scanning-Based Fourier Ptychography

Fourier ptychography (FP) is a computational microscopy technique that achieves gigapixel imaging capability with both wide FOV and high resolution in both phase and amplitude. The working principle of FP can be understood as the implementation of the regular ptychography in the Fourier conjugate planes. The setup involves scanning the illumination across different angles, as illustrated in (Fig. 6).

Consider a thin sample with transmission function $g(\mathbf{r})$. The object is illuminated by a plane wave with the spatial frequency $\mathbf{u}_\ell$, defined by $\exp(i 2 \pi \mathbf{u}_\ell \cdot \mathbf{r})$. After passing through the object, the exit wave is the product of the object and the illumination,

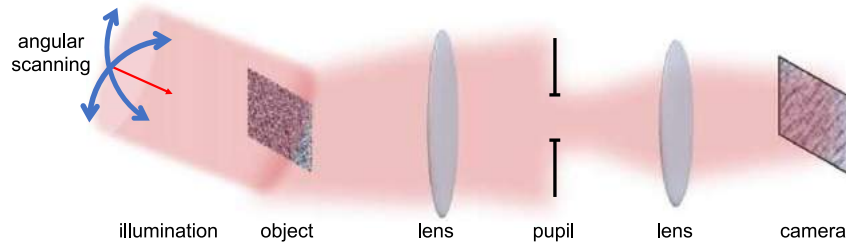


Fig. 6 Schematics of Fourier ptychography.

$g(\mathbf{r}) \exp(i2\pi\mathbf{u}_\ell \cdot \mathbf{r})$. In the Fourier space, this corresponds to a shifted version of the Fourier spectrum of the object, $G(\mathbf{u} - \mathbf{u}_\ell)$, where $G(\mathbf{u}) = \mathcal{F}\{g(\mathbf{r})\}$. This Fourier spectrum is then low-pass filtered by the pupil function, $P(\mathbf{u})$. The final intensity taken by the real-space camera is

$$I_\ell(\mathbf{r}) = |\mathcal{F}\{P(\mathbf{u})O(\mathbf{u} - \mathbf{u}_\ell)\}|^2 \quad (53)$$

The goal of the FP algorithm is to stitch together all the diffraction intensity corresponding to different areas of Fourier space in order to increase the bandwidth/resolution of the final image. The effective numerical aperture (NA) is the sum of the objective NA and illumination NA. Similar to the regular ptychography, phase is necessary to *coherently* combine all the intensity images. The first FPM algorithm adapted the sequential alternating projection algorithm in the regular ptychography. Shifted support constraints from the finite pupil size are enforced in the Fourier domain and the corresponding amplitude constraints (measured images) are applied in the real space, while letting the phase evolve as each image is stepped through sequentially.

Significant progress has been made to extend the various schemes in the regular ptychography to FP. The recovery of the aberration in the pupil function is demonstrated by adapting the probe recovery algorithm. FP mode multiplexing algorithms have been developed to allow reconstruction using much fewer images by simultaneously illuminating the object from multiple mutually incoherent illumination beams when capturing each image, as well as to recover spectrally multiplexed information. FP multi-slice algorithm has been developed to recover 3D object and improve spatial resolution in all three dimensions.

Further Reading

- Batey, D.J., Claus, D., Rodenburg, J.M., 2014. Information multiplexing in ptychography. *Ultramicroscopy* 138, 13–21.
- Chapman, H.N., Barty, A., Marchesini, S., *et al.*, 2006. High-resolution ab initio three-dimensional X-ray diffraction microscopy. *Journal of the Optical Society of America A* 23, 1179–1200.
- Clark, J.N., Huang, X., Harder, R., Robinson, I.K., 2012. High-resolution three-dimensional partially coherent diffraction imaging. *Nature Communications* 3, 993.
- Clark, J.N., Peele, A.G., 2011. Simultaneous sample and spatial coherence characterisation using diffractive imaging. *Applied Physics Letters* 99, 154103.
- Fienup, J.R., 1982. Phase retrieval algorithms: A comparison. *Applied Optics* 21, 2758–2769.
- Guigay, J.P., Langer, M., Boistel, R., Cloetens, P., 2007. Mixed transfer function and transport of intensity approach for phase retrieval in the Fresnel region. *Optics Letters* 32, 1617–1619.
- Jingshan, Z., Claus, R.A., Dauwels, J., Tian, L., Waller, L., 2014. Transport of intensity phase imaging by intensity spectrum fitting of exponentially spaced defocus planes. *Optics Express* 22, 10661–10674.
- Jingshan, Z., Tian, L., Dauwels, J., Waller, L., 2015. Partially coherent phase imaging with simultaneous source recovery. *Biomedical Optics Express* 6, 257–265.
- Maiden, A.M., Humphry, M.J., Rodenburg, J.M., 2012. Ptychographic transmission microscopy in three dimensions using a multi-slice approach. *Journal of the Optical Society of America A* 29, 1606–1614.
- Maiden, A.M., Rodenburg, J.M., 2009. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy* 109, 1256–1262.
- Marchesini, S., 2007. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments* 78, 011301.
- Nugent, K., Gureyev, T., Cookson, D., Paganin, D., Barnea, Z., 1996. Quantitative phase imaging using hard X-rays. *Physical Review Letters* 77, 2961–2964.
- Paxman, R.G., Schulz, T.J., Fienup, J.R., 1992. Joint estimation of object and aberrations by using phase diversity. *Journal of the Optical Society of America A* 9, 1072–1085.
- Pelz, P.M., Guizar-Sicairos, M., Thibault, P., *et al.*, 2014. On-the-fly scans for X-ray ptychography. *Applied Physics Letters* 105, 251101.
- Shechtman, Y., Eldar, Y.C., Cohen, O., *et al.*, 2015. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing magazine* 32, 87–109.
- Teague, M.R., 1983. Deterministic phase retrieval: A Green's function solution. *Journal of the Optical Society of America* 73, 1434–1441.
- Thibault, P., Dierolf, M., Bunk, O., Menzel, A., Pfeiffer, F., 2009. Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy* 109, 338–343.
- Thibault, P., Menzel, A., 2013. Reconstructing state mixtures from diffraction measurements. *Nature* 494, 68–71.
- Tian, L., Li, X., Ramchandran, K., Waller, L., 2014. Multiplexed coded illumination for Fourier ptychography with an LED array microscope. *Biomedical Optics Express* 5, 2376–2389.
- Tian, L., Waller, L., 2015. 3D intensity and phase imaging from light field measurements in an LED array microscope. *Optica* 2, 104–111.
- Wade, R., 1992. A brief look at imaging and contrast transfer. *Ultramicroscopy* 46, 145–156.
- Waller, L., Tian, L., Barbastathis, G., 2010. Transport of intensity phase-amplitude imaging with higher order intensity derivatives. *Optics Express* 18, 12552–12561.
- Whitehead, L.W., Williams, G.J., Quiney, H.M., *et al.*, 2009. Diffractive imaging using partially coherent X-rays. *Physical Review Letters* 103, 243902.
- Yang, C., Qian, J., Schirotzek, A., Maia, F., Marchesini, S., 2011. Iterative algorithms for Ptychographic phase retrieval. *arXiv* 1105, 5628.
- Yeh, L.-H., Dong, J., Zhong, J., *et al.*, 2015. Experimental robustness of Fourier ptychography phase retrieval algorithms. *Optics Express* 23, 33214–33240.
- Zheng, G., Horstmeyer, R., Yang, C., 2013. Wide-field, high-resolution Fourier ptychographic microscopy. *Nature Photonics* 7, 739–745.
- Zhong, J., Tian, L., Varna, P., Waller, L., 2016. Nonlinear optimization algorithm for partially coherent phase retrieval and source recovery. *IEEE Transactions on Computational Imaging* 2, 310–322.

Phase Space and Imaging

Markus Testorf, Thayer School of Engineering, Hanover, NH, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Phase-space optics (Torre, 2005) represents a framework to model optical systems and signal propagation. It introduces concepts of joint time-frequency transforms, well established for signal processing, to the domain of ray and wave optics. In contrast to many joint transforms specifically designed for signal processing applications, phase-space optics preserves a physical interpretation of the joint space – spatial frequency representation. In close analogy to the phase space of classical and quantum mechanics, the physics of optical signal propagation is encoded in the associated phase-space representation.

Optical system analysis and design usually avoid rigorous modeling of Maxwell's equations. Instead, convenient approximations are in use, which are tailored to the typical constraints of optical systems. These include ray optics and radiometry, Gaussian beam optics, Fourier optics of scalar complex amplitudes, as well as coherence theory. Phase-space optics combines most aspects of the aforementioned into a single signal representation. This is particularly true for paraxial signals and systems, where formal relationships or even the equivalence of models can be established rather straightforwardly.

This article surveys the application of phase-space optics to optical imaging. While phase-space optics may be developed as a consistent replacement of more common models of light propagation, it is generally more useful to consider it as a complementary framework, which can be used to find solutions to specific imaging related problems more easily. This refers in particular to problems involving the analysis of the space-bandwidth requirements of signals and systems. For phase imaging phase-space optics provides a unique perspective, which has motivated new signal recovery methods, namely phase-space tomography. The phase-space interpretation provides a unique platform to compare imaging modalities and to establish performance limits. Furthermore, the phase space of quasi-geometrical optics establishes highly intuitive interpretations for optical problems otherwise burdened with the overhead of mathematical formalism.

Elements of Phase Space Optics

Optical phase space may be introduced conveniently as the configuration space of generalized ray coordinates. This means, in each plane along the optical axis z , any optical ray is characterized by its transverse location x and its angle of propagation θ relative to the optical axis. Each pair of generalized coordinates corresponds to a single point in phase space. It is often convenient to associate the propagation direction of a ray with the spatial frequency $v = \sin(\theta)/\lambda$ of a coherent monochromatic plane wave propagating in the same direction, thereby associating the spatial frequency effectively with a bundle of parallel rays.

Consequently, the phase-space distribution of rays $W(x, v)$ obeys the rules of paraxial ABCD matrix optics (Saleh and Teich, 1991). This is, for a paraxial optical system, the input ray coordinates (x_{in}, v_{in}) are related to the output coordinates (x_{out}, v_{out}) by the system's ray transfer matrix,

$$\begin{pmatrix} x_{out} \\ v_{out} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_{in} \\ v_{in} \end{pmatrix} \quad (1)$$

with the shape of the phase-space distribution transforming accordingly,

$$W_{out}(x, v) = W_{in}(Dx - Bv, -Cx + Av) \quad (2)$$

If the input and the output plane of the system are located in a medium of the same refractive index, the determinate of the ray-transfer matrix is always equal to one, i.e., only three of the four matrix coefficients are independent.

For three-dimensional geometries, both x and v are two-dimensional vectors. The phase-space distribution $W(x, v)$ may be recognized as the radiance, i.e., the power directed along v and radiating from, or being received by a point x .

For partially coherent optical signals the concept of radiance may be generalized and the phase-space distribution is defined as the Fourier transform of the cross spectral density $J(x_1, x_2)$ (see chapters 1, 2 and 7 of Testorf *et al.*, 2009),

$$W(x, v) = \int_{-\infty}^{\infty} J(x + x'/2, x - x'/2) \exp(-i2\pi x'v) dx' \quad (3)$$

For the ideal case of incoherent signals, the phase space distribution is given by the irradiance distribution,

$$W(x, v) = I(x) \quad (4)$$

For coherent wavefronts, the mutual intensity factorizes and the phase-space distribution becomes the Wigner distribution function (WDF) of the complex amplitude $u(x)$,

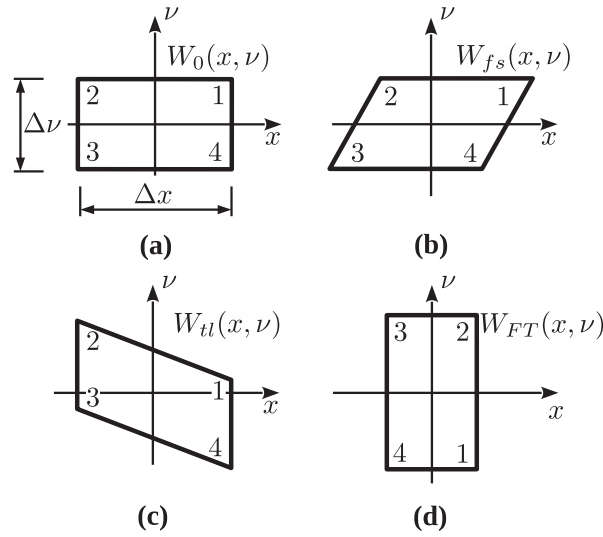


Fig. 1 Phase-space diagrams of ray transformations: (a) Phase-space diagram of a generic signal with essential support Δx and essential bandwidth $\Delta \nu$. (b) Phase-space distribution of the signal in (a) after Fresnel diffraction, (c) transmission through a thin paraxial lens, and (d) after Fourier transformation of the signal. Corners of the phase-space volume are numbered to depict the orientation of the transformed volume.

$$W(x, \nu) = \int_{-\infty}^{\infty} u\left(x + \frac{x'}{2}\right) u^*\left(x - \frac{x'}{2}\right) \exp(-i2\pi \nu x') dx' \quad (5)$$

As long as diffracting system apertures can be ignored, the ABCD transfer matrix equally applies to all definitions of $W(x, \nu)$. This implies, that the dynamics of coherent and partially coherent signals can be associated with geometrical transformations of phase space. For instance, free space propagation corresponds to a shear of phase space parallel to the spatial axis, and the transfer matrix of Fresnel diffraction reads

$$\mathbf{T}_f = \begin{pmatrix} 1 & \lambda z \\ 0 & 1 \end{pmatrix} \quad (6)$$

where z is the propagation distance. A thin parabolic lens of focal length f is given by

$$\mathbf{T}_l = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\lambda f} & 1 \end{pmatrix} \quad (7)$$

which is equivalent to a chirp modulation of the complex amplitude.

A Fourier transform step is equivalent to a rotation of phase space by 90° , and consequently, a 2-f system corresponds to a combination of 90° rotation and a magnification step. Rotation with arbitrary angles correspond to fractional Fourier transforms of the signal (Ozaktas *et al.*, 2001).

Fig. 1 depicts a set of the generic phase-space transformations in the form of phase-space diagrams. Signals are schematically represented by a finite phase-space volume with extensions equivalent to the essential signal support and bandwidth. This identifies the phase-space volume as a measure of the space-bandwidth product of the signal.

In the context of optical imaging it is important to highlight a number of properties of the WDF. The WDF is a real valued function. It is symmetric with respect to the two variables x and ν , and we find the alternative definition,

$$W(x, \nu) = \int_{-\infty}^{\infty} \tilde{u}\left(\nu + \frac{\nu'}{2}\right) \tilde{u}^*\left(\nu - \frac{\nu'}{2}\right) \exp(i2\pi \nu' x) d\nu' \quad (8)$$

with $\tilde{u}(\nu)$ being the Fourier transform of the signal

$$\tilde{u}(\nu) = \int_{-\infty}^{\infty} u(x) \exp(-i2\pi \nu x) dx \quad (9)$$

was used. The WDF is a complete signal representation in so far as the complex amplitude can be recovered from the WDF except for a constant phase factor. Intensity and power spectrum of the complex signal are calculated as projections of the WDF,

$$\int_{-\infty}^{\infty} W(x, \nu) d\nu = |u(x)|^2 \quad \text{and} \quad \int_{-\infty}^{\infty} W(x, \nu) dx = |\tilde{u}(\nu)|^2 \quad (10)$$

The WDF is a bilinear transformation of $u(x)$. This means, the sum of two signals will be transformed into the respective WDFs, with additional interference or cross terms added. For instance, a signal consisting of two plane waves, $u_{1,2}(x) = \exp(i2\pi v_1 x) + \exp(i2\pi v_2 x)$, is transformed into

$$W_{1,2}(x, v) = \delta(v - v_1) + \delta(v - v_2) + 2\cos[2\pi(v_1 - v_2)]\delta\left(v - \frac{v_1 + v_2}{2}\right) \quad (11)$$

The WDF contains a frequency, which is located halfway between the two plane wave frequencies, and which is modulated by a cosine function. The modulation frequency of the interference term reflects the spacing between the unmodulated frequency terms. This interference term effectively encodes the mutual coherence of the two plane waves.

The convolution of two signals,

$$u_c(x) = u_1(x) * u_2(x) \quad (12)$$

is transformed into a convolution of the respective WDFs with respect to x and a multiplication with respect to v ,

$$W_c(x, v) = W_1(x, v) *_x W_2(x, v) \quad (13)$$

Similarly, a multiplication of two signals, i.e., a convolution in the frequency domain, is transformed into a convolution of the WDFs with respect to v ,

$$W_M(x, v) = W_1(x, v) *_v W_2(x, v) \quad (14)$$

For slowly varying phase functions, $u(x) = \exp(i\phi(x))$, the WDF can be approximated by the instantaneous (or local) frequency of the signal,

$$W_u(x, v) \approx \delta\left(v - \frac{1}{2\pi} \frac{d\phi(x)}{dx}\right) \quad (15)$$

which may be interpreted as the quasi-geometrical approximation of WDF. For quadratic phase functions, Eq. (15) is the exact expression for the WDF, which is equivalent to an inclined line in phase space.

The WDF is related to the ambiguity function (AF) (chapter 2 of Testorf *et al.*, 2009) via a double Fourier transformation,

$$A(\bar{v}, \bar{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(x, v) \exp[-i2\pi(\bar{v}x - \bar{x}v)] dx dv \quad (16)$$

Many of the properties of the WDF carry over to the AF domain, namely the ray matrix transformation to relate input and output AF for quadratic phase systems.

Phase-Space Interpretation of Image Formation

In classical geometrical optics, perfect imaging is accomplished, if all rays originating from one point in the object plane intersect in a single point in the imaging plane. The spatial coordinates of object and image points are relate by a constant magnification factor, i.e.,

$$x_{out} = m x_{in} \quad (17)$$

Negative magnification refers to an inverted image.

In phase space, the set of all rays crossing through a single point in space define a line parallel to the frequency axis (Fig. 2). This implies that perfect imagery is accomplished with optical systems, which map vertical lines in phase space to vertical lines with a constant scaling factor in x .

This implies a ray transfer matrix with $A=m$ and $B=0$ and $D=1/m$. The forth parameter C can be chosen freely to satisfy the imaging condition. This matrix can be interpreted as a magnification step followed by a chirp modulation, i.e.,

$$\begin{pmatrix} m & 0 \\ C & 1/m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ C & 1 \end{pmatrix} \begin{pmatrix} m & 0 \\ 0 & 1/m \end{pmatrix} \quad (18)$$

While coefficient $A=m$ represents the spatial magnification factor of the imaging system, $D=1/m$ is the angular magnification factor.

This is exemplified by a single lens system, consisting of free space propagation over a distance a , a lens of focal length f , followed by propagation over a distance b , with the additional condition $1/f = 1/a + 1/b$. The associated transfer matrix reads

$$\mathbf{T}_{SLI} = \begin{pmatrix} 1 & \lambda b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/\lambda f & 1 \end{pmatrix} \begin{pmatrix} 1 & \lambda a \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{b}{a} & 0 \\ -\frac{1}{\lambda f} & -\frac{a}{b} \end{pmatrix} \quad (19)$$

The case $C=0$ describes an afocal imaging system. A parallel bundle of input rays is transformed into a parallel bundle of output rays. The 4-f system and the closely related Kepler telescope are particular examples of afocal systems.

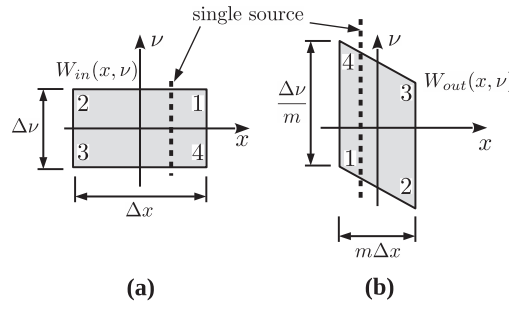


Fig. 2 Single lens imaging: (a) Phase-space diagram of input signal; (b) phase-space distribution in the image plans.

For $C \neq 0$ the chirp modulation corresponds to a shear of the phase-space distribution along the frequency axis (**Fig. 2(b)**). For the ideal case of a strictly incoherent object, with $W(x, \nu) = I(x)$, the phase space distribution is shift-invariant with respect to frequency and the vertical shear is not encoded in the phase space of the image plane. For the more general case of partially coherent or coherent signals, the shear of the phase space distribution corresponds to a modulation of the wavefront with a quadratic phase function.

For a wave optical analysis and finite lens apertures, the WDF of the aperture is convolved in ν with the signal's WDF in the aperture plane and further propagated to the output plane.

Imaging Aberrations

The close relationship between geometrical optics and phase space extends beyond the paraxial regime, and the phase-space formalism inherently is an integral part of geometrical optics theory (Wolf, 2004). This includes the analysis of optical aberrations.

One approach to understand imaging aberrations in the context of optical phase space is the interpretation of geometrical aberrations as deviations from the linear transformation, which connects input and output of the imaging systems. This means, the generalized coordinates of the output ray are no longer linear functions of the input coordinates expressed by a ray transfer matrix with constant coefficients. Instead, the matrix coefficients are functions of the input ray coordinates and the phase space distribution is distorted in both x and ν coordinate. As a consequence, the image, i.e., the projection of the phase space distribution along ν in the output plane of the system, is not longer a perfect copy of the object.

This nonlinear geometrical transform of phase space can be reversed, if the phase space distribution is known. This forms the basis for aberration correction with lightfield camera systems (see Section Lightfield Imaging).

The interpretation of aberrations as a geometrical distortion of phase space can be translated to the analysis of wavefront aberrations and their influence on the WDF of the optical signal. Wave aberrations may be interpreted as a phase-only modulation of the wavefront in the exit pupil of the imaging system. The phase $\phi(x)$ can be expressed as a power series and for a smooth aberration function the WDF of the modulating function can be approximated by the local frequency (Lohmann *et al.*, 1983),

$$W_{ab}(x, \nu) \approx \delta \left[\nu - \frac{1}{2\pi} \frac{\phi(x)}{dx} \right] \quad (20)$$

The aberrated phase space distribution is the result of a convolution of $W_{ab}(x, \nu)$ with the WDF of the ideal wavefront in ν in the plane of the exit pupil. This is identical with a local shift of the WDF along the frequency axis and proportional to the local frequency.

In the image plane, the WDF of the exit pupil appears essentially rotated due to the inherent Fourier relationship between exit pupil and image plane, and the aberrations are manifest as frequency dependent shifts along the x axis. Consequently, image points derived as the projection of the output WDF are blurred compared to the non-aberrated image point.

The Self-Imaging Condition

The phase space analysis of classical geometric optical imaging suggests generalizations imaging. The point-to-point mapping of optical rays and the associated mapping of vertical lines in phase space implies a mapping of the irradiance distribution, i.e., the imaging system is designed to reproduce the vertical projection of phase space. For incoherent imaging this reproduces all information available about the object.

For imaging applications based on coherent optical signals, it is equally possible to associate imaging with a recovery of the complex amplitude, equivalent to a full recovery of the object's phase space distribution. Furthermore, neither the recovery of the actual phase space distribution, nor its projection needs to be the result of a spatial point-to-point mapping of the signal. Instead, each image point may be assembled as an interference pattern of signals originating from more than one single object point.

The latter cannot be ensured to work for arbitrary objects, but rather limits imaging for a particular system to a subset of all possible objects.

One example of this type of imaging is the Talbot effect, also known as coherent self-imaging. In its most generic form, self-images is associated with Fresnel diffraction of periodic object waves. Images, i.e., replica of the complex amplitude of the object are observed at multiples of the Talbot distance,

$$z_T = \frac{2p^2}{\lambda} \quad (21)$$

where p is the transverse period of the object wave.

In phase space the coherent periodic signal corresponds to a WDF, which is periodic in x and discrete in frequency ν (Fig. 3(a)). In particular, for signals of the form,

$$u(x) = u(x - p) = \sum_{n=-\infty}^{\infty} a_n \exp(i2\pi x n/p) \quad (22)$$

the WDF reads,

$$W(x, \nu) = W(x - p, \nu) = \sum_{n, n'=-\infty}^{\infty} a_n a_{n'}^* \exp(i2\pi x(n - n')/p) \delta\left(\nu - \frac{n + n'}{2p}\right) \quad (23)$$

Due to cross-terms the spacing of the discrete frequencies is $\Delta\nu = 1/(2p)$.

Fresnel diffraction corresponds to a linear shear (Fig. 3(b)). The sheared WDF is identical to the initial distribution, when the lowest non-zero frequency component is shifted by p , or

$$\lambda z_T \frac{1}{2p} = p \quad (24)$$

which effectively provides a geometrical derivation of the Talbot distance (see chapter 10 of Testorf *et al.*, 2009).

The self-imaging principle can easily be generalized to account for image magnification, and other optical systems, which restore the WDF of particular types of signals (Testorf and Hennelly, 2016). Most interestingly, the phase-space interpretation highlights the duality between operation in the spatial and the spatial frequency domain. This means, for instance, the Talbot effect has a dual frequency counterpart, which is equivalent to modulating a discrete signal with a quadratic phase function. In phase-space, space and spatial frequency axis are merely two particular orientations and self-imaging can be further generalized to include discrete chips which are sheared along their axis of orientation in phase space.

Holographic Imaging

The phase-space formulation of holographic imaging provides an intuitive interpretation of critical properties, namely the space-bandwidth product requirements of the recording medium and conditions for a unique recovery of the wavefront during the reconstruction step (Lohmann *et al.*, 2004).

The phase-space analysis of the holographic principle can be demonstrated by assuming a wavefront in the recording plane composed of the complex amplitude of the object wave $u(x)$ and a plane off-axis reference wave. Fig. 4(a) schematically depicts the object wave O and the off-axis plane wave R in the recording plane. The reference wave corresponds to a line parallel to the x axis. The extensions in space Δx and frequency $\Delta\nu$ define the space-bandwidth product (SBP) of the object wave. The spatial frequency of the reference wave is ν_R . The phase-space diagram in Fig. 4 does not contain any cross-terms, since they are inconsequential for the interpretation of the holographic principle that follows.

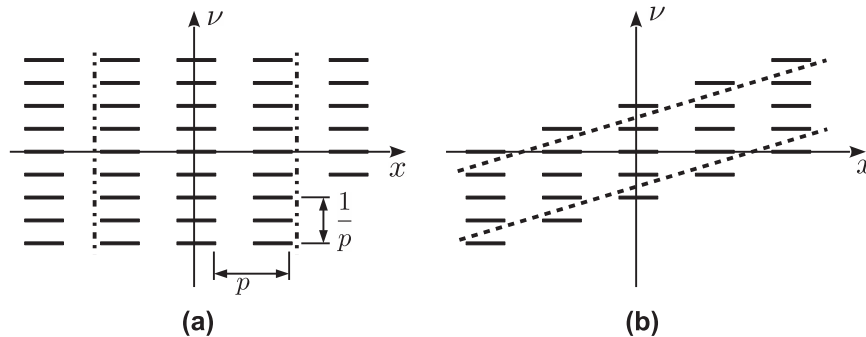


Fig. 3 Selfimaging in phase-space: (a) Phase-space diagram of periodic signal. The diagram depicts a finite number of periods and harmonics only. The vertical dashed lines mark arbitrary locations x ; (b) phase-space distribution after Fresnel diffraction over a single Talbot length.

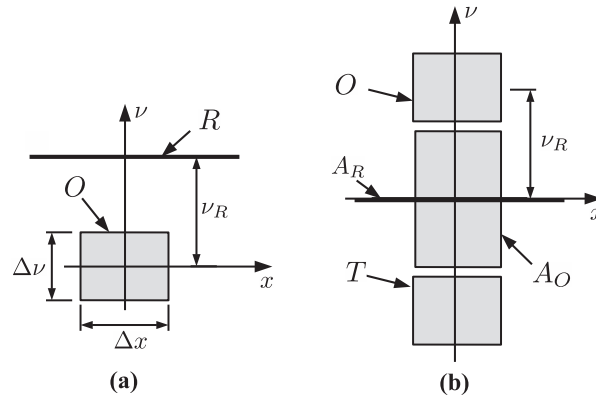


Fig. 4 Holography in phase space: (a) Phase-space diagram of the signal O and the plane reference wave R . (b) Phase-space distribution after the recording. The hologram is composed of the object O , the autocorrelation of object A_O and reference A_R , as well as the twin image T .

The holographic recording stores the squared magnitude of the complex amplitude, i.e., the transmission function of the recorded hologram reads

$$t(x) = |u(x) + \exp(i2\pi v_R x)|^2 = (1 + |u(x)|^2) + u^*(x)\exp(i2\pi v_R x) + u(x)\exp(-i2\pi v_R x) \quad (25)$$

with last term containing the object wave.

In phase space, the squared modulus of a signal is translated into the autocorrelation function of the WDF along v . The autocorrelation of object wave A_O and reference A_R are centered at the origin in phase space (Fig. 4(b)). Apart from the object wave O it is possible to observe the twin image T . All terms occupy separate volumes and do not overlap. This is necessary to recover and isolated the object term by illuminating the hologram with the reference wave. Estimating the bandwidth of the autocorrelation term to be twice the bandwidth of the object, the minimum frequency of the reference wave has to be

$$v_R \geq \frac{3}{2} \Delta v \quad (26)$$

The bandwidth of the hologram equals $2v_R + \Delta v$, which is at least four times the bandwidth of the hologram.

Fig. 4(b) can be interpreted as the phase-space diagram of image plane holography or Fourier plane holography. For the latter case, object wave and reference are recorded after a Fourier transform step equivalent to a 90° rotation of phase space. For Fresnel plane holography, the object wave is propagated in free space before it is recorded. The corresponding phase space operation is a horizontal shear, and it is possible to derive the SBP requirements of the recording medium from a geometric interpretation in phase space (Lohmann *et al.*, 2004).

Instead of analysing holography exclusively in the WDF domain, it is possible to use the AF (Situ and Sheridan, 2007). The AF of the complex amplitude may be interpreted as a generalized autocorrelation function. This means, the phase-space diagram of the AF of the complex amplitude resembles the diagram of the WDF of the recorded signal in Fig. 4(b). The actual transmission function of the hologram can be identified as the Fourier transform of the cross-section of the AF along its frequency axis.

Lightfield Imaging

The lightfield is closely related to the phase space of geometrical optics. Instead of parameterizing rays in each plane along the optical axis by their transverse location and propagation direction, the lightfield defines each ray by its spatial coordinates in two planes. While encoding the ray parameters in a different way, the lightfield and geometrical optical phase space contain the same information about the optical signal. Lightfield imaging is aimed at sampling the lightfield (or phase space) of the signal directly, rather than its projections.

This may be accomplished with a lightfield (or plenoptic) camera. The most generic implementation of the plenoptic camera is based on Lippmann's concept of integral photography. The detector array in the image plane of a camera system is replaced with a lenslet array and the detector, now located in the focal plane of the lens array (Fig. 5(a)). The aperture Δx of each lenslet defines the spatial resolution of the image as well as the spatial resolution with which phase space is sampled. The propagation direction of the ray is encoded as lateral displacement in the detector plane. Consequently, with the framework of geometrical optics, the angular resolution is determined by the pixel size Δx_θ , with an angular resolution $\Delta\theta \approx \Delta x_\theta/f$ and the resolution in frequency estimated as $\Delta v \approx \Delta x_\theta/(\lambda f)$.

This means, each detector pixel corresponds to a finite area in phase-space. The center of this finite volume in phase space indicates the position of the microlens. Each microlens projects the signal onto a subset of detector elements each in turn encoding a location along the frequency direction.

The phase-space diagram of the focal plane array is rendered in Fig. 5(b). Each ellipse schematically depicts the footprint of a single detector element. The recorded signal is the average of the signal's phase-space distribution over the volume covered by a

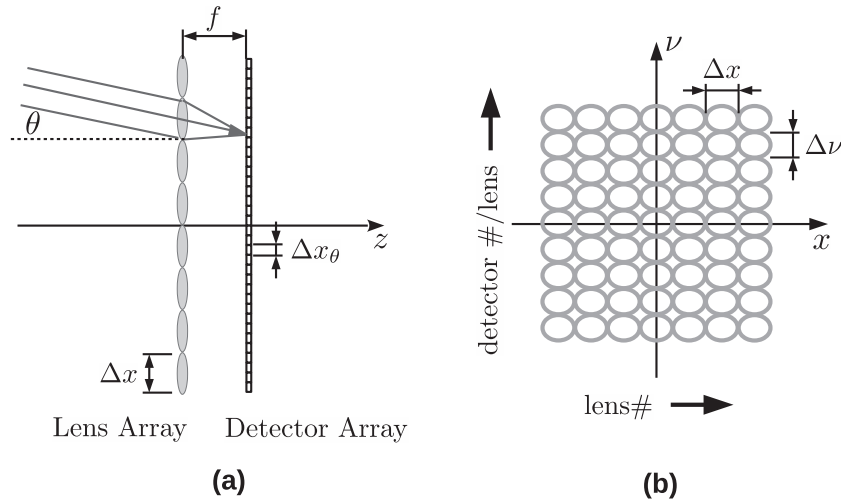


Fig. 5 Lightfield detector: (a) a focal plane array can be used to measure the radiance distribution. (b) The phase-space diagram of the focal plane array. Each ellipse corresponds to the phase-space volume recorded by a single detector element.

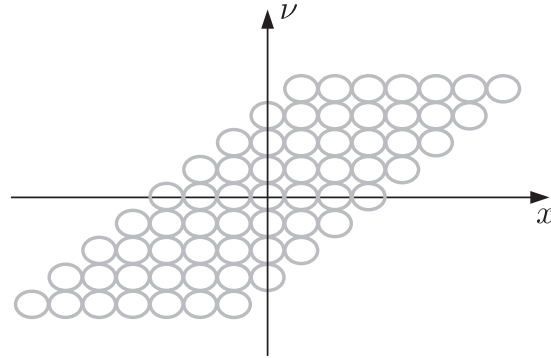


Fig. 6 Phase-space diagram of the focal plane array for correcting defocus: Each detector in phase space is associated with the location the in-focus phase-space distribution would assume in an out-of-focus plane. The in-focus image can be recovered by numerically reassigning the position of the detected phase-space locations before rendering the image from the frequency projection.

detector element. This means, as long as the phase-space distribution changes slowly on the scale of the detector's phase-space coverage, the focal plane array can be used as a phase-space detector without loss of important information. The number of detectors is equivalent to the SBP of the camera, i.e., comparing this with the SBP of the signal allows an estimation of how well the lightfield camera is recording the signal information.

Direct access to the phase-space distribution allows one to transform phase space before rendering an image by projecting along the frequency axis. This permits a change of perspective, refocusing, as well as the compensation of aberrations after recording a single lightfield. For instance, to correct defocus the signal associated with a detector is assigned to a point in phase space, which reflects the geometrical transform of phase space associated with free space propagation (Fig. 6). The defocus blur can be removed by reassigning the phase-space coordinates to the locations of the in-focus plane in Fig. 5(b), which is equivalent to rendering the image from projections at the shear angle.

The lightfield concept can also be used to design three-dimensional displays for autostereoscopic viewing. This can be understood as reversing the recording process, i.e., each pixel of the detector is replaced with a point source. This creates a virtual image, which is viewed through the lenslet array. Each of the point sources encodes one sample of the radiance distribution of the reconstructed optical signal.

The lightfield inherently assumes a geometrical optics model of imaging with a phase-space distribution that is changing slowly. In this context, the focal plane array provides sufficient information to determine the radiance distribution from the recorded samples. If diffraction effects have to be accounted for the finite phase-space volume of a single detector does not provide sufficient resolution of phase space and it is necessary to recover the actual WDF of the optical signal. In general, it is necessary to resort to a synthesis of the phase-space distribution from information acquired with multiple recordings in different diffraction planes.

Phase-Space Tomography and Phase Retrieval

Every measurement in the optical regime can be interpreted as a power reading written as a projection

$$P = |\langle u(x), h(x) \rangle|^2 \quad (27)$$

where the complex amplitude $u(x)$ is detected with an optical instrument $h(x)$. This formalism, for instance, describes the coupling of the coherent wavefront $u(x)$ incident on an optical fiber, with $h(x)$ describing one particular optical mode. Eq. (27) then refers to the power P coupled into the mode and transported to a square law detector. In phase space, this overlap integral becomes

$$P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_u(x, v) W_h(x, v) dx dv \quad (28)$$

where $W_u(x, v)$ and $W_h(x, v)$ are the normalized WDFs of the signal and the instrument, respectively.

The ideal generic detector may be modeled as a point detector, located at x_d , which receives optical power independent of the incident angle,

$$W_{pd}(x, v) = \delta(x - x_d) \quad (29)$$

The detected signal is the overlap integral,

$$P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_u(x, v) W_{pd}(x, v) dx dv = \int_{-\infty}^{\infty} W_u(x_d, v) dv \quad (30)$$

which is consistent with the projection theorem of the WDF in Eq. (10). The signal power as a function of x is obtained with a continuum of point detectors.

This also means, the focal plane array of the lightfield camera effectively provides samples of a low pass filtered WDF only. It is possible, however, to add additional measurements with different instruments $W_h(x, v)$ and recover the signal WDF from the set of all measurements.

In particular, it is possible to envision power recordings obtained after passing the wavefront through a set of fractional Fourier transform systems. The fractional Fourier transformation generalizes the Fourier transforms in that it implements rotations of phase space at arbitrary angles. The rotation angle is proportional to the order of the fractional Fourier transformation.

Measuring power after fractional Fourier transforming the signal corresponds to the recording of projections of the WDF of the original signal along the rotation angle of the fractional Fourier transform. This means, a set of measurements covering all projection angles can be used for tomographic reconstruction of the WDF. Invoking the projections slice theorem of tomography identifies the projections at each angle as one slice of the AF of the signal. From knowledge of the AF the WDF and the signal can be estimated via an inverse Fourier transform.

Fractional Fourier transforms are not necessarily required to obtain the projections for tomographic reconstruction. For instance, Fresnel diffraction patterns may be recorded instead. Fig. 7 demonstrates, how projections along vertical lines $\overline{P_1 P_2}$ of the rotated WDF can be found after a horizontal shear, equivalent to free space propagation. While mapping recordings in Fresnel diffraction planes requires numerical scaling of the recording and presents inherent difficulties to obtain the equivalent of rotation angles close to 90° , it is typically easier to move the detector array along the optical axis, rather than implementing a large set of fraction Fourier transform systems.

This approach, known as phase-space tomography (Smithey *et al.*, 1993), is remarkable in that it provides a linear inverse algorithm for phase retrieval. This advantage is partially counterbalanced by the number of measurements required for the reconstruction of the WDF. For coherent wavefronts, a signal which is represented by N samples, is transformed into nominally $4N^2$ samples of the WDF, which includes the interference terms. Without additional prior knowledge, the linear retrieval then requires the same order of measurements to uniquely recover the WDF.

For partially coherent optical signals the number of samples to represent the phase-space distribution is equivalent to the number of sampling points to describe the mutual intensity. This means, in this case, phase-space tomography fully represents the complexity of the inverse problem.

For the recovery of two-dimensional images the projection slice theorem calls for anamorphic fractional Fourier transform systems, where the order of the fractional Fourier transform system can be changed independently for the two transverse coordinates. This means, a four-dimensional phase space has to be recovered to determine a two dimensional image. For coherent imaging, however, it is possible to reduce the experimental burden and recover the 2-D image as a set of 1-D slices.

The phase-space interpretation of tomographic signal recovery can be extended to other forms of phase-retrieval, most notably deterministic phase retrieval methods (Semichaevsky and Testorf, 2004) and pytochraphic imaging (Lubk and Röder, 2015). For coherent signals, it is possible to reduce the number of power measurements by surrendering the linearity of the reconstruction algorithm. For instance, deterministic phase retrieval from two closely located Fresnel diffraction planes are inherently based on the recovery the WDF from its local first order moments. This provides access to the AF and it first derivative along the frequency axis in turn providing magnitude and phase of the signal. The phase-space interpretation then provides a way to compare underlying principles of various phase retrieval methods.

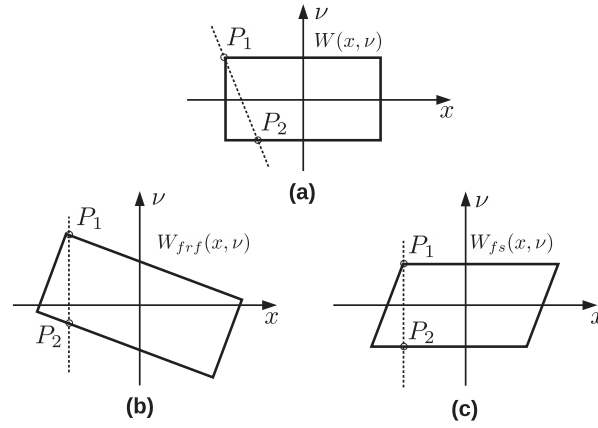


Fig. 7 Equivalence of Fresnel diffraction and fractional Fourier transform. (a) Phase-space diagram of the signal. (b) Signal after fractional Fourier transform. The line $\overline{P_1 P_2}$ describes the projection associated with the signal detected in a fraction Fourier plane. (c) The phase-space after Fresnel diffraction. The horizontal shear can be selected to provide the same vertical projections as observed for the rotated phase-space distribution.

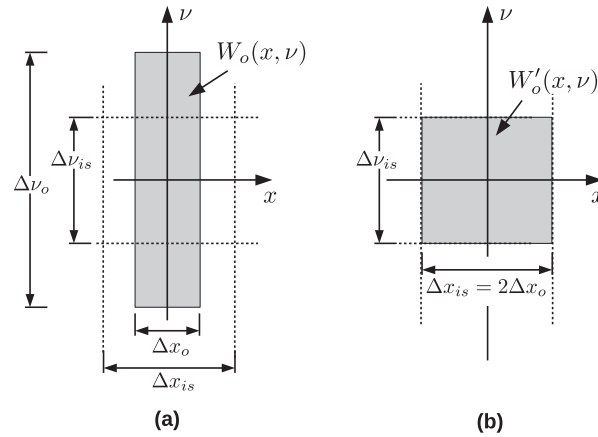


Fig. 8 Phase-space diagram of superresolution imaging: (a) The object bandwidth $\Delta\nu_o$ is twice the bandwidth supported by the optical system $\Delta\nu_{is}$. (b) Adapting the phase-space volume to the constraints of the imaging system allows transmission through the optical system without loss of information.

Superresolution Imaging

Superresolution imaging describes mathematical algorithms and imaging systems, which provide optical resolution superior to the nominal value associated with the size of the clear aperture of the system (Zalevsky and Mendlovic, 2004).

The mechanisms to achieve superresolution may be understood as a consequence of Shannon's information capacity theorem. Interpreting the imaging system as a continuous Gaussian channel of bandwidth $\Delta\nu_{is}$, field of view Δx_{is} and signal-to-noise ratio SNR , we expect to resolve

$$C = \Delta x_{is} \Delta \nu_{is} \log_2(1 + SNR) \quad (31)$$

bits of information from the image. If input signals to the system do not occupy the entire information capacity in terms of spatial support, or signal-to-noise ratio, it is possible to reassign information, which is encoded by the signal bandwidth, to either the signal extension or the SNR and transmit signals with a bandwidth exceeding the supported bandwidth of the system.

The phase-space interpretation is useful, in particular, to explore the trade-off between signal extension and bandwidth (Zalevsky et al., 2000). The generic operation of the this type of superresolution system is shown in Fig. 8. The input signal has bandwidth $\Delta\nu_o$ and a lateral extension Δx_o . The imaging systems only supports half the signal's bandwidth, i.e $\Delta\nu_{is} = \Delta\nu_o/2$, but the field of view Δx_{is} is twice the object extension. Then, a coding step can be implemented to transform the signal into the transmitted signal with half the original bandwidth, but twice the original signal extension. This type of bandwidth compression may be accomplished with quadratic phase systems which leave the phase-space volume in tact.

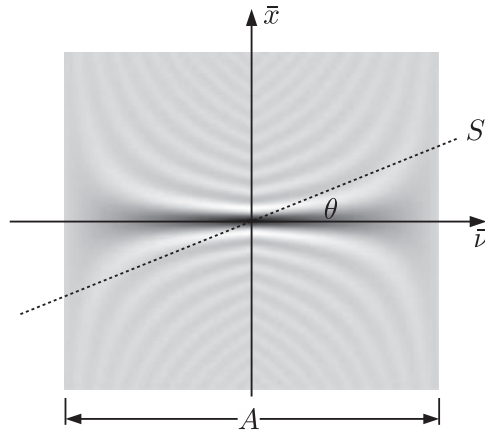


Fig. 9 Ambiguity function of the diffraction limited rectangular aperture: Cross-sections C correspond to the OTF in out-of-focus planes. The angle θ is related to the degree of defocus.

After passing the modified signal through the imaging system a decoding step is required to restore the original signal. Since the bandwidth of the transmitted signal, and consequently its resolution, is twice the nominal bandwidth of the system, this type of system effectively achieves superresolution imaging.

In order to recover a superresolved image of the object, the phase-space distribution at the output plane should not overlap with any other signal component.

Coding and decoding step may fundamentally differ depending on the imaging application. For instance, structured illumination sources may be used to move high frequencies of the object to the transmitted baseband, while numerical computation is used to assemble the image.

Pupil Design for Extended Depth of Field

The phase-space interpretation of wave aberrations can be combined with the concept of phase-space tomography to design imaging systems with increased depth of focus (see chapter 5 of [Testorf et al., 2009](#)). Conceptually, a defocus error can be described as the response of a linear system (Fresnel diffraction), which would suggest a removal of the defocus error by applying an inverse filter. However, the optical transfer function (OTF) for out-of-focus planes contains zeros, and the application of an inverse filter will typically result in images of low quality.

Alternatively, it is possible to design a lens function, for which the OTF never assumes the optimum shape of a diffraction limited imaging system, but which contains no zeros and a shape only depending weakly on the position of the object along the optical axis. This means, it is possible to apply an inverse filter numerically to the recorded low quality image to approximate more closely a diffraction limited OTF over a range of object planes superior to the depth of field predicted by the f-number of the lens. Originally, a cubic phase plate was suggested as a filter function to achieve this effect ([Dowski and Cathey, 1995](#)). Other families of filter functions were shown to provide a systems response suitable to extending the depth of field.

Recording defocused objects corresponds to the projections of the in-focus WDF after shearing relative to the x axis. This is equivalent to projections of the WDF at oblique angles (compare Section Phase-Space Tomography and Phase Retrieval). The Fourier transform of these projections are equivalent to slices of the in-focus AF at different angles. If the object is a perfect point source, the image represents the point-spread function of the system. The corresponding AF of the in-focus image, in [Fig. 9](#), displays the point-spread function along the $\bar{x}$ axis and the in-focus OTF along the $\bar{v}$ axis. Slices S represent the OTF at a plane which is out-of focus by a distance

$$\Delta z = \frac{\tan(\theta)}{\lambda} \quad (32)$$

For phase-space tomography each slice corresponds to information about the AF, which is used to reconstruct the signal. Here, the same concept can be used to evaluate pupil functions. The AF represents a simultaneous display of all out-of-focus OTFs for a particular object plane. A filter function suitable for extending the depth-of-field corresponds to an AF, with cross-sections not changing over the range of angles for which the defocus error will be removed.

References

- Dowski, E.R., Cathey, W.T., 1995. Extended depth of field through wave-front coding. *Appl. Opt.* 34 (11), 1859–1866.
 Lohmann, A.W., Ojeda-Castañeda, J., Streibl, N., 1983. The influence of wave aberrations on the Wigner distribution. *Optica Applicata* 13 (4), 465–471.

- Lohmann, A.W., Testorf, M., Ojeda-Castañeda, J., 2004. The art and science of holography: A tribute to Emmett Leith and Yuri Denisjuk. In: Caulfield, H.J. (Ed.), *Holography and the Wigner Function*. SPIE Press, pp. 129–144. (chapter **Holography and the Wigner Function**).
- Lubk, A., Röder, F., 2015. Phase-space foundations of electron holography. *Phys. Rev. A* 92, 033844.
- Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. *The Fractional Fourier Transform*. Chichester: John Wiley & Sons.
- Saleh, B.E.A., Teich, M.C., 1991. *Fundamentals of Photonics*. New York: John Wiley & Sons.
- Semichaevsky, A., Testorf, M., 2004. Phase-space interpretation of deterministic phase retrieval. *J. Opt. Soc. Am. A* 21 (11), 2173–2179.
- Situ, G., Sheridan, J.T., 2007. Holography: An interpretation from the phase-space point of view. *Opt. Lett.* 32 (24), 3492–3494.
- Smithey, D.T., Beck, M., Raymer, M.G., Faridani, A., 1993. Measurement of the Wigner distribution and the density matrix of light mode using optical homodyne tomography: Application to squeezed states and the vacuum. *Phys. Rev. Lett.* 70, 1244–1247.
- Testorf, M., Hennelly, B., 2016. Self-imaging and discrete paraxial optics. In: Healy, J.J., Kutay, M.A., Ozaktas, H.M., Sheridan, J.T. (Eds.), *Linear Canonical Transforms – Theory and Applications*. Springer, pp. 257–292.
- Testorf, M., Hennelly, B., Ojeda-Castañeda, J., 2009. *Phase-Space Optics: Fundamentals and Applications*. McGraw-Hill.
- Torre, A., 2005. *Linear Ray and Wave Optics in Phase Space*. Amsterdam: Elsevier.
- Wolf, K.B., 2004. *Geometric Optics on Phase Space*. Springer.
- Zalevsky, Z., Mendlovic, D., 2004. *Optical Superresolution*. Springer.
- Zalevsky, Z., Mendlovic, D., Lohmann, A.W., 2000. Understanding superresolution in Wigner space. *J. Opt. Soc. Am. A* 17 (12), 2422–2430.

Information Theory in Imaging

FO Huck and CL Fales, NASA Langley Research Center, Hampton, VA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Information theory is an intuitively attractive mathematical approach for combining observation with communication, or sensing with processing. So far, however, this approach has remained largely in the realm of signal processing and transmission, as first formalized by Shannon and Wiener over 50 years ago. Recent extensions of this approach to imaging have mostly addressed the interpretation of images, without explicitly accounting for the critical limiting factors that constrain image gathering and restoration. Yet, to be useful in the characterization of imaging, information theory must account for these constraints and lead to a close correlation between predicted and actual performance. The characterization that is outlined here traces the rate of information from the scene to the observer; and it pairs this rate, first, with the theoretical minimum data rate to assess the efficiency with which information can be transmitted, and, second, with the maximum-realizable fidelity to assess the quality with which images can be restored. Further details, especially for the mathematical development of figures of merit and for the design of imaging systems, can be found in the Further Reading.

Model of Imaging

Description

Modern imaging systems usually have the following three stages, as depicted in Fig. 1:

1. *Image gathering*, to capture the radiance field that is either reflected or emitted by the scene and transform this field into a digital signal;
2. *Signal coding*, to encode the acquired signal for the efficient transmission and/or storage of data and decode the received signal for the restoration of images. The encoding may be either lossless (without loss of information) or lossy (with some loss of information), if a higher compression is required;
3. *Image restoration*, to produce a digital representation of the scene and transform this representation into a continuous image for the observer.

Human vision may be included in this model as a fourth stage, to characterize imaging systems in terms of the information rate that the eye of the observer conveys from the displayed image to the higher levels of the brain.

These stages are similar to each other in one respect: each contains one or more transfer functions, either continuous or discrete, followed by one or more sources of noise, also either continuous or discrete. The continuous transfer function of the image-gathering device (and of the human eye) is followed by sampling that transforms the captured radiance field into a discrete signal with analog magnitudes. In most devices the photodetection mechanism conveys this signal serially into an analog-to-digital (A/D) converter to produce a digital signal. However, analog signal processing in a parallel structure, akin to that in human vision, has many advantages that are increasingly emulated by neural networks.

This model of imaging differs from the classical model of communication that Shannon and Wiener addressed in two fundamental ways:

1. *The continuous-to-discrete transformation in image gathering.* Whereas communication is constrained critically only by bandwidth and noise, image gathering is constrained also by the compromise between blurring and aliasing, due to limitations in the response of optical apertures and in the sampling density of photodetection mechanisms. This additional constraint requires the rigorous treatment of insufficiently sampled signals throughout the imaging system.
2. *The sequence of image gathering followed by signal coding.* Whereas the characterization of communication addresses the perturbations that occur at the encoder and decoder or during transmission, the characterization of imaging systems must also address the perturbations in the image-gathering process prior to coding. These additional perturbations require a clear

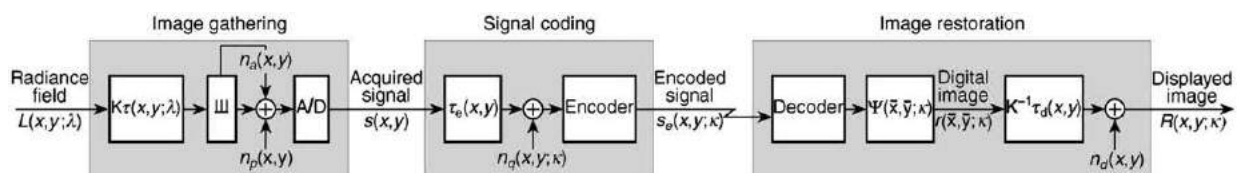


Fig. 1 Model of digital imaging systems.

distinction between the information rate of the encoded signal and the associated theoretical minimum data rate or, more generally, between information and entropy.

The mathematical model of imaging that is outlined here accounts for the appropriate deterministic and stochastic (random) properties of natural scenes and for the critical limiting factors that constrain the performance of imaging systems. The corresponding formulations of the figures of merit are sufficiently general to account for a wide range of radiance field properties, photodetection mechanisms, and signal processing structures. However, the illustrations emphasize: (a) the salient properties of natural scenes that dominate the captured radiance field in the visible and near-infrared region under normal daylight conditions; and (b) the photodetector arrays that are used most frequently as the photodetection mechanism of image-gathering devices.

Radiance Field

The radiance field $L(x, y; \lambda)$, reflected by natural scenes, may be modeled approximately as

$$L(x, y; \lambda) = \frac{1}{\pi} I(\lambda) \rho(\lambda) \rho(x, y) \quad (1)$$

where $I(\lambda)$ and $\rho(\lambda)$, respectively, represent the spectral irradiance and reflectance properties as a function of the wavelength λ , and $\rho(x, y)$ represents the random reflectance properties as a function of the spatial coordinates (x, y) . Fig. 2 illustrates a model of $\rho(x, y)$ for which the power spectral density (PSD) $\hat{\Phi}_\rho(v, \omega)$ corresponds closely to that typical of natural scenes, as given by

$$\hat{\Phi}_\rho(v, \omega) = \frac{2\pi\mu_\rho^2\sigma_\rho^2}{[1 + (2\pi\mu_\rho\zeta)^2]^{3/2}} \quad (2)$$

where $\zeta^2 = v^2 + \omega^2$ and (v, ω) are the spatial frequency coordinates of the Fourier-transform domain (see Fig. 3). This model consists of random polygons, whose boundaries are distributed according to Poisson probability with a mean separation of μ_ρ , and whose magnitudes are distributed according to independent zero-mean Gaussian statistics of variance σ_ρ^2 . The constraints associated with Eq. (1) are satisfied by letting the mean reflectance $\bar{\rho} = 0.5$ and the standard deviation $\sigma_\rho \leq 0.2$.

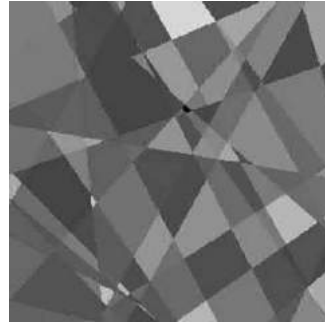


Fig. 2 Model of reflectance $\rho(x, y)$.

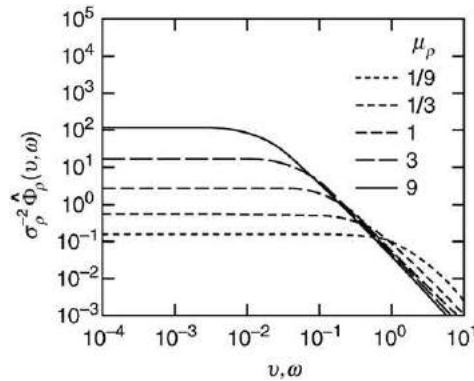


Fig. 3 Normalized PSD of the reflectance $\rho(x, y)$.

Image Gathering

Fig. 4 depicts the basic optical geometry of an image-gathering device, in which the objective lens forms an image of the captured radiance field across the photodetector array, and the array transforms this image into a discrete signal. The sampling lattice (X, Y) , which determines the resolution of this device, is formed in angular units by the separation (X_p, Y_p) of the photodetector-array apertures and the distance ℓ_p as given for one of the dimensions by

$$X = 2 \tan^{-1} \frac{X_p}{2\ell_p} \approx \frac{X_p}{\ell_p} \text{ rad} = \frac{57.3 X_p}{\ell_p} \text{ deg} \quad (3)$$

Normally, the sampling lattice is square, as shown in Fig. 5, so that $X_p = Y_p$. Then the corresponding area of the photodetector apertures is $A_p = \gamma_p^2$, and the solid angle Ω of the instantaneous field of view (IFOV) is $\Omega \approx A_p / \ell_p^2$ ster.

The sampling lattice projected onto the scene at the distance ℓ_o has the dimensions $(X_o, Y_o) \approx (X\ell_o, Y\ell_o)$. Because these dimensions represent the finest spatial detail that normally can be restored, it is convenient to measure the mean spatial detail of the scene relative to the sampling lattice, as given by $\mu = \mu_p(X_o, Y_o)^{-1/2}$, which reduces to $\mu = \mu_p X_o^{-1}$ for $X = Y$. Hence, $\mu = 1$ indicates that the sampling interval is equal to the mean spatial detail.

The image-gathering process that transforms the continuous radiance field $L(x, y; \lambda)$ into the discrete signal $s(\mathbf{x}, \mathbf{y})$ may be modeled by the expression

$$s(\mathbf{x}, \mathbf{y}) = K \rho(\mathbf{x}, \mathbf{y}) * \tau(\mathbf{x}, \mathbf{y}) + n_p(\mathbf{x}, \mathbf{y}) \quad (4)$$

where the discrete coordinates $(\mathbf{x}, \mathbf{y}) = (mX, nY)$, and the symbol $*$ denotes continuous spatial convolution. This model accounts, in the simplest form, for the reflectance-to-signal conversion gain K , the reflectance $\rho(x, y)$ of the scene, the spatial response (SR) $\tau(x, y)$ of the image-gathering device, the continuous-to-discrete transformation by the photodetector-array lattice, and the discrete noise $n_p(\mathbf{x}, \mathbf{y})$ of the photodetectors. The discrete coordinates are defined by the sampling function

$$\mathbb{I} \equiv \mathbb{I}(\mathbf{x}, \mathbf{y}) = XY \sum_{\mathbf{x}, \mathbf{y}} \delta(\mathbf{x} - \mathbf{x}', \mathbf{y} - \mathbf{y}') \quad (5)$$

where $\delta(x, y)$ is the Dirac delta function; and the gain is given by

$$K = \frac{1}{\pi} A_\ell \Omega \int I(\lambda) \rho(\lambda) r(\lambda) d\lambda \quad (6)$$

where $A_\ell = \pi D^2/4$ is the area of the objective lens with diameter D , and $r(\lambda)$ is the responsivity of the photodetectors. This gain relates the magnitude of the signal $s(\mathbf{x}, \mathbf{y})$ directly to that of the reflectance $\rho(x, y)$; and, together with the variances σ_p^2 and σ_r^2 of the reflectance and the photodetector noise, respectively, it also determines the rms-signal-to-rms-noise ratio (SNR) $K\sigma_r/\sigma_p$.

The DFT of $s(\mathbf{x}, \mathbf{y})$ is the periodic spectrum $\tilde{s}(v, \omega)$ given by

$$\tilde{s}(v, \omega) = K \hat{\rho}(v, \omega) \hat{\tau}(v, \omega) * \mathbb{I} + \tilde{n}_p(v, \omega) \quad (7)$$

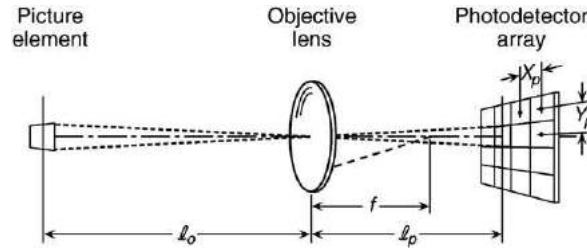


Fig. 4 Optical geometry of image-gathering device.

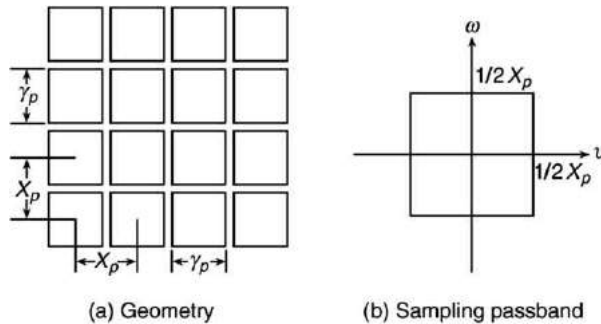


Fig. 5 Photodetector-array geometry and sampling passband.

where $\hat{\rho}(v, \omega)$ is the spectrum of the reflectance, $\hat{\tau}(v, \omega)$ is the spatial frequency response (SFR) of the image-gathering device, and $\hat{n}_p(v, \omega)$ is the spectrum of the photodetector noise. The symbol $\hat{\mathbb{I}}$ is the Fourier transform of the sampling function $\mathbb{I}$. It accounts for the sidebands generated by the sampling process and is given by

$$\hat{\mathbb{I}} \equiv \hat{\mathbb{I}}(v, \omega) = \sum_{v, \omega} \delta(v - v, \omega - \omega) \quad (8)$$

where $(v, \omega) = (m/X, n/Y)$. The corresponding sampling passband is

$$\hat{B} = \left\{ (v, \omega); |v| < \frac{1}{2X}, |\omega| < \frac{1}{2Y} \right\} \quad (9)$$

The above expression for $\hat{s}(v, \omega)$ can be rewritten more conveniently as

$$\hat{s}(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) + \hat{n}(v, \omega) \quad (10)$$

where

$$\hat{n}(v, \omega) = \hat{n}_a(v, \omega) + \hat{n}_p(v, \omega) \quad (11)$$

is the spectrum of the total noise in the acquired signal. The first term $\hat{n}_a(v, \omega)$ represents the aliased signal components caused by insufficient sampling, as given by

$$\hat{n}_a(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) * \hat{\mathbb{I}}_s \quad (12)$$

where

$$\hat{\mathbb{I}}_s \equiv \hat{\mathbb{I}}_s(v, \omega) = \sum_{v, \omega \neq (0,0)} \delta(v - v, \omega - \omega) \quad (13)$$

If, for spatial coordinates at either the scene or the photodetector array, the unit of (x, y) is meters, then the unit of (v, ω) is cycles m^{-1} ; and if, for angular coordinates centered at the objective lens, the unit of (x, y) is radians, then the unit of (v, ω) is cycles rad^{-1} . Either way, it is convenient to normalize the sampling intervals to unity (i.e., $X=Y=1$). Then the area of the sampling passband is $|\hat{B}| = 1$, the unit of (x, y) is samples, and the unit of (v, ω) is cycles $sample^{-1}$.

The SFR $\hat{\tau}(v, \omega)$ of the image-gathering device is

$$\hat{\tau}(v, \omega) = \hat{\tau}_e(v, \omega)\hat{\tau}_p(v, \omega) \quad (14)$$

where $\hat{\tau}_e(v, \omega)$ and $\hat{\tau}_p(v, \omega)$ are the SFRs of the objective lens and photodetector apertures, respectively. This SFR may be modeled approximately by the circularly symmetric Gaussian shape

$$\hat{\tau}(v, \omega) = \exp[-(\zeta/\zeta_c)^2] \quad (15)$$

where, as shown in Fig. 6, the optical-response index ζ_c controls the trade-off between the blurring that is caused by the decrease of the SFR inside the sampling passband and the aliasing that is caused by the extension of the SFR beyond this passband.

Signal Coding

Signal coding may be divided into two stages, as depicted in Fig. 1. The first stage, which is represented by the digital operator $\tau_c(x, y)$ and the quantization noise $n_q(x, y; \kappa)$, accounts for manipulations of the acquired signal that seek to extract those features of the captured radiance field considered to be perceptually most significant, which usually entails some loss of information. The second stage, which is represented by the encoder, accounts for manipulations that seek to reduce redundancies in the acquired

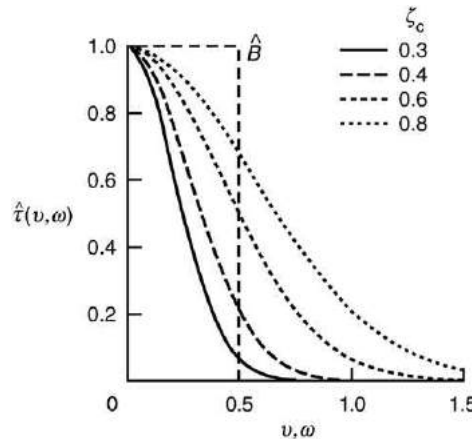


Fig. 6 SFRs $\hat{\tau}(v, \omega)$ of the image-gathering device relative to the sampling passband $\hat{B}$ for unit sampling intervals (i.e., $X=Y=1$).

signal, without loss of information. The resultant encoded signal $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ may be expressed in the same form as Eq. (4) for the acquired signal $s(\mathbf{x}, \mathbf{y})$ as

$$s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) = s(\mathbf{x}, \mathbf{y}) \otimes \tau_e(\mathbf{x}, \mathbf{y}) + n_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) \quad (16)$$

where the symbol $\otimes$, denotes discrete spatial convolution, and $n_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ is the quantization noise. If the data transmission and/or storage are error-free, as is assumed here, then the signal that leaves the decoder remains identical to the signal that enters the encoder.

The DFT of $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ is the periodic spectrum $\tilde{s}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ given in the same form as Eq. (10) by

$$\tilde{s}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = K \hat{\rho}(\mathbf{v}, \boldsymbol{\omega}) \hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega}) + \hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (17)$$

where

$$\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega}) = \hat{\tau}(\mathbf{v}, \boldsymbol{\omega}) \hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega}) \quad (18)$$

is the throughput SFR from image gathering to transmission, and

$$\hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{n}(\mathbf{v}, \boldsymbol{\omega}) \hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega}) + \tilde{n}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (19)$$

is the accumulated noise in the encoded signal. For $\log \boldsymbol{\kappa}$ -bit quantization, where $\log \boldsymbol{\kappa}$ is the number of uniformly spaced quantization levels, $\tilde{n}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ may be treated (for ≥ 4) as independent Gaussian noise with the PSD

$$\tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \sigma_q^2 = (\sigma_{s_e}/\boldsymbol{\kappa})^2 \quad (20)$$

where $\sigma_{s_e}^2$ is the variance given approximately by

$$\sigma_{s_e}^2 \approx K^2 \int_B \int \hat{\Phi}_\rho(\mathbf{v}, \boldsymbol{\omega}) |\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega})|^2 d\mathbf{v} d\boldsymbol{\omega} \quad (21)$$

Image Restoration

Image restoration, as depicted in Fig. 1, transforms the decoded signal $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ into the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ as modeled by the expression

$$r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa}) = \overline{\overline{\overline{s_e}}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) \otimes \Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa}) \quad (22)$$

The operation indicated by $\overline{\overline{\overline{s_e}}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ (called zero padding) prepares the decoded signal for image restoration with interpolation, and the operator $\Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ produces the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ on the interpolation lattice $\overline{\overline{\overline{\cdot}}}$ that is Z times denser than the sampling lattice $\overline{\overline{\cdot}}$. If the interpolation in the image restoration is sufficiently dense (normally, for $Z \geq 4$), then the blurring and raster effects of the image-display process are suppressed entirely. This condition leads not only to the best realizable image quality but also to the simplest expression for the continuous displayed image, as given in the same form as Eqs. (10) and (17) by

$$\hat{R}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\rho}(\mathbf{v}, \boldsymbol{\omega}) \hat{\Gamma}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (23)$$

where

$$\hat{\Gamma}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\tau}(\mathbf{v}, \boldsymbol{\omega}) \hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega}) \hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (24)$$

is the throughput SFR from image gathering to display, and

$$\hat{n}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = K^{-1} \hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{n}_{ar}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_{pr}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_{qr}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (25)$$

is the accumulated noise in the restored image. This noise may be expressed explicitly by substituting the noise components given above. The noise due to the granularity in image-display mediums is not accounted for here.

The Wiener filter, that minimizes the mean-square error due to the perturbations in image gathering and coding, is given by

$$\hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \frac{\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega}) \hat{\Gamma}_e^*(\mathbf{v}, \boldsymbol{\omega})}{\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega}) |\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega})|^2 + \hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})} \quad (26)$$

where $\hat{\Phi}'(\mathbf{v}, \boldsymbol{\omega}) = \sigma_\rho^{-2} \hat{\Phi}_\rho(\mathbf{v}, \boldsymbol{\omega})$ is the normalized PSD of the radiance field and

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = (K\sigma_\rho)^{-2} [\hat{\Phi}_n(\mathbf{v}, \boldsymbol{\omega}) |\tilde{\tau}_e(\mathbf{v}, \boldsymbol{\omega})|^2 + \tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})]$$

is the normalized PSD of the accumulated noise in the encoded signal. If the photodetector noise is white so that the PSD $\tilde{\Phi}_p(\mathbf{v}, \boldsymbol{\omega})$ is equal to its variance σ_p^2 and the PSD $\tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ is equal to its variance σ_q^2 , then the PSD of the noise simplifies to

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}) + \hat{\Phi}'_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (27)$$

where

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}) = [\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega}) \hat{\tau}(\mathbf{v}, \boldsymbol{\omega})]^2 * \overline{\overline{\overline{\delta}}} + (K\sigma_\rho/\sigma_\rho)^{-2} |\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega})|^2 \quad (28)$$

and

$$\hat{\Phi}'_q(v, \omega; \kappa) = (K\sigma_\rho/\sigma_{s_e})^{-2}\kappa^{-2} \quad (29)$$

These simplifying conditions permit the Wiener filter given above and the figures of merit given below to be expressed as a function of the SNR $K\sigma_\rho/\sigma_{s_e}$, which can be specified without accounting explicitly for either the gain K or the variances σ_ρ^2 and $\sigma_{s_e}^2$.

Figures of Merit

Information Rate $\mathcal{H}$

The information rate $\mathcal{H}$ (in bits sample⁻¹) of the encoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ is defined by

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \hat{\rho}(v, \omega) : (v, \omega) \in \hat{B}] \quad (30)$$

where the first term, $\mathcal{E}[\cdot]$ represents the entropy of $\tilde{s}_e(v, \omega; \kappa)$ and the second term, $\mathcal{E}[\cdot | \cdot]$ represents the conditional entropy of this signal when the reflectance spectrum $\hat{\rho}(v, \omega)$ within the sampling passband $\hat{B}$ is known. For the assumptions that have been made about image gathering and coding, the conditional entropy becomes the entropy of the accumulated noise $\hat{n}_e(v, \omega; \kappa)$, so that $\mathcal{H}$ simplifies to

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\hat{n}_e(v, \omega; \kappa) : (v, \omega) \in \hat{B}] = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\hat{\Phi}'_{n_e}(v, \omega; \kappa)} \right] dv d\omega \quad (31)$$

The theoretical upper bound of $\mathcal{H}$ is Shannon's channel capacity $\mathcal{C}$ for a bandwidth-limited system, with white photodetector noise and an average power limitation (here the variance σ_ρ^2), as given by

$$\mathcal{C} = \frac{1}{2} \log[1 + (K\sigma_\rho/\sigma_\rho)^2] \quad (32)$$

$\mathcal{H}$ can reach $\mathcal{C}$ only when

$$\begin{aligned} \hat{\Phi}_\rho(v, \omega) &= \begin{cases} \sigma_\rho^2, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \\ \hat{\tau}(v, \omega) &= \begin{cases} 1, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \end{aligned} \quad (33)$$

However, neither of these two conditions can occur in practice because both the PSD $\hat{\Phi}_\rho(v, \omega)$ of natural scenes and the SFR $\hat{\tau}(v, \omega)$ of image-gathering devices decrease gradually with increasing spatial frequency, as depicted in Figs. 3 and 6, respectively. Hence, $\mathcal{H}$ can never reach $\mathcal{C}$.

Theoretical Minimum Data Rate $\mathcal{E}$

The theoretical minimum data rate $\mathcal{E}$ (also in bits sample⁻¹) that is required to convey the information rate $\mathcal{H}$ is defined by

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \tilde{s}(v, \omega)] \quad (34)$$

where the second term, $\mathcal{E}[\cdot | \cdot]$ is the uncertainty of $\tilde{s}_e(v, \omega; \kappa)$ when $\tilde{s}(v, \omega)$ is known. For the assumptions that have been made about the coding process, the conditional entropy becomes the entropy $\mathcal{E}[\hat{n}_q(v, \omega; \kappa)]$ of the quantization noise, so that $\mathcal{E}$ simplifies to

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\hat{n}_q(v, \omega; \kappa)] = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2 + \hat{\Phi}'_{n_e}(v, \omega)}{\hat{\Phi}'_q(v, \omega; \kappa)} \right] dv d\omega \quad (35)$$

This expression for $\mathcal{E}$ represents the entropy of completely decorrelated data.

Information Efficiency $\mathcal{H}/\mathcal{E}$

The information efficiency of completely decorrelated data is defined by the ratio $\mathcal{H}/\mathcal{E}$. This ratio reaches its upper bound $\mathcal{H}/\mathcal{E} = 1$ when $\hat{\Phi}'_{n_e}(v, \omega) \ll \hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2$ and $\hat{\Phi}'_{n_e}(v, \omega) \ll \hat{\Phi}'_\rho(v, \omega; \kappa)$ because then both Eq. (31) for $\mathcal{H}$ and Eq. (35) for $\mathcal{E}$ reduce to the same expression

$$\mathcal{H} = \mathcal{E} = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\hat{\Phi}'_q(v, \omega; \kappa)} \right] dv d\omega \quad (36)$$

This upper bound can be approached with a small loss in $\mathcal{H}$ only when the aliasing and photodetector noise are small. Hence, the electro-optical design that optimizes $\mathcal{H}$ also optimizes $\mathcal{H}/\mathcal{E}$. However, a compromise always remains between $\mathcal{H}$ and $\mathcal{H}/\mathcal{E}$ in the selection of the number of quantization levels: $\mathcal{H}$ favors fine quantization, whereas $\mathcal{H}/\mathcal{E}$ favors coarse quantization.

Maximum-Realizable Fidelity $\mathcal{F}$

The maximum-realizable fidelity $\mathcal{F}$ of the images restored with the Wiener filter $\hat{\Psi}(v, \omega; \kappa)$ can be expressed as a function of the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$ given by Eq. (24) as

$$\mathcal{F} = \int \int \hat{\Phi}'_\rho(v, \omega) \hat{\Gamma}_r(v, \omega; \kappa) dv d\omega \quad (37)$$

or, equivalently, as a function of the integrand $\hat{\mathcal{H}}(v, \omega; \kappa)$ of the information rate $\mathcal{H}$ given by Eq. (31) as

$$\mathcal{F} = \int \int \hat{\Phi}'_\rho(v, \omega) [1 - 2^{-\hat{\mathcal{H}}(v, \omega; \kappa)}] dv d\omega \quad (38)$$

This relationship between fidelity and information rate can be extended to other goals. For example, the enhancement of spatial features (e.g., edges and boundaries), for robotic vision, can be accommodated by letting $\tilde{\tau}_c(v, \omega)$ be the desired feature-enhancement filter.

Performance and Design Trade-offs

Electro-optical Design

Fig. 7 presents the information rate $\mathcal{H}$ (for lossless encoding) as a function of the optical-response index ζ_c and the mean spatial detail μ . The information rate $\mathcal{H}$ versus ζ_c is given for several SNRs $K\sigma_p/\sigma_p$ and $\mu=1$, and $\mathcal{H}$ versus μ is given for the three informationally optimized designs specified in Table 1.

The curves show that $\mathcal{H}$ can reach its highest possible value only when both of the following conditions are met:

1. The relationship between the SFR $\hat{\tau}(v, \omega)$ and the sampling passband $\hat{B}$ is chosen to optimize $\mathcal{H}$ for the available SNR $K\sigma_p/\sigma_p$. This design is said to be informationally optimized.

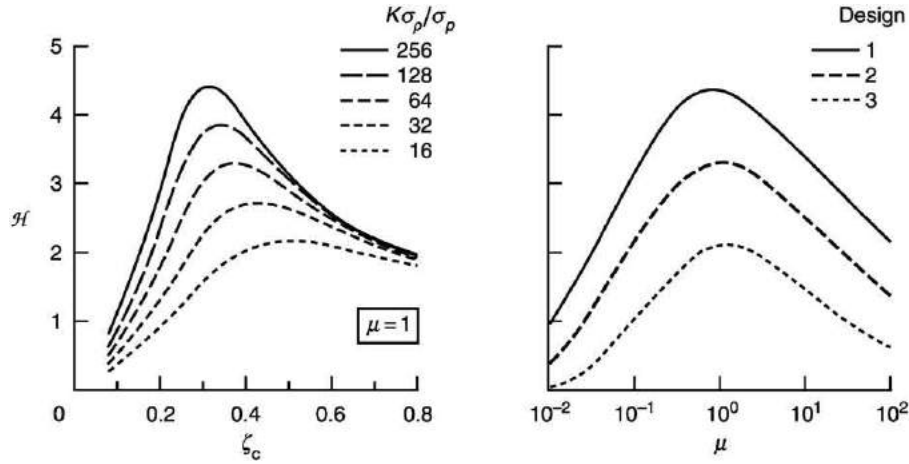


Fig. 7 Information rate $\mathcal{H}$ versus the optical-response index ζ_c and the mean spatial detail μ (relative to the sampling interval). The designs are specified in Table 1, and the SFRs $\hat{\tau}(v, \omega)$ for ζ_c are shown in Fig. 6.

Table 1 Informationally optimized designs

Design	$K\sigma_p/\sigma_p$	ζ_c	$\mathcal{H}^*$
1	256	0.3	4.4
2	64	0.4	3.3
3	16	0.5	2.2

* $\mu=1$.

2. The relationship between the sampling passband $\hat{B}$ and the PSD $\hat{\Phi}_p(v, \omega)$ is chosen to optimize $\mathcal{H}$. For the PSDs $\hat{\Phi}_p(v, \omega)$ typical of natural scenes, the best match occurs when the sampling interval is near the mean spatial detail of the scene (i.e., when $\mu \approx 1$).

These conditions are consistent with those for which the information rate $\mathcal{H}$ reaches Shannon's channel capacity $\mathcal{C}$. However, blurring and aliasing constrain $\mathcal{H}$ to values that are smaller than $\mathcal{C}$ by a factor of nearly two (i.e., $\mathcal{H}\mathcal{C}/2$). Hence, these conditions replace the role of white and band-limited signals in classical communication theory, to establish an upper bound on the information rate in imaging systems.

The above two conditions appeal intuitively when the problem of image restoration is considered:

1. The result that the SFR $\hat{\tau}(v, \omega)$ that optimizes $\mathcal{H}$ depends on the available SNR is consistent with the observation that, in one extreme, when the SNR is low, substantial blurring should be avoided because the photodetector noise would constrain the enhancement of fine spatial detail even if aliasing were negligible. In the other extreme, when the SNR is high, substantial aliasing should be avoided so that this enhancement is relieved from any constraints except those that the sampling passband inevitably imposes.
2. The result that $\mathcal{H}$ reaches its highest value when the sampling interval is near the mean spatial detail of the scene is consistent with the observation that, ordinarily, it would not be possible to restore spatial detail that is finer than the sampling interval, whereas it would be possible to restore much coarser detail from fewer samples.

Information Rate and Data Rate

Fig. 8 presents an information-entropy $\mathcal{H}(\mathcal{E})$ plot that characterizes the relationship between the information rate $\mathcal{H}$ of the encoded signal and the associated theoretical minimum data rate, or entropy, $\mathcal{E}$ for η -bit quantization. The encoder SFR $\hat{\tau}_e(v, \omega)$ is assumed to be unity. The plot illustrates the trade-off between $\mathcal{H}$ and $\mathcal{E}$ in the selection of the number of quantization levels for encoding the acquired signal. It shows, in particular, that the design that realizes the highest information rate $\mathcal{H}$ also enables the encoder to realize the highest information efficiency $\mathcal{H}/\mathcal{E}$. This result appeals intuitively, because the perturbations due to aliasing and photodetector noise that interfere with the image restoration can also be expected to do so with the signal decorrelation.

Throughput SFR

Fig. 9 illustrates the SFRs $\hat{\tau}(v, \omega)$ and $\hat{\Psi}(v, \omega; \kappa)$ of the image-gathering device and Wiener filter, respectively, and of their product, the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$ for the three designs specified in **Table 1**. As the accumulated noise becomes negligible, $\hat{\Gamma}_r(v, \omega; \kappa)$ approaches the ideal SFR of the classical communication channel, which is unity inside the sampling passband and zero outside.

Information Rate, Fidelity and Robustness

The PSD $\hat{\Phi}_p(v, \omega)$ of a scene is seldom known when images are restored. Moreover, even if the PSD were known to be the one given by **Eq. (2)**, then the mean spatial detail μ would usually remain uncertain because it depends, not only on the properties of the scene, but also on the angular resolution and viewing distance of the image-gathering device. In practice, therefore, image restorations have to rely on estimates of the statistical properties of scenes. The tolerance of the quality of the restored image to errors in these estimates is referred to as the robustness of the image restoration.

Fig. 10 presents the information rate $\mathcal{H}$ and the corresponding maximum-realizable fidelity $\mathcal{F}$ for matched and mismatched Wiener restorations. These restorations represent the digital image $R(\bar{x}, \bar{y}; \kappa)$ or, equivalently, the continuous image $R(x, y; \kappa)$ displayed on a noise-free medium. The matched restorations use the correct value of μ , whereas the mismatched restorations use wrong estimates of μ . These curves reveal that the informationally optimized designs produce the highest fidelity and robustness, and that both improve with increasing $\mathcal{H}$. Moreover, these curves reveal that, ordinarily, one cannot go far wrong in Wiener restorations by assuming that the mean spatial detail is equal to the sampling interval (i.e., $\mu = 1$). This observation appeals

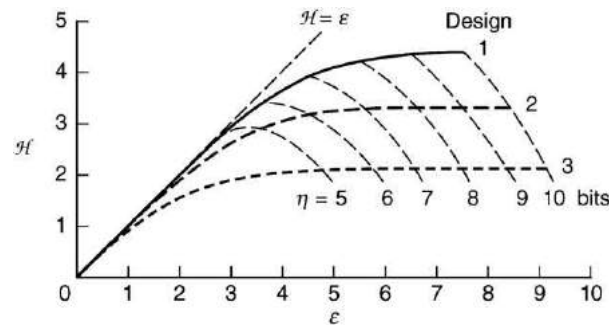


Fig. 8 The information-entropy $\mathcal{H}(\mathcal{E})$ plot of the information rate $\mathcal{H}$ versus the associated theoretical minimum data rate $\mathcal{E}$ for η -bit quantization. The designs are specified in **Table 1**.

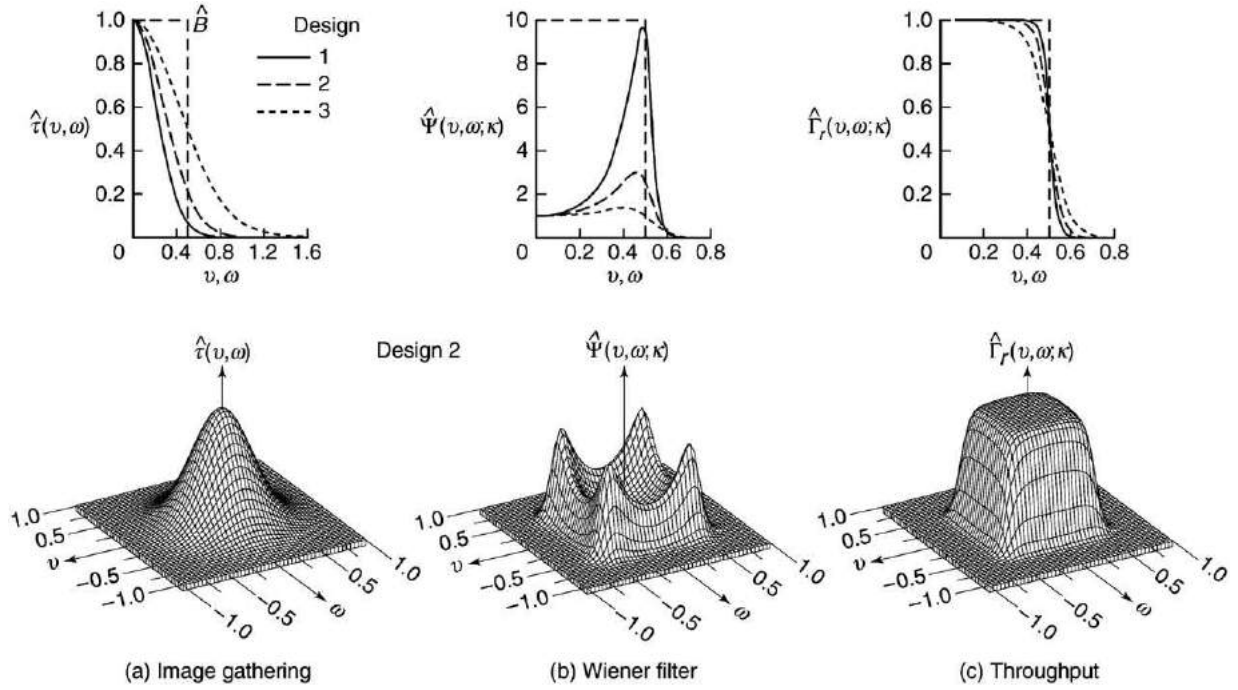


Fig. 9 SFRs of image gathering, restoration, and throughput. The designs are specified in Table 1.

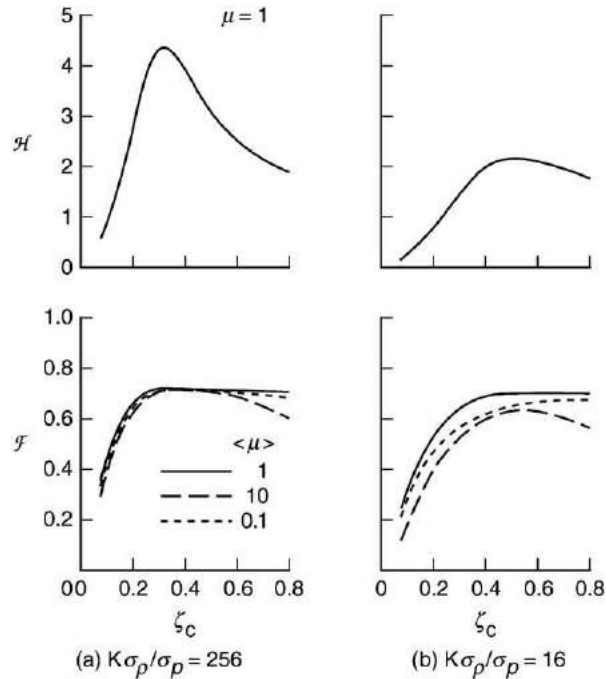


Fig. 10 Information rate $\mathcal{H}$ and maximum-realizable fidelity $\mathcal{F}$ versus the optical-response index ζ_c for two SNRs $K\sigma_\rho/\sigma_p$. $\mathcal{F}$ is given for the matched ($\langle\mu\rangle=1$) and two mismatched Wiener restorations.

intuitively because spatial detail that is much smaller cannot be resolved, and detail that is larger is not degraded substantially by blurring and aliasing.

Information Rate and Visual Quality

Figs. 11 and 12 present Wiener restorations and their three components, as given by Eq. (23), where $\rho(x, y) * \Gamma_r(x, y; \kappa)$ accounts for blurring, $n_{ar}(x, y; \kappa)$ accounts for aliasing, and $n_{pr}(x, y; \kappa)$ accounts for photodetector noise. The contrast of the illustrations of the

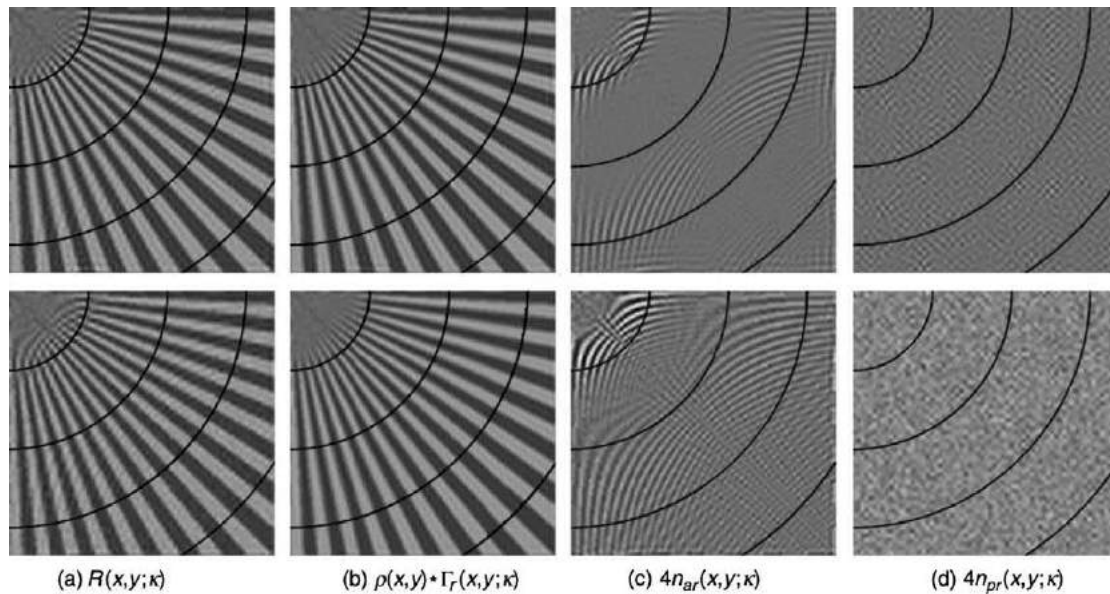


Fig. 11 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

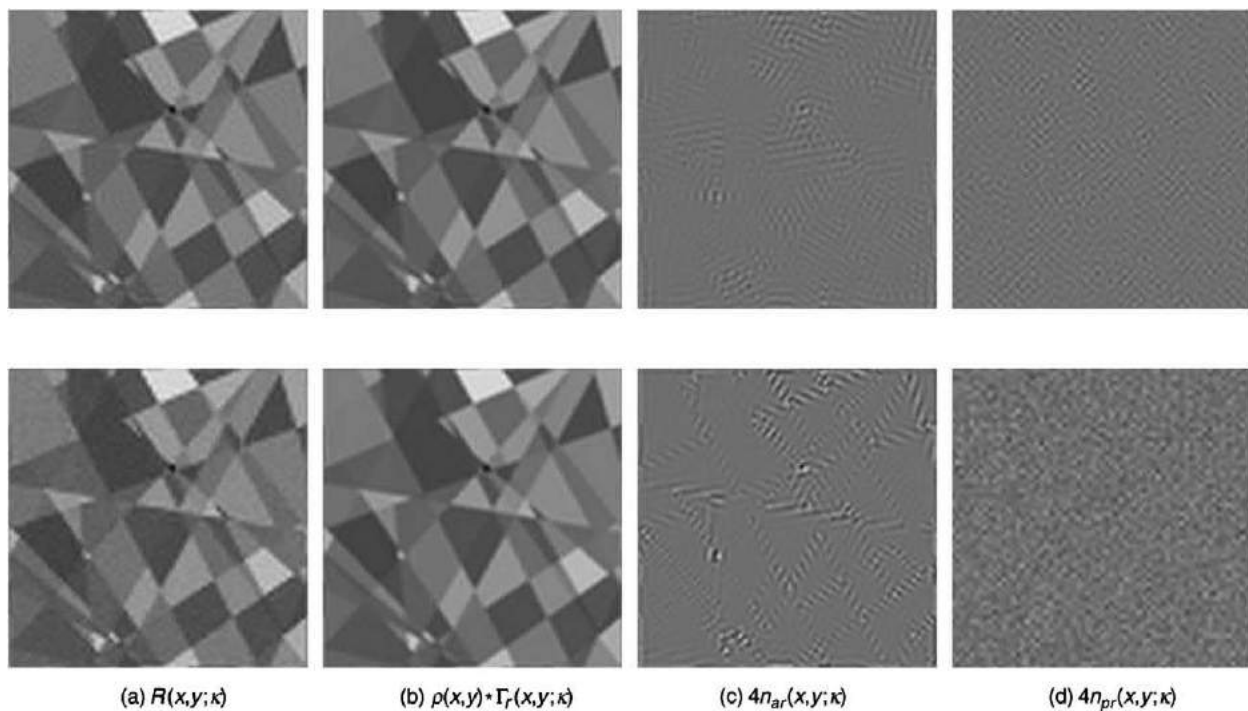


Fig. 12 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

two noise components has been increased by a factor of four to improve their visibility. The extraneous structure that aliasing produces in the images of the resolution wedges is commonly referred to as Moiré pattern. The corresponding distortion in the images of the random polygons emerges more subtly as jagged (or staircase) edges.

The images are formed from a set of 64×64 samples with a $Z=4$ interpolation and have a 4×4 cm format. If these images were displayed in a smaller format, then the jagged edges could not be resolved by the observer and would appear to be the result of blurring or photodetector noise instead of aliasing. Aliasing has, therefore, often been overlooked as a significant source of image degradation.

The Wiener restoration produces images in which fine spatial detail near the limit of resolution normally has a high contrast approaching that in the scene. However, these images also tend to contain visually annoying defects, due to the accumulated noise

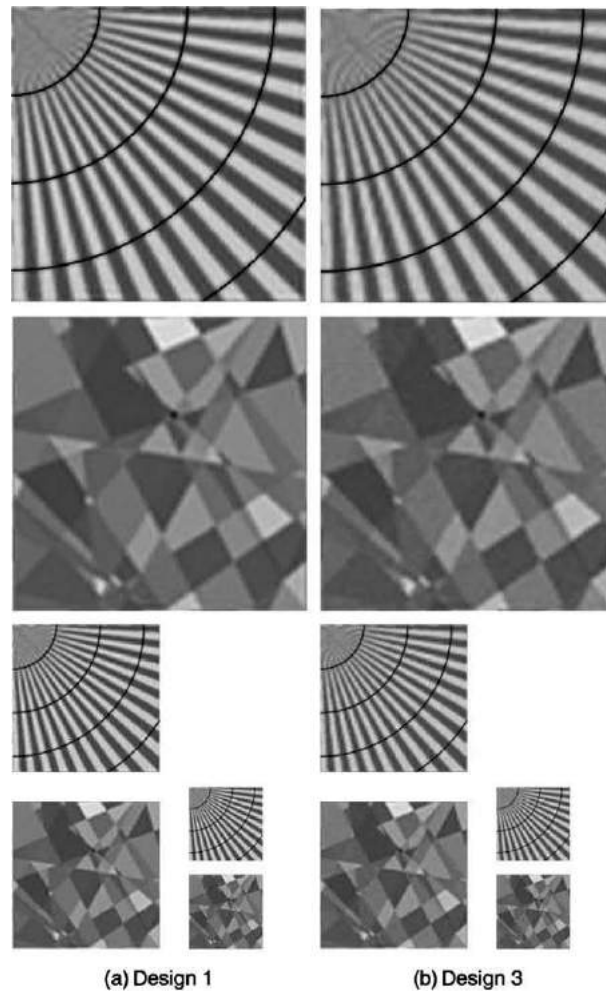


Fig. 13 Images restored with the Wiener filter and enhanced for visual quality. The large, medium, and small formats contain 16, 32 and 64 pixels/cm, respectively.

and the ringing near sharp edges (Gibbs phenomenon). Hence, it is often desirable to combine this restoration with an enhancement filter that gives the user some control over the trade-off among fidelity, sharpness and clarity.

Fig. 13 presents images for which the Wiener restorations given in **Figs. 11** and **12** are enhanced for visual quality by suppressing some of the defects due to aliasing and ringing. This enhancement improves clarity at the cost of a small loss in sharpness. The images are produced again from 64×64 pixels, but they are displayed in three different formats. The smallest format, with a density of 64 pixels/cm, corresponds to a display of 256×256 pixels in a 4×4 cm area, which is typical of the images found in the prevalent digital image-processing literature. As can be observed, this small format leads to images that are indistinguishable from each other because they hide many of the distortions that larger formats clearly reveal.

In general, if the images are displayed in a format that is large enough to reveal the finest spatial detail near the limit of resolution of the imaging system, then distortions due to the perturbations in this system become also visible. When no effort is spared to reduce these distortions without a perceptual loss of resolution and sharpness, then it becomes strongly evident that the best visual quality with which images can be restored improves with increasing information rate, even after the fidelity has essentially reached its upper bound. This improvement continues until it is gradually ended by the unavoidable compromise among sharpness, aliasing, and ringing as well as by the granularity of the image display.

Concluding Remarks

The performance of image-gathering devices and the human eye is constrained by the same critical limiting factors. When these factors are accounted for properly, then it emerges that the design of the image-gathering device that is optimized for the highest realizable information rate corresponds closely, under appropriate conditions, to the design of the human eye that evolution has

optimized for viewing the real world. This convergence of information theory and evolution toward the same design clearly supports the extension of information theory to the characterization of imaging systems.

Numerous other information-theoretic characterizations of imaging are now emerging. However, none of these characterizations accounts for the critical limiting factors that constrain image gathering and restoration, as outlined here. Hence, it is not possible for them to address the efficiency and accuracy with which images can be conveyed. Moreover, none of these characterizations combines observation with communication, or sensing with processing, in a general formalization. Instead, each of them still deals with a particular measurement or imaging problem. Hence, the role of information theory in imaging is far from mature. Many challenges remain: to develop general formalizations, to account for physical phenomena, and to correlate predictions with measurements.

Further Reading

- Fales, C.L., Huck, F.O., Alter-Gartenberg, R., Rahman, Z., 1996. Image gathering and digital restoration. *Philosophical Transactions of the Royal Society of London A354*, 2249–2287.
- Huck, F.O., Fales, C.L., 2001. *Encyclopedia of Imaging Science and Technology*, Characterization of image systems., New York: John Wiley & Sons.
- Huck, F.O., Fales, C.L., Rahman, Z., 1996. An information theory of visual communication. *Philosophical Transactions of the Royal Society of London A354*, 2193–2248.
- Huck, F.O., Fales, C.L., Rahman, Z., 1997. *Visual Communication: An Information Theory Approach*. Boston: Kluwer Academic Publishers.
- Lesurf, J.C.G., 1995. *Information and Measurement*. Bristol, UK: Institute of Physics Pub.
- O'Sullivan, J.A., Blahut, R.E., Snyder, D.L., 1998. Information-theoretic image formation. *IEEE Transactions Information Theory* 44, 2094–2123.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423. and 28: 623–656.
- Shannon, C.E., Weaver, W., 1964. *The Mathematical Theory of Communication*. Urbana, IL: U. Illinois Press.
- Wiener, N., 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons.

Inverse Problems and Computational Imaging

M Bertero and P Boccacchi, University of Genova, Genova, Italy

© 2005 Elsevier Ltd. All rights reserved.

Introduction

A digital image is an array (2D image) or a cube (3D image) of quantized numbers, providing a representation of a physical object or a map of some distributed property of the object (density of matter, energy, etc.). It can be obtained by digitizing the original image (for instance, a photograph) or can be directly formed by the outputs of the imaging system (digital camera, microscope, telescope, etc.). When visualized on a suitable image display device, a digital image must provide a representation of the physical object which can be easily interpreted and analyzed. In many cases, however, this condition is not satisfied. We give two typical examples.

The first is that of an image degraded by blurring which can be introduced by defects of the optical systems (such as aberrations) or by image motion, image defocusing, etc. The representation of the object does not have the quality which should be desirable for its interpretation.

The second case is even more important and arises when the output of the imaging system is not directly related to the physical quantity to be imaged. For instance, in X-ray tomography, the output of a detector is the attenuation of an X-ray pencil crossing the body while the quantity to be imaged is the density of the body.

In the situations described above, the desired image can be obtained by solving a number of mathematical problems. If we denote f as the object, namely the target of the imaging system, and g as the image of f provided by the system, then the problems to be solved can be summarized as follows:

- develop a physical model of the system to obtain a mathematical relationship between the object f and the image g ; the computation of g from a given f is usually called the solution of the direct problem;
- solve the problem of obtaining the physical quantity f from given outputs g of the system; this second step is usually called the solution of the inverse problem.

The digital image of f , obtained by solving an inverse problem, is a computed one, generated by the computer where the algorithm for the solution of the inverse problem has been implemented.

It may be convenient to formulate both the direct and the inverse problem in terms of functions rather than arrays, cubes, etc., because such a formulation allows the use of powerful tools of functional analysis for their investigation. Next, the results obtained in the continuous case can be used for understanding the features of the corresponding discrete problem. In most important applications, for instance, the relationship between f and g is linear so that, in the discrete version of the direct problem, one obtains g by applying a suitable matrix A to f while, in the discrete version of the inverse problem, one obtains f from g by solving a linear algebraic system. However, the solution of the inverse problem is not so simple and the difficulties can be understood by looking at the corresponding continuous problem.

The basic reason relies on the fact that inverse problems are ill-posed in the sense of Hadamard: the solution may not exist; even if it exists it may not be unique; and, even if it exists and is unique, it may not depend continuously on the data. The last statement means that a small perturbation of the data, such as that caused by noise or any kind of experimental error, can completely modify the solution.

Inverse problems are ill-posed because imaging systems do not transmit complete information about the physical object. Therefore, the problem is to extract the useful information contained in the data or, in mathematical terms, to look for approximate solutions which are stable against noise. This is the purpose of a mathematical theory introduced in its general form by the Russian mathematician, Tikhonov and known as regularization theory.

We illustrate the general framework outlined above by means of a specific example – image deconvolution, also known as image deblurring, image restoration, etc. The object f is the image which should be recorded in the absence of degradation, while g is the image corrupted by aberrations or other causes of blurring. Both f and g are functions of 2D (or 3D) space variables $\mathbf{x}$ which are the coordinates of a point in the image domain. If the imaging system is isoplanatic, then it is described by a space-invariant point spread function (PSF) $K(\mathbf{x})$, which is the image of a point source located in the center of the image domain. The PSF provides the response of the system to any point source wherever it is located. On the other hand, its Fourier transform, the transfer function $\hat{K}(\omega)$ (the hat denotes Fourier transform and ω , the spatial frequencies associated to the space variables $\mathbf{x}$), tells us how a signal of a fixed frequency is propagated through the linear system, so that the blurring can also be viewed as a sort of frequency filtering.

Under these assumptions the image g is the convolution product of object f and PSF K :

$$g(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')d\mathbf{x}' \quad (1)$$

and the solution of the direct problem is just the computation of a convolution product.

The image g , as given by Eq. (1), is the so-called noise-free image. The actually recorded image g_r is affected by a noise term and, in general, it can be written in the following form:

$$g_r(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \quad (2)$$

where $n(\mathbf{x})$ is a function describing noise contamination. It is a random function and, therefore, it is unknown even if one understands its statistical properties. We point out that Eq. (2) does not imply that the noise is additive or signal independent. For instance, in microscopy and astronomy, the contamination of the image is due both to photon noise, which satisfies Poisson statistics, and read-out noise, which is basically Gaussian and white. Therefore the noise term in Eq. (2) must only be intended as the difference between the noisy and the noise-free image.

In a first attempt at approaching the inverse problem, it seems quite natural to ignore the noise term $n(\mathbf{x})$ and to solve Eq. (1) for $f(\mathbf{x})$ with $g(\mathbf{x})$ replaced by $g_r(\mathbf{x})$. Then, by taking the Fourier transform of both sides of this equation and using the well-known convolution theorem, one finds the following relationship:

$$\hat{g}_r(\omega) = \hat{K}(\omega)\hat{f}(\omega) \quad (3)$$

which relates the Fourier transform of f and g to the transfer function of the imaging system.

The solution of this equation looks elementary. However, a first problem is due to the fact that an optical system is generally band-limited; its band Ω is the bounded domain of spatial frequencies where the transfer function is different from zero. Since the noise is not band-limited or, at least, has a band much broader than that of the optical system, outside Ω the right-hand side of Eq. (3) is zero while the left-hand side is not; in other words, the solution of the deconvolution problem, as formulated above, does not exist because no object f satisfies Eq. (3) everywhere. In this particular case, one can circumvent the problem by applying a suitable filter to $g_r(\mathbf{x})$ in order to suppress the out-of-band noise.

The second problem is the nonuniqueness of the solution of the problem with the filtered data. This is due to the so-called invisible objects, namely objects whose Fourier transform is zero on Ω , so that the corresponding images are exactly zero. If we find a solution of the deconvolution problem, by adding an arbitrary invisible object to this solution, we find another solution of the same problem. A well-defined solution can be obtained by requiring its Fourier transform to be zero outside Ω . In other words, we set to zero what is not transmitted by the imaging system.

The standard way for approaching the previous questions is to look for solutions in the least-squares sense, namely for objects which minimize the functional:

$$\varepsilon^2(f) = \|K * f - g_r\|^2 \quad (4)$$

where $*$ denotes the convolution product, as defined in Eq. (1), and the right-hand side is the square of the L^2 -norm (which can also be interpreted as energy) of the discrepancy between the recorded image g_r and the computed image $K * f$. If one looks for a least-squares solution with minimal energy then one automatically re-obtains the solution discussed above, namely an object whose Fourier transform in Ω is given by

$$\hat{f}_{\text{inv}}(\omega) = \frac{\hat{g}_r(\omega)}{\hat{K}(\omega)} \quad (5)$$

and is zero outside Ω . Such a solution is that provided by the so-called inverse filter.

However, there is an additional difficulty in that the transfer function usually tends to zero at the boundary of the band while the Fourier transform of the noise, hence that of g_r , does not, or tends to zero in a different way. It follows that $\hat{f}_{\text{inv}}(\omega)$ is divergent or very large at the boundary of the band. We conclude that its inverse Fourier transform does not exist or, if it exists, is affected by large artifacts due to noise propagation. The latter is the typical situation in the case of digital images. In Fig. 1 we give the result of a numerical simulation showing the typical effect produced by the inverse filter. The object is a beautiful picture of a galaxy recorded by the Hubble Space Telescope (HST), while the PSF is a computed one describing the blurring of HST before installation of the optics correction. The blurred image is obtained by convolving the object with the PSF and by adding white noise. The restored image produced by the inverse filter is completely corrupted by noise.

The inverse filter provides a solution which fits the data in the best possible way and therefore also fits the noise in the frequency domains where the signal is weakly transmitted by the imaging system. This can be achieved at the cost of an energy (L^2 -norm) of the solution which is too large. Therefore, such a conclusion suggests that one must search for a reasonable compromise between data fitting and energy of the solution. This is the basic idea of regularization theory, at least in its most simple form. The mathematical formulation is obtained by looking for objects which minimize the functional:

$$\Phi_\mu(f) = \|K * f - g_r\|^2 + \mu \|f\|^2 \quad (6)$$

where μ , the so-called regularization parameter, is a parameter controlling the trade-off between data fitting and energy of the solution: when μ is small, the data fitting is good and the energy is large while, when μ is large, the data fitting is poor and the energy is small.

For a given μ , the function f_μ which minimizes the functional of Eq. (6) is called the regularized solution of the problem. It is easy to show that its Fourier transform is given by

$$\hat{f}_\mu(\omega) = \frac{\hat{K}^*(\omega)}{|\hat{K}(\omega)|^2 + \mu} \hat{g}_r(\omega) \quad (7)$$

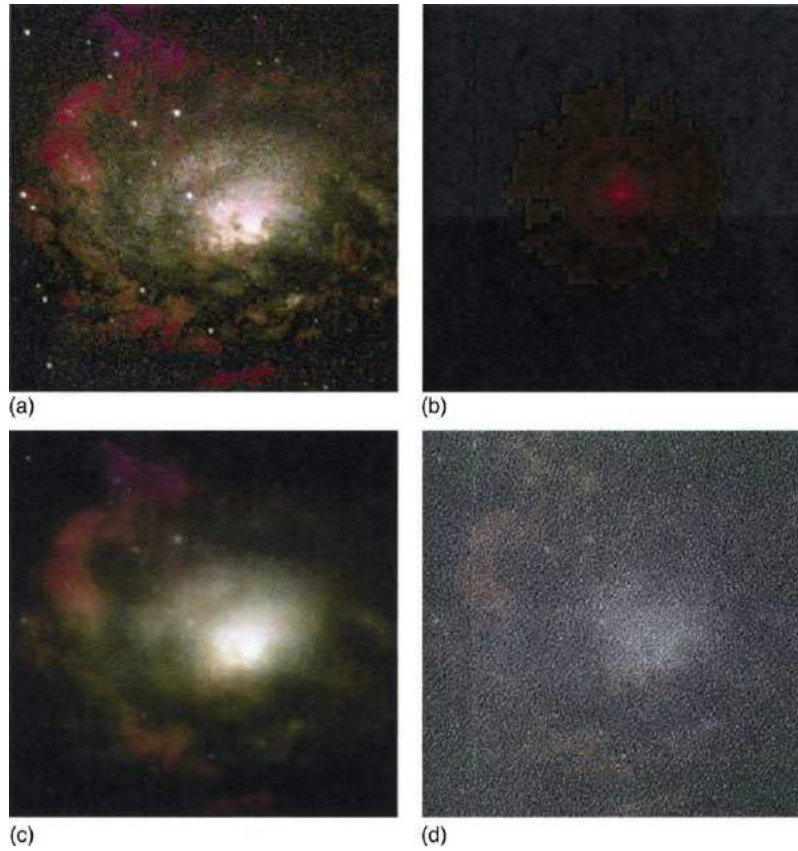


Fig. 1 Illustrating the effect of the inverse filter: (a) Image of the Circinus galaxy taken on 10 April 1999, with the Wide Field Planetary Camera 2 of the Hubble Space Telescope; (b) the PSF used for blurring the RGB components of the image (here shown in red); (c) the blurred image obtained by convolving the components of the image in (a) with the corresponding PSFs and adding noise; (d) the restoration obtained by applying the inverse filter to the three components of (c). Part (a) courtesy of NASA/Space Telescope Science Institute.

The regularized solutions form a one-parameter family of functions. When μ is small these functions are strongly contaminated by noise whose effect is gradually reduced when μ increases. For too large values of μ , one re-obtains blurred versions of the original object.

Such behavior is illustrated by the sequence of regularized solutions given in Fig. 2 and corresponding to the example of Fig. 1. The regularization parameter increases starting from left to right and from top to bottom. Its values are $\mu=0$ (inverse filter), 0.001, 0.01, 0.05, 0.1, and 0.5 (the noise is of the order of a few percent and the PSF is normalized in such a way that the sum of all its value is 1). The sequence makes evident the well-established theoretical result of the existence of an optimal value of the regularization parameter. Such an optimal value can be determined in the case of numerical simulations as that represented in Fig. 2. However, its estimation in the case of real images is more difficult. Several methods for the selection of μ have been proposed. We only mention the so-called discrepancy principle. It consists of determining the value of μ , such that the value of the discrepancy functional at f_μ , $\varepsilon^2(f_\mu)$ coincides with an estimate of the energy of the noise. In other words, the data are fitted with an accuracy comparable with their uncertainty.

As shown by Eq. (7), the regularized solution has the same band Ω of the imaging system, because $\hat{K}^*(\omega)$ is zero outside Ω . As is known, a band-limited function can be represented in terms of its samples taken at a rate which is roughly proportional to the size of the band. This is the content of the famous Shannon sampling theorem (more precisely its 2D extension). On the other hand, the sampling distance can be taken as a measure of the resolution provided by the restored image so that one concludes that the restoration method discussed above does not produce an improvement of the resolution of the imaging system. It certainly produces an improvement of the quality of the image which makes possible a visual verification of that resolution.

In many circumstances, the term super-resolution is used for describing methods which allow an improvement of resolution beyond the limit provided by the sampling theorem. Therefore, a super-resolving method must produce a restored image with a band broader than that of the imaging system. Therefore, super-resolution is related to the problem of out-of-band extrapolation.

Such an approach was already proposed in the 1960s, when it was observed that the Fourier transform of an object with a finite spatial extent (all objects have this property) is analytic; then the theorem of unique analytic continuation implies that the Fourier transform of the object can be determined everywhere from its values on the band of the imaging system.

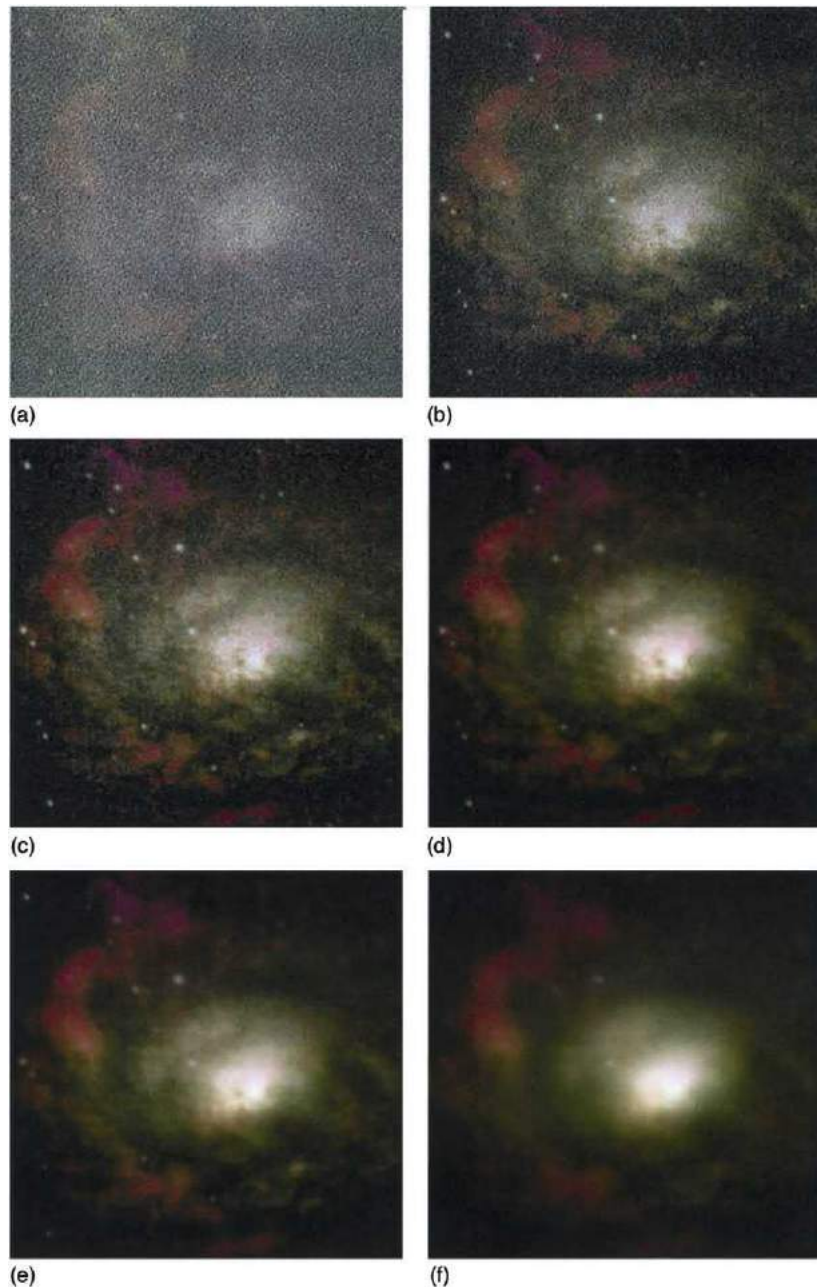


Fig. 2 Illustrating the effect of the regularization parameter. The blurred image is that shown in [Fig. 1\(c\)](#). The sequence is obtained by increasing the value of the regularization parameter: (a) $\mu=0$ (inverse filter); (b) $\mu=0.001$; (c) $\mu=0.01$; (d) $\mu=0.05$; (e) $\mu=0.1$; (f) $\mu=0.5$. It is evident that the best restoration is that shown in (c).

Unfortunately this problem is ill-posed, so that out-of-band extrapolation is not feasible in practice. Only recently it has been recognized that a considerable amount of super-resolution is possible when the spatial extent of the object is not much greater than the resolution distance of the instrument. For example, super-resolving methods could be used for detecting unresolved binary star in astronomical images.

The important point is that a super-resolving method must implement explicitly the finite extent of the object: the domain of the object must be estimated and the method must search for a solution which is zero outside this domain. This introduces an important question in object restoration and, more generally, in the theory of inverse problems, namely the design of regularized solutions satisfying additional conditions (constraints). This question is also important from the theoretical point of view: the problem is ill-posed because of insufficient information on the object as transmitted by the imaging system; the use of additional constraints reduces the class of the solutions which are compatible with the data and therefore can improve the restoration. This is the use of a priori information in the solution of inverse problems.

A constraint which has been widely investigated and used is the positivity of the solution, i.e., all the values of the restored image must be positive or zero. The physical meaning of the constraint is obvious. Its beneficial effect is to reduce ringing artifacts which affect the regularized solutions previously discussed, in cases where the object contains bright spots over a black background or sharp intensity variations from zero to some positive value. The ringing is essentially related to the well-known Gibbs effect in the truncation of the Fourier series.

The requirement of positivity is the requirement of a particular lower bound on the values of the solution. Therefore, one can consider more general constraints in requiring a given lower and/or upper bound on the values of the solution, eventually combined with a constraint on the domain to produce, for instance, a super-resolved positive solution. The most general form of these constraints requires that the solution belongs to a given closed and convex set C in some functional space. A general method producing regularized solutions in a convex set is an iterative method which is essentially a gradient method for the minimization of the least-squares functional, with a projection on the set C at each iteration. In the case of object deconvolution, the least-squares functional is that given in Eq. (4) and the method has the following form: if f_k is the result of the k -th iteration, then f_{k+1} is given by

$$f_{k+1} = P_C[f_k + \tau K^T * (g_r - K * f_k)] \quad (8)$$

where P_C is the convex projection onto the set C , τ is a relaxation parameter and $K^T(x) = K(-x)$. In general, the algorithm, known as the projected Landweber method, is initialized with $f_0 = 0$. An important property is that it has a regularization effect, in the sense that iterations must not be pushed to convergence but stopped after a suitable number of iterations to avoid strong noise contamination. In Fig. 3 we compare, in a specific example, the result obtained by means of the linear regularization method and that obtained by means of the projected iterative method with a lower and upper bound on the solution. The reduction of ringing is evident.

Among the methods producing positive solutions we must mention the most popular one in astronomy, the so-called Richardson–Lucy method, which is an iterative method for Maximum-Likelihood estimation. If the PSF has been normalized in such a way that the sum of all its values is one, it takes the following form:

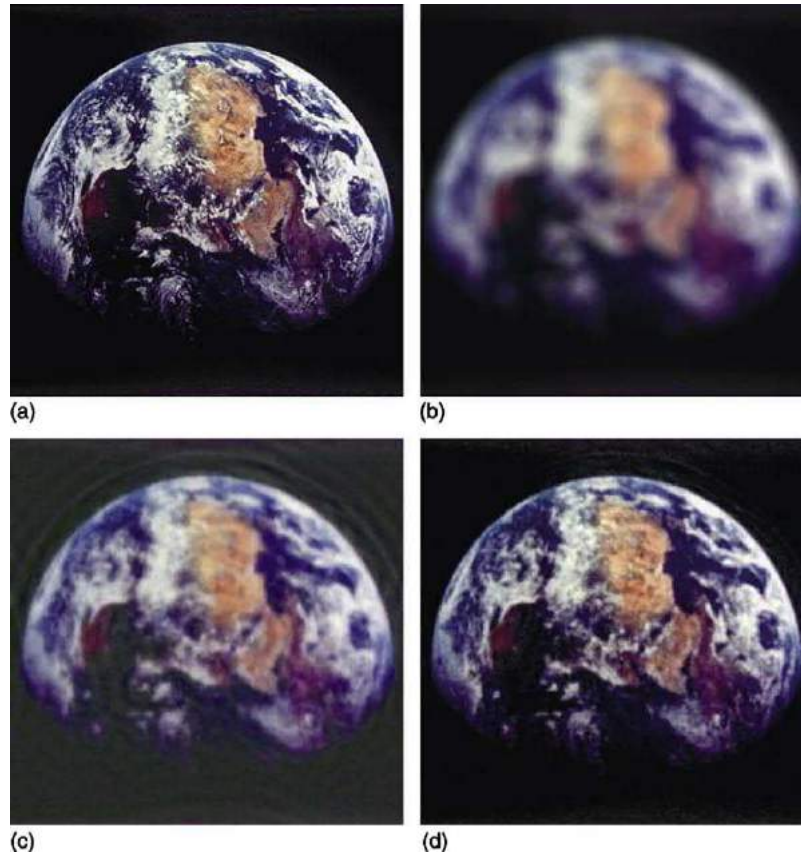


Fig. 3 Illustrating the effect of constraints on image restoration: (a) Image of the Earth taken during the Apollo 11 mission; (b) blurred image of (a) obtained by assuming out-of-focus blur and adding noise; (c) optimal restoration provided by the regularization method; the ringing around the boundary of the Earth is evident; (d) optimal restoration obtained by means of the projected Landweber method with a lower and an upper bound on the solution at each iteration. Part (a) courtesy of NASA.

$$f_{k+1} = f_k \left(K^T * \frac{g_r}{K * f_k} \right) \quad (9)$$

The method is normally initialized with a uniform image and it must not be pushed to convergence to avoid noise amplification. Another algorithm suggested by statistical arguments and producing positive solution is the so-called Maximum Entropy Method (MEM).

The image restoration methods described above have wide applications both in microscopy and astronomy. For instance, the minimization of the functional Eq. (6), with the additional constraint of positivity, has been used for deconvolving images in microscopy to reduce the effect of the missing cone. On the other hand, the Richardson–Lucy method was widely used for deconvolving the images of HST before installation of corrective optics.

Nowadays, image deconvolution is becoming more important for the ground-based telescopes equipped with adaptive optics. Indeed, even if adaptive optics is able to provide a considerable compensation of the atmospheric blur, a further improvement can be obtained by deconvolving the detected images. The PSF is provided by the image of a suitable guide star.

Finally, it is worth mentioning the Large Binocular Telescope, in construction on the Mount Graham in Arizona, because this instrument requires image restoration methods for a full exploitation of its imaging properties. The telescope (Fig. 4) consists of two 8 m mirrors on a common mount: both mirrors are equipped with adaptive optics and are combined interferometrically to reach the resolution of a 22.8 m mirror in the direction of the baseline. The telescope can be rotated to record images of the same astronomical target with different orientations of the baseline, i.e., the line joining the centers of the two mirrors. Finally the images must be processed by means of multiple images deconvolution methods to obtain a unique high-resolution image, equivalent to that produced by a 22.8 m mirror.

The applications of inverse problems with major social impact are in the area of medical imaging. The spectacular success of X-ray computerized tomography (CT), invented by GH Hounsfield at the beginning of the 1970s, has stimulated the development of other imaging modalities, based essentially on the same principle, such as positron emission tomography (PET) and single photon emission computerized tomography (SPECT). Computed images are also provided by magnetic resonance imaging (MRI). In such a case, however, the computational problem is rather simple since it consists essentially in Fourier transform inversion.

The basic principle of X-ray CT is described as a finely collimated source S (see Fig. 5(a)) emitting a pencil of X-rays which propagates through the body along a straight line L up to a well collimated receiver R . If we know both the intensity I_0 of the source and the intensity I detected by the receiver, the logarithm of the ratio I_0/I is the integral of the linear attenuation function (roughly proportional to the density function of the body) along the line L . If the source and the receiver are moved along two parallel lines, having direction θ and defining the plane Π (in practice a planar slice of the body under consideration), one obtains the integrals of the linear attenuation function $f(\mathbf{x})$ along all parallel lines orthogonal to θ . The scanning variables in the plane Π are defined in Fig. 5(b): φ is the angle formed by the unit vector θ with the x -axis and s is the signed distance between the origin and the integration line L . The integrals along all parallel lines orthogonal to θ define a function of s which is called the projection of $f(\mathbf{x})$ in the direction θ :

$$P_\theta(s) = \int_{-\infty}^{+\infty} f(s \cos \varphi - u \sin \varphi, s \sin \varphi + u \cos \varphi) du \quad (10)$$

The set of the projections of $f(\mathbf{x})$ for all possible θ is called the Radon transform of $f(\mathbf{x})$, in honor of the mathematician Johann

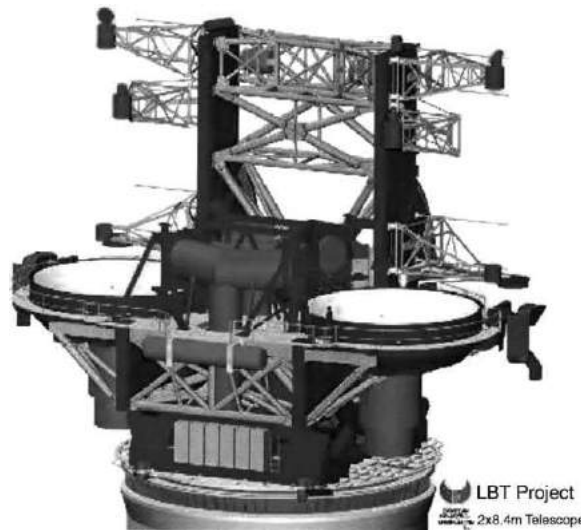


Fig. 4 Picture of the Large Binocular Telescope (LBT) in construction on the Mount Graham in Arizona. It consists of two mirrors on the same mount. The images of the two mirrors are combined interferometrically to produce a high-resolution image in the direction of the line joining the centers of the mirrors (baseline).

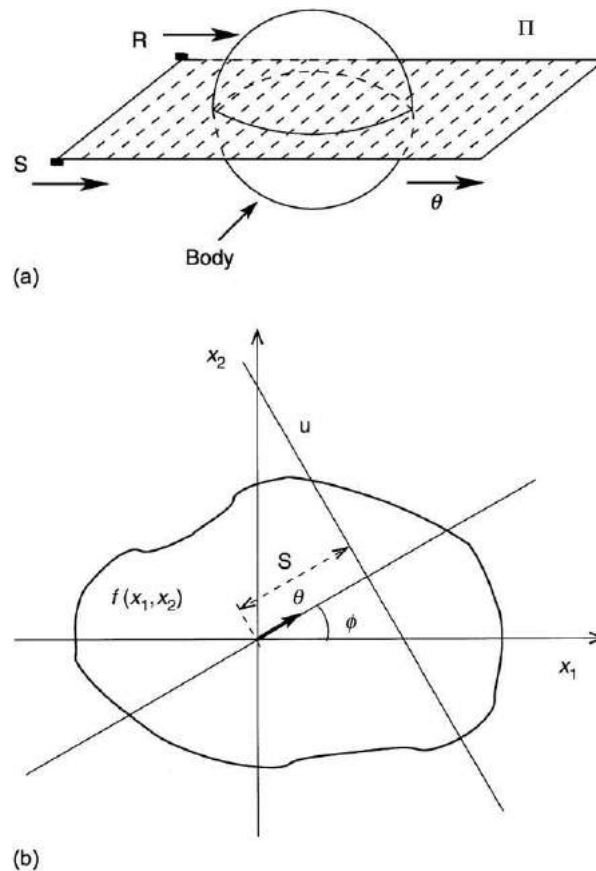


Fig. 5 (a) Scheme of the scanning procedure in X-ray tomography: the source(S)–receiver(R) pair is moved along a direction θ orthogonal to the line S–R, defining the plane Π to be imaged; when the scanning has been completed the system is rotated by a certain angle, the scanning procedure is repeated, and so on. (Reproduced with permission from Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.) (b) Definition of the variables used in the representation of the Radon transform; φ is the angle formed by the S–R line with the x_1 -axis; s is the signed distance between the origin and the integration line (orthogonal to θ); and u is the coordinate on the integration line.

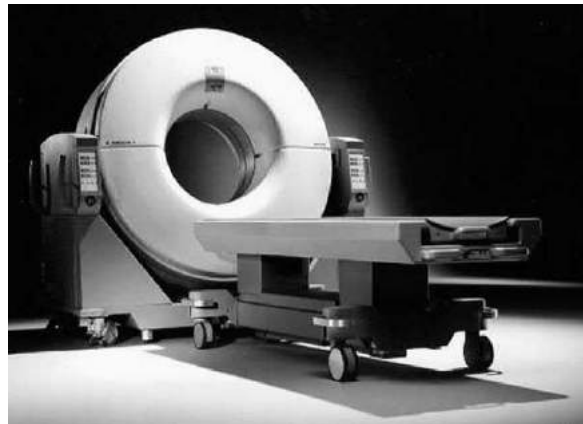


Fig. 6 Picture of a scanner for X-ray tomography. The annular part contains the acquisition and scanning systems.

Radon, who first investigated the problem of recovering a function of two variables from its line integrals. Today, such a problem is known as Radon transform inversion or object restoration from projections. Sampled values of the Radon transform, i.e., sampled values of the projections for a number of possible directions between 0 and π , are the outputs of the acquisition system of the medical equipment known as CT-scanners (Fig. 6). Their data-processing system contains the implementation of an algorithm for Radon transform inversion and the final result is a set of computed images providing maps of slices of the human body.

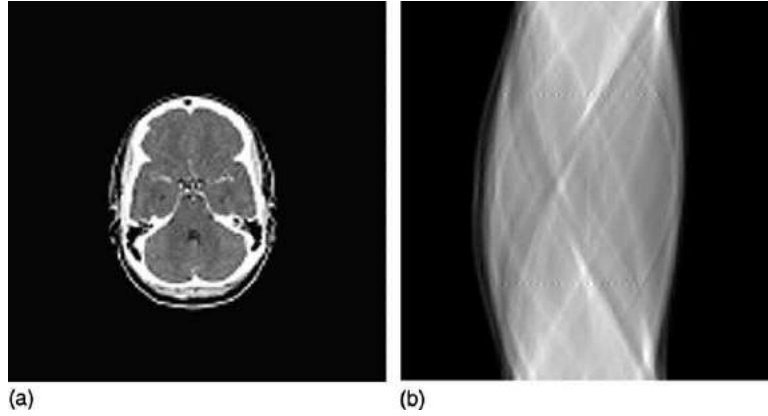


Fig. 7 (a) Picture of a brain slice; (b) the corresponding sinogram. As explained in the text, the sinogram is a gray-level representation of the Radon transform in the s, φ plane. Each horizontal line in (b) corresponds to a projection of (a).

The plot of the Radon transform of f in the (s, φ) -plane, obtained by representing its values as gray levels, is called the sinogram of f . In Fig. 7, we give the image of the sinogram of a brain slice. The horizontal rows provide representations of the projections of the phantom. In the image, the angle φ defining the direction θ takes values between 0 and 2π while the variable s takes values between $-a$ and a , if a is the radius of a disc containing the phantom. As is evident, the sinogram is symmetric with respect to the point $(0, \pi)$. Each point in the rectangle corresponds to a straight line crossing the phantom and the straight lines through a fixed point of the phantom describe a sinusoidal curve in the (s, φ) -plane. This property has given rise to the name sinogram.

The sinogram is, in a sense, the image of the body slice under investigation, as provided directly by a CT-scanner. However, an inspection of this image does not easily provide information about the body. Therefore, it must be processed to obtain a more significant image. Radon transform inversion is an example of inverse and ill-posed problem and its solution must be treated with extra care.

The crucial point in Radon transform inversion is the Fourier slice theorem, whose content is the following: the Fourier transform of the projection in the direction θ of the function f , is the Fourier transform of f on the straight line passing through the origin and having direction θ :

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\theta) \quad (11)$$

This theorem clarifies the information content of CT data: each projection provides the Fourier transform of the function f along a well-defined straight line in the plane of the spatial frequencies. It is also the starting point for deriving the basic algorithm of data inversion in tomography, the filtered back-projection (FBP). It is obtained from Fourier transform inversion in polar coordinates and is a two-step algorithm: the first step is a filtering of the projections by means of a ramp-filter; the second is the back-projection of the filtered sinogram. More precisely the two steps are as follows:

- *Filtering*: for each θ the projection $P_\theta(s)$ is replaced by a filtered projection given by

$$Q_\theta(s) = \int_{-\infty}^{+\infty} |\omega| \hat{P}_\theta(\omega) e^{i\omega s} d\omega \quad (12)$$

where the multiplication by the ramp filter $|\omega|$ derives from the Jacobian of polar coordinates;

- *Back-projection*: this operation is the dual of the projection operation: indeed the projection assigns to a straight line L with coordinates (s, φ) , hence to a point of the domain of definition of the Radon transform, the integral of f along L ; on the other hand, the back-projection assigns the value of the Radon transform at (s, φ) to all points of the straight line L with the same coordinates (a pictorial representation of projection and back-projection is given in Fig. 8); as a consequence the back-projection operator $R^\#$ assigns to each point $\mathbf{x}$ the sum (integral) of all values of the Radon transform corresponding to straight lines passing through $\mathbf{x}$:

$$R^\# P_\theta(\mathbf{x}) = \int_0^{2\pi} P_\theta(x \cos \varphi + y \sin \varphi) d\varphi \quad (13)$$

The back-projection operator, when applied directly to the projections as in Eq. (13), provides a blurred image of f ; in order to get a satisfactory restoration it must be applied to the filtered projections $Q_\theta(s)$. This point is illustrated in Fig. 9.

We point out that the ramp filter introduced in Eq. (12) amplifies the high frequency components of the projections hence the noise corrupting these components. This remark clarifies that Radon transform inversion is an ill-posed problem and requires some kind of regularization. This is obtained by introducing an additional low-pass filter which reduces the effect of the high frequencies. The most frequently used combination of ramp and low-pass filter is the so-called Shepp–Logan filter.

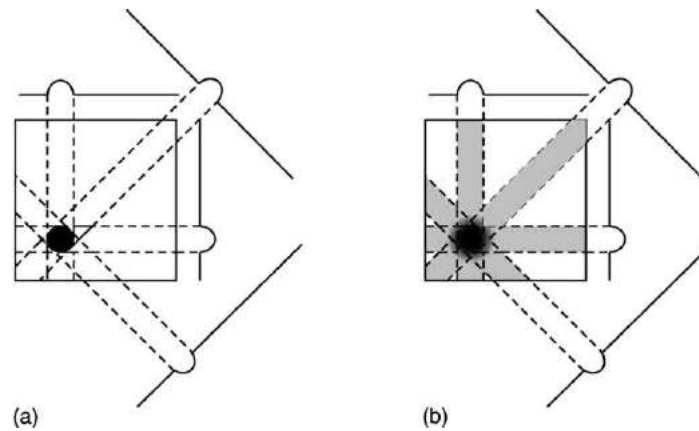


Fig. 8 (a) The effect of the projection operation is shown by drawing the projections of a circular phantom corresponding to three directions. (b) shows the result obtained by applying the back-projection operation to the three projections given in (a). It is evident that the picture provided by this operation has a maximum on the domain of the circular phantom.

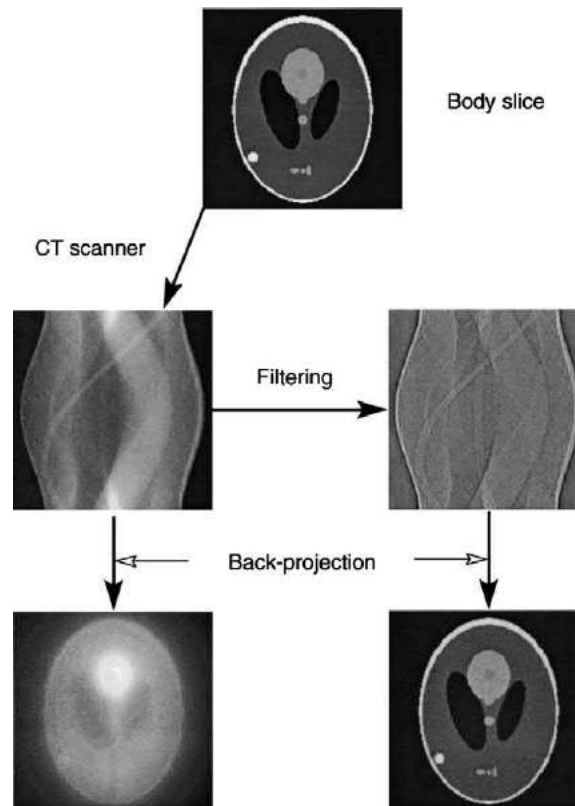


Fig. 9 Illustrating the effect of the filtered back-projection. Shown here is: a slice of the so-called Shepp–Logan phantom; the corresponding sinogram (indicated by the arrow CT scanner); the filtered sinogram (indicated by the arrow filtering); the restorations obtained by applying the back-projection operation both to the original sinogram and to the filtered sinogram.

X-ray tomography is also called transmission computerized tomography (TCT) because the image is obtained by detecting the X-rays transmitted by the body. It provides information about anatomical details of human organs because the map of the linear attenuation function is essentially the map of the density of the tissues.

A different type of information is obtained by the so-called emission computerized tomography (ECT) which is based on the administration of radio-nuclide-labeled agents known as radio-pharmaceuticals. Their distribution in the body of the patient depends on factors such as blood flow, metabolic processes, etc. Then a map of this distribution is obtained by detecting the γ -rays produced by the decay of the radio-nuclides. Therefore, ECT yields functional information, in the sense that the images produced by ECT show the function of the tissues of the organs.

Two different modalities of ECT are usually considered:

- single photon emission computerized tomography (SPECT), which makes use of radio-isotopes where a single γ -ray is emitted per nuclear disintegration;
- positron emission tomography (PET), which makes use of β^+ -emitters, where the final result of a nuclear disintegration is a pair of γ -rays, propagating in opposite directions, produced by the annihilation of the emitted positron in the tissues.

In both cases it is necessary to detect the γ -rays coming from well-defined regions of the body and, to this purpose, different methods are used in each case.

In SPECT, the discrimination of the γ -rays is obtained by means of a collimator, consisting of holes in a large lead slab covering the crystal detector face. In PET the collimation is obtained by means of pairs of detectors in coincidence. This technique, which is also called electronic collimation, is more accurate than the physical collimation used in SPECT and allows the design of systems with a great efficiency.

The basic reconstruction method in TCT, namely FBP, is often used in the reconstruction of SPECT and PET data. However, several corrections are necessary in practice. The principal ones are due to the collimator blur and to the scattering of photons in the body. If one develops a more accurate model of data acquisition by taking into account these effects, then the restoration problem implies the inversion of a very large matrix which is sparse and ill-conditioned. To this purpose iterative regularization methods are used such that the basic operation required at each step is matrix-vector multiplication. Other imaging modalities, which have been proposed for medical applications, are electrical impedance and microwave tomography. The underlying inverse problems are basically nonlinear so that the computational cost of the methods for their solution is, in general, too high for clinical applications. However the research in these areas is very active and, therefore, the situation could be improved in the near future.

Further Reading

- Bertero, M., 1989. Linear inverse and ill-posed problems. In: Hawkes, P.W. (Ed.), *Advances in Electronics and Electron Physics*, vol. 75. New York: Academic Press.
- Bertero, M., Boccacci, P., Boccacci 1998. *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.
- Bertero, M., De Mol, C., 1996. Super-resolution by data inversion. In: Wolf, E (Ed.), *Progress in Optics*, vol. XXXVI. Amsterdam: Elsevier, pp. 129–178.
- Engl, H.W., Hanke, M., Neubauer, A., 1996. *Regularization of Inverse Problems*. Dordrecht, Germany: Kluwer.
- Groetsch, C.W., 1993. *Inverse Problems in Mathematical Sciences*. Braunschweig, Germany: Vieweg.
- Herman, G.T. (Ed.), 1979. *Image Reconstruction from Projections*. Berlin: Springer.
- Herman, G.T., 1980. *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography*. New York: Academic Press.
- Huang, T.S. (Ed.), 1975. *Picture Processing and Digital Filtering*. Berlin: Springer.
- Kak, A.C., Slaney, M., 1988. *Principles of Computerized Tomographic Imaging*. New York: IEEE Press.
- Krestel E (ed.) (1990) *Imaging Systems for Medical Diagnostics*. Berlin: Siemens Aktiengesellschaft.
- Natterer, F., 1986. *The Mathematics of Computerized Tomography*. New York: Wiley.
- Tikhonov, A.N., Arsenin, V.Y., 1977. *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley.

Adaptive Optics

C Pernechele, Astronomical Observatory of Padova, Padova, Italy

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The image quality of a perfect optical system is ideally limited by its diffraction figure. In the real world there are many effects which degrade that physical limit, resulting in poorer image quality at the focal plane. These effects may be static (e.g., time-independent, as, for example, error in the alignment of the optical train) or dynamic (e.g., variable with time, such as optical misalignment introduced by thermo-mechanical distortion). Active and adaptive optics, the topics of this article, are technologies able to correct for the optical aberrations induced by dynamical effects. There are a variety of causes liable for dynamical image degradation, among these time-variable optical misalignments, pointing and tracking errors (for motorized instruments), and atmospheric perturbations. In practice, both atmospheric turbulence and misaligned instrument introduce random spatial and temporal perturbations on the optical beam, distorting the spherical (ideal) wavefront incoming on to the focal plane. This leads to loss of both the diffraction limited condition and image resolution. A typical situation for astronomical application is shown in Fig. 1. A target (a star in this case) emits light whose wavefront, being far from the observing system, can be considered plane. When this wavefront passes through the atmosphere it is distorted with a temporal frequency of the order of hundreds of Hertz. This dynamically distorted wavefront fills the aperture of the optical system and will once more be distorted by the dynamic misalignment of the optical system itself. The final wavefront will be deformed with respect to the ideal (spherical) wavefront.

Dynamic optical misalignment may depend on gravitational and thermal deformation, while pointing error may appear as very slow (drift), slow (wander), or fast (jitter) tilt of the image on the focal plane. Atmospheric perturbations are produced by air turbulence and may be natural (as in the astronomical or satellite tracking cases) or induced by the experiment itself

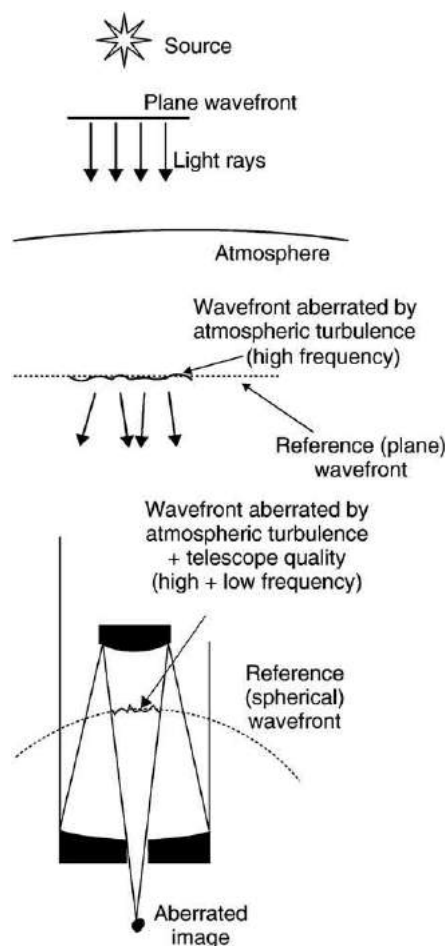


Fig. 1 Dynamical wavefront aberration.

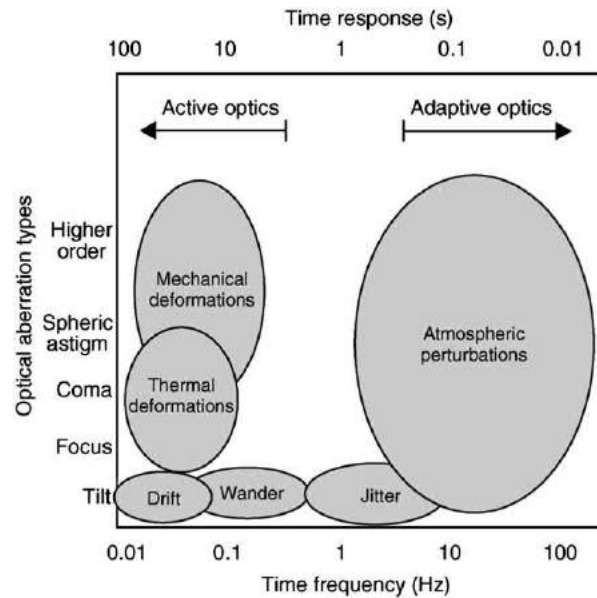


Fig. 2 Wavefront aberration time-scale and associated aberrations.

(thermal blooming in high power laser applications). All these effects appear on very different time-scales, as shown in [Fig. 2](#). As noticeable, atmospheric turbulence introduces aberrations of all orders, at a frequency above about 1 Hz. Pointing and tracking errors introduce an image tilt with different time-scale. Gravitational effects induce mechanical deformation of the optics, mainly in astronomical applications because of the massive components (the telescope primary mirror in particular). Thermal deformation produces optical misalignment among optical components mainly resulting in defocus. [Fig. 2](#) is purely illustrative, because some effects may produce wavefront aberrations at different frequencies than indicated there. For example, drift may be faster than indicated in the figure, and thermal blooming (a thermal effect) introduces aberrations at a higher frequency, and so on.

The correction of the wavefront errors is usually made by means of devices able to change the phase wave somewhere in the light path. Though the concept for correcting aberrated wavefront at slow or high frequency is the same, for example, the use of one or more deformable elements in the optical path, the technologies developed for each area are quite different: variations at slow frequency (approximately below 0.1 Hz in [Fig. 2](#)) are indicated as quasi-static errors and the techniques developed to overcome them are properly called active optics, while adaptive optics indicates the techniques to overcome the unwanted effects at high frequency (above about 1 Hz in frequency). The frequency value dividing the active from adaptive optics is not well defined and depends on the field of application and its scientific engineering communities. The important point is that a correct use of a combination of the two techniques may ultimately lead to bringing the imaging system close to its diffraction limit.

Slow Corrections: Active Optics

The analysis of the wavefront is required before its correction, and one or more deformable/movable mirrors and their control system, to operate the corrections. A typical system for low frequency correction is shown in [Fig. 3](#). A telescope contains three mirrors (M1, M2, and M3) and an instrument to detect its scientific target, usually placed on its optical axis. The dynamic deformations present in this kind of instrument are mainly third-order spherical (thermal deformation of M1), third-order astigmatism (a saddle-horse deformation of M1 due to gravity when the telescope is pointing away from the zenith), and third-order coma (due to gravitational decentering between M1 and M2). Since a telescope has a tracking system some pointing error may be present, causing motion of the image in the focal plane. The observation of an off-axis target (typically a star for astronomical applications) is made to measure the deformations of the incoming wavefront by means of a wavefront sensor (WS) system. This step is completed using some of the wavefront sensing technique described elsewhere in this encyclopedia. The WS must measure the information on the wavefront distortion with adequate spatial resolution to compensate for aberrations across the full aperture of the system. Once the wavefront distortion has been measured by the WS, the needed corrections are sent to the control system which apply the appropriate mirror deformation/movements. In the example of [Fig. 3](#) only the primary mirror (entrance pupil) is deformable and it is used for correcting spherical and astigmatism aberrations. M2 is used to compensate the decentering coma by means of lateral displacement. M3 may be tilted to correct for low-frequency pointing errors. In astronomical applications active optics is essential for telescopes with primary-mirror diameters larger than approximately four meters which cannot be manufactured rigidly at reasonable cost. Active optics requires a thin deformable or a segmented primary mirror maintained on the correct shape at every elevation. Using this method, low-frequency correction has been achieved to date in telescopes with primary mirrors of the order of ten meters.

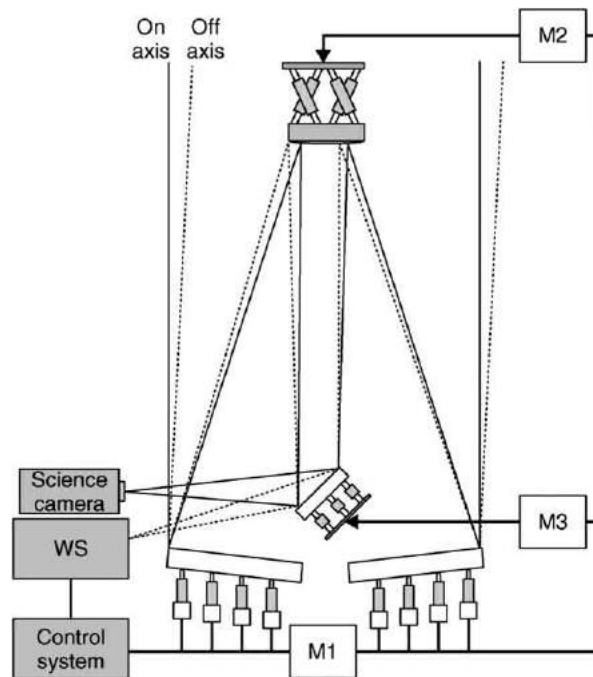


Fig. 3 A typical system for correcting low-frequency dynamical aberrations (active optics).

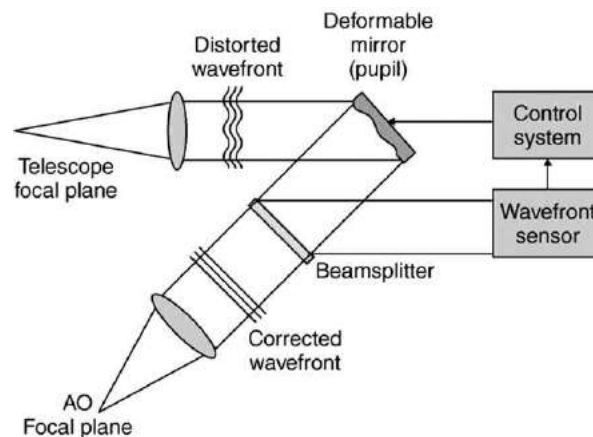


Fig. 4 A typical setup for correcting high-frequency dynamical aberrations (adaptive optics).

Fast Corrections: Adaptive Optics

It is clear that corrections at high frequency are impossible using the approach discussed above, because of the large dimensions and weight of the mirror and the low response time required. For an adaptive correction it is necessary to image a smaller system pupil in the optical train and make the active correction of the wavefront. A typical configuration is shown in Fig. 4. The light from the telescope focal plane (corrected for low frequency errors by means of active optics) is collimated and a pupil is located in this beam. Behind the active mirror a beamsplitter is used to send part of the light into the WS and measure wavefront distortions. The analysis results are sent to the control system which performs the active mirror deformation for the wavefront compensation. If the WS and the control system are sufficiently fast (typically faster than the time-variations of the wavefront distortion), this system is able to produce a plane wavefront which can be focused at the diffraction limit of the system. For this reason adaptive optics are often referred to as real-time system.

The seminal paper on the adaptive optics concept was written by H.W. Babcock of the Mount Wilson and Palomar observatories in 1953. Since then a great improvement has been made, especially in military and astronomical fields. In recent years, however, many applications of adaptive optics have developed in fields different from traditional ones, such as precision manufacturing by means of high-power lasers, telecommunications, aero-optics, medical physics, and ophthalmology.

We describe adaptive optics with particular reference to astronomical applications. We provide a brief description of other applications at the end of this article.

Imaging Through the Atmosphere

The optical effects of the atmospheric turbulence are due to local temperature variations that induce fluctuations in the refractive index of air. The refractive index changes depend on the air density as well as the range of the temperature variations. Air density is highest at sea level and decreases exponentially with altitude. Optical effects of turbulence decrease with the altitude of the site, the reason for locating astronomical telescopes on mountains. The effects of turbulence on a point source image are depicted in Fig. 5. Hereafter we consider an ideal telescope, for example, one without need for active correction at low frequencies. We consider the source 1 (continuous line). The radiation emitted by the source is a spherical wavefront that can be treated as a plane, because of the very long distance to the Earth's atmosphere.

In absence of wavefront distortion, the angular diameter of the image (Airy disk), is approximately $2.44\lambda/D$, with λ the wavelength of the observation and D the telescope aperture (the mirror diameter). In this case the image is said to be diffraction limited, and it is the best possible performance for the observation. This is the performance of a space-based ideal telescope.

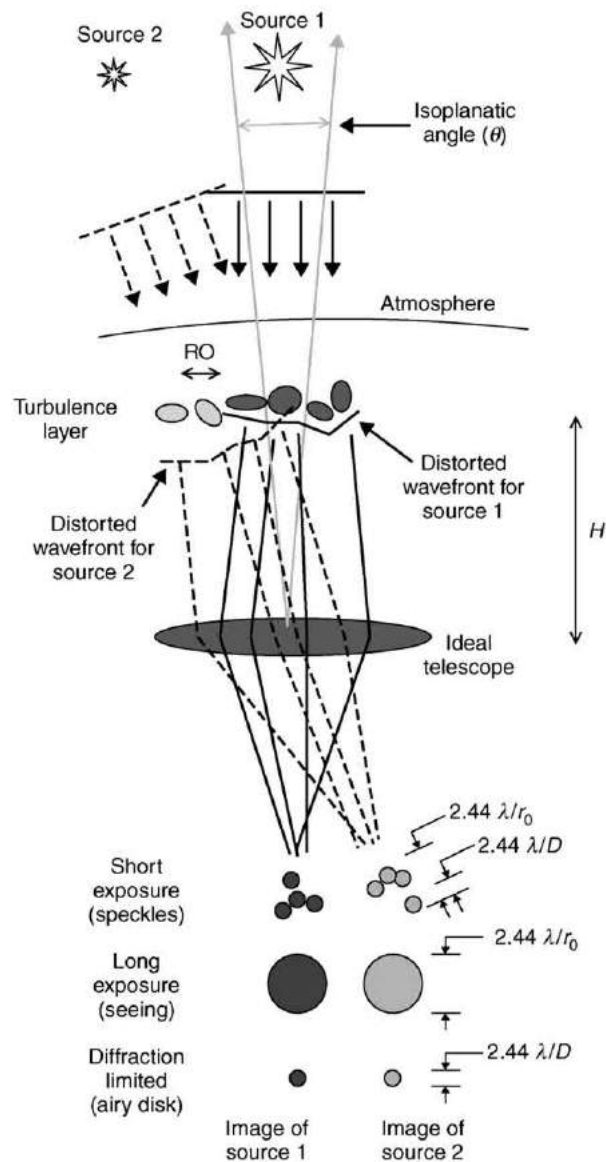


Fig. 5 Imaging through the atmosphere.

In the presence of the atmosphere the incoming wavefront is distorted by the turbulence and when it reaches the telescope aperture it is no longer a plane (see in Fig. 5). The so-called Fried's parameter r_0 (also called atmospheric coherence length) is the most common parameter used to quantify the distortion effects of atmospheric turbulence. It is defined as the distance over which the wavefront phase is highly correlated (in this way it provides a measure of the turbulence scale size). For an undistorted plane wave, r_0 is infinity, while for atmospheric distortion it typically varies between 5 cm (bad sites) and 20 cm (excellent sites). We depict Fried's parameter schematically in Fig. 5, as bubbles of dimension r_0 in the turbulent layer (referred to as Fried's cells). r_0 is a statistical parameter, which depends on time and wavelength ($r_0 = r_0(t, \lambda^{6/5})$). Its variability with time may be considerable, sometimes changing by a factor of 2 within a few seconds. A small telescope, with a diameter less than r_0 , will always operate at this diffraction limit. When the r_0 is smaller than the telescope aperture, the angular size of the image is different if we consider long- or short-time exposures. If the exposure time is very short (less than hundredths of a second) it is possible to image a structure in the focal plane (see short exposure image in Fig. 5). This structure, which varies with time, is due to high-order wavefront distortions that occur within the aperture. The short exposure image is broken into a large number of speckles of diameter $2.44\lambda/D$, which are randomly distributed over a circular region of approximately $2.44\lambda/r_0$ in angular diameter. A lower-order wavefront distortion is correlated with the size of the largest disturbances produced by atmospheric turbulence (the so-called outer scale) which varies considerably (from 1 m to over 1 km). Such disturbances produce a time-dependent tilt of the overall wavefront that causes fluctuations in the angle of arrival of the light from a star and randomly displace the image. For D of the order of the Fried parameter r_0 , image motion dominates, and a simple tip-tilt mirror (not just deformable) provides useful correction. For D greater than r_0 , speckles dominate and a deformable mirror is necessary to account for the wavefront distortion. For long exposures the dimension of the image is determined by the ratio λ/r_0 rather than $2.44\lambda/D$. The angular resolution is diminished due to the atmospheric turbulence (see Fig. 5). For a 4 m telescope, atmospheric distortion degrades the spatial resolution by more than one order of magnitude compared to the diffraction limit, and the intensity at the center of the star image is lowered by a factor of 100 or more. Even at the best astronomical sites, ground-based telescopes observing at visible wavelengths cannot, because of atmospheric turbulence alone, achieve an angular resolution better than telescopes of about 20 cm diameter.

Most of the distortions imposed by the atmosphere are in the wave-phase and an initially plane wavefront traveling 20 km through the turbulent atmosphere accumulates, across the diameter of a large telescope, phase errors of a few micrometers.

Turbulence is distributed along the propagation path through the atmosphere, and then the wavefront errors become uncorrelated as the field angle increases. This effect is known as angular isoplanatism and is one of the major limitations to the astronomical adaptive optics. The structure becomes completely uncorrelated out of the isoplanatic angle, approximately equal to r_0/H , where H is the altitude of the turbulence layer. This decorrelation is sketched in Fig. 5, where the short exposure images for objects 1 and 2 have different speckles structure.

In Earth's atmosphere significant turbulence occurs at altitudes up to 20 km, often with large peaks in the vicinity of the tropopause, at around 10 km there exist, in practice, more than one parameter H in Fig. 5, one for each turbulence layer. This three-dimensional distribution of the turbulence considerably restricts the angle over which adaptive compensation is effective and leads to the design of multiconjugate adaptive optics, as we next explain.

Adaptive System Design

As stated above, an active system is composed of at least one wavefront sensor and one deformable mirror. Typically, the wavefront is corrected by making independent tip, tilt, and piston movements of the deformable mirror surface, as shown in Fig. 6. As a rule of thumb, a deformable mirror used for atmospheric correction should have one actuator for each Fried's cell, such that the total number of actuators should be equal to the number of Fried's cells filling the telescope aperture, i.e., D^2/r_0^2 . For instance, a near-perfect correction for an observation in visible light ($0.6 \mu\text{m}$) with an 8 m telescope placed in a good astronomical site ($r_0 = 20 \text{ cm}$), would require a deformable mirror with approximately 6400 actuators in its adaptive optics system. Recalling that r_0 is strongly dependent on the wavelength (i.e., proportional to $\lambda^{6/5}$), similar performance in the infrared at $2 \mu\text{m}$ requires only 250 actuators. This is one reason why adaptive correction is easier in the infrared. A large number of Fried's cells requires a similarly large number of subapertures in the wavefront sensor. Commonly used WSs are based on modifications of the classic Hartmann screen test or of curvature type. In the former case the adaptive device is made with individual piezoelectric actuators (see Fig. 6), while in the latter the correction is achieved with a bimorph adaptive mirror, made of two bonded piezoelectric plates. With both

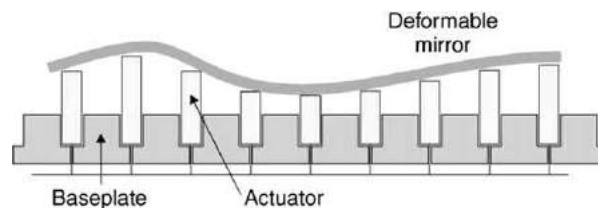


Fig. 6 A sketch of a deformable mirror.

methods, wavefront sensing is done on a reference off-axis object (Fig. 3), or even on the observed object itself (Fig. 4) if it is bright enough. In the former case the angular distance of the reference object must fall inside the isoplanatic angle and must be sufficiently luminous. In typical astronomical sites the isoplanatic angle at a wavelength of $0.6\ \mu\text{m}$ is $\sim 5\ \text{arcsec}$, while it can be $\sim 20\ \text{arcsec}$ at $2\ \mu\text{m}$. This is an additional reason why the adaptive optics system is most efficient in the infrared range. Correction in the visible requires a reference star approximately 25 times brighter than in the infrared. Most current astronomical systems are designed to provide diffraction-limited images in the near-infrared ($1\ \text{to}\ 2\ \mu\text{m}$) with the capability for partial correction in the visible. However, some military systems for satellite observations do provide full correction in the visible on at least $1\ \text{m}$ class telescopes.

The control system is generally a specialized computer that calculates, from the WS measurements, the commands to send the actuators of the deformable mirror. The calculation must be performed quickly (within $0.5\ \text{to}\ 1\ \text{ms}$), otherwise the state of the atmosphere may have changed making the wavefront correction inaccurate. The required computing power can exceed several hundred million operations for each command set sent to a 250 actuator deformable mirror. As in active optics systems, zonal or modal control methods are used. In zonal control, each zone or segment of the mirror is controlled independently by wavefront signals that are measured for the subaperture of that zone. In modal control, the wavefront is expressed as the best-fit linear combination of polynomial modes of the wavefront distorted by the atmospheric perturbations.

Future Development

As explained above, a convenient method for measuring the wavefront distortion is to look for an off-axis reference star (also known as natural guide star, NGS) within the isoplanatic angle. However, it is generally impossible to find a bright reference star close to an arbitrary astronomical target. Conditions are far better in the infrared than in the visible because atmospheric turbulence (and in particular its high spatial frequencies) is, for a given AO correction, less pronounced at longer wavelengths. The spatial and temporal sampling of the disturbed wavefront can be reduced, which in turn can permit the use of fainter reference stars. Coupled with the larger isoplanatic angle in the infrared, this gives a much better outlook for AO correction than in the visible.

Nevertheless, even for observations at $2.2\ \mu\text{m}$, the sky coverage achievable by this technique (equal to the probability of finding a suitable reference star in the isoplanatic patch around the chosen target) is of the order of $0.5\ \text{to}\ 1\%$. It is therefore unsurprising that most scientific applications of AO have to date been limited to the study of objects which provide their own reference object. The most promising way to overcome the isoplanatic angle limitation is by the use of artificial reference stars, also referred to as laser guide stars (LGS). These are patches of light created by the back-scattering of pulsed laser light of sodium atoms in the high mesosphere or by molecules and particles located in the low stratosphere. The laser beam is focused at an altitude of about $90\ \text{km}$ in the first case (sodium resonance) and $10\ \text{to}\ 20\ \text{km}$ in the second case (Rayleigh diffusion). Such an artificial reference star can be created as close to the astronomical target as desired, and a wavefront sensor measuring the scattered laser light is used to correct the wavefront aberrations on the target object.

Military laboratories have reported the successful operation of adaptive optics devices at visible wavelengths using laser guide star on both a $60\ \text{cm}$ telescope and a $1.5\ \text{meter}$ telescope, achieving an image resolution of $0.15\ \text{arcsec}$. Nevertheless, there are still physical limitations of a LGS. The first problem, focus anisoplanatism, or cone-effect, was realized long ago. Since the artificial star is created at a relatively low altitude, back-scattered light collected by the telescope forms a conical beam, which does not cross the same turbulence-layer areas as light from the astronomical target. This leads to a phase estimation error, which in principle can be solved by the simultaneous use of several laser guide stars surrounding the targeted object. A more significant problem is image motion or tilt determination. Since the paths of the outgoing and returning light rays are the same, the centroid of the artificial light appears to be stationary in the sky, while the apparent position of an astronomical source suffers lateral motions (also known as tip/tilt). The simplest solution is to supplement the AO system and LGS with a tip/tilt corrector set on a (generally) faint close NGS. Performance is then limited by poor photon statistics in the tip/tilt correction.

Adaptive optics with a multicolor laser probe is another concept that has been studied to solve the tilt determination problem of laser beacon based AO. Only applicable to sodium resonant scattering at $90\ \text{km}$, it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength. Its main drawback is the limited returned flux, due to the saturation of mesospheric sodium layer.

The quality of a site for adaptive optics is determined not only by the median seeing, but also by the height and thickness of the turbulent layers (often only a few thin layers dominate). Some adaptive-optics systems can be adapted to correct for turbulence originating at a particular height by using several reference stars. This technique is known as multiconjugate adaptive optics (MCAO) and is still in its infancy. This concept deals with a 3-D view of the atmosphere, and for this reason it is also called turbulence tomography. Several wavefront sensors are needed for tomography. Turbulence tomography is also useful to correct for the cone effect when using LGS. When several deformable mirrors are placed on the appropriate conjugate planes, as shown in Fig. 7, anisoplanatism reduces and the size of the corrected field of view increases. The situation is roughly depicted in Fig. 7, where each turbulent layer (indicated as TL) has a proper conjugated plane (C-TL).

Another challenging optical solution is one of utilizing large secondary adaptive mirrors. This application seems especially interesting at thermal wavelengths, where any additional mirror raises the significant thermal background detected by instruments.

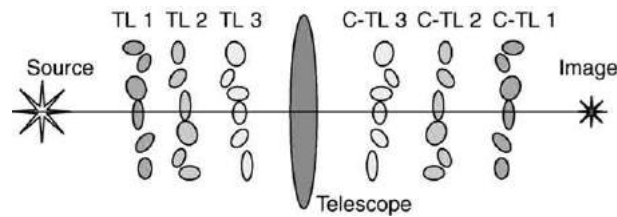


Fig. 7 Conjugated planes of turbulent layers.

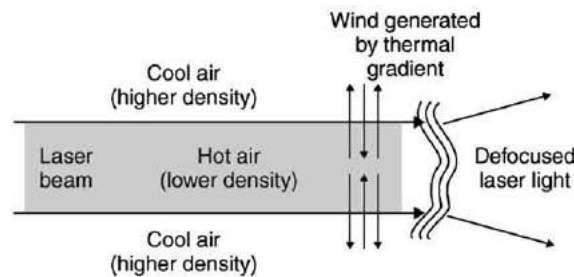


Fig. 8 Thermal blooming effect in high-power laser applications.

Other Adaptive Optics Applications

In recent years the adaptive optics technique has been developed in many applications beside astronomy and the military, for example, as in high-power laser welding or cutting, in ophthalmology and in aero-optics.

In high-power laser applications, thermal blooming takes place when laser light is absorbed by the gases and particles that make up the atmosphere. Since energy is conserved, the energy lost by the laser (its intensity diminishes) is gained by the atmosphere that heats up. This inhomogeneity at the boundary of the laser beam generates wavefront distortions by turbulence, as shown in [Fig. 8](#). This means that laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. The net effect is that, in the case of a Gaussian laser beam, light following a wavefront which passed at an earlier time, 'sees' a concave lens that deflects the successive rays out of the beam (away from the axis) causing the beam to appear to grow in radius or 'bloom'. Adaptive optics techniques for real-time compensation of blooming in laser beams were developed in the late 1960s. Given the high coherence and monochromaticity of a laser, the technique is more straightforward than for more common white-light applications of the astronomical applications.

Density fluctuations in a compressible shear layer are sufficiently high to cause a time varying index of refraction of the layer. This situation is common in a turbulent flowfield, and the study of the wavefront aberrations induced by this flowfield, when it is of relatively short propagation length in the near-field of the optical system, is referred to as aero-optics. Turbulent flowfield induces various types of degradations in the performances of an optical system. For example the jitter and defocusing of laser-based weapons cause a reduction in the amount of energy delivered to the target. Tracking systems experience bore-site errors and imaging systems experience blurring. Recent work suggests that adaptive optics compensation may successfully overcome these problems.

In ophthalmology the adaptive technique finds application in studying the retina of the eye. In this case the eye itself produces wavefront aberration, and the retina is seen as blurred. Techniques in ophthalmic laser surgery are being developed to use adaptive optics to correct for detected distortions.

In the field of telecommunications, studies are in progress for using adaptive techniques in data transmission between ground-based stations and satellites.

See also: Nonlinear Optical Phase Conjugation. Phase Conjugation and Image Correction

Further Reading

- Beckers, J.M., 1993. Adaptive optics for astronomy: principles, performance and application. *ARAA* 31, 13–62.
 Fried, D.L., 1965. Statistics of a geometric representation of wavefront distortion. *JOSA* 55, 1427–1435.
 Hardy, J.W., 1998. *Adaptive Optics for Astronomical Telescopes*. New York: Oxford University Press.
 Jumper, E.J., Fitzgerald, E.J., 2001. Recent advances in aero-optics. *Progress in Aerospace Sciences* 37, 299–339.
 Liang, J., Williams, D.R., Miller, D.T., 1997. Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSA A* 14 (11), 2884–2892.

- Lukin, V.P., 1995. *Atmospheric Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.
- Ragazzoni, R., Marchetti, E., Valente, G., 2000. Adaptive-optics corrections available for the whole sky. *Nature* 403 (6765), 54–56.
- Roddier, F., 1999. *Adaptive Optics in Astronomy*. Cambridge: Cambridge University Press.
- Roggeman, M., Welsh, B., 1996. *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.
- Tyson, R.K., 2000. *Introduction to Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.
- Welford, W., 1974. *Aberrations of the Symmetrical Optical System*. New York: Academic Press.
- Wilson, R.W., Jenkins, C.R., 1996. Adaptive optics for astronomy: theoretical performance and limitations. *MNRAS* 268, 39.

Phase Conjugation and Image Correction

EN Leith, University of Michigan, Ann Arbor, MI, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Phase conjugation of a wave is a process that has long been recognized. However, the technique for producing high-quality phase conjugation is relatively new. Also, the principal use of phase conjugation, the correction of phase errors in waves that have propagated through phase-distorting media, is relatively new, having originated in holography in the 1960s. Today, phase conjugation is an important area of modern optics. This article describes the conjugation of a wavefield by the process of holography and the holographic phase conjugation method of removing phase distortions from a wavefield, thereby permitting high-quality images to be produced from phase-distorted waves.

The basic meaning of a conjugated wave is explained with the aid of Fig. 1. Fig. 1(a) shows a wavefield propagating to the right. For simplicity, we show only one wavefront, i.e., one contour of constant phase. It is understood that a wavefield consists of a sequence of such wavefronts, with each wavefront differing in phase by 2π from the adjacent one, and separated by the wavelength λ . Of the two points, A and B, on the wavefront, we say that the wavefront at A lags that at B, since when the wavefront crosses a plane normal to the direction of travel, point A of the wavefront crosses that plane behind point B. The wavefield produces, at the plane P, a field $ae^{i\phi}$, where a is the amplitude and ϕ is the phase of the wavefront at the plane P.

We place a phase conjugator in the path of the wave. The wave passes through the conjugator, emerging as a phase conjugated wave, in which the point B, which formerly led point A in phase, now lags behind it by a corresponding amount. This illustrates the basic meaning of the term 'phase conjugation'. This device could be considered as a forward phase conjugator, since the phase conjugated wave travels in the same direction as the original wave.

Next (Fig. 1(b)) we describe a backward phase conjugator. The conjugate wave travels in a direction opposite to that of the incident wave. The conjugate wave emitted by the phase conjugator seems to be identical to the incident wave; we show the two waves to be in coincidence, or more precisely, for purposes of clarity, almost in coincidence. The points A on the two wavefronts coincide, as do the points B. As before, on the incident wave, point A lags behind point B. On the exiting wave, since the wave is traveling in the opposite direction, point B is now lagging point A, i.e., is lagging in phase. In either case, forward or backward conjugation, the wave emerging from the conjugator has the complex amplitude $a(x, y)e^{-i\phi(x, y)}$.

A phase conjugator (backward variety) is like a mirror, in that it obeys the basic mirror law, that the angle of reflection equals the angle of incidence, but in a rather different way, as Fig. 2 shows. In fact, the reflected ray travels back on itself. A simple way to generate a conjugate wave, and one that has been known for a very long time, is with the use of an array of retroreflectors, or corner cubes. As shown in Fig. 3, each ray undergoes two reflections and emerges traveling in exactly the opposite direction, although with some displacement. If we considered the third dimension, there could be three reflections. The amount of displacement depends on where the ray strikes; if it strikes near the apex of the corner cube, the ray displacement will be slight; if at the edge, the displacement is greatest. Thus, the image formed in reflection is degraded. This geometrical degradation is greater for larger corner

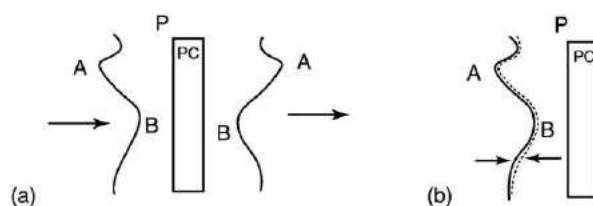


Fig. 1 A phase conjugator. (a) shows a forward phase conjugator; (b) shows a backward phase conjugator. For clarity, in (b) the conjugated wave is shown with dotted lines. PC is phase conjugator, and P is the entrance plane of the conjugator.

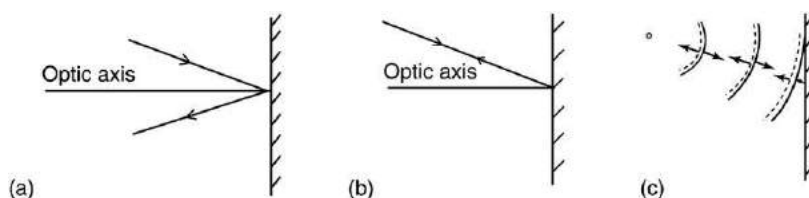


Fig. 2 Light reflected from mirrors. (a) Light falling on a conventional mirror and being reflected; (b) light reflected from a conjugating mirror; (c) as in (b), except that illustration is with wavefronts instead of rays.

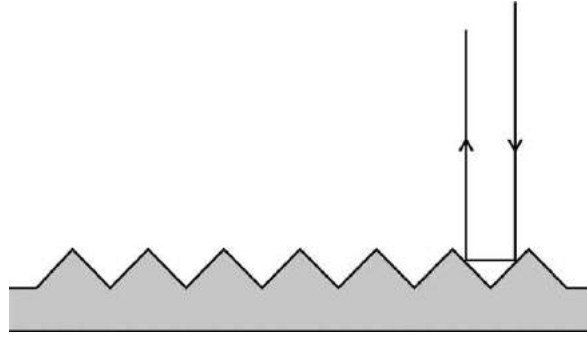


Fig. 3 A ray of light reflected from a retroreflector (or corner cube).

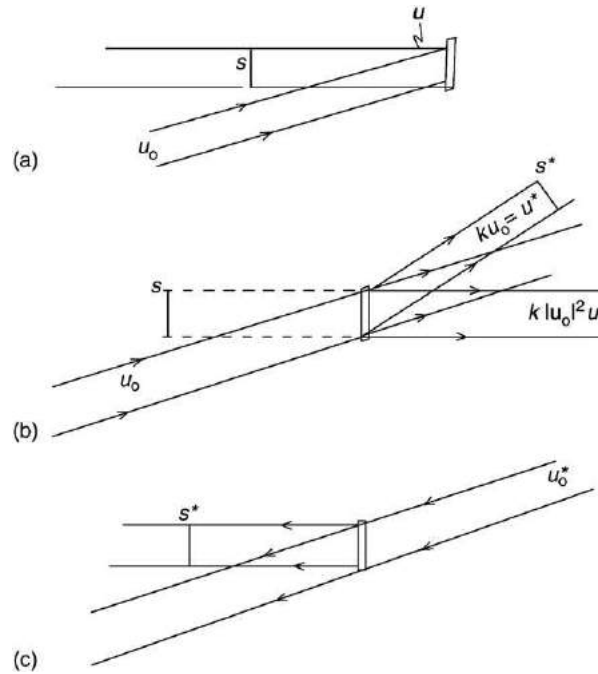


Fig. 4 The holographic process. (a) Making the hologram; (b) reading out the hologram, using a readout beam that duplicates reference beam; (c) reading out with reversal of reference beam.

cubes. Thus, why not make the cubes very small? Then, however, the image is degraded by diffraction effects, since each corner cube constitutes a small aperture. Thus, there is an optimum size for the cubes; if too large, geometrical factors degrade the image; if too small, diffraction effects degrade the image.

Phase Conjugation by Holography

With holography, however, came a method of conjugating a wavefront that yields high-resolution imagery. The holographic process inherently produces a phase conjugate image, as was noted by Gabor in his early papers on holography. The holographic process is illustrated in Fig. 4. For simplicity, we assume the object to be a planar transparency $s(x, y)$. A coherent light wave, generally a plane or spherical wave, passes through the transparency and a diffracted field u is produced at the recording plane P , as in Fig. 4(a). The field u is typically the Fresnel diffraction pattern of the object s , but if the propagation distance is sufficiently great, u could be the Fraunhofer pattern of s . Of course, lenses can be placed between the object and recording planes, thus modifying the diffraction pattern.

The field u is combined with a reference wave u_o , which is generally an off-axis planar or spherical wave. The two beams produce an interference pattern with intensity

$$I = |u_o + u|^2 = |u_o|^2 + |u|^2 + 2\text{Re } u_o^* u \quad (1)$$

where the final term can be written as $u_o^*u + u_o u^*$; the $*$ denotes complex conjugate. The process of forming the interference of the two beams, i.e., forming the intensity of their sum, has generated the conjugate term u^* . When the intensity is recorded, as on photographic film, for example, both terms are stored on the film. Suppose for simplicity that u_o is a plane wave $a_o e^{i2\pi f_o x}$, where a_o is a constant, and $f_o = (\sin \theta)/\lambda$, where θ is the angle that the reference wave makes with the optic axis and λ is the wavelength of the light. Also, u can be written as a product of an amplitude term a and a phase term $e^{i\phi}$, or $u = a e^{i\phi}$. Let the photographic plate, after exposure and development, have a transmittance t proportional to the intensity I , or $t = kI$ with k a constant; this is an oversimplification of the recording process, but preserves the essence of the holographic process. Now let the photographic plate, which after development becomes a hologram, be illuminated with a readout wave that, for simplicity, we let be u_o , a duplicate of the reference wave. Then, the field emerging from the hologram is

$$u_e = u_o k [|u_o|^2 + |u|^2 + u_o^* u + u_o u^*] \quad (2)$$

Only the last two terms are of interest. These can be written as $k|u_o|^2 u$ and $ku_o^* u^*$, respectively, or alternatively, as $ka_o^2 a e^{i\phi}$ and $ka_o^2 a e^{i(4\pi f_o x - \phi)}$. The term $k|u_o|^2 u$ represents a wave that is, to within a constant, identical to the object wave u ; the other term represents a wave that is, to within a constant and a linear phase term, a conjugate of the original wave. The factor $e^{i4\pi f_o x}$ has the physical meaning that the wave is propagating in a direction 2θ given by $\sin 2\theta/\lambda = 2f_o$. The readout waves are shown in Fig. 4(b).

The so-called true wave u and the conjugate wave u^* thus propagate angularly away from each other and from the direct transmitted beam, which carries the uninteresting components $|u_o|^2$ and $|u|^2$.

However, a more suitable way to form the conjugate wave is to illuminate the hologram with a readout wave that is, again, a duplicate of the original reference wave, but propagating in the opposite direction, as in Fig. 4(c). The conjugate wave is now propagating along the z axis, along the path of the original object wave. It is readily shown that this counterpropagating wave is identical in shape to the original object wave recorded by the hologram, except that it travels in the opposite direction, exactly the situation portrayed in Fig. 2. This wave propagates so as to form an image that is, to within a constant, identical to the original object, except for the counterpropagation, which produces a conjugate image $s^*(x, y)$. The image intensity, however, is $|s|^2$, which is identical to the original image intensity.

This conjugate wave was, in the early days of holography, a detriment, since without the use of the off-axis method (reference wave introduced at an angle to the object wave) the two images were inseparable, and the true image could be observed only against the background of the defocused conjugate image. Thus, an out-of-focus image overlaid the true image, making it noisy.

In 1962, the same year that the off-axis reference wave for separating the two images was introduced, a possible use for the conjugate image was suggested. If the original propagation path of the object wave were aberrated, as would occur if the wave propagated through an irregular medium, then the conjugate wave propagating through the medium would undergo a correction and would arrive at the original object position perfectly corrected, resulting in a sharp image, a process sometimes called wavefront healing. The idea lay dormant for 3 years, until in 1965, experimental demonstration of this idea was reported. In one case, the object was a transparency containing opaque lettering against a transparent background and the inhomogeneity was a piece of frosted glass, which resulted in extremely severe aberrations, so that the object wave, having traversed the diffuser, was no longer capable of forming an image. However, when the distorted wave was recorded holographically, conjugated, and sent back through the same diffuser, the result was a highly legible image.

The principle of wavefront healing is illustrated in Fig. 5, when the inhomogeneity is an extended, or three-dimensional, structure. A point object O sends a spherical wave into the inhomogeneity. As the wavefront propagates, it becomes distorted, and loses its sphericity. When it emerges from the inhomogeneity, the wavefront is highly distorted. The wave is conjugated at the conjugator C and emerges as a back-propagating wave that retraces the original wave path through the inhomogeneity, gradually losing its aberrations and looking more like a spherical wave. When it emerges from the inhomogeneity, it is a perfect spherical wave that converges to a point image that coincides with the original object. At each plane throughout the inhomogeneity, the back-propagating conjugate wave (dashed lines) has the same shape as did the original wave when it crossed that plane (solid line).

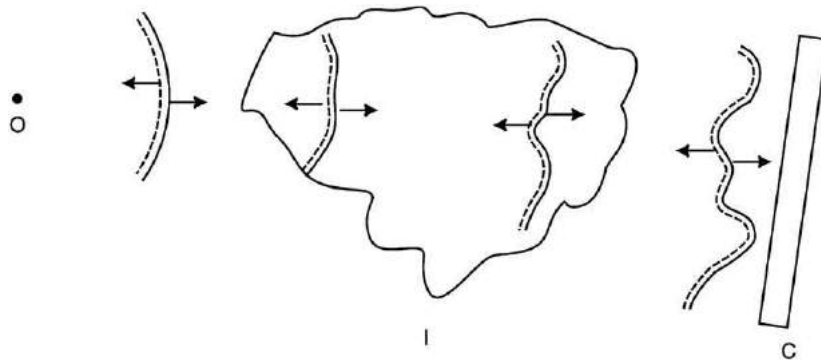


Fig. 5 The wavefront healing process. C is the conjugator, I is the inhomogeneity, and O is a point object.

This interesting wavefront healing property of holographic phase conjugation was the beginning of a new branch of optics, called phase conjugation, giving rise to a number of applications. However, the holographic method had the problem that it did not occur instantaneously, but had to await the processing of the photographic film on which the hologram was recorded. This delay limited the range of potential applications. Later, in the 1970s, nonlinear crystals became available that did the equivalent of holography, but in real time, or near real time – a process known as four wave mixing. There are yet other types of real-time phase conjugators with nonlinear devices.

Pseudoscopic Image

As long as the object is only two-dimensional, a function of x and y only, with no extent along the axial dimension, z , the intensity distribution of the true and conjugate images is identical. The conjugateness, being a pure phase phenomenon, disappears when the observable is the image intensity $|s|^2$, and not the complex amplitude s . Since the intensity is always the observable, whether the image is viewed by eye, or with a camera, the conjugateness generally is of no significance.

However, when the object for holography is three-dimensional, the situation is different, and the conjugate image exhibits some strange properties. These arise from conflicting cues pertaining to the three-dimensionality of the image. To explain, we review some basics about how a three-dimensional world is perceived by a viewer. First, there is the stereo effect, resulting from the viewer having two eyes, with each eye seeing an object from slightly different positions, a phenomenon called binocular disparity. The mind uses this stereo effect to perceive three-dimensionality. But there are other cues that also contribute to perception of three-dimensionality. Another is parallax; as the viewer moves his head, objects farther from the observer change their angular position more slowly than objects close by. The observer can sense the three-dimensionality of an object, even with only one eye, by moving the head, thereby inducing parallax. Also, there is interposition, whereby near object portions block the viewing of object portions lying behind them, again giving a sense of depth. There are still other cues, more subjective, such as perspective, parallel lines converging toward infinity, shadowing, etc. With these observations in mind, we consider the hologram under illumination and the two images that it forms (Fig. 6). As depicted in Fig. 6, these two images lie in mirror symmetrical positions relative to the hologram. Both images have the same side facing the hologram. It must be this way, because when the hologram was recorded, it was side *a* that faced the recording plane. And both images must therefore have side *a* closer to the hologram.

When a viewer observes the true image (a virtual image), he views it looking through the hologram, and he sees the image at the position where the original object was when the hologram was recorded, and it is, in its visual properties, indistinguishable from the actual object; it has all of the three-dimensional and the parallax properties that the original object would exhibit. However, when the viewer observes the conjugate image, he must view it from the side of the image opposite from the hologram, i.e., he views the image from the back side. When a viewer observes the conjugate image, he sees it from the back side (side *b*). All of the above cited cues are in agreement about the orientation of the image, viz., that side *b* is closer to the observer than is side *a*. That is, all cues but one. The conjugate image is formed as the hologram recording plate saw it, with side *a* being closer. Thus, all interpositions are formed with the same relation that the object had with respect to the recording plate. The side *a* is the part of the object that was closer to the hologram, so that when interpositions occur, side *a* blocks side *b*. Since the observer thus sees side *a* partially obscuring side *b*, he must conclude that side *a* is closer than is side *b*. However, this observation is in direct conflict with the other cues, especially the stereo and the parallax cues, which are themselves strong cues, as is the interposition cue. The conflict of these cues, with two of them interpreting the image as being formed with one depth orientation, the other with the opposite, gives a number of strange effects. For example, as the viewer moves his head so as to induce parallax, the object appears to rotate. If the image is that of a human head, for example, the observed effect is that as the viewer changes his position, the head appears to

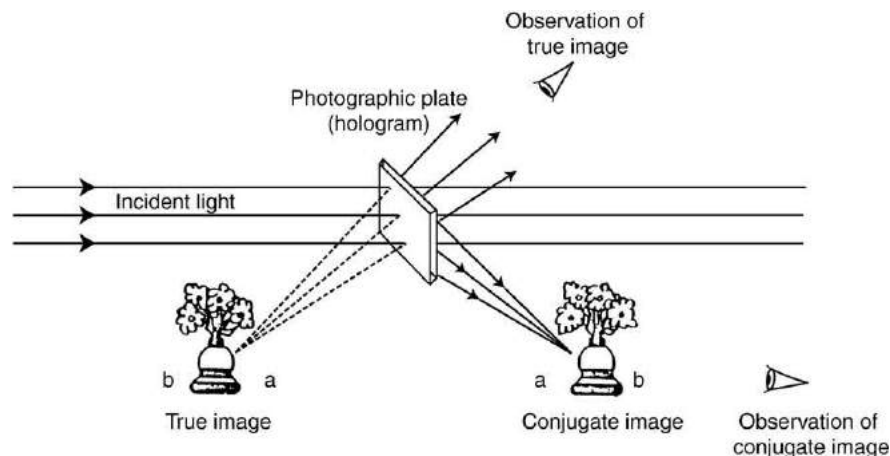


Fig. 6 Formation of the true and conjugate images from a hologram.

rotate in synchronism with the observer, as if the image had its gaze fixed on the viewer, an effect that is often perceived as being eerie.

Similarly, if the object were two flat plates, (Fig. 7) in the situation depicted, the planes have no overlap, i.e., no interposition, as seen by the observer. And the observer unequivocally perceives plate A as being closer than plate B. However, if the viewer moves his head to a position where interposition occurs, the conflict of cues arises, A becomes partially blocked by B, which is quite correct as observed at the position of the hologram, but is quite incorrect from the viewer's vantage point. The result of this conflict is that the viewer, who previously perceived the three-dimensionality, now sees the three-dimensionality perception suddenly collapse, and the sense of three-dimensionality is lost. Other strange perception anomalies can also be produced. Such images, with conflicting depth cues, are called pseudoscopic, as opposed to orthoscopic images, where the cues do not conflict.

One Way Phase Conjugation

A significant limitation of phase conjugation aberration correction is that the image and object are on the same side of the inhomogeneity. One would usually want them to be on opposite sides. In a modified form of the phase conjugation process, shown in Fig. 8, this result is achieved. The first step of the process is carried out by passing a plane wave, instead of the object wave, through the inhomogeneity. This plane wave originates from a point source located at the focal plane of a collimating lens.

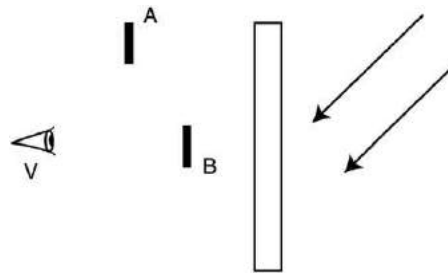


Fig. 7 Pseudoscopic conjugate image formed from an object consisting of two plates. The viewer V perceives plate A as being closer than B and there are no conflicting cues, since there is no interposition. However, if the observer moves his head upward, there is partial interposition, thus conflicting cues, and the sense of three-dimensionality is lost.

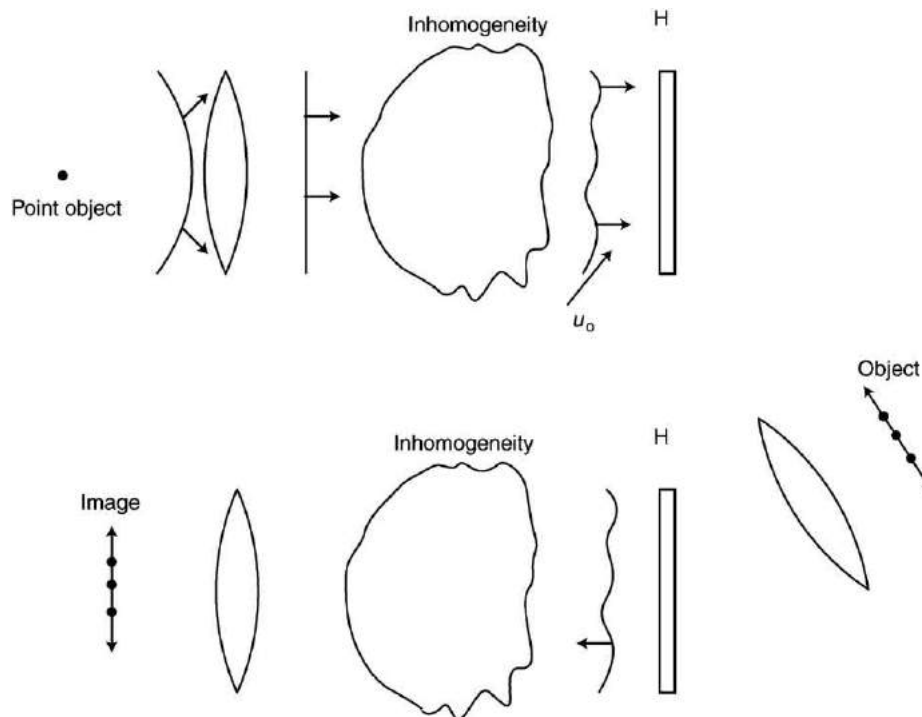


Fig. 8 One-way phase conjugation. Top: Making the hologram. Bottom: The viewing process. We show three points on the object. The center point forms a well-corrected image, but the other two points produce imperfectly corrected images.

The plane wave thus carries information about the inhomogeneity to the recording plate, where it is combined with a reference wave and a hologram of the distorted plane wave is recorded. Thus, information about the phase defect is stored on the hologram.

Next, the hologram is read out, not with a counterpropagating replica of the reference wave as before, but instead with an object-bearing wave that passes through the hologram. This wave emerges from the hologram and now has on it the recorded aberration information. As it counterpropagates along the path of the original object wave, it passes through the inhomogeneity, whereupon the aberration is cancelled and a corrected image forms at the focal plane of the collimating lens, where previously the point source had originated. The goal is achieved; the object and the corrected image are now on opposite sides of the inhomogeneity.

This correction is only approximate. To explain, we think of the object as being a collection of many points. The radiation from each point is converted into a plane wave by the lens, with different object points producing plane waves propagating at different angles. These plane waves then impinge on the hologram and emerge with each one bearing the aberration stored on the hologram. They continue on their paths and enter the inhomogeneity. The plane wave component, whose path direction matches that of the reference wave used in recording the hologram, will emerge from the inhomogeneity as a perfectly corrected wave, thus forming a well-corrected point image at the position of the original point radiator. However, another plane wave component, traveling in an incorrect direction, will be laterally displaced from the proper position for entering the inhomogeneity, and the aberration correction will be imperfect, the degree of imperfection being in direct relation to this displacement error. The displacement error is the product of the angular error and the propagation distance between hologram and inhomogeneity. If the inhomogeneity is slowly varying as a function of x and y , and if the object does not have a large extent, so that the range of angles is small, the image degradation will be small.

If the last two conditions do not apply, two options are available. First, the distance between inhomogeneity and hologram can be reduced, either by bringing the hologram closer to the inhomogeneity, or by imaging the hologram onto the inhomogeneity. The correction is now valid over a wider extent of the image, especially if the inhomogeneity can be treated as planar rather than extending over a volume. Alternatively, a scanning process can be carried out on the object beam falling on the hologram, so that in sequence, each plane wave component assumes the proper angle to counterpropagate along the path for perfect correction, and will form a well-corrected image point. This output spot then falls on a pinhole positioned to separate it from other, imperfectly corrected points. Then, over time, all points of the object will form well-corrected image points, and the entire object will thus be well-imaged. The one way phase conjugation method was first used to correct the spherical aberration of a lens, and the result was quite successful.

Phase conjugation has become an important branch of modern optics. The process is often carried out using photographic materials. More often nowadays, it is carried out using nonlinear crystals. Some of the crystal methods are essentially holography in real time, but others, such as phase conjugation by Brillouin scattering, go beyond holography.

See also: Adaptive Optics. Nonlinear Optical Phase Conjugation

Further Reading

- Caulfield, H.J. (Ed.), 1979. *Handbook of Optical Holography*. New York: Academic Press.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill, chap. 9. This chapter is devoted to holography, with section 9.12.3 dealing with phase conjugation aberration correction.
- Kogelnik, H., 1965. Holographic image projection through inhomogeneous media. *Bell Systems Technology Journal* 48, 2451–2452.
- Leith, E.N., Upatnieks, J., 1965. Holograms: their properties and uses. *SPIE Journal* 4, 3–6. This is the earliest publication demonstrating the principle of wavefront healing by holographic phase conjugation.
- Smith, H.M., 1975. *Principles of Holography*, 2nd edn. New York: Wiley.

Hyperspectral Imaging

ML Huebschman, RA Schultz, and HR Garner, University of Texas, Dallas, TX, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Hyperspectral imaging techniques on a microscopic level have paralleled the hyperspectral remote astronomical sensing which has developed over the years. It is especially apparent in remote sensing of our own planet's land masses and atmosphere, such as in the LANDSAT satellite program. In biology, biochemistry, and medicine, hyperspectral imaging and multispectral imaging are often interchanged, although there is a distinct difference in the two techniques.

A multispectral imager produces images of each of a number of discrete but relatively wide spectral bands. Multispectral imagers employ several wide bandpass filters to collect the data for the images. A hyperspectral imager produces images over a contiguous range of spectral bands and, thus, offers the potential for spectroscopic analysis of data. Hyperspectral imagers may use gratings or interferometers to produce the contiguous range of spectral bands to record.

Fluorescent dyes (also referred to as fluorophores or fluorochromes) and recently quantum dots, as markers, are a very powerful tool for analyzing organic tissues for researchers in cytology (study of cell function), cytogenetics (study of the human chromosomes), histology (study of structure and function of tissues), microarray scanning (analysis of concentrations of multiple varied probes), and biological material science. Without the correlation of the spectral and spatial data in biological images, there would be little useful information for the researcher. Therefore, hyperspectral imaging in the biological sciences must combine imaging, spectroscopy, and fluorescent probe techniques. Such imaging requires high spatial resolution to present a true physical picture, and the spectroscopy must also have high spectral resolution in order to determine the quantities of biochemical probes present. The correlation must provide the color-position relation for identification of meaningful information.

This description will begin with an initial review of multispectral imaging to lead into a discussion of hyperspectral imaging. The discussion will examine aspects of hyperspectral imaging based on an interferometric method and on a grating dispersion method similar to the methods in multispectral imaging.

Image Parameters

The analysis of fluorophores involves the excitation of the dye with a shorter wavelength and thus higher energy than the emission spectrum. The use of the appropriate excitation wavelength will yield a maximum emission intensity at a characteristic wavelength. The quality of a fluorescent image generally combines contrast and brightness. Contrast depends on the ratio of emission intensity to that of background light. Consequently, the use of filters to block the excitation wavelength and prevent it from overwhelming the field is critical in providing high contrast for the detection of fluorophore emission. However, the use of narrow filters to restrict the background can severely limit the brightness of the fluorescent image. This is due to the barrier filters allowing only a portion of the actual emission spectrum to pass. Brightness is an intrinsic property of each individual fluorophore, but is also dependent on the ability to illuminate with the peak excitation wavelength for that fluorophore.

The balance between contrast and brightness may be further complicated by applications designed to facilitate the evaluation of multiple fluorophores. The selection of appropriate excitation and emission filters within a single set to allow for the simultaneous viewing of multiple fluorophores can be very inhibiting. Although quad-band beamsplitters are currently available, commercial advances to date have yielded only triple-pass combinations for simultaneous detection, such as one of which permits the simultaneous detection of the widely used fluorophore combination fluorescein-5-isothiocyanate (FITC), Texas Red and 4',6-diamidino-2-phenylindole (DAPI) (e.g., Chroma Technology Corp.).

Although the capture of digital images has propelled this field, efforts to collect multicolored images as digital files also suffer from technical limitations. Color CCD cameras often utilize filters or liquid crystals to segregate the red, green, and blue wavelengths onto the imaging chip. The employment of these cameras is restricted to resolving colors as discriminated by the filters used. For scientific applications, then, the greatest sensitivity and selectivity has been achieved through black and white CCD cameras with optimized filters for a specific application. While this approach is widely used, it suffers from the inherent limits of the spectra of data that can be analyzed since it is defined by the characteristics of the optical filters employed. Thus, the system is restricted to the use of fluorophores that can be effectively resolved from each other by filters, and the filters selected limit the data collected to a portion of the total emission spectrum for any one fluorophore.

Multispectral Imaging

Multispectral imaging systems in biomedical research are composed of a microscope, filters and/or interferometer for spectral selection, a charged coupled device (CCD) camera to record the data, and computer control for acquisition, analysis, and display

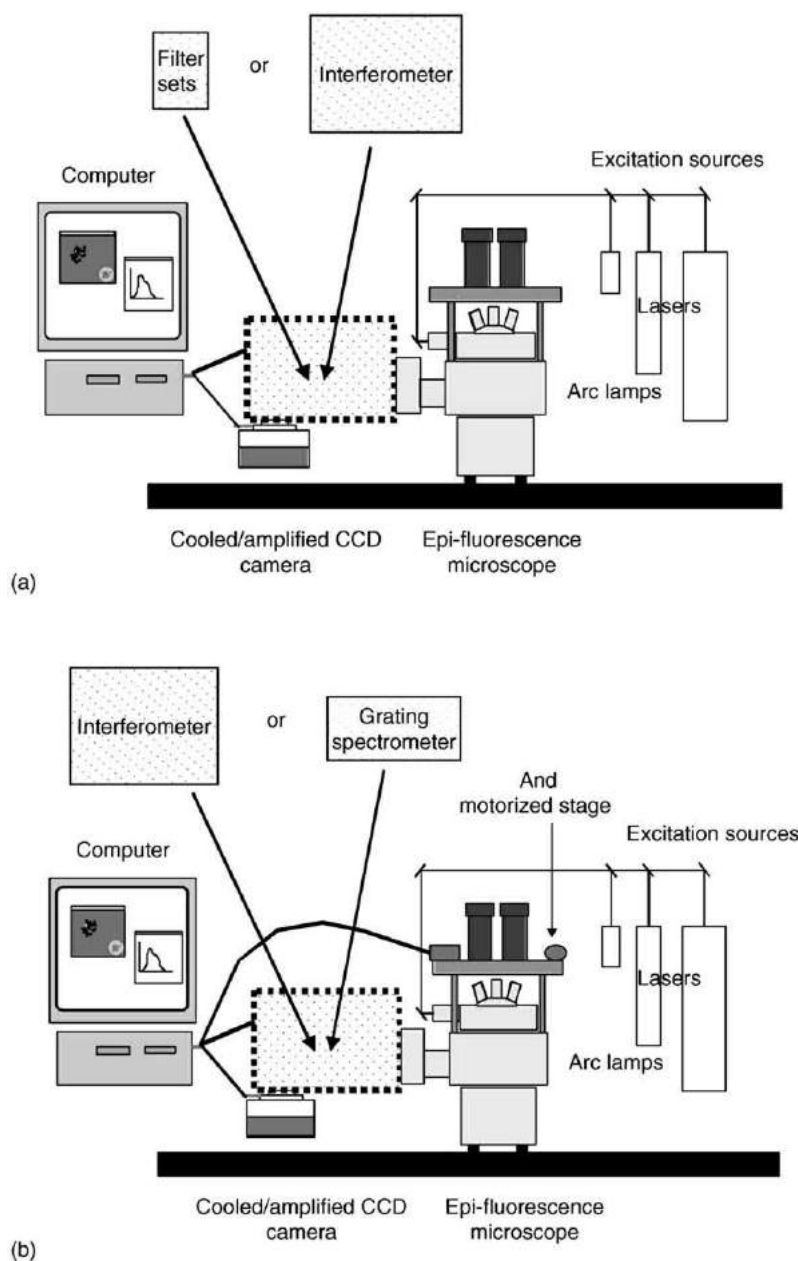


Fig. 1 The instruments for multispectral and hyperspectral imaging are similar. (a) Multispectral imagers use interferometers or bandpass filters to collect noncontiguous emission wavelengths. (b) Hyperspectral imagers employ interferometers or gratings to collect a contiguous wavelength band. Computers, microscopes, CCD cameras, and excitation light sources are common to all the techniques. The systems which use filter sets or interferometers use motorized filter placement systems or automated interferometers. The gratings systems employ a motorized stage for the sample movement. Adapted with permission from Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

(Fig. 1(a)). The introduction of CCD cameras has offered a means to document results and capture images as digital computer files. This has facilitated the analysis of results with more and more complex probe sets and has allowed the use of more sophisticated computer algorithms for data analysis. The other major component of the multispectral imaging system is a light source to excite the sample's fluorescent dyes whose emissions are selected by filters or an interferometer for recording on the CCD camera.

An example of the use of multispectral imaging is in the identification of the 24 human chromosomes. A scheme was presented for the analysis of human chromosomes by differentially labeling each chromosome with a unique combination of 5 fluorophore-labeled nucleotides (building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group). These complex probes are simultaneously hybridized to metaphase chromosomes and analyzed to determine which of the

fluorophores are present. Combinations of one, two, or three of the five fluorophores are enough to distinguish a particular chromosome from the others.

An interferometric multispectral system used to perform this type of analysis is the Spectral Karyotyping (chromosome typing) or SKY System (Fig. 1(a)). This technology does not represent hyperspectral imaging, but is presented for comparison. Illumination light passes through a triple band-pass dichroic filter which allows for the simultaneous excitation of all the fluorochromes in, for instance, Speicher's scheme. Modern multi-band pass filters add little to image distortion while effectively blocking unwanted autofluorescence or excitation illumination. The fluorescence emissions from the sample are split into two separate beams by the interferometer. The beams are then recombined with an optical path difference (OPD) between them such that the two beams are no longer in register and interfere with each other. By adjusting the OPD, a particular wavelength band can be selected for, which results in constructive interference and thus can be collected by the CCD camera pixels. After adjusting the OPD to record the 5 dye wavelengths, each pixel's 5 records are then compared to the known probe-labeling scheme to identify the corresponding chromosome material. The image of the metaphase chromosome spread can be karyotyped using specific pseudo-colors for maximal visual discrimination of the chromosomes.

Another nonhyperspectral method is the Multi-Fluorescence *In Situ* Hybridization or M-FISH System. It uses five single-band-pass excitation/emission filter sets where each filter set is specific for one of the five fluorochromes used in the combinatorial labeling process (Fig. 1(a)). Again, five separate exposures are taken (one exposure for each filter set) and combined to create a multicolor image. Most M-FISH systems now have motorized filter set selectors which minimize any image registration shifts that may occur between exposures. For each pixel, the presence/absence of signal from each of the five fluorochromes is assessed and then matched to the probe-coding scheme to determine which chromosome material is present. As with the SKY System, an image of the metaphase spread is converted into a karyotype and specific pseudo-colors are used to maximize visual discrimination of the chromosomes.

Hyperspectral Imaging

As has been stated, the main difference between multi- and hyper-spectral imaging is the collection of a broad contiguous band of wavelengths. The apparatus and methods for hyperspectral imaging are similar to multispectral imaging (Fig. 1). Instead of collecting sections of the emission spectrum using filters or interferometers, hyperspectral imaging collects a continuous emission spectrum by dispersing the wavelength of light using gratings and measuring the wavelength intensity, or by dispersing Fourier components of light using interferometers and measuring interference intensity (Fig. 1(b)). Multispectral images are limited to measuring total intensity within the selected wavelength band. Determining the abundance of more than one dye per pixel is less accurate when only a small segment of each dyes' spectrum is passed by the emission filter. This has not been considered a limitation in the past because most research depended on just identifying the presence and not the abundance of several spectrally broad and overlapping dyes which also occupied relatively large spatial areas. With the advent of smaller and smaller featured microarrays, the desire to use more dyes in identification schemes, and the use of quantum-dot markers which can be used to indicate both genetic makeup and quantity, the limitations of multispectral imaging has led to a move to hyperspectral imaging.

In contrast to the fluorescence multispectral imaging approaches, a hyperspectral image is fully three-dimensional. Every XY spatial pixel of the image has an arbitrarily large number of wavelengths allied with it. An intensity cube of data is collected. Two of the axes of the cube are spatial and one axis is wavelength (see Fig. 2). At each of the 3D points, there is an associated intensity of emitted energy. Therefore, it is a four-dimensional information space. The individual intensity signature of differing emitting bodies can then be separated to give a complete view. Spectral decomposition is a mathematical/algorithmical process by which individual emission spectra of the dyes can be uniquely separated from a composite (full spectral) image by knowing the signature of each dye emission spectra component measured separately. Spectral decomposition may be performed using standard algorithms with codes which are readily available.

Interferometric Hyperspectral Imaging

The system architecture of the interferometric hyperspectral imager resembles that of the SKY System (Fig. 1). But, instead of just sequencing the interferometer to select a set of specific wavelengths to record as in the SKY System, the interferometer in the hyperspectral imager is sequenced through discrete optical path differences to record the intensity at each pixel for each OPD. Fourier spectroscopy techniques are used to transform the OPD intensities at each pixel into the wavelength intensities at each pixel. Each pixel may be thought of as a Fourier spectrometer which is not dependent on any other pixel. Then, spectral decomposition algorithms can be run to determine which dyes are present (see the next section).

Like the multifilter methods, the interferometer approach provides a real two-dimensional image that can be observed, focused, and optimized prior to measurement. But like the filtering methods, the excitation source must be filtered for sufficient contrast to view the fluorescent emissions. Interferometers have high optical throughput, relative to other methods, and can have high and variable spectral resolution. The spectral range can cover from the ultraviolet to the infrared. The high and variable spectral resolution allows the user to trade off sensitivity, spectral resolution, and acquisition time in a way that is not possible with any other technique. And interferometers do not cause polarization of light as a filter may.

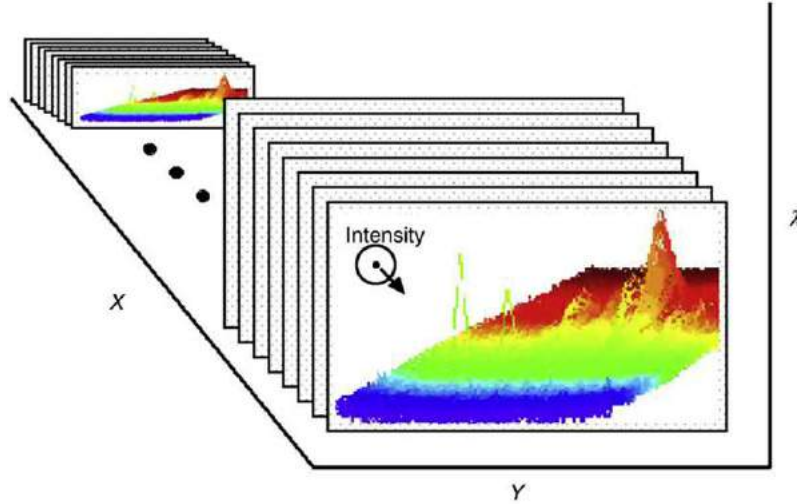


Fig. 2 Intensity data cube. In this data cube, each CCD exposure captures a measurement in intensity (number of counts) by wavelength for a small spatial area in Y and ΔX . Many displacements in X with subsequent CCD acquisitions yield a 'cube' of data in X , Y , and λ . The intensity counts may be thought of as another dimension, $I(x,y,\lambda)$. Another form of the data cube is with each CCD acquisition comprising intensity counts for a spatial X and Y area for a small $\Delta\lambda$ and multiple displacements in λ for each acquisition.

However, interferometric instruments are susceptible to image registration and lateral coherence difficulties. They are limited by the extreme precision required of the optical components which results in the continuous need for monitoring and adjustment, since aberrations are not tolerated in interferometric instruments. The throughput is reduced by loss due to the beamsplitter design and from the mirrors. Additionally, the Fourier transform instrument deconvolutes based on a two-step process; the correlation of the light passed by an interferometer followed by an inverse Fourier transform in software.

To mitigate the extreme precision and sensitivity of the most interferometric systems, Sagnac interferometers have been employed because of their common path intrinsic stability. The common path compensates any misalignment, vibration, or temperature affect. The OPD is provided by the Sagnac affect. When the Sagnac interferometer is rotated, the path traveled in one direction will be different from that of the light on the common path in the other direction. Although, this interferometer method is trading one difficulty for another, it is somewhat easier to manage the rotation of a physical system than to maintain the exacting precision of mirror alignments in a standard interferometer.

Spatial Resolution

The spatial resolution of a Sagnac Hyperspectral Imager is set by the size of the pixels of the CCD camera and the microscope system magnification, since the interferometric hyperspectral viewer sees the entire field of view at once. Let the x - and y -dimensions of the CCD pixel be δx and δy , respectively, and the magnification be M , then the spatial resolutions in the x and y directions are $2\delta x/M$ and $2\delta y/M$, respectively. For CCD pixel sizes of 10 microns square, binning together two pixels in the x - and two pixels in the y -direction, and an overall magnification of 60, the x and y resolutions are both 0.66 microns.

Field of View

For N pixel rows and C pixel columns, the field of view will be $N\delta y/M$ and $C\delta x/M$ for the length of the area along the y - and x -axis, respectively. For a CCD array of 1536 columns by 1024 rows, and the parameters above, the field of view is 171 microns by 256 microns, x and y , respectively.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution will depend on the ability to vary the OPD and the Fourier transform. While the interferometer is used basically as a filter in the multispectral imager to select a wavelength, in the hyperspectral imager it is used to vary the OPD for all the wavelengths whose energy is integrated at the CCD camera pixels.

The electric field, E , for emissions at one wavelength, λ , or wave number, $k = 1/\lambda$, radial frequency, ω , and amplitude, A , can be written as

$$E(k) = A(k)e^{-i(2\pi kx - \omega t)} \quad (1)$$

and the intensity which is measured is the magnitude squared of the electric field:

$$I = E^* E \quad (2)$$

Emissions from fluorescent sources are usually noncoherent and composed of a range of wavelengths, even with only one dye present. The intensity measured at a pixel is the superposition of many wave numbers. After traveling a certain OPD, L , the intensity integrated over all wave numbers may be written as

$$I_{\text{pxl}}(L) = \frac{1}{2} \left[\int I(k) dk + \int I(k) \cos(2\pi k L) dk \right] \quad (3)$$

or

$$I_{\text{pxl}}(L) = C_0 + C_1 \cdot \text{Re}\{\text{FT}[I(k)]\} \quad (4)$$

where FT stands for Fourier transform.

The inverse transform is

$$\begin{aligned} \int I_{\text{pxl}}(L) e^{-i2\pi k' L} dL &= C_0 \int e^{-i2\pi k' L} dL + C_1 \int \int I(k) e^{-i2\pi k L'} dk \cdot e^{i2\pi k' L} dL \\ \int I_{\text{pxl}}(L) e^{-i2\pi k' L} dL &= C_3 + C_4 I(k) \end{aligned} \quad (5)$$

The integration constant, C_3 , can be ignored for large wave numbers, as in the case of the visible range, since C_3 will be proportional to the inverse of the wave number. The intensity as a function of wave number is then

$$I(k) = C_5 \int I_{\text{pxl}}(L) e^{-i2\pi k L} dL \quad (6)$$

For absolute quantitative measurements, the constant C_5 can be determined by calibration with a known emission. For discrete measurements made by changing the optical path length, then the working formula is

$$I_{\text{pxl}}(k) = \sum_{L_i} I_{\text{pxl}}(L_i) e^{-i2\pi k L_i} \Delta L \quad (7)$$

This is the discrete Fourier transform which must be performed for each pixel to determine the intensity as a function of wavelength. Per the Nyquist sampling tenets:

$$\Delta L = \frac{1}{(N/2)\Delta k} \quad \text{or} \quad \Delta k = \frac{2}{N\Delta L} \quad (8)$$

where Δk is the discrete wave number interval and N the number of different OPDs. Sagnac interferometers can make rotations which give at least a change of $\Delta L = \lambda/4$, so substituting in [Eq. \(8\)](#):

$$\begin{aligned} \Delta k = k_1 - k_0 &= \frac{1}{\lambda_1} - \frac{1}{\lambda_0} = \frac{\Delta \lambda}{\lambda^2 + \lambda \Delta \lambda} = \frac{8}{N\lambda} \\ \frac{\Delta \lambda}{\lambda^2} &\approx \frac{8}{N\lambda} \quad \text{or} \quad \frac{\Delta \lambda}{\lambda} \approx \frac{8}{N} \end{aligned} \quad (9)$$

A 5% resolution in wavelength (or 20 nm at 400 nm) requires $N=160$ and, of course, a 2.5% resolution requires $N=360$. Therefore, about 400 OPD changes and acquisitions of the intensity fringes will give calculated resolutions of less than 10 nm in wavelength.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Software techniques for displaying data will be described in the next section.

Grating-Dispersion Hyperspectral Imaging

A hyperspectral imaging microscope collects the complete visible emission spectrum from a microscope slide by using a grating spectrometer to provide the wavelength dispersion. This microscope correlates spatial and spectral information with minimal use of optical filters. A hyperspectral microscope system is composed of a standard epi-fluorescence microscope (Olympus IX70) equipped with objectives lens (including $1.25\times$ and $10\times$ Plan Fluorite, $40\times$ Plan Apochromat Dry, and $60\times$, and $100\times$ Plan Apochromat Oil), as well as a set of standard filter cubes (Chroma Technology) to allow for traditional visualization of dyes through the eyepiece. (Prototype typical components are listed in parentheses.) Multiple excitation sources can be utilized including standard 100 W mercury and xenon arc lamps, tungsten brightfield illumination, a 532 nm solid state laser (Brimrose), a helium neon laser (Uniphase), an argon ion laser (Uniphase), and a pulsed-doubled nitrogen dye laser (Laser Science, Inc.) ([Fig. 1\(b\)](#)).

One side port of microscope is optically coupled to a spectrograph (Acton Research SpectraPro[®] 556i). A narrow entrance slit into the spectrograph allows for only one line (ΔX) of the image to be captured at a time. A 50-micron slit permits 80% of maximum throughput for the spectrometer. The spectrograph separates the polychromatic light illuminated from this single line into its light components. Normal operation uses a grating with 50 grooves/mm and a 600 nm-blaze wavelength, which allows

wavelengths of approximately 400–780 nm to be measured. Different spectrograph gratings can be used for special applications from near UV to near IR. The center wavelength and grating can be shifted under computer control. The output of the spectrograph is then captured with an air/liquid-cooled CCD camera (Photometrics-Quantix) and an optional image intensifier (Video Scope International VS4-1845). The camera has a resolution of 1536×1024 pixels with a depth of 12 bits per pixel. The 1536 pixels are aligned in the spatial direction to allow the greatest spatial resolution. During normal operation the wavelength resolution is 'binned' to 128 pixels, sufficient to resolve ~ 8 different spectra in a single excitation scan. Higher resolution can be employed for special applications.

A line of excitation light can be provided by a line generator when laser excitation is used or by a combination of slits and cylindrical lenses when using traditional lamps or burners (mercury or xenon). This line of light illuminates the sample in the same position that maps the resulting emission light through the imaging spectrograph and onto the camera. Some advantages of illumination with a line of light include decreased sample bleaching, reduced background due to scattering, and efficient use of available excitation power. A traditional bandpass, a low-pass dichroic filter, or a linear variable filter (Raynard Corp.) may be employed to select or restrict the excitation wavelength during specific applications.

Linear Variable Filter (LVF)

The linear variable filter (LVF) (Raynard Corporation) can be used to replace the standard excitation filter cube to provide a capability for continuous excitation wavelength band selection. The LVF is $1'' \times 2.25''$ and is coated such that it continuously filters light from ~ 400 nm to ~ 800 nm along its length. Custom filters that span other wavelengths are available. The coating and therefore wavelength is uniform along a short distance, making it an ideal method for selecting an excitation wavelength band to paint across the slide, excite the slide's contents, and map the contents' emission into the slit of the imaging spectrograph. The bandwidth passed is about 21 nm wide. The LVF introduces $\sim 50\%$ attenuation of the intensity.

Spatial Resolution

The spatial resolution of a grating spectrograph Hyperspectral Imager is set by the size of the pixels of the CCD camera in the y -direction and the microscope system magnification. However, differing from the interferometric system, the spatial resolution in the x -direction depends on the spectrometer slit width and the microscope system magnification. As before, let the x - and y -dimensions of the CCD pixel be δx and δy , respectively, and the magnification be M . Let w_s be the slit width of the spectrometer, and note that the slit width is always going to be larger than a few δx . Then the spatial resolution in the y -direction is again $2\delta y/M$, but the spatial resolution in the x -direction is now $2w_s/M$. Again, for CCD pixel sizes of 10 microns square, binning together two pixels in the y -direction, and an overall magnification of 60, the y -resolution is 0.66 microns, as it was for the interferometer method. For a slit width of 50 microns, the x -resolution is 1.67 microns, independent of the number of pixels binned in the x -direction. The control software moves the stage the right number of steps to minimize the overlap or skipped areas. The slit width can be reduced, but at a loss of throughput or increased exposure time. The control software can compensate for different spatial resolution to display the image as 1:1.

Field of View

For N pixel rows (y -dimension) and C pixel columns (x -dimension), the field of view will be $N\delta y/M$ for the length of the area along the y -axis. If the spectrometer slit length is less than the height of the CCD, then the length is just h_s/M , where h_s is the slit length. We orient our CCD array so that the 1536-pixel dimension is along the y -axis. The y -view is thus 256 microns as in the interferometric example. To achieve the same field of view as the Sagnac method imager, we would need to take 103 stage movements and acquisitions. This highlights the difference in the two hyperspectral methods. The grating method collects the complete spectrum on each acquisition with no spectrometer adjustment but needs to step down the slide to produce the full spatial image. The interferometer method collects the complete spatial image on each acquisition with no position adjustment but needs to step the interferometer through a range of optical path differences to acquire multispectral content.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution depends on spectrometer instrument function which includes the slit width, the dispersion of the grating, the CCD pixel size, and number of pixels binned together. The dispersion for normal operation was determined to be ~ 40 nm/mm or 0.4 nm/pixel in the 10 micron CCD pixel example. For example, binning of 8 pixels yields a spectral resolution of 3.2 nm at all wavelengths. To achieve this spectral resolution, the interferometer method would require 1000 OPD increments and acquisitions.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Although the spatial resolutions differ considerably in the two methods, as long as the dye peaks are separated by the spectral resolution spacing, the spectral decomposition algorithms will find all dyes within the spatial resolution area. We have demonstrated that the analysis algorithms can resolve 5 dyes with peaks separated by more than 6 nm in various combinations and known quantity ratios against the dye standards. The program successfully identifies the dyes present with a 91% accuracy in the quantities.

Software

Software (*HyperScope*, a 'C' language program) controls all the hardware components for acquisition. It includes features for analyzing and displaying the resultant $X - Y$ images and spectral information for the grating spectrometer Hyperspectral Imager. During a scan, individual pictures are taken in the $Y - \lambda$ plane, while the stage is incremented in the X -direction. A collection of such $Y - \lambda$ scan files, which can total from 50–1000, is merged to build an image cube (Fig. 2). Another advantage of the single-line acquisition mode is that time-dependent features can be monitored continuously. The software was designed to permit repeated collection of a single line at variable time points. The utility of this feature has been demonstrated by capturing images of fluorescent objects moving through UV-transparent capillaries.

The software was designed to display the resulting $X - Y$ plane extracted in various ways from the image cube. A graph of the emission spectrum for any pixel within that image may be viewed by clicking on any pixel of interest in the $X - Y$ image (Fig. 3(c)). A standard linear curve-fitting algorithm is used to determine the contribution of individual dyes to the measured emission spectrum. The software also features a windowing technique (integration of the emission spectra over a fixed window) which emulates standard filtering. An overlay feature within the software allows the curve fitting results to be displayed in false color and compared by viewing the composite for multiple fluorophores in a single image. Additionally, it permits the contribution of a fluorophore to be emphasized for visualization purposes by either displaying the contribution of a single fluorescent signature in the absence of the others that were present in the image or by scaling individual display intensity values up or down. Overlapping pseudo-colors assigned to represent each fluorophore (either by curve fitting or by windowing) may be displayed as added, averaged, maximum or minimum values. For each new fluorophore–excitation source combination, the characteristic emission spectra is acquired and added to a library of available spectra that can be used as components in the spectral decomposition. This permits the user to introduce new fluorophores at any time. Moreover, background spectra (defined as a region of interest by the user) can be acquired and used as one element of the decomposition. This background component can be turned off in the false color view (as any of the spectral components may be) resulting in enhanced noise reduction and visual contrast improvement.

Description of Acquisition

In order to acquire a full image, a slide is placed on the microstage mover (Ludl Electronics) that allows for precise movement of the sample across the image acquisition plane. The user can position the stage via a joystick or via software to move to pre-determined coordinates. The image is then scanned, with the number of scans being selected at the time of the image acquisition.

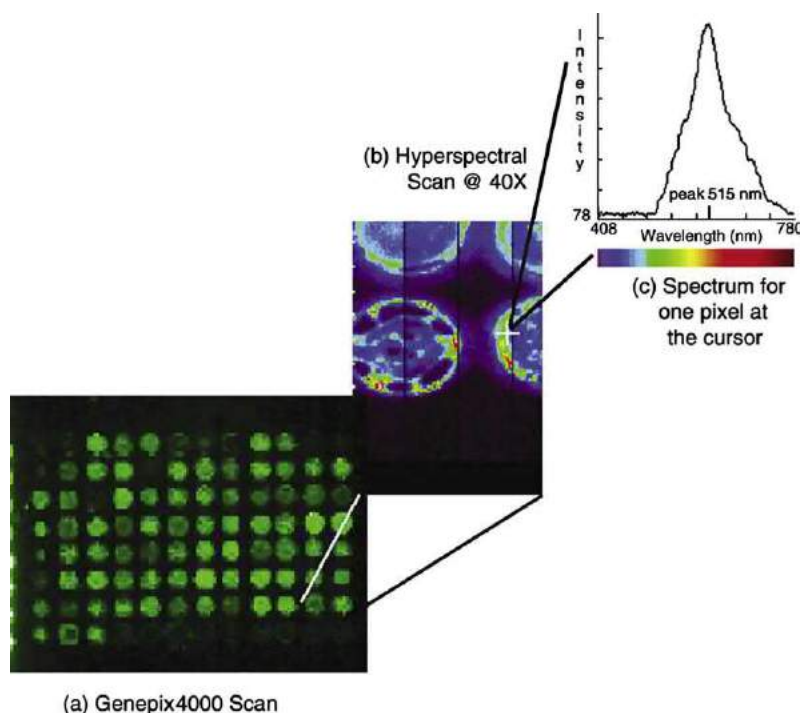


Fig. 3 The benefit of spectral imaging is demonstrated in a spatial high-resolution scan of an expression microarray. (a) A portion of the scan of an expression microarray on a microscope slide using the Genepix4000 scanner. It records the intensity emitted by the Cy3 dye when excited with a 532 nm laser. (b) A portion of the scan of the same microarray with the Hyperspectral Imaging Microscope using a 40X objective and Hg excitation lamp. False color is displayed to indicate the intensity variations. (c) A *HyperScope* computer program display of the spectrum recorded when the cursor is placed over one pixel in a scan.

Furthermore, the sample can be scanned at one magnification or resolution and then rescanned at another. Each $Y - \lambda$ image is saved as a single $Y - \lambda$ file for each increment along the sample. Thus, each image file corresponds to the excitation line of light and the emission collection region that maps through the imaging spectrograph to the CCD camera. Saving individual files permits unlimited $Y - \lambda$ image acquisition at the same Y coordinate, valuable for time-dependent studies. This feature also provides some robustness by enabling a partial $X - Y$ image to be displayed should acquisition of one or more Y scans fail for any reason. After acquiring each image, the stage is moved automatically, one or more steps in the X direction, to acquire emissions at the next line and generate the next $Y - \lambda$ file. Therefore, only the fluorescent probes in a line at a single X -location are examined at each acquisition time. Step sizes can be adjusted by the software to give: (i) the best resolution; (ii) the best 1:1 image; (iii) the minimum overlap; or (iv) the fast acquisition time.

After a variable number of rows of associated emission have been captured (typically 50–1000 scans), an image cube is loaded into memory comprising all the $Y - \lambda$ data files. A graph of the spectrum from any pixel (or bin) in the $X - Y$ plane may be displayed in a separate window by clicking on the cube (Fig. 3(c)). A typical acquisition scan of 250 $Y - \lambda$ files, which are used to construct a $250 \times 768 \times 128$ ($\times 12$ bit) image cube, consists of approximately 100 Mb of information.

MicroArray Scanning and Customized Excitation

Another use of this technology is to analyze spotted microarrays. The Hyperspectral Imager has been compared to commercial two-color scanners with comparable sensitivity at approximately 0.5 fluors per square micron (Fig. 3). Increased sensitivity (500–10000 gain) can be provided by using a microchannel plate amplifier (Model VS4-1845, Videoscope International, Ltd.) installed between the CCD camera and the imaging spectrograph.

Since the hyperspectral imaging microscope collects images that reflect only the $Y - \lambda$ plane, it contains the potential to use a single line of light as a source for image illumination. Therefore, lights systems being developed, which are capable of generating a line of light composed of variable or complex wavelengths, could be employed to precisely control excitation. The incorporation of such an excitation source in hyperspectral imaging microscopy could be useful in many clinical and basic research applications.

Further Reading

- Anguiano, A., 2000. Fluorescence in situ hybridization (FISH): Overview and medical applications. *Journal of Clinical Ligand Assay* 23, 33–42.
- Berland, K., Jacobson, K., French, T., 1998. Electronic cameras for low-light microscopy. *Methods in Cell Biology* 56, 19–44.
- Bernheim, A., Vagner-Capodano, A.M., Couturier, J., 2000. From cytogenetics to cytogenomics of tumors. *S-Medical Science* 16, 528–539.
- Bezrookove, V., Hansson, K., van der Burg, M., *et al.*, 2000. Individuals with abnormal phenotype and normal G-banding karyotype: improvement and limitations in the diagnosis by the use of 24-colour FISH. *Human Genetics* 106, 392–398.
- Boland, M.V., Murphy, R.F., 1999. Automated analysis of patterns in fluorescence-microscope images. *Trends in Cell Biology* 9, 201–202.
- Cardullo, R.A., Alm, E.J., 1998. Introduction to image processing. *Methods in Cell Biology* 56, 91–115.
- Garini, Y., Katzir, N., Cabib, D., Buckwald, R.A., Soenksen, D., Malik, Z., 1996. Spectral bio-imaging. In: Wang, X.F., Herman, B. (Eds.), *Fluorescence Imaging Spectroscopy and Microscopy*. New York: John Wiley and Sons, pp. 87–122.
- Garini, Y., Macville, S., du Manoir, *et al.*, 1996. Spectral Karyotyping. *Bioimaging* 4, 65–72.
- Gothot, A., Grosdent, J., Palulus, J., 1996. A strategy for multiple immunophenotyping by image cytometry: Model studies using latex microbeads labeled with seven Streptavidin-bound fluorochromes. *Cytometry* 24, 214–225.
- Gribble, S.M., Roberts, I., Grace, C., Andrews, K.M., Green, A.R., Nacheva, E.P., 2000. Cytogenetics of the chronic myeloid leukemia-derived cell line K562: Karyotype clarification by multicolor fluorescence *in situ* hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenetics* 118, 1–8.
- Han, H., Gao, X., Su, J., Nie, S., 2001. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nat Biotech.* 19, 631–635.
- Huebschman, M.L., Schultz, R.A., Garner, H.R., 2002. Characteristics and capabilities of the hyperspectral imaging microscope. *Eng in Med and Bio.* 21, 104–117.
- Inoue, T., Gliksmann, N., 1998. Techniques for optimizing microscopy and analysis through digital image processing. *Methods in Cell Biology* 56, 63–90.
- Jain, A.K., 1997. In: *Fundamental of Digital Image Processing*. New Jersey: Prentice-Hall, pp. 80–128.
- Jalal, S.M., Law, M.E., 1999. Utility of multicolor fluorescent in situ hybridization in clinical cytogenetics. *Genet Med.* 1, 181–186.
- Johannes, C., Chudoba, I., Obe, G., 1999. Analysis of X-ray-induced aberrations in human chromosome 5 using high-resolution multicolour banding FISH (mBAND). *Chromosome Research* 7, 625–633.
- Kruse, F., 1993. The Spectral Image Processing System (SIPS) – Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environment* 44, 145–163.
- Lu, Y.J., Morris, J.S., Edwards, P.A.W., Shipley, J., 2000. Evaluation of 24-color multifluor-fluorescence *in-situ* hybridization (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47. *Chromosome Research* 8, 127–132.
- Malik, Z., Cabib, D., Buckwald, R.A., Talmi, A., Garini, Y., Lipson, S.G., 1996. Fourier transform multipixel spectroscopy for quantitative cytology. *Journal of Microscopy* 182, 133–140.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1986. *Numerical Recipes, the Art of Scientific Computing*. New York: Cambridge University Press.
- Raap, A.K., 1998. Advances in fluorescence in situ hybridization. *Mutat. Res. Fundam. Mol. Mech. Mutagen* 400, 287–298.
- Sakamoto, M., Pinkel, D., Mascio, L., *et al.*, 1995. Semiautomated DNA probe mapping using digital imaging microscope: II. system performance. *Cytometry* 19, 60–69.
- Schrock, E., duManoir, S., Veldman, T., *et al.*, 1996. Multicolor spectral karyotyping of human chromosomes. *Science* 273, 494–497.
- Schultz, R.A., Nielsen, T., Zavaleta, J.R., Ruch, R., Wyatt, R., Garner, H.R., 2001. Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43, 239–247.
- Schwartz, S., 1999. Molecular cytogenetics: show me the colors. *Genet Med.* 1, 178–180.
- Speicher, M.R., Ballard, S.G., Ward, D.C., 1996. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12, 368–375.
- Wang, Y.L., 1998. Digital deconvolution of fluorescence images for biologists. *Methods in Cell Biology* 56, 305–315.
- Zhao, L., Hayes, K., Glassman, A., 2000. Enhanced detection of chromosomal abnormalities with the use of RxFISH multicolor banding technique. *Cancer Genet Cytogenetics* 118, 108–111.

Imaging Through Scattering Media

AC Boccara, Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

We are often faced with the problem of seeing through scattering media such as smoke, fog, clothes, blood, tissues, ceramics, walls, etc. Tremendous effort has been made to introduce techniques which will allow us to overcome the intrinsic difficulty of selecting 'good' photons which are able to provide valuable information on structures which appear fuzzy or, for most of the time, completely hidden.

Although a number of concepts and techniques that we will discuss in this article could be applied to a large number of scattering samples, we will restrict ourselves to biological tissues.

Studying light propagation through biological tissues is becoming more and more popular among physicists, biologists, engineers, and doctors. Indeed this research domain appears to be growing rapidly as seen by the number of published articles, conferences, and international meetings. Optical techniques are attractive and are starting to compete with other well-established techniques (magnetic resonance imaging, X-ray imaging, ultrasonic imaging, positron emission tomography, etc.), because they are nonionizing, nondestructive, they can reach high resolution and are usually much cheaper. Optical techniques reveal an optical contrast which is valuable in providing morphologic as well as functional information.

Most biological tissues are weakly absorbing in the deep red and near infrared region of the spectrum; the main difficulty in performing imaging arises from the high level of scattering. Fig. 1 shows the picture of a small flash light bulb placed in the mouth: it appears fuzzy and red.

In Fig. 2 we consider various photon trajectories between a short pulsed source S and a fast detector D.

Unscattered photons are called 'ballistic' photons, their number decreases exponentially, and they are completely damped after one or two millimeters.

Then there are scattered photons which propagate close to the ballistic trajectory reaching the detector; these are named 'snake-like' photons and can be used for imaging purposes.

Most of the energy which reaches D is related to the scattered photons which may exhibit very long trajectories (typically ~ 10 times the distance S-D).

The description of various experimental methods used to overcome the difficulty introduced by scattering will be associated with the various photon families that we have mentioned: ballistic, snake-like, and scattered photons. We will first give a few orders of magnitude of the parameters which will dictate the way light interacts with the tissues. Let us emphasize that these parameters also carry the information related to the contrast mechanism. The ultimate goal is to realize an 'optical biopsy', noninvasive but as close as possible to the regular biopsy followed by histopathology.



Fig. 1 Biological tissues are highly scattering but the absorption is low in the red (as seen in this digital visible photography) and the near infrared.

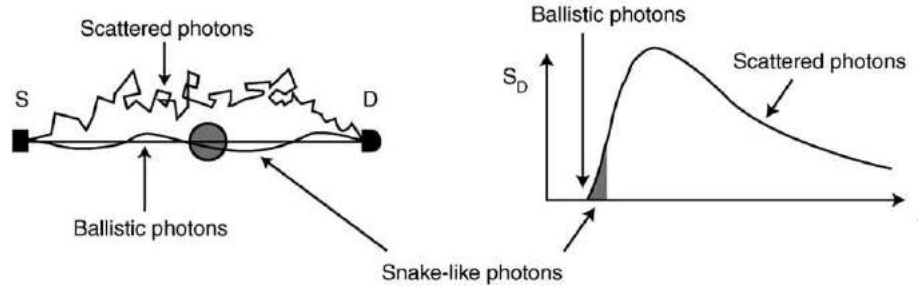


Fig. 2 Between an ultrafast light source S and a detector D can be distinguished various classes of photons which can be used for imaging.

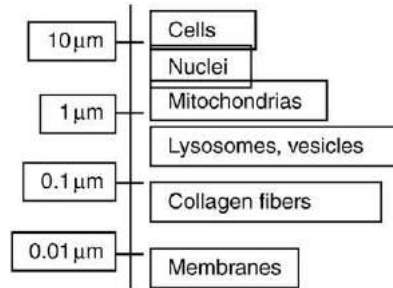


Fig. 3 Various sizes of scatterers may be found in cell structures.

Optical Parameters

Scattering Cross-Section

As shown in **Fig. 3**, biological tissues are constituted of structures which exhibit various characteristic scales: from a few nanometers in membranes to about ten micrometers for an entire cell.

As we know from scattering theories these various scales correspond to various scattering properties. For small particles ('small' means much smaller than the optical wavelength λ) we are in the Rayleigh regime which is based on the evaluation of a time-varying dipole induced by a uniform time-varying (at the light frequency scale) electric field. This dipole radiates in the far field and its scattering cross-section, which is the ratio of the scattered power to the incoming irradiance for a sphere of radius r , is given by

$$\sigma = \frac{\langle P \rangle}{I} = \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{8\pi}{3} r^6 k^4 \quad (1)$$

n being the refractive index of the sphere divided by the index of the surrounding medium and $k = 2\pi/\lambda$. The scattering diagram for such particles is isotropic if we use unpolarized light. This cross-section is much smaller than the 'geometrical' section S of the particle; for instance, if $n = 1.5$, $a = 10$ nm, $\lambda = 0.6$ μm , $\sigma = 1510^{-8}$ μm^2 , which is much smaller than the actual section (3×10^{-4} μm^2).

For larger particles the sphere (or the ellipse) interacting with a plane polarized wave is rather complicated because we are not restricted to the dipolar approximation (uniform field), but induced multipolar moments and their radiating fields have to be taken into account. This theory was proposed by Mie and can be found in optics textbooks. As an example, **Fig. 4** shows the scattering cross-section of a micron-size particle as a function of the wavelength: a number of 'resonances' can be observed.

The scattering diagram is no longer isotropic, it exhibits lobes, and forward scattering is much higher than back-scattering. Scattering is characterized by a scattering coefficient, μ_s , which is the inverse mean free path l_s of photons between scattering events:

$$l_s = 1/\mu_s = \sigma \rho \quad (2)$$

where σ is the scattering cross-section and ρ the number of scatterers by unit volume.

The scattering coefficients μ_s of tissues are found within the range 100–1000 cm^{-1} . After a path length L , the number of ballistic photons will be damped by a factor $\exp(-L/l_s)$, so tissue imaging cannot rely only on ballistic photons as they survive only to a path length larger than about one mm. One needs to consider in more detail the scattering properties of the scatterers.

The dominant sizes of the scatterers in tissue (cells) demand consideration of the scattering in the Mie regime. Such a scattering event does not result in isotropic scattering angles. Instead, the scattering in tissue is biased in the forward direction.

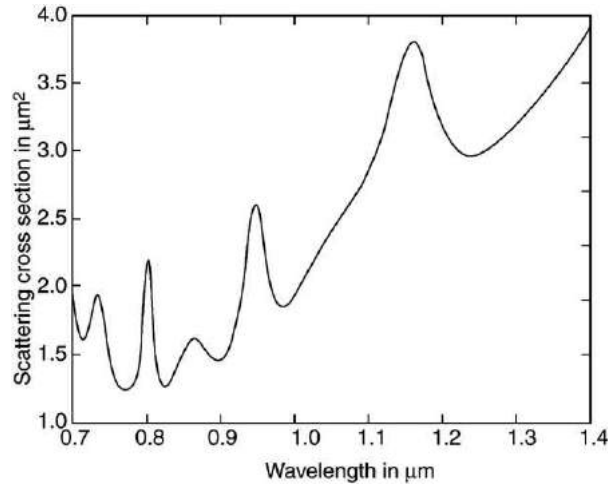


Fig. 4 In the Mie scattering regime resonances are observed when varying the wavelength (here a 1 μm sphere, refractive index: 2).

Anisotropic scattering is quantified in a coefficient, g , which is defined as the mean cosine of the scattering angle θ , $p(\theta)$ being the probability of a particular scattering angle:

$$g \equiv \langle \cos(\theta) \rangle \equiv \frac{\int_0^\pi p(\theta) \cos(\theta) \sin(\theta) d\theta}{\int_0^\pi p(\theta) \sin(\theta) d\theta} \quad (3)$$

For isotropic scattering, $g=0$. For complete forward scattering, $g=1$. In tissues, g varies from 0.7 to 0.95.

Photon migration, especially in the so-called diffusion regime, is often based on isotropic scattering such as for phonons (heat) or charge carriers scattering. For diffusion-like models, it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ'_s , and obtain equivalent results using:

$$\mu'_s = \mu_s(1 - g) \quad (4)$$

One can consider $l'_s = 1/\mu'_s$ as the length over which the incoming photon loses the memory of its initial direction, μ'_s is usually called the transport-corrected scattering coefficient.

Absorption Cross-Section

The absorption coefficient, μ_a (cm^{-1}) represents the inverse mean path length of photons before absorption. $1/\mu_a$ is also the distance in a medium at which intensity is attenuated by a factor of $1/e$ (Beer's Lambert law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In the visible range, the blood absorption is relatively high, whereas it is much weaker in the near infrared. Water absorption is low in the visible and NIR region and increases above 1 μm , with tissues turning completely opaque above 2 μm . Thus, for greatest penetration of light in tissue, wavelengths in the 700–1300 nm spectrum are used. This region of the spectrum is called 'the therapeutic window'. Different spectra of chromophores allow one to separate the contribution of varying functional species in tissue such as quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation.

Finally, when light is propagating in scattering media of thickness L , the damping of the incoming energy involves a combination of its absorption and scattering properties, and is given by

$$\exp(-L/a) \quad \text{with} \quad a^2 = 1/3\mu_a(\mu'_s + \mu_a) \quad (5)$$

This formula reflects the fact that the photon's pathlength is increased by scattering, leading to an increase of the damping, which is related to absorption.

These few examples provide a 'static' view of light behavior in scattering media. Studying in a more detailed way, the spatial and temporal distribution of light, is more difficult.

Photon Migration in Scattering Media

Photon migration models have been inspired by other fields, such as astrophysics, soft matter physics, and neutronics, where physical media are random in both space and time. To avoid difficulties linked to the microscopic complexity such investigations rely on a statistical basis.

The most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla} \cdot (D \vec{\nabla} \Phi(\vec{r}, t)) - \mu_a \Phi(\vec{r}, t) = \frac{1}{c} \frac{\partial \Phi(\vec{r}, t)}{\partial t} - S(\vec{r}, t) \quad (6)$$

where $\vec{r}$ and t are spatial and temporal coordinates, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients by the formula:

$$D = \frac{1}{3[\mu_a + \mu'_s]} \quad (7)$$

The quantity $\Phi(\vec{r}, t)$ is called the fluence, defined as the power divided by the unit area. This equation does not reflect the scattering angular dependence; the use of μ'_s instead of μ_s ensures that isotropic scattering has been reached. Indeed, for the use of the diffusion theory which implies anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient; $S(\vec{r}, t)$ is the source term.

The diffusion equation has been solved analytically for different types of measurements such as reflection and transmission modes, assuming that the optical properties remain invariant throughout the tissue. To incorporate the finite boundaries, the method of images has been used.

The diffusion approximation equation in the frequency-domain is the Fourier transform of the time-domain equation. This leads to the equation:

$$\vec{\nabla} \cdot (D \vec{\nabla} \Phi(\vec{r}, \omega)) - \left[\mu_a + \frac{i\omega}{c} \right] \Phi(\vec{r}, \omega) + S(\vec{r}, \omega) = 0 \quad (8)$$

Here the time variable is replaced by the frequency ω which is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the so-called 'photon density wave' in addition to a DC component:

$$\Phi(\vec{r}, \omega) = \Phi_{AC}(\vec{r}, \omega) + \Phi_{DC}(\vec{r}, 0) = I_{AC} \exp(i\theta) + \Phi_{DC}(\vec{r}, 0) \quad (9)$$

where θ is the phase shift of the diffusing wave whose wavelength is:

$$2\pi \sqrt{\frac{2c}{3\mu'_s \omega}} \quad (10)$$

Although analytical techniques provide a rapid understanding of most phenomena involved in static or dynamic situations, real biological samples are much more complex than homogeneous isotropic media. In these cases, direct and inverse algorithms map the spatially varying optical properties.

Monte Carlo simulations are very popular nowadays; a number of useful programs are available on the websites of various groups. Results may include time or polarization dependence of the photons along their paths.

The other method used in tissue optics is the random walk theory (RWT) on a lattice. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Imaging in the Ballistic Regime: Taking Advantage of Various Coherence Properties

Shallow Tissue Imaging Through Selection of Ballistic Photons

When imaging thin (less than 2 mm) tissue samples, it is possible to use photons which have not been scattered, called the ballistic photons which can be used to form high-resolution images in the same manner as if no scattering had taken place.

Their number, however, decreases exponentially with propagation distance and this ballistic signal is usually hidden by multiply scattered photons that obscure the image. For relatively shallow tissue depths it is possible to use various spatial filtering techniques to block the multiply scattered photons and there is much current research that aims to extend optical imaging and biopsy to depths of a few mm.

Confocal microscopy rejects much of the scattered light through spatial filtering and has been shown to form 3-D images at tissue depths of up to 200–300 μm , both *in vitro* and *in vivo*, which corresponds to about 50 dB of damping for the ballistic photons.

Two-photon fluorescence microscopy is particularly interesting in this regime since it ensures that all the detected fluorescence photons originate from the desired image plane. This technique also ensures a better penetration (using IR excitation) and a better resolution (nonlinear response) than single photon confocal microscopes.

Imaging to much greater tissue depths does not appear to be practical using optical microscopy and spatial filtering alone and more sophisticated techniques must also be employed to discriminate in favor of the ballistic photons. These techniques exploit either the fact that the ballistic signal photons retain spatial coherence of incident light or the fact that the ballistic photons keep the optical phase memory that multiply diffused photons have lost.

One of the most successful ballistic light-imaging techniques based on the latter filtering approach is optical coherence tomography (OCT) which uses coherence gating using low temporal coherence light and an interferometric detection. Coherent

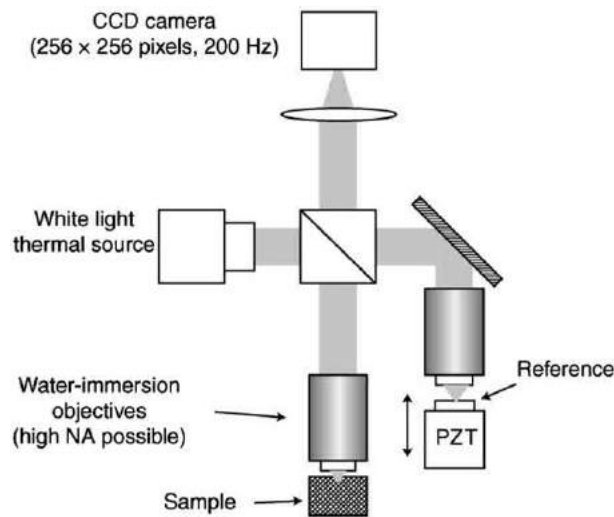


Fig. 5 The full-field OCT setup based on a Linnik interferometer. The PZT actuator is used to modulate the path difference between the two arms.

detection is used to discriminate against the multiply scattered photons that scatter back into the trajectories of ballistic photons. The ballistic light signal is measured by interference with a reference beam and the resulting pattern is detected with high sensitivity using homodyne or heterodyne detection. By using low coherence length sources and matching the time-of-flight of the reference beam signal to that of the desired ballistic light, OCT can acquire depth-resolved images in scattering media such as biological tissues. For powers in the mW range for tissue irradiation in the infrared, the unscattered ballistic (or more precisely single back scattered) component of the light will be limited by the shot noise detection limit after propagating through approximately 25 scattering mean free paths (MFP) of a scattering medium (or more than 100 dB). This corresponds to a 'typical' tissue thickness of about 1 mm tissue depth for the reflection geometry which is potentially useful for clinical applications such as screening for skin cancer, or for retinal examination.

OCT has proved clinically useful as a means of acquiring depth-resolved images in the eye and is being developed for application in strongly scattering biological tissue. Although the image acquisition requires pixel-by-pixel scanning, the use of new high-power superluminescent diodes in the 100 mW range or of ultrafast lasers to provide high average power, low coherence length radiation (sub-two cycle lasers) has resulted in an *in vivo* OCT system providing real-time imaging.

There are a number of other approaches to coherent imaging, including whole-field 2-D acquisition techniques which remove the need to scan transversely pixel by pixel.

Whole-field imaging using a temporally and spatially incoherent source and a well-balanced interferometric microscope intrinsically provides higher image acquisition rates and can take advantage of inexpensive spatially incoherent, broadband sources such as high-power LEDs, and white light thermal sources.

As an example, Fig. 5 shows Linnik-type white light interferometer with optical path modulation and Fig. 6 an image of an *ex-vivo* rat retina where one can recognize the main feature expected from histology.

Holography, which works in the Fourier space rather than in the object or image space, is also an interferometric technique able to discriminate between ballistic and scattered photons. Since the development of electronic holography, which takes advantage of available arrays of detectors with multimillion pixels, this approach is rapidly growing.

Real-time 3-D imaging systems based on photorefractive holography in multi-quantum-wells devices, which can potentially acquire depth-resolved images at thousands of frames per second, have been demonstrated.

These ballistic light techniques can extend the depth of tissue imaging to the mm range when imaging in reflection. When it is necessary to image through cms, rather than mms of tissue, however, there will be no detectable ballistic light signal and one must extract useful information from the scattered photons.

Imaging in the Snake-Like Regime: Taking Advantage of Time (or Frequency) Gating

The first approach is to use time gating to select the earliest arriving scattered light, which will provide the most useful image information.

For moderate tissue depths one can exploit the fact that biological tissue is highly forward scattering, with most photons being only slightly deviated from their original direction upon each scattering event. After a few cm of tissue, there can be a significant number of photons that have followed reasonably well-defined 'snake-like' paths about the original direction through the tissue, arriving at the detector after the ballistic light but before the fully diffuse photons whose trajectories are effectively randomized.

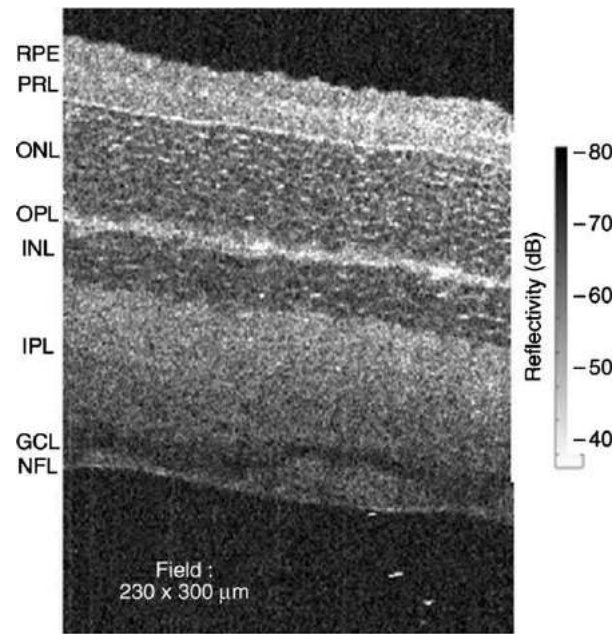


Fig. 6 OCT rat retinal image: retinal pigmented epithelium. PRL: photoreceptor layer. ONL: outer nuclear layer. OPL: outer plexiform layer. INL: inner nuclear layer. IPL: inner plexiform layer. GCL: ganglion cell layer. NFL: nerve fiber layer.

This early arriving light can be selectively gated using ultrafast lasers and a variety of detectors such as streak cameras, time-gated optical image intensifiers, and time-gated fast photon counting systems, which provide picoseconds time gates.

An alternative, sometimes cheaper, approach to temporal gating is the use of (high) frequency signal discrimination (which is the Fourier transform of the pulse experiment). The periodic solution of the diffusion equation with a periodic source term gives rise to the so-called 'photon density waves'. Such diffusive waves, like thermal waves, are used in the near-field regime (their amplitudes are damped by $\exp(-2\pi)$ after propagating over a single wavelength path) and they can reveal subsurface structures through their refractive and diffusive behavior. Coherence gating and polarization gating can be used when the early arriving light still retains some degree of coherence of the incident light. Indeed polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, from the photons which experience multiple scattering resulting in a complete loss of memory of their initial polarization state.

As an example, Fig. 7 shows the image of sub-mm light sources (LEDs) taken through 1 cm of polystyrene solution ($\mu'_s \sim \mu_s \sim 10 \text{ cm}^{-1}$) by selecting photons which kept the memory of their initial circular polarization (here a few less than 1%). One can see the sources quite clearly.

For many important biomedical applications, such as mammography or imaging brain function, it is necessary to penetrate through 5–10 cm of tissue, after which all the detected signal is diffuse. Detecting earlier arriving photons can provide more information but, if the time window is too narrow, the detected signal becomes too weak to use with acceptable data acquisition times.

This problem is, however, being addressed with some success using statistical models of photon transport, with different degrees of approximation ranging from full Monte Carlo simulation of photon propagation to the diffusion equation. The approach is to address the inverse problem, i.e. to measure the scattered light signal as comprehensively as is practical and to calculate what distribution of absorption and scattering properties would have produced the measured signal. This provides a means to quantify the optical properties of biological tissue, averaged over a volume corresponding to a particular distribution of photon paths, and to form relatively low-resolution tomographic images.

Multiplying the number of sources and detectors and correlating the various intensity distributions over the whole structure under examination gives, as expected, better precision in localizing the size and position of the local structure to be identified.

Despite the multiple input and multiple output approaches, real-life situations with the previously mentioned problem of the inhomogeneity of the optical properties of human tissues means that the results are still poor in term of resolution and sub-centimeter structures can hardly be observed in breast imaging, for instance.

One important exception to this limited success in finding precisely the position, the size, and the optical properties of hidden structures is the case of 'activation' studies. This corresponds to a 'differential' situation in which an unknown structure is locally changing with time. Here the goal is not to reach a full description of the unknown structure but only to characterize the local changes induced by the activation.

The main field of application of this approach is certainly in brain activation studies. A matrix of light emitters (usually at two wavelengths which differentiate oxy- and deoxy-hemoglobin, for instance 780 and 840 nm) is coupled to a matrix of light

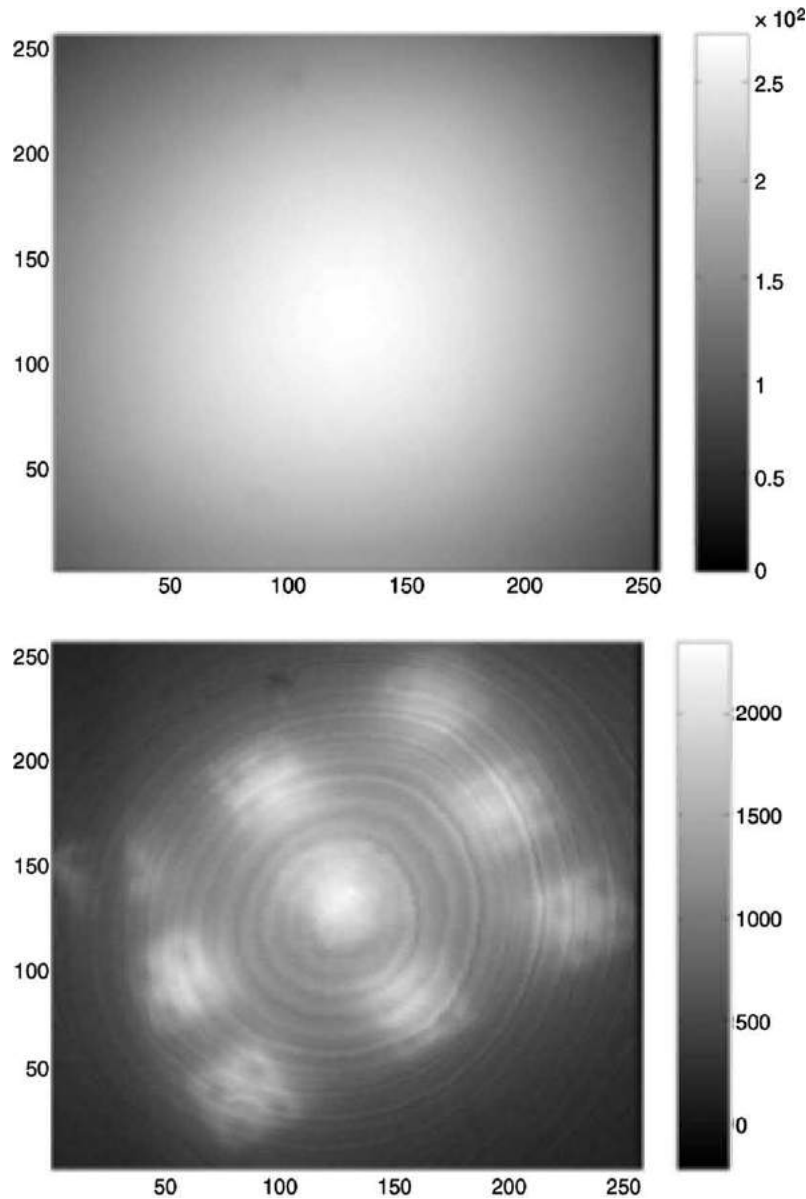


Fig. 7 Polarization imaging: here the degree of circular polarization is measured in the upper image. The lower image shows the field without polarization filtering.

detectors. The signals are usually balanced and any local change which breaks the symmetry of the scattered photon paths induces a differential signal, easy to detect, because there is no background signal before activation takes place.

Moreover it has been demonstrated that for weak perturbations (weak variations of the optical parameters), the inverse problem can be rigorously solved. This geometrical selection provides a cheap and simple setup with rather spectacular results. Sometimes a modulation is added and the phase shift of the photon density wave is enhanced by the differential approach.

Imaging in the Multiple Scattering Regime: Taking Advantage of all the Photons

Light Only: Transillumination and Fluorescence Imaging

The optical systems used for CW measurements are rather simple and require only the alignment of a number of light sources and detectors. Compared to time-gated techniques due to multiple scattering a much stronger path length distribution is present in the signal and the result is a loss of localization and resolution. In the so-called transillumination geometry, collimated detection is used to isolate multiply scattered photons close to the normal emergence angle.

Direct imaging with the required resolution using CW techniques in very thick tissues, such as breast tissue, has not been established even with sophisticated multi-sources/multi-detectors combinations (typically a few tens of sources and detectors at three different wavelengths).

Despite rapid progress in the modeling of light behavior in scattering media it has not yet been demonstrated that CW imaging can provide images with suitable resolution in thick tissue and time-dependent measurement techniques discussed earlier are dominantly used.

For fluorescence imaging, one has to account for the strong attenuation of light as it passes through tissue. Fluorescent dyes (porphyrin, indo-cyanine, and more recently quantum dots) have been developed that are excited and re-emit in the 'biological window' at near-infrared (NIR) wavelengths, where scattering and absorption coefficients are lower, which is favorable for deep fluorescence imaging. Fluorescent intensity in deep tissues depends on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, and on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. In order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon path lengths which depend on all these different parameters. The amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore.

Using fluorescence contrast has the unique feature of carrying two spectroscopic contrasts related to both excitation and fluorescent photons.

The next two techniques take advantage of another way to handle the optical information by coupling optics and acoustics.

Coupling Light and Sound: Acousto-Optic and Photo-Acoustic

Although the two approaches that we will now describe are different in their basic principles as well as in the experimental setups, their names are similar – opto-acoustics (sometimes called photo-acoustics) and acousto-optics.

Opto- (or Photo-)Acoustics

In this technique a pulsed (or modulated) electromagnetic beam irradiates the structure under examination. Despite their tortuous paths (about 10 to 100 cm) photons propagate through the volume in a few nanoseconds. Local absorbing centers then absorb the electromagnetic energy, they experience a fast thermal expansion, and become sources of acoustic waves.

The optical problem of tomographic reconstruction is then transformed into a simpler problem of acoustic reconstruction of source distribution because the speed of the sound is much slower than the speed of the light. More precisely we can compute the signal generated by a short light pulse impinging on a scattering medium which contains an absorber (thickness e) as can be seen on Fig. 8, when $e\mu_a \ll 1$.

The amount of power per unit surface which is deposited in the absorbing slice is $E\mu_a$ (E is the fluence in W/m^2). When the light pulse duration τ is short enough, despite the propagation in the scattering medium ($\tau \ll e/v$ where v is the speed of the sound), we find a variation of the mass by volume unit:

$$\Delta\rho \sim \rho\beta\Delta T \sim \rho \frac{\beta E\mu_a e}{C_v e} \quad (11)$$

The corresponding pressure is:

$$\Delta p = \frac{v^2}{\gamma} \Delta\rho \quad \text{or} \quad \Delta p \sim \rho \frac{v^2 \beta E}{C_p} \mu_a \quad (12)$$

Although this approach is new, compared to the purely optical transillumination approach, it has led to a large number of studies in which time gating brings new information to the accurate determination of the US emitter depth. It has been applied to breast tumor examination to complement standard imaging techniques such as echography and X-ray mammography. Recently this technique has demonstrated submillimeter resolution for *in vivo* blood distribution in rat brain imaging.

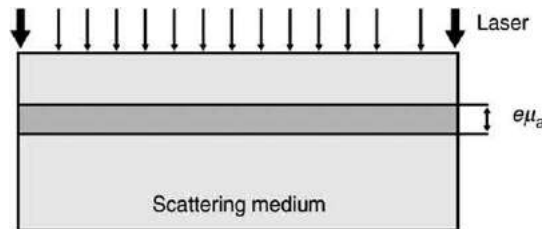


Fig. 8 Photo acoustic signal generation: 1-D model.

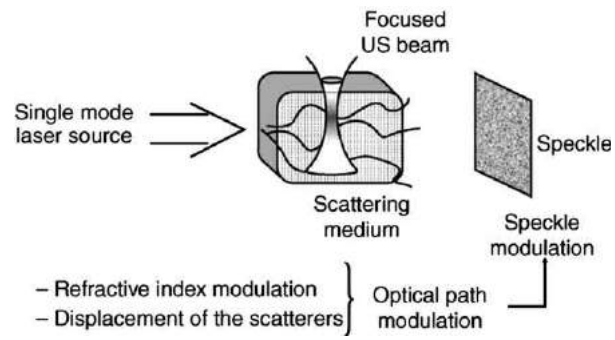


Fig. 9 Acousto-optic imaging principle: photons are tagged by the acoustic field.

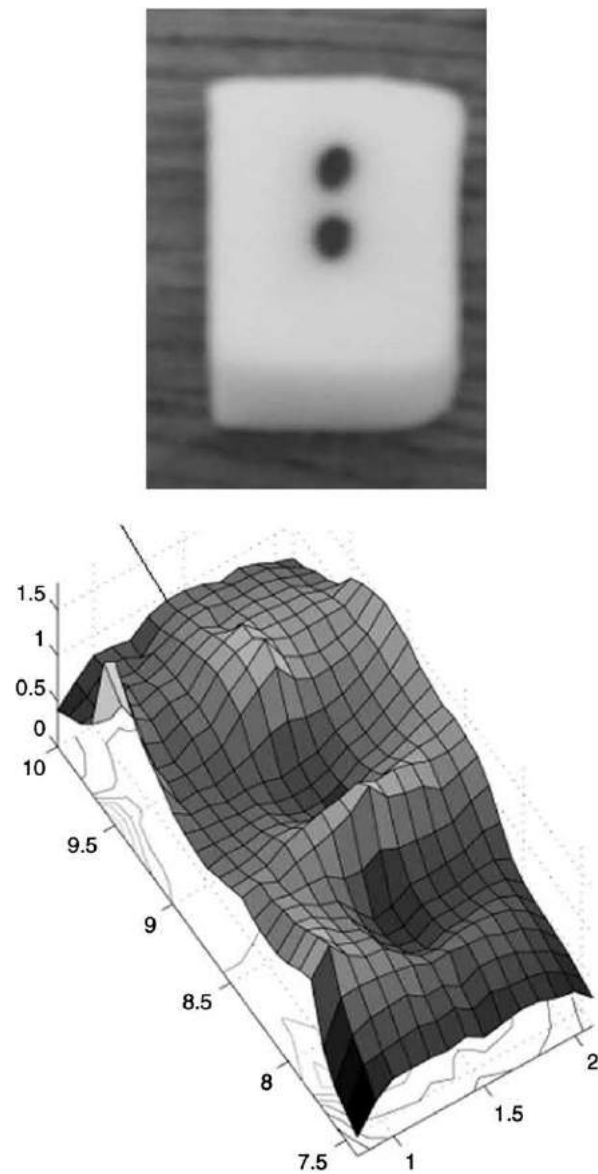


Fig. 10 Acousto-optic imaging (lower image) of two absorbing spheres embedded in a scattering medium (upper image shows a section of the gel sample).

Acousto-Optic

Here a DC single frequency laser is used to illuminate the sample. For a typical biological sample a few cm thick the trajectories are distributed between the sample thickness value and more than ten times this value. Due to the good temporal coherence of the source all the scattered photons are likely to interfere when they emerge from the sample volume. These interferences are seen as a speckle field which can be observed at a suitable distance from the sample surface.

As seen in [Fig. 9](#), adding an ultrasonic field will mainly induce a small (smaller than the optical wavelength) displacement of the scatterers (it also induces local variation of the refractive index) which will cause a speckle modulation at the ultrasonic frequency.

If the zone irradiated by the ultrasonic field is optically absorbing, less photons will emerge from this zone than from a nonabsorbing one and thus less modulation will be seen when scanning across the sample volume. This modulation can be detected by selecting a single speckle grain and averaging over a number of uncorrelated speckle distributions to obtain a suitable average value (for instance with latex particles). This is not always possible, in particular because the overall geometry is rather stable in a semisolid tissue. Using a single detector which receives many speckle grains gives less contrast. A multiple detector and a parallel processing of many speckle grains improve greatly the speed and the signal-to-noise ratio: a typical image of absorbing spheres, embedded in a scattering phantom, is shown in [Fig. 10](#).

Conclusion

The complexity associated with multiple scattering in turbid media, such as tissues of the human body, prevents an easy use of light as a routine tool for in-depth examination, whereas superficial examinations have always been sources of diagnostics.

Optical techniques will obviously progress to better characterization of tissues' optical properties, improvement of laser properties and detection techniques, as well as new mathematical approaches of the inverse problem.

We believe that this field of research is still open to new ideas, and new experimental schemes leading to new breakthroughs in the delicate problem of using light in highly scattering media.

Acknowledgements

I would like to warmly thank all the research staff (Arnaud Dubois, Benoit Forget, François Ramaz, Vincent Lorient) and PhD students (Michael Atlan, Kate Grieve, Gael Moneron, Juliette Selb) at ESPCI whose results have been used to illustrate this article. I also acknowledge discussions with my colleagues Pr Paul French (Imperial College London) and Dr Amir Gandjbakhche (NIH Bethesda).

See also: Wavefront Sensors and Control (Imaging Through Turbulence)

Further Reading

- Boccara, A.C., Fink, M. (Eds.), 2001. *Compte Rendu de l'Académie des Sciences, Physics and Astrophysics série IV tome 2, no. 8. Optical and Acoustical Imaging of Biological Media*. Paris: Elsevier.
- Born, M., Wolf, E., 1997. *Principles of Optics*. New York: Oxford University Press.
- Goodman, J.W., 2000. *Statistical Optics*. New York: Wiley-Interscience.
- Hecht, E., Zajac, A., 1979. *Optics*. Reading, MA: Addison-Wesley.
- Sebbah, P. (Ed.), 2001. *Waves and Imaging through Complex Media*. Dordrecht: Kluwer Academic Publishers.
- Van de Hulst, H.C., 1981. *Light Scattering by Small Particles*. New York: Dover.
- Van Tiggelen, B., Skipetrov, S. (Eds.), 2003. *Wave Scattering in Complex Media: From Theory to Applications*. Dordrecht: Kluwer Academic Publishers.

Infrared Imaging

K Krapels and RG Driggers, US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

Published by Elsevier Ltd.

Introduction

All objects above 0 Kelvin emit electromagnetic radiation associated with the thermal activity on the surface of the object. For terrestrial temperatures (around 300 Kelvin), objects emit a good portion of the electromagnetic flux in the infrared part of the electromagnetic spectrum. The visible band spans wavelengths from 0.4 to 0.7 micrometers (infrared engineers typically use micrometer/ μm for wavelength rather than the nanometer or wavenumber more commonly used in other fields). Infrared imaging devices convert energy in the infrared portion of the electromagnetic spectrum into images that can be created by the human eye. The unaided human eye cannot image infrared energy because the lens of the eye is opaque to infrared radiation.

The infrared spectrum begins at the red end of the visible spectrum where the eye can no longer sense energy. It spans from 0.7 μm to 100 μm . The infrared spectrum is, by common convention, broken into five different bands (this may vary according to the application). The bands are typically defined as: near infrared (NIR) from 0.7 to 1.0 μm ; short wavelength infrared (SWIR) from 1.0 to 3.0 μm ; mid-wavelength infrared (MWIR) from 3.0 to 5.0 μm ; long wavelength infrared (LWIR) from 8.0 to 14.0 μm ; and far infrared (FIR) from 14.0 to 100 μm . These bands are depicted graphically in Fig. 1, which shows the atmospheric transmission for a 1-kilometer horizontal ground path for a 'standard' day in the United States. These types of transmission graphs can be tailored for any conditions, using sophisticated atmospheric models such as MODTRAN (from www.ontar.com). Note that there are many atmospheric 'windows', such that an imager designed with such a band selection can see through the atmosphere.

The primary difference between a visible spectrum camera and an infrared imager is the physical phenomenology of the radiation from the scene being imaged. The energy used by a visible camera is predominantly reflected solar, or some other illuminating energy, in the visible spectrum. The energy imaged by infrared imagers (Forward Looking Infrared, commonly known as FLIRs) in the MWIR and LWIR bands, is primarily self-emitted radiation. From Fig. 1, the MWIR band has an atmospheric window in the 3 to 5 μm region and the LWIR band has an atmospheric window in the 8 to 12 μm region. The atmosphere is opaque in the 5 to 8 μm region, so it would be pointless to construct a camera that responds to this waveband.

Figs. 2 and 3 are images that show the difference in the source of the radiation sensed by the two types of cameras. The visible image (Fig. 2) is light that was provided by the Sun, propagated through Earth's atmosphere, reflected off the objects in the scene, traversed through a second atmospheric path to the sensor, and then imaged with a lens and a visible band detector. A key here is that the objects in the scene are represented by their reflectivity characteristics. The image characteristics can also change by any change in atmospheric path or source characteristic change. The atmospheric path characteristics from the Sun to the objects change frequently, because the Sun angle changes throughout the day and also weather and cloud conditions change. The visible imager characterization model is therefore a multipath problem that is extremely difficult.

The LWIR image given in Fig. 3 is obtained primarily by the emission of radiation by objects in the scene. The amount of electromagnetic flux depends on the temperature and emissivity of the objects. A higher temperature and a higher emissivity correspond to a higher flux. The image shown is 'white-hot' or the whiter a point in the image the higher the flux leaving the object. It is interesting to note that trees have a natural self-cooling process since a high temperature can damage foliage. Objects that have absorbed a large amount of solar energy are hot and are emitting large amounts of infrared radiation. This is sometimes called 'solar loading'.

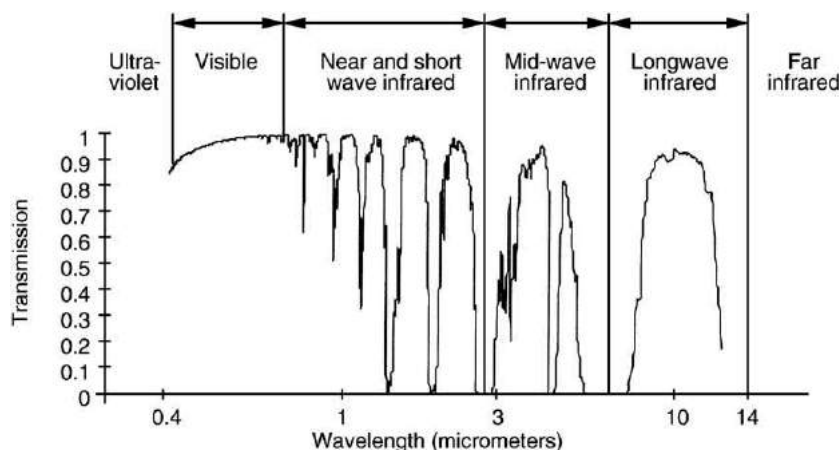


Fig. 1 Visible and infrared spectra.



Fig. 2 Visible image—reflected energy.

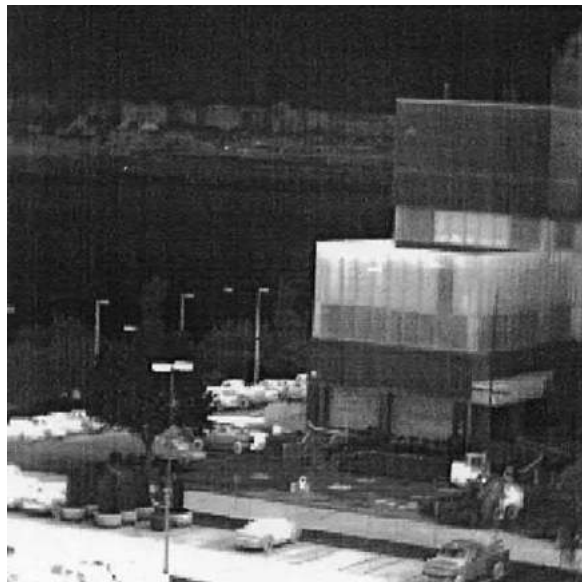


Fig. 3 Infrared image—self-emitted energy.

The characteristics of the infrared radiation emitted by an object are described by Planck's black-body law in terms of spectral radiant emittance

$$M_\lambda = \varepsilon(\lambda) \frac{c_1}{\lambda^5 (e^{c_2/\lambda T} - 1)} \text{ W/cm}^2 \mu\text{m} \quad (1)$$

where c_1 and c_2 are constants of $3.7418 \times 10^4 \text{ W } \mu\text{m}^4/\text{cm}^2$ and $1.4388 \times 10^4 \text{ } \mu\text{m K}$. The wavelength, λ , is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A black-body source is defined as an object with an emissivity of 1.0, so it is a perfect emitter. Source emissions of black-bodies at typical terrestrial temperatures are shown in Fig. 4. Often, in modeling, the terrestrial background temperature is assumed to be 300 K. The source emittance curves are shown for other temperatures for comparison purposes. One curve corresponds to an object colder than the background and two curves correspond to temperatures hotter than the background.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to the left (shorter wavelengths) as the body temperature increases. If the temperature of a black-body is increased to that of

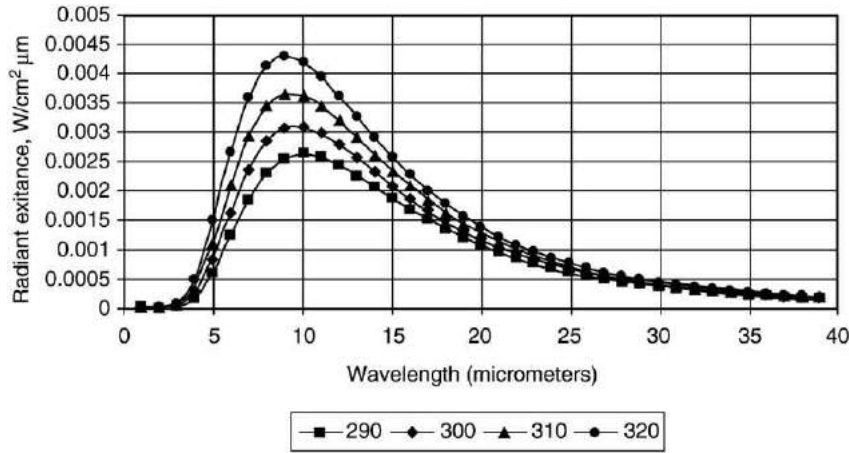


Fig. 4 Planck's blackbody radiation curves for four temperatures from 290 K to 320 K.

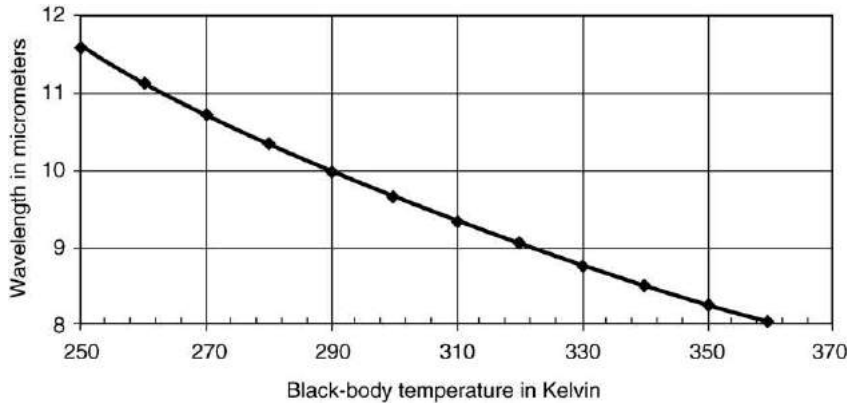


Fig. 5 Location of peak of black-body radiation, Wien's Law.

the Sun (5900 Kelvin), the peak of the spectral shape would decrease to 0.55 μm or green light. This peak wavelength is described by Wien's displacement law

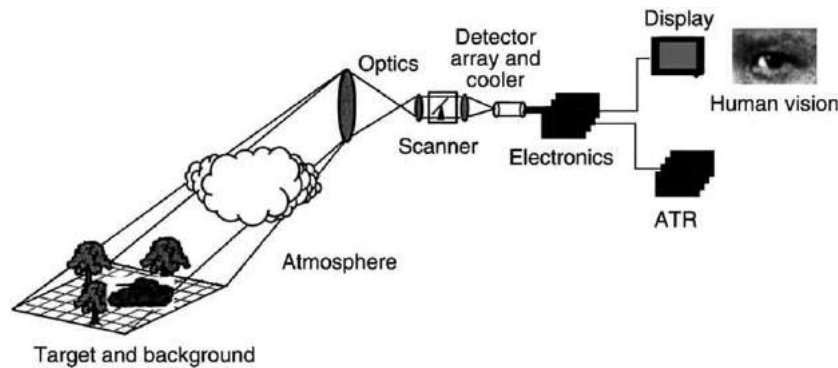
$$\lambda_{\text{max}} = 2898/T, \mu\text{m} \quad (2)$$

Fig. 5 shows the peak wavelength as a function of temperature through the longwave region. It is important to note that the difference between the black-body curves is the 'signal' in the infrared bands. For an infrared sensor, if the background is at 300 K and the target is at 302 K, the signal is the difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. However, in the visible band this is not the case. For example, consider the case of a white target on a black background. The black background is generating no signal, whereas the white target is generating a maximum signal (given that the sensor gain has been adjusted). Dynamic range may be fully utilized in a visible sensor. In the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux never has a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors. **Table 1** summarizes this characteristic numerically for the 300 K background and targets shown in **Fig. 4**.

A typical infrared imaging system scenario is depicted in **Fig. 6**. The scene consists of two major components, the target and the background. In an IR scene, the majority of the energy is remitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. All of these contributors, which are not the target of interest, essentially reduce one's ability to discriminate the target. So, at the entrance aperture for the sensor, the target signature is reduced from the values observed at the target. Then, the signal is degraded by the optics and scanner (as applicable) of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for

Table 1 Signal/dynamic range limitation in IR bands

Blackbody temperature	Exitance in 8–12 μm band (W/cm^2)	Contrast
290 ($-10\Delta T$ K target)	$1.25\text{E}-02$	– 8.48%
300 (background)	$1.48\text{E}-02$	
310 ($+10\Delta T$ target)	$1.74\text{E}-02$	7.97%
320 ($+20\Delta T$ target)	$2.02\text{E}-02$	15.39%

**Fig. 6** Typical infrared imaging scenario.

straightforward modifications to a performance model for changes in the sensor or environment when performing trade-off analyses.

History of Infrared Imaging

Night vision systems began to be developed extensively during World War II. Satisfying the requirement to image at night was approached with two different methodologies. The first method was image intensification (I^2). This method involves amplification of any small light that was available and displaying it directly to the eye. Typically I^2 devices are sensitive to visible light and a small portion of the short-wave IR (SWIR, 0.7–3.0 μm) band. They are often classified (along with visible band imagers) as electro-optical (EO) imagers. With an I^2 imager, there must be some source of illumination for them to function well (as little as starlight is sufficient for approx 20/40 acuity with current systems). The second type of night vision device is the infrared or thermal imager (also known as FLIRs). FLIRs are sensitive to the radiation in either the mid-wave IR (MWIR, 3–5 μm) or long-wave IR (LWIR, 8–14 μm) bands.

The first infrared cameras used photographic film that was sensitive to infrared wavelengths. Intelligence and reconnaissance imaging from aircraft drove the infrared system development as a night augmentation to existing visible cameras. These imagers used either continuous moving film or stationary film, like an ordinary camera. A slit aperture was used in the continuous moving film cameras and aircraft motion provided the scan of the scene along the film to build a continuous image. The introduction of television, as electronic imaging in the visible spectrum, led to the development of electronic imagers in the infrared band. Electronic infrared imagers were not restricted to looking down and using aircraft motion for scan, hence these devices came to be known as FLIRs (Forward Looking InfraRed). Today, an electronic infrared imager may be called a FLIR or a thermal imager, with the trend being to use FLIR to describe military applications.

Early FLIRs were often accompanied by illumination devices. They were not very sensitive compared to modern FLIRs so active illumination was necessary. Infrared searchlights or illuminators were used to get higher signal-to-noise ratios. These types of FLIRs are called 'active' IR imagers, while FLIRs that do not use illuminators are 'passive' imagers. Generally, active imagers are used in civil applications and are declining in military use. Broadly, applications for thermal imagers fall into two categories; military and commercial. Table 2 lists some of the purposes and users for modern thermal imagers. The design and performance criteria vary widely for some of the applications.

Types of Infrared Imagers

Infrared imagers are classified by different characteristics: scan type, detector material/cooling requirements, and detector physics. The scan type refers to the method used to construct the electronic image. The camera may use a single detector which is raster scanned over the input scene to build an image. Alternatively, a parallel scan uses a linear array of detectors scanned across a scene

Table 2 Applications of infrared imaging

<i>Category</i>	<i>User</i>	<i>Application</i>
Military	Intelligence analyst	Intelligence surveillance and reconnaissance
	Pilot, driver	Navigation/pilotage
	Gunner, weapons officer	Targeting/fire control
	Air defense	Search and track
Commercial	<i>Civil government</i>	
	Police	Surveillance, fugitive tracking
	Fire	Rescue, hot-spot location
	<i>Environmental</i>	
	EPA	Emission tracking
	Interior department	Resource management
	NIMA/USGS	Mapping
	NOAA/MWS	Weather forecasting
	<i>Industrial</i>	
	Manufacturers	Machine vision
	Maintenance	Non-destructive testing
	<i>Medical</i>	
	Doctors	Diagnostic imaging

to build an image. The latest advances in materials have led to staring arrays of detectors. In a staring system, a detector is present in two dimensions to represent each image pixel and no mechanical motion of the focal plane is necessary to construct the image. There are some hybrid FLIR types that combine the different imaging techniques. Usually this combination leads to an improvement in signal-to-noise ratio or to reduce undersampling.

The second classification, by detector material/cooling requirements, usually describes FLIRs built using differing materials. For example, a typical LWIR FLIR material is mercury cadmium telluride (HgCdTe or MCT). In order to achieve high sensitivity, these devices are cooled to decrease dark current. Usually the cooling is based on liquid nitrogen or a cryogenic cooler, and the detectors operate at 77 K. Another common detector material is indium antimonide (InSb), which is used for MWIR FLIRs and is also cooled. A newer class of infrared cameras is not cooled, being referred to as 'uncooled' FLIRs. These uncooled FLIRs are micro-bolometers (resistive elements) or pyrometers (capacitive elements) and have less sensitivity than cooled imagers. Typically, the cooled imagers are comprised of photon detectors while the uncooled imagers are based on thermal detectors, the uncooled FLIRs being used in lower-performance applications.

Infrared Imager Performance

There are three general categories of infrared sensor performance characterizations. The first is sensitivity and the second is resolution. When end-to-end, or human-in-the-loop (HITL), performance is required, the third type of performance characterization describes the visual acuity of an observer through a sensor. The former two are both related to the hardware and software that comprises the system, while the latter includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene/environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis.

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range/quantization of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In infrared systems, noise equivalent temperature difference (NETD) is often a first-order description of the system sensitivity. The 3D noise model describes more detailed representations of sensitivity parameters (see Further Reading).

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in infrared systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution.

Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity. Typically, visible band systems have plenty of sensitivity and are resolution-limited, whereas infrared imagers have been more sensitivity-limited. With staring infrared sensors, the sensitivity has seen significant improvements.

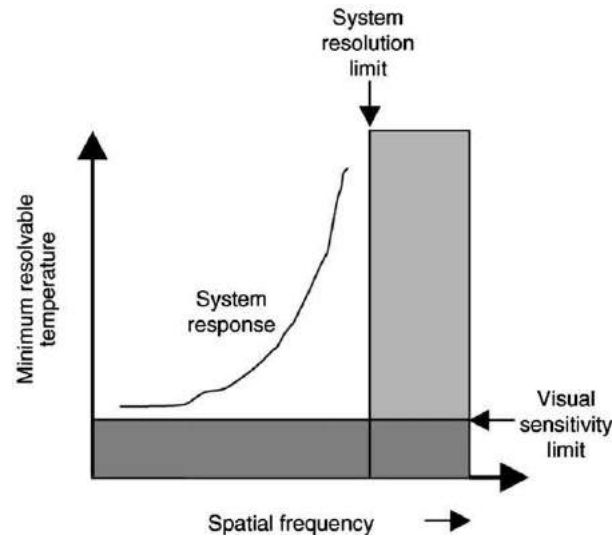


Fig. 7 IR imager performance.

Often metrics, such as NETD and MTF, are considered to be separable. However, in an actual sensor, sensitivity and resolution performance are not independent. As a result, minimum resolvable temperature difference (MRT or MRTD) has become the primary performance metric for infrared systems. MRT is a quantitative performance measure in terms of both sensitivity and resolution and a simple MRT curve is shown in Fig. 7. The performance is bounded by the sensor's limits and the observer's limits. The temperature difference (or thermal contrast) required to image smaller details in a scene increases with detail size. The inclusion of observer performance yields a single sensor performance characterization, MRT. It describes sensitivity as a function of resolution, and includes the human visual system.

Many different sensor characteristics may be considered to increase the fidelity of a sensor model. A list of typical infrared imaging system parameters is given in Fig. 8. When considering performance in the system sense, the groupings are often blocks in the system diagram, each with separate transfer functions. This approach works well unless the linear shift invariance assumption is not valid.

General Characteristics of Infrared Imagers

There are a number of imaging characteristics that make infrared systems a little different than conventional visible imaging systems. These include source flux levels, detector charge storage and sensitivity, detector size, diffraction blur, sampling, and uniformity characteristics.

First, we consider the source flux levels. The daytime and night-time flux levels (in photons per second per square centimeter) on Earth in the visible (0.3 to 0.7 μm) is 1.5×10^{17} and around 1×10^{10} , respectively. In the mid-wave (3 to 5 μm), the daytime and night-time flux levels are 4×10^{15} and 2×10^{15} μm where the flux is a combination of emitted and solar reflected flux. In the longwave, the flux is primarily emitted where both day and night yield a 2×10^{17} μm frequency, the AMOP drops rapidly to zero. The effect is to extend the prediction of the MRT beyond the half sample rate and introduce a parameter called minimum temperature difference perceived (MTDP) to replace MRT in the range performance estimate and also in the lab to characterize an undersampled system.

The concept of MTDP follows from observation of standard 4-bar images as the pattern frequency passes the imager half sample rate. The image changes from four bars to three, two, and finally one (can be perceived). The phase of the target image on the detector array may need to be adjusted to observe this progression. The standard MRT measurement requires that all four bars be resolvable by the observer during the measurement process. For use in the lab, the MTDP is defined as the minimum temperature difference at which two, three, or four bars can be resolved in the image of the standard 4-bar test pattern by an observer, with the test pattern positioned at optimum phase. The optimum phase is the phase at which two, three, or four bars are 'best perceived'. TRM3 uses the standard definition of MRT as for adequately sampled imagers. The TRM3 Approach Model is implemented if the imager is considered undersampled as defined when the prefilter MTF at the half-sample frequency is larger than 10%. The MTDP equation is given by

$$\text{MTDP}(f) = \frac{\frac{\pi}{2} \text{SNR}_{\text{th}} \Psi(f)}{\text{AMOP}(f)} \quad (3)$$

where SNR_{th} is a threshold signal-to-noise ratio and Ψ is the total system filtered noise. MTDP is used in the same manner as MRTD in range calculations.

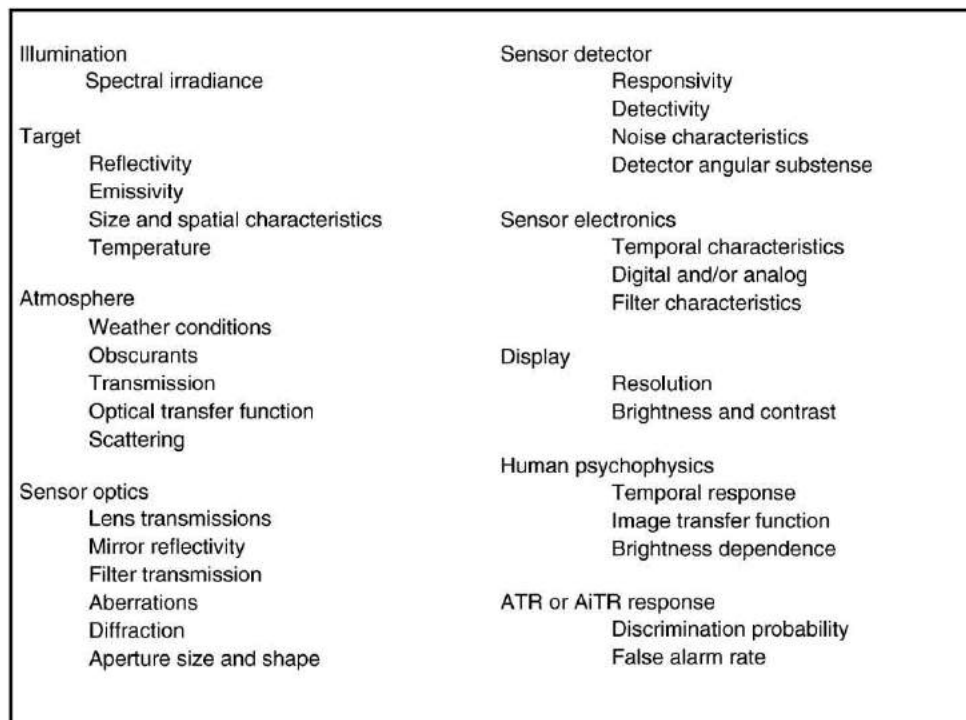


Fig. 8 Factors in infrared sensor modeling.

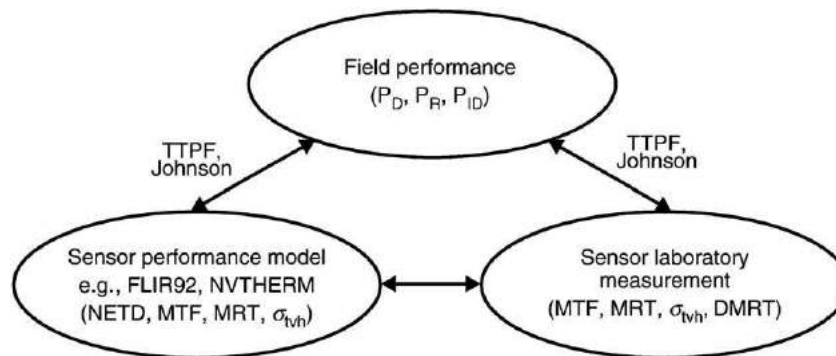


Fig. 9 Modeling, testing and performance triangle.

Another approach is the triangle orientation discrimination (TOD) threshold. In the TOD, the test pattern is a (nonperiodic) equilateral triangle in four possible orientations (apex up, down, left, or right), and the measurement procedure is a robust 4AFC (four alternative forced-choice) psychophysical procedure. In this procedure, the observer has to indicate which triangle orientation he sees, even if he is not sure. Variation of triangle contrast and/or size leads to a variation in the percentage correct between 25% (complete guess) and 100%, and by interpolation the exact 75% correct threshold can be obtained. A complete TOD curve (comparable to an MRTD curve) is obtained by plotting the contrast thresholds as a function of the reciprocal of the triangle angular size (Fig. 9).

The TOD method has a large number of theoretical and practical advantages: it is suitable for under-sampled and well-sampled electro-optical and optical imaging systems in both the thermal and visual domains, it has a close relationship to real target acquisition, and the observer task is easy. The results are free from observer bias and allow statistical significance tests. The method may be implemented in current MRTD test equipment with little effort, and the TOD curve can be used easily in a target acquisition (TA) model such as ACQUIRE. Two validation studies with real targets show that the TOD curve predicts TA performance for under-sampled and well-sampled imaging systems very well. Currently, a theoretical sensor model to predict the TOD (comparable to NVTherm or TRM3) is under development. The lab measurement and field performance appear sound, but the model is not yet available.

Testing Infrared Imagers: NETD, MTF, and MRTD (Sensitivity, Resolution and Acuity)

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. For infrared imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. The system intensity transfer function (SITF) can be used to estimate the NETD, which is the system noise rms voltage over the noise differential output. NETD is the smallest measurable signal produced by a large target (extended source).

Eq. (4) describes NETD as a function of noise voltage and the system intensity transfer function. The measured values are determined from a line of video stripped from the image of a test target, as depicted in Fig. 10. A square test target is placed before a black-body source. The ΔT is the difference between the black-body temperature and the mask. This target is then placed at the focal point of an off-axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Fig. 10(b). The SITF slope for the scan line in Fig. 10(c) is the $\Delta S/\Delta T$ where ΔS is the signal measured for a given ΔT . The N_{rms} is the background signal on the same line.

$$\text{NETD} = \frac{N_{\text{rms}}[\text{volts}]}{\text{SITF_Slope}[\text{volts/K}]} \quad (4)$$

Resolution is a general term that describes the size of resolvable features in an imager's output. With infrared imaging systems, the resolution is described by the system modulation transfer function (MTF). Consider Fig. 11 for the determination of MTF, from either point spread function or edge spread function (psf or esf). In Fig. 11(a), a test target is placed at the focal point of a collimator. The target may be a point, edge, or line as shown in Fig. 11(b). The IR system images this target. The output of a detector scan, or a line of detectors in the staring array case, is taken for analysis, in Fig. 11(c). For the edge spread function, the spatial derivative is taken to get the psf. The Fourier transform is then calculated to give the system MTF as shown in Fig. 11(d). The point spread function is really the impulse response of the system, so a smaller psf is desirable, where a wide MTF is desirable. Such a psf/MTF combination gives a higher resolution.

The MRTD of an infrared system is defined as the differential temperature of a four bar target that makes the target just resolvable by a particular observer. It is a measure of observer visual acuity when a typical observer is using the infrared imager. It results in a descriptor, which is a function not just of a single value. It is a plot of sensitivity as a function of spatial frequency. This parameter can be extended to field performance using Johnson's criteria.

As a laboratory measurement, it uses a 7:1 aspect ratio bar target. The procedure is shown in Fig. 12. A bar target mask, with a black-body illuminator, is placed at the focal plane of a collimator (Fig. 12(a) and 12(b)). The temperature is increased from a small value until the bars are just resolvable to a trained observer (Fig. 12(c)). Then the temperature difference is decreased until the bars are no longer visible. These are averaged for a particular target. The same procedure is performed with a negative contrast bar target. The average of the positive and negative values is the MRT for a particular spatial frequency. These data points are plotted

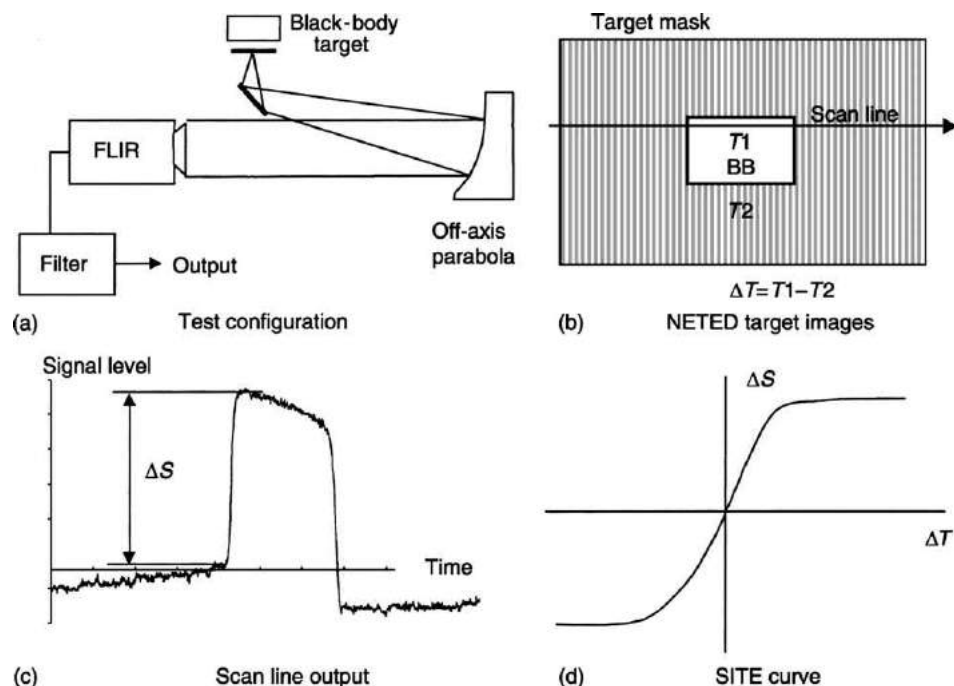


Fig. 10 NETD measurement.

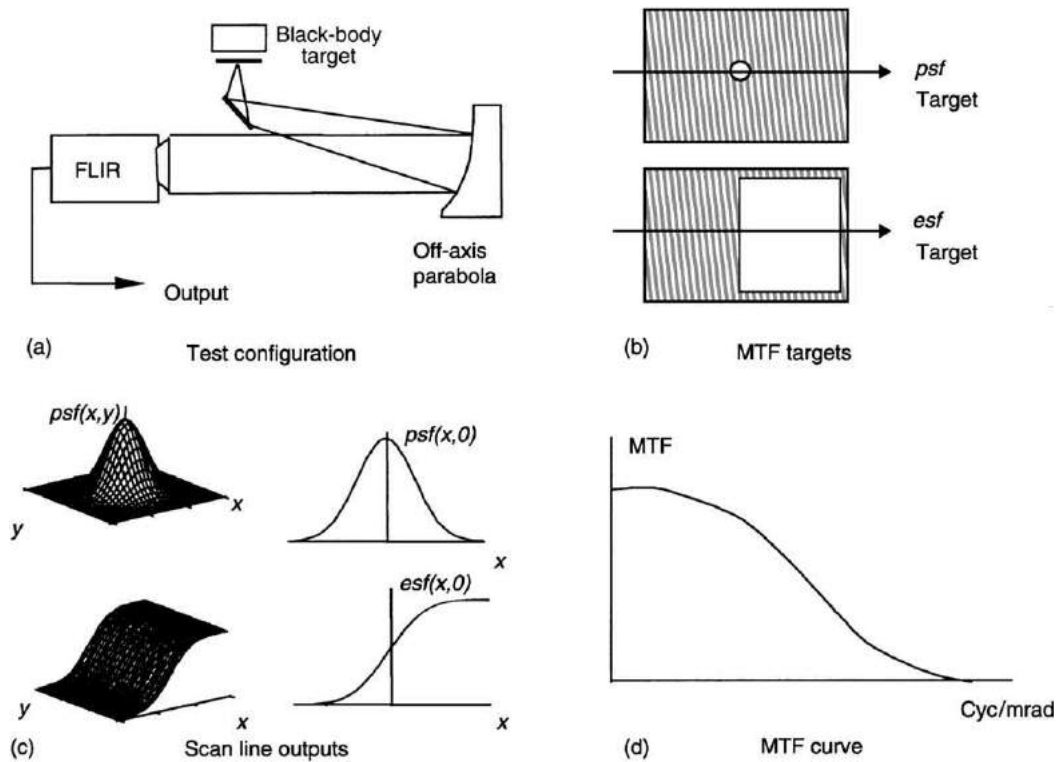


Fig. 11 MTF measurement.

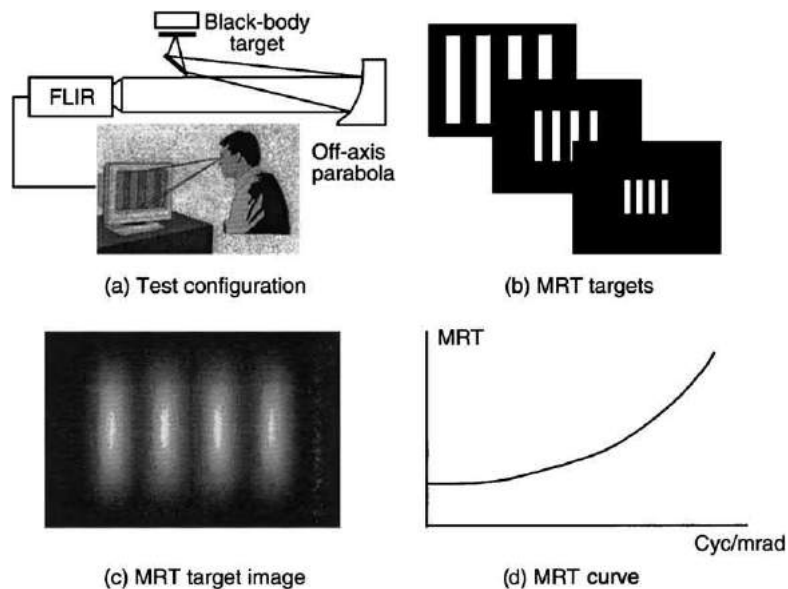


Fig. 12 MRTD measurement.

and a curve may be fitted to interpolate/extrapolate performance at other than the discrete spatial frequencies of the targets, as in Fig. 12(d).

Summary

We have provided a general description of infrared imaging systems in terms of characteristics, modeling, field performance, and performance measurement. The characteristics of infrared imagers continue to change. Significant changes have occurred in the

past five years, to include the development of higher performance uncooled imagers and ultra-narrow field of view photon systems. Large format detector arrays are commercially available in the mid-wave and are becoming more available in the longwave. Still in the research phase are dual-band focal plane level. At first sight, it appears that the longwave flux characteristics are as good as a daytime visible system; however, there are two other factors limiting performance. First, the energy bandgaps of infrared photons are much smaller than those photons in the visible, so the detectors suffer from higher dark current. The detectors are usually cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%. In the infrared, all terrestrial objects are emitting with an emission temperature of around 300 Kelvin. Typically, a two-degree equivalent black-body difference in photon flux is considered a high contrast target to background flux. The flux difference between two blackbodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in Fig. 4. This is the difference in signal that provides an image, so note the difference in signal compared to the ambient background flux. In the longwave, the signal is three percent of the mean flux and in the mid-wave, it is six percent of the mean flux. This means that there is a large flux pedestal associated imaging in the infrared. Unfortunately, there are two problems accompanying this large pedestal. First, photon noise is the dominant noise and is determined by the mean of the pedestal and is compared to the small signal differences. Second, the charge well storage in infrared detectors is limited to around 10^7 charge carriers. A longwave system in a hot desert background would generate 10^{10} charge carriers in a 33 millisecond integration time. Smaller F-numbers, spectral bandwidths, and integration times are used so that the charge well does not saturate. This results in SNR of 10 to 30 times below the ideal. It has been suggested that some on-chip compression may be a solution to the well pedestal problem. In many scanning FLIR systems, the pedestal is eliminated by AC-coupling the detector signals. Infrared focal plane array (IRFPA) readout circuits have been previously designed and fabricated to perform charge skimming and charge partitioning.

Another major difference in infrared systems, from that of visible systems, is the size of the detector and diffraction blur. For infrared photon detectors, typical sizes range from 25 to 50 μm , where visible systems can be fabricated with 3 μm detectors. Also, the diffraction blur for the longwave is more than ten times larger than the visible blur and mid-wave blur is eight times larger than visible blur. Therefore, the image blur, due to diffraction and detector size, is much larger in an infrared system than a visible system. Also, it is common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance.

Finally, both infrared photon and thermal detectors in staring arrays have responsivities that vary dramatically and it is current practice to correct for the resulting nonuniformity (nonuniformity correction or NUC). The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise (especially in static imaging or stabilized target acquisition imaging).

Modeling Infrared Imagers

Otto Schade Sr. developed the earliest imaging system models and performance measures. His work describing television in the 1950s and 1960s, pioneered the way for the characterizations of imaging sensors used by engineers today. His performance measure for still pictures, moving pictures and television systems was based on an observer resolving a three bar periodic target.

John Johnson developed the technique of relating the acquisition of military targets in the field to laboratory measurements of resolvable bar targets. He divided the discrimination tasks into four categories: detection, orientation, recognition, and identification. This technique was termed an 'equivalent bar pattern approach'. It related performance on a simple test target to performance with complex object targets. Johnson viewed scale models and bar targets in the laboratory against a bland background. The smallest discernible barchart target yielded the maximum resolvable bar pattern frequency. These results, tabulated in Table 3 as cycles across the minimum dimension, were the basis for the discrimination methodology in widespread use today. Note that detection took only one cycle, but as the tasks got more specific or difficult, the requirement went as high as 8 cycles across the target minimum dimension.

Table 3 Johnson's criteria

Target	Resolution per minimum dimension			
	Detection	Orientation	Recognition	Identification
Broadside view				
Truck	0.9	1.25	4.5	8
M-48 Tank	0.75	1.2	3.5	7
Stalin Tank	0.75	1.2	3.5	6
Centurion Tank	0.75	1.2	3.5	6
Half-track	1	1.5	4	5
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier (standing)	1.5	4.8	3.8	8
105 howitzer	1	1.5	4.8	6
Average	1 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Table 4 Moser's data

Discrimination task	Line pairs/critical dimension
Detect ship	Aperiodic treatment
Classify as combatant	4
Recognize type	10

Recognizing that typical vehicle aspect ratios were somewhat limited, Johnson and Lawson conducted further experiments. Paul Moser conducted some of his own, reanalyzed Johnson and Lawson's, and developed the concept that this task was related to the average or critical dimension instead of the minimum dimension. This led to the conversion of line pairs to resolution elements, introducing the second dimension. Additionally, he performed a similar experiment on marine targets, which is summarized in **Table 4**. Lloyd and Sendall developed the MRT metric, which combines both resolution and sensitivity characteristics. This measurement of MRT is described in detail in a following article, but there is a theoretical model that is used to evaluate new sensor designs and concepts. It relates the minimum temperature difference between bar pairs at which they are just resolvable as a function of spatial frequency. MRT describes the infrared imager sensitivity as a function of resolution. More specifically, MRT is a measure of thermal contrast sensitivity as a function of spatial frequency.

Ratches *et al.* developed the NVL (US Army Night Vision and Electronic Sensors Directorate formerly known as the Night Vision Laboratory) Static Performance Model, which predicted the end-to-end performance. It started with the target signature and carried through to the observer. The MRT, a laboratory measurement or a modeled value, was related to the probabilities of discrimination through the target transfer probability function, cumulative percentages related to Johnson's criteria.

D'Agostino and Webb created the 3D noise model, both spatial and temporal considerations, for analyzing system noise. This technique divided the noise components into manageable and understandable components. It also simplified the integration of these effects into models.

The introduction of focal plane array (FPA) imagers has created a requirement to improve the model performance to account for the difference between scanning arrays and staring arrays. This change has made sampling an important area for improvement. A new model called NVTHERM is a result of the NVESD model improvement effort, which addresses undersampled imaging system performance.

Sensor characterization is seen in three ways: theoretical models, field performance, and laboratory measurements (**Fig. 9**). Theoretical models are developed that describe sensor sensitivity, resolution, and human performance for the purpose of evaluating new conceptual sensors. Acquisition models, and other models, are developed to relate the theoretical models to field performance. This link allows theoretical models to be converted to field performance quantities (e.g., probabilities of detection, recognition, and identification). Field performance is measured in the field so that the theoretical models can be refined and become more accurate with advanced sensor developments. Since field performance activities are so expensive, methods for the direct measurement of sensor performance are developed for the laboratory. Field performance testing of every infrared sensor built, including buy-off, acceptance, and life-cycle testing, is ridiculous and out of the question. Laboratory measurements of sensor performance are developed such that, given these measurements, the field performance of a system can be predicted. Sensor characterization programs require accurate sensor models, field performance estimates with acquisition models, and repeatable laboratory measurements.

There are a few alternatives to the US NVTherm model. One candidate approach to undersampled imager modeling is Germany's TRM3, or the MDTP approach. The impact of undersampling in the image of a 4-bar target can readily be seen by simply observing the change in the image as a function of spatial frequency and phase. The spatial frequency is defined as line pairs per angular extent of the target, and the phase specifies the relative location of the target image to the detector array raster. These effects are obviously not independent of each other, but for each target there can be found an optimal phase where the observer can see the maximum amplitude modulation in the image of the target. MRT measurements in the past have utilized this variation with phase by allowing the observer to optimize the displayed image through target or system line-of-sight changes during the measurement process. For undersampled imagers, TRM3 addresses the problem of the MRT calculation's inability to predict beyond the half sample rate of the sensor, by replacing the MTF in the denominator of the MRT equation with an appropriately scaled term called the average modulation at optimum phase (AMOP). AMOP is the average signal difference in the image of the 4-bar standard pattern, with the test pattern positioned at optimum phase. AMOP oscillates between the pre-sample MTF and the bar modulation. Beyond a frequency of 0.8 times the sample rate or 1.6 times the half sample arrays and quantum well detector systems. These systems will find their place in applications to include both military and commercial sectors.

Further Reading

- Bijl, P., Valeton, M.J., 1998. Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* 37 (7), 1976–1983.
- D'Agostino, J., Webb, C., 1991. 3-D analysis framework and measurement methodology for imaging system noise. *Proceedings of SPIE* 1488, 110–121.
- Driggers, R.G., Cox, P., Edwards, T., 1998. *Introduction to Infrared and Electro-Optical Systems*. Boston, MA: Artech House, p. 8.
- Holst, G.C., 1995. *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 347.
- Holst, G.C., 1995. *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 432.

- Johnson, J., 1958. Analysis of image forming systems. Proceedings of IIS. 249–273.
- Johnson J and Lawson W (1974) Performance modeling methods and problems. *Proceedings of IRIS*.
- Ratches JA (1973) *NVL Static Performance Model for Thermal Viewing Systems*. USA Electronics Command Report ECOM 7043, AD-A011212.
- Schade, O.H., 1948. Electro-optical characteristics of television systems. RCA Review IX (1–4),
- Sendall, R., Lloyd, J.M., 1970. Improved specifications for infrared imaging systems. Proceedings of IRIS 14 (2), 109–129.
- Vollmerhausen, R.H., Driggers, R.G., 1999. NVTHERM: next generation night vision thermal model. Proceedings of IRIS Passive Sensors 1.
- Wittenstein, W., 1998. Thermal range model TRM3. Proceedings of SPIE 3436.

Interferometric Imaging

DL Marks, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005 Elsevier Ltd. All rights reserved.

Interferometric imaging is image formation by measuring the interference between electromagnetic signals from incoherently emitting or illuminated sources.

Interferometric imaging attempts to improve the resolution of images by interferometrically combining the signals from several apertures. By doing so, a resolution can be achieved that is superior to that achievable through standard image formation by any of the individual apertures. This is especially useful in astronomy, where resolving more distant and smaller stellar objects is always desirable. It is also finding increasing use in microscopy. It has several advantages over conventional noninterferometric imaging methods. The method can combine the light gathered from several imaging instruments to form an image superior to that which can be formed by any one of the instruments. It can also obviate the need to produce a single very large aperture to achieve the equivalent resolution, which in many cases is of impractical size. In addition, because it can produce a phase-resolved measurement, the data can be processed more flexibly to form a computed image estimate. There are also considerable disadvantages to interferometric imaging compared to conventional imaging methods. The amount of signal gathered is far less than could be gathered with an equivalent single aperture. It is difficult to mechanically or feedback stabilize the time delay between two widely separated apertures to obtain an accurate phase measurement. Often only a small bandwidth of the source light can be utilized, further reducing the available signal. The interference component of the combined signal can be very small. Typically an interferometric image must be computed from the data, rather than being directly measured on photographic film or an electronic focal plane array. Even with these disadvantages, interferometric imaging is frequently more feasible than building a single large aperture to achieve the highest available resolution images.

In contrast to interferometric imaging, a standard image forming instrument such as a telescope has a resolution limited by the telescope mirror aperture in the absence of aberrations and atmospheric turbulence. As telescope mirrors become very large, they become bulky, extremely heavy, and difficult to manipulate, as well as deform under their own weight and temperature gradients. However, adaptive optics is being more frequently utilized to dynamically correct for instrument aberrations as well as atmospheric turbulence. The images captured by telescopes are typically intensity-only images, which are amenable to image processing and deconvolution, but nevertheless lack phase information. Interferometric imaging is an alternative, where the resolution of a large aperture can be attained by measuring the interference between individual points of the electromagnetic field within an area equivalent to the large aperture extent. The interferometric combining of light from sub-apertures to achieve the resolution of a larger aperture is called aperture synthesis.

Unlike image formation by a lens, interferometric images are never physically formed. Conventional images are formed when diverging spherical waves emanating from sources are focused by a lens to converging spherical waves on a photographic film or an electronic sensor. In this case, the image information is directly contained in the exposure at each position on the sensor. Interferometric measurements contain the image information as statistical correlations between the electromagnetic fields at pairs of spatial points. The statistical properties of electromagnetic waves are modeled by optical coherence theory. Interferometric imaging works by measuring the statistical correlations between various pairs of points in the electromagnetic field, and then uses optical coherence theory to infer what the image of the source of the radiation is.

The light that emanates from most natural and many artificial sources is incoherent. Incoherent radiation is optical random noise. It is analogous to the sound of static heard on a radio receiving no signal. Because it is random, the fluctuations of the electromagnetic field produced by an incoherent source are completely unrelated to the fluctuations of any other source. The origin of the randomness is usually from microscopic processes such as the thermal motions of electrons or the spontaneous emission of an excited atom. Because these microscopic processes tend to act independently of one another, the light produced by different sources tends to be mutually incoherent. Because incoherent sources do not radiate a deterministic field, the average amount of power they radiate is usually specified rather than the field itself, which is random.

However, as the fields propagate away from their sources, the fields can become partially coherent. To see this, consider the light waves propagating away from a single point-like incoherent source. The field produced at the source point consists of random fluctuations in time, but these fluctuations travel outward in a spherical wave at the speed of light. The fluctuations of the field of any two points an equal distance from the source will be the same, or perfectly correlated, because they arrive from the source at the same time. This is illustrated in [Fig. 1\(a\)](#), where a random wave is emanated from a point incoherent source. Because of the complete correlation, we can say that these two points are fully spatially coherent. So while the source itself produces random noise, the fluctuations of the field at points distant from the source become correlated because they originate from the same source.

Partially coherent waves occur when the measured field contains a superposition of the fields from more than one incoherent source. If there are now two point sources instead of one, each field point in space will receive radiation from both sources. The amount of the contribution from each source to each point in space depends on the distance between the field point and the source. In general, because both sources can be different distances from the two field points, the fluctuations at the field points will not be the same. For example, two field points that are the same distances from both sources will be perfectly correlated, or

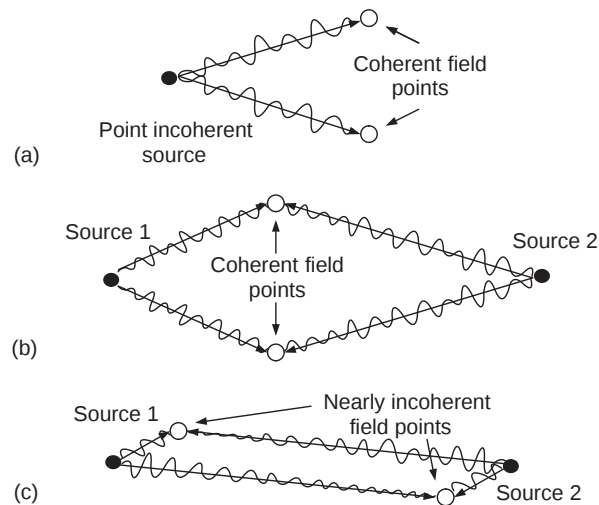


Fig. 1 The distance between incoherent sources and two field points determines the coherence between the field points. (a) Because there is only one source, the field everywhere is determined by the radiation of that source. Therefore, the same fluctuations of the wave arrive at two points an equal distance from the source at the same time, and produce identical fields at both points, and so are completely coherent. (b) Two field points are now the same distance from two incoherent sources. Because both sources contribute identical fields to both points (because the fluctuations of the wave from both sources arrive at both points at the same time), the superposition of the fields from both sources at both field points are identical, and so the fields are coherent. (c) Two field points are now near two sources. Because each field point receives radiation mostly from its nearby source and very little from the other source, the fields at each field point are almost completely mutually incoherent.

coherent, because they receive the same combination of fields from both sources. This situation is illustrated in [Fig. 1\(b\)](#). However, if one considers two points that are near both respective sources, the points will usually receive light from the nearby source. Because the sources are incoherent, and each point is mostly receiving radiation from its closer source, these points will be nearly uncorrelated, as shown in [Fig. 1\(c\)](#).

A spatial distribution of incoherent sources radiating various amounts of power will create a pattern of correlations in the field remote from the source. By using interferometry, the correlation between pairs of field points remote from the sources can be measured. The interference between two points in the electromagnetic field is achieved by relaying the two fields (e.g., with mirrors) to a common location that has an equal travel time for the fields from both points, where the fields are superimposed. The power of the superimposed beams is then detected, by a photocathode, for example. If the two signals being combined are incoherent, then no interference will take place, and the total power measured on the photodetector is simply the sum of the power of the two signals. However, if the two signals are partially coherent, there will be a deviation in the measured power from the sum of the power of the two signals. The magnitude of this deviation relative to the power of the constituent signals indicates the degree of partial coherence of the signal.

An interferometer, such as depicted in [Fig. 2](#), gathers the light from two areas of the electromagnetic field remote from an object and combines them together at the same point to form an interferogram. In this example, two telescopes of a known separation gather light from a star. The primary mirror of each telescope is of insufficient size to resolve the star as anything but a point. However, a telescope is used in practice to collect enough light to produce a measurable interference signal. Each telescope primary mirror forms a point image of the star. At the image plane of each telescope a mirror directs the point image to a measurement plane equidistant between the two telescopes. This line that joins the two telescopes is called the baseline, and it is the baseline length rather than the individual telescope mirror size that determines the minimally resolvable size. The image from each telescope is superimposed by mirrors (or perhaps beamsplitters) and imaged onto a focal plane, to get an image similar to that of [Fig. 3](#). If the light arriving from one telescope is blocked, then the image formed looks like an Airy ring pattern formed by an image, a point-like object, which is the incoherent image of [Fig. 3](#). However, when both signals are allowed to interfere, the image will have vertical stripes superimposed on it. If these stripes are very prominent, so that the image is fully darkened inside the stripes, then an image similar to the rightmost image of [Fig. 3](#) is obtained, complete interference is occurring and the two field points at which the telescopes are situated are coherent or completely correlated. If the stripes disappear, then no interference is occurring and the two field points at which the telescopes reside are incoherent. These examples are shown in [Fig. 3](#), with images of two superimposed points with varying degrees of coherence on the top row, and a line plot of the measured field through the center of the images in the bottom row. The degree to which the contrast of the stripes is enhanced by coherence is referred to as the 'modulation' and indicates the degree or magnitude of partial coherence.

In addition to having a degree of coherence, the correlation between two electromagnetic field points also has a relative phase shift. This phase shift is given by the relative position of the interference maxima, as shown in [Fig. 4](#). As the relative phase between the fields collected by the telescopes changes (which is a function of the path length difference between the light arriving at the two telescopes), the peaks of the interferogram between the two telescope images shift position. By observing the depth of the formed

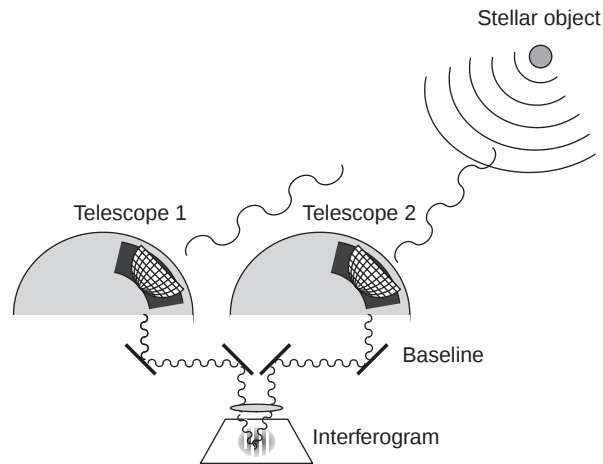


Fig. 2 Two telescopes receive radiation from a stellar object. The light is collimated and directed down the baseline to a point halfway between the telescopes, where the two images are interfered together onto a focal plane.

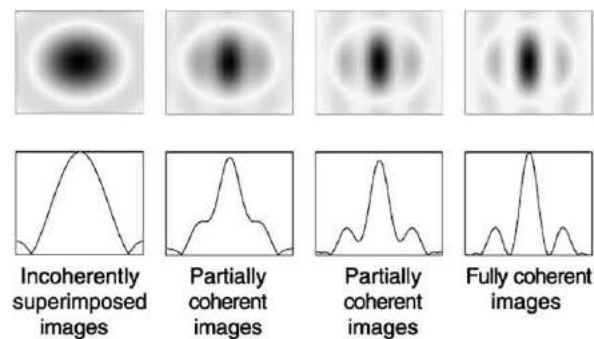


Fig. 3 The image of two interfering wavefronts from a point object with various states of coherence.

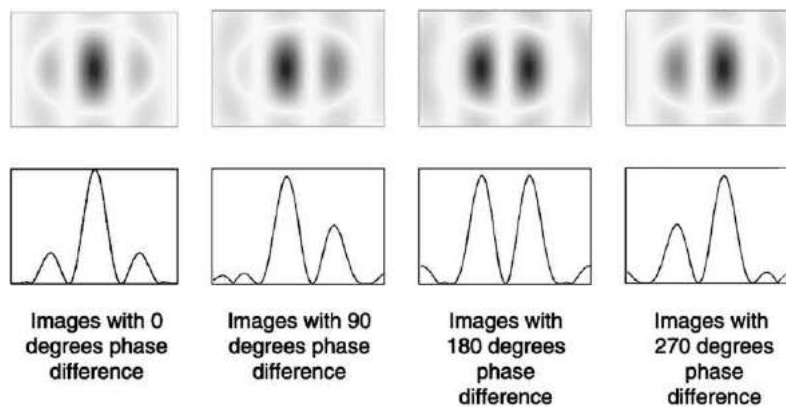


Fig. 4 The image of two interfering wavefronts from a point object with various phase differences between the two wavefronts.

interference fringes and the position of the fringes, both the amplitude and the phase of the correlation between the two field points can be simultaneously measured. The amplitude of the modulation and the phase of the interference are used to compute the complex degree of coherence that is used in the computation of interferometric images.

In practice, the phase of the complex degree of coherence is much more difficult to measure than its magnitude. This is because while the amplitude tends to change very slowly as the positions of the two correlated points change, the phase is sensitive to changes in relative path length on the order of an optical wavelength. Even though a long baseline produces better resolution, the baseline length must be known to a small fraction of a wavelength to obtain a meaningful phase measurement. Minute vibrations

in the positions of the mirrors on the baseline can easily cause fluctuations in the path length greater than a wavelength. This is observed as wild oscillations in the position of the fringes on the image plane. In addition, atmospheric turbulence can also produce large random phase shifts. Achieving a stable phase measurement is one of the most challenging engineering aspects of interferometric imaging.

To model realistic sources, we must determine the coherence produced by sources of a finite size. Finite size incoherent radiators can be regarded as an arrangement of an infinite number of point-like incoherent sources radiating various amounts of power. When an interferogram is made of a field which has two superimposed incoherent sources, the power of the interferograms that each would make alone is added together, rather than the fields of the interferograms. This is because when the fields of the two sources interfere, the phases of both sources are varying randomly over time. The two fields will randomly vary between constructively and destructively interfering with each other. Over a long time, the constructive and destructive interference averages out to no interference. We can then calculate what the interferogram of many incoherent sources combined is by computing the interferogram of a point source and adding the power of many such interferograms together.

To figure out what the power of an interferogram of an arbitrary object is, first consider what the interferogram of a single point source is. Consider a situation depicted in Fig. 5, but with a single point illuminator at the source plane positioned at (x, y, z) with radiant power I . We will have a receiving plane at a distance z from the source plane with two telescopes located symmetrically about the origin at $(\Delta x/2, \Delta y/2, 0)$ and $(-\Delta x/2, -\Delta y/2, 0)$. The fields received at the telescopes from the point will be interfered together at the center of the baseline, which is at the origin. We will assume that the point source produces a narrow band of radiation centered at wavelength λ . The point source produces a spherical wave that is received at both telescopes. The interference power is proportional to the two intensities measured at each telescope with the relative phase at the two telescopes causing constructive or destructive interference:

$$P \propto I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \phi \quad (1)$$

where I_1 and I_2 are the intensities of the field at each telescope, and ϕ is the phase difference between the two fields at each telescope. We will assume that the point (x, y, z) is very far away from the telescopes, and so the intensities at the two telescopes are equal and proportional to the point intensity I , so that $P \propto I(1 + \cos \phi)$. The phase ϕ is given by the difference in distance from the the source point at (x, y, z) to the receivers at $(\Delta x/2, \Delta y/2, 0)$ and $(-\Delta x/2, -\Delta y/2, 0)$. The phase is given by the difference in the Euclidean distances:

$$\phi = \frac{2\pi}{\lambda} (\sqrt{z^2 + (x + \Delta x/2)^2 + (y + \Delta y/2)^2} - \sqrt{z^2 + (x - \Delta x/2)^2 + (y - \Delta y/2)^2}) \approx \frac{2\pi(x\Delta x + y\Delta y)}{z\lambda} \quad (2)$$

where the second term gives the difference in the distance to a first-order approximation in x and y . The interference intensity relation is then $P \propto I(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda))$.

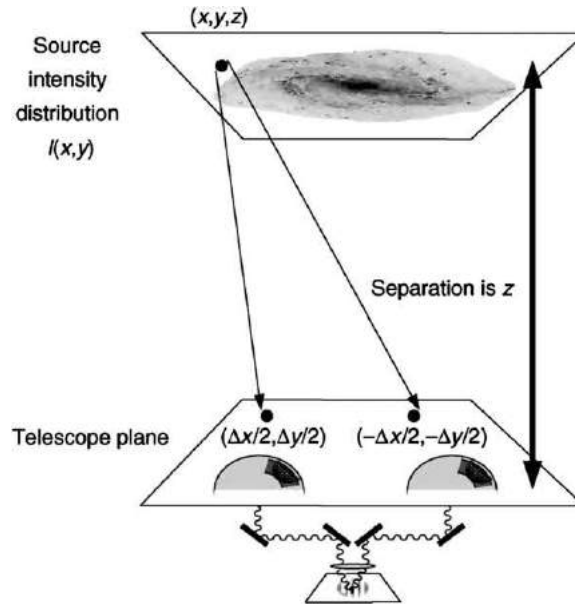


Fig. 5 Diagram of source and telescope plane used for derivation of Van Cittert-Zernike theorem. A source at a distance z from the telescope plane has a power density $I(x, y)$. The telescope plane has two movable apertures symmetric about the origin of the telescope plane separated by $(\Delta x, \Delta y)$. The light from both apertures is directed to the center of the baseline at the origin of the telescope plane, where it is interfered on the sensor.

If the source is now of a finite size, we can regard it as radiating an intensity (power per unit area) $I(x, y)$. To compute the total intensity of the interferogram, we integrate over all the sources in the source plane:

$$P(\Delta x, \Delta y) \propto \int I(x, y) \left(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda) \right) dx dy = \int I(x, y) dx dy + \int I(x, y) \cos(2\pi(x\Delta x + y\Delta y)/z\lambda) dx dy \quad (3)$$

This result can be interpreted as follows. The intensity of the interference when the two telescopes are separated by a distance $(\Delta x, \Delta y)$ is proportional to the real part of the two-dimensional Fourier transform of the intensity distribution $I(x, y)$ evaluated at spatial frequencies $(\Delta x/z\lambda, \Delta y/z\lambda)$, plus a constant intensity. This result, related to the Van Cittert–Zernike theorem, is a commonly used result in interferometric imaging.

By moving the telescopes to various horizontal and vertical separations and measuring the interference intensity, the real part of the Fourier transform of the intensity is inferred. In practice the imaginary part can also be found by varying the baseline optical path length difference by a quarter wavelength. With these samples of the Fourier transform, the inverse Fourier transform computes the image intensity distribution $I(x, y)$. This is how simple computational image formation occurs with interferometric data.

One application of the Van Cittert–Zernike theorem and interferometric imaging is the measurement of the diameter of stars. When observing coherence, the correlations between field points tend to decrease as the separation between the points grows. By observing this decrease of the coherence, the angular size of a star can be measured. This method was pioneered by Albert Michelson in 1890 with the invention of the Michelson Stellar Interferometer. It is needed because the aperture size required to measure the angular diameter of most nearby stars in the Milky Way is larger than the size of most telescope apertures. Stars usually have a circular profile that produces a circularly symmetric pattern of correlations. The coherence between field points as the distance between them is increased does not decrease monotonically, but disappears at certain separations, and has many revivals. The first separation distance at which interference disappears (and the degree of coherence is zero) for a circular profile object is $d_0 = 0.61\lambda/\alpha$, where λ is the wavelength of the light measured from the star, α is the angular diameter (in radians), and d_0 is the first separation from zero at which the interference disappears. In practice, each wavelength produces a null in interference at a different separation, so all but a small bandwidth of the star's light is filtered out before measurement to achieve a suitably strong interference null. For example, for the star Sirius, which has an angular diameter of 0.0068 arc seconds viewed from the Earth, measuring at a wavelength of $\lambda = 500$ nm (blue-green in the visible spectrum), the first separation distance at which no interference will be observed is at 18.5 m. This is larger than the largest available telescope aperture, and therefore an image cannot be directly formed that can resolve the size of Sirius.

Modern examples of stellar interferometry include the European Southern Observatory in Paranal, Chile, the CHARA Array at Mt Wilson, California, USA, and the NASA Jet Propulsion Laboratory Keck Interferometer in Hawaii, USA. The ESO main telescope system is the Very Large Telescope Interferometer (VLTI) consisting of four 8.2 m diameter telescopes as well as several smaller ones that can be moved independently, and will achieve milliarcsecond resolution through the interferometric combination of the telescopes. These telescopes are connected by underground tunnels through which the collected light from each telescope is combined. To measure the phase of the interference, extreme positioning precision is required in the relative delay between the light collected by the telescopes along the baseline, of 50 nm over 120 m, or 1 part in 2400 million, equivalent to 1.6 cm over the circumference of the Earth. Achieving this requires combinations of electronic and mechanical feedback systems with laser stabilized interferometric position measurement. The CHARA array has a very large 330-meter baseline, with 1-meter telescope apertures. The Keck Interferometer consists of two 10 m diameter telescopes with a baseline separation of 85 m. One of the main goals of achieving such high resolution is to be able to observe planets and planet formation around distant stars, to identify possible signs of life in other solar systems.

A similar principle is used for interferometric image formation in radio astronomy. The chief differences are the methods used to collect, detect, and correlate the electromagnetic radiation. Radio astronomical observatories, such as the Very Large Array in New Mexico, USA, and MERLIN at the Jodrell Bank Observatory in Cheshire, UK, use collection dishes and radio antennas to directly measure the electric field, which is not feasible at optical frequencies. These fields are converted to electrical voltages, which can then be directly correlated using electronic mixers, rather than detecting the correlation indirectly through interference. Because the wavelength is much larger, the stability requirements are greatly relaxed. Radio and optical frequency interferometric images are formed in essentially the same way, using the Van Cittert–Zernike theorem.

Interferometric imaging methods have also become of interest for microscopy applications. Aperture synthesis is not needed in microscopy, because the objects of interest are very small and therefore single apertures can be used to achieve diffraction-limited resolution. However, interferometric detection of partially coherent light affords greater flexibility because an image may be computed from the data rather than being directly detected. Typically this is achieved by creating an incoherent hologram of the object, typically through shearing interferometry. For example, by using a Rotational Shearing Interferometer, one can measure an incoherent hologram that can be used to produce a representation of an object with a greatly increased depth-of-field. In addition, direct detection of the partial coherence of sources can allow the methods of computed tomography to be used to produce three-dimensional images of volumes of incoherently emitting objects. Because of the versatility that interferometric detection can provide, interferometric imaging will surely be increasingly used outside of astronomy.

See also: Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

Further Reading

- Baldwin, J.E., Haniff, C.A., 2002. The application of interferometry to optical astronomical imaging. *Philosophical Transactions of the Royal Society of London A* 360, 969–986.
- Beckers, J.M., 1991. Interferometric imaging with the very large telescope. *Journal of Optics* 22, 73–83.
- Born, M., Wolf, E., 1997. *Principles of Optics*. New York: Cambridge University Press.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*. New York: McGraw-Hill Inc.
- Goodman, J.W., 2000. *Statistical Optics*. New York: Wiley-Interscience.
- Mandel, L., Wolf, E., 1995. *Optical Coherence and Quantum Optics*. New York: Cambridge University Press.
- Marks, D.L., Stack, R.A., Brady, D.J., Munson Jr, D.C., Brady, R.B., 1999. Visible cone-beam tomography with a lensless interferometric camera. *Science* 284, 2164–2166.
- Monnier, J.D., 2003. Optical interferometry in astronomy. *Reports on Progress in Physics* 66, 789–857.
- Potluri, P., Fetterman, M.R., Brady, D.J., 2001. High depth-of-field microscopic imaging using an interferometric camera. *Optics Express* 8, 624–630.
- Thompson, A.R., Moran, J.M., Swenson, G.W., 2001. *Interferometry and Synthesis in Radio Astronomy*. New York: Wiley-Interscience.

Multiplex Imaging

A Lacourt, Université de Franche-Comté, Besancon, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The ability to project an image onto a screen in a darkened room has been known since antiquity and Leonardo da Vinci (1515) precisely described principles of the Camera oscura. When one exposes a paper sheet directly to sunlight, it is uniformly lighted. However, if one pierces a little hole through the paper sheet then an image of the sun can be displayed on a screen beyond the hole (Fig. 1). Similarly, the shape of the light spot we can see in a dark room, from the rays passing through a small aperture of a closed shutter, is not that of the aperture. Generally, the spot has a circular shape, as an image of the sun. Actually, it is neither a shadow of the hole nor an image of the sun but a combination of both. Light incident on an aperture contains the entire information about the light distribution of the object, but this information has to be 'deblurred', or decoded. The aperture acts as an elementary spatial 2D-filter that selects and preserves the direction and intensity of light rays coming from each point of the scene. The projection on the screen of each pencil of light passing through the hole is a small light spot, the geometry of which is the same as the aperture, and the intensity is proportional to that of the emitting point. The smaller the hole, the more detail the image exhibits, but with less luminosity. Thus, we must adopt a compromise between the resolution and the illumination of the image.

Now, if the mask has several holes, the amount of light reaching the screen, being proportional to the whole aperture area, is extensively increased. On the other hand, as each hole displays an image on the screen, the overlapping of these multiple images gives a new but blurred image. However, the information exists and we should be able to retrieve it if we know the characteristics of the complex aperture. This is the basic concept of multiplex imaging. The crucial question is: what is a good aperture design and a good way to reconstruct images having high resolution and with high illumination levels? The suitable masks with this aim are named synthetic apertures.

Porta (1589) and Kepler (1611) found the perfect solution, by setting up a lens in front of a large hole. Indeed, one can regard a lens as a wide aperture made of continuously juxtaposed microprisms (Fig. 2). The lens obtains such refractive properties that the image displayed by each of these microprisms is moved toward a center. Moreover, as the angle of the microprisms varies continuously, it focuses on a single point, all the rays coming from a given object point. Thus, the blurring is suppressed and the illumination is increased. The way was opened toward the invention of optical instruments and photography and the 'pinhole' has fallen into oblivion for a long time.

Two new problems awake interest in synthetic apertures, around 1960. First, the domain of imaging has extended out of visible light: infrared and ultraviolet light, γ -rays, X-rays, ultrasound waves, and others. Materials suitable for making lenses or mirrors, operating with such waves, are expensive or often do not exist. For example, most of the transparent materials have a refractive index close to 1 for γ -rays and X-rays. However, some are opaque enough to enable the use of masks.

The second challenge appeared with progress in astronomy. In order to explore the universe more, further from Earth and closer to the Big Bang, astrophysicists needed telescopes providing very high luminosity and high resolution, since distant stars, nebulae, and other galaxies are extremely weak and small luminous objects. This exploration can be achieved by increasing the diameter of the primary mirror of telescopes. Indeed, due to diffraction of light by the pupil of the primary mirror, the theoretical limit of resolution of a telescope (i.e., the smaller distance between two points that one can discriminate), is inversely proportional to the diameter of the mirror. In addition, the illumination is proportional to its area. However, the working of large mirrors is a technical prowess that is very expensive and has high risks. Defect of a fraction of micrometer when polishing the surface can have dramatic consequences, e.g., Hubble. Also, such mirrors, being heavy, can bend out of shape under their own weight. Nevertheless, outstandingly large telescopes have been implemented, based on Adaptive optics: the deformations of the mirror are measured in

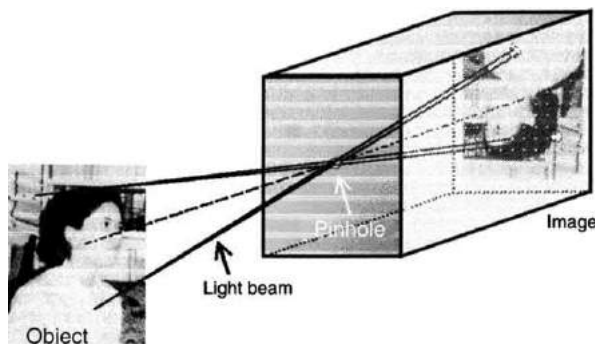


Fig. 1 Principle of darkroom.

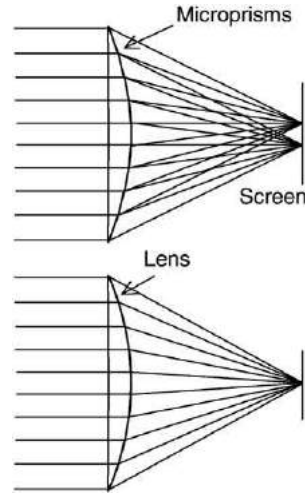


Fig. 2 A lens can be modeled as a juxtaosition of microprisms with continuously variable angles.

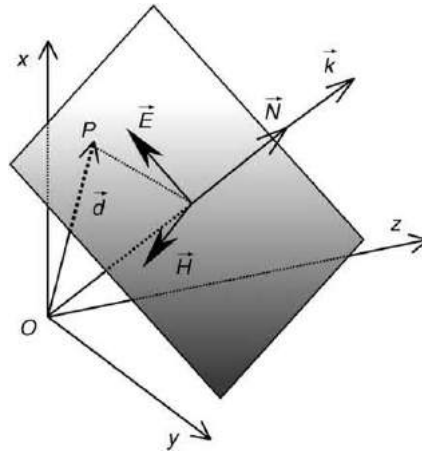


Fig. 3 Electromagnetic wave propagation.

real time by an interference method and compensated by activating a mosaic of jacks (New Technology Telescope – NTT – European program, 1989). Another way consists in replacing the solid-state mirror by a mosaic of 36 mirrors accurately directed (Keck telescope, USA). However, in either case, the mirror diameter does not exceed 8 to 10 meters. Synthetic apertures could be a solution to that difficulty, as earlier demonstrated by Fizeau (1868) and Michelson (1890).

In the first part of this article, any fundamentals are briefly recalled about properties of light and imaging. The second part describes recent experimental methods and results obtained in synthetic aperture imaging. The last part reports the important programs in progress today relating to stellar interferometry.

Background

Electromagnetic Waves Propagation

Light power, which is involved in image formation, is carried by the electric field of an electromagnetic wave (Fig. 3). The electric field $\vec{E}$ has properties of a vector orthogonal to the direction of propagation $\vec{N}$ of the light ray. It is characterized by its amplitude $\vec{E}_0$, the angular time frequency of vibration ω , and the wavevector $\vec{k}$ which specifies the propagating direction of the ray. Let $P(\vec{d})$ be a current point on a plane wavefront, defined by the position vector $\vec{d}$ from the origin of the reference frame. The electric field propagation of a monochromatic wave is expressed as:

$$\vec{E}(\vec{d}, t) = \vec{E}_0 \exp[i(\vec{k} \cdot \vec{d} - \omega t)] \quad (1)$$

The light power carried by the electromagnetic field is proportional to $I = \|\vec{E}_0\|^2 = \vec{E} \cdot \vec{E}^*$ where $\vec{E}^*$ is the complex conjugate of $\vec{E}$. The time frequency $\nu = \omega/2\pi$, or the wavelength $\lambda = c/\nu$, where c is the electromagnetic wave celerity, characterize the hue of the

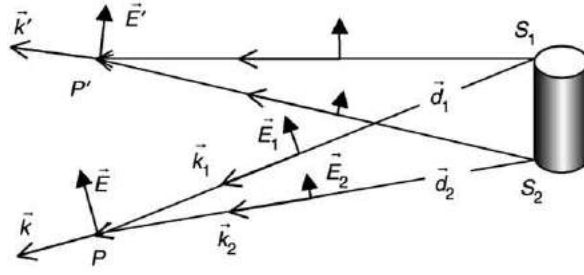


Fig. 4 Spatially coherent light.

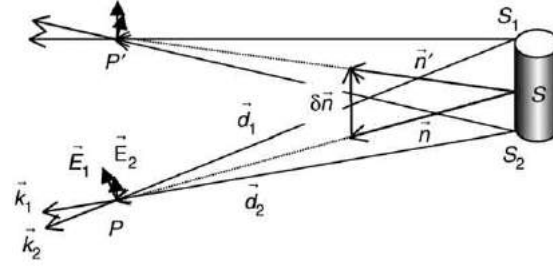


Fig. 5 Spatially incoherent light.

light. The wavevector $\vec{k} = \vec{N}\omega/c$ where $||\vec{N}|| = 1$ is an angular space frequency vector, as $\vec{n} = \vec{N}/\lambda = \vec{N}v/c$ is a space frequency vector. We have $\vec{k} = 2\pi\vec{n}$. As λ is a constant for a given monochromatic light, $\vec{n}$ defines the direction of propagation of a ray.

Spatially Coherent Light

Consider a very small source stated as the point source S_1 , emitting uniform lightwaves in any direction (Fig. 4). The wave reaching a farfield point P , from S_1 , propagates along a direction defined by the wavevector $\vec{k}_1$. The vectors $\vec{k}_1$ and $\vec{d}_1 = S_1P$ are collinear. Another point source, S_2 is characterized from P by the wavevector $\vec{k}_2$ such that $\vec{k}_2$ and $\vec{d}_2 = S_2P$ are collinear. When the distances d_1 and d_2 are close to each other, the amplitudes $\vec{E}_{01}$ and $\vec{E}_{02}$ of the waves at P are proportional to those of the corresponding emitting sources, S_1 and S_2 . According to the principle of superposition, the resulting field at the point P , $\vec{E} = \vec{E}_1 + \vec{E}_2$ and the instant wavevector $\vec{k}$ are unique vectors.

If S_1 and S_2 are two synchronous sources, the amplitude $\vec{E}_0$ of $\vec{E}$ and the wavevector $\vec{k}$ are constant versus time at any point P . This is still the case when an extended object is made of an arrangement of synchronous points. That situation defines a spatially coherent object. Thus, one cannot recover multiple data about the object's distribution, from the field framework at a single point P , since the relationship between the two is multi-unequivocal. Nevertheless, due to diffraction, one can retrieve them by knowing the complex amplitude distribution of light on a set of wavevectors available over an extensive area of a screen. Let $\vec{E}_0(\vec{r})$ be the amplitude of the electric field at the object point $M(\vec{r})$ and let $\vec{E}_0(\vec{n})$ be the one diffracted by the object to farfield toward the direction corresponding to the space frequency $\vec{n}$. Fourier transforms give the connection between $\vec{E}_0(\vec{r})$ and $\vec{E}_0(\vec{n})$:

$$\left\{ \begin{array}{l} \vec{E}_0(\vec{n}) = \int \vec{E}_0(\vec{r}) \exp[i2\pi\vec{n} \cdot \vec{r}] d\vec{r}, \\ \vec{E}_0(\vec{r}) = \int \vec{E}_0(\vec{n}) \exp[-i2\pi\vec{n} \cdot \vec{r}] d\vec{n}. \end{array} \right\} \quad (2)$$

This is the situation encountered when an object is illuminated by a laser or by a point source such as a single star.

Spatially Incoherent Light

When the point sources S_1 and S_2 are independent – spatially incoherent – both wavevectors $\vec{k}_1$ and $\vec{k}_2$ coexist at the point P , each characterizing the angular position of S_1 and S_2 (Fig. 5). If S_1 and S_2 are part of an extended intensity distribution $I_0(\vec{r})$ fans of wavevectors and fields are mixed at P , each coming from an independent point source S . Together, they carry the whole data on the object distribution. Those data are partially uncorrelated at any two points P and P' on a screen, related to the different optical paths from any point S to P and P' and to the finite time of coherence of light. The correlation between data available to farfield in two directions $\vec{n}$ and $\vec{n}'$ from the object, is given by the mutual intensity function, $\Gamma(\delta\vec{n}) = \int \vec{E}(\vec{n}) \cdot \vec{E}^*(\vec{n} + \delta\vec{n}) d\vec{n}$ where $\delta\vec{n} = \vec{n}' - \vec{n}$. Fourier transforms give the relation between $I_0(\vec{r})$ and $\Gamma(\delta\vec{n})$:

$$\begin{cases} \Gamma(\delta\vec{n}) = \int I(\vec{r}) \exp[i2\pi\delta\vec{n} \cdot \vec{r}] d\vec{r}, \\ I(\vec{r}) = \int \Gamma(\delta\vec{n}) \exp[-i2\pi\delta\vec{n} \cdot \vec{r}] d\delta\vec{n} \end{cases} \quad (3)$$

Therefore, by drawing a map of $\Gamma(\delta\vec{n})$ one can recover the object distribution $I_0(\vec{r})$ by a Fourier transformation. Most of time, one uses the spatial degree of coherence function, $\gamma(\delta\vec{n}) = \Gamma(\delta\vec{n})/\Gamma(\vec{0})$. When the object illuminates a screen, at a large distance d , the degree of coherence of light between two points P and P' on the screen is $\gamma(\delta\vec{p}/\lambda d)$ with $\delta\vec{p} = \vec{PP'}$. The 2D extension of $\gamma(\delta\vec{p}/\lambda d)$ defines the area of coherence of light on the screen. The width of the area of coherence, $\Delta\rho$, and that of the object distribution, Δr , are inversely proportional:

$$\Delta\rho \cdot \Delta r \sim \lambda d \quad (4)$$

If a mask perforated by two pinholes, separated by a distance $\delta\vec{p}$ replaces the screen, one can observe Young fringes of interference beyond the mask, the visibility of them being precisely $\gamma(\delta\vec{p}/\lambda d)$. Therefore, one can build a map of $\gamma(\delta\vec{p}/\lambda d)$ versus $\delta\vec{p}$ from a set of 2D measurements of visibilities of fringes, then reconstruct the object distribution by a Fourier transformation.

Notice that the entire data of the object distribution is available from a single coherence area. That is, a pinhole covering exactly one area of coherence would select the entire available data from the object and would display to farfield the most accurate image possible. Generally, such a pinhole is so small that practically no light passes through it. However, we can regard each area of coherence as an independent channel of information. Then an imaging device based on a juxtaposition of such redundant channels can be considered as a multiplex imaging device.

Self-luminous objects such as extended or double stars, and artificial thermic sources are spatially incoherent sources.

Multiplex Imaging

Methods of multiplex – or lensless – imaging are performed either with coherent or incoherent light. The very different properties of the two types of light leads to completely different methods, even if all are related to holography. Holography with coherent light is an interference method for local recording of both the direction and modulus of the wavevectors and the amplitude of the field contained in many juxtaposed grains of speckle, scattered by the object. Each grain is an independent channel of imaging.

Synthetic Apertures with Spatially Incoherent Objects

As interferences are impossible by this method with incoherent light, we must improve other methods of imaging. Synthetic aperture is a concept consisting of using a redundant multichannel system to increase the luminosity of images without loss of resolution. The basic principle is that of a multiple holes darkroom. The processing is in two steps: recording and reading the image. The recording step consists in illuminating a photosensitive receptor with the object through a mask (Fig. 6). Generally, the receptor is a photographic plate. Obviously, the light distribution on the plate is blurred. Indeed, any elementary transparencies on the mask project on the receptor inverted images of the object located the shadow of the mask. Such a distribution is expressed by a convolution. Let $O(x,y)$ be the conical projection of the object on the receptor and $Z(x,y)$ that of the mask. The shadows distribution on the receptor is (the variable being normalized):

$$I'(x',y') = Z(x,y) * O(x,y) = \int Z(x,y) O(x' - x, y' - y) dx dy \quad (5)$$

The reading step is a decorrelation processing from the recorded intensity (Fig. 7). It consists in achieving a new correlation of I' with an arbitrary function T :

$$I''(x,y) = T(x,y) * [Z(x,y) * O(x,y)] = [T(x,y) * Z(x,y)] * O(x,y) \quad (6)$$

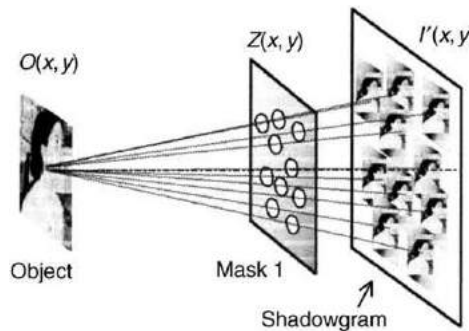


Fig. 6 Principle of synthetic aperture: recording.

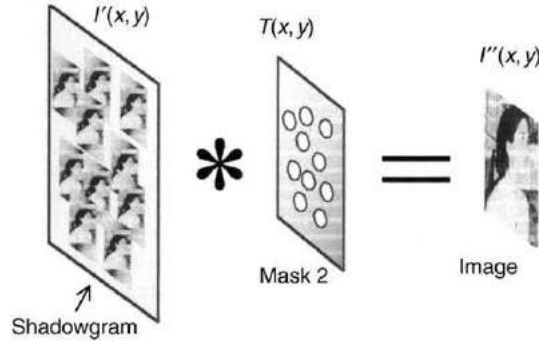


Fig. 7 Synthetic aperture: reading.

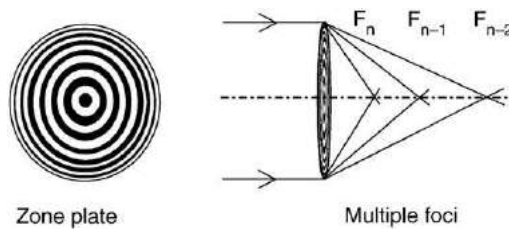


Fig. 8 Properties of FZP.

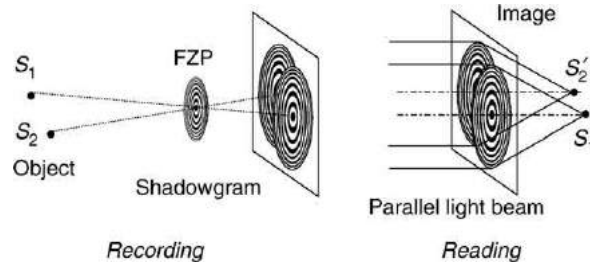


Fig. 9 Recording and analogical reading with a FZP.

$R[x, y] = [T(x, y) * Z(x, y)]$ is the system point spread function (SPSF) of the processing. It expresses the image distribution of a single-point. The perfect image is obtained when $R(x, y)$ is a 2D Dirac – or delta– distribution, $\delta(x, y)$:

$$R(x, y) = \delta(x, y), \text{ as } \int \delta(x, y) O(x' - x, y' - y) dx dy = O(x', y') \quad (7)$$

Fresnel zone plates

First Mertz and Young used Fresnel zone plate (FZP) as a synthetic aperture for X-ray imaging purposes. FZPs are made of alternate transparent and opaque concentric crowns of equal area. The radius of the n th circle is $r_1 \sqrt{n}$, r_1 being the smallest. When illuminated by a cylindrical light beam, the FZP is a multiple foci diffractive lens (Fig. 8).

The reading step is either analogical or numerical. Through the analogical method, we consider the recorded plate as a pseudohologram (Fig. 9). When the 'shadowgram' is lighted with a parallel beam of laser light, the FZPs recorded around each point concentrates the light on its foci. Then an image of the object is reconstructed on the focal planes; a method that is simple and elegant. Nevertheless, the processing is not linear since we record an intensity distribution at the recording step, while the reading step by diffraction involves amplitude. Moreover, the S/N ratio decreases as the amount of data increases, recorded on the photographic plate because of its finite dynamic.

Mertz and Young had the idea of using a second FZP to perform the deconvolution processing by optical methods. In this case, the SPSF, $R(x, y)$, is the autocorrelation function of the FZP transparency (Fig. 10). It presents a narrow central peak, emerging from a pyramidal ground (Fig. 11). The optical setup is the same as the recording step. The FZP scale factor only must be adjusted to compensate for the magnification ratio generated by the conical projection at the recording step.

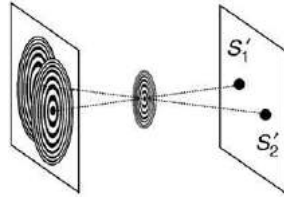


Fig. 10 Optical decorrelation.

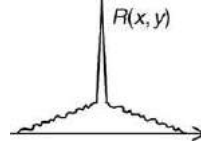


Fig. 11 System point spread function with a Fresnel zone plate.

In either digital or analogical deconvolution, the image reconstructed is accurate, with good sharpness, if the number of zones is sufficient. Ceglio *et al.* obtained relevant images of 8 μm targets with FZP having 100 zones. Nevertheless, the multiple foci of FZP induces decreasing of the S/N ratio which is an important limitation. Besides, we pointed out above that the larger the object, the more the decrease in the S/N ratio.

Random and uniformly redundant arrays

Dicke and Ables substituted a random array to the FZP. The mask is made of randomly distributed holes and the expected advantage is doubled. First, one can reduce the transparent area of the mask by less than 50% in order to increase the S/N ratio when the object is wide. Second, the autocorrelation function of such an array is theoretically a Dirac distribution when its width is infinite. Thus, one can hope to narrow the system point response. However, when the width is finite, the pyramidal ground of the point response mentioned above is present and the advantage is not clearly demonstrated. In order to correct this effect, Chris Brown used a 'mismatched' function, for the postprocessing deconvolution. This function is calculated, first by achieving the complement of the recording mask function, then subtracting from it. Thus, the 1 transparencies result in $+1$ while the 0 transparencies are changed by -1 . Brown shows images digitally reconstructed of X-ray sources, 30 sec. of arc wide. Fenimore and Cannon conceived a uniformly redundant array. The center random motif of the mask is half replicated on each side with only the center area being used for the recording step. The deconvolution is calculated by taking into account the whole area in a mismatched way. The result is an exact delta function. The authors show images of 200 μm -wide microspheres with a similar resolution but a signal 7100 times stronger than with a comparable pinhole camera.

Theta Rotating Interferometer

The theta shearing interferometer enables us to record complete data from the Fourier transform of a spatially incoherent object (Fig. 12). The source can be any natural one, but monochromatic. This is easily practicable using a good filter or monochromator.

The device is based on a Michelson interferometer, the plane mirrors being replaced by two roof prisms. One of the roof prisms is fixed so the other can rotate around the axis of the interferometer. As a result, we have two images of the object. Let θ be the angle between the two edges of the prisms. The angle between the images is 2θ . Since the original object is noncoherent, and two homologous points being mutually coherent, the light they emit interferes, resulting in fringes.

The intensity, space frequency, and direction of the fringes characterize the intensity and the position of the point source. One shows that the farfield visibility distribution of fringes $V(\delta\vec{n})$ and the object distribution $O(\vec{r})$ are related by Fourier transforms:

$$\begin{cases} V(\delta\vec{n}) = \int O(\vec{r}) \exp \left[i 2\pi \frac{\delta\vec{n} \cdot \vec{r}}{2\sin\theta} \right] \frac{d\vec{r}}{2\sin\theta}, \\ O(\vec{r}) = \int V(\delta\vec{n}) \exp \left[-i 2\pi \frac{\delta\vec{n} \cdot \vec{r}}{2\sin\theta} \right] d\delta\vec{n} \end{cases} \quad (8)$$

Then, one can see that the visibility of the interference pattern is an image of the degree of coherence $\gamma(\delta\vec{n})$ with a scale factor $1/2 \sin \theta$.

Such a pattern recorded on a photographic plate is a Fourier hologram of the noncoherent object. An image can be reconstructed by diffraction at infinity or near a focus (see Fig. 13, examples performed in 1972 by the author). The technique was successfully applied to the determination of the modulation transfer function of optical systems. C Roddier and F Roddier applied

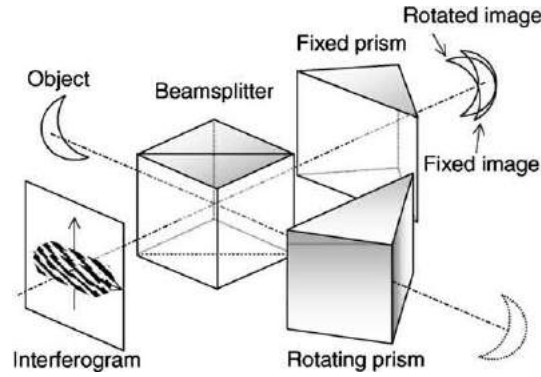


Fig. 12 Theta shearing interferometer.

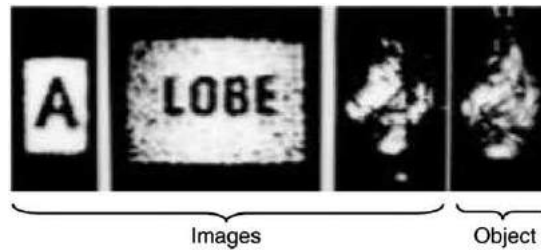


Fig. 13 Incoherent light holograms.

the method to astronomy. For instance, they examined Alpha Orionis and discovered the evolution of its dust envelope and the possible existence of a companion.

High Resolution Imaging in Astronomy

In astronomy, the limit of resolution depends chiefly on two factors. The first is a technical one. Adaptive optics and very large telescopes (VLT) brought spectacular progress to clarity and resolution of telescopes. The 8.2 m VLT of Kueyen set up in 1989, shows images easily resolved of HIC 69495 Centauri, separated by 0.12 sec. of arc. Nevertheless, working of large mirrors for telescopes reaches its limits. Therefore, their clarity remains too weak. On the other hand, atmospheric turbulences make the refractive index of air unstable and the position of the images in the focal plane of telescopes unsettled. Thus, the point response is a granular spot named 'speckle'. Short exposure times relieve that inconvenience, but they prohibit the observation of weak stars. In 1970, Labeyrie proposed a method of speckle interferometry consisting of a statistical analysis of a set of speckles. The result is a display of a diffraction-limited image. He resolved many tens of single and double stars. Only Hubble, above the atmosphere, is able to do it by direct imaging. However, the primary mirror of Hubble being only 2.40 m wide, has clarity much lower. Stellar interferometry is a very promising alternative solution.

Principles of Stellar Interferometry

Stellar interferometry is based on Eq. (3). The light emitted by double or wide stars is noncoherent, so that one can access an image of them through 2D measurements of the degree of coherence of light. We showed above that one could achieve this by using interference devices such as Young holes. Eq. (3) shows that if we measure the visibility of Young fringes at increasing distances $\delta\rho$ between the pinholes, the distance $\Delta\rho$ from which the visibility is zero gives the angular width of the star, $\Delta\alpha$, under the formula:

$$\Delta\rho \cdot \Delta\alpha \sim \lambda \quad (9)$$

where $\Delta\alpha = \Delta r/d$. In the case of a double star, the visibility varies periodically versus $\delta\rho$ when the pinholes are aligned along the axis of two stars. Let $\Delta\alpha$ be the angular distance between the stars: the smallest distance $\Delta\rho$ for which the visibility is zero is given by:

$$\Delta\rho \cdot \Delta\alpha = \lambda/2 \quad (10)$$

The larger the uttermost distance between the holes, the more the limit of resolution of the processing is potentially sharp.

In order to exploit this idea, Stephan placed and rotated a mask with two holes in front of the objective of a refractive telescope. He established that the diameter of the stars he observed was much less than 0.16 sec. of arc. Then, Michelson built up, on top of a

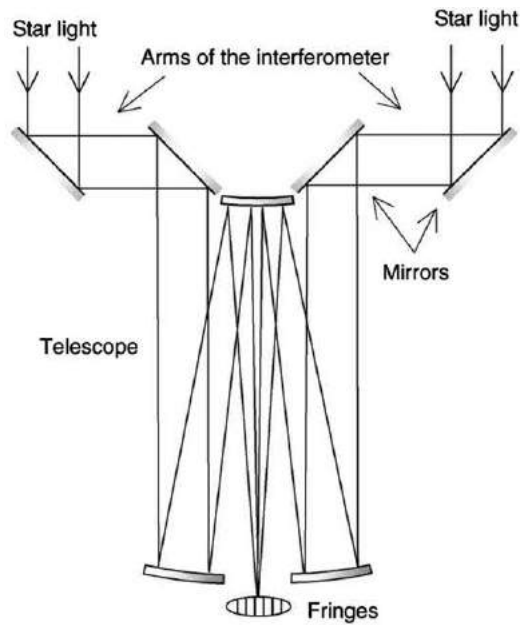


Fig. 14 Stellar interferometer of Michelson.

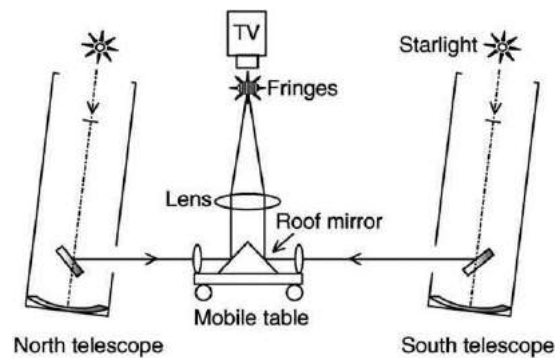


Fig. 15 The two telescopes interferometer of Labeyrie.

30 cm refractive telescope, a stellar interferometer consisting of two arms catching the light at 7 m of distance from each other (Fig. 14). He measured correctly the diameter of the satellites of Jupiter. Hale carried the distance of the base up to 15 m in 1920.

The Two Telescopes Interferometer of Labeyrie

In 1975, Labeyrie improved the idea by setting up two separate 25 cm diameter Cassegrain-coudé telescopes on a 12 m baseline, oriented north–south (Fig. 15). The setting and mechanical stability of the device are extremely sensitive. The tolerances do not exceed any μm . A mobile table near the common focal plane of the two telescopes compensates for the optical path differences due to stars tracking. In spite of the acute difficulties, he succeeded in observing fringes on Vega. He obtained a resolution of 0.01 sec. of arc versus 0.5 sec. with a single 25 cm telescope. Later, one telescope was rendered mobile bringing the base width variable between 4 and 67 m. The resolution reaches up to 0.00015 sec. of arc. Note that the rotation of Earth also enables east–west analysis with resolution of 0.005 sec., and today, the 25 cm telescopes are replaced by 1.5 m ones.

The Very Large Telescope Interferometer (VLTi)

The two telescopes interferometer of Labeyrie is a first step toward a very ambitious project: the telescope at synthetic aperture and very large base. Proofs of reliability of such a principle was already given by the very large array (VLA) experimented in radio astronomy. Labeyrie with Lena, and the team of the ESO (Ecole Supérieure d', Paris), imagined a very large telescope interferometer (VLTi) made of 6 telescopes of 2 m, in a 2D structure on a base of 100 m. The European VLTi was built at Cerro Paranal. It is made of four 8.2 m telescopes corrected in real time by adaptive optics (NTT technology) and three 1.8 m telescopes. The

fourth 8.2 m telescope was put in service in 2001. Some days, each of the 8.2 m telescopes used separately gives a resolution close to that of Hubble. However, the clarity is much better because of the larger primary mirror diameter (8.2 m against 2.4 m). In May 2003, the team of J-G Cuby observed a galaxy the farthest away ever seen, only 900 millions years old since the Big Bang. Obviously, a drastic gain of resolution (by a factor 10) is expected from the interference configuration.

Further Reading

Labeyrie, A., 1976. High resolution techniques in optical astronomy. In: Wolf, E. (Ed.), *Progress in Optics*, vol. XIV. New York: North-Holland, pp. 49–85.
Mertz, L., 1965. *Transformations in Optics*. Amsterdam, Oxford: John Wiley & Sons, Inc.

Photon Density Wave Imaging

V Toronov, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

μ_a [cm^{-1}] Absorption coefficient
Alternating part of PDW, alternating current (ac)
 g_1 Average cosine of the scattering angle
 ω [s^{-1}] Circular modulation frequency
Computed tomography (CT)
 $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$
 $[\text{cm}^{-1}]$ Difference between the estimate and actual optical properties
 $D = v/(3\mu_a + 3\mu'_s)$ [cm^2/s] Diffusion coefficient
Diffusion equation (DE)
 $S(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-3} \text{steradian}^{-1}$] Distribution of light source power
 ∇ [cm^{-1}] Divergence operator
 $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ [cm^{-1}] Estimate of optical properties
 F General nonlinear operator
 J Jacobian
 Δ [cm^{-2}] Laplace operator
Magnetic resonance imaging (MRI)
 $\{\mu_a^m, \mu_s^m\}$ [cm^{-1}] Measured bulk optical properties (absorption and reduced scattering coefficients)
 ν [s^{-1}] Modulation frequency
Non-alternating part of PDW, direct current (dc)

$\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}$ [cm^{-1}] Optical properties of the medium (absorption and reduced scattering coefficients) as functions of spatial coordinate
 $\phi(\mathbf{r})$ Phase as a function of source–detector distance
 $f(\hat{\Omega}, \hat{\Omega}')$ Phase function
 $U(\mathbf{r}, t)$ [J cm^{-3}] Photon density
Photon density wave imaging (PDWI)
Photon density wave (PDW)
 l [cm] Photon mean free path
 $L(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-2} \text{steradian}^{-1}$] Radiance
 $\Phi(\mathbf{r}, t)$ [W cm^{-2}] Radiation fluence
 $J(\mathbf{r}, t)$ [W cm^{-2}] Radiation flux
 μ'_s [cm^{-1}] Reduced scattering coefficient
 μ_s [cm^{-1}] Scattering coefficient
 G Set of measured values (ac, dc, and phase)
 G_e Set of predicted values (ac, dc, and phase)
 $Q(\mathbf{r}, t)$ [$\text{J cm}^{-3} \text{s}^{-1}$] Source term
 r [cm] Source–detector distance
 $\mathbf{r}$ [cm] Spatial coordinate
 v [cm/s] Speed of light in medium
 $Y_{nm}(\hat{\Omega})$ Spherical harmonics
Transport equation (TE)
 $h(\mathbf{r})$ Weight function
 $x = \omega/(\nu\mu_a)$

Photon density wave imaging (PDWI) employs diffused light, usually in the near-infrared band, for imaging inhomogeneous structures in a highly scattering medium. The other names for PDWI are diffuse optical tomography and photon migration imaging. The term photon density wave (PDW) refers to the wave of light density spreading in the highly scattering medium from the source whose power changes in time. Therefore, rigorously speaking, photon density wave imaging uses intensity-modulated light sources, such as pulsed lasers, or sources whose power is harmonically modulated. However, many applications employ steady-state light sources. Most applications of PDWI are in the field of imaging of biological tissues, including medicine and cognitive neuroscience.

As in the other imaging methods, such as x-ray, computed tomography (CT) or ultrasound imaging, the problem of PDWI consists in the reconstruction of the internal structure of the medium from the measurements made outside the medium (on its surface). In the case of PDWI the internal structure of the medium is represented by the spatial distribution of its optical properties. The reconstruction procedure is essentially an iterative mathematical algorithm, which includes the modeling of the light propagation inside the medium assuming a certain distribution of the optical properties, and the correction of this distribution to match the model prediction with the actual measurement. The modeling part of the imaging problem is usually referred to as the forward problem, and the correction procedure is known as an inverse problem.

Forward Problem: Photon Density Waves

Unlike optically transparent media, in a highly scattering medium the coherence of light from the coherent source very rapidly decays with distance. This happens due to the interference of many randomly and multiply scattered light waves at every point of the medium that is distanced from the source much farther than the photon mean free path l , which is the mean distance at which a light wave undergoes a single scattering event. For biological tissues, where light scattering occurs on cell membranes and other heterogeneities of the index of refraction, l is typically about 1 mm. Although high scattering eliminates the coherence of the electromagnetic waves at the short time intervals corresponding to oscillation at optical frequencies ($\sim 10^{-15}$ s), in the case of an intensity-modulated light source, the macroscopic wave of the density of light energy, called 'photon density wave', forms in the highly scattering medium. The properties of PDW depend on the type of

the source modulation and on the local optical properties of the medium, which are the scattering coefficient μ_s [cm^{-1}] and the absorption coefficient μ_a [cm^{-1}].

Because of the complexity and incoherence of the macroscopic electromagnetic field in highly scattering media, i.e. media with $\mu_a \ll \mu_s$, the fundamental Maxwell equations of electrodynamics do not provide a convenient approach to the solution of the problem of light propagation in such media at distances much larger than l . Therefore, statistical approaches have been developed, such as the photon transport model and its diffusion approximation. In the photon transport model the light is characterized by the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-2} \text{steradian}^{-1}$] (where the steradian is the unit solid angle), which is the power of radiation near the point $\mathbf{r}$ propagating within a small solid angle around the direction $\hat{\Omega}$ at time t . The photon transport model accounts for the possible nonisotropy of the radiation flux. For the given spatio-temporal distribution of the optical properties, μ_a and μ_s , and for the given spatial and angular distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-3} \text{steradian}^{-1}$] the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ can be found from the linear radiation transfer equation, also called the photon transport equation, or simply the transport equation (TE):

$$\frac{1}{v} \frac{\partial L(\mathbf{r}, \hat{\Omega}, t)}{\partial t} + \nabla \cdot L(\mathbf{r}, \hat{\Omega}, t) \hat{\Omega} + (\mu_s + \mu_a) L(\mathbf{r}, \hat{\Omega}, t) = \mu_s \int L(\mathbf{r}, \hat{\Omega}', t) f(\hat{\Omega}, \hat{\Omega}') d\hat{\Omega}' + S(\mathbf{r}, \hat{\Omega}, t) \quad (1)$$

where $f(\hat{\Omega}, \hat{\Omega}')$ is the phase function, which represents the portion of light energy scattered from the direction $\hat{\Omega}$ to the direction $\hat{\Omega}'$ and v is the speed of light in the medium. TE is equivalent to the Boltzmann transport equation describing the transport of particles undergoing scattering.

Eq. (1) is difficult to solve even numerically. Therefore, a hierarchy of approximations to the PTM was developed. This hierarchy, known as the P_N approximation, is based on the expansion of the radiance in a series of spherical harmonics $Y_{nm}(\hat{\Omega})$ truncated at $n=N$. In the P_1 approximation the coefficients of the series are proportional to the radiation fluence

$$\Phi(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t) d\hat{\Omega} \quad (2)$$

and the flux

$$J(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t) \hat{\Omega} d\hat{\Omega} \quad (3)$$

both having units of [W cm^{-2}]. By substituting the P_1 series for $L(\mathbf{r}, \hat{\Omega}, t)$ and $S(\mathbf{r}, \hat{\Omega}, t)$ into Eq. (1) and assuming isotropic light sources and sufficiently slow change of source intensity, one can obtain the equation that includes only $\Phi(\mathbf{r}, t)$ as the unknown. Usually, this equation

$$v\mu_a U(\mathbf{r}, t) + \frac{\partial U(\mathbf{r}, t)}{\partial t} - D\Delta U(\mathbf{r}, t) = Q(\mathbf{r}, t) \quad (4)$$

is written in terms of the value $U(\mathbf{r}, t) = \frac{1}{v} \Phi(\mathbf{r}, t)$ [J cm^{-3}], which has units of energy density, and is called the photon density. In Eq. (4) $D = v/(3\mu_a + 3\mu'_s)$ [cm^2/s] is the light diffusion coefficient, and $\mu'_s = \mu_s(1 - g_1)$ [cm^{-1}] is the reduced scattering coefficient; g_1 is the average cosine of the scattering angle. The source term $Q(\mathbf{r}, t)$ [$\text{J cm}^{-3} \text{s}^{-1}$] in the right-hand side of the Eq. (4) describes the spatial distribution of light sources (in the isotropic approximation, i.e., neglecting the dipole and higher momenta of $S(\mathbf{r}, \hat{\Omega}, t)$) and the temporal modulation of source intensity.

Eq. (4) is mathematically equivalent to the equation describing the density of the medium consisting of the particles undergoing linear diffusion. Therefore, it is called the diffusion equation (DE), and the P_1 approximation is also known as the diffusion approximation. The diffusion approximation is valid under the following three conditions: (a) the light sources are isotropic or the distance between the source and the detector is much larger than l ; (b) characteristic source modulation frequencies ν are low enough that $v\mu'_s/(2\pi\nu) \gg 1$; and (c) $\mu_a \ll \mu'_s$. In the biomedical applications of PDWI the condition (a) is fulfilled when the source-detector distance is larger than 1 cm. The condition (b) is valid when the characteristic modulation frequencies of the light source are less than 1 GHz. The condition (c) is fulfilled in most biological tissues, for which the scattering is highly isotropic, i.e. $g_1 \ll 1$. Thus, the forward problem of PDWI, i.e., the modeling of light propagation, can be solved using the diffusion approximation in a large variety of situations.

PDWs described by Eq. (4) are the scalar damped traveling waves, characterized by the coherent changes of light energy (photon density) in different locations on time intervals from microseconds to nanoseconds. Usually pulsed or harmonically modulated light sources are used to generate PDWs. The properties of the PDWs from a harmonic source are fundamental, because the PDW from the source of any type of modulation can be considered as a superposition of harmonic waves from sources having different frequencies using the Fourier transform.

In order to get an idea of the basic properties of PDW, one can consider the analytical solution of (4) for the case of a harmonically modulated point-like source placed in an infinite homogeneous scattering medium. This solution is

$$U(r, t) = U_{dc}(r) + U_{ac}(r) \exp[-i\omega t + i\phi(r)] \quad (5)$$

where U_{ac} , U_{dc} , and ϕ are the ac, dc and phase, respectively, of the wave measured at the distance r from the source, and $\omega = 2\pi\nu$. U_{ac} , U_{dc} , and ϕ depend on the medium optical properties, source intensity Q_0 , modulation amplitude A_0 , and initial phase ψ_0 as

$$U_{ac}(r) = \frac{Q_0 A_0}{4\pi D} \frac{\exp\left[-r\left(\frac{v\mu_a}{2D}\right)^{1/2} \left((1+x^2)^{1/2} + 1\right)^{1/2}\right]}{r} \quad (6a)$$

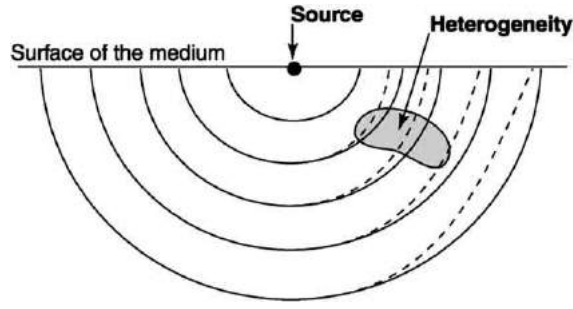


Fig. 1 Equiphasic surfaces of the PDW induced by a point source in a semi-infinite homogeneous medium (solid lines) and in a medium including heterogeneity (dashed lines).

$$U_{dc}(r) = \frac{Q_0}{4\pi D} \frac{\exp\left[-r\left(\frac{\nu\mu_a}{D}\right)^{1/2}\right]}{r} \quad (6b)$$

$$\phi(r) = r\left(\frac{\nu\mu_a}{2D}\right)^{1/2} \sqrt{(1+x^2)^{1/2} - 1} \quad (6c)$$

where $x = \omega/(\nu\mu_a)$. Eqs. (5)–(6) describe a spherical wave whose wavelength at $\mu_a = 0.1 \text{ cm}^{-1}$ and $\mu'_s = 8 \text{ cm}^{-1}$ (typical for biological tissues) and at modulation frequency $\nu = 100 \text{ MHz}$, is close to 35 cm, so that the phase of the wave changes approximately by 10° per cm. In the same time, the wave decays exponentially at a distance of about 3 cm, so that PDWs are strongly damped. A similar analytical solution in the form of a spherical wave can be obtained for the case of a semi-infinite homogeneous medium and point source placed on the surface of the medium (see Fig. 1). These analytical solutions are employed to measure μ_a and μ'_s of the homogeneous medium. The method is based on fitting measured ac, dc, and phase of the PDW as functions of source–detector distance by the analytical solution.

Inhomogeneities in the optical properties of a turbid medium cause distortion of the PDW from sphericity (see Fig. 1). PDWs in homogeneous media of limited size or with strong surface curvature also usually have complex structure. For a limited number of configurations one can obtain solutions to the DE using analytical methods of mathematical physics, such as the method of Green's functions. However, in practice the complexity of the inhomogeneities and surfaces usually requires numerical solution of the DE. This can be achieved using numerical methods for the solution of partial differential equations, such as the finite element method and the finite difference method. The finite difference method can also be used to solve the TE. This may be necessary in situations when the DE is not valid, for example to accurately account for the influence of the cerebrospinal fluid (which has relatively low scattering) on light transport in the human skull. However, the solution of the TE requires formidable computational resources, which increase proportionally to the volume of the medium. Therefore, hybrid DE-TE methods were developed, in which DE is used to describe light propagation in most of the medium except small low-scattering areas described by the TE.

One should note that although the term 'photon' suggests quantum properties of light, no consistent application of quantum theory to problems related to PDWI has been developed so far. Also, it should be noted that the notion of a photon as a particle of light having certain energy, momentum, and localization at the same time is not completely consistent with quantum electrodynamics. However, since the theoretical models of light transport in a scattering medium (such as TE and DE) are mathematically equivalent to the models describing systems of randomly moving particles, the notion of a photon as a 'particle of light' may be accepted as adequate and is widely used in the limits of PDWI.

Inverse Problem and Imaging Algorithm

Mathematically the forward problem can be expressed in the form of a general nonlinear operator F :

$$G = F[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}] \quad (7)$$

where G is the set of measured values (ac, dc, and phase), $\mathbf{r}$ is the coordinate vector of a medium element (voxel), and $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}$ denotes the set of tissue optical properties corresponding to whole collection of voxels. The problem of finding the spatial distribution of medium optical properties can be formalized as

$$\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} = F^{-1}[G] \quad (8)$$

Neither the transport model nor the diffusion model allow inversion in a closed analytical form for a general boundary configuration and spatial distribution of optical properties. The basic reason for the complexity of the inverse problem is in the complexity of the light paths in the highly scattering medium. This situation is opposite to that in CT, where the x-ray radiation propagates through the medium along straight lines almost without scattering, which significantly simplifies the forward problem and allows its analytical inversion. From the mathematical point of view, the complexity of the inverse problem of PDWI is due to its nonlinearity and ill-posedness. Only approximate methods of inversion have been developed so far.

The simplest method is called backprojection by analogy with the method of the same name used in CT. The method attempts to reconstruct the actual spatial distribution of the optical properties μ_a and/or μ'_s from the 'measured' bulk values μ_a^m and μ'_s^m obtained at a variety of locations assuming homogeneity of the medium. The measured values are considered as a result of the volumetric 'projection'

$$\{\mu_a^m(\mathbf{r}), \mu'_s^m(\mathbf{r})\} = \int \{\mu_a(\mathbf{r}'), \mu'_s(\mathbf{r}')\} h(\mathbf{r} - \mathbf{r}') d^3\mathbf{r}' \quad (9)$$

of the actual distribution, where $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the measurement and reconstruction, respectively. A weight function $h(\mathbf{r})$ accounts for the contribution of the particular region to a measured value and should be derived from the forward model. Different versions of the backprojection method use different weight functions $h(\mathbf{r})$. In the backprojection method the values of μ_a and/or μ'_s are obtained by inverting the convolution (9). In practice the convolution integral is usually estimated by a sum over a limited number of voxels for a limited number of measurements. Therefore, the inversion of the convolution reduces to the solution of the set of algebraic equations corresponding to (9). The advantage of the method is high speed, which allows real-time imaging of the dynamic processes, for example, functional hemodynamic changes in the brain. However, the backprojection method provides a phenomenological rather than a self-consistent inversion of the forward problem, and its quantitative accuracy is low.

A more consistent method is based on the perturbation approach to the solution of the inverse problem (8). If we have an estimate $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ that is close to the ideal solution, then $G_e = F[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e]$ (see Eq. (7)) is close to S . Then the difference $G - G_e$ can be written as a Taylor series:

$$G - G_e = J[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e] \{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} + \dots \quad (10)$$

where $J[\dots]$ is the Jacobian, and $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$. Neglecting terms after the first one in (10), the problem of finding the correction $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\}$ to the estimate $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ reduces to the calculation of the Jacobian and to the solution of a linear algebraic equation.

This approach provides a basis for the PDWI algorithm, which is the following. First one makes an initial guess at the optical properties of the medium. Then one calculates the ac, dc, and phase of the PDW using a model of light propagation through the tissue and the geometry of the source–detector configuration. The calculated values are compared with the measured ones. If the error is above a set tolerance level, one calculates corrections to the current values of μ_a and μ'_s by solving Eq. (10). The algorithm is completed when the required error level is achieved. Thus, the PDWI algorithm iteratively applies both the forward and inverse problems. Since at each of many iterations one should solve the forward problem, which in the simplest case is a partial differential equation, the algorithm typically requires considerable computational resources.

One should note that the success of the method is largely dependent on the closeness of the initial guess to the actual values of the optical properties. In many studies the 'first guess' of the tissue optical properties was obtained by measuring the bulk absorption and reduced scattering coefficients. However, because of the ill-posedness of the inverse problem, homogeneous initial conditions did not always lead to the correct reconstruction. A more productive approach to the problem of the initial conditions is to assign approximate initial values of the optical properties known for example from *in vitro* measurements to different structures revealed by nonoptical imaging techniques, such as magnetic resonance imaging (MRI) or CT.

Applications

PDWI was proposed and developed as a noninvasive biomedical imaging technique. Although the spatial resolution of PDWI cannot compete with that of CT or MRI, PDWI has the advantages of low cost, high temporal resolution, and high biochemical specificity. The latter is based on the optical spectroscopic approach. Biological tissues are relatively transparent to light in the near-infrared band between 700 and 1000 nm (a near-infrared window). The major biological chromophores in this window are oxy- and deoxyhemoglobin, and water. One can separate contributions of these chromophores into light absorption by employing light sources at different wavelengths. Multi-wavelength PDWI can also be used to measure the redox state of the cytochrome oxidase, which is the terminal electron acceptor of the mitochondrial electron transport chain and responsible for 90% of cellular oxygen consumption.

One of the hot topics in cognitive neuroscience is the hemodynamic mechanism of the oxygen supply to neurons and removal of products of cerebral metabolism. High temporal resolution and biochemical specificity make PDWI a unique noninvasive tool for studies of functional cerebral hemodynamics, although light can penetrate no deeper than the very surface of the adult brain. Fig. 2 displays a typical arrangement of light sources and detectors used in PDWI of functional cerebral hemodynamics, and the images of the activated area in the brain.

Because of its low cost, noninvasiveness, and compact size of the optical sensors, PDWI is also considered to be a promising tool for mammography and structural imaging of anomalies in the brain. PDWI is an especially promising tool for neonatology, because the infant tissues have low absorption.

PDWI instruments employ lasers, light-emitting diodes and gas-discharge tubes as sources of light. Among the detectors there are photomultiplier tubes, avalanche photodiodes, and CCD cameras. Both sources and detectors can be coupled with the tissue by means of optical fibers. A number of commercially produced near-infrared instruments for PDWI are available on the market.

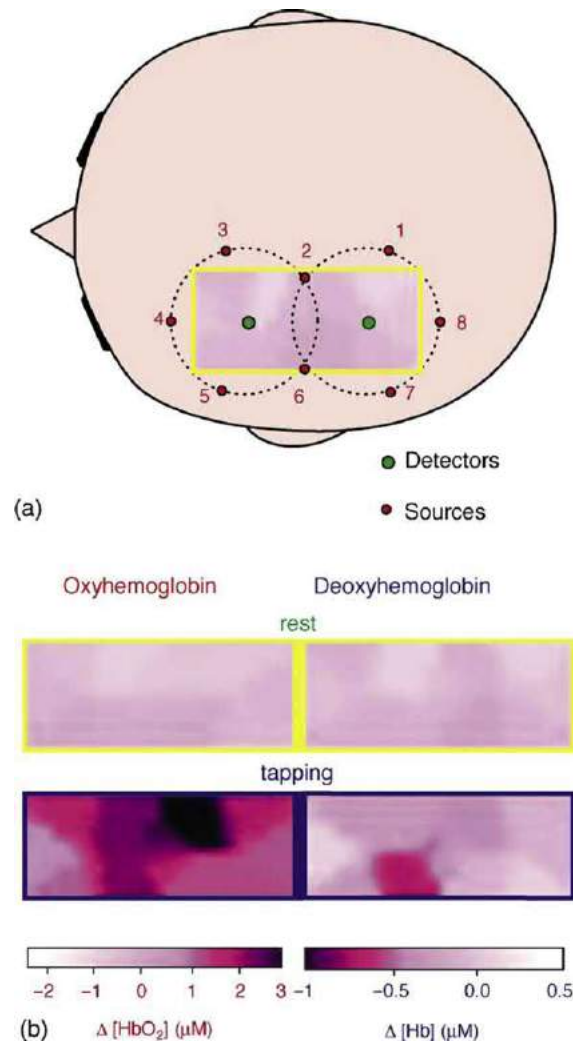


Fig. 2 (a) Geometrical arrangement of the 16 sources and two detectors on the head of a human subject. Each one of the eight small red circles (numbered 1–8) represents one pair of source fibers at 758 and 830 nm. Two larger green circles represent detector fiber bundles. The rectangle inside the head outline is the $4 \times 9 \text{ cm}^2$ imaged area. (b) Maps of oxyhemoglobin ($[\text{HbO}_2]$, left panels) and deoxyhemoglobin ($[\text{Hb}]$, right panels) concentration changes at rest (top) and at the time of maximum response during the third tapping period (bottom). The measurement protocol involved hand-tapping with the right hand (contralateral to the imaged brain area). Adapted with permission from Franceschini MA, Toronov V, Filiaci M, Gratton E, Fantini S (2000) On-line optical imaging of the human brain with 160-ms temporal resolution. *Optics Express* 6(3), 49–57 (Figs. 1 and 4).

According to the type of source modulation, there are three groups of instruments: frequency domain (harmonically modulated sources), time domain (pulsed sources), and steady state. Steady-state instruments are simpler and cheaper than frequency-domain and time-domain ones, but their capabilities are much lower. Time-domain instruments can theoretically provide a maximum of information about the tissue structure, but technically it is difficult to build a time-domain instrument with high temporal resolution and high signal-to-noise ratio. Frequency-domain instruments provide temporal resolution up to several milliseconds and have high quantitative accuracy in the determination of optical properties.

Further Reading

- Arridge, S.R., 1999. Optical tomography in medical imaging. *Inverse Problems* 15, R41–R93.
- Boas, D.A., 1996. Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media: Theory and Biomedical Applications. PhD Thesis, University of Pennsylvania.
- National Research Council (US), Institute of Medicine (US), 1996. Mathematics and Physics of Emerging Biomedical Imaging. Washington, DC: National Academy Press.
- O’Leary, M.A., 1996. Imaging with Diffuse Photon Density Waves. PhD Thesis, University of Pennsylvania.
- Special Issue, 1997. Near-infrared spectroscopy and imaging of living systems. *Philosophical Transactions of the Royal Society B* 352 (1354).
- Tuchin, V., 2002. Handbook of Optical Biomedical Diagnostics. Bellingham, WA: SPIE Press.
- Tuchin, V.V., Luo, Q., Ulyanov, S.S., 2001. Imaging of Tissue Structure and Function. In: Proceedings of SPIE 4427. Bellingham, WA: SPIE Press.

- Villringer, A., Chance, B., 1997. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience* 20, 435–442.
- Villringer, A., Dirnagl, U., 1993. *Optical Imaging of Brain Function and Metabolism*. New York: Plenum Press.
- Villringer, A., Dirnagl, U., 1997. *Optical Imaging of Brain Function and Metabolism 2: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. New York: Plenum Press.

Three-Dimensional Field Transformations

R Piestun, University of Colorado, Boulder, CO, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Conventional optical imaging is understood as an isomorphic mapping from an object space onto an image space. The object space is composed of scattering or emitting objects. The scattering objects can be opaque, transparent, or partially absorbing, while emitting objects can be as diverse as stars in astronomical imaging or fluorescent molecules in microscopic imaging. The image space is an optical field representation of the object.

In most classical applications of optical imaging the goal is to optimize the mapping between one plane in the object space and one plane in the image space. Modern optical systems, though, are aimed at acquiring three-dimensional information regarding the object. The image is thus a three-dimensional field representation of the object.

While many imaging systems are aimed at obtaining visually appealing images, many modern imaging systems register the field information in some type of sensitive device such as film or CCD cameras for later processing. This processing is typically performed electronically by a computer and the final image is ultimately a mathematical digital representation of the object. This type of imaging is called computational imaging. An example of computational imaging is optical tomography, which involves the registration of a set of images that do not necessarily resemble the object, and off-line digital processing to obtain a three-dimensional representation of the object.

In this article we will focus on imaging in the conventional sense but will also focus on the three-dimensional aspects of the image.

Geometrical Optics Transformations

In many optical problems it is appropriate to consider the wavelength of the electromagnetic wave as infinitely small. This approximation is the origin of the field of geometrical optics in which the propagation of light is formulated in terms of geometrical laws. Accordingly, the energy is transported along light rays and their trajectories are determined by certain differential or integral equations that can be directly derived from Maxwell's equations. One of these formulations uses Fermat's principle that states the principle of the shortest optical path. The optical path between two points P_1 and P_2 is defined as:

$$OP = \int_{P_1}^{P_2} n(P) ds \quad (1)$$

where n is the refractive index of the medium. Fermat's principle states that the optical path of a real ray is always a local minimum.

In ideal optical imaging systems, every point P_o of the so-called object space will generate a sharp image point P_i , that is a point where infinite rays emerging from P_o pass through P_i . Perfect imaging occurs when every curve defined by object points is geometrically similar to its image. Perfect imaging can occur between three-dimensional spaces or between surfaces.

An interesting result of geometrical optics is Maxwell's theorem. This theorem states that if an optical instrument produces a sharp image of a three-dimensional domain, the optical length of any curve in the object is equal to the optical length of its image. More explicitly, if the object space has index n_o and the image space index n_i , the theorem states that:

$$\int_{C_o} n_o ds_o = \int_{C_i} n_i ds_i \quad (2)$$

where C_i is the image of the object curve C_o .

An important consequence of Maxwell's theorem is a theorem by Carathéodory which states that any imaging transformation has to be a projective transformation, an inversion, or a combination of them. Another consequence of this theorem is that if the two media are homogeneous, perfect imaging can only occur if the magnification is equal to the ratio of the refractive indices in the object and image space. If the media have the same index, the imaging is trivial with unit magnification. The simplest example of a perfect imaging instrument is a planar mirror. Therefore, if we want to achieve nontrivial imaging between homogeneous spaces with equal index, we must give up either the perfect imaging (similarity) or the sharpness of the image.

The geometrical optics approximation is very useful in the design of optical imaging systems. However, it fails to provide a complete description of wave optical effects, notably diffraction. Therefore, the following sections provide a description of field transformations based on wave optics.

Systems Approach to 3D Imaging

In this article we deal with linear imaging systems. Hence the relation between three-dimensional (3D) input (object) signals and 3D output (image) signals can be completely specified by the 3D impulse response or point spread function (PSF) of the system $h(x, y, z; x', y', z')$. This function specifies the response of the system at image space coordinates (x, y, z) to a δ -function input (point source) at object coordinates (x', y', z') . The optical system object and image signals are related by the superposition integral:

$$i(x, y, z) = \int \int \int_{\Omega} o(x', y', z') h(x, y, z; x', y', z') dx' dy' dz' \quad (3)$$

where Ω is the object domain. It is clear that for a generic linear imaging system the image signal is completely specified by the response to input impulses located at all possible points of the image domain.

If the system is shift invariant, the PSF due to a point source located anywhere in the object domain will be a linearly shifted version of the response due to a reference point source. This approximation is in general not valid and the system's PSF is a function of the position of the point source in the object domain. Shift invariance is traditionally divided into radial and axial shift invariance. Radial shift variance is the change in the image of a point source as a function of its radial distance from the optical axis. Radial shift variance is caused by lens aberrations. Axial shift variance is the change in the image of a point source as a function of its axial position, z' , in the object space. Axial shift variance is inherent to any rotationally symmetric lens system because they can only have one pair of conjugate surfaces, i.e., object-image corresponding surfaces.

In most practical cases the image is detected at a single plane, $z = z_1$ transverse to the optical axis Oz . In those situations, we need to specify the impulse response on this plane:

$$h(x, y, z_1; x', y', z') = h_z(x, y; x', y') \quad (4)$$

The condition for lateral shift invariance is then expressed as $h_z(x, y; x', y') = \tilde{h}_z(x - x', y - y')$. In this case, the superposition integral is a two-dimensional (2D) convolution and it is possible to define the transfer function of the system that is the 2D Fourier transform of the impulse response $\tilde{h}_z(x - x', y - y')$.

Point Spread Function for Coherent, Incoherent, and Partially Coherent Illumination

The input-output relations of an optical imaging system relate different magnitudes according to the type of illumination. If the illumination is coherent, the system is linear in the complex amplitude, while if the illumination is fully incoherent the system is linear in intensity. The 3D impulse response of the incoherent (h_{Inc}) and coherent (h_{Coh}) systems are related as follows:

$$h_{\text{Inc}}(x, y, z; x', y', z') = |h_{\text{Coh}}(x, y, z; x', y', z')|^2 \quad (5)$$

Notwithstanding, many optical systems cannot be considered fully coherent or fully incoherent. In these situations the system is said partially coherent. If the optical path differences within the optical system are much shorter than the coherence length of the illumination, the light effectively has complete temporal coherence and the magnitude of interest is the mutual intensity. The mutual intensity is defined as the statistical correlation of the complex wavefunction $U(\mathbf{r}, t)$ between different points in space:

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1, t) U(\mathbf{r}_2, t) \rangle \quad (6)$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$, $i = 1, 2$. The illumination is then called quasi-monochromatic. For the case of partially coherent quasi-monochromatic illumination, the system is linear in the mutual intensity. The partially coherent impulse response (h_{PC}) is also related to the coherent impulse response as follows:

$$h_{\text{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) = h_{\text{Coh}}^*(\mathbf{r}_1; \mathbf{r}'_1) h_{\text{Coh}}(\mathbf{r}_2; \mathbf{r}'_2) \quad (7)$$

The 3D superposition integral that relate object and image mutual intensities is now:

$$J_{\text{image}}(\mathbf{r}_1, \mathbf{r}_2) = \int \int \int_{\Omega} h_{\text{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) \times J_{\text{object}}(\mathbf{r}'_1, \mathbf{r}'_2) d\mathbf{r}'_1 d\mathbf{r}'_2 \quad (8)$$

Therefore, we conclude that knowledge of the coherent PSF is enough to characterize the optical system for illumination of various degrees of coherence.

Spatial Frequency Content of the 3D Point Spread Function

As previously stated, the coherent impulse response of the optical system completely defines its imaging characteristics. Therefore, it is justified to wonder what the limitations are in the specification of this response. In this section we analyze the frequency domain of coherent and incoherent impulse responses.

The impulse response of the optical system will satisfy the corresponding wave equations regardless of the particulars of the optical system. In particular, assuming a homogeneous isotropic medium in the imaging domain, the coherent response $h(\mathbf{r})$ to an

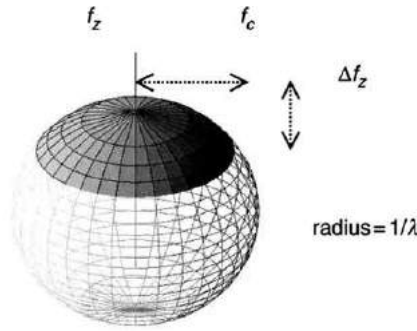


Fig. 1 Three-dimensional spatial-frequency domain of the coherent point spread function.

impulse at $\mathbf{r}'$ will satisfy the homogeneous Helmholtz equation in the image domain:

$$\nabla^2 h(\mathbf{r}) + k^2 h(\mathbf{r}) = 0 \quad (9)$$

where $h(\mathbf{r})$ is the complex amplitude and $k = 2\pi/\lambda$, with λ the wavelength in the medium. If the field at a given origin of coordinates $z = 0, f(x, y)$, is known, $h(\mathbf{r})$ can be calculated using the angular spectrum representation as follows:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(f_x, f_y) \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y \quad (10)$$

where $f_z = \sqrt{\lambda^{-2} - (f_x^2 + f_y^2)}$ and $F(f_x, f_y) = \text{FT}\{f(x, y)\}$; FT represents a 2D Fourier transformation and f_x, f_y are the spatial frequencies along the x and y axes respectively.

Eq. (10) can be written as a 3D integral as follows:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(f_x, f_y) \cdot \delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y df_z \quad (11)$$

where δ represents the dirac impulse. Next we represent the 3D impulse response $h(x, y, z)$ as a 3D Fourier decomposition:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(f_x, f_y, f_z) \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y df_z \quad (12)$$

where $H(f_x, f_y, f_z) = \text{FT}^{3D}\{h(x, y, z)\}$. Comparing Eqs. (11) and (12) we can infer that:

$$H(f_x, f_y, f_z) = F(f_x, f_y) \cdot \delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \quad (13)$$

This relation implies that the domain of the Fourier transform of the impulse response is confined to a spherical shell of radius $1/\lambda$. Alternatively, $H(f_x, f_y, f_z)$ is zero everywhere, except from the surface of the sphere.

The previous derivation did not consider any limitation in the range of spatial frequencies allowed through the system. We now assume a finite aperture that will introduce an effective radial cutoff frequency $f_c < 1/\lambda$ and will further limit the spatial spectral domain of $h(x, y, z)$. This is depicted in Fig. 1. Note that the transverse and longitudinal spatial frequencies are related, hence the longitudinal spatial frequencies are located within a bandpass of width $\Delta f_z = 1 - \sqrt{\lambda^{-2} - f_c^2}$.

Let us now consider the frequency content of possible incoherent PSFs. From Eq. (5) we derive the 3D Fourier transform of the incoherent impulse response:

$$\begin{aligned} H_{\text{Inc}}(f_x, f_y, f_z) &= \text{FT}\{h_{\text{Inc}}(x, y, z; x' y', z')\} \\ &= \text{FT}\{|h_{\text{Coh}}(x, y, z; x' y', z')|^2\} \\ &= H \otimes H(f_x, f_y, f_z) \end{aligned} \quad (14)$$

where $\otimes$ represents a 3D correlation. The domain of H_{Inc} now becomes a volume of revolution as shown in Fig. 2. Moreover, because h_{Inc} is real, H_{Inc} is conjugate symmetric, i.e., $H_{\text{Inc}} = H_{\text{Inc}}^*$. Now, the transverse cutoff frequency is $f_c^{\text{Inc}} = 2f_c$ while the longitudinal cutoff frequency is $f_{\text{cl}}^{\text{Inc}} = 2\Delta f_z$.

Diffraction Calculation of the 3D Point Spread Function

The 3D-image field created by a point source in the object space can be predicted by proper modeling of the optical system. In the absence of aberrations, the distortion of the PSF is originated in the diffraction created by the limiting aperture of the optical system. Diffraction is thus the ultimate limitation to the resolution of the optical system. Under the design conditions of the

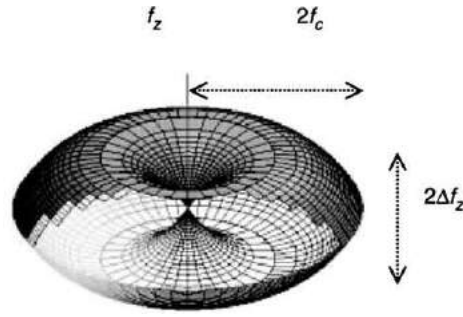


Fig. 2 Three-dimensional spatial-frequency domain of the incoherent point spread function.

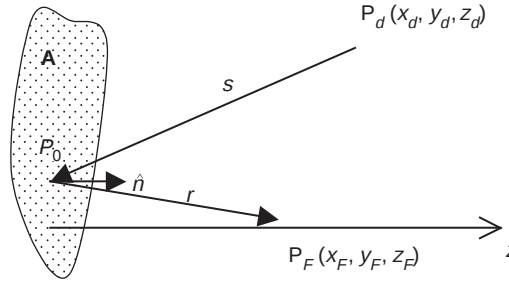


Fig. 3 Diffraction of a spherical wave by a planar aperture.

system, i.e., assuming all aberrations are corrected, the image of a point source is the diffraction pattern produced when a spherical wave converging to the (geometrical) image point is diffracted by the limiting aperture of the system. This is called a diffraction limited image.

There are many formulations to describe the 3D diffraction pattern obtained with a converging wave through an aperture. Here we review the Kirchhoff formulation. Let us consider the aperture A and the field at a point P_d originated from a converging spherical wave focused at point P_F (see **Fig. 3**). We define the vectors $\mathbf{r}$ joining a generic point on the aperture P_0 and P_F , and $\mathbf{s}$ joining P_0 and P_d . The Kirchhoff formulation states that the complex amplitude $U(P_d/P_F)$ is the integral over the aperture of the complex amplitude at each point propagated to P_d by a spherical wave $\exp(iks/s)$ modified by the directional factor $\cos(\hat{\mathbf{n}}, \mathbf{s}) - \cos(\hat{\mathbf{n}}, \mathbf{r})$. $\hat{\mathbf{n}}$ is the unit vector normal to the aperture:

$$U(P_d/P_F) = \frac{-iC}{2\lambda} \int_A \frac{\exp[-ik(r-s)]}{rs} [\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})] dS \quad (15)$$

where $r = \|\mathbf{r}\|$, $s = \|\mathbf{s}\|$, and C is the amplitude of the spherical wave.

The directional factor varies slowly and in many situations can be approximated by a constant equal to 2 for small image-field angles. The amplitude over the aperture is also slowly varying and can be approximated by a constant, i.e., $C/r \approx C_0$. Different formulations of the Kirchhoff integral differ in the estimates of the factor $1/s$ and the phase $k(r-s)$. For example, one such formulation leads to the following expression for the 3D diffraction pattern for a focal point on the axis of an optical system with circular aperture:

$$U(P_d/P_F) = \frac{-ika^2 C_0}{z_d} \exp\left[ik\left(z_d - z_F + \frac{r_d^2}{2z_d}\right)\right] \times \int_0^1 J_0\left(\frac{k\rho r_d}{z_d}\right) \exp\left[\frac{-ika^2 \rho^2 (z_d - z_F)}{2z_d z_F}\right] \rho d\rho \quad (16)$$

where a is the radius of the aperture, r_d is the radial coordinate of P_d , and ρ is the radial coordinate of P_0 .

To illustrate the use of **Eq. (16)**, we calculated the intensity around the focus of a spherical wave diffracted through a circular aperture of radius $a=0.5$ cm, focusing at distance $z_F=4$ cm from the screen. We assumed the wavelength to be $\lambda=0.5$ μ . **Fig. 4** depicts the magnitude of the complex amplitude in a meridional plane near the focus of the converging spherical wave.

Models that take into account off-axis focal points have also been developed. For the most accurate calculations, it is necessary to consider the full electromagnetic nature of the optical field.

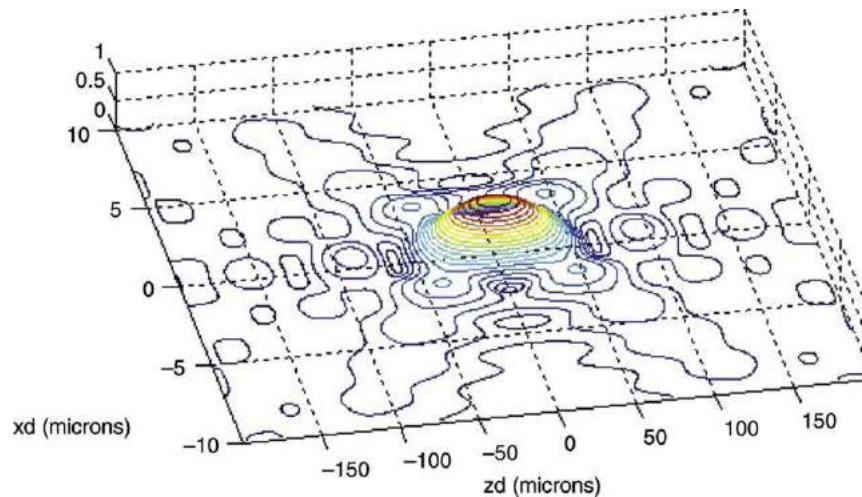


Fig. 4 Lines of equal intensity in a meridional plane near the focus of a converging spherical wave diffracted at a circular aperture ($\lambda = 0.5 \mu$, $a = 0.5$ cm, $z_F = 4$ cm).

Conclusions

Classical imaging theory can be applied to the 3D imaging transformation from object to image. The systems approach to the imaging problem implies that knowledge of the 3D coherent PSF suffices to predict the 3D response of the system under illumination of varying degrees of coherence. The PSF response can be predicted using well-known formulations of diffraction.

Further Reading

- Born, M., Wolf, E., 1999. *Principles of Optics*, 7th edn. New York: Cambridge University Press.
- Gibson, S.F., Lanni, F., 1989. Diffraction by a circular aperture as a model for three-dimensional optical microscopy. *Journal of the Optical Society of America A* 6, 1357–1367.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*. New York: McGraw-Hill.
- Li, Y., Wolf, E., 1984. Three-dimensional intensity distributions near the focus in systems of different Fresnel numbers. *Journal of the Optical Society of America A* 1, 801–808.
- Piestun, R., Shamir, J., 2002. Synthesis of 3D light fields and applications. *Proceedings of the IEEE* 90, 222–244.
- Saleh, B.E.A., Teich, M.C., 1991. *Fundamentals of Photonics*. New York: Wiley.
- Stamnes, J.J., 1986. *Waves in Focal Regions*. Bristol: Hilger.

Volume Holographic Imaging

G Barbastathis, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Published by Elsevier Ltd.

Volume holographic imaging (VHI) refers to a class of imaging techniques which use a volume hologram in at least one location within the optical path. The volume hologram acts as a 'smart lens,' which processes the optical field to extract spatial information in three dimensions (lateral as well as longitudinal) and spectral information.

Fig. 1 is a generic VHI system. The object is either illuminated by a light source (e.g., sunlight or a pump laser) as shown in the figure, or it may be self-luminous. Light scattered or emitted by the object is first transformed by an objective lens and then illuminates the volume hologram. The role of the objective is to form an intermediate image which serves as input to the volume hologram. The volume hologram itself is modeled as a three-dimensional (3D) modulation $\Delta\epsilon(\mathbf{r})$ of the dielectric index within a finite region of space. The light entering the hologram is diffracted by $\Delta\epsilon(\mathbf{r})$ with efficiency η , defined as

$$\eta = \frac{\text{Power diffracted by the volume hologram}}{\text{Power incident to the volume hologram}} \quad (1)$$

We assume that diffraction occurs in the Bragg regime. The diffracted field is Fourier-transformed by the collector lens, and the result is sampled and measured by an intensity detector array (such as a CCD or CMOS camera).

Intuitively, we expect that a fraction of the illumination incident upon the volume hologram is Bragg-matched and is diffracted towards the Fourier-transforming lens. The remainder of the incident illumination is Bragg-mismatched, and as result is transmitted through the volume hologram un-diffracted. Therefore, the volume hologram acts as a filter which admits the Bragg-matched portion of the object and rejects the rest. When appropriately designed, this 'Bragg imaging filter' can exhibit very rich behavior, spanning the three spatial dimensions and the spectral dimension of the object.

To keep the discussion simple, we consider the specific case of a transmission geometry volume hologram, described in **Fig. 2**. The volume hologram is created by interfering a spherical wave and a planewave, as shown in **Fig. 2(a)**. The spherical wave originates at the coordinate origin. The planewave is off-axis and its wavevector lies on the xz plane. As in most common holographic systems, the two beams are assumed to be at the same wavelength λ , and mutually coherent. The volume hologram results from exposure of a photosensitive material to the interference of these two beams.

First, assume that the object is a simple point source. The intermediate image is also approximately a point source, that we refer to as 'probe,' located somewhere in the vicinity of the reference point source. Assuming the wavelength of the probe source is the same as that of the reference and signal beams, volume diffraction theory shows that:

- (i) If the probe point source is displaced in the y direction relative to the reference point source, the image formed by the volume hologram is also displaced by a proportional amount. Most common imaging systems would be expected to operate this way.
- (ii) If the probe point source is displaced in the x direction relative to the reference point source, the image disappears (i.e., the detector plane remains dark).
- (iii) If the probe point source is displaced in the z direction relative to the reference point source, a defocused and faint image is formed on the detector plane. 'Faint' here means that the fraction of energy of the defocused probe transmitted to the detector plane is much smaller than the fraction that would have been transmitted if the probe had been at the origin.

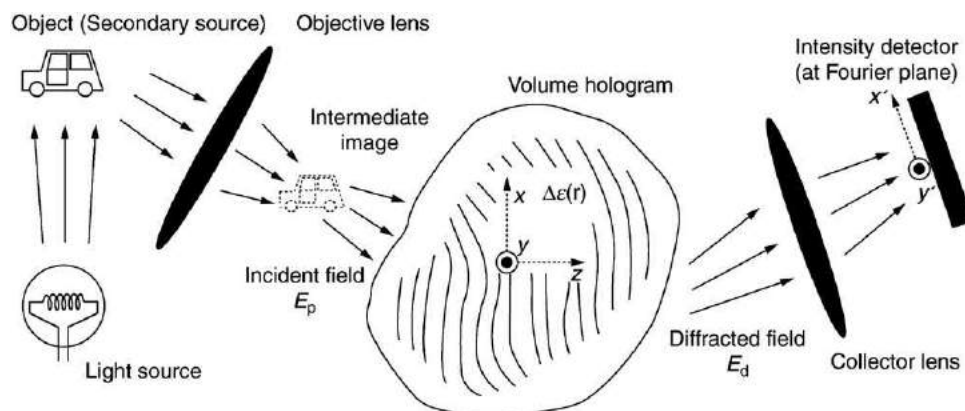


Fig. 1 Volume holographic imaging (VHI) system.

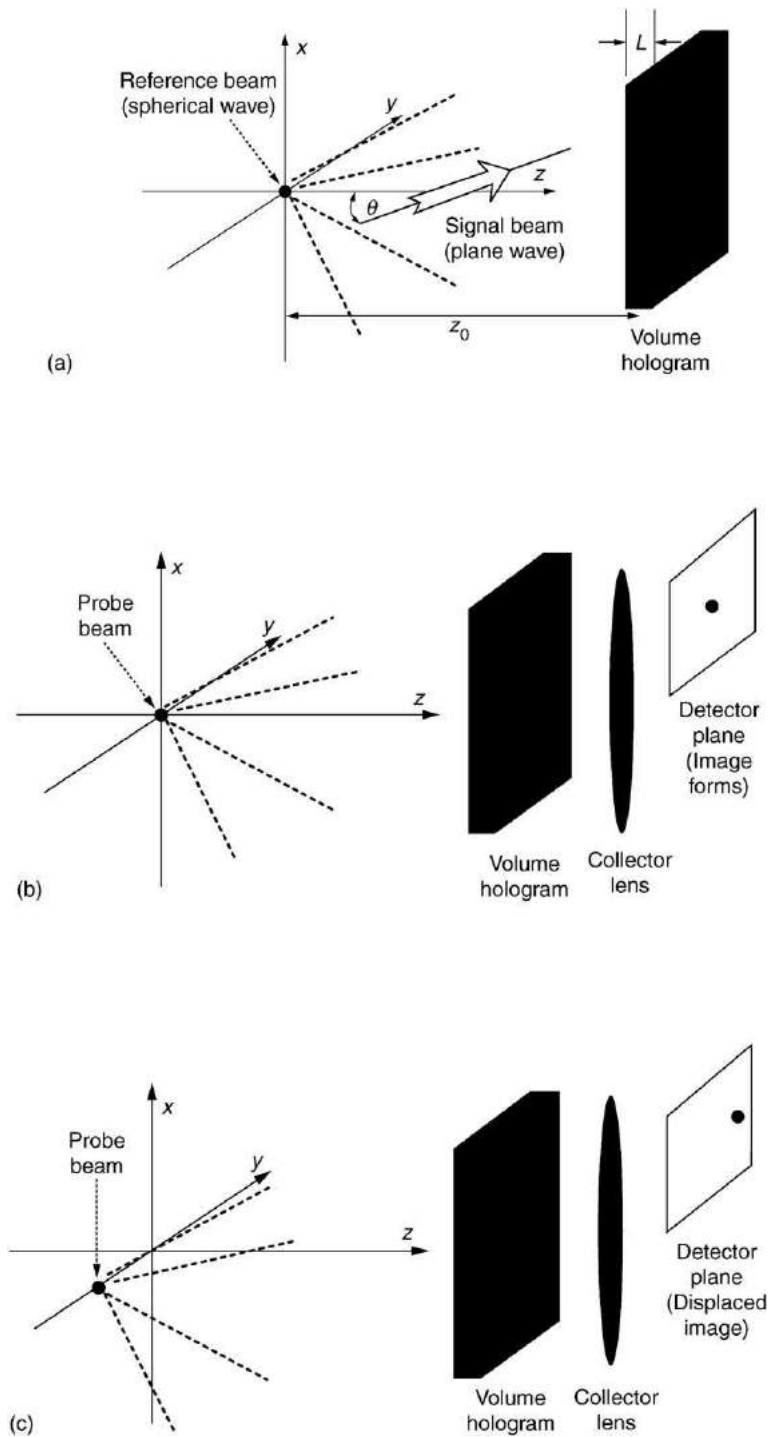


Fig. 2 (a) Recoding of a transmission-geometry volume hologram with a spherical wave and an off-axis plane wave with its wave-vector on the xz plane. (b) Imaging of a probe point source that replicates the location and wavelength of the reference point source using the volume hologram recorded in part (a). (c) Imaging of a probe point source at the same wavelength but displaced in the y direction relative to the reference point source. (d) Imaging of a probe point source at the same wavelength but displaced in the x direction relative to the reference point source. (e) Imaging of a probe point source at the same wavelength but displaced in the z direction relative to the reference point source. (f) 'Bragg slitting:' Imaging of an extended monochromatic, spatially incoherent object using a volume hologram recorded as in (a). (g) Joining Bragg slits from different colors to form rainbow slices: Imaging of an extended polychromatic object using a volume hologram recorded as in (a). (h) Multiplex imaging of several slices using a volume hologram formed by multiple exposures.

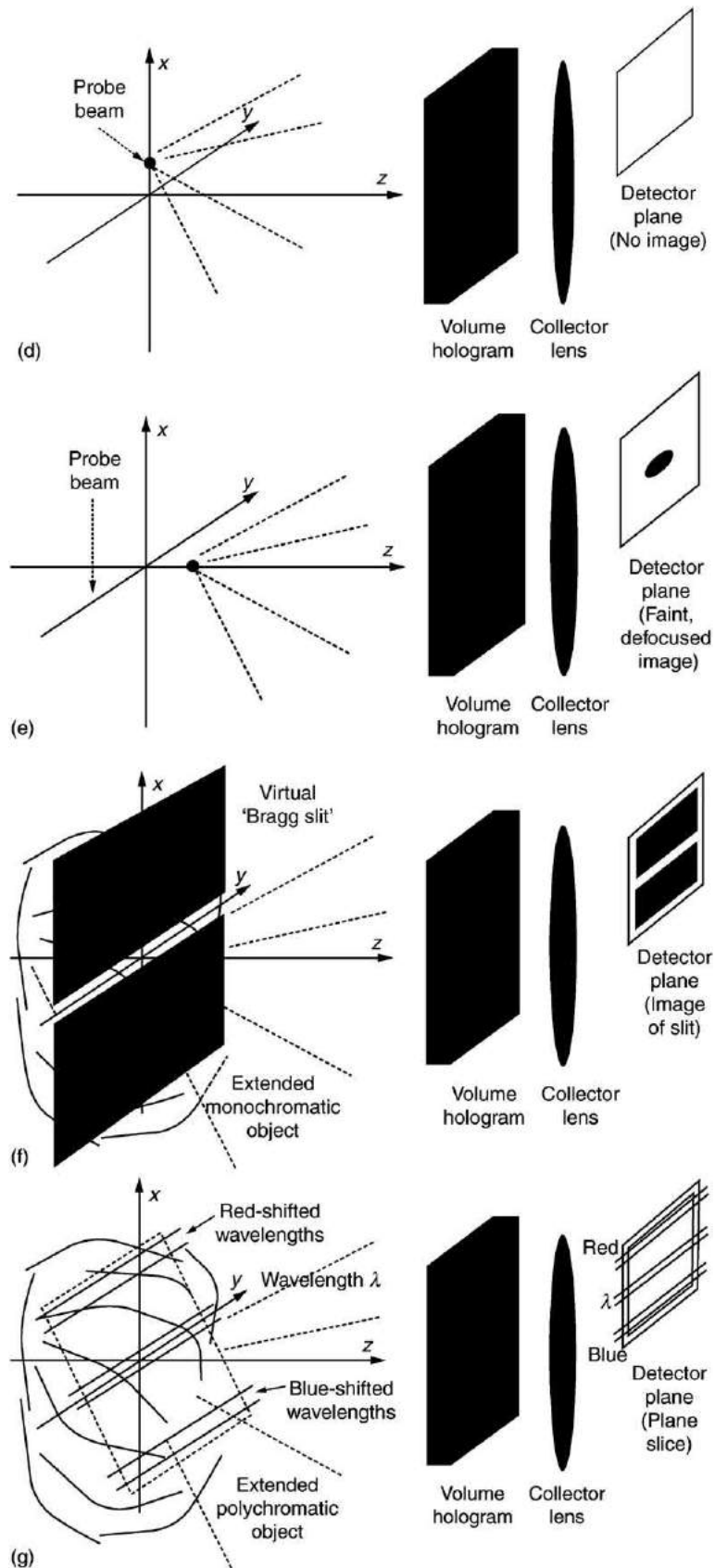


Fig. 2 (Continued).

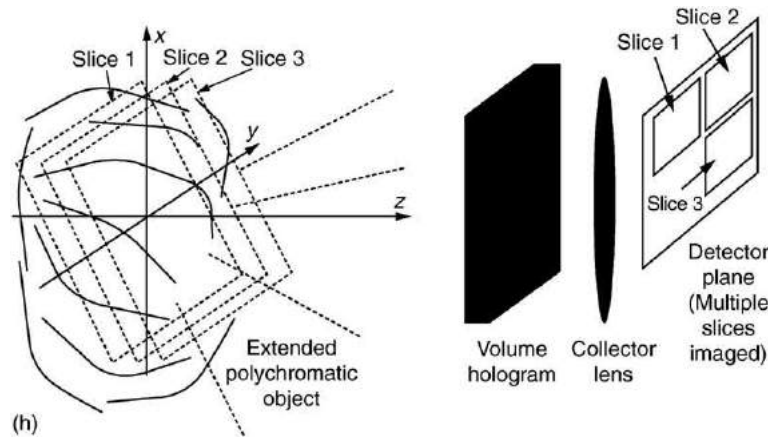


Fig. 2 (Continued).

Now consider an extended, monochromatic, spatially incoherent object and intermediate image, as in Fig. 2(f). According to the above description, the volume hologram acts as a 'Bragg slit' in this case: because of Bragg selectivity, the volume hologram transmits light originating from the vicinity of the y axis, and rejects light originating anywhere else. For the same reason, the volume hologram affords depth selectivity (like the pinhole of a confocal microscope). The width of the slit is determined by the recording geometry, and the thickness of the volume hologram. For example, in the transmission recording geometry of Fig. 2(a), where the reference beam originates a distance z_0 away from the hologram, the planewave propagates at angle θ with respect to the optical axis z (assuming $\theta \ll 1$ radian), and the hologram thickness is L (assuming $L \ll z_0$), the width of the slit is found to be

$$\Delta x \approx \frac{\lambda z_0}{L\theta} \quad (2)$$

The imaging function becomes richer if the object is polychromatic. In addition to its Bragg slit function, the volume hologram exhibits then dispersive behavior, like all diffractive elements. In this particular case, dispersion causes the hologram to image simultaneously multiple Bragg slits, each at a different color and parallel to the original slit at wavelength λ , but displaced along the z axis. Light from all these slits finds itself in focus at the detector plane, thus forming a 'rainbow image' of an entire slice through the object, as shown in Fig. 2(g).

To further exploit the capabilities of volume holograms, we recall that in general it is possible to 'multiplex' (superimpose) several volume gratings within the same volume by successive exposures. In the imaging context, suppose that we multiplex several gratings similar to the grating described in Fig. 2(a) but with spherical reference waves originating at different locations and plane signal waves at different orientations. When the multiplexed volume hologram is illuminated by an extended polychromatic source, each grating forms a separate image of a rainbow slice, as described earlier. By spacing appropriately the angles of propagation of the plane signal waves, we can ensure that the rainbow images are formed on nonoverlapping areas on the detector plane, as shown in Fig. 2(h). This device is now performing true four-dimensional (4D) imaging: it is separating the spatial and spectral components of the object illumination so that they can be measured independently by the detector array. Assuming the photon count is sufficiently high and the number of detector pixels is sufficient, this 'spatio-spectral slicing' operation can be performed in real time, without need for mechanical scanning.

The complete theory of VHI in the spectral and spatial domains is given in the Further Reading section below. Experimental demonstrations of VHI have been performed in the context of a confocal microscope where the volume hologram performs the function of a 'Bragg pinhole' as well as a real-time 4D imager. The limited diffraction efficiency η of volume holograms poses a major concern for VHI systems. Holograms, however, are known to act as matched filters. It has been shown that the matched filtering nature of volume holograms as imaging elements is superior to other filtering elements (e.g., pinholes or slits) in an information-theoretic sense. Efforts are currently underway to strengthen this property by design of volume holographic imaging elements which are more elaborate than described here.

See also: Overview: Holography

Further Reading

- Barbastathis, G., Brady, D.J., 1999. Multidimensional tomographic imaging using volume holography. *Proceedings of the IEEE* 87 (12), 2098–2120.
 Barbastathis, G., Sinha, A., 2001. Information content of volume holographic images. *Trends in Biotechnology* 19 (10), 383–392.
 Barbastathis, G., Balberg, M., Brady, D.J., 1999. Confocal microscopy with a volume holographic filter. *Optics Letters* 24 (12), 811–813.

- Barbastathis, G., Sinha, A., Neifeld, M.A., 2001. Information-theoretic treatment of axial imaging. In: OSA Topical Meeting on Integrated Computational Imaging Systems, pp. 18–20. Albuquerque, NM.
- Liu, W., Psaltis, D., Barbastathis, G., 2002. Real time spectral imaging in three spatial dimensions. *Optics Letters* 27 (10), 854–856.
- van-der Lugt, A., 1964. Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10 (2), 139–145.

Wavefront Sensors and Control (Imaging Through Turbulence)

CL Matson, Air Force Research Laboratory, Kirtland, NM, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

In many situations, distant objects are imaged using optical or near-infrared imaging systems. Examples include terrestrial surveillance from space, tactical surveillance from airborne imaging systems, and ground-based astronomical imaging. Ideally, the resolution in these images is limited only by the wavelength of light λ used to create the images divided by the imaging system's effective diameter D . The ratio λ/D is called the diffraction limit of the imaging system. Unfortunately, when the imaging system is located in the Earth's atmosphere, it is often atmospheric turbulence fluctuations that limit the level of detail in the long-exposure images, not the diffraction limit of the telescope. For example, the Keck telescopes, located on top of Mauna Kea in Hawaii, are 10 m diameter telescopes that have a diffraction-limited angular resolution of $0.2 \mu\text{rad}$ for an imaging wavelength of $2 \mu\text{m}$. However, atmospheric turbulence limits the achievable resolution in long-exposure images to the order of $1\text{--}5 \mu\text{rad}$ at this wavelength.

The important discovery by Anton Labeyrie in 1970 that short-exposure images of objects retain meaningful diffraction-limited information about the object, where 'short' means that the exposure time is less than the atmospheric turbulence coherence time, has revolutionized imaging through turbulence. Atmospheric coherence times depend upon a number of factors, including wind speeds, wind altitudes, and observation wavelength. At astronomical observatories for visible observation wavelengths, a rule of thumb is that atmospheric turbulence stays essentially constant for exposure times of one to ten milliseconds under average clear-sky weather conditions. A variety of techniques that exploit this discovery have been developed. These techniques can be grouped into three different categories: techniques that remove blurring from the detected intensity images (post-detection compensation), techniques that modify the optical field prior to detecting the light (pre-detection compensation), and hybrid techniques that use both post-detection and pre-detection compensation. In this article, a variety of these techniques are presented that are representative of the myriad of methods that have been developed during the past decades. To simplify the discussion, it is assumed that the blurring due to atmospheric turbulence is the same at every point in the image (isoplanaticity), that the images are formed using incoherent light as emitted by the sun and stars, and that a single filled aperture is used to create the images. Interferometric techniques that use multiple apertures to estimate the mutual coherence function of the optical field and calculate an image from it.

The emphasis in this article is on identifying and removing atmospheric blurring in images. As part of this process, regularization techniques must be included in the image reconstruction algorithms in order to obtain the best possible resolution for an acceptable level of noise. Although regularization techniques are not discussed in this article, information can be found in the Further Reading section at the end of this article.

Imaging and Atmospheric Turbulence

In Fig. 1, a conceptual diagram illustrates the impact that atmospheric turbulence has upon resolution in classical imaging systems. A distant star is being imaged by a telescope/detector system. In the absence of atmospheric turbulence, the field emitted by the star has a planar wavefront over the extent of the lens since the star is a long distance from the imaging system and thus unresolved by the telescope. When imaged by the telescope, the diffraction pattern of the telescope is obtained, resulting in a diffraction-limited image (Fig. 2). However, in the presence of atmospheric turbulence near the telescope, the planar wavefront is randomly distorted by the turbulence, resulting in a corrugated wavefront with a spatial correlation scale given by the Fried parameter r_o . For visible wavelengths, at good astronomical observing sites, r_o ranges from 5 cm to 20 cm. When this corrugated wavefront is converted to an image by the telescope, a distorted image is obtained (Fig. 3). Because atmospheric turbulence is time-varying, the distortions in the image also change over time. If the imaging shutter on the telescope remains open for many occurrences of atmospheric turbulence, the resulting image will be an average of all the individually distorted images, resulting in what is known as the seeing disk (Fig. 4). It is the seeing disk that typically limits spatial resolution in long-exposure images, not the diffraction pattern of the telescope. As a result, the effective spatial resolution in long-exposure images in the presence of atmospheric turbulence is given by λ/r_o , not λ/D .

Atmospheric turbulence distorts a field's wavefront by delaying it in a way that is spatially and temporally random. The origin of these delays is the random variation in the index of refraction of the atmosphere as a result of turbulent air motion brought about by temperature variations. A Fourier optics imaging model is useful in characterizing the effects of atmospheric turbulence on spatial resolution in images. Let $i(\vec{x})$ be an image of an object $o(\vec{x})$ where $\vec{x}$ is a two-dimensional spatial vector, and let $I(\vec{f})$ and $O(\vec{f})$ be their spatial Fourier transforms, where $\vec{f}$ is a two-dimensional spatial frequency vector. In the absence of atmospheric turbulence, the diffraction-limited Fourier transform $I(\vec{f})$ is given by

$$I_d(\vec{f}) = O(\vec{f})H_d(\vec{f}) \quad (1)$$

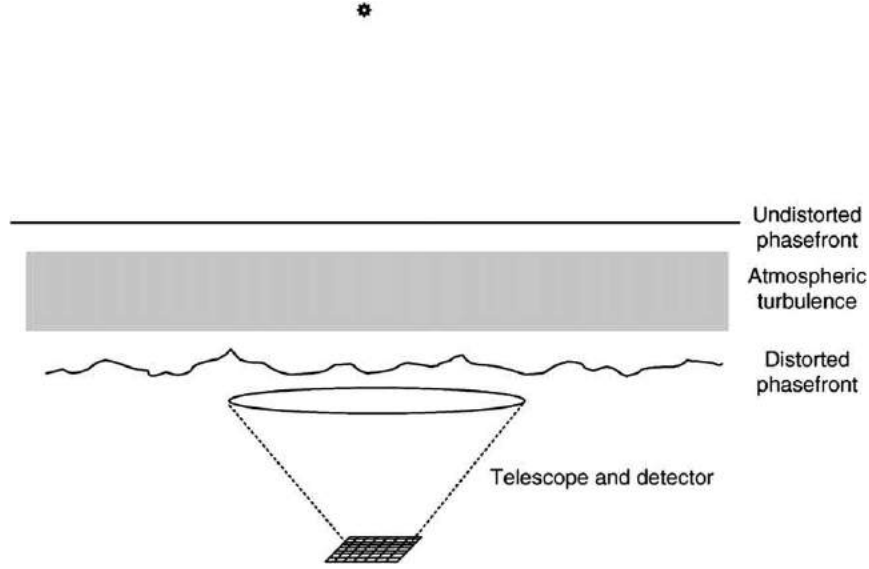


Fig. 1 Conceptual diagram of imaging through atmospheric turbulence.

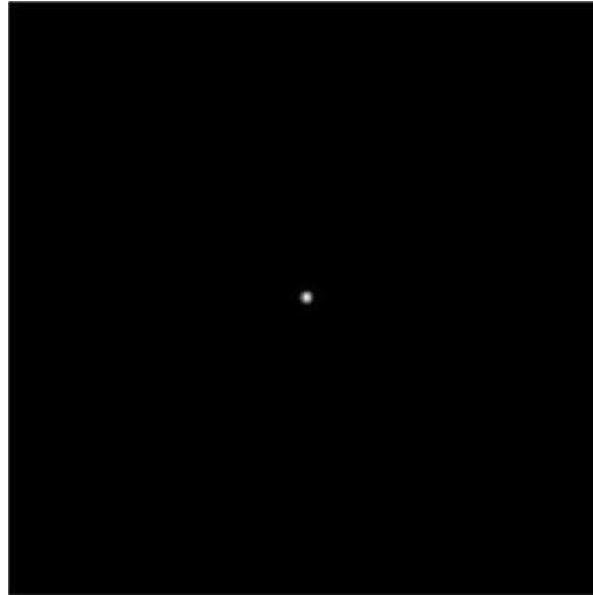


Fig. 2 Computer-simulated diffraction-limited image of a point source.

where the subscript 'd' denotes the diffraction-limited case and where $H_d(\vec{f})$ is the diffraction-limited telescope transfer function. In the presence of atmospheric turbulence, the Fourier transform of a long exposure image $i_{le}(\vec{x})$ is given by

$$I_{le}(\vec{f}) = O(\vec{f})H_d(\vec{f})H_{le}(\vec{f}) = I_d(\vec{f})H_{le}(\vec{f}) \quad (2)$$

where the subscript 'le' denotes long exposure. If the turbulence satisfies Kolmogorov statistics, as is typically assumed, then:

$$H_{le}(\vec{f}) = \exp \left[-3.44 \left(\frac{\lambda d |\vec{f}|}{r_o} \right)^{5/3} \right] \quad (3)$$

where d is the distance between the imaging system's exit pupil and detector plane. As can be seen from Eq. (2), the effect of atmospheric turbulence on a long-exposure image is a multiplication of the diffraction-limited image's Fourier transform by a transfer function due to atmospheric turbulence. In Fig. 5, an amplitude plot of a diffraction-limited transfer function is displayed along with an amplitude plot of a long-exposure transfer function for a 2.3 m diameter telescope, an imaging wavelength of

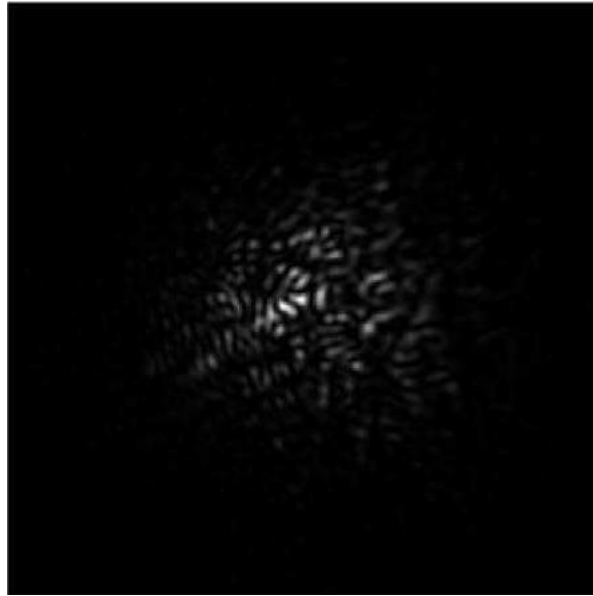


Fig. 3 Computer-simulated short-exposure image of a point source through atmospheric turbulence.

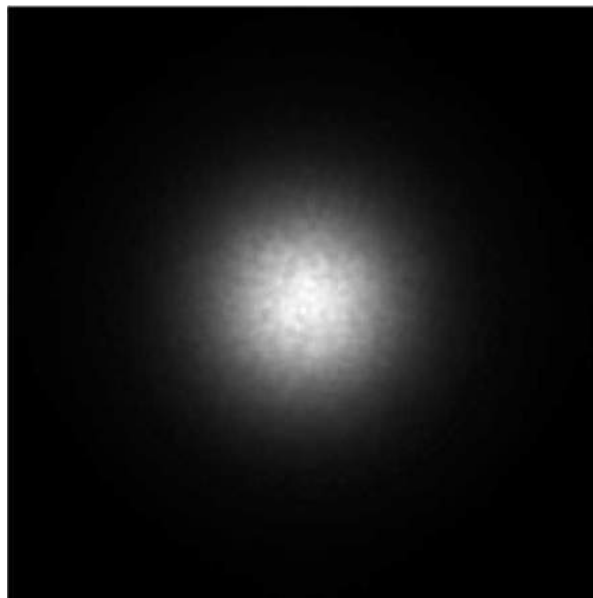


Fig. 4 Computer-simulated long-exposure image of a point source through atmospheric turbulence.

0.5 μm , and an r_0 value of 10 cm. Notice that the long exposure transfer function significantly reduces the amount of high spatial frequency information in an image. Diffraction-limited and long-exposure turbulence-limited images of a binary star are shown in [Figs. 6](#) and [7](#), illustrating the significant loss of spatial resolution brought on by atmospheric turbulence.

Post-Detection Compensation

Speckle Imaging

Labeyrie's 1970 discovery was motivated by his analysis of short-exposure star images. In these images (see [Fig. 3](#)), a number of speckles are present, where the individual speckle sizes are roughly the size of the diffraction-limited spot of the telescope. From this, he realized that high spatial frequency information is retained in short-exposure images, albeit distorted. He then showed that

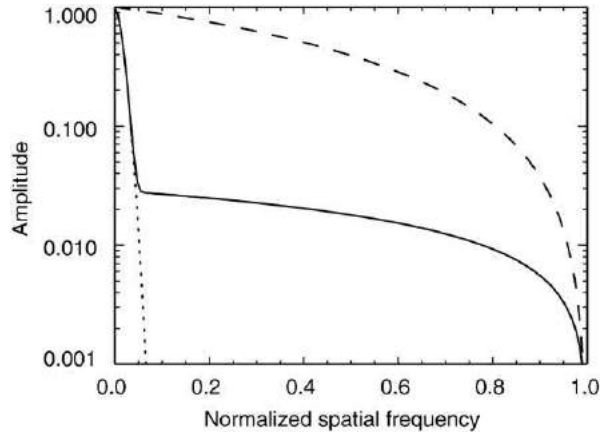


Fig. 5 Plots of the Fourier amplitudes of the diffraction-limited transfer function (dashed line), Fourier amplitudes of the long-exposure transfer function (dotted line), and the square root of the average short exposure energy spectrum. The spatial frequency axis is normalized to one at the diffraction limit of the telescope.

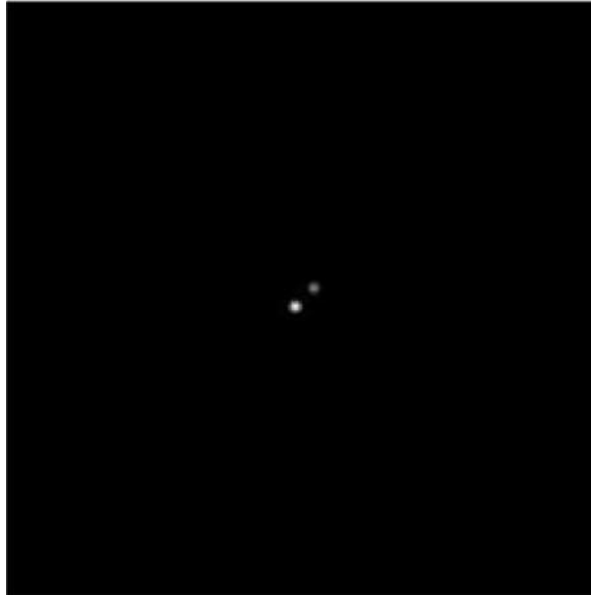


Fig. 6 Computer-simulated diffraction-limited image of a binary star.

the energy spectra $E_1(\vec{f})$ of the individual short-exposure images can be averaged to obtain an estimate of the average energy spectrum that retains high signal-to-noise ratio (SNR) information at high spatial frequencies. The square root of the average energy spectrum (the amplitude spectrum) of a star, calculated using Labeyrie's method (also known as speckle interferometry), is shown in **Fig. 5**, along with the diffraction-limited and long-exposure amplitude spectra. Notice that meaningful high spatial frequency information is retained in the average amplitude spectrum out to the diffraction limit of the telescope. Because the atmospherically corrupted amplitude spectrum estimate is significantly lower than the diffraction-limited amplitude, a number of short exposure images must be used to build up the SNRs at the higher spatial frequencies.

To reconstruct an image, both the Fourier amplitudes and Fourier phases of the image must be known. In fact, it is well known that the Fourier phase of an image carries most of the information in the image. Because only the Fourier amplitudes are available using Labeyrie's method, other techniques were subsequently developed to calculate the Fourier phase spectrum from a sequence of short-exposure images. Techniques that combine amplitude spectrum estimation using Labeyrie's method along with phase spectrum estimation are called speckle imaging techniques. The two most widely used phase spectrum estimation methods are the Knox-Thompson (or cross spectrum) and bispectrum algorithms. Both of these algorithms calculate Fourier-domain quantities that retain high spatial frequency information about the Fourier phase spectrum when averaged over multiple short-exposure images. The cross spectrum and bispectrum techniques calculate the average cross spectrum $C(\vec{f}, \Delta\vec{f})$ and bispectrum $B(\vec{f}_1, \vec{f}_2)$

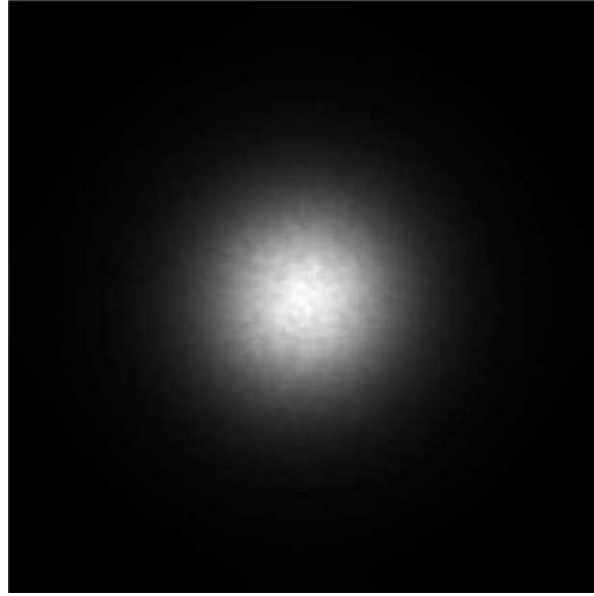


Fig. 7 Computer-simulated long-exposure image of a binary star.

respectively, that are given by

$$C(\vec{f}, \Delta\vec{f}) = I(\vec{f})I^*(\vec{f} + \Delta\vec{f}) \quad (4)$$

$$B(\vec{f}_1, \vec{f}_2) = I(\vec{f}_1)I(\vec{f}_2)I^*(\vec{f}_1 + \vec{f}_2) \quad (5)$$

where the superscript * denotes complex conjugate. Although the cross spectrum is defined for all values of $\vec{f}$ and $\Delta\vec{f}$ only for values of $\Delta\vec{f}$ or $\vec{f}$ less than r_o/λ are the cross spectrum elements obtained with sufficiently high SNRs to be useful in the reconstruction process. The equivalent constraint on the bispectrum elements is that either $\vec{f}_1, \vec{f}_2$ or $\vec{f}_1 + \vec{f}_2$ must be less than r_o/λ . Once the average cross spectrum or bispectrum is calculated from a series of short-exposure images, the phase spectrum $\phi_1(\vec{f})$ of the underlying image can be calculated from their phases using one of a number of algorithms based upon the following equations:

$$\phi_1(\vec{f} + \Delta\vec{f}) = \phi_1(\vec{f}) - \phi_c(\vec{f} + \Delta\vec{f}) \quad (6)$$

for the cross spectrum and:

$$\phi_1(\vec{f}_1 + \vec{f}_2) = \phi_1(\vec{f}_1) + \phi_1(\vec{f}_2) - \phi_b(\vec{f}_1 + \vec{f}_2) \quad (7)$$

for the bispectrum. Because the phases are only known for units of 2π , phasor versions of Eqs. (6) and (7) must be used. For the cross spectrum technique, each short-exposure image must be centroided prior to calculating its cross spectrum and adding it to the total cross spectrum. The bispectrum technique is insensitive to tilt, so no centroiding is necessary; however, the bispectrum technique requires setting two image phase spectrum values to arbitrary values to initialize the phase reconstruction process. Typically, the image phase spectrum values at $\vec{f} = (0, 1)$ and $\vec{f} = (1, 0)$ are set equal to zero, which centroids the reconstructed image but does not affect its morphology.

To obtain an estimated Fourier transform $I_e(\vec{f})$ of the image from the energy and phase spectrum estimates, the square root of the average energy spectrum is combined with the average phase spectrum calculated from either the bispectrum or cross spectrum. The result is given by

$$I_e(\vec{f}) = [E_1(\vec{f})]^{1/2} \exp[j\phi_1(\vec{f})] \quad (8)$$

where $j = \sqrt{-1}$. The result in Eq. (8) includes atmospheric attenuation of the Fourier amplitudes, so it must be divided by the amplitude spectrum of the atmosphere to obtain the desired object spectrum estimate. The atmospheric amplitude spectrum estimate is obtained by using Labeyrie's technique on a sequence of short-exposure images of a nearby unresolved star. No phase spectrum correction is necessary because it has been shown theoretically and verified experimentally that atmospheric turbulence has a negligible effect upon $\phi_1(\vec{f})$.

Blind Deconvolution

Speckle imaging techniques are effective as well as easy and fast to implement. However, it is necessary to obtain a separate estimate of the atmospheric energy spectrum using an unresolved star in order to reconstruct a high-resolution image. Usually, the star is not co-located with the object and the star images are not collected simultaneously with the images of the object. As a result,

the atmospheric energy spectrum estimate may be inaccurate. At the very least, collecting the additional star data expends valuable observation time. In addition, there is often meaningful prior knowledge about the object and the data, such as the fact that measured intensities are positive (positivity), that many astronomical objects are imaged against a black background (support), and that the measured data are band-limited, among others. Speckle imaging algorithms do not exploit this additional information. As a result, a number of techniques have been developed to jointly estimate the atmospheric blurring effects as well as an estimate of the object being imaged using only the short-exposure images of the object and available prior knowledge. These techniques are called blind deconvolution techniques. Two of the most common techniques are known as multi-frame blind deconvolution (MFBD) and projection-based blind deconvolution (PBBD).

The MFBD algorithm is based upon maximizing a figure of merit to determine the best object and atmospheric short-exposure point spread function (PSF) estimates given the data and available prior knowledge. This figure of merit is created using a maximum likelihood problem formulation. Conceptually, this technique produces estimates of the object and the atmospheric turbulence short-exposure PSFs that most likely generated the measured short-exposure images. Mathematically, the MFBD algorithm obtains these estimates by maximizing the conditional probability density function (pdf) $p[i_m(\vec{x})|_o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M]$, where $i_m(\vec{x})$ is the m th of M short exposure images and is given by

$$i_m(\vec{x}) = h_m(\vec{x}) * o(\vec{x}) + n_m(\vec{x}) \quad (9)$$

where $h_m(\vec{x})$ and $n_m(\vec{x})$ are the atmospheric/system blur and detector noise functions for the m th image, respectively, and the asterisk denotes convolution. The noise is assumed to be zero mean with variance $\sigma^2(\vec{x})$. The conditional pdf $p[i_m(\vec{x})|_o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M]$ is the pdf of the noise with a mean equal to $h_m(\vec{x}) * o(\vec{x})$ and, as a function of $o(\vec{x})$ and $h_m(\vec{x})$, it is known as a likelihood function. For spatially independent Gaussian noise, as is the case for camera read noise, the likelihood function is given by

$$L_G[o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M] = \prod_m \prod_{\vec{x}} \frac{1}{\sqrt{2\pi\sigma^2(\vec{x})}} \exp\left\{-[i(\vec{x}) - (o * h_m)(\vec{x})]^2 / 2\sigma^2(\vec{x})\right\} \quad (10)$$

For Poisson noise, the likelihood function is given by

$$L_P[o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M] = \prod_m \prod_{\vec{x}} \frac{[(o * h_m)(\vec{x})]^{i(\vec{x})} \exp\{-(o * h_m)(\vec{x})\}}{i(\vec{x})!} \quad (11)$$

In practice, logarithms of Eqs. (10) and (11) are taken prior to searching for the estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that maximize the likelihood function.

In general, there are an infinite number of estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that reproduce the original dataset when convolved together. However, when additional prior knowledge is included to constrain the solution space, such as positivity, support, and the knowledge that $h_m(\vec{x})$ is band-limited, along with regularization, the MFBD algorithm almost always converges to the correct estimates of $o(\vec{x})$ and $h_m(\vec{x})$.

The PBBD technique also uses prior knowledge constraints to jointly estimate $o(\vec{x})$ and $h_m(\vec{x})$. However, instead of creating and maximizing a likelihood function, the PBBD technique uses the method of convex projections to find estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that are consistent with the measured short-exposure images and all the prior knowledge constraints. The data consistency constraint and the prior knowledge constraints are mathematically formulated as convex sets and the PBBD technique repeatedly takes the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ and projects them onto the convex constraint sets. As long as all the sets are truly convex, the algorithm is guaranteed to converge to a solution that is in the intersection of all the sets as long as the algorithm doesn't stagnate at erroneous solutions called traps. These traps are the projection equivalent of local minima in cost function minimization algorithms.

A key component of this algorithm is the ability to formulate the measured short-exposure images and prior knowledge constraints into convex sets. As an example, consider the convex set $C_{\text{pos}}[u(\vec{x})]$ that corresponds to the positivity constraint. This set is given by

$$C_{\text{pos}}[u(\vec{x})] \triangleq \{u(\vec{x}) \text{ such that } u(\vec{x}) \geq 0\} \quad (12)$$

It is straightforward to show that $C_{\text{pos}}[u(\vec{x})]$ is a convex set. Other constraints can be converted into convex sets in a similar manner. Although all constraints can be formulated into sets, most but not all of these sets are convex.

The projection process is straightforward to implement. For example, the projections of the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ onto $C_{\text{pos}}[u(\vec{x})]$ are carried out by zeroing out the negative values in these estimates. Similarly, the projections of $o(\vec{x})$ and $h_m(\vec{x})$ onto the convex sets that correspond to their support constraints are carried out by zeroing out all their values outside of their supports. In practice, some versions of PBBD algorithms have been formulated to use minimization routines as part of the projection process in order to minimize the possibility of stagnation.

Deconvolution from Wavefront Sensing

A key distinguishing feature of post-detection compensation algorithms is the type of additional data required to estimate and remove atmospheric blurring in the measured short-exposure images. Blind deconvolution techniques need no additional data. Speckle imaging techniques estimate average atmospheric blurring quantities and need an estimate of the average atmospheric

energy spectrum. Because average quantities are calculated when using speckle imaging techniques, only an estimate of the average atmospheric energy spectrum is needed, not the specific energy spectra that corrupted the measured short-exposure images. As a result, the additional data need not be collected simultaneously with the short-exposure images. The technique described in this section, deconvolution from wavefront sensing (DWFS), seeks to deconvolve the atmospheric blurring in each short-exposure image by estimating the specific atmospheric field phase perturbations that blur each short-exposure image. Thus, the additional data must be collected simultaneously with the short-exposure images. The atmospheric field phase perturbations are calculated from data obtained with a wavefront sensor using a different region of the optical spectrum than used for imaging so that no light is taken from the imaging sensor. Most commonly, the wavefront sensor data consists of the first derivatives of the phasefront (phase differences) as obtained with Shack–Hartmann or shearing interferometer wavefront sensors, but curvature sensing wavefront sensors are also used that produce data that are proportional to the second derivatives of the phasefront.

The original algorithm to implement the DWFS technique produced an estimate $\hat{O}(\vec{f})$ of the Fourier transform $O(\vec{f})$ of the object being imaged using the following estimator:

$$\hat{O}(\vec{f}) = \frac{\sum_m I_m(\vec{f}) \tilde{H}_m^*(\vec{f})}{\sum_m |\tilde{H}_m(\vec{f})|^2} \quad (13)$$

where $\tilde{H}_m(\vec{f})$ is the overall system transfer function (including the telescope and atmospheric effects) for the m th short-exposure image estimated from the wavefront sensor data. To create $\tilde{H}_m(\vec{f})$ the atmospheric field phase perturbations corrupting the m th image must be calculated from the wavefront sensor measurements. Then $\tilde{H}_m(\vec{f})$ is obtained using standard Fourier optics techniques. Although Eq. (13) is easily implemented given an existing system that includes a wavefront sensor, it suffers from the fact that it is a biased estimator. An unbiased version of Eq. (13) is given by

$$\hat{O}(\vec{f}) = \frac{\sum_m I_m(\vec{f}) \tilde{H}_m^*(\vec{f})}{\sum_n H_{n,\text{ref}}(\vec{f}) \tilde{H}_{n,\text{ref}}^*(\vec{f})} \quad (14)$$

where $\tilde{H}_{n,\text{ref}}(\vec{f})$ and $\tilde{H}_{n,\text{ref}}^*(\vec{f})$ are estimates of the overall system transfer function obtained from data collected on a reference star using wavefront sensor data and image-plane data, respectively. Thus, to obtain an unbiased estimator, two more datasets must be collected. The number of frames in the reference star dataset need not be the same as for the imaging dataset. As for speckle imaging, the atmospheric statistics must be the same for the reference star and imaging datasets.

Another technique that uses at least one additional data channel along with the measured image is called phase diversity. In the additional data channels, images are also recorded, but with known field phase aberrations intentionally introduced to add known diversity to the unknown field phase corrupting the measurements. The most commonly introduced field phases aberration in an additional channel is defocus due to its ease of implementation. In the case of defocus, the additional data channel along with the imaging channel can be viewed as samples of the focal volume of the imaging system. Both the noise-free object and the unknown field phase aberrations can be recovered from the measured data using optimization techniques to enforce data consistency along with other prior knowledge constraints. These types of systems can use higher-order aberrations as well. The phase diversity concept has been generalized to an approach known as computational imaging, where an imaging system is designed by taking into account its intended use, the expected image degradation mechanisms, and the fact that the image will be the product of both an optical system and a postprocessing algorithm.

Pre-Detection Compensation

The primary pre-detection-only compensation technique in use today employs dynamical electro-optical systems to sense and remove the phases in the impinging wavefront due to atmospheric turbulence. This technique, known as adaptive optics, provides much higher-quality data than post-detection compensation techniques, when there is enough light from the object being imaged or a guidestar (artificial or natural) to drive the adaptive optics system at a bandwidth high enough to keep up with the temporal changes in atmospheric turbulence. It consists of a wavefront sensor that measures the portion of the phasefront in the pupil of the telescope that is due to atmospheric turbulence, a wavefront reconstructor that generates the deformable mirror drive signals from the wavefront sensor data, and a deformable mirror that applies the conjugate of the atmospheric phase distortions to a reimaged pupil. Some of the light reflected off the deformable mirror is then sent to the imaging sensor. Here, it suffices to say that no adaptive optics system produces imagery that is diffraction limited in quality, although, under the right conditions, the images can be near diffraction limited. However, atmospheric seeing conditions fluctuate significantly over a night as well as seasonally. Therefore, virtually all adaptive optics images benefit from applying post-compensation techniques, as is discussed next.

Hybrid Techniques

Adaptive Optics and Post-Processing

Often, an adaptive optics system does not fully correct for atmospheric turbulence. This can occur when atmospheric conditions become worse than the system is designed to handle. Also, an adaptive optics system may be intentionally designed to only

partially correct for atmospheric turbulence in order to lower the cost of the system. As a result, there still may be significant atmospheric blurring in systems that use adaptive optics. For this reason, additional post-processing of adaptive optics imagery can result in significant improvement in image quality. All of the techniques discussed in the section on post-detection compensation can be applied to adaptive optics imagery. Applying these techniques should never degrade image quality, and will improve the quality if the adaptive optics system doesn't fully correct for atmospheric turbulence. Generally speaking, the amount of improvement decreases as the performance of the adaptive optics system increases. Because the adaptive optics system's performance parameters can vary as atmospheric conditions vary, the blind deconvolution algorithms are especially applicable to processing adaptive optics imagery since they estimate the overall system transfer function from the imaging data. Because speckle imaging techniques require a separate estimate of the overall system transfer function, it can be difficult and time-consuming to obtain all of the additional data necessary to apply these techniques to adaptive optics data. Since these post-detection compensation techniques require short-exposure images, the adaptive optics system needs to have a camera capable of collecting short-exposure images. Finally, even if an adaptive optics system fully corrects for the atmosphere, deconvolution techniques that remove the diffraction-limited transfer function amplitude attenuation can improve image quality.

Partially Redundant Pupil Masking

Although adaptive optics systems can function very well, currently they are expensive and require significant maintenance. The primary benefit of adaptive optics systems is that they remove atmospheric turbulence phase distortions in the pupil of the telescope, thus permitting light from all areas of the pupil to interfere in phase when creating an image. An interferometric view of imaging reveals that multiple regions of the light in the pupil contribute to a single spatial frequency in the image. When the phasefront in the pupil is distorted by the atmosphere, these multiple regions tend to not interfere in phase, reducing the amplitude of the image's Fourier component at that spatial frequency. This amplitude reduction effect occurs for all spatial frequencies greater than the limit imposed by the atmosphere, r_o/λ . Therefore, techniques have been developed to modify the wavefront by masking part of the pupil prior to measuring the field with an imaging sensor. This class of techniques has various names including optical aperture synthesis, pupil masking, and aperture masking. Here this class of techniques will be referred to as partially redundant pupil masking (PRPM).

PRPM techniques function by masking the pupil in such a way as to minimize the amount of out-of-phase interference of light on the imaging detector. Early efforts used pupil masks that were opaque plates with r_o -sized holes in them. Because the atmospheric coherence length is r_o , all the light from an r_o -sized hole is essentially in phase. By choosing the holes in the mask so that only two holes contribute to a single spatial frequency in the image, out-of-phase interference problems are removed. This type of mask is called a nonredundant pupil mask. Of course, a significant amount of light is lost due to this masking, so only bright objects benefit from this particular implementation. Often, the pupil mask needs to be rotated in order to obtain enough spatial frequencies in the image to be able to reconstruct a high-quality estimate of the object. To increase the light levels in the image, masks have been made to allow some out-of-phase interference for a given spatial frequency, thus trading off the amplitude of the spatial frequencies allowed in the image with the amount of light in the image. Masks with this property are called partially redundant pupil masks.

One version of PRPM is called the variable geometry pupil (VGP) technique. It is envisioned to work in conjunction with an adaptive optics system that does not fully correct the wavefront for atmospheric phase distortions. It employs a spatial light modulator or a micro-electromechanical mirror between the deformable mirror and the imaging sensor. Either of these two devices is driven by signals from the wavefront sensor in the adaptive optics system in such a way that the portions of the pupil where the phasefront has a sufficiently large second derivative is blocked, while the remainder of the phasefront is transmitted.

All PRPM techniques benefit from post-compensation processing. For the case of nonredundant pupil masks, post-processing techniques are required to convert the discrete-spatial-frequency information into an image with reasonable quality. These techniques are essentially the same techniques as used in interferometric imaging, not those described in this article. When partially redundant pupil masks are used (including VGP masks), the image may include enough spatial frequency content that the methods described in the post-detection compensation section can be applied.

See also: Imaging Through Scattering Media

Further Reading

- Ayers, G.R., Northcott, M.J., Dainty, J.C., 1988. Knox–Thompson and triple-correlation imaging through atmospheric turbulence. *Journal of the Optical Society of America A* 5, 963–985.
- Bertero, M., Boccacci, P., 1998. *Introduction to Inverse Problems in Imaging*. Bristol, UK: Institute of Physics Publishing.
- Craig, I.J.D., Brown, J.C., 1986. *Inverse Problems in Astronomy*. Bristol, UK: Adam Hilger Ltd.
- Dainty, J.C., 1984. Stellar speckle interferometry. In: Dainty, J.C. (Ed.), *Topics in Applied Physics*, vol. 9: *Laser Speckle and Related Phenomena*, 2nd edn Berlin: Springer-Verlag, pp. 255–320.
- Goodman, J.W., 1985. *Statistical Optics*. New York: John Wiley and Sons.

- Haniff, C.A., Buscher, D., 1992. Diffraction-limited imaging with partially redundant masks I: Infrared imaging of bright objects. *Journal of the Optical Society of America A* 9, 203–218.
- Katsaggelos, A.K., Lay, K.T., 1991. Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In: Katsaggelos, A.K. (Ed.), *Digital Image Restoration*. Berlin: Springer-Verlag, pp. 143–176.
- Nakajima, T., Haniff, C.A., 1993. Partial adaptive compensation and passive interferometry with large ground-based telescopes. *Publications of the Astronomical Society of the Pacific* 105, 509–520.
- Roggemann, M.C., Welsh, B., 1996. *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.
- Schulz, T.J., 1993. Multiframe blind deconvolution of astronomical images. *Journal of the Optical Society of America A* 10, 1064–1073.
- Tyson, R.K., 1991. *Principles of Adaptive Optics*. San Diego, CA: Academic Press.
- Youla, D.C., Webb, H., 1982. Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Transactions on Medical Imaging* M1-1, 81–94.



Encyclopedia of Modern Optics

Second Edition

Editors in Chief

Bob Guenther
Duncan Steel

**ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION**

ENCYCLOPEDIA OF MODERN OPTICS SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 4

Communications ■ Optical Modulation ■ Holography
■ Optical Processing ■ Transformation Optics ■ Optical Interconnects
■ Optics of Semiconductors and 2D Materials ■ Nonlinear Optics ■ Laser Radar



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 4

Editorial Board	v
List of Contributors for Volume 4	ix
Editors in Chief	xiii
Introduction	xv

VOLUME 4

Basic Concepts of Optical Communication Systems	<i>S Lee and AE Willner</i>	1
Historical Development of Optical Communication Systems	<i>G Keiser</i>	13
Dispersion Management	<i>AE Willner, Y-W Song, J McGeehan, Z Pan, and B Hoanca</i>	20
All-Optical Multiplexing/Demultiplexing	<i>Z Ghassemlooy and G Swift</i>	34
Pulse Characterization Techniques	<i>DJ Kane</i>	48
Optical Time Division Multiplexing	<i>LP Barry</i>	60
Lightwave Transmitters	<i>JG McNerney</i>	67
Broadband Passive Optical Access Networks	<i>Elaine Wong, Maluge P Imali Dias, Zhengxuan Li, and Lilin Yi</i>	73
Holographic Recording Media and Devices	<i>Pierre-Alexandre Blanche</i>	87
Colour Holography: Perception Versus Technical Reality	<i>Andrew Pepper</i>	102
High-Resolution Underwater Holographic Imaging	<i>John Watson</i>	106
Digital Holographic Display	<i>Daping Chu, Jia Jia, and Jhensi Chen</i>	113
Holography: Computer Generated Holograms	<i>WJ Dallas and AW Lohmann</i>	130
Module: Digital Holography	<i>Wolfgang Osten</i>	139
Overview: Holography	<i>C Shakher and AK Ghatak</i>	151
The Fractional Order Fourier Transform and Fresnel Diffraction	<i>Pierre Pellat-Finet and Yezid Torres Moreno</i>	157
Ambiguity Function in Optics	<i>JP Guigay</i>	164
Phase Space Tomography in Optics	<i>Tatiana Alieva, José A Rodrigo, and Antonio Picón</i>	174
Coordinate Transformations and the Hough Transform	<i>Filippus S Roux</i>	182
Single-Pixel Imaging Using the Hadamard Transform	<i>Fernando Soldevila, Pere Clemente, Enrique Tajahuerce, and Jesús Lancis</i>	193
Linear Canonical Transforms	<i>Kurt B Wolf</i>	199
Phase-Space Representations of Freeform Optical Systems	<i>Alois M Herkommer</i>	205
Silicon Photonics; Ring Modulator Transmitters	<i>M Ashkan Seyedi and Marco Fiorentino</i>	216
Optical Switches	<i>Dritan Celo, Dominic J Goodwill, and Eric Bernier</i>	224
Indium Phosphide Photonic Integrated Circuits	<i>Yuliya Akulova</i>	242
CMOS Transceiver Circuits for Optical Interconnects	<i>Samuel Palermo</i>	254

Foundations of Coherent Transients in Semiconductors	<i>Torsten Meier and Stephan W Koch</i>	264
Nonlinear Optics in Disordered Media: Anderson Localization	<i>Arash Mafi</i>	278
Second-Harmonic Generation in Two-Dimensional Materials	<i>Myrta Grüning</i>	284
Parity-Time Symmetry in Optics	<i>Mercedeh Khajavikhan, Mohammad-Ali Miri, Andrea Alu, and Demetrios N Christodoulides</i>	291
Laser-Induced Damage in Optical Materials	<i>Wolfgang Rudolph and Luke A Emmert</i>	302
Four-Wave Mixing	<i>L Canioni and L Sarger</i>	310
Kramers–Krönig Relations in Nonlinear Optics	<i>M Sheik-Bahae</i>	317
Nonlinear Optical Phase Conjugation	<i>BY Zeldovich</i>	322
Photorefraction	<i>M Cronin-Golomb and B Kippelen</i>	327
Ultrafast and Intense-Field Nonlinear Optics	<i>AL Gaeta and RW Boyd</i>	335
Electromagnetically Induced Transparency	<i>JP Marangos</i>	339
Nonlinear Optics, Basics: Nomenclature and Units	<i>MP Hasselbeck</i>	347
Raman Spectroscopy	<i>R Withnall</i>	354
Spatial Heterodyne	<i>Mark F Spencer</i>	369
3D Metrics for Airborne Topographic Lidar	<i>Shea T Hagstrom and Myron Z Brown</i>	401
Multiple Input, Multiple Output, MIMO, Active Electro-Optical Sensing	<i>Paul F McManamon and Jeffrey R Kraczek</i>	407
InGaAs Linear-Mode Avalanche Photodiodes	<i>Andrew S Huntington</i>	415
Very High Range Resolution Lidars	<i>Zeb W Barber</i>	430
Multi-Dimensional Laser Radars	<i>Vasyl V Molebny</i>	444
A Review of Laser Range Profiling for Target Recognition	<i>Ove Steinvall</i>	474
Micro-Lidars for Short Range Detection and Measurement	<i>Vasyl V Molebny</i>	496

EDITORS IN CHIEF



Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.



Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Basic Concepts of Optical Communication Systems

S Lee and AE Willner, University of Southern California, Los Angeles, CA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Since modern lightwave communication started in the 1970s, extensive efforts have been made to increase data rate and transmission distance. Recently, the bit rates for single-channel and multichannel systems have exceeded 40 and 2 Tb/s on single fibers respectively, for transoceanic distances. One revolutionary development in lightwave systems has been the deployment of erbium doped fiber amplifiers (EDFAs). These amplifiers facilitate transmission of an optical signal over long distances, by providing periodic analog-like amplification rather than digital-like regeneration. The wide bandwidth provided by EDFAs has made possible the increased use of wavelength division multiplexing (WDM) in which multiple lightwave signals, each having a different wavelength, are co-propagated on a single fiber. EDFAs and WDM techniques can enhance the lightwave system capacity, both in terms of obtainable transmission distance, and total number of data rates (see Fig. 1).

Optical Fiber

Most telecommunication fibers are made of silica, therefore following the attenuation of silica material. Fig. 2 shows the silica fiber attenuation as a function of wavelength. The downward slope towards 1.6 μm and upward slope away from 1.6 μm , are the theoretical limits due to Rayleigh scattering and absorption of silica material, respectively. The absorption peak near 1.4 μm is due to water molecule absorption. There are two low attenuation regions: $\sim 1.3 \mu\text{m}$ and 1.55 μm , both of which are used for communications in general, and some other wavelengths are also used for shorter distance communications. Fig. 2 also shows the comparison between gain bandwidth of EDFA ($\sim 3 \text{ THz}$) and the low attenuation window around 1.55 μm ($\sim 25 \text{ THz}$). The typical values of attenuation at 1.3 μm and 1.55 μm for single-mode fiber (SMF) are $\sim 0.35 \text{ dB/km}$, and 0.2 dB/km , respectively.

When an electromagnetic wave interacts with the bound electrons of a silica fiber, the medium response depends upon optical frequency (ω). This property, referred to as material dispersion, manifests itself through the frequency dependence of the refractive

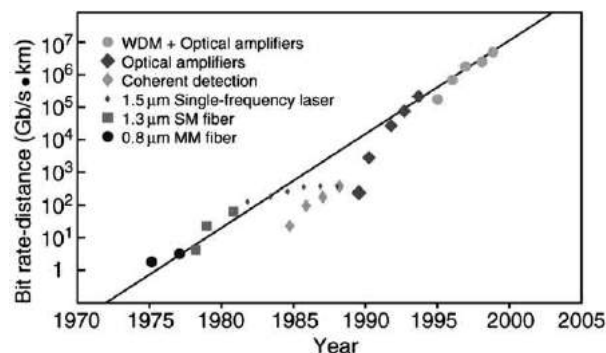


Fig. 1 Bit rate product. Data from T. Li (private communication).

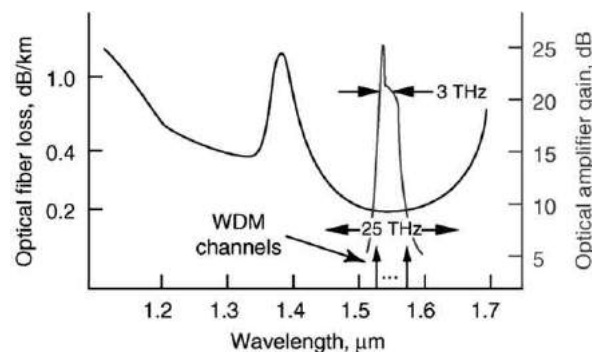


Fig. 2 Silica fiber attenuation versus wavelength.

index $n(\omega)$. Fiber material dispersion plays a critical role in propagation of short optical pulses since different spectral components associated with the pulse travel at different speeds, given by $c/n(\omega)$. Consequently, the optical pulse at the output of the fiber is broadened in time. The commonly used system parameter is called the dispersion parameter D (the negative D value is called normal dispersion, and the positive D value is called anomalous dispersion), and is measured in $\text{ps}/(\text{nm} \cdot \text{km})$; D for SMF is $\sim +17 \text{ ps}/(\text{nm} \cdot \text{km})$ at $1.5 \mu\text{m}$. However, because of dielectric waveguiding, the effective mode index is slightly lower than the material index $n(\omega)$, with the reduction itself being ω -dependent (waveguide dispersion). This results in a waveguide contribution that must be added to the material contribution to obtain the total dispersion (see Fig. 3).

Generally, the waveguide contribution to dispersion D is to shift zero dispersion wavelength (λ_0) to longer wavelengths. Furthermore, the waveguiding can be tailored to shift the λ_0 from $1.3 \mu\text{m}$ to $1.5 \mu\text{m}$, the wavelength at which the fiber has the minimum attenuation loss. The fiber having λ_0 in the neighborhood of $1.5 \mu\text{m}$, is called dispersion-shifted fiber (DSF); D for DSF is usually between -2.5 and $+2.5 \text{ ps}/\text{nm} \cdot \text{km}$, at a wavelength of $1.5 \mu\text{m}$. The typical dispersion parameter, D , as a function of a wavelength for both SMF and DSF, is shown in Fig. 4.

In addition to chromatic dispersion, optical fiber has another dispersion properties related to polarization. An optical wave of arbitrary polarization can be represented as the superposition of two orthogonally polarized modes. In an ideal fiber, these two modes are indistinguishable, and have the same propagation constants owing to the cylindrical symmetry of the waveguide. However, in real fibers there is some residual anisotropy due to unintentional circular asymmetry, usually caused by noncircular waveguide geometry or asymmetrical stress around the core, as shown in Fig. 5.

In either case, the loss of circular symmetry gives rise to two distinct orthogonally-polarized modes with different propagation constants (differential phase velocity) responsible for polarization mode dispersion (PMD) in the fiber, and can be related to the difference in refractive indices (birefringence) between the two orthogonal polarization axes. The differential phase velocity indicated is accompanied by a difference in the group velocities for the two polarization modes, therefore the pulse at the end of

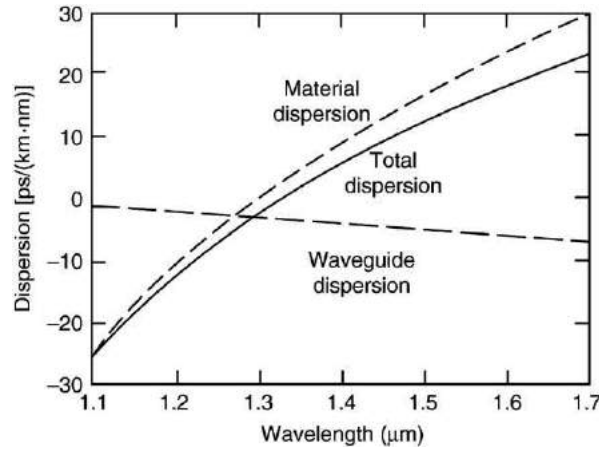


Fig. 3 Total dispersion (D) and contribution of material dispersion and waveguide dispersion in conventional SMF.

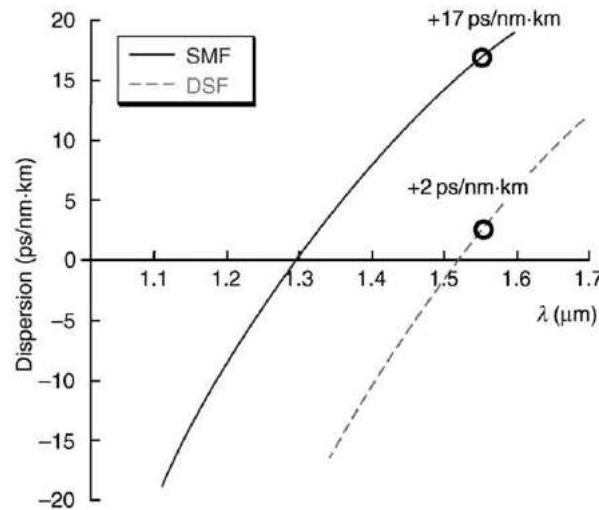


Fig. 4 Dispersion versus wavelength for conventional SMF and DSF.

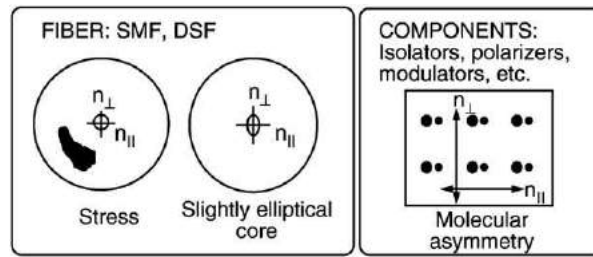


Fig. 5 Origin of PMD.

the transmission fiber is broadened by this differential group delay (DGD). DGD is usually expressed in units of ps/km for a short length (1 m to 1 km) of birefringent fiber. However, DGD does not accumulate along a long fiber link in a linear fashion. Instead, because of random variations in the perturbations along a fiber span, the DGD in one section may either add to or subtract from another section of the fiber. As a result, average DGD in long fiber spans accumulates in a random-walk-like process that leads to a square root of transmission-length dependence. Therefore, average DGD is expressed in ps/km^{1/2} in long fiber spans, referred to as the PMD of the fiber, and the typical PMD parameter has a value of 0.01 to 10 ps/km^{1/2}. Because of the many perturbations that act on a real-world fiber (e.g., temperature, vibration, etc.), transmission properties typically vary with time. Therefore, the PMD of the fiber link fluctuates randomly, thus causing random fluctuations in system performance.

It is well known that optical fibers show nonlinear behavior under conditions of high power and long interaction length. The power-times-distance products for amplified transmission systems can be large enough to make fiber nonlinear effects the dominant factor in the design of long-distance systems.

There are two categories of fundamental optical nonlinear effects: stimulated scattering effects and refractive index effects. Stimulated scattering effects arise from parametric interactions between light and acoustic or optical phonons in the fiber. Two nonlinear effects fall into this category: stimulated Brillouin scattering (SBS) and stimulated Raman scattering (SRS). The main difference between the two is that optical phonons participate in SRS, while acoustic phonons participate in SBS. In a simple quantum-mechanical picture, a photon of the incident field is annihilated to create a photon at a longer wavelength. The new photon is co-propagated and counter-propagated along the original signal in SRS, while counter-propagated in SBS. Refractive index effects are caused by modulation due to changes in light intensity. There are three types of refractive index effects: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). SPM introduces the change of optical phase by its own intensity, and, this leads to frequency chirp of the pulse, depending on the relative position from the peak. When this nonlinear frequency chirp interacts with the fiber dispersion, the pulse either broadens or compresses, depending on the sign and the amount of dispersion along the fiber. In XPM, the phase of the signal in one wavelength channel is modulated by the intensity fluctuation of the other wavelength channels. In a WDM system, XPM imposes far more damaging effects than SPM because XPM is stronger by a factor of two than SPM when the channel power is the same, and large number of WDM channels can contribute to XPM. FWM is the generation of modulation sideband at new frequencies, due to the phase modulation of channels between lights at different frequencies in multichannel system. FWM causes penalties in a WDM system if the newly generated frequency is either equal to or close to the frequency of existing WDM channels.

Fiber nonlinearities impose significant degradation in optical transmission system, and limit the system on allowable number of channels, channel power, and channel spacing (see Fig. 6). But, the effects can be well suppressed by carefully designing the fiber dispersion profile (i.e., dispersion map) in the transmission system, and controlling the optical launch power for each fiber span.

Communication Components

Distributed feedback laser (DFB) is widely used as a light source for metro, long-haul, and undersea applications, due to its narrow spectral width, and wavelength stability. Fabry–Perot (FP) lasers, and vertical cavity surface emitting lasers (VCSEL) are used for local area networks (LAN), and for access applications, because these are cheaper than DFB lasers. For an application, where the cost is the most important parameter with low data modulation speed, light-emitting diode (LED) is used as a light source.

In optical data modulation, the simplest and easiest technique is turning ON or OFF the laser, depending on the binary logic '1' or '0' (direct modulation). This type of modulation is simple and cheap compared to external modulation technique. But, its application is limited by the bandwidth of modulation, and the induced frequency chirp (i.e., difference in optical frequency at the turn ON state and just before the turn OFF state of laser) that limits the transmission distance. Two different types of external modulators are widely used for digital optical data transmission. Electro absorption modulator (EAM) uses the Franz–Keldysh effect, where it is observed that the wavelength of the optical absorption edge in a semiconductor is lengthened by applying an electric field. Electro-optic modulator (EOM) utilizes the linear electro-optic effect, called the Pockels effect, which changes the refractive index of the material caused by and proportional to an applied electric field, so therefore changes the phase of the optical signal. This phase change is used to modulate the intensity of a lightwave through a Mach–Zehnder (MZ) interferometer (see Fig. 7). Because the Pockels effect exists only in crystal, lithium niobate crystal (LiNbO₂) or electro-optic polymers are used for EOM.

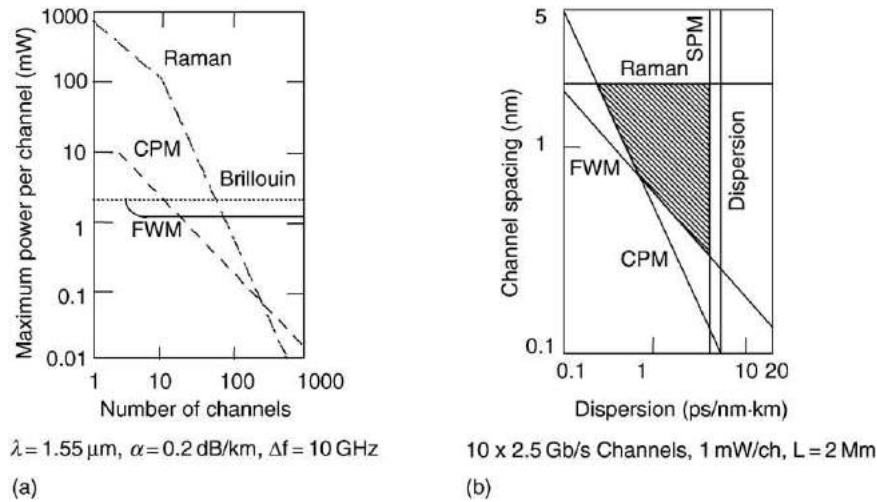


Fig. 6 Limitation due to fiber nonlinearities in (a) maximum channel power and number of channels, and (b) channel spacing and dispersion. Reproduced with permission from Chraplyvy AR (1990) Limitation on lightwave communications imposed by optical-fiber nonlinearities. *IEEE Journal of Lightwave Technology* 8:1548–1557. Copyright © 1990 IEEE.

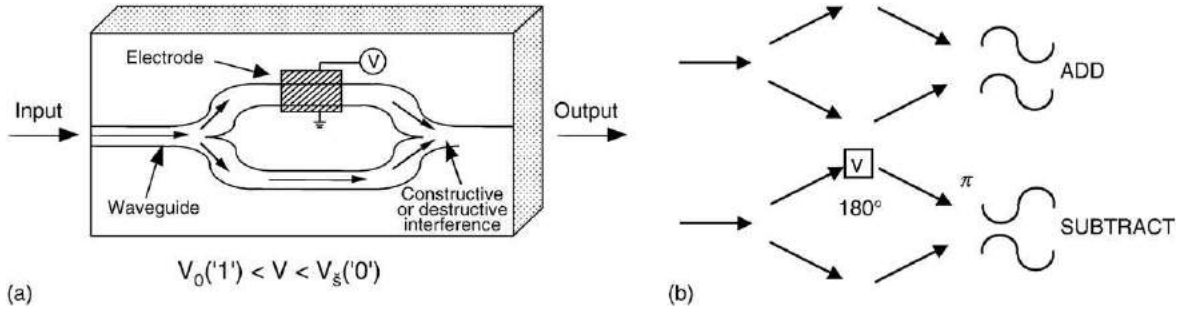


Fig. 7 (a) EOM: MZ interferometer on LiNbO2, (b) operation of EOM; constructive and destructive interference resulting in intensity modulation.

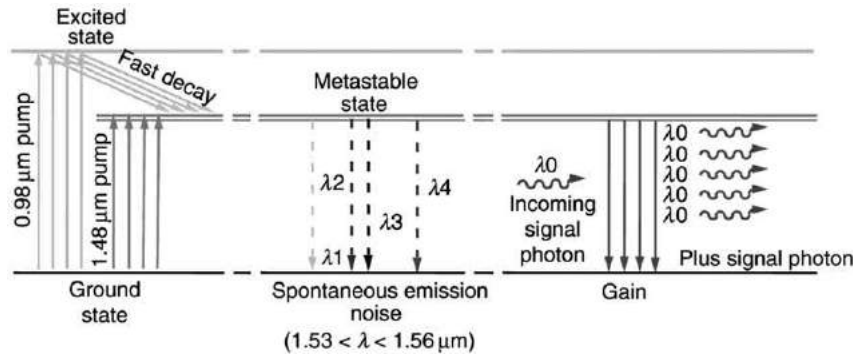


Fig. 8 Energy level and transition diagram of EDFA. Reproduced with permission from Willner AE (1997) Mining the optical bandwidth for a terabit per second. *IEEE Spectrum* Apr.: 32–41. Copyright © 1997 IEEE.

EDFA is a wideband optical amplifier that has merits in that: (i) erbium ions (Er^{3+}) emit light in the $1.55 \mu\text{m}$ loss-minimum band of optical fiber, (ii) a circular fiber-based amplifier is inherently compatible with a fiber optics system; (iii) it provides amplification independent of bit-rate, modulation format, power, and wavelength; and (iv) it has low distortion and low noise during amplification. EDFA contains a gain medium (i.e., erbium-doped fiber) that must be inverted by a pump source. A signal initiates stimulated emission, resulting in gain, and spontaneous emission occurs naturally, which results in noise.

Fig. 8 shows the energy level and transition diagram of EDFA. A $0.98 \mu\text{m}$ pump photon is absorbed and excites a carrier (Er^{3+} ions) into higher excited states, and the excited carrier decays rapidly to the first excited state that has a very long lifetime of $\sim 10 \text{ ms}$ (metastable state). In contrast, a $1.48 \mu\text{m}$ pump photon is absorbed and excites a carrier into a metastable state.

This long metastable state is one of the key advantages of EDFA, and intersymbol distortion and interchannel crosstalk are negligible as a result of these slow dynamics. Depending on the external optical excitation signal, this carrier will decay in a stimulated or spontaneous fashion to the ground state and emit a photon. Both absorption and emission spectra have an associated bandwidth depending on the spread in wavelengths that can be absorbed or emitted from a given energy level. This is highly desirable because (i) a pump laser does not need to be an exact wavelength; and (ii) the signal may be at one of several wavelengths, especially in a WDM system. Fig. 9 shows the bandwidth in the 1.48 μm absorption and 1.55 μm fluorescence spectrum of a typical erbium-doped amplifier.

When the noise characteristics of EDFA are considered, noise figure (NF) is an important parameter, which is defined by $NF \equiv SNR_{in}/SNR_{out}$, where SNR_{in} (SNR_{out}) is the electrical equivalent signal-to-noise ratio (SNR) of the optical wave going into (coming out) of the amplifier, if it were to be detected. The typical value of NF in commercially-available EDFA is 4.5–6 dB. The importance of NF in optical communication systems is presented later in this chapter, when the required optical SNR (OSNR) for error-free optical data transmission is considered. Fig. 10 shows an example of EDFA diagram.

Recent development of EDFA in longer wavelength regimes or L-band (1570–1620 nm) has doubled the useful transmission bandwidth over the conventional band or C-band (1525–1265 nm). L-band gain of EDFA is generally achieved by proper arrangement of pump lasers through longer lengths of erbium-doped fiber than conventional C-band EDFA. Fig. 11(a) shows the concept of wideband EDFA. The incoming WDM signal is split into two bands (C- and L-bands), amplified separately, and combined at the output. Gain and NF characteristics of this wideband EDFA is shown at Fig. 11(b).

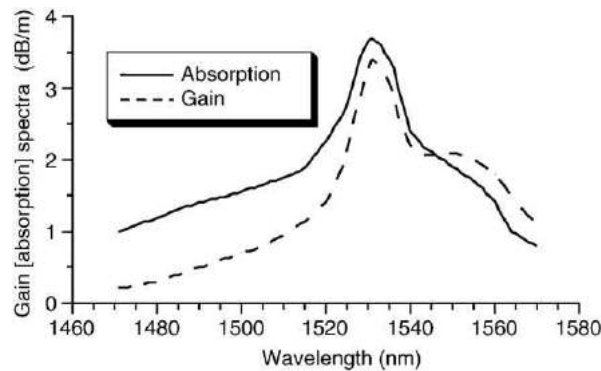


Fig. 9 The absorption and fluorescence spectra for erbium near 1.5 μm . Reproduced with permission from Miniscalco WJ (1991) Erbium-doped glasses for fiber amplifiers at 1550nm. *IEEE Journal of Lightwave Technology* 9: 234–250. Copyright © 1991 IEEE.

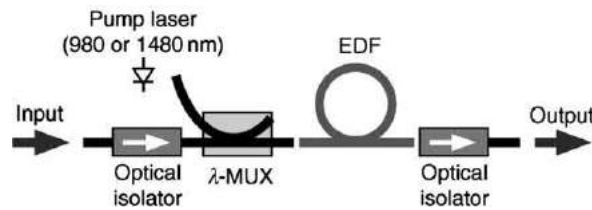


Fig. 10 Typical EDFA diagram.

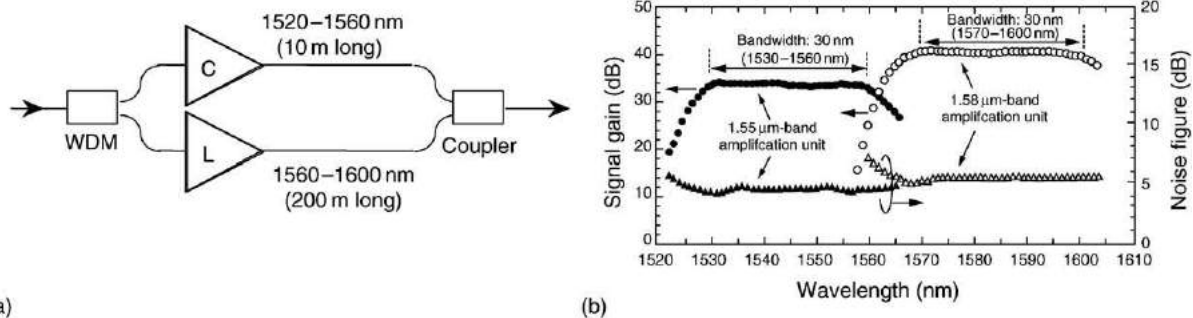


Fig. 11 Wide band EDFA. Reproduced with permission from Yanada M (1997) *Electronics Letters*. Copyright © 1997 IEEE.

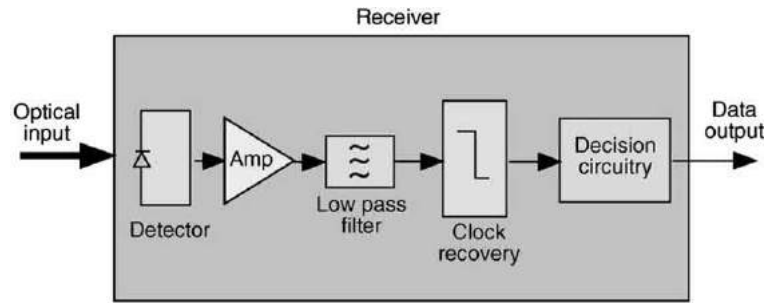


Fig. 12 Block diagram of functions performed in advanced receiver packages.

Because SRS transfers the energy of a lower wavelength channel into the higher wavelength channel in the optical fiber, the transmission fiber itself can be used as a gain medium on specific channel (or channels) when the proper pump signal (or signals) are provided in the fiber. Even though Raman gain is much lower than the gain of EDFA (i.e., < 0.1 dB gain/mW pump in Raman amplifier (RA) compared to a few dB gain/mW pump in EDFA), RA has several advantages over EDFA (i) RA has the capability to provide gain at any signal wavelengths, (ii) the amplification window is expandable by combining multipump wavelengths; and (iii) RA offers improved noise performance because the signal is amplified over transmission fiber (i.e., distributed amplifier). Especially improved noise performance increases system OSNR, which can be used to extend system reach or create more wavelength channels on the system. The Raman pumps are typically located ~ 100 nm shorter than the wavelength of the signal to be amplified, and a few 100s mW pump lasers are used to get a 10–15 dB gain.

In general, telecom receiver (RCVR) consists of photo diode (PD) with transimpedance amplifier (TIA), followed by limiting amplifier (LA) or automatic gain control amplifier (AGC), low pass filter, and clock data recovery circuit (CDR) (see Fig. 12).

Vertically illuminated positive-intrinsic-negative (PIN) photodiode and avalanche photo diode (APD) are the most common PD that is used in RCVR. In PIN, incoming light is absorbed in the n^- (low doped or intrinsic) region and generates an electron-hole pair, which drifts to a depletion region and collected n^+ and p^+ regions, therefore generating electric current. APD is used for high-sensitivity detection up to 10 Gb/s. It is structured as absorption, grading, and multiplication layers. The multiplication layer of APD provides additional transit time delay compared to PIN, therefore the bandwidth of APD is smaller than PIN in general. TIA is a broadband low noise amplifier that is used to convert and amplify the weak current from PD to the desired output voltage. A low pass filter is used to suppress the noise and data distortion from high frequency components with a typical cut-off frequency ~ 0.7 bitrate. LA and AGC are amplifiers that limit the output voltage as a constant. LA is a nonlinear amplifier that makes a decision either to make logic '1' or '0' based on input voltage level, and limits output voltage as fixed values. AGC is a linear amplifier that limits the output voltage by adjusting the gain of the amplifier based on input voltage level. CDR consists of a phase lock loop (PLL) with a voltage controlled oscillator (VCO) and data flip-flop circuit. The PLL is used to synchronize the VCO to the incoming high-speed bit stream, therefore recovering the system clock. The recovered clock is used in D-FF to retiming and regenerate the data.

Data Modulation Formats

In digital optical communications, numerous modulation formats have been developed and utilized, in order to achieve higher spectral efficiency (i.e., ratio of individual channel bit rate to dense WDM (DWDM) separation), and to improve system performance over fiber impairments. A simple amplitude modulation of lightwave is still the most widely used method for optical communication. The easiest amplitude modulation technique is to using a nonreturn-to-zero (NRZ) format. Fig. 13(a) shows an example of NRZ binary sequence in time domain, and the corresponding frequency domain spectrum. In NRZ, data '1' and '0' correspond to the ON and OFF state of a lightwave transmitter, respectively, occupying entire bit time (i.e., binary symbol time). When the ON or OFF state of a lightwave transmitter does not occupy the entire bit time (T_B) (i.e., when the data pulse duration (T_d) is less than (T_B) it is called the return-to-zero (RZ). Fig. 13(b) shows a binary sequence of RZ, and its frequency spectrum when $T_d \approx 0.5T_B$. The Fourier transform of the square wave is the Sinc function. Therefore, the frequency spectrum of an ideal NRZ and RZ follows the pattern of a Sinc square function. Note that the first null of frequency spectrum of RZ is about twice as high as the null of NRZ.

Even though NRZ is better in spectral efficiency than RZ, RZ has an advantage over NRZ in tolerance of fiber impairments. Fig. 14 shows an example on eye closure penalty comparison between NRZ and RZ data formats at a 16-channel WDM system at 40 Gb/s data rate. It is clearly seen that the penalty of RZ is much less severe than the penalty of NRZ, as the transmission distance increases. (Note, nonlinearities and dispersion increases as the transmission distance increases.)

There has been much effort to increase the spectral efficiency, to improve the system performance over fiber impairments, or to achieve both simultaneously. Fig. 15 shows some of examples of these. Chirped RZ (CRZ) is the modulation format inducing frequency chirp on conventional RZ. CRZ is robust on fiber impairment at the cost of losing spectral efficiency. In carrier-suppressed RZ (CS-RZ), a strong un-modulated optical carrier is removed from conventional RZ, therefore suppressing the nonlinear effects. Duo-binary is a three-level signal modulation format generated by a series of delay and added circuits in the

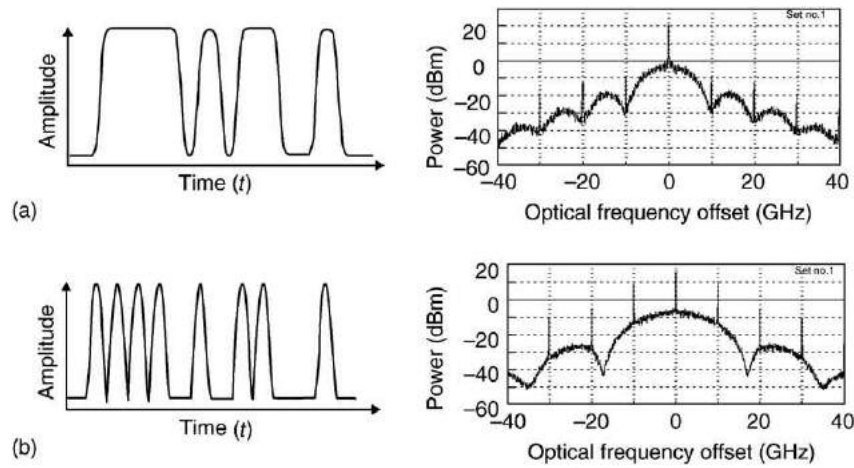


Fig. 13 Amplitude modulation (a) nonreturn-to-zero (NRZ), and (b) return-to-zero (RZ).

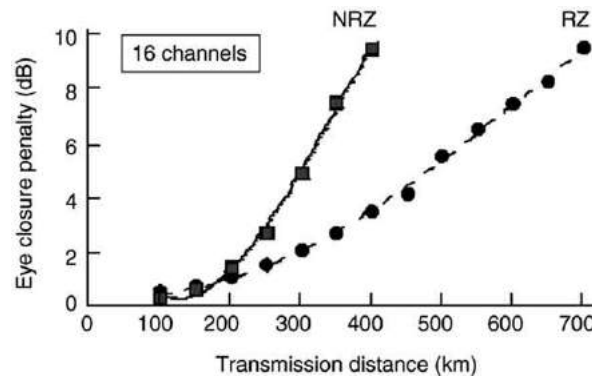


Fig. 14 System performance of different modulation format (NRZ vs. RZ).

transmitter. The advantage of duo-binary is that its bandwidth is reduced to one half of conventional NRZ, therefore having enhanced spectral efficiency, and has a high tolerance in chromatic dispersion, and suppression of fiber nonlinearities.

Single-sideband (SSB) or vestigial singleside band (VSB) can be generated into either NRZ or RZ, by either optically utilizing optical filtering with a sharp cut-off filter, or electrically applying tapped-delay-line filter approximation on both amplitude and phase modulators. SSB (or VSB) has about a two times higher spectral efficiency and higher tolerance on chromatic dispersion than conventional modulation formats.

Data Multiplexing

At their most basic, optical networks can imitate electrical networks in which time division multiplexing (TDM) is overwhelmingly used for digital data transmission. A fiber can carry many time-multiplexed channels, in which each channel can transmit its data in an assigned time slot. A typical TDM link is shown in **Fig. 16**, in which N transmitters are sequentially polled by a fast multiplexer to transmit their data. The time-multiplexed data are sequentially and rapidly demultiplexed at the receiving node.

Time multiplexing and demultiplexing functions can be performed either electrically with an electrical time (DE)MUX switch and an optical transmitter/receiver, or optically with multiple optical transmitters/receivers and an optical time (DE)MUX switch. The major advantage of TDM is that there is no output-port contention problem (each data bit occupies its own time slot and there is only a single high-speed signal present at any given instant). The disadvantage of TDM is that the scheme requires a ultra-high-speed switching component if the individual signals are themselves high-speed and if there are many users.

In WDM, multiple wavelength channels are transmitted through a single fiber, therefore enabling the fiber to carry more throughput. By using wavelength-selective devices for the ON and OFF ramps, independent signal routing also can be accomplished. **Fig. 17(a)** shows a simple point-to-point WDM system in which several channels are multiplexed at one node, the combined signals transmitted across a distance of fiber, and the channels demultiplexed at a destination node. As shown in

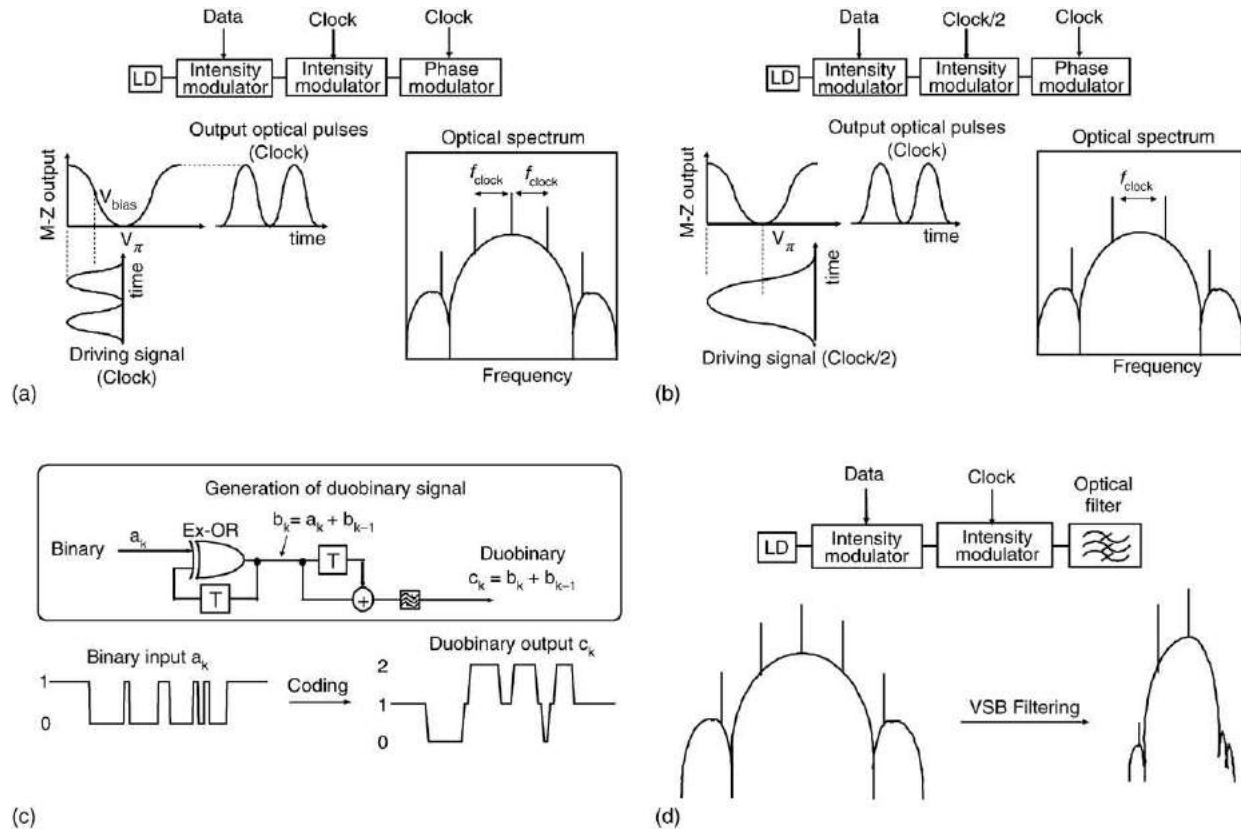


Fig. 15 Examples of bandwidth efficient modulation formats. (a) Chirped RZ (CRZ) generation, (b) carrier suppressed-RZ (CS-RZ) generation, (c) duo-binary generation, and (d) vestigial single side band (VSB) RZ generation.

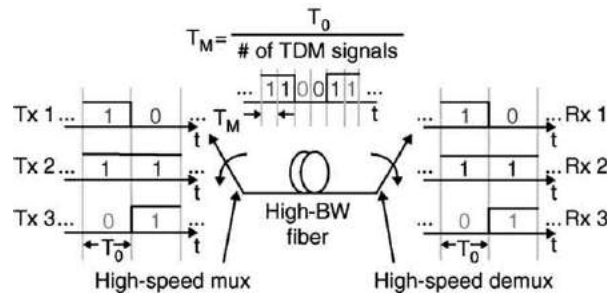


Fig. 16 Concept of bit-interleaving TDM.

Fig. 17(b), the wavelength becomes the signature address for either the transmitter or the receivers, and the wavelength will determine the routing path through an optical network.

Many interesting challenges face the eventual implementation of WDM systems and networks. Several of these challenges involve the control and management of the data through this novel high-speed network. **Fig. 18** shows a small subset of critical component technologies typically required for a WDM system, including multiple-wavelength transmitters, multiport star couplers, passive and active wavelength routers, EDFAs, and tunable optical filters.

Some generic goals that a WDM-device technologist aims to achieve include; large wavelength tuning range; multi-user capability; wavelength selectivity and repeatability; low cross-talk; high extinction ratio; minimum excess losses; fast wavelength tunability; high-speed modulation bandwidth; low residual chirp; high finesse; low noise; robustness; high yield; potential low cost. Depending on system requirements, device availability and cost, WDM technologies divide into two; course-WDM (CWDM) and dense-WDM DWDM. **Fig. 19** shows the wavelength allocation for CWDM and DWDM. CWDM technology uses an International Telecommunication Union (ITU) standard 20 nm spacing between the wavelengths from 1310 nm to 1610 nm. Also DWDM technology uses an ITU standard 100 GHz or 200 GHz between wavelengths, arranged in several bands at around 1500–1600 nm.

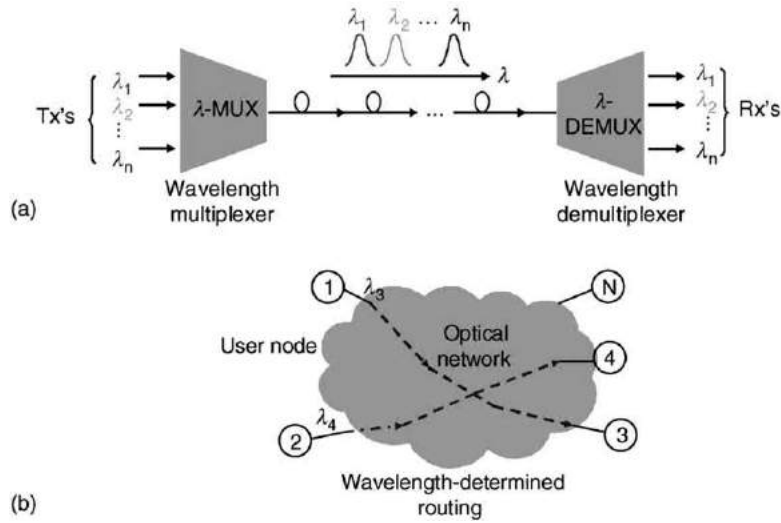


Fig. 17 (a) A simple point-to-point WDM transmission system; (b) a generic multiuser network in which the communication links and routing paths are determined by the wavelengths used within the optical switching fabric.

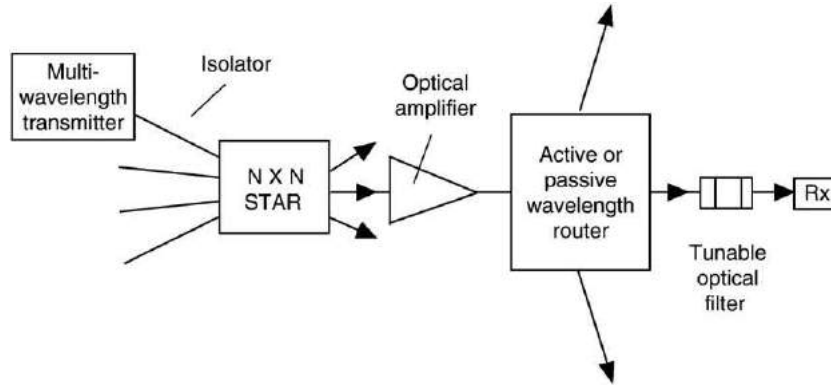


Fig. 18 Schematic of a small subset of enabling device technologies for a WDM system.

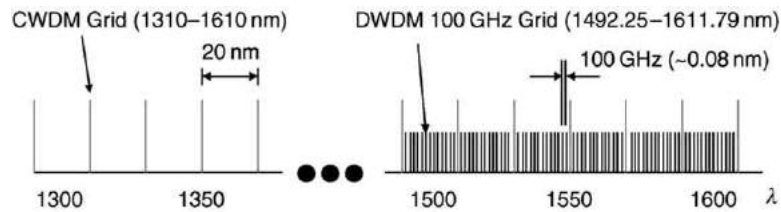


Fig. 19 WDM wavelength allocation.

Dispersion Management and Compensation

As mentioned previously, chromatic dispersion is one of the most basic characteristics of fiber, although it is possible to manufacture fiber that induces zero chromatic dispersion. But such fiber is incompatible with the deployment of a WDM system since harmful nonlinear effects are generated, therefore chromatic dispersion must exist, and must be compensated for. The effect of chromatic dispersion is cumulative and increases quadratically with the data rate (see Fig. 20).

The quadratic dependence of dispersion with the data rate is a result of two effects. First, a doubling of the data rate will double the Fourier-transformed frequency spectrum of the signal, thereby doubling the effect of dispersion. Second, the same doubling of the data rate makes the data pulse only half as long in time and therefore twice as sensitive to temporal spreading due to dispersion. The conventional wisdom for the maximum distance over which data can be transmitted is to consider a broadening of

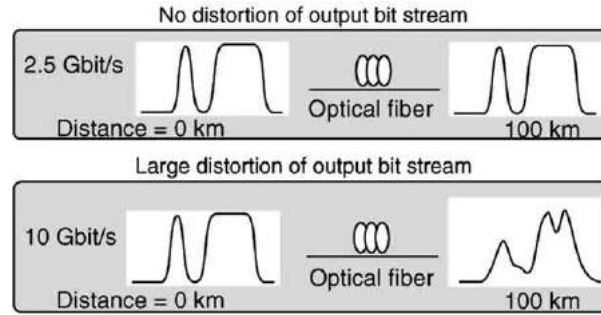


Fig. 20 Pulse broadening at two different data rates (2.5 Gbit/s, and 10 Gbit/s) as a result of quadratic nature of chromatic dispersion.

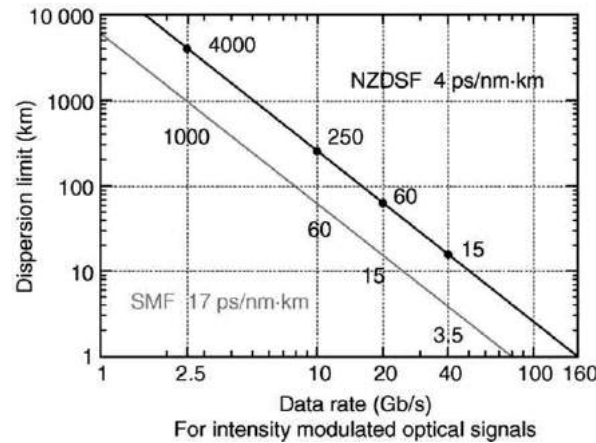


Fig. 21 Dispersion-limited (uncompensated) transmission distance in single-mode fiber (SMF) as a function of data rate for intensity-modulated optical signals. Courtesy of L.D. Garrett.

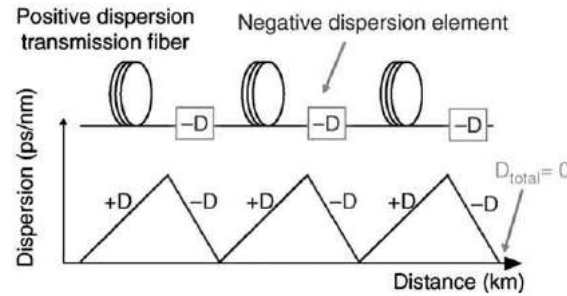


Fig. 22 Typical static management of chromatic dispersion.

the pulse equal to the bit time period. For a bit period B a dispersion value D and a spectral width $\Delta\lambda$ the dispersion-limited distance is given by $L_D = 1/(D \cdot B \cdot \Delta\lambda)$ (see Fig. 21).

In theory, compensation of chromatic dispersion for high-speed or long-distance systems can be fixed in value if each link's dispersion value is known. A simple yet elegant solution is to create a dispersion 'map', in which the designer of a transmission link alternates elements that produce positive and then negative dispersion (see Fig. 22). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, effectively eliminating FWM and XPM, but the total accumulated dispersion at the end of the fiber link is zero, so that minimal pulse broadening is induced. The specific system design, as to the periodicity of management, depends on several variables, but a typical number for SMF as the embedded base is a compensation at every 80 km in a 10 Gb/s system. Dispersion-compensating fiber (DCF) has been generally used as a dispersion compensating element.

However, there are several important aspects of optical systems and networks that make tunable dispersion compensation solutions attractive, including: (i) it significantly reduces the inventory of different required types of compensation modules; (ii) it tunes to adapt to routing path changes in a reconfigurable network, (iii) it tracks dynamic changes in dispersion due to environment, and (iv) it achieves a high degree of accuracy necessary for 40 Gb/s channels (see Fig. 23).

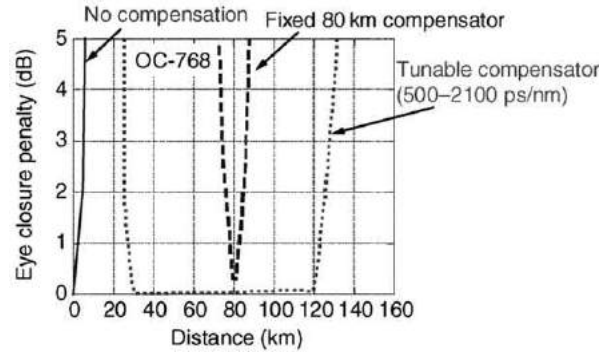


Fig. 23 Effect of tunable compensation at 40 Gb/s (OC-768); it allows a wide range of transmission distance. On the contrary, the 80 km fixed compensator allows only small range of transmission around 80 km.

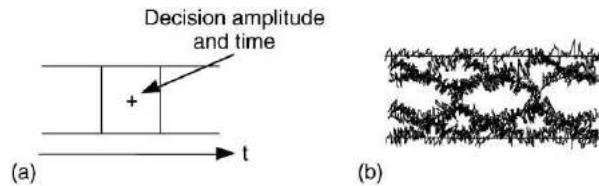


Fig. 24 (a) A perfect square eye diagram, (b) a closed eye diagram due to bandwidth, fiber impairments, noise, and timing jitter.

One brute-force method to achieve tunable dispersion compensation is to build a module with optical switches used to add or remove sections of fixed DCF to achieve a discrete set of dispersion compensation. Many other elegant, yet viable, approaches have been developed for tunable dispersion compensation, and these approaches include linearly chirped FBG with nonuniform heating, nonlinearly chirped Fiber Bragg grating (FBG) with a simple mechanical stretcher, virtually imaged phase array, electronic tap delay filter with weights, multiple stage all pass filter, etc.

System Performance Parameters

An eye diagram provides a simple and useful tool to visualize intersymbol interference between data bits. **Fig. 24(a)** shows a perfect eye diagram. A square bit stream (i.e., series of symbol '1's and '0's) is sliced into sub-bit stream with predetermined eye intervals (i.e., several bit periods), and displayed through bit analyzing equipment (e.g., digital channel analyzer), overlapping the sliced sub-bit stream in order to obtain the eye diagram. When a perfect transmitter and receiver (i.e., infinite receiver bandwidth with zero noise characteristics and jitter), and a perfect transmission media (i.e., no dispersion and nonlinearities) are used, the received eye diagram is shaped as a perfect rectangular. In reality, the transmitter and receiver have a limited bandwidth with noise and jitter, and the transmission media (i.e., optical fiber) has dispersion and nonlinearities. Therefore, the eye diagram deviates from the perfect rectangular shape. **Fig. 24(b)** shows the eye diagram close to a real situation. The shape of the eye is generally broadened and distorted (i.e., eye is closed) due to limited bandwidth and fiber impairments, and noise and timing jitter are added onto this broadened and distorted eye shape.

The Q-factor (q) is also an important system parameter widely used in long-distance optical transmission system design. It is defined as the electrical signal-to-noise ratio before the decision circuit at receiver. The Q-factor is directly related to bit-error-rate (BER: the percentage of bits that has an error relative to total number of bits received in a specific time) by: $BER = 0.5 \cdot \text{erfc}(q/2^{0.5})$, where $\text{erfc}(x)$ is the complementary error function. The Q-factor is a unitless linear ratio, and is expressed in dB by $20 \cdot \log(q)$. As a conventional relationship, Q-factor of 15.6 dB (i.e., linear ratio 6) is required to achieve $BER = 10^{-9}$. **Fig. 25** shows the relationship of eye diagram, Q-factor, and BER.

Power penalty (PP) is one of the most important system parameters, and is defined as the received optical power difference in dB with and without signal impairments at a specified BER (conventionally 10^{-9}), from measured BER versus optical power curve. Therefore, 1 dB PP means that a system with signal impairments requires 1 dB more optical power at the receiver in order to achieve the same BER performance compared to the system without signal impairments. **Fig. 26** shows the typical BER curve as a function of received optical power and PP. Note that PP reduces from ~ 3.5 dB to ~ 1 dB with dispersion compensation.

In long-haul and undersea transmission system with many EDFA chains, OSNR is an important system parameter to design and characterize the optical transport system. OSNR is defined as the ratio of the power of signal channel to the power of ASE in a specified optical bandwidth (conventionally 0.1 nm). The OSNR (in dB) of a signal channel at the end of the system is approximated by: $OSNR \text{ (dB)} = 58 + P_{\text{out}} - L_{\text{span}} - NF - 10 \log(N_{\text{amp}})$, when the system has N_{amp} fiber spans, span loss L_{span}

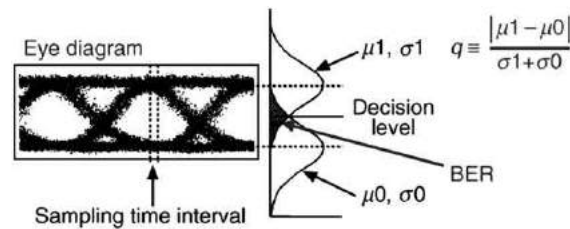


Fig. 25 The relationship of eye diagram, Q-factor, and BER (μ_1, μ_2 : mean levels of logic '1' and '0', σ_1, σ_2 : standard deviation of logic '1' and '0').

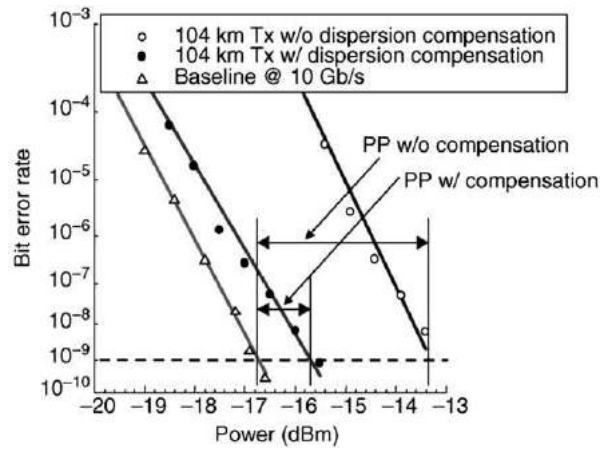


Fig. 26 BER curve and power penalty.

(in dB) followed by an optical amplifier with output power P_{out} (in dBm) per channel launched into the span and noise figure (NF) (in dB). As an example, a typical OSNR requirement for $\text{BER}=10^{-9}$ is about 17 dB at 10 Gb/s data rate. The required OSNR should be increased by 6 dB when the data rate is increased by a factor of 4. Improving the amplifier's NF can increase OSNR, and the improved OSNR can be used to increase the system reach, reducing the channel power in order to suppress the nonlinearities, etc.

See also: Historical Development of Optical Communication Systems. Measuring Fiber Characteristics

Further Reading

- Emmanuel, D., 1994. Erbium-Doped Fiber Amplifiers. New York: Wiley.
 Govind, P.A., 1995. Nonlinear Fiber Optics. San Diego, CA: Academic Press.
 Ivan, K., Tingye, L., 2002. Optical Fiber Telecommunications IV A Components. San Diego, CA: Academic Press.
 Ivan, K., Tingye, L., 2002. Optical Fiber Telecommunications IV B Systems and Impairments. San Diego, CA: Academic Press.
 Leonid, K., Sergio, B., Alan, W., 1996. Optical Fiber Communication Systems. Norwood, MA: Artech House Inc.

Historical Development of Optical Communication Systems

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

This article describes the development of optical fiber communication systems. After describing some of the motivations for using optical fiber communications and the advantages of this technology, the key milestones and the principal people involved in developing optical fibers and compatible light sources are presented. Following this, the article looks at the evolution of fielded systems and the use of optical fiber links in undersea applications.

One of the principal needs of people since antiquity has been to communicate. This need created interests in devising communication systems for sending messages from one distant place to another. Many forms of such systems have appeared over the years. The basic motivations behind each new one were either to improve the transmission fidelity, to increase the data rate so that more information could be sent, or to increase the transmission distance between relay stations. Prior to the nineteenth century, all communication systems operated at a very low information rate and involved only optical or acoustical means, such as signal lamps or horns. One of the earliest known optical transmission links, for example, was the use of a fire signal by the Greeks in the eighth century BC for sending alarms, calls for help, or announcements of special events.

The invention of the telegraph by Samuel FB Morse in 1838 ushered in the era of electrical communications. In the ensuing years an increasingly larger portion of the electromagnetic spectrum, shown in Fig. 1, was utilized for the conveying of information from one place to another. The reason for this trend is that, in electrical systems, the data usually are transferred over the communication channel by superimposing the information onto a sinusoidally varying electromagnetic wave, which is known as the carrier. When reaching its destination the information is removed from the carrier wave and processed as desired. Since the amount of information that can be transmitted is related directly to the frequency range over which the carrier operates, increasing the carrier frequency theoretically increases the available transmission bandwidth and, consequently, provides a larger information capacity. Thus the trend in electrical communication system developments was to employ progressively higher frequencies (shorter wavelengths), which offer corresponding increases in bandwidth or information capacity. This activity led to the birth of communication mechanisms such as radio, television, microwave, and satellite links.

Another important portion of the electromagnetic spectrum encompasses the optical region shown in Fig. 1. In contrast to electrical communications, transmission of information in an optical format is carried out, not by frequency modulation of the carrier but by varying the intensity of the optical power. Similar to the radio-frequency spectrum, two classes of transmission medium can be used: an atmospheric channel or a guided-wave channel. Whereas transmission of optical signals through the atmosphere was done thousands of years ago, the use of a guided-wave optical channel such as an optical fiber is a fairly recent application.

Fiber optic communication systems have a number of inherent advantages over their copper-based and radio-transmission counterparts. Fiber optic cable can transmit at a higher capacity, thereby reducing the number of physical lines and the amount of equipment needed for a given transmission span. The lower weight and smaller size of optical fibers offer a distinct advantage over heavy, bulky copper cables in crowded underground cable ducts, in ceiling-mounted cable trays, and in mobile platforms such as aircraft and ships. Their dielectric composition is an especially important feature of optical fibers, since this ensures freedom from electromagnetic interference between adjacent fibers, eliminates ground loops, and results in extremely low fiber-to-fiber crosstalk.

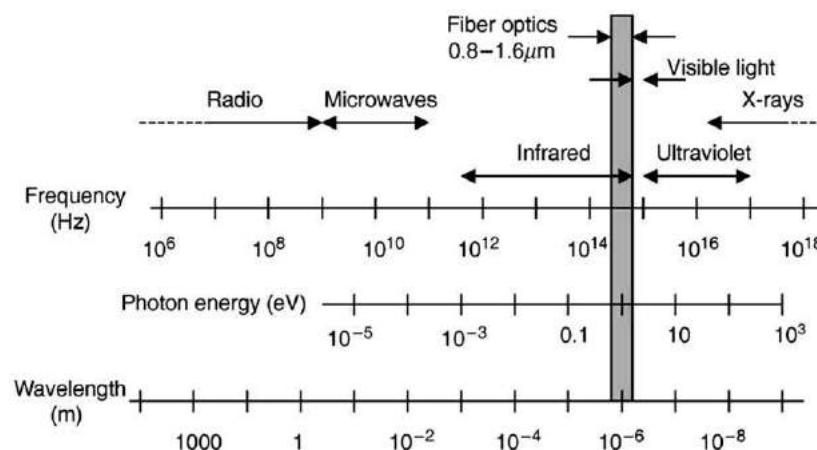


Fig. 1 Electromagnetic spectrum for electrical and optical communications.

In addition, an optical fiber affords a high degree of data security since the optical signal is well confined within the optical waveguide.

Optical Fiber Development

Since an atmospheric channel requires a line-of-sight link and because it can be adversely affected by weather conditions, a guided-wave channel is the preferred approach in most cases for communication system applications. A challenge in using an optical fiber channel is to have a flexible, low-loss medium that transfers the optical signal over long distances without significant attenuation and distortion. Glass is an obvious material for such applications. The earliest known glass was made around 2500 BC and glass already was drawn into fibers during the time of the Roman Empire. However, such glasses have very high attenuations and thus are not suitable for communication applications. One of the first known attempts of using optical fibers for communication purposes was a demonstration in 1930 by Heinrich Lamm of image transmission through a short bundle of optical fibers for potential medical imaging. However, no further work was done beyond the demonstration phase since the technology for producing reasonably low-loss fibers with good light confinement was not yet mature.

Further work and experiments continued on using optical fibers for image transmission and by 1960 glass-clad fibers had attenuations of about one decibel per meter. This attenuation allowed fibers to be used for medical imaging, but it was still much too high for communications. Optical fibers had attracted the attention of researchers at that time, because they were analogous in theory to plastic dielectric waveguides used in certain microwave applications. In 1961, Elias Snitzer demonstrated this similarity by drawing fibers with cores so small they carried light in only one waveguide mode. He published a classic theoretical description of single-mode fibers in May 1961. However, to be useful for communication systems, optical fibers would need to have a loss of no more than 10 or 20 decibels per kilometer.

As a reminder, decibels measure the ratio of the output power to the input power on a logarithmic scale. The abbreviation for a decibel is 'dB.' The power ratio in decibels is given by the expression ' $10 \log (\text{power out}/\text{power in})$.' As an example, a power loss of 20 dB over a 1-km distance in an optical fiber means that only 1% of the light injected into the fiber comes out of the other end. **Table 1** gives some typical examples of various power ratios and their decibel equivalents.

In the early 1960s Charles Kao pursued the idea of using a clad glass fiber for an optical waveguide, building on optical waveguide research being done at the Standard Telecommunication Laboratories in England. After he and George Hockman painstakingly examined the transparency properties of various types of glass, Kao made a prediction, in 1966, that losses of no more than 20 dB/km were possible in optical fibers. In July 1966, Kao and Hockman presented a detailed analysis for achieving such a loss level. Kao then went on to actively advocate and promote the prospects of fiber communications, which generated interest in laboratories around the world to reduce fiber loss. It took four years to reach Kao's predicted goal of 20 dB/km, and the final solution was different from what many had expected.

To understand the process of making a fiber, consider the schematic of a typical fiber structure shown in **Fig. 2**. A fiber consists of a solid dielectric (glass or plastic) cylinder of refractive index n_1 called the core. This is surrounded by a dielectric cladding which has a refractive index n_2 , that is less than n_1 , in order to achieve light guiding in the fiber. This structure is then encapsulated by a buffer coating to protect the fiber from abrasion and outside contaminants. The first step in making a fiber is to form the clear glass

Table 1 Examples of some optical power ratios and their decibel equivalents

Power ratio (P_{out}/P_{in})	Decibel equivalent (dB)
0.01	-20
0.10	-10
0.50	-3
1	0
2	3
10	10
100	20

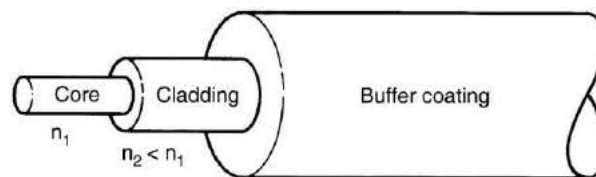


Fig. 2 Typical optical fiber structure.

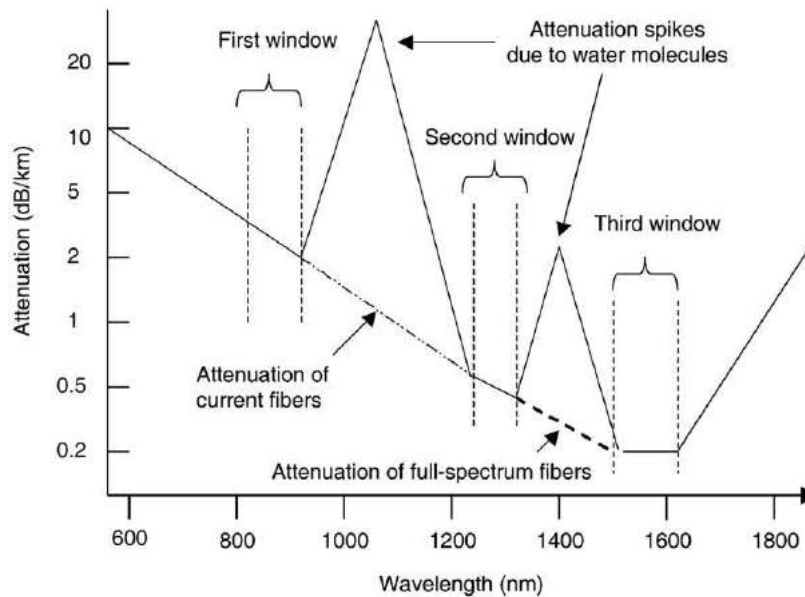


Fig. 3 Attenuation versus wavelength approximation for silica fibers.

rod or tube called a preform. This normally is done by a vapor-phase oxidation process. The preform has two distinct regions that correspond to the core and cladding of the eventual fiber. Fibers are made from the preform by precision feeding it into a circular furnace that softens the end of the preform to the point where it can be drawn into a very thin filament which becomes the optical fiber.

Most researchers had tried to purify the compound glasses used for standard optics, which are easy to melt and draw into fibers. A different approach was taken at the Corning Glass Works where Robert Maurer, Donald Keck, and Peter Schultz started with fused silica. This material can be made extremely pure, but has a high melting point and a low refractive index. Silica has the approximate attenuation versus wavelength characteristic shown in Fig. 3. Note that for early silica fibers there are regions of low attenuation around 850, 1310, and 1550 nm, which the literature refers to as the first, second, and third windows, respectively. The large attenuation spikes in the 1000- and 1400-nm spectral regions are due to absorption by residual water molecules in the glass. Although ultra-pure material processing techniques can eliminate these spikes to produce what are known as full-spectrum or low-water-peak fibers, most installed fibers still have a relatively large attenuation between 1350 and 1500 nm.

Note that since the attenuation spikes no longer are present in various types of fibers, the idea of operating windows has been replaced by the concept of spectral bands. The International Telecommunications Union (ITU) has designated six spectral bands for use in intermediate-range and long-distance optical fiber communications within the 1260 to 1675-nm region. The regions, which are known by the letters O, E, S, C, L, and U, are defined as follows:

- Original Band (O-Band): 1260 to 1360 nm
- Extended Band (E-Band): 1360 to 1460 nm
- Short Band (S-Band): 1460 to 1530 nm
- Conventional Band (C-Band): 1530 to 1565 nm
- Long Band (L-Band): 1565 to 1625 nm
- Ultra-long Band (U-Band): 1625 to 1675 nm

The Corning team made cylindrical preforms by depositing purified materials from the vapor phase. To produce a fiber that has light guiding properties they carefully added controlled trace levels of titanium to the core to make its refractive index slightly higher than that of the cladding without raising the attenuation significantly. In September 1970, they announced the fabrication of single-mode fibers with an attenuation of 17 dB/km at the 633-nm helium-neon line. The fibers were fragile, but independent tests at the British Post Office Research Laboratories facility in Martlesham Heath, England confirmed the low loss.

This dramatic breakthrough was the first among the many developments that opened the door to fiber optic communications. The ensuing years saw further reductions in optical fiber attenuations. By mid-1972, Maurer, Keck, and Schultz at Corning had made multimode germania-doped fibers with a 4-dB/km loss and much greater strength than the earlier brittle titania-doped fibers. In order to couple a sufficient amount of optical power into a fiber, early applications used multimode fibers with a refractive-index gradient between core and cladding of around 2% and core diameters of 50 or 62.5 micrometers.

Single-mode fibers have much smaller core diameters on the order of 9 micrometers in order to allow only one propagation mode. This type of fiber has a much higher transmission capacity since the effect of modal dispersion is eliminated. The first single-mode fibers were optimized to have a zero dispersion value at 1310 nm, since the silica material used at that time exhibited a low

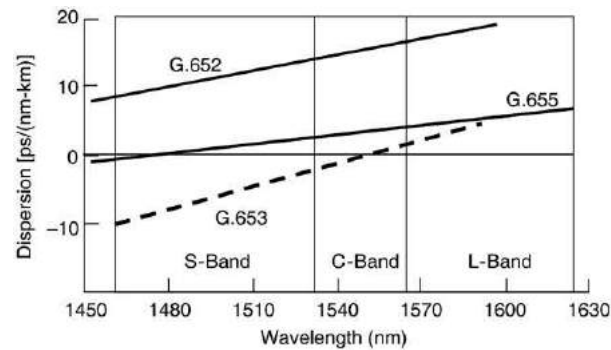


Fig. 4 Dispersion characteristics of three types of standard optical fibers.

loss within a spectral band around this wavelength. **Fig. 4** shows the dispersion characteristic of this type of fiber as a function of wavelength in the S-, C-, and L-bands. As a result of its widespread use in early single-mode transmission systems, this fiber design has been standardized by the International Telecommunication Union (ITU-T) under the designation Recommendation G.652.

Standard G.652 silica fibers provide the lowest attenuation at 1550 nm, but have a much larger signal dispersion at this wavelength than at 1310 nm. Since system designers wanted to use fibers at a 1550-nm wavelength, in order to transmit high-speed data over longer distances, fiber manufacturers overcame the larger signal dispersion limitation by creating the so-called dispersion-shifted fibers. This was done through a clever manipulation of the core and cladding designs that allowed the zero-dispersion point to shift to longer wavelengths. In particular, the ITU-T created a specification for operation at 1550 nm which is designated as recommendation G.653. **Fig. 4** also shows the dispersion characteristic of this type of fiber as a function of wavelength.

Although the G.652 and G.653 fibers work well for single-wavelength operation, a different type of fiber, having non-zero dispersion within a broad spectral range, is needed when implementing systems that use many independent light sources simultaneously within a particular wavelength band. This led to the specification of the non-zero dispersion shifted fiber (NZDSF) in the ITU-T Recommendation G.655 that are designated to operate around 1550 nm. The main purpose of having a positive dispersion value over the entire operating spectrum is to mitigate a nonlinear optical effect called four-wave mixing (FWM), which is analogous to intermodulation distortion in electrical systems. **Fig. 4** shows the dispersion characteristic of the G.655 fiber in the S-, C-, and L-bands.

Light Sources

The fiber by itself is not practical unless there is a compatible optical source for launching light signals into it. The most suitable device for this is a semiconductor light-emitting diode (LED) or a laser diode. In the 1960s, a great deal of effort took place to achieve laser action in pn-junction diodes. The early devices were GaAs and GaAsP lasers that operated at a temperature of 77 K, emitted at a wavelength around 850 nm, and had high lasing threshold current densities. To make devices that were more application-friendly by operating at room temperature, structures consisting of sandwiched layers of AlGaAs and GaAs were investigated. Finally, in 1970, researchers at Bell Laboratories and a team at the Ioffe Physical Institute in Leningrad made the first semiconductor diode lasers based on a layered AlGaAs/GaAs/AlGaAs structure that were able to emit continuous-wave light in the 850-nm region at room temperature.

Major improvements in laser diode performance and reliability followed this achievement during the next decade. In addition, around 1976, room-temperature laser diodes operating at longer wavelengths, in the 1100 to 1600-nm region, started to appear. Of particular interest were GaInAsP/InP-based laser diodes emitting in the 1310-nm and 1550-nm low-loss windows of optical fibers. The progressive development of ever-improving devices during the 1980s and 1990s, included single-frequency emission with narrow linewidths under continuous operation, low levels of chirp under direct modulation, high output power, and the ability to tune specially constructed laser diodes over a wavelength range of up to 30 nm.

Fielded Systems

The bit rate-distance product $B \times L$ where B is the transmission bit rate and L is the repeater spacing, measures the transmission capacity of optical fiber links. Since the inception of optical fiber communications in the mid-1970s, the link transmission capacity has experienced a tenfold increase every four years. Several major technology advances spurred this growth. Among the technology developments are laser diodes emitting over an extremely narrow spectral band, optical amplifiers, fibers with low losses and low dispersions, and the concepts of wavelength division multiplexing.

Some of the initial telephone-system field trials in the USA were carried out in 1977 by GTE in Los Angeles and by AT&T in Chicago. These transmission links were largely for the trunking of telephone lines, which are digital links consisting of

Table 2 Commonly used SONET and SDH transmission rates

<i>SONET level</i>	<i>Electrical level</i>	<i>Line rate (Mb/s)</i>	<i>SDH equivalent</i>	<i>Common rate name</i>
OC-1	STS-1	51.84	–	
OC-3	STS-3	155.52	STM-1	155 Mb/s
OC-12	STS-12	622.08	STM-4	622 Mb/s
OC-24	STS-24	1244.16	STM-8	
OC-48	STS-48	2488.32	STM-16	2.5 Gb/s
OC-96	STS-96	4976.64	STM-32	
OC-192	STS-192	9953.28	STM-64	10 Gb/s
OC-768	STS-768	9,813.12	STM-256	40 Gb/s

time-division-multiplexed 64-kb/s voice channels. Similar demonstrations were carried out in Europe and Japan. Applications ranged from 45 to 140 Mb/s with repeater spacings around 10 km.

With the advent of high-capacity fiber optic transmission lines in the 1980s, service providers established a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. These standards define a synchronous frame structure for sending multiplexed digital traffic over optical fiber trunk lines. The basic building block and first level of the SONET signal hierarchy is called the synchronous transport signal–level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an optical carrier–level N (OC- N) signal. Thus the OC- N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems, the fundamental building block is the 155.52-Mb/s synchronous transport module–level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM- N signal. Table 2 shows commonly used SDH and SONET signal levels.

The development of optical sources and photodetectors capable of operating at 1310 nm allowed a shift in the transmission wavelength from 850 nm to 1310 nm. This resulted in a substantial increase in the repeaterless transmission distance for long-haul telephone trunks, since optical fibers exhibit lower power loss and less signal dispersion at 1310 nm. Intercity applications first used multimode fibers, but in 1984 switched exclusively to single-mode fibers, which have a significantly larger bandwidth. Bit rates for long-haul links typically range between 155 and 622 Mb/s (OC-3 and OC-12), and in some cases up to 2.5 Gb/s (OC-48), over repeater spacings of 40 km. Both multimode and single-mode 1310-nm fibers are used in local area networks, where bit rates range from 10 Mb/s to 100 Mb/s over distances ranging from 500 m to tens of kilometers.

In the next step of system evolution, links operating in the low-loss window around 1550 nm attracted much attention for high-capacity, long-span terrestrial and undersea transmission links. These links routinely carry traffic at 2.5 Gb/s over 90-km repeaterless distances. By 1996, advances in high-quality lasers and receivers allowed single-wavelength transmission rates of 10 Gb/s (OC-192).

In 1989, the introduction of optical amplifiers gave a major boost to fiber transmission capacity. Although there are GaAlAs-based solid-state optical amplifiers for the first window and InGaAsP amplifiers for the second window, the most successful and widely used devices are erbium-doped fiber amplifiers (commonly called EDFAs) operating in the 1530 to 1560-nm range and Raman fiber amplifiers that are used for operation in the 1560 to 1600-nm region.

During the same time period, impressive demonstrations of long-distance high-capacity systems were made using optical soliton signals. A soliton is a nondispersive pulse that makes use of nonlinear dispersion properties in a fiber to cancel out chromatic dispersion effects. As an example, solitons at rates of 10 Gb/s have been sent over a 12200-km experimental link using optical amplifiers and special modulation techniques.

Entrance of Wavelength Division Multiplexing

The use of wavelength division multiplexing (WDM) offers a further boost in fiber transmission capacity. The basis of WDM is to use multiple sources operating at slightly different wavelengths to transmit several independent information streams over the same fiber. Although researchers started looking at WDM in the 1970s, during the ensuing years it generally turned out to be easier to implement higher-speed electronic and optical devices than to invoke the greater system complexity called for in WDM. However, a dramatic surge in WDM popularity started in the early 1990s, as electronic devices neared their modulation limit and high-speed equipment became increasingly complex.

Fig. 5 shows the concept of implementing many closely spaced wavelengths within a spectral band centered around 1552.524 nm. This scheme is referred to as dense WDM or DWDM. Conceptually, the DWDM scheme is the same as frequency division multiplexing (FDM) used in microwave radio and satellite systems. Just as in FDM, the wavelengths (or optical frequencies) in a DWDM link must be properly spaced to avoid interference between channels. In an optical system this interference may arise from the fact that the center wavelength of laser diode sources and the spectral operating characteristics of other optical components in the link may drift with temperature and time, thereby giving rise to the need for a guard band between wavelength channels.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the ITU developed DWDM standards that specify channel spacings in terms of frequency. The ITU-T Recommendation G.694.1, which is entitled ‘Dense Wavelength

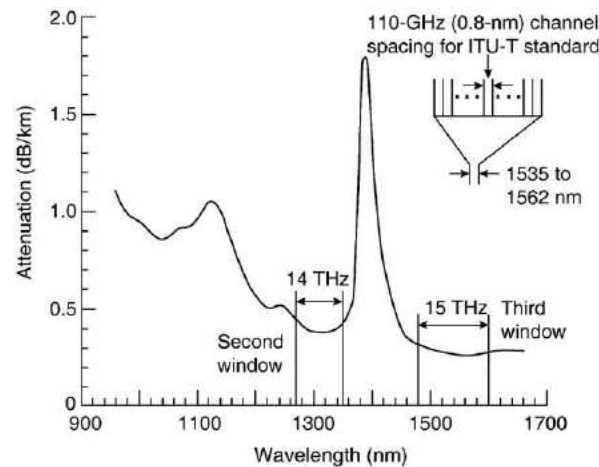


Fig. 5 Wavelength division multiplexing (WDM) concept.

Division Multiplexing (DWDM), specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm). This implementation requires the use of stable, high-quality, temperature-controlled and wavelength-controlled (frequency-locked) laser diode light sources.

With the production of full-spectrum (low-water-content) fibers, the development of relatively inexpensive optical sources, and the desire to have low-cost optical links operating in metro and local area networks, came the concept of coarse WDM (CWDM). In 2002, the ITU-T released a standard aimed specifically at CWDM. This is Recommendation G.694.2, which is entitled 'Coarse Wavelength Division Multiplexing (CWDM)'. The CWDM grid is made up of 18 wavelengths defined within the range 1270 nm to 1610 nm (O-through L-bands) spaced by 20 nm with wavelength-drift tolerances of ± 2 nm. This can be achieved with inexpensive light sources that are not temperature-controlled. The targeted transmission distance for CWDM is 50 km on single-mode fibers, such as those specified in ITU-T Recommendations G.652, G.653, and G.655.

Wavelength tunability of a source is an important property of WDM systems. Obviously it is not desirable or practical to maintain an inventory of dozens of lasers that emit at different wavelengths for WDM applications. The ideal tunable laser should be adjustable to emit at a specific wavelength across a broad spectral range. One such device is a distributed Bragg reflector (DBR) laser diode that can be tuned over a 10 to 20 nm spectral range. Work on perfecting such devices are still underway.

Starting in the mid-1990s, a combination of EDFAs and WDM was used to boost fiber information capacity to even higher levels and to increase the transmission distance. A major system consideration in these super-high capacity links is to ensure that there is appropriate link and equipment redundancy, so that alternate paths are available in case of disruptions in communications resulting from cable ruptures (for example, caused by errant digging from a backhoe) or equipment failures at an intermediate node. Such disruptions otherwise could have a devastating effect on a large group of users.

Undersea Optical Cable Systems

The first transoceanic fiber optic cable systems were installed in the Atlantic and Pacific Oceans in 1988 and 1989. Initially these systems operated at 280 Mb/s per fiber pair using 1310-nm lasers and single-mode fibers. The links consisted of a series of point-to-point optical fiber segments between electronic-based undersea regeneration points that were located nominally 60 km apart. Later the transmission capacity of these links was upgraded to 2.5 Gb/s and the regenerator spacing was increased to 100 km by converting the 1310-nm multiple-frequency light sources to 1550-nm single-frequency laser diodes. Later, the regenerator spacing was increased to 140 km.

Although these cable systems significantly improved the quality of the international telephone service, the optical-to-electrical conversion process at each regeneration point remained a capacity bottleneck. The introduction of erbium-doped optical fiber amplifiers (EDFA) eliminated this bottleneck from undersea lightwave systems by amplifying signals directly in the optical domain. Since EDFAs operate over a 30-nm wavelength band, they are well suited for use with undersea WDM links, which have provided a further capacity increase. These undersea optical amplifiers are typically spaced about 45 km apart.

One example of the many worldwide installations of optically amplified WDM networks is the SEA-ME-WE-3 Cable System. This undersea network runs from Germany to Singapore, connecting more than a dozen countries in between. Hence the name SEA-ME-WE, which refers to Southeast Asia (SEA), the Middle East (ME), and Western Europe (WE). The network has two pairs of undersea fibers with a capacity of eight STM-16 wavelengths per fiber.

See also: Basic Concepts of Optical Communication Systems

Further Reading

- Bergano, N.S., 1997. Undersea amplified lightwave system design. In: Kaminow, I.P., Koch, T.L. (Eds.), *Optical Fiber Telecommunications*, vol. IIIA. New York: Academic, pp. 302–335.
- Goralski, W.J., 2002. *SONET/SDH*, 3rd edn. New York: McGraw-Hill.
- Hecht, J., 1999. *City of Light*. New York: Oxford University Press.
- Joyner, C.H., 1997. Semiconductor laser growth and fabrication technology. In: Kaminow, I.P., Koch, T.L. (Eds.), *Optical Fiber Telecommunications*, vol. IIIB. New York: Academic, pp. 163–199.
- Kaiser, P., Keck, D.B., 1988. Fiber types and their status. In: Miller, S.E., Kaminow, I.P. (Eds.), *Optical Fiber Telecommunications II*. New York: Academic, pp. 29–54.
- Kao, K.C., Hockman, G.A., 1966. Dielectric-fibre surface waveguides for optical frequencies. In: *Proceedings IEE*, vol. 113. . pp. 1151–1158.
- Keiser, G., 2000. *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.
- Keiser, G., 2003. *Optical Communications Essentials*. New York: McGraw-Hill.
- Mollenauer, L.F., Gordon, J.P., Mamyshev, P.V., 1997. Solitons in high bit rate long-distance transmission. In: Kaminow, I.P., Koch, T.L. (Eds.), *Optical Fiber Telecommunications*, vol. IIIA. New York: Academic, pp. 373–460.
- Nyman, B., Farries, M., Si, C., 2001. Technology trends in dense WDM demultiplexers. *Optical Fiber Technology* 7 (4), 255–274.
- Ramaswami, R., Sivarajan, K.N., 2002. *Optical Networks*, 2nd edn. San Francisco, CA: Morgan Kaufmann.
- Snitzer, E., 1961. Cylindrical dielectric waveguide modes. *Journal Optical Society of America* 51, 491–498.
- Trischitta, P.R., Marra, W.C., 1998. Applying WDM technology to undersea cable networks. *IEEE Communication Magazine* 36 (2), 62–66.

Dispersion Management

AE Willner, Y-W Song, J McGeehan, and Z Pan, University of Southern California, Los Angeles, CA, USA
B Hoanca, University of Alaska Anchorage, Anchorage, AK, USA

© 2005 Elsevier Ltd. All rights reserved.

Glossary

Chromatic dispersion The time-domain pulse broadening in an optical fiber caused by the frequency dependence of the refractive index. This results in photons at different frequencies traveling at different speeds [ps/nm/km].

Chromatic dispersion slope Different wavelengths have different amounts of dispersion values in an optical fiber [ps/nm²/km].

Conventional single mode fiber (SMF) Fiber that transmits only a single optical mode by virtue of a very small core diameter relative to the cladding. Provides lowest loss in the 1,550 nm region and has zero dispersion at 1,300 nm.

Cross-phase modulation (XPM) A nonlinear Kerr effect in which a signal undergoes a nonlinear phase shift induced by a copropagating signal at a different wavelength in a WDM system.

Dispersion compensating fiber (DCF) Optical fiber that has both a large negative dispersion and dispersion slope around 1,550 nm.

Fiber Bragg gratings (FBGs) A small section of optical fiber in which there is a periodic change of refractive index along the core of the fiber. An FBG acts as a wavelength-selective mirror, reflecting only specific wavelengths (Bragg wavelengths) and passing others.

Four-wave mixing (FWM) A nonlinear Kerr effect in which two or more signal wavelengths interact to generate a new wavelength.

Kerr effect Nonlinear effect in which the refractive index of optical fiber varies as a function of the intensity of light within the fiber core.

Modulation The process of encoding digital data onto an optical signal so it can be transmitted through an optical network.

Nonlinear effects Describes the nonlinear responses of a dielectric to intense electromagnetic fields. One such effect is the intensity dependence of the refractive index of an optical fiber (Kerr effect).

Nonzero dispersion-shifted fiber (NZDSF) Optical fiber that has a small amount of dispersion in the 1,550 nm region in order to reduce deleterious nonlinear effects.

Polarization mode dispersion (PMD) Dispersion resulting from the fact that different light polarizations within the fiber core will travel at different speeds through the optical fiber [ps/km^{0.5}].

Self-phase modulation (SPM) A nonlinear Kerr effect in which a signal undergoes a self-induced phase shift during propagation in optical fiber.

Wavelength division multiplexing (WDM) Transmitting many different wavelengths down the same optical fiber at the same time in order to increase the amount of information that can be carried.

Introduction

Optical communications systems have grown explosively in terms of the capacity that can be transmitted over a single optical fiber. This trend has been fueled by two complementary techniques, those being the increase in data-rate-per-channel coupled with the increase in the total number of parallel wavelength channels. However, there are many considerations as to the total number of wavelength-division-multiplexed (WDM) channels that can be accommodated in a system, including cost, information spectral efficiency, nonlinear effects, and component wavelength selectivity.

Dispersion is one of the critical roadblocks to increasing the transmission capacity of optical fiber. The dispersive effect in an optical fiber has several ingredients, including intermodal dispersion in a multimode fiber, waveguide dispersion, material dispersion, and chromatic dispersion. In particular, chromatic dispersion is one of the critical effects in a single mode fiber (SMF), resulting in a temporal spreading of an optical bit as it propagates along the fiber. At data rates ≤ 2.5 Gbit/s, the effects of chromatic dispersion are not particularly troublesome. For data rates ≥ 10 Gbit/s, however, transmission can be tricky and the chromatic dispersion-induced degrading effects must be dealt with in some way, perhaps by compensation. Furthermore, the effects of chromatic dispersion rise quite rapidly as the bit rate increases – when the bit rate increases by a factor of four, the effects of chromatic dispersion increase by a factor of 16! This article will only deal with the management of chromatic dispersion in single mode fiber.

One of the critical limitations of optical fiber communications comes from chromatic dispersion, which results in a pulse broadening as it propagates along the fiber. This occurs as photons of different frequencies (created by the spreading effect of data modulation) travel at different speeds, due to the frequency-dependent refractive index of the fiber core. Compounding the problems caused by chromatic dispersion is the fact that as bit rates rise, chromatic dispersion effects rise quadratically with respect to the increase in the bit rate. One can eliminate these effects using fiber with zero chromatic dispersion, known as dispersion shifted fiber (DSF). However, with zero dispersion, all channels in a wavelength-division-multiplexed (WDM) system travel at the

same speed, in-phase, and a number of deleterious nonlinear effects such as cross-phase modulation (XPM) and four-wave mixing (FWM) result. Thus, in WDM systems, some amount of chromatic dispersion is necessary to keep channels out-of-phase, and as such chromatic dispersion compensation is required. Any real fiber link may also suffer from 'dispersion slope' effects, in which a slightly different dispersion value is produced in each WDM channel. This means that while one may be able to compensate one channel exactly, other channels may progressively accumulate increasing amounts of dispersion, which can severely limit the ultimate length of the optical link and the wavelength range that can be used in a WDM system. This article will address the concepts of chromatic dispersion and dispersion slope management followed by some examples highlighting the need for tunability to enable robust optical WDM systems in dynamic environments. Some dispersion monitoring techniques are then discussed and examples given.

Chromatic Dispersion in Optical Fiber Communication Systems

In any medium (other than vacuum) and in any waveguide structure (other than ideal infinite free space), different electromagnetic frequencies travel at different speeds. This is the essence of chromatic dispersion. As the real fiber-optic world is rather distant from the ideal concepts of both vacuum and infinite free space, dispersion will always be a concern when one is dealing with the propagation of electromagnetic radiation through fiber. The velocity in fiber of a single monochromatic wavelength is constant. However, data modulation causes a broadening of the spectrum of even the most monochromatic laser pulse. Thus, all modulated data have a nonzero spectral width which spans several wavelengths, and the different spectral components of modulated data travel at different speeds. In particular, for digital data intensity modulated on an optical carrier, chromatic dispersion leads to pulse broadening – which in turn leads to chromatic dispersion limiting the maximum data rate that can be transmitted through optical fiber (see Fig. 1).

Considering that the chromatic dispersion in optical fibers is due to the frequency-dependent nature of the propagation characteristics, for both the material (the refractive index of glass) and the waveguide structure, the speed of light of a particular wavelength λ will be expressed as follows, using a Taylor series expansion of the value of the refractive index as a function of the wavelength:

$$v(\lambda) = \frac{c_0}{n(\lambda)} = \frac{c_0}{n_0(\lambda_0) + \frac{\partial n}{\partial \lambda} \delta \lambda + \frac{\partial^2 n}{\partial \lambda^2} (\delta \lambda)^2} \quad (1)$$

Here, c_0 is the speed of light in vacuum, λ_0 is a reference wavelength, and the terms in $\partial n / \partial \lambda$ and $\partial^2 n / \partial \lambda^2$ are associated with the chromatic dispersion and the dispersion slope (i.e., the variation of the chromatic dispersion with wavelength), respectively. Transmission fiber has positive dispersion, i.e. longer wavelengths result in longer propagation delays. The units of chromatic dispersion are picoseconds per nanometer per kilometer, meaning that shorter time pulses, wider frequency spread due to data modulation, and longer fiber lengths will each contribute linearly to temporal dispersion. Higher data rates inherently have both shorter pulses and wider frequency spreads. Therefore, as network speed increases, the impact of chromatic dispersion rises precipitously as the square of the increase in data rate. The quadratic increase with the data rate is a result of two effects, each with a linear contribution. On one hand, a doubling of the data rate makes the spectrum twice as wide, doubling the effect of dispersion. On the other hand, the same doubling of the data rate makes the data pulses only half as long (hence twice as sensitive to dispersion). The combination of a wider signal spectrum and a shorter pulse width is what leads to the overall quadratic impact. Moreover, the data modulation format used can significantly affect the sensitivity of a system to chromatic dispersion. For example, the common nonreturn-to-zero (NRZ) data

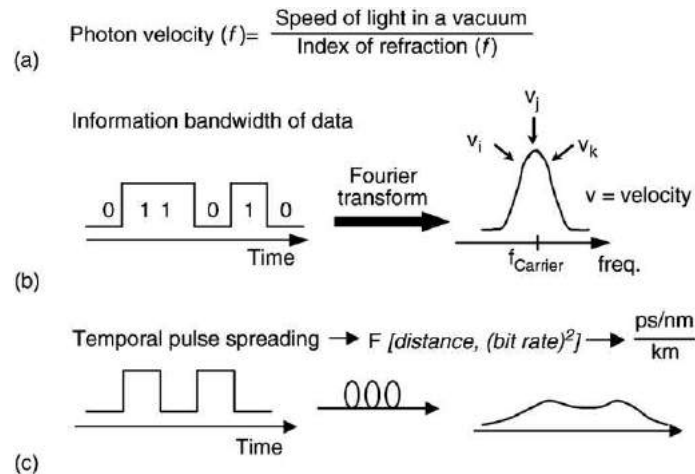


Fig. 1 The origin of chromatic dispersion in data transmission. (a) Chromatic dispersion is caused by the frequency-dependent refractive index in fiber. (b) The nonzero spectral width due to data modulation. (c) Dispersion leads to pulse broadening, proportional to the transmission distance and the data rate.

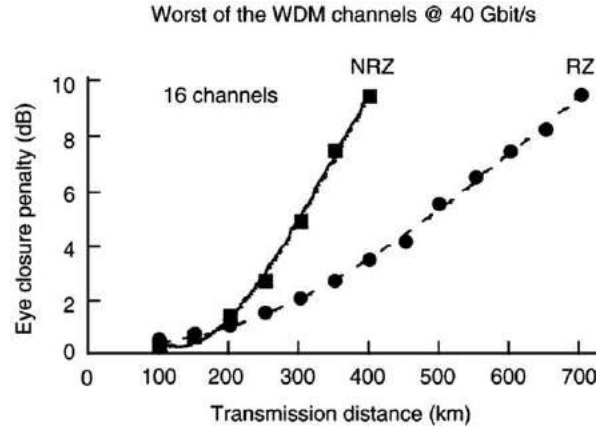


Fig. 2 Performances of RZ and NRZ formats in a real fiber transmission link. Reproduced with permission from Hayee I and Willner AE (1999) NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems. *IEEE Photon. Tech. Lett.* 11(8): 991–993.

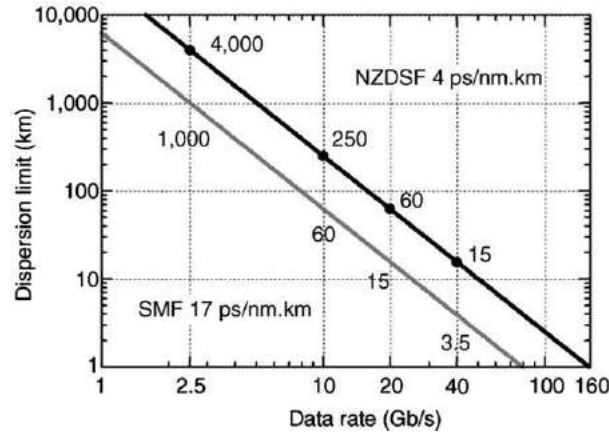


Fig. 3 Transmission distance limitations due to uncompensated dispersion in SMF as a function of data rate for intensity modulated optical signals. Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.

format, in which the optical power stays high throughout the entire time slot of a '1' bit, is more robust to chromatic dispersion than is the return-to-zero (RZ) format, in which the optical power stays high in only part of the time slot of a '1' bit. This difference is due to the fact that RZ data have a much wider channel frequency spectrum compared to NRZ data, thus incurring more chromatic dispersion. However, in a real WDM system, the RZ format increases the maximum allowable transmission distance by virtue of its reduced duty cycle (compared to the NRZ format), making it less susceptible to fiber nonlinearities as can be seen in Fig. 2.

A rule for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit period. For a bit period B , a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by

$$L_D = \frac{1}{D \cdot B \cdot \Delta\lambda} = \frac{1}{D \cdot B \cdot (cB)} \propto \frac{1}{B^2} \quad (2)$$

(see Fig. 3). For example, for single mode fiber, $D=17$ ps/nm/km, so for 10 Gbit/s data the distance is $L_D=52$ km. In fact, a more exact calculation shows that for 60 km, the dispersion induced power penalty is less than 1 dB (see Fig. 4). The power penalty for uncompensated dispersion rises exponentially with transmission distance, and thus to maintain good signal quality, dispersion compensation is required.

Chromatic Dispersion Management

Optical Nonlinearities as Factors to be Considered in Dispersion Compensation

Even though it is possible to manufacture fiber with zero dispersion, it is not practical to use such fiber for WDM transmission, due to large penalties induced by fiber nonlinearities. Most nonlinear effects originate from the nonlinear refractive index of fiber, which is not only dependent on the frequency of light but also on the intensity (optical power), and is related to the optical

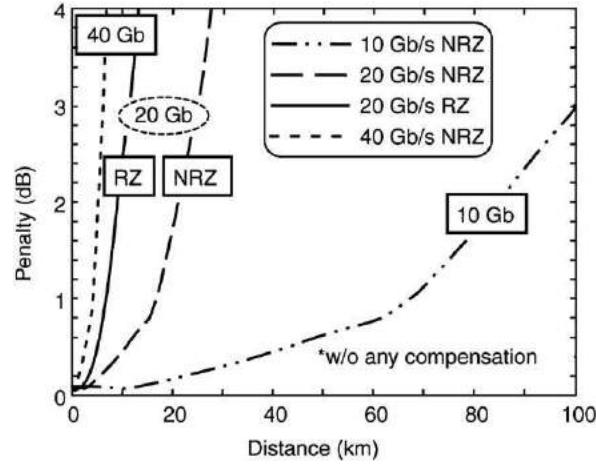


Fig. 4 Power penalties due to uncompensated dispersion in single mode fiber (SMF) as a function of distance and data rate. Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.

power as:

$$\bar{n}(f, P) = n(f) + n_2 \frac{P}{A_{\text{eff}}} \quad (3)$$

where $n(f)$ is the linear part of the refractive index, P is the optical power inside the fiber, and n_2 is the nonlinear-index coefficient for silica fibers. The typical value of n_2 is $2.6 \times 10^{-20} \text{ m}^2/\text{W}$. This number takes into account the averaging of the polarization states of the light as it travels in the fiber. The intensity dependence of the refractive index gives rise to three major nonlinear effects.

Self-phase modulation (SPM)

A million photons 'see' a different glass than does a single photon, and a photon traveling along with many other photons will slow down. SPM occurs because of the varying intensity profile of an optical pulse on a single WDM channel. This intensity profile causes a refractive index profile and, thus, a photon speed differential. The resulting phase change for light propagating in an optical fiber is expressed as:

$$\Phi_{\text{NL}} = \gamma P L_{\text{eff}} \quad (4)$$

where the quantities γ and L_{eff} are defined as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}} \quad \text{and} \quad L_{\text{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \quad (5)$$

where A_{eff} is the effective mode area of the fiber and α is the fiber attenuation loss. L_{eff} is the effective nonlinear length of the fiber that accounts for fiber loss, and γ is the nonlinear coefficient measured in $\text{rad}/\text{km}/\text{W}$. A typical range of values for γ is between 10–30 $\text{rad}/\text{km}/\text{W}$. Although the nonlinear coefficient is small, the long transmission lengths and high optical powers, that have been made possible by the use of optical amplifiers, can cause a large enough nonlinear phase change to play a significant role in state-of-the-art lightwave systems.

Cross-phase modulation (XPM)

When considering many WDM channels co-propagating in a fiber, photons from channels 2 through N can distort the index profile that is experienced by channel 1. The photons from the other channels 'chirp' the signal frequencies on channel 1, which will interact with fiber chromatic dispersion and cause temporal distortion. This effect is called cross-phase modulation. In a two-channel system, the frequency chirp in channel 1, due to power fluctuation within both channels, is given by

$$\Delta B = \frac{d\Phi_{\text{NL}}}{dt} = \gamma L_{\text{eff}} \frac{dP_1}{dt} + 2\gamma L_{\text{eff}} \frac{dP_2}{dt} \quad (6)$$

where, dP_1/dt and dP_2/dt are the time derivatives of the pulse powers of channels 1 and 2, respectively. The first term on the right-hand side of the above equation is due to SPM, and the second term is due to XPM. Note that the XPM-induced chirp term is double that of the SPM-induced chirp term. As such, XPM can impose a much greater limitation on WDM systems than can SPM, especially in systems with many WDM channels.

Four-wave-mixing (FWM)

The optical intensity propagating through the fiber is related to the electric field intensity squared. In a WDM system, the total electric field is the sum of the electric fields of each individual channel. When squaring the sum of different fields, products emerge that are beat terms at various sum and difference frequencies to the original signals. Fig. 5 depicts that if a WDM channel exists at

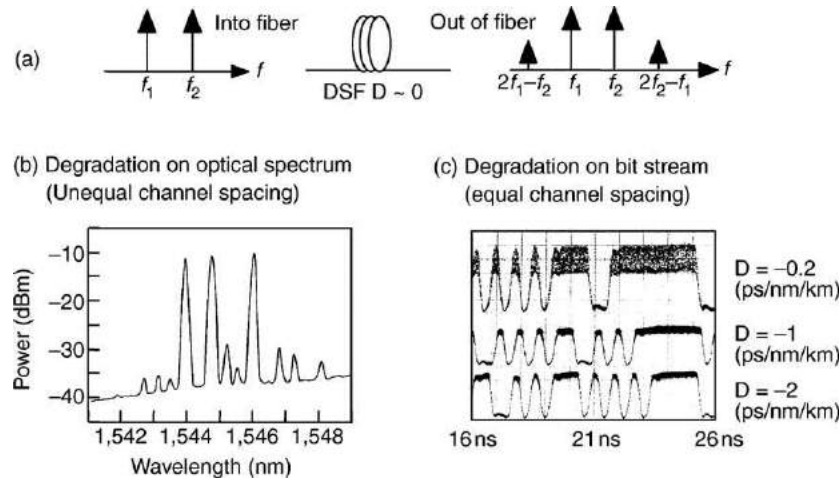


Fig. 5 (a) and (b) FWM induces new spectral components via nonlinear mixing of two wavelength signals. (c) The signal degradation due to FWM products falling on a third data channel can be reduced by even small amounts of dispersion. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnauck AH and Derosier RM (1995) Four-photon mixing and high-speed WDM systems. *Journal of Photon Technology* 13(5): 841–849.

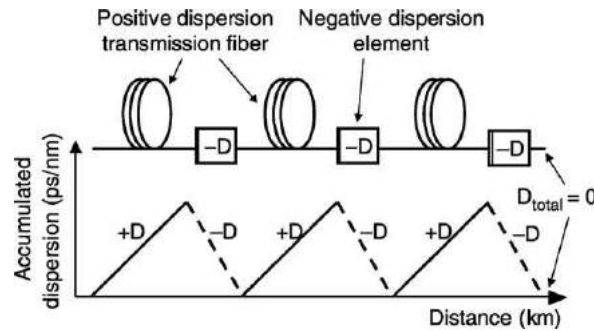


Fig. 6 Dispersion map of a basic dispersion managed system. Positive dispersion transmission fiber alternates with negative dispersion compensation elements such that the total dispersion is zero end-to-end.

Table 1 Commercially available fibers and their characteristics

	Dispersion @ 1,550 nm [ps/nm/km]	Dispersion slope [ps/nm ² /km]	Attenuation [dB]	Mode field diameter [μm]	PMD [ps/km ^{0.5}]
SMF	18	0.08	≤ 0.25	11	≤ 0.5
Tera Light	8.0	0.058	≤ 0.2	9.2	≤ 0.04
TW-RS	4.4	0.043	≤ 0.25	8.4	≤ 0.03
LEAF	4.0	0.085	≤ 0.25	9.6	≤ 0.08
Standard DCF	− 90	− 0.22	≤ 0.5	5.2	≤ 0.08

one of the four-wave-mixing beat-term frequencies, then the beat term will interfere coherently with this other WDM channel and potentially destroy the data.

Dispersion Maps

While zero-dispersion fiber is not a good idea, a large value of the accumulated dispersion at the end of a fiber link is also undesirable. An ideal solution is to have a ‘dispersion map,’ alternating sections of positive and negative dispersion as can be seen in [Fig. 6](#). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, eliminating FWM and XPM, but the total dispersion at the end of the fiber link is zero, so that no pulse broadening is induced ([Table 1](#)). The most advanced systems require periodic dispersion compensation, as well as pre- and post-compensation (before and after the transmission fiber).

The addition of negative dispersion to a standard fiber link has been traditionally known as ‘dispersion compensation,’ however, the term ‘dispersion management’ is more appropriate. SMF has positive dispersion, but some new varieties of nonzero

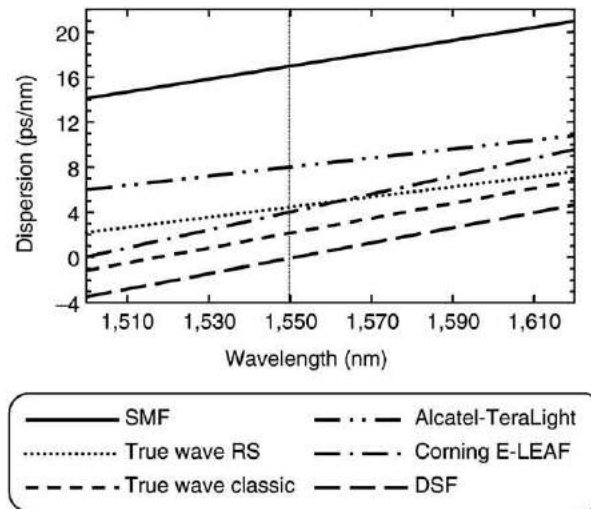


Fig. 7 Chromatic dispersion characteristics of various commercially available types of transmission fiber.

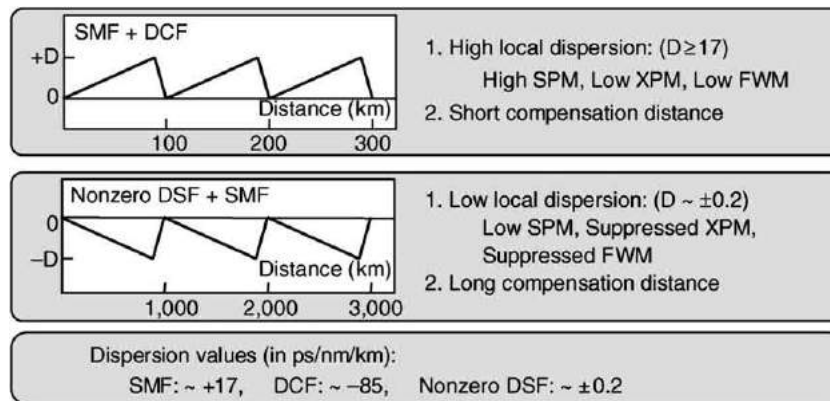


Fig. 8 Various dispersion maps for SMF-DCF and NZDSF-SMF.

dispersion-shifted fiber (NZDSF) come in both positive and negative dispersion varieties. Some examples are shown in [Fig. 7](#). Reverse dispersion fiber is also now available, with a large dispersion comparable to that of SMF, but with the opposite sign. When such flexibility is available in choosing both the magnitude and sign of the dispersion of the fiber in a link, dispersion-managed systems can be fully optimized to the desired dispersion map using a combination of fiber and dispersion compensation devices (see [Fig. 8](#)).

Dispersion is a linear process, so first-order dispersion maps can be understood as linear systems. However, the effects of nonlinearities cannot be ignored, especially in WDM systems, with many tens of channels, where the launch power may be very high. In particular, in systems deploying dispersion compensating fiber (DCF), the large nonlinear coefficient of the DCF can dramatically affect the dispersion map.

Corrections to Linear Dispersion Maps

Chromatic dispersion is a necessity in WDM systems, to minimize the effects of fiber nonlinearities. A chromatic dispersion value as small as a few ps/nm/km is usually sufficient to make XPM and FWM negligible. To mitigate the effects of nonlinearities but maintain small amounts of chromatic dispersion, NZDSF is commercially available. Due to these nonlinear effects, chromatic dispersion must be managed, rather than eliminated.

If a dispersion-management system was perfectly linear, it would be irrelevant whether the dispersion along a path is small or large, as long as the overall dispersion is compensated to zero (end to end). Thus, in a linear system the performance should be similar, regardless of whether the transmission fiber is SMF, and dispersion compensation modules are deployed every 60 km, or the transmission fiber is NZDSF (with approximately a quarter of the dispersion value of SMF) and dispersion compensation modules are deployed every 240 km. In real life, optical nonlinearities are very important, and recent results seem to favor the use of large, SMF-like, dispersion values in the transmission path and correspondingly high dispersion compensation devices. A recent

study of performance versus channel spacing showed that the capacity of SMF could be more than four times that of NZDSF. This is because the nonlinear coefficients are much higher in NZDSF than in SMF, and for dense WDM the channel interactions become a limiting factor. A critical conclusion is that not all dispersion compensation maps are created equal: a simple calculation of the dispersion compensation, to cancel the overall dispersion value, does not lead to optimal dispersion map designs.

Additionally, several solutions have been shown to be either resistant to dispersion, or have been shown to rely on dispersion itself for transmission. Such solutions include chirped pulses (where prechirping emphasizes the spectrum of the pulses so that dispersion does not broaden them too much), dispersion assisted transmission (where an initial phase modulation tailored to the transmission distance leads to full-scale amplitude modulation at the receiver end due to the dispersion), and various modulation formats robust to chromatic dispersion and nonlinearities.

Dispersion Management Solutions

Fixed Dispersion Compensation

From a systems point of view, there are several requirements for a dispersion compensating module: low loss, low optical nonlinearity, broadband (or multichannel) operation, small footprint, low weight, low power consumption, and clearly low cost. It is unfortunate that the first dispersion compensation modules, based on DCF only, met two of these requirements: broadband operation and low power consumption. On the other hand, several solutions have emerged that can complement or even replace these first-generation compensators.

Dispersion compensating fiber (DCF)

One of the first dispersion compensation techniques was to deploy specially designed sections of fiber with negative chromatic dispersion. The technology for DCF emerged in the 1980s and has developed dramatically since the advent of optical amplifiers in 1990. DCF is the most widely deployed dispersion compensator, providing broadband operation and stable dispersion characteristics, and the lack of a dynamic, tunable DCF solution has not reduced its popularity.

As can be seen in Fig. 9, the core of the average dispersion compensating fiber is much smaller than that of standard SMF, and beams with longer wavelengths experience relatively large changes in mode size (due to the waveguide structure) leading to greater propagation through the cladding of the fiber, where the speed of light is greater than that of the core. This leads to a large negative dispersion value. Additional cladding layers can lead to improved DCF designs that can include negative dispersion slope to counteract the positive dispersion slope of standard SMF.

In spite of its many advantages, DCF has a number of drawbacks. First, it is limited to a fixed compensation value. In addition, DCF has a weakly guiding structure and has a much smaller core cross-section, $19 \mu\text{m}^2$, compared to the $85 \mu\text{m}^2$ of SMF. This leads to higher nonlinearity, higher splice losses, as well as higher bending losses. Last, the length of DCF required to compensate for SMF dispersion is rather long, about one-fifth of the length of the transmission fiber for which it is compensating. Thus DCF modules induce loss, and are relatively bulky and heavy. The bulk is partly due to the mass of fiber, but also due to the resin used to hold the fiber securely in place. One other contribution to the size of the module is the higher bend loss associated with the refractive index profile of DCF; this limits the radius of the DCF loop to 6–8 inches, compared to the minimum bend radius of 2 inches for SMF.

Traditionally, DCF-based dispersion compensation modules are usually located at amplifier sites. This serves several purposes. First, amplifier sites offer relatively easy access to the fiber, without requiring any digging or unbraiding of the cable. Second, DCF

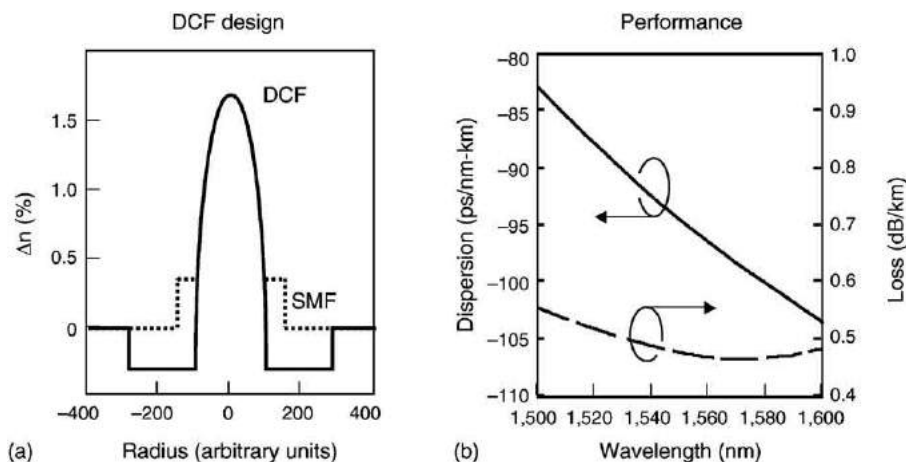


Fig. 9 Typical DCF (a) refractive index profile and (b) dispersion and loss as a function of wavelength. Δn is defined as refractive index variation relative to the cladding.

has high loss (usually at least double that of standard SMF), so a gain stage is required before the DCF module to avoid excessively low signal levels. DCF has a cross-section four times smaller than SMF, hence a higher nonlinearity, which limits the maximum launch power into a DCF module. The compromise is to place the DCF in the mid-section of a two-section EDFA. This way, the first stage provides pre-DCF gain, but not to a power level that would generate excessive nonlinear effects in the DCF. The second stage amplifies the dispersion compensated signal to a power level suitable for transmission through the fiber link. This launch power level is typically much higher than could be transmitted through DCF without generating large nonlinear effects. Many newer dispersion compensation devices have better performance than DCF, in particular lower loss and lower nonlinearities. For this reason, they may not have to be deployed at the mid-section of an amplifier. Fig. 10 shows the real demonstration results using the DCF.

Chirped fiber Bragg gratings

Fiber Bragg gratings have emerged as major components for dispersion compensation because of their low loss, small footprint, and low optical nonlinearity. Bragg gratings are sections of single-mode fiber in which the refractive index of the core is modulated in a periodic fashion, as a function of the spatial coordinate along the length of the fiber. When the spatial periodicity of the modulation matches what is known as a Bragg condition with respect to the wavelength of light propagating through the grating, the periodic structure acts like a mirror, reflecting the optical radiation that is traveling through the core of the fiber. An optical circulator is traditionally used to separate the reflected output beam from the input beam.

When the periodicity of the grating is varied along its length, the result is a chirped grating which can be used to compensate for chromatic dispersion. The chirp is understood as the rate of change of the spatial frequency as a function of position along the grating. In chirped gratings the Bragg matching condition for different wavelengths occurs at different positions along the grating length. Thus, the roundtrip delay of each wavelength can be tailored by designing the chirp profile appropriately. Fig. 11 compares the chirped FBG with uniform FBG. In a data pulse that has been distorted by dispersion, different frequency components arrive

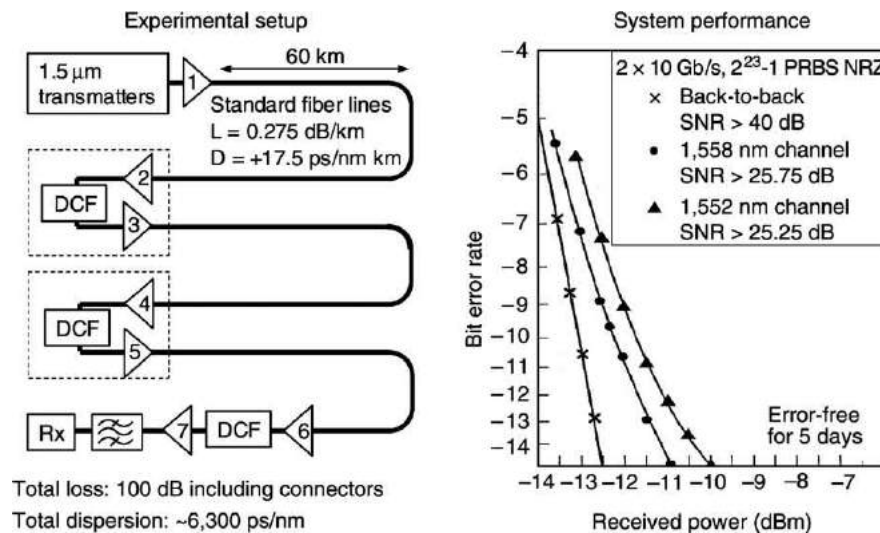


Fig. 10 System demonstration of dispersion compensation using DCF. Reproduced with permission from Park YK, Yeates PD, Delavaux J-MP, *et al.* (1995) A field demonstration of 20-Gb/s capacity transmission over 360 km of installed standard (non-DSF) fiber. *Photon. Technol. Lett.* 7 (7): 816–818.

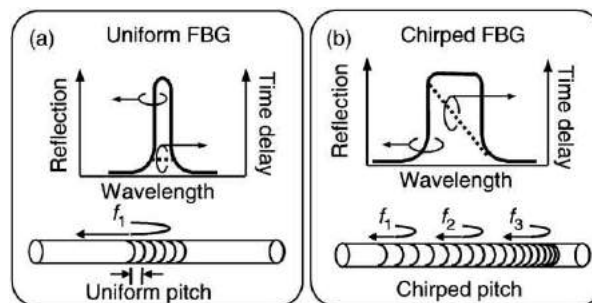


Fig. 11 Uniform and chirped FBGs. (a) A grating with uniform pitch has a narrow reflection spectrum and a flat time delay as a function of wavelength. (b) A chirped FBG has a wider bandwidth, a varying time delay and a longer grating length.

with different amounts of relative delay. By tailoring the chirp profile such that the frequency components see a relative delay which is the inverse of the delay of the transmission fiber, the pulse can be compressed back. The dispersion of the grating is the slope of the time delay as a function of wavelength, which is related to the chirp.

The main drawback of Bragg gratings is that the amplitude profile and the phase profile as a function of wavelength have some amount of ripple. Ideally, the amplitude profile of the grating should have a flat (or rounded) top in the passband, and the phase profile should be linear (for linearly chirped gratings) or polynomial (for nonlinearly chirped gratings). The grating ripple is the deviation from the ideal profile shape. Considerable effort has been expended on reducing the ripple. While early gratings were plagued by more than 100 ps of ripple, published results have shown vast improvement to values close to ± 3 ps.

Higher order mode dispersion compensation fiber

One of the challenges of designing standard DCF is that high negative dispersion is hard to achieve unless the cross-section of the fiber is small (which leads to high nonlinearity and high loss). One way to reduce both the loss and the nonlinearity is to use a higher-order mode (HOM) fiber (LP_{11} or LP_{02} near cutoff instead of the LP_{01} mode in the transmission fiber).

Such a device requires a good-quality mode converter between LP_{01} and LP_{02} to interface between the SMF and HOM fiber. HOM fiber has a dispersion per unit length greater than six times that of DCF. Thus, to compensate for a given transmission length in SMF, the length of HOM fiber required is only one sixth the length of DCF. Thus, even though losses and nonlinearity per unit length are larger for HOM fiber than for DCF, they are smaller overall, because of the shorter HOM fiber length. As an added bonus, the dispersion can be tuned slightly by changing the cutoff wavelength of LP_{02} (via temperature tuning). A soon to be released HOM fiber-based commercial dispersion compensation module is not tunable, but can fully compensate for dispersion slope.

Tunable Dispersion Compensation

The need for tunability

In a perfect world, all fiber links would have a known, discrete, and unchanging value of chromatic dispersion. Network operators would then deploy fixed dispersion compensators periodically along every fiber link to exactly match the fiber dispersion. Unfortunately, several vexing issues may necessitate that dispersion compensators are tunability, that they have the ability to adjust the amount of dispersion to match system requirements.

First, there is the most basic business issue of inventory management. Network operators typically do not know the exact length of a deployed fiber link nor its chromatic dispersion value. Moreover, fiber plants periodically undergo upgrades and maintenance, leaving new and nonexact lengths of fiber behind. Therefore, operators would need to keep in stock a large number of different compensator models, and even then the compensation would only be approximate. Second, we must consider the sheer difficulty of 40 Gbit/s signals. The tolerable threshold for accumulated dispersion for a 40 Gbit/s data channel is 16 times smaller than at 10 Gbit/s. If the compensation value does not exactly match the fiber to within a few percent of the required dispersion value, then the communication link will not work. Tunability is considered a key enabler for this bit rate (see Figs. 12 and 13). Third, the accumulated dispersion changes slightly with temperature, which begins to be an issue for 40 Gbit/s systems and 10 Gbit/s ultra long-haul systems. In fiber, the zero-dispersion wavelength changes with temperature at a typical rate of 0.03 nm/°C. It can be shown that a not-uncommon 50 °C variation along a 1,000 km 40 Gbit/s link can produce significant degradation (see Fig. 14). Fourth, we are experiencing the dawn of reconfigurable optical networking. In such systems, the network path, and therefore the accumulated fiber dispersion, can change. It is important to note that even if the fiber spans are compensated span-by-span, the pervasive use of compensation at the transmitter and receiver suggests that optimization and tunability based on path will still be needed.

Other issues that increase the need for tunability include: (i) laser and (de)mux wavelength drifts for which a data channel no longer resides on the flat-top portion of a filter, thereby producing a chirp on the signal that interacts with the fiber's chromatic dispersion; (ii) changes in signal power that change both the link's nonlinearity and the optimal system dispersion map; and (iii) small differences that exist in transmitter-induced signal chirp.

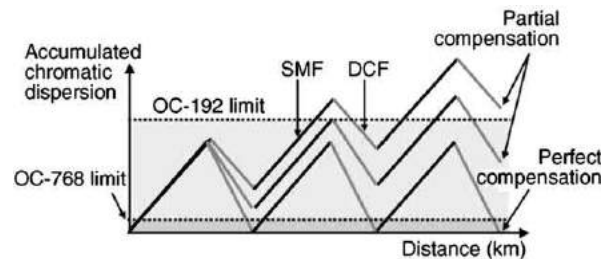


Fig. 12 The need for tunability. The tolerance of OC-768 systems to chromatic dispersion is 16 times lower than that of OC-192 systems. Approximate compensation by fixed in-line dispersion compensators for a single channel may lead to rapid accumulation of unacceptable levels of residual chromatic dispersion.

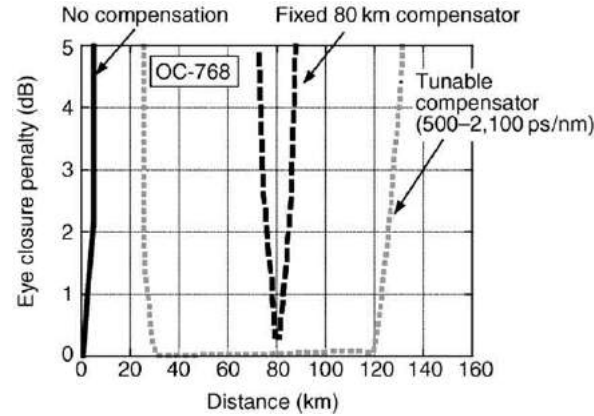


Fig. 13 Tunable dispersion compensation at OC-768 (40 Gb/s) is essential for achieving a comfortable range of acceptable transmission distances (80 km for tunable, only ~4 km for fixed compensation).

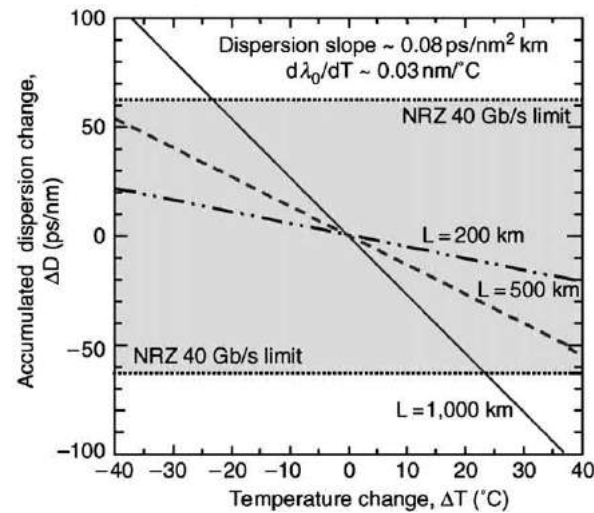


Fig. 14 Accumulated dispersion changes as a function of the link length and temperature fluctuation along the fiber link.

Approaches to tunable dispersion compensation

A host of techniques for tunable dispersion compensation have been proposed in recent years. Some of these ideas are just interesting research ideas, but several have strong potential to become viable technologies.

Fiber gratings offer the inherent advantages of fiber compatibility, low loss, and low cost. If a FBG has a refractive-index periodicity that varies nonlinearly along the length of the fiber, it will produce a time delay that also varies nonlinearly with wavelength (see Fig. 15). Herein lies the key to tunability. When a linearly chirped grating is stretched uniformly by a single mechanical element, the time delay curve is shifted towards longer wavelengths, but the slope of the ps-vs.-nm curve remains constant at all wavelengths within the passband. When a nonlinearly-chirped grating is stretched, the time delay curve is shifted toward longer wavelengths, but the slope of the ps-vs.-nm curve at a specific channel wavelength changes continuously. Ultimately, tunable dispersion compensators should accommodate multichannel operation. Several WDM channels can be accommodated by a single chirped FBG in one of two ways: fabricating a much longer (i.e., meters-length) grating, or using a sampling function when writing the grating, thereby creating many replicas of transfer function of the FBG in the wavelength domain (see Fig. 16).

One free-space-based tunable dispersion compensation device is the virtually imaged phased array (VIPA), based on the dispersion of a Fabry–Perot interferometer. The design requires several lenses, a movable mirror (for tunability), and a glass plate with a thin film layer of tapered reflectivity for good mode matching. Light incident on the glass plate undergoes several reflections inside the plate. As a result, the beam is imaged at several virtual locations, with a spatial distribution that is wavelength-dependent.

Several devices used for dispersion compensation can be integrated on a chip, using either an optical chip media (semiconductor-based laser or amplifier medium) or an electronic chip. One such technology is the micro-ring resonator, a device that, when used in a structure similar to that of an all-pass filter (see Fig. 17), can be used for dispersion compensation on a chip-scale. Although these technologies are not yet ready for deployment as dispersion compensators, they have been used in other applications and have the potential to offer very high performance at low cost.

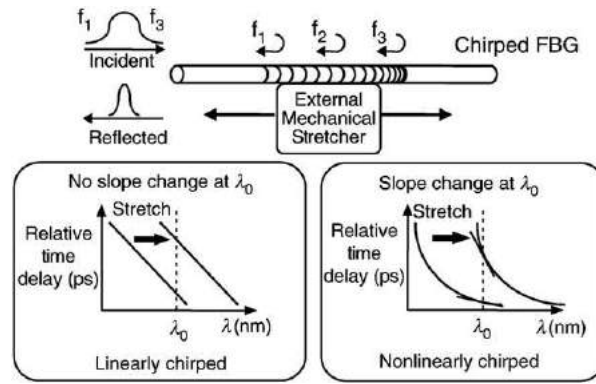


Fig. 15 Tuning results for both linearly and nonlinearly chirped FBGs using uniform stretching elements. The slope of the dispersion curve at a given wavelength λ_0 is constant when the linearly chirped grating is stretched, but changes as the nonlinearly chirped grating is stretched.

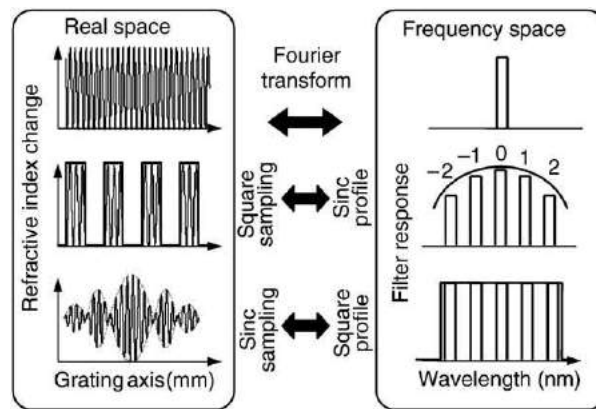


Fig. 16 The concept of 'sampled' FBGs, where a superstructure is written on top of the grating that produces a Fourier transform in the frequency domain, leading to multiple grating passbands.

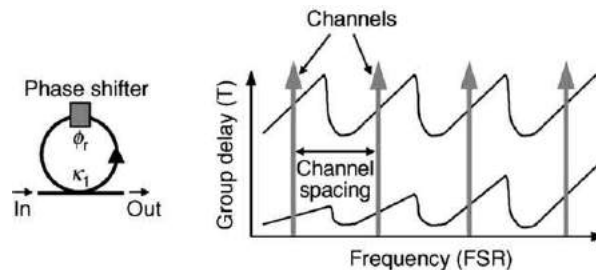


Fig. 17 Architecture of an all-pass filter structure for chromatic dispersion and slope compensation. Reproduced with permission from Madsen CK, Lenz G, Bruce AJ, *et al.* (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11 (12): 1623–1625.

As the ultimate optical dispersion compensation devices, photonic bandgap fibers (holey fibers) are an interesting class in themselves (see Fig. 18). These are fibers with a hollow structure, with holes engineered to achieve a particular functionality. Instead of being drawn from a solid preform, holey fibers are drawn from a group of capillary tubes fused together. This way, the dispersion, dispersion slope, the nonlinear coefficients could in principle all be precisely designed and controlled, up to very small or very large values, well outside the range of those of the solid fiber.

Dispersion Slope Mismatch

Transmission fiber, especially fiber with dispersion compensation built in, may suffer from a dispersion slope in which a slightly different dispersion value is produced for each WDM channel (see Fig. 19). Even though the compensator would be able to cancel

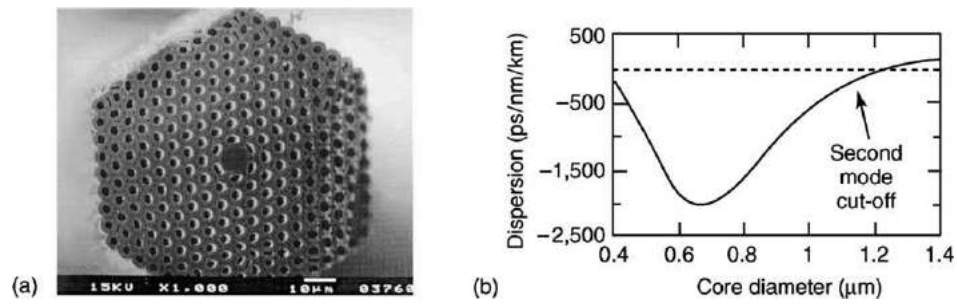


Fig. 18 (a) SEM image of a photonic crystal fiber (holey fiber), and (b) net dispersion of the fiber at 1550 nm as a function of the core diameter. Reproduced with permission from Birk TA, Mogilets D, Knight JC and Russell PST.J (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(6): 674–676.

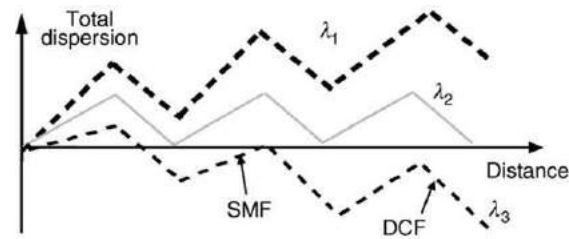


Fig. 19 The chromatic dispersion slope mismatch caused by the different slopes of transmission fiber (SMF or NZDSF) and DCF.

the dispersion of the fiber at the design wavelength, there will be residual dispersion left at the other wavelength channels unless the compensator can match the slope of the dispersion curve of the fiber as well. Some solutions for dispersion slope compensation are described in this section.

First, DCF, with negative dispersion slope, is a prime candidate for deployment as a dispersion slope compensator even though it cannot easily be made tunable. By designing the DCF with the same ratio of dispersion to dispersion slope as that of a real fiber link, new types of DCF can be used to compensate for both dispersion and dispersion slope, much like DCF is used for dispersion compensation today. DCF's popularity, wideband functionality, and stable dispersion characteristics make this a particularly attractive solution. Designing the DCF to match the dispersion characteristics of the transmission fiber is the critical engineering challenge in this slope compensation scheme. Second, third-order nonlinearly chirped FBG can act as a tunable dispersion slope compensator. A simple modification of the nonlinearly chirped FBG allows tuning of the compensated dispersion slope value via stretching the grating. The grating is prepared such that the time delay as a function of wavelength has a cubic profile that covers several WDM channels over a continuous bandwidth of many nanometers. Since the resulting dispersion curve is quadratic over the grating bandwidth, the dispersion slope experienced by the WDM channels can be tuned by stretching the grating using a single mechanical element (see Fig. 20). Third, combining the VIPA, with either a 3D mirror or diffraction grating, can also provide tunable free-space dispersion slope compensation. Slope tuning is achieved by dynamically controlling the MEMS-based 3D mirror or the diffraction grating. Fourthly, using an FBG with many spaced thin-film heater sections can also enable tunable dispersion slope. Each heater can be individually electrically controlled, allowing the time delay profile of the grating to be dynamically tuned via changing the temperature along the length of the grating. Advanced applications of this technique can allow alterations to the entire time delay profile of the grating, providing a truly flexible dispersion and dispersion slope compensation mechanism.

Chromatic Dispersion Monitoring

Another important issue related to dispersion management is dispersion monitoring techniques. In a reconfigurable system, it is necessary to reconfigure any tunable chromatic dispersion compensation modules on the fly as the network changes. An in-line chromatic dispersion monitor can quickly measure the required dispersion compensation value while data are still being transmitted through the optical link. This is very different from the more traditional chromatic dispersion measurement techniques where dark fiber is used and the measurement is done off-line over many hours (or days).

Chromatic dispersion monitoring is most often done at the receiving end, where the Q-factor of the received data or some other means is employed to assess the accumulated dispersion. Existing techniques that can monitor dispersion in-line, fast, and with relatively low cost include: (i) general performance monitoring using the bit-error rate (BER) or eye opening, but this approach cannot differentiate among different degrading effects; (ii) detecting the intensity modulation induced by phase modulation at the transmitter; (iii) extracting the bit-rate frequency component (clock) from photo-detected data and monitoring its RF power (see Fig. 21); (iv) inserting a subcarrier at the transmitter and subsequently monitoring the subcarrier power degradation;

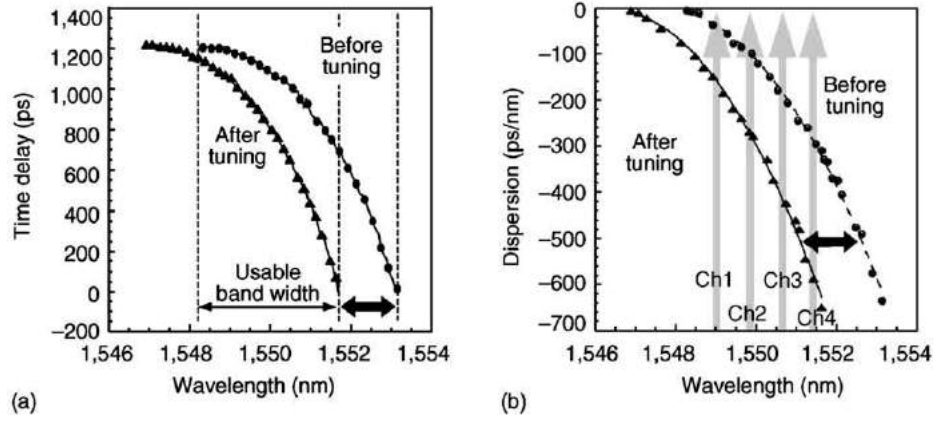


Fig. 20 Tunable dispersion slope compensation using a third-order nonlinearly chirped FBG. (a) Cubic time delay curves of the grating, and (b) quadratic dispersion curves showing the change in dispersion in each channel before and after tuning. Reproduced with permission from Song YW, Motoghian SMR, Starodubov D, *et al.* (2002) Tunable dispersion slope compensation for WDM systems using a non-channelized third-order-chirped FBG. *Optical Fiber Communication Conference 2002*. Paper ThAA4.

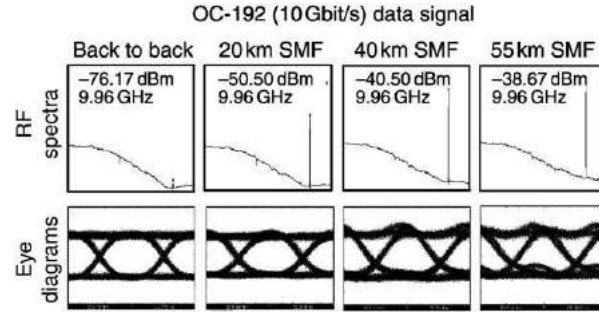
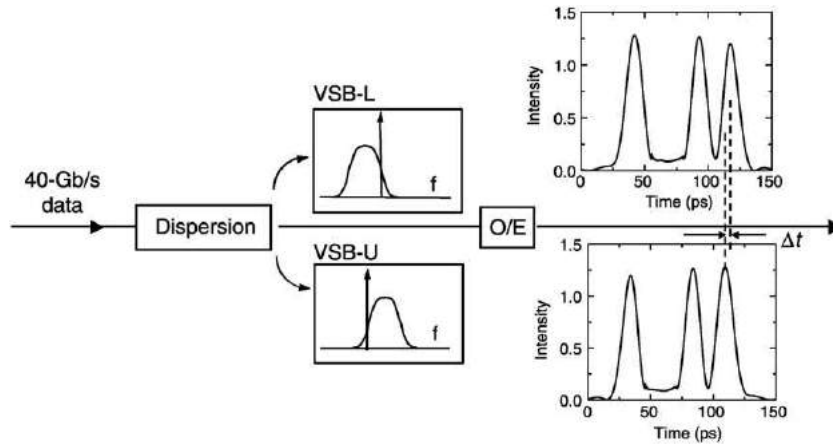


Fig. 21 Clock regenerating effect due to chromatic dispersion for NRZ data. As the amount of residual dispersion increases, so does the amount of power at the clock frequency. This power can be used to monitor the amount of uncompensated chromatic dispersion. Reproduced with permission from Pan Z, Yu Q, Xie Y, *et al.* (2001) Chromatic dispersion monitoring and automated compensation for NRZ and RZ data using clock regeneration and fading without adding signalling. *Optical Fiber Communication Conference 2001*. Paper WH5.



$$\text{Clock phase shift} = 2\pi \times R_b \times (\text{Time delay}) = f(\text{dispersion})$$

Fig. 22 Chromatic dispersion monitoring using the time delay (Δt) between two VSB signals, which is a function of chromatic dispersion. Reproduced with permission from Yu Q, Yan L-S, Pan Z and Willner AE (2002) Chromatic dispersion monitor for WDM systems using vestigial-sideband optical filtering. *Optical Fiber Communication Conference 2002*. Paper WE3.

(v) extracting several frequency components from the optical data using very narrow-band optical filters and detecting the optical phase; (vi) dithering the optical-carrier frequency at the transmitter and measuring the resultant phase modulation of the clock extracted at receiver with an additional phase-locked loop; and (vii) using an optical filter to select the upper and lower vestigial sideband (VSB) signals in transmitted optical data and determine the relative group delay caused by dispersion (see [Fig. 22](#)).

Conclusion

Chromatic dispersion is a phenomenon with profound implications for optical fiber communications systems. It has negative effects, broadening data pulses, but it also helps reduce the effects of fiber nonlinearities. For this reason, managing dispersion, rather than trying to eliminate it altogether, is the key. Fixed dispersion components are suitable for point-to-point OC-192 systems. Tunable dispersion components are essential for dispersion management in reconfigurable and OC-768 systems. These dispersion compensation elements must have low loss, low nonlinearity, and must be cost effective.

Although several technologies have emerged that meet some or all of the above requirements, no technology is a clear winner. The trend is towards tunable devices, or even actively self-tunable compensators, and such devices will allow system designers to cope with the shrinking system margins and with the emerging rapidly reconfigurable optical networks.

See also: Dispersion

Further Reading

- Agrawal, G.P., 1997. *Fiber-Optic Communication Systems*. Rochester, NY: John Wiley & Sons.
- Agrawal, G.P., 1997. *Nonlinear Fiber Optics*, 2nd edn. Academic Press.
- Gowar, J., 1993. *Optical Communication Systems*, 2nd edn. Prentice Hall.
- Kaminow, I.P., Koch, T.L., 1997. *Optical Fiber Telecommunications IIIA & IIIB*. Academic Press.
- Kashyap, R., 1999. *Fiber Bragg Gratings*. Academic Press.
- Kazovsky, L., Benedetto, S., Willner, A., 1996. *Optical Fiber Communication Systems*. Artech House.
- Willner, A.E., Hoanca, B., 2002. Fixed and tunable management of fiber chromatic dispersion. In: Kaminow, I.P., Li, T. (Eds.), *Optical Fiber Telecommunications IV*. Academic Press.
- Yariv, A., Yeh, P., 1984. *Optical Waves in Crystals*. John Wiley & Sons.

All-Optical Multiplexing/Demultiplexing

Z Ghassemloo, Northumbria University, Newcastle upon Tyne, UK
G Swift, Sheffield Hallam University, Sheffield, UK

© 2005 Elsevier Ltd. All rights reserved.

Introduction

A_i	i th data bit in a packet	P_{ds}	Data signal launch power
c	Speed of light in a vacuum	t_{asy}	The window width is given by $t_{asy} = 2\Delta x/c_f$
c_f	Speed of light within the fiber	T	Incoming bit interval
E_c	Electric fields of control signals	T_w	Walk off time per unit length between control and data pulses
E_{ds}	Electric fields of data signals	T_x	Transmittance of the NOLM
f_{FWM}	Frequency due to four-wave mixing	$\delta(\bullet)$	Optical carrier pulse shape
$G_{SLA}(t)$	Gain profiles of SLA	$\Delta\phi(t)$	Time-dependent phase difference
I_c	Intensity of the control signal	Δx	SLA offset from the fiber loop center
I_D	Intensity of the data signal	ϵ_0	Vacuum permittivity of the medium
$\bar{I}_m$	Average photo current for a mark	λ_s	Data signal wavelength
$\bar{I}_s$	Average photo current for a space	$\sigma_{amp,m}$	Variances for optical pre-amplifier for a mark
k	Packet length in bits	$\sigma_{amp,s}$	Variances for optical pre-amplifier for a space
L	Fiber length	$\sigma_{rec,m}$	Variances for receiver for a mark
L_d	Distance between two SLAs	$\sigma_{rec,s}$	Variances for receiver for a space
L_E	Fiber effective length	$\sigma_{RIN,m}$	Variances of RIN for a mark
n	Number of stages	$\sigma_{RIN,s}$	Variances of RIN for a space
n_2	Kerr coefficient	τ	Outgoing bit interval
N_{SLA}	SLA index	$\phi(t)$	Phase profile of SLA
P	Polarization	$\chi^{(3)}$	Third-order non-linear susceptibility of an optical fiber
P_c	Control signal launch power		

Introduction

The ever-increasing aggregate demand of electrically based time division multiplexing systems should have coped with the steady growth rate of voice traffic. However, since 1990, the explosive growth of the Internet and other bandwidth-demanding multimedia applications, has meant that long-haul telecommunication traffic has been increasingly dominated by data, not voice traffic. Such systems suffer from a bandwidth bottleneck due to speed limitations of the electronics. This limits the maximum data rate to considerably less than the THz bandwidth offered by an optical fiber. Optical technology is proposed as the only viable option and is expected to play an ever increasing role in future ultrahigh-speed links/networks. There are a number of multiplexing techniques, such as space division multiplexing (SDM), wavelength division multiplexing (WDM), and optical time division multiplexing (OTDM), that are currently being applied to effectively utilize the bandwidth of optical fiber as a means to overcome the bandwidth bottleneck imposed by electrical time division multiplexing (TDM). In SDM a separate optical fiber is allocated to each channel, but this is the least preferred option for increasing channel numbers. In WDM, a number of different data channels are allocated to discrete optical wavelengths for transmission over a single fiber. Dense WDM technology has been improving at a steady rate in recent years, with the latest systems capable of operating at a data rate of > 1 T bps, using a large number of wavelengths over a single fiber link. However, there are a number of problems associated with the WDM systems such as:

- Performance of WDM is highly dependent on the nonlinearities associated with fiber, i.e.:
 - *Stimulated Raman scattering*: degrades the signal-to-noise (SNR) as the number of channels increases;
 - *Four-wave mixing*: limits the channel spacing;
 - *Cross-phase modulation*: limits the number of channels.
- Relatively static optical paths, thus offering no fast switching with high performance within the network;
- Switching is normally carried out by separating each wavelength of each fiber onto different physical outputs. Space switches are then used to spatially switch the separated wavelengths, an extremely inefficient way of utilizing network resources;
- The need for amplifiers with high gain and flat spectra.

In order to overcome these problems, OTDM was introduced that offers the following:

- Flexible high bandwidth on demand (> 1 Tbit/s compared to the bit rates of 2.5–40 Gbit/s per wavelength channel in WDM systems);
- The total bandwidth offered by a single channel network is equal to DWDM;

- In a network environment, OTDM provides potential improvements in:
 - Network user access time, delay and throughput, depending on the user rates and statistics.
 - Less complex end node equipment (single-channel versus multichannels).
- Self-routing and self-clocking characteristics;
- Can operate at second- and third-transmission windows:
 - 1500 nm (like WDM) due to Erbium doped fiber amplifier (EDFA);
 - 1300 nm wavelengths.
- Offers both broadcast and switched based networks.

Principle of OTDM

Fig. 1 show the generic block diagram of a point-to-point OTDM transmission link, where N optical data channels, each of capacity M Gbps, are multiplexed to give an aggregate rate of $N \times M$ Gbps. The fundamental components are a pulsed light source, an optical modulator, a multiplexer, channel, add/drop unit, and a demultiplexer. The light source needs to have good stabilities and be capable of generating ultrashort pulses (< 1 ps). Direct modulation of the laser source is possible but the preferred method is based on external modulation where the optical signal is gated by the electronic data. The combination of these techniques allows the time division multiplexed data to be encoded inside a subnanosecond time slot, which is subsequently interleaved into a frame format. Add/drop units provide added versatility (see Fig. 2) allowing the 'adding' and 'dropping' of selected OTDM channels to interchange data at chosen points on the link. At the receiving end, the OTDM pulse stream is demultiplexed down to the individual channels at the initial M Gbps data rate. Data retrieval is then within the realm of electronic devices and the distinction between electronic and optical methods is no longer relevant. Demultiplexing requires high-speed all optical switching and can be achieved using a number of methods, which will be discussed in more detail below.

The optical multiplexing (or interleaving) can be carried out at the bit level (known as bit interleaving) or at the packet level (known as packet interleaving), where blocks of bits are interleaved sequentially. This is in accord with the popular conception of packet switched networks. The high data rates required and consequently narrow time slots necessitate the need for strict tolerances at processing nodes, e.g., switches. As such, it is important that the duration of the optical pulses is chosen to be significantly shorter than the bit period of the highest multiplexed line rate, in order to reduce the crosstalk between channels.

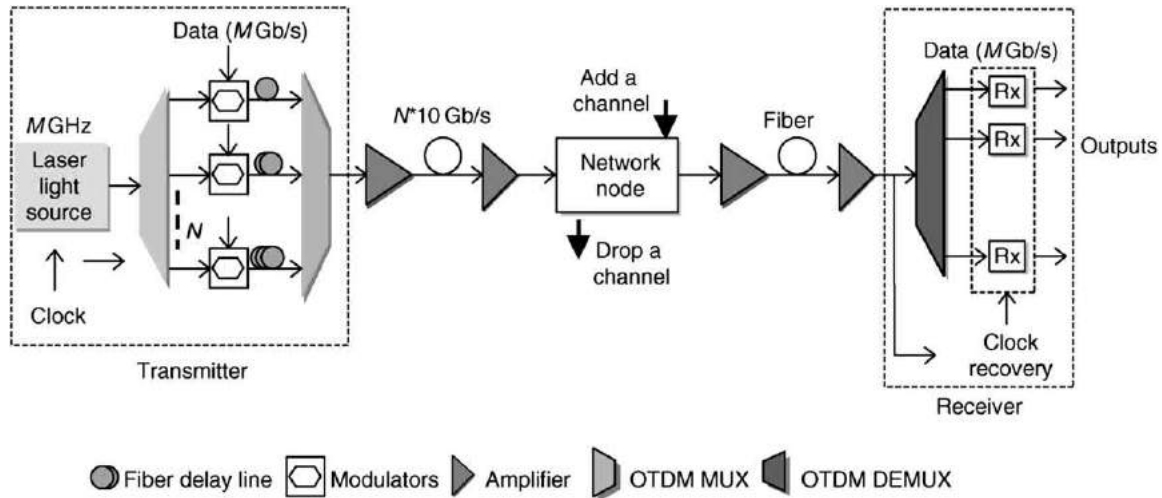


Fig. 1 Block diagram of a typical OTDM transmission system.

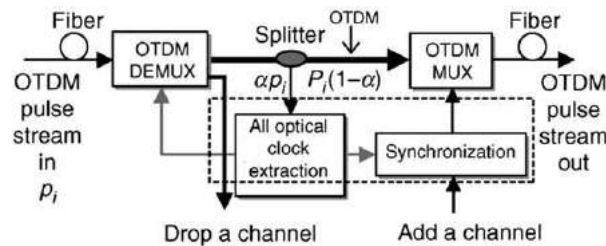


Fig. 2 Add/drop unit in an OTDM network node.

Bit Interleaved OTDM

A simple conceptual description of a bit interleaved multiplexer is shown in Fig. 3. It uses a number of different length optical fiber delay lines (FDL) to interleave the channels. The propagation delay of each FDL is chosen to position the optical channel in its corresponding time slot in relation to the aggregate OTDM signal. Prior to this, each optical pulse train is modulated by the data stream. The output of the modulators and an undelayed pulse train, labeled the framing signal, are combined, using a star coupler or combiner, to produce the high bit rate OTDM signal (see Fig. 3(b)). As shown in Fig. 3(b), the framing pulse has a higher intensity for clock recovery purpose. At the demultiplexer, the incoming OTDM pulse train is split into two paths (see Fig. 4(a)). The lower path is used to recover the framing pulse by means of thresholding, this is then delayed by an amount corresponding to the position of the i th (wanted) channel (see Fig. 4(b)). The delayed framing pulse and the OTDM pulse stream are then passed through an AND gate to recover the i th channel. The AND operation can be carried out all optically using, for example, a nonlinear loop mirror, a terahertz optical asymmetric demultiplexer, or a soliton-trapping gate.

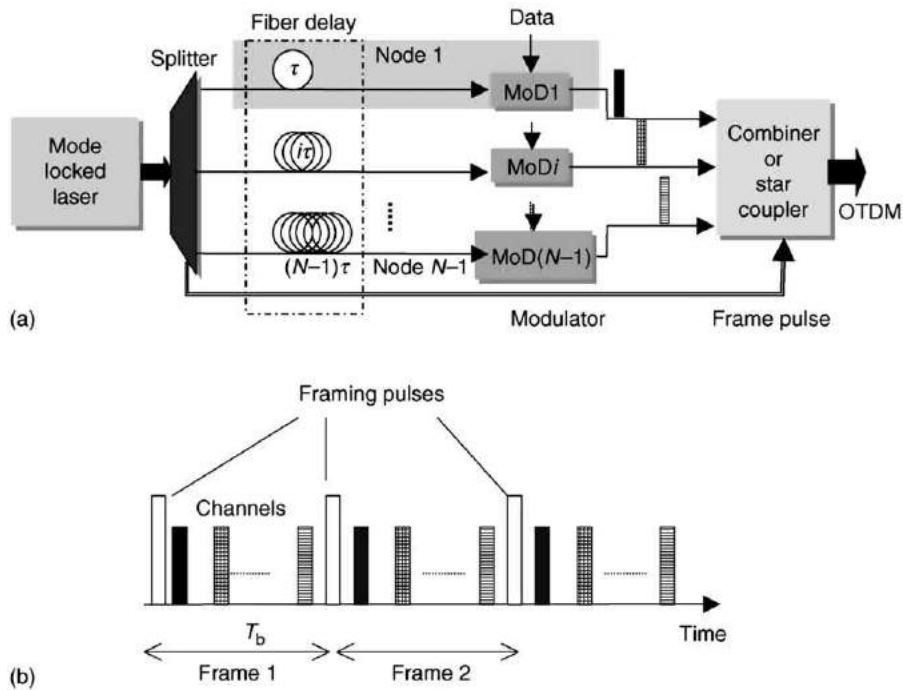


Fig. 3 Bit interleaved OTDM; (a) block diagram and (b) timing waveforms.

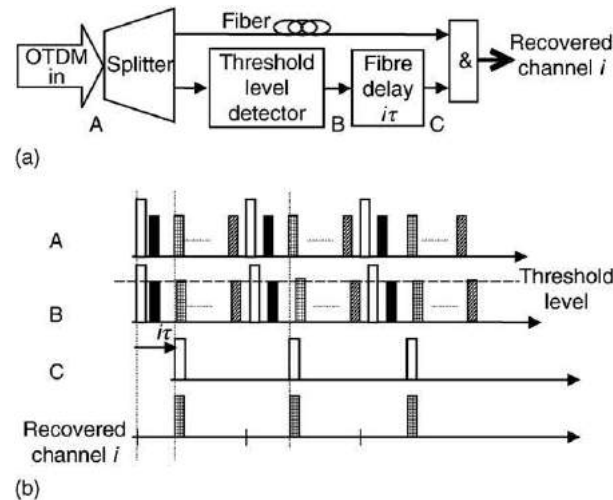


Fig. 4 Bit interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

Packet Interleaved OTDM

Fig. 5(a) shows the block diagram of a system used to demonstrate packet interleaving. Note that the time interval between successive pulses now needs to be much less than the bit interval T . This is achieved by passing the low bit rate output packet of the modulator into a compressor, which is based on a feed forward delay line structure. The feed forward delay lines are based on a cascade of passive M-Z interferometers with unequal arm lengths. This configuration generates the delay times required, i.e., $T - \tau, 2(T - \tau) \dots (2^{n-1})(T - \tau)$, etc., where $n = \log_2 k$ is the number of stages, T and τ are the incoming bit duration and the outgoing bit duration, respectively, and k is the packet length in bits. Pulse i 's location at the output is given by $(2^n - 1)(T - \tau) + (i - 1)\tau$. The signal at the input of the compressor is:

$$I_{in}(t) = \sum_{i=0}^{k-1} \delta(t - iT) A_i \quad (1)$$

Where $\delta(\cdot)$ is the optical carrier pulse shape, and A_i is the i th data bit in the packet.

As shown in Fig. 5(b), each bit is split, delayed, and combined to produce the four bit packets O_i :

$$O_i(t) = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} I_i[t - i(T - \tau)] \quad (2)$$

Note the factor of $1/(2^{n-1})$ is due to the signal splitting at the 2×2 3-dB coupler at the input of each stage.

The combined signal is shown in Fig. 5(b), and is defined as:

$$I_{out} = \sum_{i=0}^{k-1} O_i = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_i[t - (i+j)T + j\tau] A_i \quad (3)$$

For a packet delay of $(i+j)T$ packets are being built up. For the case $(i+j=3)$, four bits are compressed into a packet (see Fig. 5(b)). An optical gating device is used to select the only complete usable compressed copy of the incoming packets from the unwanted copies.

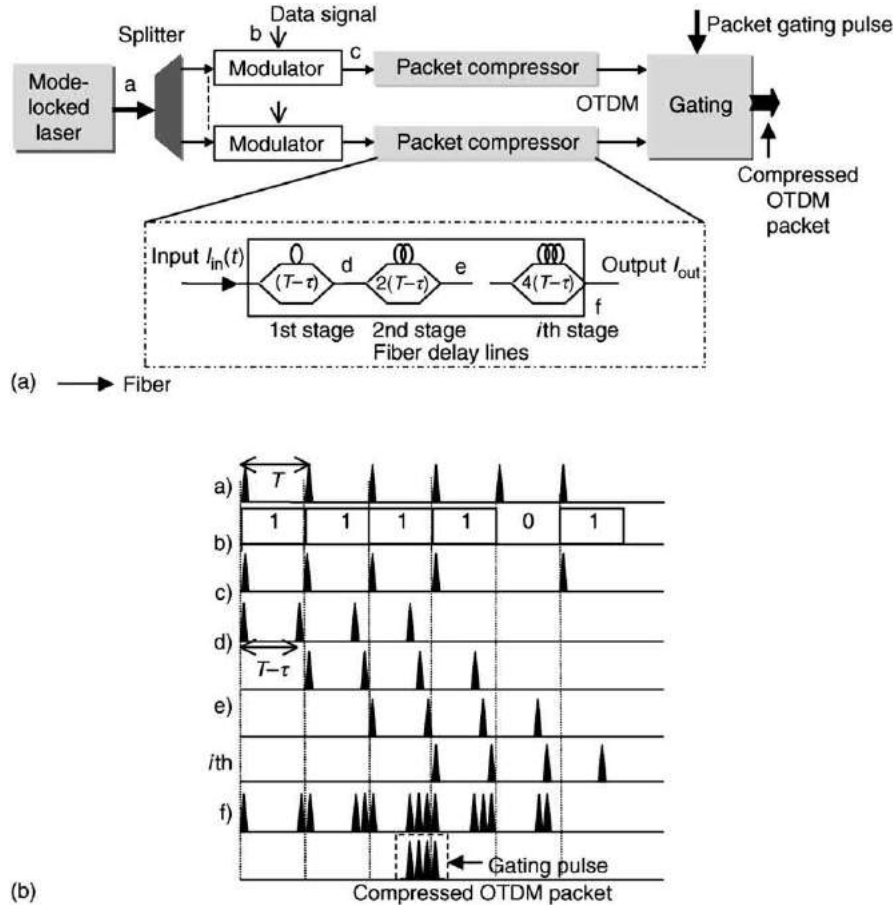


Fig. 5 Packet interleaving multiplexer: (a) block diagram and (b) typical waveforms.

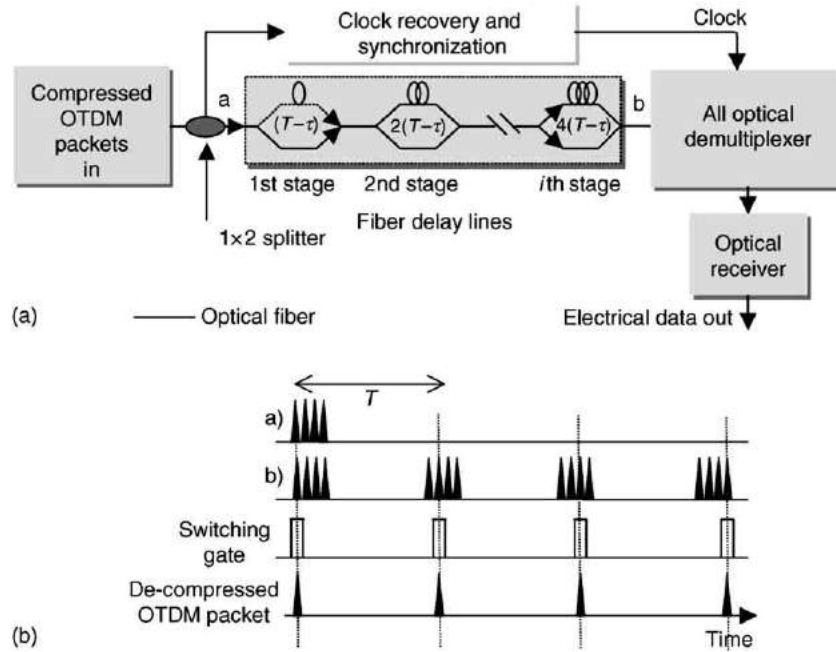


Fig. 6 Packet interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

At the receiving end, demultiplexing is carried out by decompressing the OTDM packet stream using the same delay line structure (see Fig. 6). Multiple copies of the compressed packet are generated, and each is delayed by $(T-\tau)$. An optical demultiplexer placed at the output of the delay lines generates a sequence of switching window at the rate of $1/T$. By positioning the switching pulses at the appropriate position, a series of adjacent packet (channel) bits from each of the copied compressed packets is extracted (see Fig. 6(b)). The output of the demultiplexer is the decompressed packet signal, which can now be processed at a much slower rate using electronic circuitry.

Components of an OTDM System

Optical Sources

In an ultrahigh speed OTDM system, it is essential that the optical sources are capable of generating transform-limited sub-nanosecond pulses having low duty cycles, tuneability, and a controllable repetition rate for synchronization. A number of suitable light sources are: gain-switch distribute feedback laser (DFB), active mode locked lasers (MLL) (capable of generating repetitive optical pulses), and harmonic mode-locking Erbium-doped fiber (EDF) lasers. Alternative techniques include super-continuum pulse generation in a dispersion shifted fiber with EDF pumping, adiabatic soliton compression using dispersion-flattened dispersion decreasing fiber, and pedestal reduction of compressed pulses using a dispersion-imbalanced nonlinear optical loop mirror. As shown in Fig. 1, the laser light source is power split to form the pulse source for each channel. These are subsequently modulated with an electrical data signal (external modulation). External modulation is preferred in an OTDM system as it can achieve narrow carrier linewidth, thus reducing the timing jitter of the transmitted pulse. For ultrahigh bit rate OTDM systems, the optical pulses emerging from the external modulators may also need compressing. One option is to frequency-chirp the pulses and pass them through an anomalous dispersive medium. Using this approach, it is essential that the frequency chirp be linear throughout the duration of the pulse, in addition the midpoint of the linear-frequency chirp should coincide with the center of the pulse. When a frequency-chirped pulse passes through a dispersive medium, different parts of the pulse travel at different speeds, due to a temporal variation of frequency. If the trailing edge travels faster than the leading edge, the result would be pulse compression. An optical fiber cable or a semiconductor laser amplifier (SLA) can be used to create the frequency chirp.

Multiplexers

Multiplexing of the pulses generated by the optical sources can be implemented either passively or actively. The former method is commonly implemented using a mono-mode optical fiber. This method has the advantage of being simple and cost-effective. The latter method uses devices more complex in nature, for example, electro-optic sampling switches, semiconductor optical amplifiers, and integrated optics.

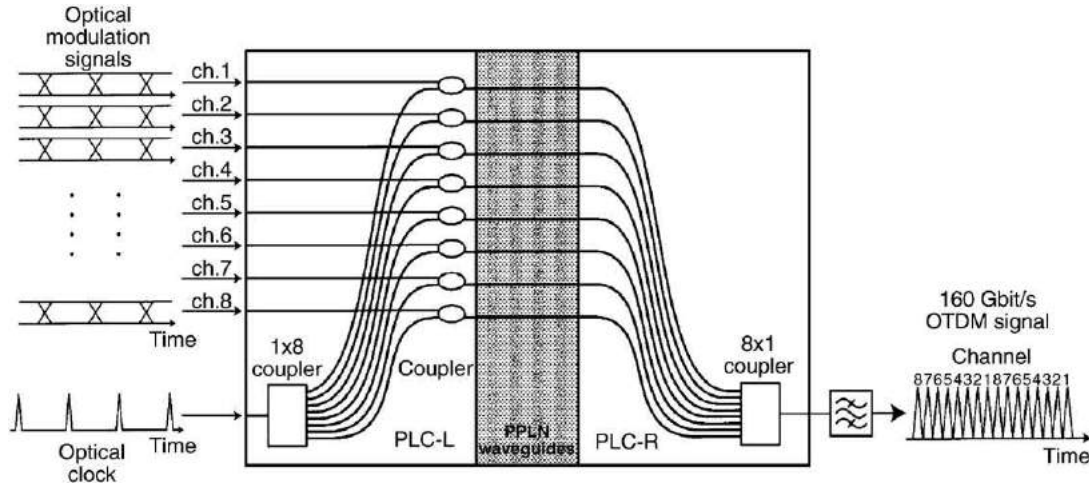


Fig. 7 Configuration of a PLC-OTDM-MUX. Reproduced with permission from Ohara T, Takara H and Shake I, *et al.* (2003). 160-Gb/s optical-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304. © 2003 IEEE.

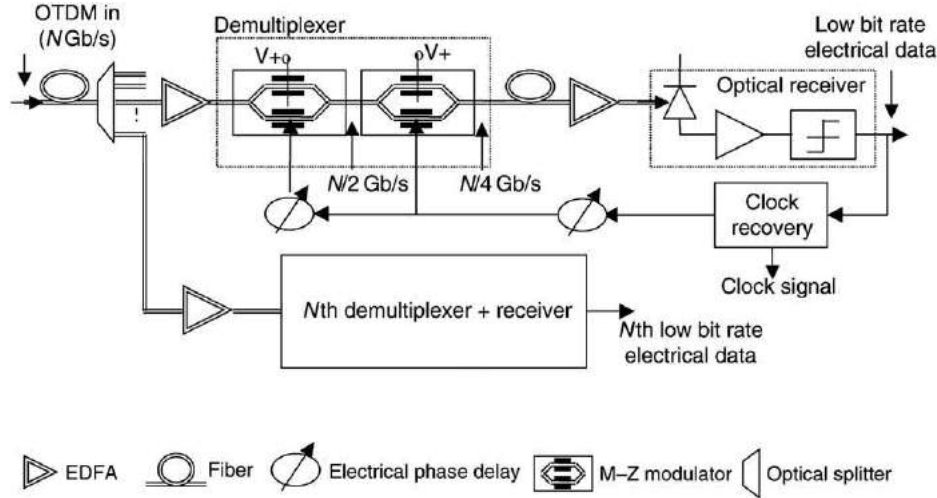


Fig. 8 Electro-optics demultiplexer and receiver system block diagram.

An integrated active multiplexer can be made by integrating semiconductor laser amplifiers (SOA) into a hybrid planar light-wave circuit (PLC). If the optical path length is kept short then temperature control is made easier. An alternative to this approach is to use lithium niobate (LiNb) technology. Periodically poled LiNb (PPLN) based OTDM multiplexers offer compact size and low noise (due to the absence of amplifier spontaneous emission noise and pattern effect, which is a feature of SOA based devices). **Fig. 7** is a schematic of a PLC-based OTDM multiplexer composed of two PLCs (R and L), composed of 1×8 and 8×1 couplers, eight 2×1 couplers, and eight different path length PPLN waveguides. The input clock pulse is split into eight by the 1×8 coupler in PLC-L, and are combined with the modulated optical pulse trains using the 2×1 couplers. The outputs of the 2×1 couplers are then passed through a PPLN waveguide to generate a return-to-zero (RZ) optically modulated signal. These are then combined, using the 8×1 coupler in PLC-R. When the path-length difference between the waveguides is set to one time slot (equal to $1/8$ of the base data rate), then the output of the 8×1 coupler is the required high-bit rate OTDM pulse.

Demultiplexers

In contrast to multiplexing, demultiplexing must be performed as an active function. It can be implemented electro-optically or optically. The former method needs to complete demultiplexing of all channels in order to extract a single channel (see **Fig. 8**). The demultiplexer in **Fig. 8** uses two LiNb Mach-Zehnder (M-Z) modulators in tandem. The first and second modulators are driven with a sinusoidal signal of amplitudes $2V_\pi$ and V_π , respectively, to down-convert the N -bit rate to $N/2$ and $N/4$, respectively. Channels can be selected by changing either the DC-bias $V+$ to the M-Zs, or the electrical phase delay. At ultrahigh speed

implementation of an electro-optics demultiplexer becomes increasingly difficult due to the higher drive voltage requirement by the M-Z. An alternative is to use all optical demultiplexing based on the nonlinear effect in a fiber and optical active devices offering switching resolution in the order of picoseconds. There are a number of methods available to implement all optical demultiplexing. The most popular methods that use fast phase modulation of an optical signal are based on M-Z and Sagnac interferometers. Four-wave mixing is another popular method.

Mach-Zehnder (M-Z) Interferometers

The key to interferometric switching is the selection of an appropriate material for the phase modulation. Semiconductor materials, often in the form of an SLA, are suitable devices for this purpose. The refractive index of an SLA is a function of the semiconductor carrier density, which can be modulated optically resulting in fast modulation. If a high-intensity pulse is input to an SLA the carrier density changes nonlinearly, as shown in Fig. 9. Phase modulation is affected via the material chirp index (dN/dn), which represents the gradient of the refractive index carrier density curve for the material. The phase modulation is quite strong and, in contrast to the intensity-based nonlinearity in an optical fiber, (see Kerr effect below), is a consequence of a resonant interaction between the optical signal and the material. The nonlinearity is relatively long-lived and would be expected to limit the switching speed when using semiconductors. Semiconductors, when in excitation from an optical field, tend to experience the effect quite quickly (ps) with a slow release (hundreds of picoseconds) time. Advantage can be taken of this property by allowing the slow recovery to occur during the time between channels of an OTDM signal, as this time may be of the order of hundreds of pico-seconds or more. A Mach-Zehnder configuration using two SLAs, one in each arm of the interferometer placed asymmetrically, is shown in Fig. 10. An optical control pulse entering the device via port 3 initiates the nonlinearity. The transmission equation relating the input signal (port 1) to the output port is given by

$$\frac{I_{out}(t)}{I_{in}(t)} = 0.25(G_{SLA1}(t) + G_{SLA2}(t) \pm 2\sqrt{G_{SLA1}(t)G_{SLA2}(t)}\cos\Delta\phi(t)) \quad (4)$$

where $G_{SLA1}(t)$ and $G_{SLA2}(t)$ refer to the gain profiles of the respective SLAs, and $\Delta\phi(t)$ is the time-dependent phase difference between them. Assuming that the gain and phase profiles of an excited amplifier are given by $G(t)$ and $\phi(t)$, respectively, then the

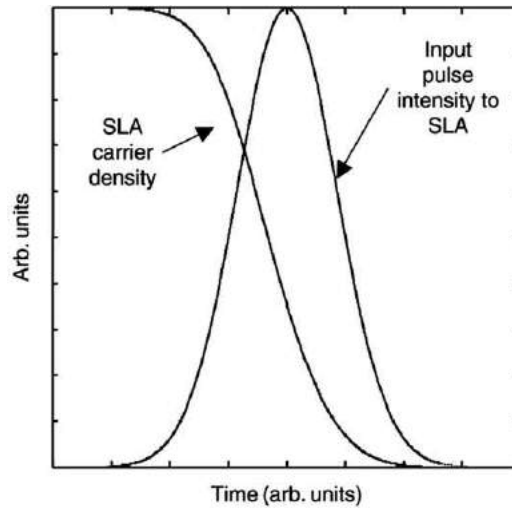


Fig. 9 SLA carrier density response to input pulse.

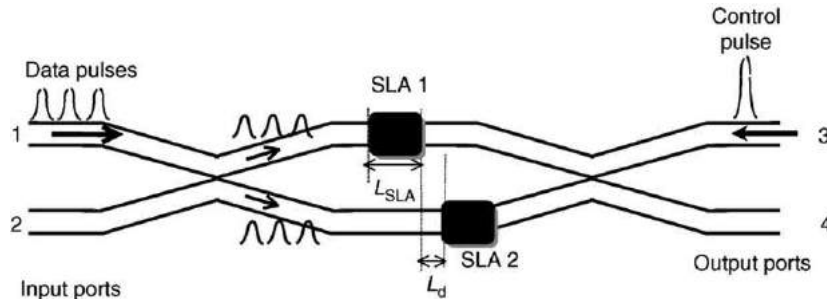


Fig. 10 Asymmetric TWSLA Mach-Zehnder devices.

signals passing through the upper and lower arms experience optical properties of the material given by:

$$G(t), G(t - T_d), \phi(t) \text{ and } \phi(t - T_d) \quad (5)$$

where T_d is given by $2L_d N_{SLA}/c$, L_d is the distance between SLA1 and SLA2, N_{SLA} is the SLA index, and c is the speed of light in a vacuum. As the width of the $\Delta\phi(t)$ profile is dependent on the distance between the SLAs, placing them in close proximity allows high-resolution switching. Less emphasis has been placed on the SLA gain as this is considered to be less effective when phase differences of π are reached. It only remains for the gain and phase modulation to recover, and as indicated previously, this is allowable over a time-scale commensurate with tributary rates.

Sagnac Interferometers

There are two main types; the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD).

NOLM

In this method of switching the inherent nonlinearity of an optical fiber known as the Kerr effect is used. The phase velocity of any light beam passing through the fiber will be affected by its own intensity and the intensity of any other beams present. When the intrinsic third-order nonlinearity of silica fibers is considered via the intensity-dependent nonlinear refraction component, then the signal phase shift is

$$\Delta\phi_{\text{signal}} = \frac{2\pi}{\lambda_{ds}} n_2 L I_{ds} + 2 \frac{2\pi}{\lambda_{ds}} n_2 L I_c \quad (6)$$

where n_2 is the Kerr coefficient, I_{ds} is the intensity of the data signal to be switched, I_c is the intensity of the control signal used to switch the data signal, λ_{ds} is the data signal wavelength, and L is the fiber length. The optical loop mirror consists of a long length of mono-mode fiber formed into a fiber coupler at its free ends (see Fig. 11). The input to the loop comprises the high-frequency data stream plus a control pulse at the frame rate. The data split at the coupler and propagate around the loop in contra directions (clockwise E_{CW} and counter-clockwise E_{CCW}) recombining back at the coupler. In the absence of a control pulse, the pulse exits via port 1. If a particular pulse in the loop (in this example E_{CW}) is straddled by the control pulse (see Fig. 12), then that pulse experiences cross-phase modulation, according to the second term on the right-hand side of Eq. (6), and undergoes a phase change relative to E_{CW} . The difference in the phase between E_{CW} and E_{CCW} causes the pulse to exit via port 2. The phase shift profile experienced by the co-propagating pulse is

$$\Delta\phi(t) = \frac{4\pi}{\lambda_{ds}} n_2 \int_0^L I_c(t) dx \quad (7)$$

Assuming unity gain around the loop, the transmittance of the NOLM is

$$T_x = 1 - \cos^2\left(\frac{\Delta\phi}{2}\right) = \Delta\phi(t) = 1 - \cos^2\left(\frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_c(t) dx\right) \quad (8)$$

As it stands, the switching resolution is determined by the width of the control pulse; however, it is possible to allow the signal and control to 'walk off' each other, allowing the window width to be increased by an amount determined by the 'walk off'. The phase shift is now

$$\Delta\phi(t) = 2 \frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_c(t - T_w x) dx \quad (9)$$

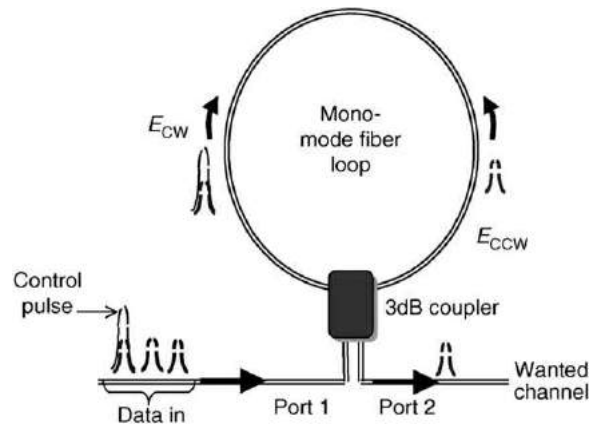


Fig. 11 Nonlinear optical loop mirror.

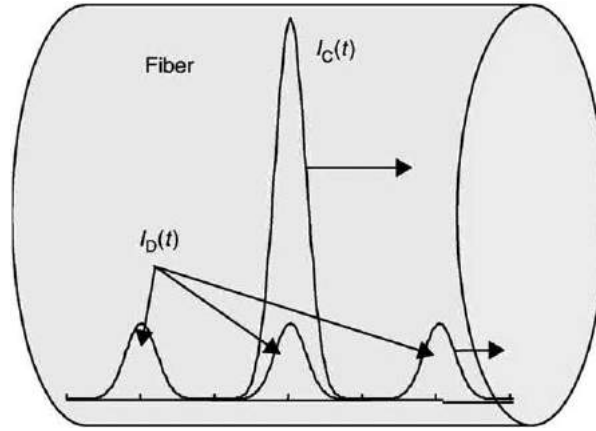


Fig. 12 Control and data pulses propagation in the fiber.

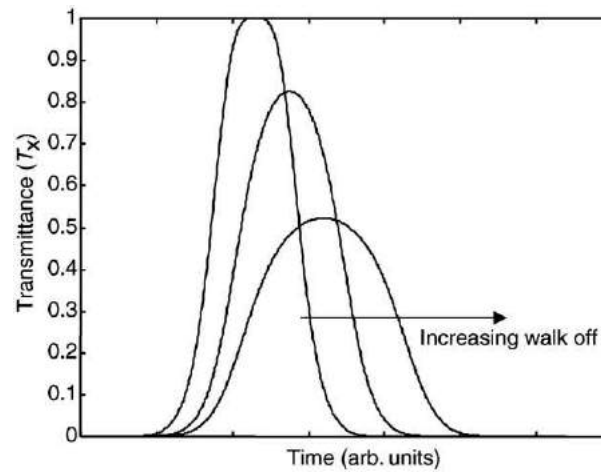


Fig. 13 Transmittance profiles with the walk-off time as a parameter.

where the parameter T_w is the walk off time per unit length between control and signal pulses. The increased window width is accompanied by a lower peak transmission (see Fig. 13).

TOAD

The TOAD uses a loop mirror architecture that incorporates an SLA and only needs a short length of fiber loop (see Fig. 14(a)). The SLA carrier density is modulated by a high-intensity control pulse, as in the M-Z-based demultiplexer. The operation of the TOAD is similar to the NOLM, where the loop transmittance is determined by the phase difference between CW and CCW traveling pulses. Strictly speaking, the gain of the SLA must also be taken into account; however, without any loss of generality the effect is adequately described by considering only the phase property. Fig. 14(a) shows the timing diagram associated with a TOAD demultiplexer. The control pulse, shown in Fig. 14(b1), is incident at the SLA at a time t_1 ; the phase profiles for the CCW and CW data pulses are shown in Figs. 14(b2) and (3), respectively. The resulting transmission window is shown in Fig. 14(b4). The window width is given by $t_{asy} = 2\Delta x/c_f$, where Δx is the SLA offset from the loop center and c_f is the speed of light in the fiber. As in NOLM, if the phase difference is of sufficient magnitude, then data can be switched to port 2. The switch definition is, in principle, determined by how close the SLA is placed to the loop center when the asymmetry is relatively large. However, for small asymmetries the switching window is asymmetric, which is due to the CW and CCW gain profiles being different (see Fig. 15). Assuming the phase modulation dominates the transmission then the normalized transmission for a small asymmetry loop is as depicted in Fig. 16. The gain response (not shown) would have a similar shape and temporal position as the phase response.

Four-Wave Mixing (FWM)

FWM demultiplexing uses a concept whereby two optical signals of different wavelengths are mixed together in a nonlinear medium to produce harmonic components. The nonlinearity in this case arises from the third-order nonlinear susceptibility $\chi^{(3)}$ of

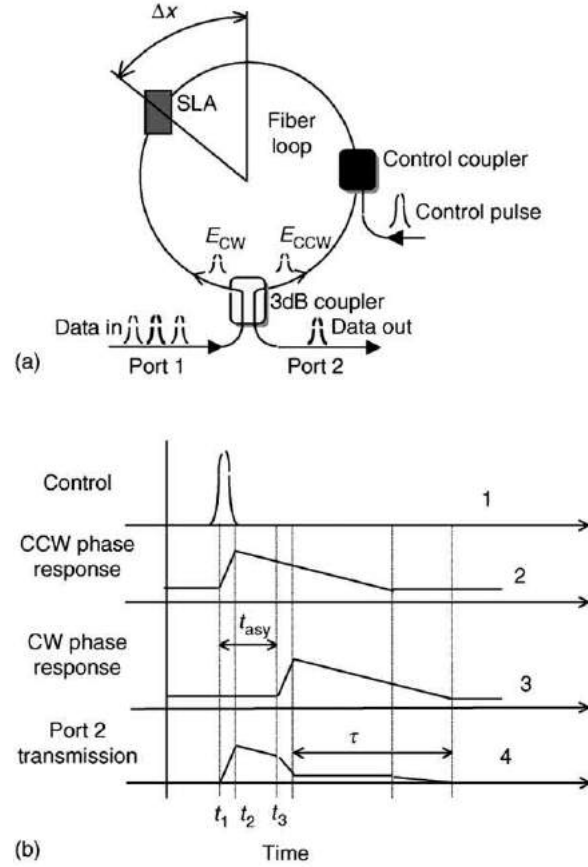


Fig. 14 TOAD: (a) architecture and (b) timing diagrams.

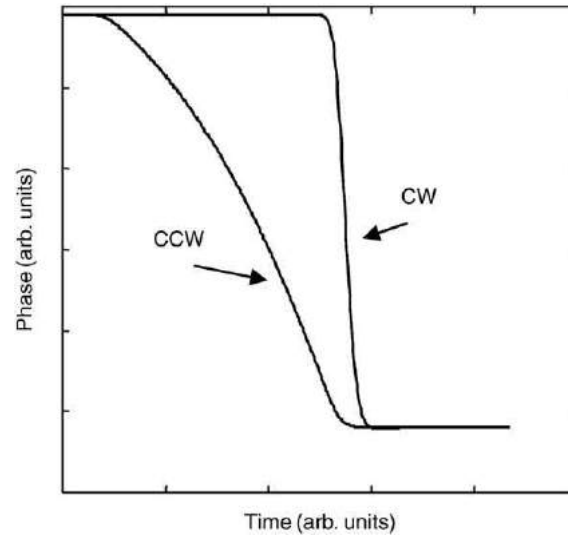


Fig. 15 Phase responses for CW and CCW components of TOAD.

an optical fiber, such that the polarization P induced on a pair of electric fields propagating through the fiber is

$$P = \epsilon_0 \chi^{(3)} (E_{ds} + E_c)^3 \quad (10)$$

where ϵ_0 is the vacuum permittivity of the medium, E_{ds} and E_c are the electric fields of data and control signals, respectively. The mixing of the two signals takes place in a long length of fiber in which the control signal propagates alongside the channel to be demultiplexed (see Fig. 17). The control signal is of sufficient intensity to cause the fiber to operate in the nonlinear regime.

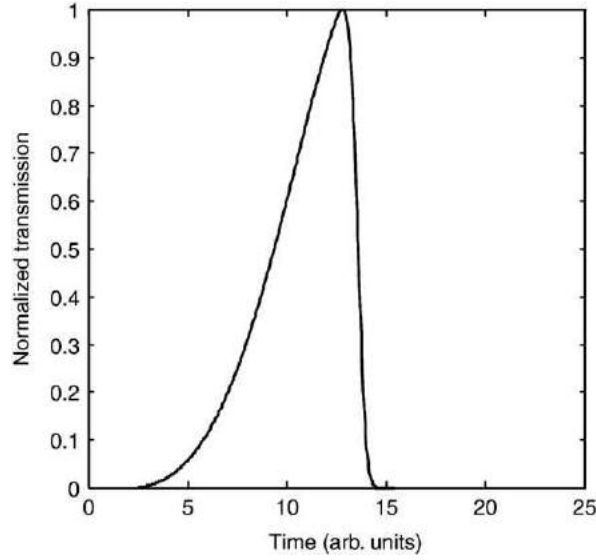


Fig. 16 TOAD transmission window profile for small asymmetry loop.

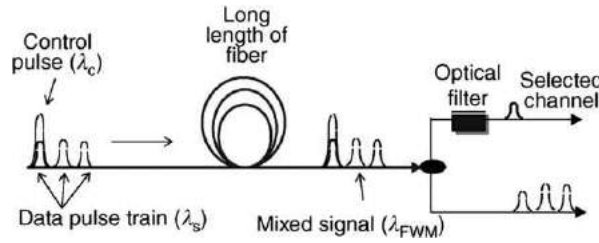


Fig. 17 Block diagram of FWM demultiplexer.

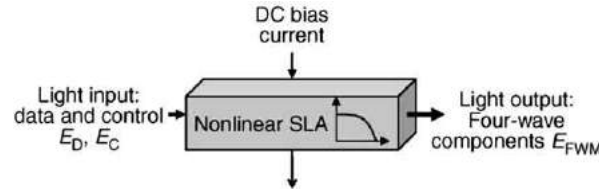


Fig. 18 Four-wave mixing in SOA.

The nonlinear relationship causes a number of frequencies to be generated, with the ones of interest having a frequency given by

$$f_{\text{FWM}} = 2f_c - f_{\text{ds}} \quad (11)$$

An optical filter is then used to demultiplex the required channel from the composite signal. The nonlinearity is detuned from the resonant frequency of the fiber glass and as such tends to be weak, requiring long lengths of fiber to give a measurable effect. The power in the demultiplexed signal, for given data and control signal wavelengths and fiber material, depends on the input power and the fiber length according to

$$P_{\text{FWM}} = kP_{\text{ds}}P_cL_E^2 \quad (12)$$

where k is a constant, P_{ds} is the signal launch power, P_c is the control signal launch power, and L_E is the fiber effective length. FWM is essentially an inefficient method as power is wasted in the unused frequency components of the four-wave signal. More in line with integrated structures, the nonlinear properties of a semiconductor laser amplifier can be used. Here a relatively high-power optical signal is input to the SLA (see Fig. 18). This enables saturation and operation in the nonlinear regime (see Fig. 19). Operating the SLA in saturation allows the nonlinear effects to produce the FWM components as in the fiber method.

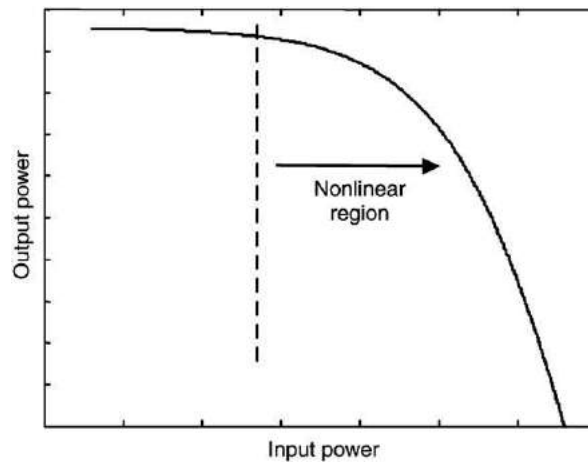


Fig. 19 SLA output-input power curve.

Clock and Data Synchronization in OTDM

In common with electronically based TDM systems, clock recovery is fundamental to the recovery of data in ultrahigh-speed OTDM systems. Two main methods are proposed: (i) clock signal transmitted with the OTDM signal (i.e., multiplexed); and (ii) extraction of clock signal from the incoming OTDM pulse stream. In a packet-based system, synchronization between the clock and the data packet is normally achieved by sending a synch pulse with each packet. However, techniques based on optical-phased locked loops are proving popular and remove the need for a separate clocking signal.

Clock Multiplexing

- (i) *Space division multiplexing*: This is conceptually the simplest to implement, where the clock signal is carried on a separate fiber from the data. However, it is susceptible to any differential delay between the different paths taken by clock and data due to temperature variation. It is difficult to justify in systems where the installed fiber base is at a premium.
- (ii) *Wavelength division multiplexing*: Here different wavelengths are allocated to the clock, and payload. It is only really practical for predetermined path lengths between nodes in single-hop networks such as point-point links or broadcast-and-select star networks. It also suffers from random delays between the clock and the payload, which is problematic in an asynchronous packet switched based network, where the optical path length a packet may take is nondeterministic.
- (iii) *Orthogonal polarization*: This is suitable for small links, where separate polarizations are used for the clock and data. However, in large networks it is quite difficult to maintain the polarization throughout the transmission link due to polarization mode dispersion and other nonlinear effects.
- (iv) *Intensity division multiplexing*: This uses higher-intensity optical clock pulses to differentiate it from the data pulses as discussed above. However, in long-distance transmission links, it is difficult to maintain both the clock intensity and its position, due to the fiber nonlinearity.
- (v) *Time division multiplexing*: In this scheme a single clock pulse, which has the same wavelength, polarization, and amplitude as the payload pulses, is separated in time, usually ahead of the data pulses.

Synchronization – Optical Phased Locked Loops (PLL)

The PLL is a common technique used for clock recovery in electronic TDM systems. However, the speed of conventional electronic PLLs tends to be limited by the response of the phase comparators used. There are a number of approaches based on opto-electronic PLL. Opto-electronic PLLs based on four-wave mixing in a traveling wave laser amplifier are complex and can suffer from frequency modulation in the recovered clock. However, others based on balanced photo-detectors, result in low timing jitter and good phase stability. In contrast, a number of all optical methods clock recovery scheme exist, one technique based on the TOAD (see above) is depicted in Fig. 20. The high-speed data stream enters the TOAD at a rate $n \times R$ where n is the number of channels per OTDM frame, and R is the frame rate. A pulse generator, such as a mode locked fiber laser (MLFL) clocked by a local oscillator (LO), is used as the TOAD control input (MLFL-C). The OTDM data is switched by the TOAD at a frequency of say, $R + \Delta f$ Hz. Thus, the switching window samples data pulses at a rate higher or lower than $n \times R$ Hz and uses this signal for cross correlation in the PLL unit. The output of the phase comparator is used to regulate a voltage controlled oscillator (VCO) running at R Hz, which in turn feeds the control signal. The PLL circuit locks into the clock frequency, generating the clock signal $S_c(t)$.

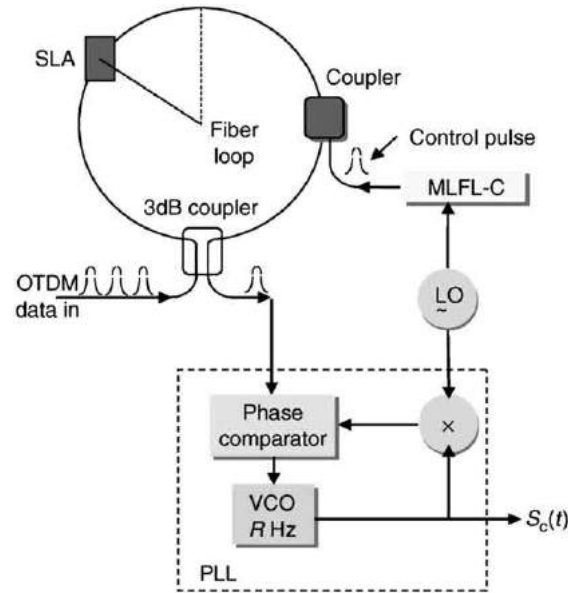


Fig. 20 Ultrafast clock recovery using TOAD.

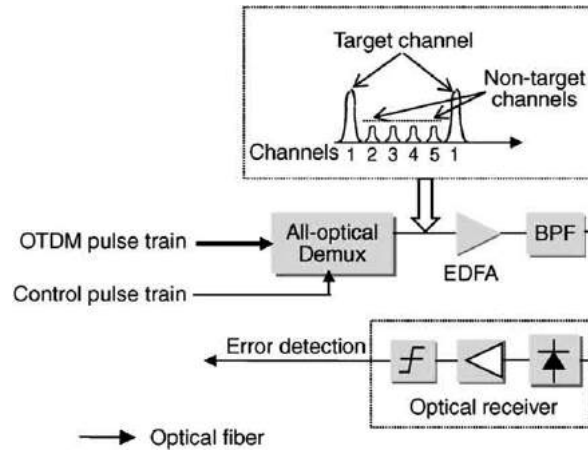


Fig. 21 Block diagram of OTDM receiver.

OTDM Bit-Error Rate (BER) Performance

Fig. 21 shows a typical block diagram of a high-speed optical receiver for an OTDM system, composed of a NOLM or a TOAD demultiplexer, an optical pre-amplifier, an optical bandpass filter, and a conventional optical receiver using a PIN photodiode. Due to the crosstalk introduced in the demultiplexing process, the demultiplexed optical signal contains not only the target channel but may also contain nontarget channels with reduced amplitude (see the inset in **Fig. 21**). The intensity of the demultiplexed optical signal is boosted by the optical pre-amplifier (EDFA). Amplified spontaneous emission (ASE) from the optical preamplifier adds a wide spectrum of optical fields onto the demultiplexed optical signal. Although an optical bandpass filter (BPF) can reduce the ASE, it still remains one of the major noise sources. The function of the optical filter is to reduce the excess ASE to within the range of the signal spectrum. The PIN photodiode converts the received optical power into an equivalent electrical current, which is then converted to a voltage signal by an electrical amplifier. Finally, the output voltage is sampled periodically by a decision circuit for estimating the correct state (mark/space) of each bit.

The BER performance of an OTDM system deteriorates because of the noise and crosstalk introduced by the demultiplexer and is:

$$\text{BER} = \frac{1}{\sqrt{2\pi}} \frac{\exp(-Q^2/2)}{Q} \quad (13)$$

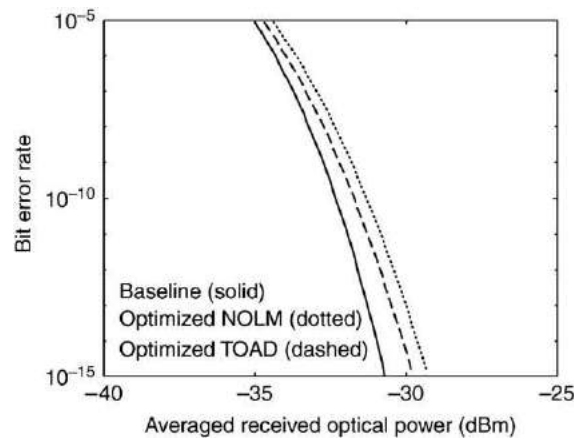


Fig. 22 BER versus the average received optical power for 100 Gb/s (10 channels) OTDM system.

where Q is defined as:

$$Q = \frac{\bar{I}_m - \bar{I}_s}{\sqrt{\sigma_{\text{RIN},m}^2 + \sigma_{\text{RIN},s}^2 + \sigma_{\text{amp},m}^2 + \sigma_{\text{amp},s}^2 + \sigma_{\text{rec},m}^2 + \sigma_{\text{rec},s}^2}} \quad (14)$$

and $\bar{I}_m$ and $\bar{I}_s$ are the average photocurrents for a mark and a space, respectively. $\sigma_{x,i}$ are the variances of the relative intensity noise (RIN), further optical pre-amplifier and receiver for a mark and a space.

For 100 Gb/s (10 channels) OTDM system, the BER against average received optical power for optimized NOLM and TOAD demultiplexers, is shown in Fig. 22.

See also: Wavelength Division Multiplexing

Further Reading

- Agrawal, G.P., 2002. *Fiber-Optic Communication Systems*, 3rd edn. New York: J Wiley and Sons Inc.
- Chang, K. (Ed.), 2003. *Handbook of Optical Components and Engineering*, 2nd edn. New York: Wiley and Sons Inc.
- De Marchis, G., Sabella, R., 1999. *Optical Networks Design and Modelling*. Dordrecht: Kluwer Academic.
- Hamilton, S.A., Robinson, B.S., Murphy, T.E., Savage, S.J., Ippen, E.P., 2002. 100 Gbit/s optical time-division multiplexed networks. *Journal of Lightwave Technology* 20 (12), 2086–2100.
- Jhon, Y.M., Ki, H.J., Kim, S.H., 2003. Clock recovery from 40 gbps optical signal with optical phase-locked loop based on a terahertz optical asymmetric demultiplexer. *Optics Communications* 220, 315–319.
- Kamatani, O., Kawanishi, S., 1996. Ultrahigh-speed clock recovery with phase locked loop based on four wave mixing in a traveling-wave laser diode maplifier. *Journal of Lightwave Technology* 14, 1757–1767.
- Nakamura, S., Ueno, Y., Tajima, K., 2001. Femtosecond switching with semiconductor-optical-amplifier-based symmetric-Mach–Zehnder-type all-optical switch. *Applied Physics Letters* 78 (25), 3929–3931.
- Ohara, T., Takara, H., Shake, I., *et al.*, 2003. 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15 (2), 302–304.
- Ramaswami, R., Sivarajan, K.N., 2002. *Optical Network; a Practical Perspective*. Morgan Kaufman.
- Reman, E., 2001. Trends and evolution of optical networks and technology. *Alcatel Telecommun. Rev.* 3rd Quarter 173–176.
- Sabella, R., Lugli, P., 1999. *High Speed Optical Communications*. Dordrecht: Kluwer Academic.
- Seo, S.W., Bergman, K., Prucnal, P.R., 1996. Transparent optical networks with time division multiplexing. *Journal of Lightwave Technology* 14 (5), 1039–1051.
- Sokoloff, J.O., Prucnal, P.R., Glesk, I., Kane, M., 1993. A terahertz optical asymmetric demultiplexer (TOAD). *IEEE Photonics Technology Letters* 5 (7), 787–790.
- Stavdas, A. (Ed.), 2001. *New Trends in Optical Network Design and Modelling*. Dordrecht: Kluwer Academic.
- Ueno, Y., Takahashi, M., Nakamura, S., Suzuki, K., Shimizu, T., Furukawa, A., Tamanuki, T., Mori, K., Ae, S., Sasaki, T., Tajima, K., 2003. Control scheme for optimizing the interferometer phase bias in the symmetric-Mach–Zehnder all-optical switch (invited paper). *Joint special issue on recent progress in optoelectronics and communications. IEICE Trans. Electron.* E86-C (5), 731–740.

Pulse Characterization Techniques

DJ Kane, Southwest Sciences, Inc., Santa Fe, NM, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

$A(t)$	Full electric field	$S_E(\omega, \tau)$	Spectrogram
$E(t)$	Complex electric field without carrier frequency	$S_E(\Omega, T)$	Sonogram
$E_{\text{sig}}(t, \tau)$	Complex electric field of the FROG trace versus time and time delay	$S(\omega)$	Spectral interferogram
$E_{\text{sig}}(t, \Omega)$	Fourier transform of $E_{\text{sig}}(t, \tau)$ with respect to τ	t, T	Variables for time
$\tilde{E}(\omega)$	Complex electric field in the frequency domain	τ	Fourier transform
$f(t)$	Correlation product	v_g	Group velocity
$f(\omega)$	Fourier transform of the correlation product	$\phi(t)$	Time domain phase
$g(t - \tau)$	Time gate or filter	Δw	Beam waist
G	FROG trace error	Z	Distance metric between the FROG signal field generated by the data constraint and the FROG signal field generated by the mathematical form constraint
$h(\omega - \Omega)$	Frequency gate or filter	Φ	Geometric angle between two beams
$I(t)$	Intensity profile of the electric field in the time domain	$\Delta\phi$	Phase difference
$\tilde{I}(\omega)$	Spectral intensity	$\prod^{(n)}$	n th order nonlinear susceptibility
$I_{\text{FROG}}(\omega, \tau)$	Intensity of the FROG trace	ω	Primary variable for frequency
$\Theta(t)$	Heaviside FUNCTION=0 $t < 0$, =1 $t \geq 0$	$\Delta\omega$	Difference in center frequency between two pulses
$\phi(\omega)$	Frequency domain phase	Ω	Secondary variable for frequency
		$\Omega(t)$	Instantaneous frequency

Introduction

Ultrashort optical pulses have very short time durations, typically less than a few tens of picoseconds. As a result, these pulses are spectrally broad. Because the index of refraction of materials is a function of wavelength, different wavelengths of light travel at different speeds in optical materials, causing the properties of ultrashort optical pulses to change as they propagate. However, the shape of the pulse can influence how the pulse itself interacts with materials. Thus, the goal of ultrafast laser pulse measurement is to obtain not only the intensity profile of the pulse, but also the actual variation of the frequencies that make up the pulse. This is called measuring the intensity and phase of the pulse, respectively.

Even though the development of techniques to characterize ultrashort optical pulses has not been easy, a myriad of pulse measurement techniques have been developed. In this chapter, we will confine our discussion to the most well-known pulse and generally accepted pulse characterization methods in an effort to provide a basic working knowledge of pulse measurement techniques and a fundamental understanding of the principles behind pulse measurement. After a brief discussion of the mathematical representation of ultrashort optical pulses, we will discuss pulse measurement in (roughly) chronological order, starting with autocorrelation methods. The next section introduces a useful method for the measurement of the relative phase between two pulses called spectral interferometry (SI). Following SI, we introduce the notion of the time-frequency representation of ultrashort optical pulses where techniques such as frequency-resolved optical gating (FROG) are discussed. Following the time-frequency section, a relatively new technique known as spectral phase interferometry for direct electric field reconstruction (SPIDER) is discussed.

Mathematical Representation of an Optical Pulse

The time-dependent variations of the pulse are embodied in the pulse electric field, $A(t)$, which can be written:

$$A(t) = \text{Re}[E(t)e^{i\omega_0 t}] \quad (1)$$

where ω_0 is the carrier frequency and Re refers to the real part. While we can use $A(t)$ as it stands for calculations, it is much easier to remove the rapidly varying ω_0 part, $e^{i\omega_0 t}$, and use a slowly varying envelope together with a phase term that contains only the frequency variations, not the rapidly varying carrier frequency:

$$E(t) = [I(t)]^{1/2} e^{i\phi(t)} \quad (2)$$

where $I(t)$ and $\phi(t)$ are the time-dependent intensity and phase of the pulse. (Note that $E(t)$ is complex.) The frequency variation, $\Omega(t)$, is the derivative of $\phi(t)$ with respect to time:

$$\Omega(t) = -d\phi(t)/dt \quad (3)$$

The pulse field can be written equally well in the frequency domain by taking the Fourier transform of Eq. (2):

$$\tilde{E}(\omega) = [\tilde{I}(\omega)]^{1/2} e^{-i\phi(\omega)} \quad (4)$$

where $\tilde{I}(\omega)$ is the spectrum of the pulse, and $\phi(\omega)$ is its phase in the frequency domain. The spectral phase contains time versus frequency information; that is, the derivative of the spectral phase with respect to frequency yields the time arrival of the frequency – the group delay.

Obtaining the intensity and phase, $I(t)$ and $\phi(t)$ (or $\tilde{I}(\omega)$ and $\phi(\omega)$) is called full characterization of the pulse. Common phase distortions include linear chirp, where the phase (either the time domain or frequency domain) is parabolic. When the frequency is increasing in time, the pulse is said to have positive linear chirp; negative linear chirp is when the high frequencies lead the lower frequencies. Higher-order chirps are common, but for these, differentiation between spectral and temporal chirp is required because spectral phase and temporal phase are not interchangeable.

Autocorrelation

Traditionally, we measure events using shorter events. Unfortunately, for the ultrafast researcher, shorter events do not exist and modern electronics are not fast enough. Therefore, because the shortest event we have is the event we wish to measure, traditional pulse measurement methods use the pulse itself to determine the approximate duration of the pulse in question giving birth to the ubiquitous intensity autocorrelation (see Fig. 1). While autocorrelations cannot be used to fully characterize ultrashort optical pulses, the methods used in autocorrelations are fundamental to all pulse measurement schemes.

The intensity autocorrelation is measured by combining a pulse and a delayed replica of a pulse in a nonlinear medium such as a second harmonic generation (SHG) crystal. A pulse is sent onto a beamsplitter to produce the two replicas. One pulse is delayed relative to the other and both are focused together into a SHG crystal. As the delay of one pulse relative to the other is varied, the intensity of the second-harmonic signal is recorded. The intensity autocorrelation is not limited to using SHG; any nonlinearity may be utilized. Indeed, two-photon absorption in a semiconductor LED or photodiode often acts as a convenient nonlinearity. Sometimes a third-order nonlinearity is used to provide some direction of time information about the pulse. When the generation of the signal involves phase matching, such as second harmonic generation, care must be taken to use a thin crystal. Typically, 10 μm to 1 mm thick SHG crystals are used, depending on the bandwidth of the pulse to be measured. The exact thickness depends on the SHG crystal, the pulse width, and the requirements of the measurement.

Regardless of the nonlinearity used, an autocorrelation yields only the approximate duration and shape of the pulse. The structure of the pulse is smeared, producing a smooth, featureless profile for even a complex pulse. Thus, the intensity autocorrelation alone does not determine the intensity profile of the pulse, $I(t)$. Only the interaction of the intensity of the pulse is recorded, producing no phase information for the pulse. Some qualitative information can be gleaned from the spectrum and the autocorrelation, but usually only that the pulse is chirped, and higher-order chirps and complex pulse structures elude such an analysis.

Another useful form of intensity autocorrelation is the single-shot autocorrelator. In this case, fairly large beams are used, and these beams are set to intersect at an angle of 2Φ , tilting the pulse fronts, so that delay is mapped onto a spatial coordinate (see Fig. 2). The beams are then focused into a nonlinear medium using a cylindrical lens. The interaction region in the SHG crystal is imaged onto a linear array or CCD camera. Typically, when second-harmonic generation is used, the SHG crystal is oriented for Type I phase matching. The time window is proportional to $\Delta w(\sin \Phi)/\Delta v_g$, where Δw is the beam waist at the SHG crystal, and v_g is the group velocity of the pulse in the crystal. Like all SHG interactions, care must be exercised to insure that the phase matching is sufficient to mix the entire bandwidth of the pulse.

Jean-Claude Diels improved the intensity autocorrelation by the development of the interferometric autocorrelation. In this configuration, a Michelson interferometer is typically used so that the beams are collinear when arriving at the SHG crystal, which allows the beams to interfere. As a result, fringes appear on the autocorrelation. By examining the shape and extent of the fringes,

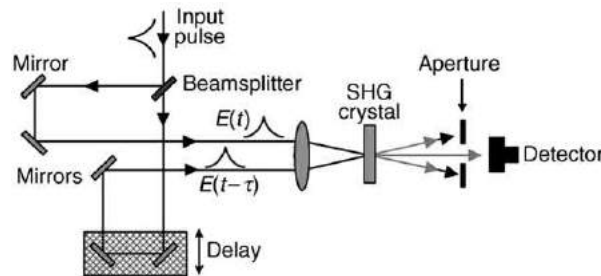


Fig. 1 An intensity autocorrelator used to determine the approximate duration of a pulse. The pulse is split into two replicas that are sent into delay lines. The outputs from both delay lines are focused into a doubling crystal. The second harmonic output is monitored as a function of delay between the two pulses. A plot of the signal versus time is the intensity autocorrelation of the pulse.

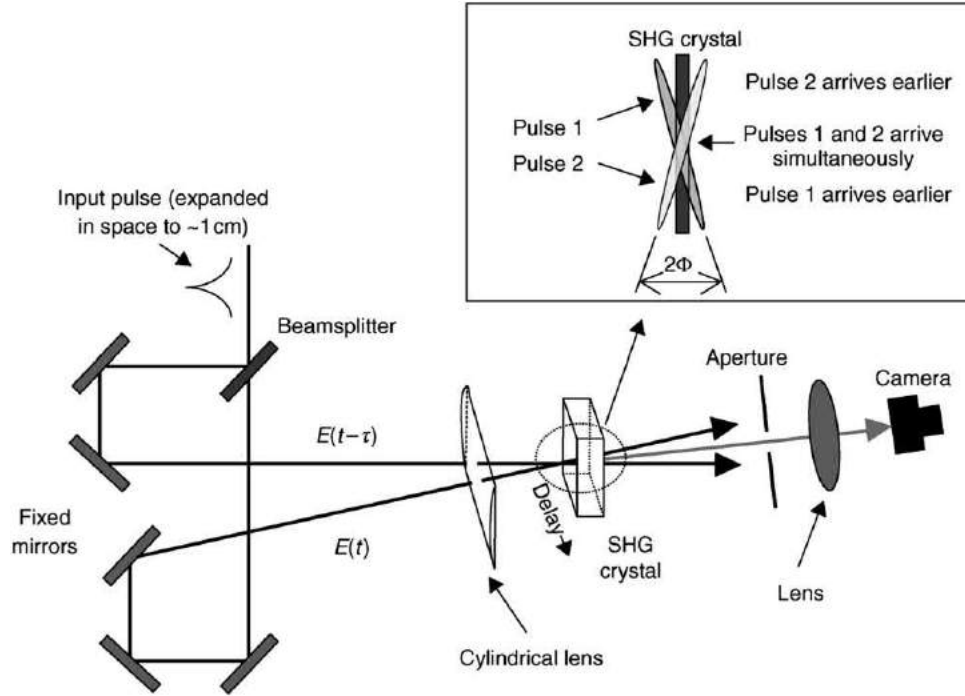


Fig. 2 A single shot intensity autocorrelator. For this type of autocorrelator, the beam size is large, on the order of 1 cm in diameter, and the pulse fronts are tilted with respect to each other in order to map relative delay to position. The two beams are focused using a cylindrical lens and the output is recorded using an array detector. The inset shows how delay is mapped into position. Inset: the tilted pulse fronts cause time delay to be mapped spatially.

some information about the pulse chirp can be obtained. While this does add to the information provided by an intensity autocorrelation, the added information is only the spectrum of the second harmonic. Unfortunately the addition of the second harmonic spectrum does not provide enough information for full retrieval of the intensity and phase.

Even though autocorrelation alone does not completely determine pulse intensity and phase, it is a very simple and useful technique to determine approximate pulse duration. Sometimes, cross-correlation can be used to determine the duration of a long pulse, if a shorter one is available, or a pulse too weak (or at an inappropriate wavelength) to measure with autocorrelation. The addition of the spectrum and/or the second harmonic spectrum can provide more, albeit, incomplete information about pulse chirp. More importantly, the experimental techniques used by correlation methods are fundamental to all pulse measurement methods.

Spectral Interferometry – Relative Phase Measurements

While determination of the absolute intensity and phase of an ultrashort laser pulse is difficult, determination of the relative phase is not. Invented in the late 1800s, spectral interferometry (SI), measures the relative phase between two pulses. SI, a linear technique, can also be modified (albeit, a some-what complex modification) to be self-referencing for full characterization of the ultrashort optical pulse (See the section on SPIDER below). In addition, when combined with a full pulse characterization technique, SI, because it is a linear technique, provides a method by which very weak optical pulses can be characterized (see below).

Typically, an SI apparatus uses a Mach–Zender interferometer (see Fig. 3). A pulse is split into two replicas; one is sent through a region or medium to measure and the other through a known path. The two pulses remain separated by some known time and are sent, collinearly, into a spectrometer; no time scanning is required. An analysis of the fringe pattern yields the relative phase between the two pulses. In other words, SI provides $\phi_{\text{unk}}(\omega) - \phi_{\text{ref}}(\omega)$, where $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase versus frequency. The spectrum of the two pulses in the frequency domain is

$$\tilde{I}_{\text{SI}}(\omega) = |\tilde{E}_0(\omega) + \tilde{E}_{\text{unk}}(\omega)|^2$$

$$\tilde{I}_{\text{SI}}(\omega) = \tilde{I}_0(\omega) + \tilde{I}_{\text{unk}}(\omega) + 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \quad (5)$$

where $\tilde{I}_{\text{SI}}(\omega)$ is the output spectrum, $\tilde{E}_0(\omega)$ is the reference pulse electric field, $\tilde{E}_{\text{unk}}(\omega)$ is the unknown pulse electric field, $\tilde{I}_0(\omega)$ is the reference pulse spectrum, $\tilde{I}_{\text{unk}}(\omega)$ is the unknown pulse spectrum, $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase, $\phi_0(\omega)$ is the reference

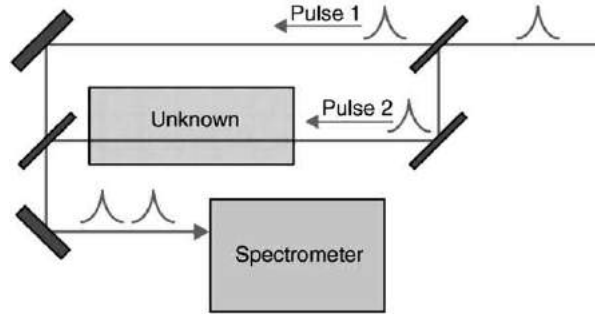


Fig. 3 A Mach-Zehnder interferometer is the typical arrangement for spectral interferometry. One path acts as the reference arm and the sample to be measured is placed in the other arm. Unlike standard interferometry, the delay is fixed in a spectral interferometry experiment.

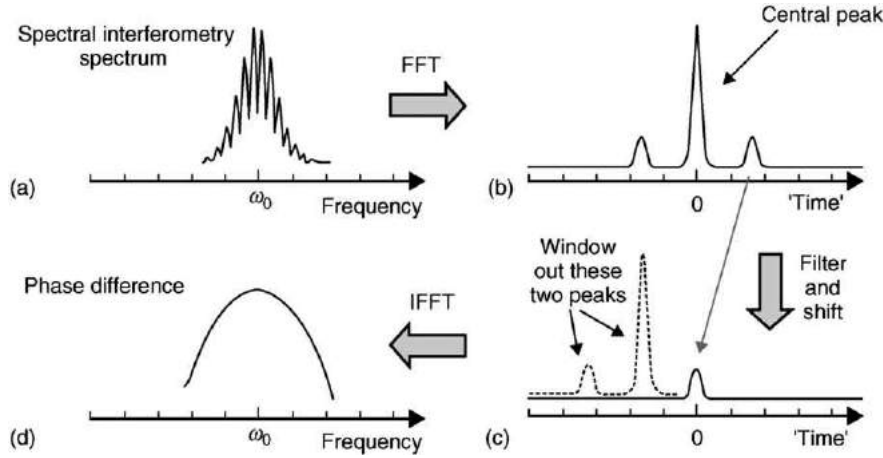


Fig. 4 The steps required for the analysis of spectral interferograms. Part (a) shows a sample SI interferogram. The dominant frequency of the fringes is the delay multiplied by the speed of light. The actual phase differences of interest are the perturbations on this frequency. Part (b) is the Fourier transform of the SI interferogram in Part (a). (Note that this is not the true time domain.) The central peak contains only spectral information. All of the phase information is contained in the satellite peaks. The next step is to mask out the central peak and one of the satellite peaks (Part (c)). The phase of the inverse Fourier transform of Part (c) yields the relative phase between the two pulses (Part (d)).

pulse phase, and τ is the delay between the two pulses. To facilitate extracting the phase difference from the SI output, τ is chosen to yield fringes in the sum spectrum. The spectral fringes, which have a period inversely proportional to the optical path difference between the two beams (see Fig. 4), contain all of the phase difference information.

Fig. 4 shows the steps required to obtain the phase difference between the two pulses. The first step is to subtract out the spectra of the individual pulses in order to isolate the spectral interferogram, $S(\omega)$ (Fig. 4(a)), where

$$S(\omega) = 2\sqrt{I_0(\omega)}\sqrt{I_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \quad (6)$$

By Fourier transforming $S(\omega)$, we obtain

$$\mathfrak{F}^{-1}[S(\omega)] = f(t - \tau) + f(-t - \tau) \quad (7)$$

where $f(t)$ is the correlation product between the reference and the unknown electric field (see Fig. 4(b)). The Fourier transform of Eq. (7) multiplied by a Heaviside function, $\theta(t)$ (to remove $f(-t - \tau)$), recovers the amplitude and phase of $f(\omega)$, the Fourier transform of $f(t - \tau)$, which contains the relative spectral phase between the unknown pulse and the unmodified pulse. Some care, however, needs to be taken to correctly remove the linear phase due to the time delay between the two pulses, $\omega\tau$. Also, for best results, the spectrometer should be well calibrated and care should be taken to properly window the spectrum before taking the Fourier transform.

Spectral interferometry is a straightforward and simple technique that always provides a relative phase. For this very reason, care must be taken to make sure the beams are mode-matched as well as collinear; also, interferometric stability must be maintained over the course of the measurement. Furthermore, SI is nongating; that is, CW background in the laser can add to the interferogram, masking transient effects. Nevertheless, spectral interferometry is a useful technique that has been applied to the measurement of the linear and nonlinear spectral phase introduced by optical fibers. More recently, SI has been applied to phase-locking and to phase-resolved pump probe experiments.

Time-Frequency Representation

In 1971, E.B. Treacy laid the foundations for a revolution in pulse measurement by introducing the idea of measuring the intensity versus time for different spectral slices of a pulse. Because his measurements provided both time and frequency information simultaneously, these measurements could be thought of as applying in a hybrid time-frequency domain. At first, this may seem confusing, but similar methods have been used to visualize sounds patterns (a musical score, for example) and speech. A time-frequency plot of this type, where spectral slices of a pulse are plotted versus time, is called a sonogram. The mathematical formalism of a sonogram is:

$$S_E(\Omega, T) = \left| \int_{-\infty}^{\infty} \tilde{E}(\omega) h(\omega - \Omega) e^{-i\omega T} d\omega \right|^2 \quad (8)$$

where $\tilde{E}(\omega)$ is the electric field of the pulse to be measured in the frequency domain and $h(\omega - \Omega)$ is a frequency gate that varies with frequency. The magnitude squared of the inverse Fourier transform of spectral slices yields the sonogram.

In 1991, Chilla and Martinez showed that the sonogram could be used to reconstruct the full intensity and phase of the pulse. That is, under certain conditions, the approximate group delay could be determined as a function of frequency from the sonogram by finding the peak time arrival of each spectral slice; integration of the group delay yields the spectral phase, which together, with the pulse spectrum, provides the intensity and phase of the pulse in the frequency domain. They labelled this technique as frequency domain phase measurement (FDPM).

An experimental diagram of an FDPM apparatus is shown in Fig. 5. The pulse is split into two replicas, and one of the pulse replicas is spectrally filtered. The spectrally filtered pulse is cross-correlated with the original pulse to find the 'time arrival' of the spectrally filtered pulse – defined as the peak of the cross-correlated pulse. Because the spectrally filtered pulse has a much longer time duration than the original pulse, the small perturbation caused by the finite length of the original pulse is neglected.

The main difficulty of the Chilla–Martinez method is that the frequency gate must be very narrow, reducing the filtered pulse energy significantly, which greatly reduces the measured signal strength. If the spectral phase is not well behaved, the filtered pulse may not be much longer than the original pulse. In addition, for a given frequency, the sonogram may have two or more peaks in time, that is, the group delay may not be a function. However, a method called 2D phase retrieval, discussed in the next section, can alleviate many of these issues.

A spectrogram is a relative of the sonogram. Rather than determining the time arrival of spectral slices of a pulse, a spectrogram is obtained when the spectrum of time slices of the pulse are measured. The spectrogram is experimentally much easier to obtain than the sonogram; however, the measured quantity is not quite as mathematically useful as the sonogram because the time slicing of the pulse occurs in the time domain. Thus, if a narrow time gate is used, the time domain phase is obtained. The time domain phase together with the intensity of the pulse, $I(t)$, provides the full intensity and phase of the pulse. Unfortunately, the intensity profile is not as readily measurable as the pulse spectrum. Consequently, to obtain the full intensity and phase of a pulse from its spectrogram requires an iterative 2D phase retrieval algorithm. This is the basis of a pulse measurement technique called frequency-resolved optical gating (FROG).

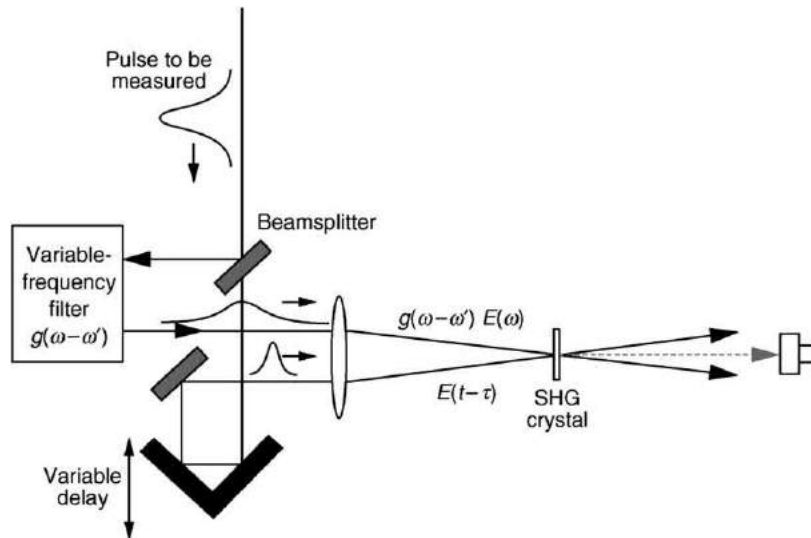


Fig. 5 Measuring a sonogram requires determining the time arrival of spectral slices of the pulse. The pulse is split into two replicas using a beamsplitter. One replica is spectrally filtered using a tunable filter. The other pulse is cross-correlated with the spectrally filtered pulse to determine the time arrival.

Frequency-Resolved Optical Gating

Frequency-resolved optical gating (FROG), developed by Kane and Trebino, measures the spectrum of a particular temporal component of the pulse (see Figs. 6 and 7) by spectrally resolving the signal pulse in an autocorrelation-type experiment using an instantaneously responding nonlinear medium. As shown in Fig. 6, FROG involves splitting a pulse and then overlapping the two

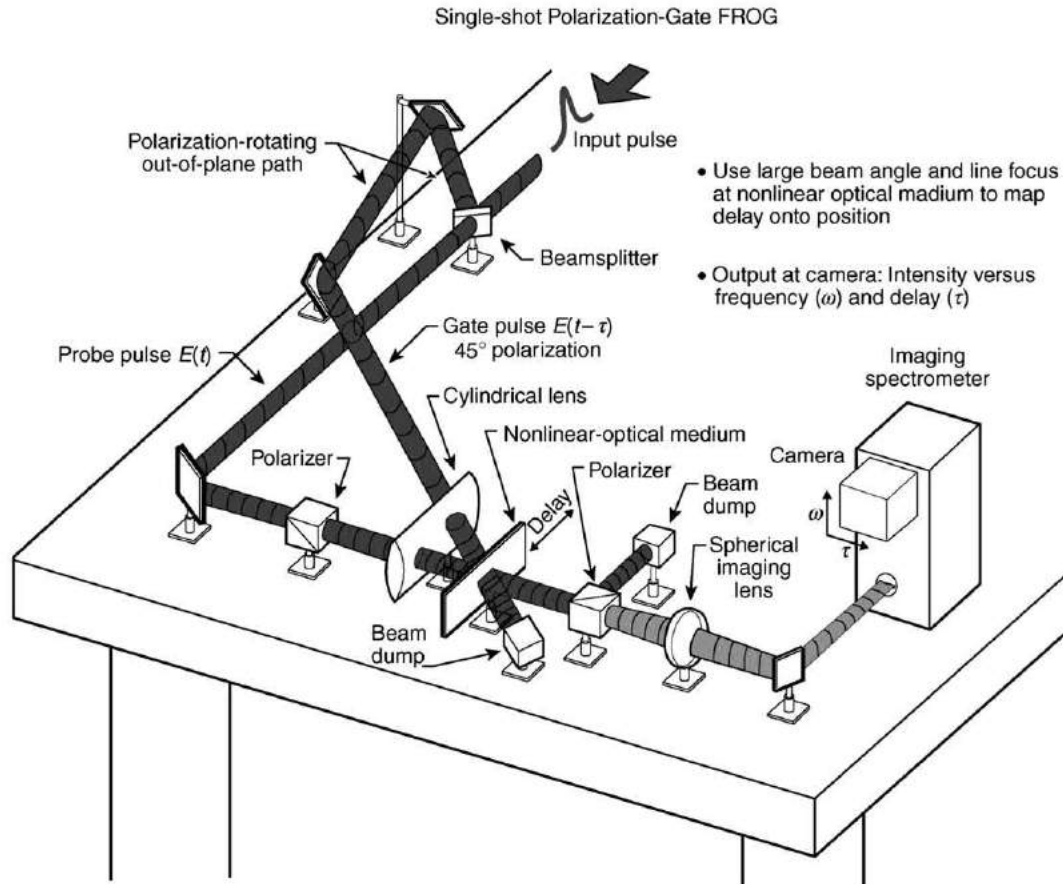


Fig. 6 Measuring the spectrogram of a pulse is easier than measuring its sonogram – a spectrogram is a spectrally resolved autocorrelation. In this figure, the optical Kerr-effect is used (polarization-gate) as the nonlinearity. Adapted with permission from Kane DJ and Trebino R (1993) Single shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating. *Optics Letters* 18(10): 823–825.

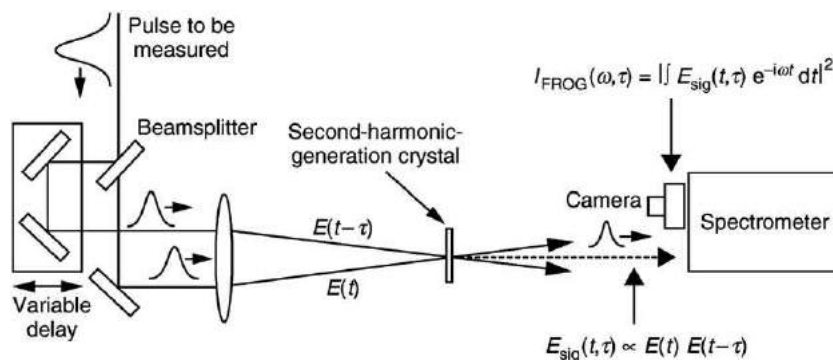


Fig. 7 This figure shows a schematic of an SHG FROG device. The SHG signal from the autocorrelation is spectrally resolved. SHG FROG is very simple and sensitive, but it has a direction-of-time ambiguity. For example, if the pulse has chirp, only the magnitude of the chirp is determined – the sign of the chirp remains unknown.

resulting pulses in an instantaneously responding $\chi^{(3)}$ or $\chi^{(2)}$ medium. Even though any instantaneous nonlinear interaction may be used to implement FROG, perhaps the most intuitive is the polarization-gating configuration. In this case, induced birefringence, due to the electronic Kerr effect, is used as the nonlinear-optical process. In other words, the ‘gate’ pulse causes the $\chi^{(3)}$ medium, which is placed between two crossed polarizers, to become slightly birefringent. The polarization of the ‘gated’ probe pulse (which is cleaned up by the first polarizer) is rotated slightly by the induced birefringence allowing some of the ‘gated’ pulse to leak through the second polarizer. This is referred to as the signal. Because most of the signal emanates from the region of temporal overlap between the gate pulse and the probe pulse, the signal pulse contains the frequencies of the ‘gated’ probe pulse within this overlap region. The signal is then spectrally resolved, and the signal intensity is measured as a function of wavelength and delay time τ . The resulting trace of intensity versus delay and frequency is a spectrogram, a time- and frequency-resolved transform that intuitively displays time-dependent spectral information of a waveform.

The spectrogram can be expressed as:

$$S_E(\omega, \tau) = \left| \int_{-\infty}^{\infty} E(t)g(t-\tau)e^{-i\omega t} dt \right|^2 \quad (9)$$

where $E(t)$ is the measured pulse’s electric field, $g(t-\tau)$ is the variable-delay gate pulse, and the subscript E on S_E indicates the spectrogram’s dependence on $E(t)$. The gate pulse $g(t)$ is usually somewhat shorter in length than the pulse to be measured, but not infinitely short. This is an important point: an infinitely short gate pulse yields only the intensity $I(t)$ and conversely, a CW gate pulse yields only the spectrum $I(\omega)$. On the other hand, a finite-length gate pulse yields the spectrum of all of the finite pulse segments with duration equal to that of the gate. While the phase information remains lacking in each of these short-time spectra, having spectra of an infinitely large set of pulse segments compensates for this loss. The spectrogram has been shown to nearly uniquely determine both the intensity $I(t)$ and phase $\varphi(t)$ of the pulse, even if the gate pulse is longer than the pulse to be measured (although if the gate is too long, sensitivity to noise and other practical problems arise).

In FROG, when using optically induced birefringence as the nonlinear effect, the signal pulse is given by:

$$E_{\text{sig}}(t, \tau) \propto E(t)|E(t-\tau)|^2 \quad (10)$$

So the measured signal intensity $I_{\text{FROG}}(\omega, \tau)$, after the spectrometer is:

$$I_{\text{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E(t)|E(t-\tau)|^2 e^{-i\omega t} dt \right|^2 \quad (11)$$

We see that the FROG trace is a spectrogram of the pulse $E(t)$ although the gate, $|E(t-\tau)|^2$, is a function of the pulse itself.

To see that the FROG trace essentially uniquely determines $E(t)$ for an arbitrary pulse, it is first necessary to observe that $E(t)$ is easily obtained from $E_{\text{sig}}(t, \tau)$. Then it is simply necessary to write Eq. (11) in terms of $E_{\text{sig}}(t, \Omega)$, the Fourier transform of the signal field $E_{\text{sig}}(t, \Omega)$, with respect to delay variable τ . We then have what appears to be a more complex expression, but one that will give us better insight into the problem:

$$I_{\text{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\text{sig}}(t, \Omega) e^{-i\omega t - i\Omega \tau} d\Omega dt \right|^2 \quad (12)$$

Eq. (12) indicates that the problem of inverting the FROG trace $I_{\text{FROG}}(\omega, \tau)$, to find the desired quantity $E_{\text{sig}}(t, \Omega)$, is that of inverting the squared magnitude of the two-dimensional (2D) Fourier transform of $E_{\text{sig}}(t, \Omega)$. This problem, which is called the 2D phase-retrieval problem, is well known in many fields, especially in astronomy, where the squared magnitude of the Fourier transform of a 2D image is often measured. At first glance, this problem appears unsolvable; after all, much information is lost when the magnitude is taken. It is well known that the one-dimensional (1D) phase retrieval problem is unsolvable (for example, infinitely many pulse fields give rise to the same spectrum). Intuition fails in this case, however; two- and higher-dimensional phase retrieval essentially always yields unique results.

The Simplified FROG – GRENOUILLE

A simplified version of the FROG device, known as grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE), can be constructed using the doubling crystal as both the gating medium and the spectrometer (see Fig. 8). In this case, the phase matching condition in certain thick doubling crystals causes a wavelength dependent output from the doubling crystal. A cylindrical lens is used in a Fourier transform configuration to image the wavelength variation onto a CCD camera. A cylindrical lens set at 90 degrees to the Fourier transform lens images the time axis onto the CCD camera. A Fresnel biprism replaces the traditionally used Mach-Zender interferometer to produce two pulses propagating at an angle with respect to each other. Presently, the GRENOUILLE technique works with BBO, in the wavelength region around 500 nm–1200 nm, and with proustite in the wavelength region around 1200 nm–2000 nm.

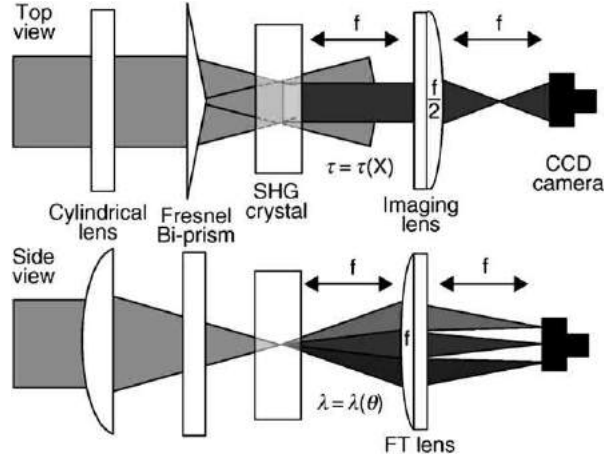


Fig. 8 Fig. showing a schematic of a GRENOUILLE FROG device which uses a thick doubling crystal as both the gating nonlinearity and the spectrometer. Like single-shot autocorrelators, a cylindrical lens focuses a spatially large input beam into an SHG crystal in only one dimension. Unlike most autocorrelators, the Mach–Zender interferometer is replaced with a Fresnel biprism that forces each half of the input beam to propagate, at an angle, toward the SHG crystal. An imaging cylindrical lens images the spatial dependence of the delay onto a CCD camera. A Fourier transform cylindrical lens maps the angular dependence of wavelength from the SHG crystal to a spatial coordinate on the CCD camera. Adapted with permission from O'Shea P, Kimmel M, Xun G and Trebino R (2001) Highly simplified device for ultrashort-pulse measurement. *Optics Letters* 26(12): 932–934.

FROG Inversion Algorithms

An iterative 2D phase retrieval algorithm is required to extract the pulse information from the measured FROG trace (see Figs. 9 and 10). This algorithm converges to a pulse that minimizes the difference between the measured and the calculated FROG trace. While this aspect of FROG has been its Achilles heel in the past, in reality, new generalized projections algorithms (together with faster computers) converge quickly and can track pulse changes at rates up to 30 Hz and beyond, making FROG a true, real-time pulse measurement technique. Indeed, excellent algorithms for the analysis of FROG traces in real-time are commercially available.

The original FROG inversion algorithm, commonly referred to as the vanilla algorithm, is simple and iterates quickly, but tends to stagnate, giving erroneous results, especially for geometries that use a complex gate function such as SHG or self-diffraction. An improved algorithm, incorporating several different algorithms including brute force minimization, was developed to alleviate stagnation programs at the expense of both speed and convergence time. By the mid-1990s, a significant advance in both speed and stability was made by the addition of a numerical method called generalized projections. This algorithm determines the next guess by constructing a projection – minimizing the error (distance) between the FROG electric field, $E_{\text{sig}}(t, \tau)$, obtained immediately after the application of the intensity constraint, and the FROG electric field calculated from the mathematical form constraint.

The first generalized projections algorithm used a standard minimization procedure to find the electric field for the next iteration, which can still be slow. For the most common FROG geometries, PG and SHG, a different algorithm that determines the next iteration directly has been developed. This algorithm, called the principal components generalized projections (PCGP) algorithm, converts the generalized projections algorithm into an eigenvector problem. Using this algorithm, pulse measurement rates of at least 30 Hz have been achieved.

The goal of the phase retrieval algorithm is to find the $E(t)$ that satisfies two constraints. The first is the FROG trace itself which is the magnitude squared of the 1D Fourier transform of $E_{\text{sig}}(t, \tau)$:

$$I_{\text{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\text{sig}}(t, \tau) e^{-i\omega t} dt \right|^2 \quad (13)$$

The other constraint is the mathematical form of the signal field, $E_{\text{sig}}(t, \tau)$, for the nonlinear interaction used. The signal forms for a variety of FROG beam geometries are:

$$E_{\text{sig}}(t, \tau) \propto \begin{cases} E(t)|E(t-\tau)|^2 & \text{PG FROG} \\ E(t)^2 E^*(t-\tau) & \text{SD FROG} \\ E(t)E(t-\tau) & \text{SHG FROG} \\ E(t)^2 E(t-\tau) & \text{THG FROG} \end{cases} \quad (14)$$

where PG is polarization gate, SD is self-diffraction, SHG is second harmonic generation, and THG is third harmonic generation FROG.

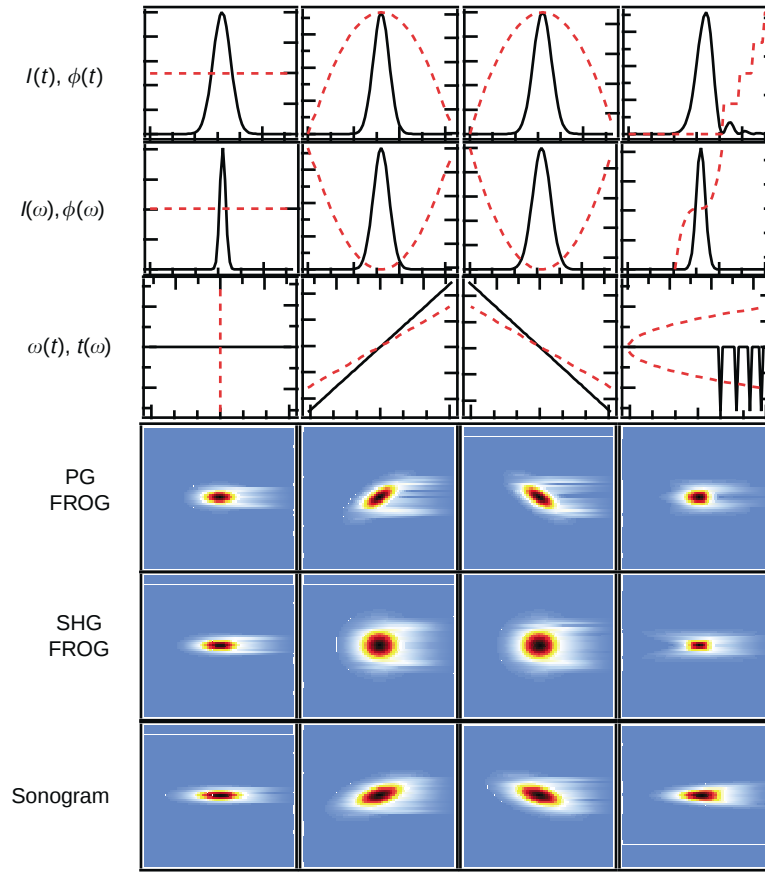


Fig. 9 Example FROG traces and sonograms are shown for four different pulses: a transform limited pulse, a positively chirped pulse, and a negatively chirped pulse. The top series of plots are the time domain representation of the pulse intensity (solid line) and phase (dashed line). The second row of four plots is the frequency domain representation of the sample pulse's intensity and phase. The next four plots show the instantaneous frequency and the group delay. The vertical axis is frequency and the horizontal axis is time. The remaining rows show the PG FROG traces, the SHG FROG traces and sonograms of the pulse shown in the first two rows.

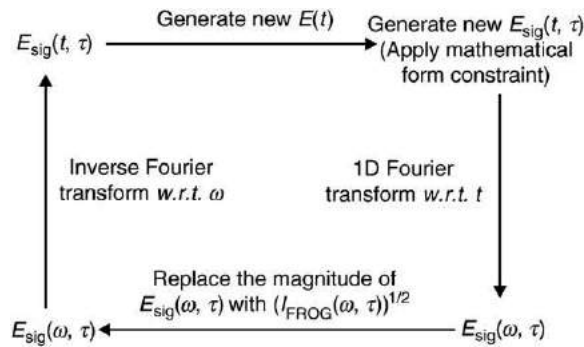


Fig. 10 Phase retrieval algorithm for the inversion of FROG spectrograms. Start with an initial guess for the pulse to generate an initial $E_{\text{sig}}(t, \tau)$. A 1D Fourier transform generates the FROG trace. The next step is to replace the magnitude of the calculated FROG trace with the square root of the measured FROG trace. Inverse Fourier transform with respect to ω to produce the new signal field and generate a new guess for $E(t)$. Interestingly, it is only the step that produces $E(t)$ from $E_{\text{sig}}(t, \tau)$ that differentiates all the FROG algorithms.

All FROG algorithms work by iterating between two different data sets: the set of all signal fields that satisfy the data constraint, $I_{\text{FROG}}(\omega, \tau)$, and the set of all signal fields that satisfy Eq. (14). The difference between the FROG algorithms is how the iteration between the two sets is completed. In the case of generalized projections, the $E(t)$'s are chosen such that the distance between the $E(t)$ on the magnitude set and the $E(t)$ on the mathematical form set is minimized. This is accomplished by minimizing

the following equation:

$$Z = \sum_{i,j=1}^N |E_{\text{sig(DC)}}^{(k)}(t_i, \tau_j) - E_{\text{sig(MF)}}^{(k+1)}(t_i, \tau_j)|^2 \quad (15)$$

where $E_{\text{sig(DC)}}^{(k)}(t_i, \tau_j)$ is the signal field generated by the data constraint, and $E_{\text{sig(MF)}}^{(k+1)}(t_i, \tau_j)$ is the signal field produced from one of the beam geometry equations in Eq. (14). For the normal generalized projection, the minimization of Z is completed using a standard steepest descent algorithm; the derivative of Z with respect to the signal field is computed to determine the direction of the minimum. The computation of the derivatives are tedious; they are tabulated elsewhere.

An alternative to an algorithm that minimizes Eq. (15) is the PCGP algorithm. This algorithm converts the computation of the next guess into an eigenvector problem, reducing the computation of the next guess to simple matrix–vector multiplications. This algorithm works for both the PG and SHG beam geometries, it is robust and the fastest FROG algorithm. Indeed, the PCGP algorithm was used in the original real-time FROG work by D.J. Kane.

Self Checks in FROG Measurements

Unlike any other pulse measurement techniques, FROG can provide a great deal of feedback about both the quality of the measurement (systematic errors) and the quality of the algorithm's performance. The most common check for convergence is the FROG trace error together with a visual comparison between the retrieved FROG trace and the measured FROG trace. The FROG trace error is given by:

$$G = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^N |I_{\text{FROG}}(\omega_i, \tau_j) - \alpha I_{\text{FROG}}^{(k)}(\omega_i, \tau_j)|^2} \quad (16)$$

where α is a renormalization constant, I_{FROG} is the measured trace, and $I_{\text{FROG}}^{(k)}(\omega, \tau)$ is computed from the retrieved electric field. Typically, the FROG trace error of a PG measurement should be less than 2% for a 64×64 pixel trace, while the FROG trace error of a 64×64 pixel SHG FROG trace should be about 1% or less. Acceptable FROG trace errors decrease as FROG trace size increase for smaller FROG traces. These values are only rules of thumb only; for example, acceptable retrievals of large and very complicated FROG traces can produce larger FROG trace errors.

In a good FROG measurement, the spectrum of the retrieved pulse should faithfully reproduce the salient features of the pulse's measured spectrum. SHG FROG even provides an additional check called the frequency marginal. The sum of an SHG FROG trace along the time axis yields the autoconvolution of the pulse's spectrum, providing two ways the FROG measurement can be checked. First, the autoconvolution of an independently measured spectrum can be compared to the sum of the FROG trace along the time axis, providing an indication of how well the measurement was made. For example, if the doubling crystal was too thick in the pulse measurement, the FROG trace's frequency marginal will be narrower than the autoconvolution of the measured spectrum. Second, comparing the autoconvolution of the retrieval pulse spectrum with the FROG trace marginal can provide a test of algorithm convergence in addition to a test of the measurement. Phase matching problems appear as a mismatch between the FROG trace frequency marginal and the autoconvolution of the retrieved spectrum.

Because FROG is a spectrally resolved autocorrelation, summing any FROG trace along the frequency axis yields the autocorrelation of the measured pulse. This autocorrelation can be compared to an independently measured autocorrelation, or a comparison can be made between the frequency sum of the FROG trace and the autocorrelation calculated from the retrieved pulse to determine algorithm convergence and the quality of the measurement.

Measuring Pulses Directly – SPIDER

SPIDER is a novel technique that measures the pulse directly – no iterative phase retrieval method is required. To accomplish this feat, spectral interferometry is conducted on two pulse replicas that are shifted in frequency with respect to one another. The original pulse is split into two replicas. The two pulse replicas are sent into a phase modulator that shifts the center frequency of each pulse slightly. The frequency-shifted pulses are sent into a spectrometer and a standard spectral interferometry analysis yields the derivative of the phase. Integration of the phase derivative, together with the spectrum of pulse, provides the full intensity and phase.

Thus, analysis of the spectral interferogram gives:

$$\Delta\phi = (\phi(\omega_1) - \phi(\omega_2))\Delta\omega \quad (17)$$

where $\Delta\phi$ is the measured phase difference from the spectral interferogram, $\Delta\omega$ is the difference in the center frequency between the two pulses, and $\phi(\omega)$ is the pulse phase as a function of frequency. Consequently, the phase difference must be divided by the frequency difference between the two pulses to determine the true derivative. It is also important to subtract out the linear phase difference resulting from the delay between the two pulses (the delay produces the fringes in the resulting interferogram) as this integrates to a linear chirp.

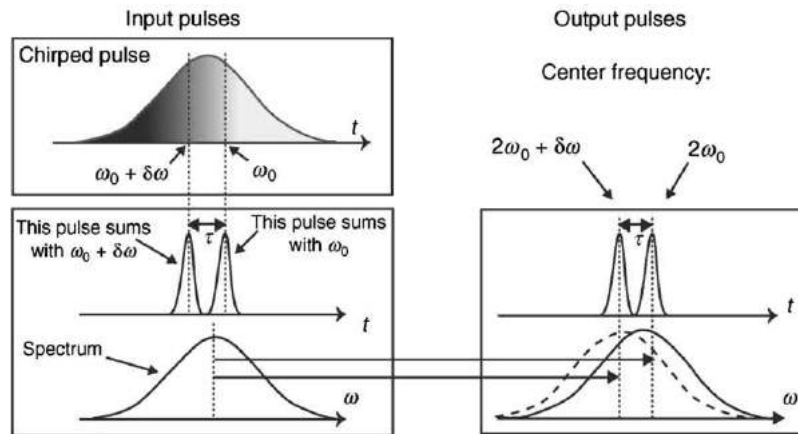


Fig. 11 Fig. showing how a SPIDER device works. The highly chirped pulse has a frequency that varies with time. Mixing the chirped pulse with two time delayed replicas produces two pulses, separated in time, that have slightly different frequencies, but are otherwise identical. By interfering the two pulses in a spectrometer, the relative phase between $\omega_0 + \delta\omega$ portions of the pulse and ω_0 portions of the pulse are determined.

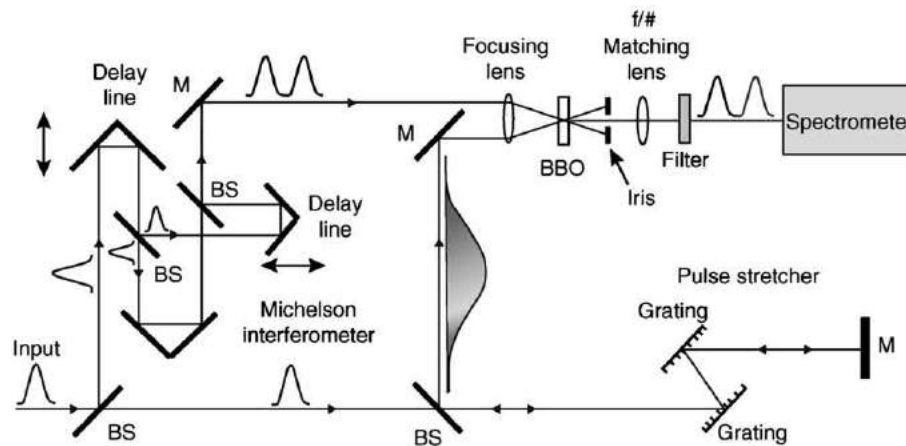


Fig. 12 SPIDER apparatus. The input pulse is first split into two pulses. One pulse is stretched in a pulse stretcher. The other pulse is sent into a Michelson interferometer to produce two time-delayed pulses. The two time-delayed pulses are mixed with the stretched pulse to produce two pulses at a slightly different center frequency. The two pulses are sent into a spectrometer to produce a spectral interferogram. Analysis of the spectral interferogram yields the derivative of the pulse phase.

Because phase modulators are too slow to be used in the femtosecond region, a different method is used to shift the frequency of the two pulses – they are mixed with a highly chirped pulse (see Fig. 11). Because the pulses are slightly time delayed, each one aligns with a slightly different frequency in the chirped pulse. The chirped pulse must be so highly chirped that the pulse frequency does not change significantly over the duration of the pulse to be measured.

A schematic of a SPIDER device for the measurement of femtosecond laser pulses is shown in Fig. 12. The input pulse is first split into two replicas. The first replica is sent to a pulse stretcher to produce to highly chirped pulse. The other replica is sent into a Michelson interferometer (or a simple étalon may sometimes suffice) to create two time-delayed replicas. The time-delayed replicas are mixed with the chirped pulse. Because each replica is mixed with a slightly different frequency in the chirped pulse, each replica produces a pulse with a slightly different center frequency. When the time-delayed replicas are spectrally resolved, each frequency of the pulse interferes with a frequency-shifted portion of the same pulse. Thus, the measured phase difference is proportional to the derivative of the phase of the original pulse.

Measuring Weak Pulses

Because virtually all the pulse measurement techniques, that do not require a reference pulse, use a nonlinearity, measuring weak ultrafast pulses directly is difficult. Fortunately, because most weak pulses are generated from a stronger pulse, spectral interferometry can be used to determine the relative phase between the weaker pulse and the stronger pulse (for regions where there is spectral overlap between the two pulses). If the stronger pulse is characterized using another pulse measurement technique, then

the weaker pulse phase can be determined from the relative phase measurement and the known phase of the stronger pulse. For example, the Trebino group termed the combination of FROG and SI as TADPOLE (Temporal Analysis by Dispersing a Pair Of Light E-fields).

Which Pulse Measurement Method Should I Use?

Choosing a pulse measurement method depends on the required measurement as well as the precision and accuracy required. All the pulse measurement techniques have advantages and disadvantages, but excellent measurements have been made with all of them. Autocorrelation is adequate when all that is required is the approximate pulse duration without any phase information. Spectral interferometry is quite useful for measuring transient changes in a sample in pump-probe experiments. FROG is perhaps the most commonly used and general-purpose pulse measurement technique, providing a great deal of feedback about the quality of the pulse measurement. Inexpensive and convenient FROG devices, together with software, are available commercially, adding to their acceptance. While SPIDER is experimentally complex and does not have the cross checks that FROG has, it provides the quickest answer, requiring no iterative phase retrieval algorithm to obtain pulse intensity and phase.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant numbers DMI-9801116 and DMI-0091454. The author would also like to acknowledge Rick Trebino for the use of figures from his lectures.

Further Reading

- Diels, J.C., Rudolph, W., 1996. *Ultrashort Laser Pulse Phenomena*. San Diego, CA: Academic Press Inc.
- Dorner, C., Belabas, N., Likhomanov, J.-P., Joffe, M., 2000. Spectral resolution and sampling issues in Fourier-transform spectral interferometry. *Journal of the Optical Society B* 17, 1795.
- Kane, D.J., Weston, J., Chu, K.-C.J., 2003. Real-time inversion of polarization gate frequency-resolved optical gating spectrograms. *Applied Optics* 42, 1140–1144.
- Iaconis, C., Walmsley, I.A., 1999. Self-referencing spectral interferometry for measuring ultrashort optical pulses. *IEEE Journal of Quantum Electronics* 35, 501–509.
- Lepetit, L., Cheriaux, G., Joffe, M., 1995. Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *Journal of the Optical Society of America B* 12, 2467–2474.
- Rulliere, C. (Ed.), 2000. *Femtosecond Laser Pulses*. New York: Springer Verlag.
- Shuman, T.M., Anderson, M.E., Bromage, J., Iaconis, C., Waxer, L., Walmsley, I.A., 1999. Real-time SPIDER: ultrashort pulse characterization at 20 Hz. *Optics Express* 5, 134–143.
- Treacy, E.B., 1971. Measurement and interpretation of dynamic spectrograms of picosecond light pulses. *Journal of Applied Physics* 42, 3848–3858.
- Trebino, R., 2000. *Frequency Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*. Boston, MA: Kluwer Academic Publisher.
- Trebino, R., DeLong, K.W., Fittinghoff, D.N., Sweetser, J.N., Krunbugel, M.A., Richman, B.A., Kane, D.J., 1997. Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. *Review of Scientific Instruments* 68, 3277–3295.
- Various authors, 1999. Feature issue on ultrashort-laser-pulse measurement and applications. *IEEE Journal of Quantum Electronics* 35: 4.

Optical Time Division Multiplexing

LP Barry, Dublin City University, Dublin, Ireland

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

Dispersion [ps km⁻¹ nm⁻¹]

Transmission data rate [bit s⁻¹]

ADM Add drop multiplexer

CW Continuous wave

FWM Four-wave mixing

NOLM Nonlinear optical loop mirror

OTDM Optical time division multiplexing

SDH Synchronous digital hierarchy

SOA Semiconductor optical amplifier

STM Synchronous transport module

TOAD Terahertz optical asymmetric demultiplexer

WDM Wavelength division multiplexing

Introduction

As the demand for bandwidth continues to grow, driven by the massive increase in Internet usage, so will the necessity to have communications networks that can handle very high data rates. The use of optical fiber networks is the clear choice for such systems given the huge available bandwidth of the fiber transmission medium. However, in a basic optical communication system comprising a laser transmitter, an optical fiber transmission medium, and a receiver, the capacity is essentially limited by the speed at which light can be modulated at the transmitter. To overcome this limitation, which is basically due to the speed of available electronics, it is necessary to use optical multiplexing techniques, and one such technique, known as optical time division multiplexing (OTDM), will be the subject of this article.

Principles of Time Division Multiplexing

Time division multiplexing (TDM) has long been the traditional method for electrically combining information channels. If we take the most fundamental data rate to be that of a simple voice call at 64 kbit/s, then the transmission of data at higher bit rates is achieved by electrically multiplexing a large number of 64 kbit/s channels in the time domain. With the evolution of standards, a number of different data rates have been specified as standard transmission rates. In Europe the synchronous digital hierarchy (SDH) has a basic data rate of 155.52 Mbit/s, which is known as the synchronous transport module – Level 1 (STM-1). This particular data rate is essentially obtained by electrically multiplexing over 2000 voice calls in the time domain (with some of the capacity required for overhead information). By subsequently multiplexing a number of STM-1 channels together we can obtain transmission at the higher standard data rates of STM-4 (622 Mbit/s), STM-16 (2.48 Gbit/s), and STM-64 (9.88 Gbit/s), as outlined in [Table 1](#).

[Fig. 1](#) illustrates how basic electrical TDM is used in standard optical communication systems. Clearly as we approach the higher data rates of STM-64, a serious level of electrical multiplexing is required, and as the data rates increase so does the cost and complexity of the electrical equipment at the transmitter and receiver. Indeed the present state of the art in electronics seems to suggest 40 Gbit/s as the limit for electrically multiplexed communication systems. However, as we stated earlier, communications traffic has been growing explosively over the last decade and will continue to do so. In order to meet this demand for capacity, and better exploit the massive available bandwidth of optical fiber, it is necessary to use optical multiplexing techniques for communications systems. The two main optical multiplexing techniques available are wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM). WDM essentially involves transmitting data at a number of different wavelengths on the same fiber. Although electrical multiplexing may be limited to data transmission rates of around 40 Gbit/s using one laser, by multiplexing together N different wavelength channels each carrying 40 Gbit/s, we can achieve overall data rates up to and beyond a terabit/s.

The second of these optical multiplexing techniques, OTDM, is the subject of this article. Whereas WDM multiplexes optical data channels in the wavelength domain, the basic principle of OTDM communications is to increase the system

Table 1 Standard data rates for SDH transmission systems

<i>SDH standard</i>	<i>Data rate (Mbit/s)</i>
STM-1	155.52
STM-4	622.08
STM-16	2488
STM-64	9953
STM-256	39 813

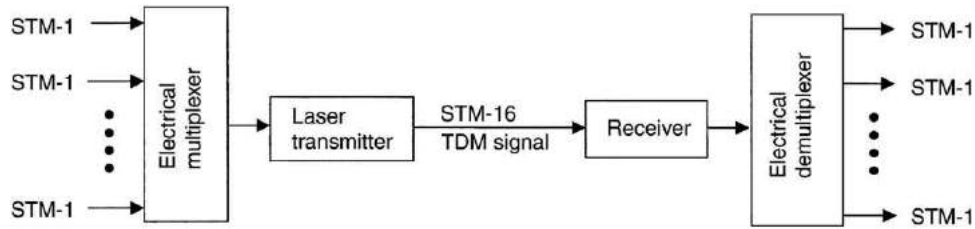


Fig. 1 Electrical time division multiplexing (ETDM) in a basic optical communication system.

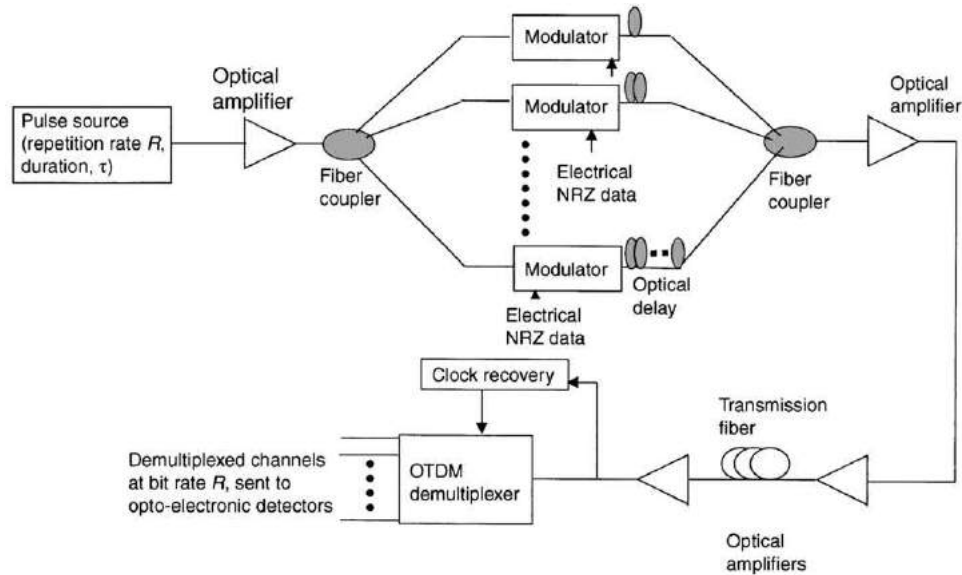


Fig. 2 Basic configuration for an OTDM transmission system comprising transmitter, transmission fiber, and receiver.

capacity by multiplexing optical data channels in the time domain. This multiplexing is usually achieved by a process known as bit-interleaving. This process can be explained by considering Fig. 2, which shows a basic schematic of an OTDM based transmission system. The main component of the overall system is a source of ultrashort optical pulses (pulse duration τ) at a certain repetition rate, R . The optical pulse source is initially split into N channels using a passive fiber coupler, and each pulse train is subsequently modulated by electrical data which is at a data rate of R . The resulting output from each modulator is essentially an optical data channel where the data are represented using ultrashort optical pulses (return-to-zero data format). The data from each modulator then passes through a fixed fiber delay line which delays each channel by a time equal to $1/RN$ relative to its adjacent channel (as shown in Fig. 2). The N modulated and delayed optical data channels are then recombined in another passive fiber coupler to form the OTDM data signal. We can consider that the delay lines essentially assign each data channel to a specific bit slot (of width $1/RN$) in the overall multiplexed signal. The multiplexed data signal may then be transmitted over optical fiber before arriving at the receiver which is responsible for demultiplexing the optical signal into its discrete channels.

The duration of the optical pulse source is extremely important in determining the maximum overall data rate which can be achieved. The overall data rate is basically defined by the temporal separation between channels in the multiplexor, but in order to avoid cross-channel interference, this separation must be significantly greater than the pulse duration. Thus to increase the overall data rate we must use shorter optical pulses. However, as we reduce the optical pulse width to raise the data rate we need to take into account the problems that may be encountered as this high-speed data signal propagates over optical fiber, and also the difficulty in demultiplexing a high-speed OTDM data signal. The following section will look in greater detail at these key elements of the OTDM communication system, namely the optical pulse source, the transmission of the ultrahigh capacity signal, and the demultiplexing at the receiver.

Key Elements of an OTDM Communication System

Ultra-Short Optical Pulse Sources

As stated earlier, the optical pulse source is a key element in any OTDM-based communication system. The important characteristics of the pulse source that will affect its usefulness in an OTDM system are the pulsewidth, the spectral width, and the

temporal jitter. The pulse duration clearly has to be short enough to support the desired overall transmission rate. For example, if we wish to design an OTDM system with an aggregate data rate of 100 Gbit/s (which will have a delay of 10 ps between each channel in the multiplexed signal), then the pulsewidth should normally be less than 30% of the channel spacing to avoid cross-channel interference (i.e. around 3 ps). For terabit/s OTDM systems we would require pulsewidths less than 0.5 ps. The spectral width of the pulse source is also important, as it will have a major impact on how the pulse will evolve during propagation in the fiber. A standard figure of merit which is employed is the time–bandwidth product, and ideally we require the pulse source to be transform limited, which implies that the spectral width is as small as possible for the associated pulsewidth. The impact of temporal jitter on the pulses can be easily understood by considering the above example of a 100 Gbit/s system, employing 3 ps pulses, with the multiplexed channels spaced by 10 ps. Obviously if the jitter on the pulses becomes significant in relation to the channel spacing, then this can also lead to interference between adjacent channels in the overall OTDM system. A large number of techniques have been employed to develop ultrashort pulse sources suitable for use in OTDM systems, but the three main methods which will be described below are active mode-locking, gain-switching, and external modulation of a CW light signal.

Active mode-locking of laser diodes normally involves modulating the amplitude of the optical field inside the laser cavity at a frequency which is equal to the mode spacing of the laser. This can be achieved by applying an electrical sinusoidal signal at the correct frequency, and results in the generation of optical pulses at the repetition rate of the applied signal. This technique has been successfully employed in the generation of subpicosecond pulses at repetition rates up to and beyond 40 GHz, with excellent spectral and temporal jitter characteristics. However, an inherent problem in all mode-locked pulse sources is the difficulty in synchronizing the mode-locked frequency to a specific SDH standard data rate.

Gain switching of semiconductor laser diodes is probably the simplest and most reliable technique to generate optical pulses. The technique, which is presented in Fig. 3, involves applying a high-power electrical pulse (or electrical sinusoidal signal) to the laser in conjunction with a certain bias current. By ensuring that the electrical pulse signal and the bias signal have the correct level, the relaxation oscillation phenomenon of the laser results in the production of optical pulses with durations between 10 and 30 ps, at the repetition rate of the applied electrical signal. The frequency of the electrical signal applied to the laser is essentially arbitrary (provided it is not larger than the modulation bandwidth of the diode), thus making it straightforward to synchronize the optical pulse train to a SDH line-rate. The main problem with this technique is that the spectral width of the pulses generated is such that the pulses are far from transform-limited, which would affect their subsequent propagation in the fiber. In addition, temporal jitter on gain-switched pulses can also be a problem. However, by using novel arrangements such as external injection into the gain-switched laser, this difficulty can be overcome.

The third pulse generation technique mentioned above involves external modulation of a cw light signal with an electro-absorption modulator. The experimental configuration for this pulse generation technique is shown in Fig. 4. By biasing the modulator around its null point, and applying an electrical sinusoidal signal to it, the cw light passing through the modulator becomes shaped into optical pulses. The optical pulse train which is generated due the nonlinear response of the electro-absorption modulator is at a repetition rate of twice the applied electrical signal. The pulses generated are normally transform limited with extremely low temporal jitter, and the repetition rate is arbitrary as in the gain-switching technique (with the limit being ultimately determined by the modulation bandwidth of the modulator). This method may be readily used to generate pulses at repetition rates up to 40 GHz with pulsewidth around 5 ps, and it has the advantage that the optical source and the modulator can be integrated in a single device to form a compact pulse source suitable for OTDM communication systems.

It should also be noted that in addition to the various techniques that have been outlined, it is possible to use pulse compression in order to reduce the pulse width. The main issues concerned with pulse compression are the shape and spectral width of the optical pulses after compression. One of the most attractive methods of achieving pulse compression involves using nonlinear compression in dispersion-decreasing fiber, with the main advantage of this technique being the ability to maintain a transform limited pulse after compression. By employing this compression scheme, optical pulse sources at 10 GHz with pulsewidths below 200 fs have been developed, and such pulsewidths would be suitable for use in Tbit/s OTDM systems.

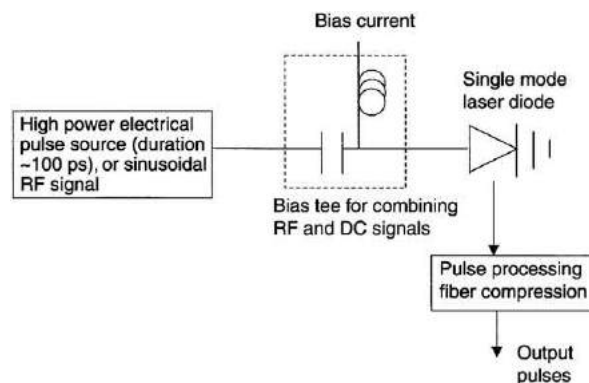


Fig. 3 Pulse generation using the gain-switching technique followed by pulse compression.

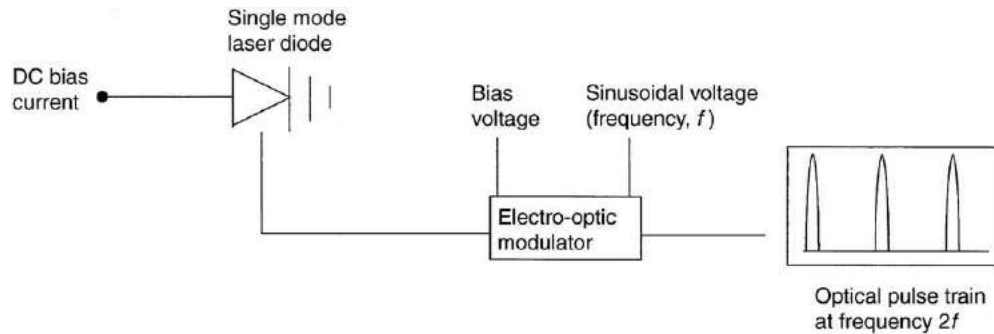


Fig. 4 Pulse generation based on shaping of cw light using the nonlinear response of an external modulator biased about its null point.

Transmission of an OTDM Signal over Fiber

The transmission performance of an OTDM data signal is vital in determining the distance over which the data can be transmitted successfully. The main fiber parameters, which will affect the signal transmission, are attenuation and dispersion. If we consider OTDM systems operating at a wavelength of around 1550 nm (minimum loss wavelength), to overcome the fiber attenuation and maintain a suitable optical power budget for the system, optical amplifiers are normally employed. If we thus assume that the amplifiers overcome the fiber loss problem, then the maximum transmission distance will be limited by the fiber dispersion. In a very basic OTDM system operating at 1550 nm, with transmitter and receiver linked using standard fiber (dispersion parameter of about $16 \text{ ps km}^{-1} \text{ nm}^{-1}$), the maximum transmission distance will be limited by the overall data rate, and the pulsewidth and spectral width of the optical pulse source. For example, consider a 40 Gbit/s OTDM data signal which is formed using 8 ps optical pulses with a spectral width of 40 GHz (0.32 nm). From the spectral width we can calculate the signal broadening due to dispersion to be around 5 ps km^{-1} , and as the pulses broaden and spread into adjacent channels of the OTDM signal then this interference will make it increasingly difficult to correctly detect the signal at the receiver. In this case after propagation through 5 km of fiber, the signal pulses will have already dispersed to around 25 ps duration, the same value as the temporal bit slot into which each channel is placed. This will clearly result in serious loss of signal integrity. A straightforward possibility to increase the transmission distance of OTDM systems is to employ dispersion shifted fiber; the dispersion parameter is now around $1\text{--}2 \text{ ps km}^{-1} \text{ nm}^{-1}$, at an operating wavelength of 1550 nm. This reduction in dispersion will obviously increase the allowed transmission distance by about an order of magnitude for the example described above. However, to develop ultrahigh-speed, long-haul OTDM communications we require more complex transmission schemes. Two possibilities for this include dispersion compensation, and soliton transmission techniques.

Dispersion compensation basically involves compensating for the dispersion that has been encountered during transmission by using some fiber with a total dispersion of opposite sign but equal magnitude to the transmission fiber. Dispersion compensation may also be achieved using a suitably designed fiber grating. For a long-haul OTDM system, a dispersion compensator may be used every 50 or 60 km, in conjunction with the optical amplifier, thus allowing us to compensate both fiber loss and attenuation periodically along the link. The main limitation, however, with the dispersion compensation technique is caused by the dispersion slope of the fiber, as for ideal compensation it is necessary to compensate completely for the dispersion slope in addition to the overall fiber dispersion. The second technique that may be employed to greatly extend the transmission distance of high-speed OTDM communication systems is the use of soliton transmission. The basic principle of soliton transmission is to use optical data pulses with a particular shape, pulsewidth, and peak power, such that as the pulse propagates, the effects of fiber dispersion and nonlinearity counterbalance to allow the signal to propagate undistorted. By using optical amplifiers to ensure that the optical pulse peak power does not vary too much along the transmission link, OTDM transmission at data rates greater than 100 Gbit/s, over distances approaching 500 km have been demonstrated.

Demultiplexing of OTDM Signal at Receiver

For OTDM communication systems with data rates above 40 Gbit/s, it is not feasible to use electrical switching. Indeed for ultrahigh-speed systems operating at 100 Gbit/s and beyond, the only solution for demultiplexing is to use all-optical switching in which optical control signals are used to switch the OTDM data signal. All-optical switches normally use nonlinear optical effects either in optical fiber or semiconductors. A typical scenario (as presented in Fig. 5) involves the injection of both the OTDM data signal and an optical control signal into the nonlinear device. The control signal consists of high-power ultrashort pulses at the repetition rate of the individual channels within the temporally multiplexed signal, and by synchronizing it with one of the OTDM channels it is possible to demultiplex this channel from the high-speed OTDM signal.

Two of the most popular all-optical demultiplexers are the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD). The NOLM is based on the nonlinear refractive index of optical fiber. It essentially consists of a 2×2 fiber coupler with its two outputs joined using a certain length of fiber. When an OTDM signal is injected into an input port of the coupler, it splits into two counterpropagating components in the fiber loop, and when these components recombine and

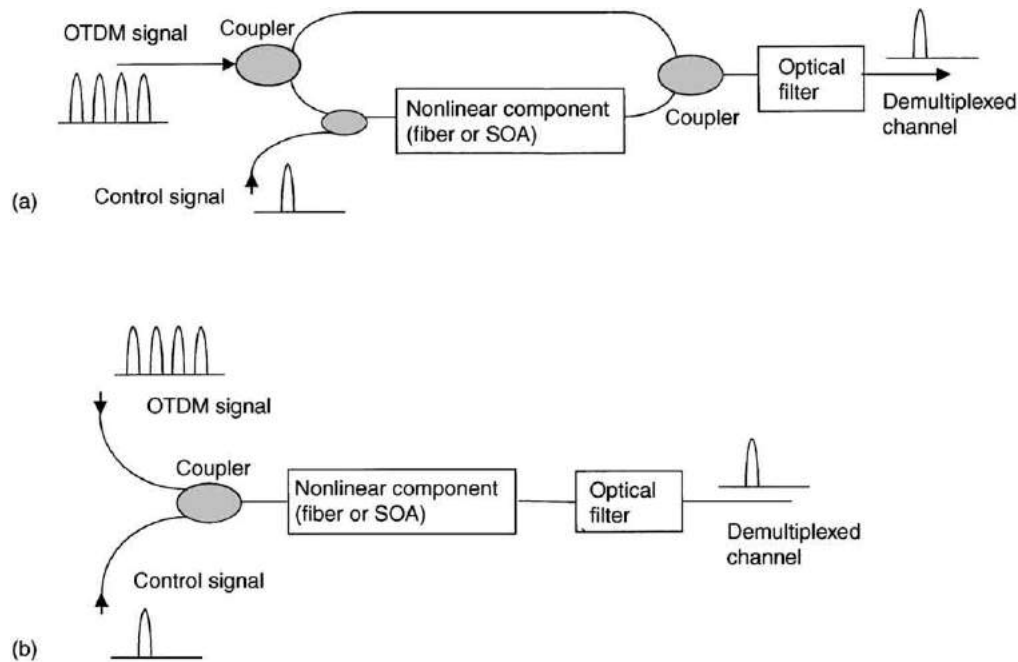


Fig. 5 Configuration of OTDM demultiplexers using (a) an interferometer based on nonlinear phase shift in fiber or semiconductor optical amplifiers, and (b) four-wave mixing in fiber or semiconductor optical amplifiers.

interfere at the coupler, the overall signal is output through its initial input port. If we now inject a high-power control signal directly into the loop, such that it propagates unidirectionally, and is synchronized with one of the OTDM data channels, then the phase shift induced by the control on the copropagating signal channel (via the nonlinear refractive index of the fiber) results in that particular channel being switched out to the second input port of the coupler. The remaining data channels of the OTDM signal are once again output through the same port that they entered the coupler. As this particular nonlinear effect in fiber (known as the Kerr effect) has a femtosecond response time, it should be possible to realize extremely high-speed switching with the NOLM. Demultiplexing at data rates in excess of 100 Gbit/s has already been achieved using this technique. One disadvantage of this method is that the nonlinear index coefficient of standard fiber is very small, so in order to achieve the required phase shift from the control pulse, it is necessary to use a fiber loop of around 1 km in length (depending on the power of the control pulse used). This length may affect the overall stability of the demultiplexer and its ability to be readily integrated into high-speed OTDM systems. However, it should be noted that recent developments in the design of high-nonlinearity fibers may reduce this problem.

The second all-optical switch mentioned above is known as a TOAD. Whereas the NOLM is based on the nonlinear refractive index of the fiber, the TOAD is based on a nonlinear optical effect in semiconductor optical amplifiers (SOA). The overall setup for a TOAD is very similar to that for a NOLM, except that a semiconductor amplifier, as shown in Fig. 5, replaces the length of fiber in the loop. As for the NOLM, the OTDM data signal is injected into the TOAD using one input of the coupler, such that it splits into two counterpropagating signals, while the control signal propagates unidirectionally in the loop containing the SOA. By synchronizing the control with one channel in the OTDM signal, this particular channel is switched out. Although the response time of a TOAD-based switch may not be as fast as the NOLM, it does have the major advantage that it can be developed into a very compact demultiplexer, suitable for deployment in OTDM-based communication systems.

In addition to the NOLM and the TOAD structures, which are both interferometric-based switches, it is also possible to use four-wavemixing (FWM) in optical fiber or semiconductor optical amplifiers to carry out the demultiplexing of OTDM signals. In this case, the high-power control pulse is synchronized with one of the OTDM channels such that the wavelength of this channel is shifted as it passes through the nonlinear device. Optical filtering may then be used to select out the demultiplexed channel.

As we discussed in the section on demultiplexing, whatever specific all-optical switching technique is employed, we need to generate an optical clock signal for use as the control pulses. This implies that it is necessary to extract a clock at the base rate of the individual data channels from the high-speed OTDM signal. Depending on the base data rate, which may be anywhere from 2.5 to 40 Gbit/s for typical OTDM systems, different clock recovery techniques have been demonstrated. These range from basic electronic clock recovery schemes to more advanced techniques based on self-pulsating laser diodes and optical phase-locked loops.

OTDM Networking Issues

The previous section has predominantly dealt with the basic elements required to implement a high-speed OTDM transmission link. However, if we are to use OTDM technology for high-speed networking, then additional elements will be required. One of

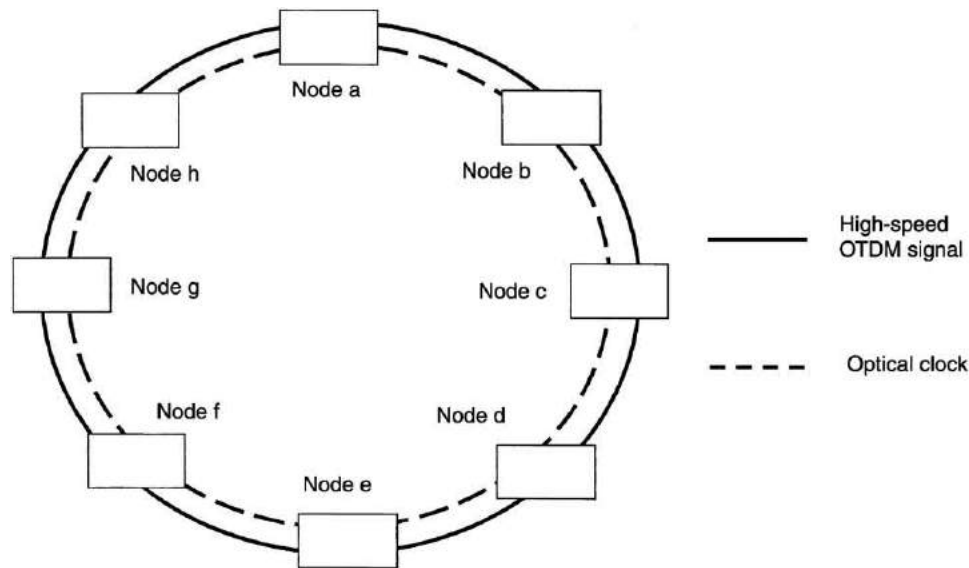


Fig. 6 Schematic representation of a ring network employing OTDM with a separate fiber for clock distribution.

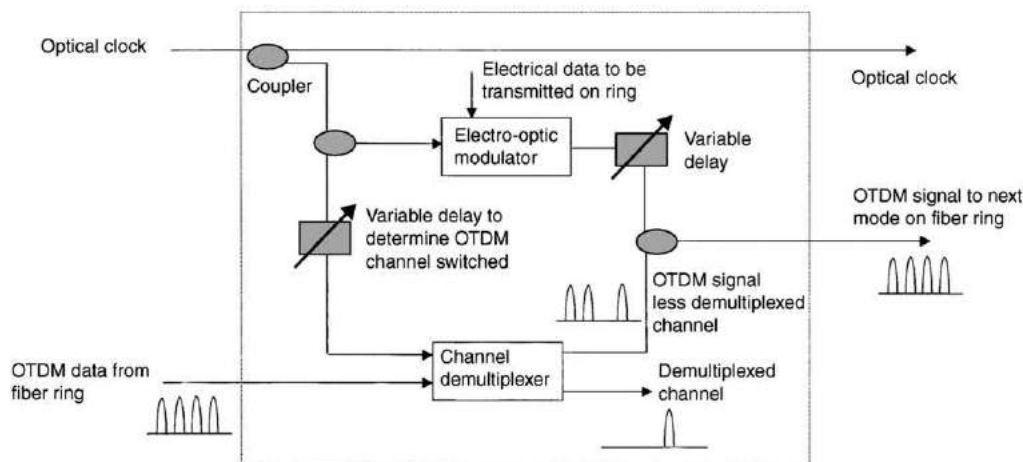


Fig. 7 Configuration of an add/drop multiplexing node for use in OTDM-based communication systems.

the most important of these is an add/drop multiplexer (ADM). The ADM allows us to switch out one channel from the OTDM signal, and then insert another data channel for transmission in the vacant bit slot. If we consider employing OTDM in a ring configuration network, then a typical architecture for the network will be as shown in Fig. 6. The high-speed OTDM data signal propagates around the closed fiber loop, and each node in the network is responsible for switching out the data it requires, in addition to inserting the data it wishes to transmit onto the fiber ring. Add/drop multiplexing is thus imperative in developing such an OTDM ring network.

The configuration for each node in an OTDM ring network is shown in Fig. 7. At each node it is initially necessary to recover the base rate clock signal. The recovered clock is then used in conjunction with the OTDM signal to switch out the required data channel from the multiplexed signal. When an add/drop multiplexer is used for this purpose we obtain two outputs, one being the demultiplexed data, and the other being the OTDM signal less the switched-out data channel. A simple fiber coupler can subsequently be used to allow this particular node to insert the data it wishes to send into the free bit slot of the overall OTDM signal. This obviously requires control of the relative delay between the OTDM signal and the local data channel to be inserted onto the fiber ring, to ensure that the local data are inserted into the vacant bit slot. A number of techniques have been used for implementing add/drop multiplexers in OTDM-based networks. These have used either fiber interferometers, or interferometers using semiconductor optical amplifiers, in which we obtain both the demultiplexed data signal and the OTDM signal less the switched-out channel (to which the local data channel is then added). The continuing development and enhancement of ADM devices for OTDM systems will be vital for the future implementation of OTDM communication systems.

Table 2 Main elements of an OTDM communications system

<i>Device</i>	<i>Principal function</i>	<i>Technical implementation</i>
Optical pulse source	Generate ultrashort optical pulses suitable for transmission in high-speed OTDM systems	Gain-switching, mode-locking, or external modulation techniques
Multiplexer	Multiplex individual pulse data channels in the temporal domain using appropriate delays to achieve an OTDM signal	Passive couplers with fixed fiber delay lines, or planar waveguide circuits (for accurate control of delays)
Optical amplifier	Amplify the OTDM signal during fiber transmission to overcome fiber loss	Fiber doped with rare earth materials (e.g. erbium) in conjunction with pump laser
Transmission fiber	Transmit the OTDM data signal from transmission site to required receiver	Standard single-mode fiber, dispersion shifted fiber, or other specialty fiber may be used
Clock recovery	Extract clock (at base rate of individual channels) from OTDM data signal at receiver, for use in demultiplexing	Electronic or optical phase-locked loops, self-pulsating lasers, or mode-locked ring laser techniques
Demultiplexer	Demultiplex one of the channels from the overall OTDM signal	Interferometric or four-wave mixing techniques based on fiber or semiconductor nonlinearity
Add/drop demultiplexer	Extract one channel from OTDM data signal, and insert a local channel into the empty time slot for transmission	Similar technical implementations to demultiplexer described above

Conclusion

In this article we have explained the main ideas involved in building OTDM communication systems, and also introduced the principal technologies which are required to do so. OTDM is an extremely wavelength-efficient technique for delivering high-capacity data signals, and it may be used both for long-haul transmission and networking around metropolitan areas. In addition, unlike WDM, it does not require a different wavelength source for each channel in the multiplexed signal. Despite these and other advantages, OTDM is still way behind WDM when it comes to commercial maturity, with no OTDM systems currently available in the telecommunications market. However, the technologies required to implement high-speed OTDM systems (outlined in [Table 2](#)), including ultrashort pulse sources, clock extraction circuitry, and demultiplexing devices, are developing rapidly. It therefore seems likely that in the future, OTDM technology may be used for enhancing the overall capacity of optical communication systems.

One possibility for employing OTDM systems would be for upgrading installed WDM networks. The current method for improving WDM networks involves using more wavelength channels, but it may also be feasible to increase the overall capacity by developing hybrid WDM/OTDM communications systems. In such a system, each wavelength channel may carry aggregate data rates in excess of 100 Gbit/s, achieved using OTDM. The design of these hybrid networks would permit switching of the optical data signals to be carried out at two different levels in the overall network. The WDM signals could be switched coarsely using passive filtering devices, and the OTDM data channels could be switched with fine granularity using one of the all-optical switching techniques available.

See also: Lightwave Transmitters

Further Reading

- Agrawal, G.P., 1997. *Fibre Optic Communications*, 2nd edn. New York: Wiley.
- Cole, M., 2000. *Introduction to Telecommunications: Voice, Data, and the Internet*. Englewood Cliffs, NJ: Prentice Hall, chapter 5.
- Islam, M., 1992. *Ultrafast Fiber Switching Devices and Systems*. Cambridge, UK: Cambridge University Press.
- Kawanishi, S., 1998. Ultrahigh-speed optical time-division-multiplexed transmission technology based on optical signal processing. *IEEE Journal of Quantum Electronics* 34, 2064–2079.
- Midwinter, J.E., Guo, Y.L., 1992. *Optoelectronics and Lightwave Technology*. Chichester, UK: Wiley.
- Mukherjee, B., 1997. *Optical Communications Networks*. New York: McGraw-Hill.
- Nakazawa, M., Kubota, H., Suzuki, K., Yamada, E., Sahara, A., 2000. Ultrahigh-speed long distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6, 363–396.
- Prucnal, P.R., Santoro, M.A., Sehgal, S.K., 1986. Ultrafast all-optical synchronous multiple access fiber networks. *IEEE Journal of Selected Areas of Communication* 4, 1484–1493.
- Spirit, D.M., Blank, L.C., Ellis, A.D., 1995. Optical time division multiplexing for future high capacity network applications. In: Smith, D.W. (Ed.), *Optical Networking Technology*. London: Chapman and Hall, pp. 54–77.
- Spirit, D.M., Ellis, A.D., Barnsley, P.E., 1994. Optical time division multiplexing: systems and networks. *IEEE Communications Magazine* 32 (12), 56–62.
- Tucker, R.S., Einstein, G., Korotky, S.K., 1988. Optical time-division multiplexing for very high bit rate transmission. *IEEE Journal of Lightwave Technology* 6, 1737–1749.

Lightwave Transmitters

JG McInerney, National University of Ireland-Cork, Cork, Ireland

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Optical fiber telecommunications has changed human society forever, providing the capacity for affordable and ubiquitous communication. It provides data rates and transport economies far in excess of those available using purely electronic means. It is doubtful whether the Internet and worldwide web, or the vast infrastructure of wireless cellular telephony, would be viable without optical fiber technology. Its rapid development and acceptance has been driven largely by contemporaneous successes in manufacturable low-loss optical fiber cables, sensitive pin-FET photoreceivers, and reliable high-performance laser diode-based transmitters. The earliest developments in silica glass fibers at Standard Telecommunication Laboratories, UK, (then part of the US-based ITT Corporation) and Corning Glass Works (USA) in the 1960s and 1970s, were pivotal in determining the directions of modern fiber technology, as were later key inventions of rare-earth doped fiber amplifiers at the University of Southampton (UK) and AT&T Bell Laboratories (USA) in the 1980s and early 1990s. However, perhaps the greatest contributors to the success of fiber optics have been semiconductor injection lasers, or laser diodes, which in various forms have made up the vast majority of all fiber optic transmitters, and the entirety of long-haul transmitters, as well as pump sources for the doped-fiber amplifiers which enable modern networks. The fascinating story of the development of fiber optics is to a large extent driven by the dramatic progress in the development of high-performance, reliable semiconductor lasers. These emitted at wavelengths near 1300 and 1550 nm, which respectively correspond to the dispersion and attenuation minima of single-mode silica optical fibers.

Semiconductor Laser Principles

Semiconductor lasers were first demonstrated in research laboratories at General Electric, IBM Corporation and the MIT Lincoln Laboratory (all USA) as early as 1962, although it took almost two decades for the basic science and engineering of materials, fabrication, reliability, and performance design issues to be developed sufficiently for their use in practical communication systems. Early successes were obtained using the GaAs/AlGaAs lattice-matched material system, in which a lightly doped GaAs or ternary compound AlGaAs active layer is sandwiched between two lattice-matched heterojunctions with n- and p-doped AlGaAs layers with higher Al fractions than the active layer, so that their refractive indices are lower and their bandgaps higher, providing optical and charge-carrier confinement. In all cases the material is epitaxially grown, meaning that the entire laser chip forms a single crystal. The gain, and hence the laser emission, occurs at the direct bandgap of the active layer, at wavelengths in the range 750–870 nm depending on the Al fraction. Later evolutions used the quaternary compound GaInAsP clad by InP guiding layers, where the In and P fractions are chosen to match the lattice constant to the InP substrate.

In general, optical gain in a laser diode is generated by creating an electron-hole plasma in the vicinity of the forward-biased junction, a situation corresponding to population inversion in a conventional laser, and with sufficient forward current the material becomes transparent, in that the gain exactly equals the absorption and scattering losses, at a particular wavelength. When this arrangement is enclosed within a suitable optical cavity, and the forward current is increased further, the system begins to lase, that is to oscillate continuously to produce coherent optical radiation at a wavelength which satisfies two basic conditions: it must be within the set of wavelengths at which the material produces sufficient gain, and it must be close to one of the electromagnetic modes at which the optical cavity is resonant. These conditions, which define lasing threshold, are characterized by the equality of the optical gain and the total losses, that is the material losses (absorption, scattering, and free-carrier plasma) and the cavity losses which include the light output.

Early laser diodes in both the GaAs/AlGaAs and GaInAsP/InP systems were of the edge-emitting type, in which the light is emitted from the edge of the laser chip, perpendicular to the growth direction in the plane of the wafer. Although this geometry complicates laser production flow, particularly the testing function which requires that the wafer be scribed into bars to define the laser output facets, edge-emitting lasers (EELs) make up the vast majority of optical fiber transmission sources. An alternative laser diode structure is the vertical cavity surface-emitting laser (VCSEL), in which emission occurs in the growth direction perpendicular to the plane of the wafer. VCSELs offer major advantages in spectral stability, beam quality, manufacturability and cost, but their output powers are low (a few mW) and they are not yet available at the key telecommunication windows around 1300 nm and 1550 nm, due to difficulty in forming the necessary high-reflectivity Bragg mirrors in the GaInAsP/InP material system. Major initiatives are currently underway to produce long-wave VCSELs in the GaAs/AlGaAs system using InAs quantum dots, or dilute nitrides such as InGaAsN, or hybrid approaches such as InP-based gain media fused to high-reflectivity mirrors using GaAs/AlGaAs, dielectric coatings, air gaps, and others. At present, however, all long-haul optical fiber transmission is based on edge-emitting laser diodes fabricated in GaInAsP/InP.

General Structure and Requirements of Optical Fiber Communication Systems

The first optical fiber communication links were simple point-to-point affairs, consisting essentially of series-connected transmitters, fibers, and receivers. The transmitters were simple Fabry–Perot laser diodes, onto which data were encoded by direct digital modulation of the injection current at rates ~ 100 Mbit/s. The first commercial optical fiber link was built in 1976, a single fiber cable linking two switching centers of the Illinois Bell Telephone Co just 2.5 km apart in the Chicago metropolitan area, using 850 nm GaAs laser technology. In 1988, the first transatlantic fiber link, AT&T's TAT-8, was completed linking endpoints in New Jersey, England, and France with three fiber pairs carrying signals generated by 1300 nm GaInAsp single mode lasers. In these early systems, data were regenerated frequently by repeaters each consisting of a receiver, signal conditioner and transmitter. Each fiber in TAT-8 carried a single optical frequency modulated at 280 Mbit/s and the construction cost of the system was $\sim \$50$ k/km, resulting in an economic figure of merit of over $\$600$ k/Mbit/s. Twenty-five years later, current systems cost the same per kilometer to build in absolute dollars despite inflation, a major effective cost reduction enabled largely by replacement of the expensive repeaters by doped-fiber optical amplifiers. Moreover, each fiber is now highly multiplexed, with the potential for hundreds of wavelengths each carrying 10 Gbit/s signals generated by 1550 nm laser transmitters, so that the unit cost has plummeted to $\sim \$200$ /Mbit/s, a reduction of three and a half orders of magnitude.

In the remainder of this article, we will describe transmitters for both short- and long-haul systems where the channel spacing is only tens of GHz in the optical carrier frequency. Modern short-haul systems either feed client networks as part of bi-directional transceivers, or drive metro and local access networks connected directly by photonic switches to terminals of the long-haul systems.

DWDM Transceivers

Modern practice uses modular units containing transmitters and receivers allowing bi-directional data flow. In the send mode, data are applied to a long-haul transmitter at the local (client) source and coupled through the external fiber network to the remote long-haul receiver. In the receive mode, data arrive from a remote long-haul transmitter and are coupled (via a regenerator including signal conditioning, clock recovery, and error correcting steps) to a local short-haul transmitter leading to the client's internal network. Each transceiver card therefore contains a long-haul transmitter and receiver set (known collectively as dense wavelength division multiplexing (DWDM) transport devices) and a short-haul transmitter and receiver set (known as client interfaces or local transport devices).

Transmitter Requirements

Short-haul transmitters may be simple Fabry–Perot edge-emitting laser diodes, or vertical cavity laser diodes, or even superluminescent LEDs for sub-Gbit/s data rates. They are generally not required to be at one of the minimum-loss telecom windows near 1550 nm, as the transmission distances are short. However, as client networks grow into large metropolitan networks, it is likely that 1300 nm sources will be used to minimize dispersion. Power requirements are modest, a few mW peak, at data rates ~ 1 –10 Gbit/s. Single mode laser emission and polarization control are not required by present day systems, although polarization-mode dispersion may limit transmission at higher data rates. Both return-to-zero (RZ) and nonreturn-to-zero (NRZ) coding may be used, and direct modulation of the laser amplitude is the norm.

For long-haul transmitters the situation is dramatically different. Operation in one of the standard erbium-doped fiber amplifier (EDFA) bands near 1550 nm is mandatory: the C (conventional) band extends from 1530–1565 nm, the S (short-wave) band from 1460–1530 nm) and the L (long-wave) band from 1565–1625 nm. In the C-band, output powers ~ 10 mW are generally sufficient, and powers above ~ 30 mW are limited by nonlinear effects such as self-phase modulation and four-wave mixing in the fiber. In the S- and L-bands, where the available gain from EDFAs is less, lasers may need to operate close to the nonlinear limit, several tens of mW, although various schemes (differential pumping, gain-flattening filters, etc.) are being evaluated to flatten the global EDFA gain spectrum.

In current practice, a single fiber utilizing DWDM, using only the C-band, can carry up to 1 Tbit/s comprising 100 channels at 10 Gbit/s on each channel, the current standard. For such performance a channel spacing of 50 GHz in optical frequency is required. Using all three bands with similar density would enable ~ 3 Tbit/s, which could be doubled by using 25 GHz channel spacings. Ultimately, even higher channel densities and hence overall system data rates are possible, with a practical limit ~ 10 Tbit/s. The channel spacing is a key parameter in that it determines the laser design, specifically its spectral stability and linewidth, and the overall transmitter design in that laser chirp (modulation induced dynamic spectral shift) must be less than half the channel spacing. In practice, only single-wavelength and coarse wavelength division multiplexing (CWDM) systems may be modulated directly. All modern DWDM systems require external modulation, for example using electro-absorption or Mach–Zehnder interferometric devices outside the laser cavity. For the latter cases, stabilization of the laser wavelength is required for example by on-chip gratings to form distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, or by external fiber gratings.

Directly Modulated Lasers

Direct modulation of semiconductor lasers is convenient and effective: by modulating the laser injection current, one modulates the carrier density and hence the optical gain, resulting in modulation of the laser output up to ~ 10 Gbit/s. The actual modulation bandwidth is determined by interactions between photons and carriers, whose respective decay lifetimes, τ_s and τ_p , are determined by cavity losses and total recombination rates. The simplest form of such interactions is described by the photon and carrier rate equations for the injected carrier density N and photon density S in a single lasing mode:

$$dN/dt = J/ed - gNS - N/\tau_s \quad (1)$$

$$dP/dt = gNS + \beta N/\tau_s - S/\tau_p \quad (2)$$

where J is the injected current density, e the electronic charge, d is the thickness of the active region (hence J/ed is the rate of injection of carriers per unit volume), g is the gain coefficient per unit length and β is the fraction of the spontaneous emission coupled into the lasing mode, typically $\sim 10^{-2}$ – 10^{-5} for edge-emitting lasers (product of geometric and spectral overlap factors). These equations simply state that the rate of change of the carriers or photons is given by the rate of generation less the loss rate. Additional terms are required in the presence of optical feedback or coupling (in which case the photon density is replaced by the complex amplitude and phase of the optical electric field) or for extremely rapid modulation where the traveling wave nature of the disturbances in the photon fields is significant. These so-called traveling-wave rate equations for the photons are of the form:

$$dS_+/dt + cdS_+/dz = gNS_+ + \beta N/\tau_s \quad (3)$$

$$dS_-/dt - cdS_-/dz = gNS_- + \beta N/\tau_s \quad (4)$$

with the same carrier rate equation as [1] using the total photon number S made up of the forward-traveling component S_+ and the backward-traveling component S_- : $S = S_+ + S_-$.

Both sets of rate equations are based on approximations such as homogeneous gain broadening, neglect of transverse field effects, and diffusion. These are valid in most situations but care is required when analyzing lasers with large transverse apertures, or when ultrashort (picosecond) pulses are involved.

Small-Signal Modulation

Small-signal modulation may be analyzed by a linear perturbation analysis of the rate equations, leading to a conjugate pair of poles in the response function, with damped relaxation oscillations at frequency:

$$f_R = (1/2\pi)\sqrt{(gS_0/\tau_p)} \quad (5)$$

or, in terms of the injected current density J (assuming $S \sim G(J - J_{th})/ed$ with Γ the electromagnetic confinement factor of the mode):

$$f_R = (1/2\pi)\sqrt{[\Gamma g(J - J_{th})/ed]}$$

f_R can be regarded as the resonant frequency for the interaction between the carriers and photons. As a practical matter, lasers can be modulated up to $\sim 2f_R$ but the modulation response rolls off rapidly with increasing frequency above f_R . Using explicit expressions for the threshold gain, we can write f_R as:

$$f_R = (1/2\pi)\sqrt{[(GN_T g\tau_p + 1)(J/J_{th} - 1)/\tau_s\tau_p]} \quad (6)$$

where N_T is the transparency carrier density $\sim 10^{18} \text{ cm}^{-3}$. f_R can take values in excess of 10 GHz for a well-designed laser. It should be noted, however, that high modulation bandwidth almost always requires high differential gain dg/dN and short τ_p . Increasing dg/dN requires p-doping (which increases optical loss) and/or some special quantum-confining structure such as wells, wires, or dots. Decreasing τ_p inevitably leads to greater optical loss and thus a higher threshold.

The damping rate of the relaxation oscillations is also important in limiting modulation bandwidth, and also (when increased) in reducing the sensitivity of the laser to optical feedback. For bulk or quantum well lasers, the condition for critical damping is $A^2 = 4B$ with $A = (S/g\tau_s - \sigma[N - N_T])/g\tau_p$ and $B = \sigma[1 + (S_0/g\tau_s) - (1 - \beta)(N - N_T)]$. For quantum dot lasers there are additional damping terms due to carrier transport and thermalization which may restrict small-signal modulation bandwidths to a few GHz.

Experimentally, small-signal modulation properties may be determined using a simple sampling oscilloscope, low-noise tunable signal generator and spectrum analyzer. A combination of these elements with automated frequency sweeping may be found in a scalar network analyzer acting as an s -parameter test set. The modulated laser is connected to port 1 and a high speed photodetector connected to port 2, then a swept-frequency measurement of the transfer characteristic s_{21} is performed. In addition to the limitations on modulation response due to the carrier-photon resonance, practical lasers are limited by RC parasitics in the laser chip (junction impedance) and its package. For the highest modulation speeds, packages need to be designed as microwave transmission line components with effective impedance matching and low back reflections, as characterized by the voltage standing wave ratio (VSWR).

Large-Signal Modulation

Large-signal direct modulation can be analyzed by numerical integration of the rate equations. In general, its results are beyond the scope of this review, but for effective high-speed response, digital modulation of laser diodes should be carried out with a pre-bias close to threshold. Modulation bandwidth also increases with laser power, which is limited for systems considerations by fiber nonlinearities, and for reasons of laser reliability. In pulse code modulation, care must be taken that the laser is near critical damping, to minimize thermal patterning effects due to long strings of ones (high power) or zeros (low power). Experimentally, large signal modulation is analyzed by constructing eye diagrams, in which pulse traces on a fast oscilloscope are continuously overlaid when the laser is modulated using pseudorandom binary pulses. When the signal quality is high, the high and low digital levels are easily distinguishable, resulting in an open 'eye' in the accumulated traces.

Requirements for Externally Modulated Lasers

The requirement for external modulation occurs when the laser chirp exceeds half the desired channel spacing in the communication system. Chirp occurs due to changes in the refractive index, and hence the optical phase, during modulation. This dynamic phase shift then results in an instantaneous frequency shift. For direct modulation we have instantaneous wavelength

$$\lambda(t) = \lambda(0)\mu(t)/\mu(0) \quad (7)$$

Where $\mu(t)$ is the refractive index at time t . μ shifts with carrier density N due to combinations of several factors (free carrier plasma dispersion, thermal bandgap shrinkage, bandgap renormalization, dynamic Burstein–Moss effects) which give $d\mu/dN \sim -3 \times 10^{-20} \text{ cm}^3$, which for large signal digital modulation can result in tens of GHz. The actual degree of chirp depends on the so-called antiguiding or linewidth enhancement factor α given by

$$\alpha = -2k_0(d\mu/dN)/(dg/dN) \quad (8)$$

where $k_0 = 2\pi/\lambda_0$ is the free-space wavenumber. α as defined is positive, typically in the range 2–7 depending on the material, operating carrier density, and wavelength relative to the gain peak. Distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, which use integrated gratings to control the lasing wavelength, can have lower chirp by detuning from the gain peak but therefore are likely to require careful temperature stabilization. Quantum dot lasers, based on 3D nanoclustered active regions, have the potential for very low chirp, narrow linewidths, and reduced wavelength shifts, so that directly modulated DWDM transmitters may be possible.

Integration of laser diodes with electro-absorption or Mach–Zehnder type modulators is achieved in the transmitter optical subassembly and may in some cases be accomplished by monolithic integration on the same chip. Using separate modulators allows each device to be optimized independently but requires optical assembly, alignment, and retention. Mach–Zehnder modulators in X-cut lithium niobate, for example, have essentially zero chirp and can operate to 40 Gbit/s and above, with dynamic extinction ratios (ratio between fully on and fully off) of ~ 20 dB. Integrated semiconductor M-Z modulators have chirp an order of magnitude less than those of directly modulated lasers, so that channel spacings ~ 25 GHz are possible in DWDM systems with suitable temperature and wavelength controls. Although the details of such modulators are beyond the scope of this article, the laser requirements are to generate ~ 10 mW of continuous power with stable center wavelength and narrow linewidth. Such lasers are usually mounted in hermetic packages (e.g., the current 14-pin butterfly standard) with integrated thermo-electric cooler, power monitor photodiode, an optical isolator to suppress optical backreflections which cause instabilities and self-pulsing, and optional wavelength locking optics.

Reliability of Lasers in Fiber Optic Systems

From the early days, when laser diode lifetimes were measured in seconds even at cryogenic temperature, enormous progress has been made in achieving materials purity and reducing crystalline defects. Today's fiber optic laser transmitters have projected lifetimes of decades. In common with electronic devices, diode lasers fail at a rate given by a 'bathtub' curve, that is the failure rate $r(t)$ defined as the probability of failure per unit time at time t , has relatively high values at low t (early failures) and high t (wearout failures) and very low values in between. In terms of the population of lasers $n(t)$ we have:

$$r(t) = (-1/n(t))dn(t)/dt \quad (9)$$

If Δn is the number of samples which fail in time Δt , then assuming Δt begins at $t=0$, the effective or average failure rate r_{eff} over the interval is

$$r_{\text{eff}}(t) = \Delta t \Delta n / n(0) \quad (10)$$

In reliability science it is customary to define r_{eff} in FITs (failures integrated in time) with the time-span chosen to yield statistically significant numbers. For electronic and photonic devices, it is customary to select the time interval $\Delta t = 10^9 \text{ h}$ (1 billion operating hours), so that if 1% of the devices fail in 10 years ($\sim 10^5$ hours) we have $r_{\text{eff}} \sim 100$ FITs, which is the order of magnitude required by modern fiber optic laser sources.

In terms of actual statistical models, failure rates can be estimated using failure probability density functions $f(t)$. This is related to $r(t)$ by $f(t) = r(t)S(t)$ where $S(t)$ is the cumulative probability of surviving until t . When $r(t)$ is nearly constant, as in early failures due to material defects or process errors, the statistics are approximately exponential:

$$f(t) + (1/\tau)\exp(-t/\tau) \quad (11)$$

or Weibull-distributed (an exponential is a Weibull distribution with $\beta=1$):

$$f(t) = (\beta/t^\beta)t^{\beta-1}\exp\left[-(t/\tau)^\beta\right] \quad (12)$$

but this description is not suited to wearout failures for which $f(t)$ shows a significant increase as the population ages. For such cases, the lognormal distribution is often applicable:

$$f(t) = (1/\sigma t \sqrt{2\pi})\exp[-(\ln t - \ln \tau)^2/2\sigma^2] \quad (13)$$

where in each case τ is approximately the mean time to failure (MTTF).

Laser diodes undergo standardized qualification tests to demonstrate sufficient reliability prior to being used in commercial systems. Operationally, failure statistics must be tabulated and the best fit obtained. Burn-in procedures are used to screen early failures. Chip and module failures should be distinguished clearly, as the latter includes many additional factors such as thermal and power management, light coupling, mechanical or chemical integrity.

Given that effective lifetimes of decades are required, it is clearly necessary to accelerate aging to produce statistically meaningful failure rates in reasonable times $\sim 10^3$ h. For the most common modes of laser chip degradation, due to recombination-induced aggregation of defects in the crystalline epitaxial material, these are thermally activated and current driven, so that it is customary to use a modified Arrhenius law at temperature T :

$$\tau \sim (1/J^n)\exp(E_a/k_B T) \quad (14)$$

where J is the operating current density, n is an index ~ 2 , E_a is the activation energy, and k_B is Boltzmann's constant. In terms of thermal acceleration, we have a factor $F = \tau(T_A)/\tau(T_B)$ with:

$$F = \exp[(E_a/k_B)(1/T_A - 1/T_B)] \quad (15)$$

Typical values of E_a for laser diodes are ~ 1 eV so that acceleration factors $\sim 10^3$ are possible.

Finally, it should be noted that laser diodes are generally subject to catastrophic failure in the event of overdriving or static discharge, the failure mode being thermal facet damage at ~ 1 – 10 MW/cm² for continuous-wave operation, the exact value depending on the material, surface preparation and specifically surface state absorption and its temperature coefficient.

Conclusions

Laser diodes are ideal sources for optical fiber communication systems and have propelled the development of fiber optics from its origins in the 1960s to the present day. They are compact, rugged, efficient, and reliable sources of light at key wavelength ranges such as 1300–1310 nm (short haul high speed links) and 1500–1600 nm (long haul amplified systems). They are capable of direct modulation at gigabit rates for simple systems but require external modulation and careful wavelength control for dense wavelength division multiplexed terabit systems. Transmitter lasers for real systems must satisfy stringent reliability conditions.

See also: Basic Concepts of Optical Amplifiers

Further Reading

- Adams, M.J., Thomas, B., 1977. Detailed calculations of transient effects in semiconductor injection lasers. *IEEE Journal of Quantum Electronics* QE-13, 580.
- Agrawal, G.P., 2002. *Fiber Optic Communication Systems*, 3rd edn. New York: Wiley.
- Alferov, Z.I., Andreev, V.M., Portnoi, E.L., Trukan, M.K., 1970. AlAs-GaAs heterojunction injection lasers with a low room-temperature threshold. *Soviet Physics – Semiconductors* 3, 1107.
- Basov, N.G., Krokhin, O.N., Popov, Y.M., 1961. Production of negative-temperature states in p-n junctions of degenerate semiconductors. *Soviet Physics – JETP* 13, 1320.
- Bernard, M.G.A., Duraffourg, G., 1961. Laser conditions in semiconductors. *Physica Status Solidi* 1, 699.
- Bhattacharya, P., 1966. *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.
- Cheng, J., Dutta, N.K. (Eds.), 2000. *Vertical Cavity Surface-Emitting Semiconductor Lasers: Technology and Applications*. London: Taylor & Francis.
- Chuang, S.-L., 1995. *Physics of Optoelectronic Devices*. New York: Wiley.
- Coldren, L.A., Corzine, S.W., 1995. *Laser Diodes and Photonic Integrated Circuits*. New York: Wiley.
- Dupuis, R.D., 1987. An introduction to the development of the semiconductor laser. *IEEE Journal of Quantum Electronics* QE-23, 651.
- Fukuda, M., 1991. *Reliability and Degradation of Semiconductor Lasers and LEDs*. Norwood, MA: Artech House.
- Fukuda, M., 1998. *Optical Semiconductor Devices*. Wiley.
- Guekos, G. (Ed.), 1998. *Photonic Devices for Telecommunications*. New York: Springer.
- Hall, R.N., Fenner, G.E., Kingsley, J.D., Soltys, T.J., Carlson, R.O., 1962. Coherent light emission from GaAs junctions. *Physical Review Letters* 9, 366.
- Hayashi, I., Panish, M.B., Foy, P.W., Sumski, S., 1970. Junction lasers which operate continuously at room temperature. *Applied Physics Letters* 17, 109.
- Holonyak Jr, N., Bevacqua, S.F., 1962. Coherent (visible) light emission from Ga(As_{1-x}P_x) junctions. *Applied Physics Letters* 1, 82.

- Ito, R., Nakashima, H., Kishino, S., Nakada, O., 1975. Degradation sources in GaAs-AlGaAs double-heterostructure lasers. *IEEE Journal of Quantum Electronics* QE-11, 551.
- Kaminow, I.P., Koch, T.L. (Eds.), 1997. *Optical Fiber Telecommunications III* (two parts). London: Academic Press.
- Kaminow, I.P., Li, T. (Eds.), 2002. *Optical Fiber Telecommunications IV* (two parts). Amsterdam: Elsevier.
- Kapon, E., 1998. *Semiconductor Lasers*, (two parts). London: Academic.
- Kartalopoulos, S.V., 1999. *Introduction to DWDM Technology*. New York: Wiley.
- Kartalopoulos, S.V., 1999. *Understanding SONET/SDH and ATM*. IEEE Press.
- Kasap, S.O., 2001. *Optoelectronics and Photonics: Principles and Practices*. Englewood Cliffs, NJ: Prentice-Hall.
- Kazovsky, L., Benedetto, S., Willner, A., 1996. *Optical Fiber Communication Systems*. Norwood, MA: Artech House.
- Keiser, G., 1991. *Optical Fiber Communications*. New York: McGraw-Hill.
- Kressel, H., Lockwood, H.F., 1974. A review of gradual degradation phenomena in electroluminescent diodes. *Journal de Physique* 35, C3-223.
- Lasher, G., Stern, F., 1964. Spontaneous and stimulated recombination radiation in semiconductors. *Physical Review* 133, A553.
- Milonni, P.W., Eberly, J.H., 1988. *Lasers*. New York: Wiley.
- Morthier, G., Vankwikelberge, P., 1997. *Handbook of Distributed Feedback Laser Diodes*. Norwood, MA: Artech House.
- Nathan, M.I., Dumke, W.P., Burns, G., Dill Jr, F.H., Lasher, G., 1962. Stimulated emission of radiation from GaAs p-n junctions. *Applied Physics Letters* 1, 62.
- Numai, T., 2004. *Fundamentals of Semiconductor Lasers*. New York: Springer.
- Paoli, T.L., 1977. Changes in the optical properties of cw (AlGa) as junction lasers during accelerated aging. *IEEE Journal of Quantum Electronics* QE-13, 351.
- Paoli, T.L., Ripper, T.E., 1970. Direct modulation of semiconductor lasers. *Proceedings of IEEE* 58, 1457.
- Pearsall, T.P., 2003. *Photonics Essentials: An Introduction with Experiments*. New York: McGraw-Hill.
- Piprek, J., 2003. *Semiconductor Optoelectronic Devices: Introduction to Physics and Simulation*. London: Academic.
- Quist, T.M., Rediker, R.H., Keyes, R.J., Krag, W.E., Lax, B., McWhorter, A.L., Zeiger, H.J., 1962. Semiconductor maser of GaAs. *Applied Physics Letters* 1, 91.
- Reinhart, F.K., Hayashi, I., Panish, M.B., 1971. Mode reflectivity and waveguide properties of double-heterostructure injection. *Lasers* 42, 4466.
- Saleh, B.E.A., Teich, M.C., 1991. *Fundamentals of Photonics*. New York: Wiley.
- Senior, J.M., 1992. *Optical Fiber Communications*, 2nd edn. Englewood cliffs, NJ: Prentice-Hall.
- Siegman, A.E., 1986. *Lasers*. University Science Books.
- Telcordia Technologies (formerly Bellcore): document family FR-796 on Reliability and Quality Generic Requirements for telecommunications equipment. Latest issue is 003 dated Jul 2003. The specific document related to laser reliability is GR-468 – Generic Reliability Assurance Requirements for Optoelectronic Devices used in Telecommunications Equipment. Standards for other network photonic, electronic and mechanical components may be found in the document family FR-2063 Network Equipment-Building System Family of Requirements, issue 003 dated Feb 2003 and specifically FR-357 – Network Equipment-Building System Family of Requirements: Component Design for Reliability.
- Verdeyen, J.T., 1994. *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
- Yariv, A., 1989. *Quantum Electronics*, 3rd edn. New York: Wiley.
- Zachos, T.H., Ripper, J.E., 1969. Resonant modes of GaAs junction lasers. *IEEE Journal of Quantum Electronics* QE-5, 29.

Broadband Passive Optical Access Networks

Elaine Wong and Maluge P Imali Dias, The University of Melbourne, Parkville, VIC, Australia
Zhengxuan Li and Lilin Yi, Shanghai Jiao Tong University, Shanghai, P.R. China

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The exponential growth in Internet and mobile traffic along with bandwidth-intensive applications, has continued to drive the deployment of fiber networks deeper into the access network segment (Wong, 2012). In future-proofing against the forecasted increase in bandwidth demands, Fiber-To-The-X (X=home, curb, building) networks have been deployed in various parts of the world. Offering direct fiber connection to or close to the home, FTTx deployments can assume point-to-multipoint (PtMP) or point-to-point (PtP) physical topologies, as illustrated in Fig. 1 (Wong, 2012). Driven by low cost and low energy consumption per bit, most deployed FTTx models are based on the passive optical network (PON) topology (Wong, 2012). Typically, a PON has a PtMP topology and comprises a central office (CO) which houses multiple optical line terminals (OLTs), optical network units (ONUs) residing in subscriber premises, and a passive splitting optical distribution network (ODN) consisting of a feeder fiber, passive optical splitter(s), and multiple distribution fibers.

Commercial PONs such as Broadband PON (BPON) (ITU-T, 2005), Gigabit PON (GPON) (ITU-T, 2008a) and Gigabit Ethernet PON (GE-PON) (IEEE, 2004) are deployed around the world. GPON systems are specified by the ITU-T G.984 Gigabit PON (GPON) standard (ITU-T, 2008a) whereas GE-PON is specified by the IEEE 802.3ah Gigabit Ethernet PON (GE-PON) standard (IEEE, 2004). Both GPON and GE-PON are classified as time division multiplexed/time division multiple access (TDM/TDMA) PONs. In TDM/TDMA PONs, encrypted information is broadcast to all ONUs in timeslots at a line-rate of 2.5 Gb/s for GPON and 1.25 Gb/s for GE-PON in the downstream direction. Due to the power-splitting ODN, all time slots are received by each ONU but only particular time slots addressed to the ONU through its media access control (MAC) destination address, are detected and received. In the upstream direction, burst mode TDMA is used between ONUs to share the aggregate 1.25 Gb/s upstream bandwidth (for both GPON and GE-PON). If required, a dedicated radio frequency (RF) channel can be added to broadcast TV/video data (Wong, 2012).

One major disadvantage of the power-splitting nature PONs is the limitation in further scaling of bandwidth, reach, and subscriber count. Specifically, the splitting loss of the ODN is a strong function of the number of supported subscribers thereby constraining any possible increase in user count, reach, and/or average user data rate. To overcome this limitation and to support future business, residential, back- and front-hauling bandwidth needs, next-generation PONs commonly referred to as NG-PON1, with a maximum line rate of 10 Gb/s, were recently standardized. Defined by both IEEE (2009) and ITU-T (2010), these standards allowed for backward compatibility and co-existence with the current-generation GPONs and GE-PONs, thus enabling progressive upgrades with minimal capital investment on the external plant and operational impact on existing users.

The PtP network can overcome the limitation in bandwidth, reach, and subscriber count of power-splitting PONs. In a PtP network, dedicated fiber and hence bandwidth is assigned between the CO and each ONU. Other beneficial features include enhanced data privacy and security. Nonetheless, PtP suffers from high fiber count and terminations at the CO (one per ONU) and thus requires high density fiber management. While the focus of this article is on PONs that adopt the PtMP topology, PtP Ethernet

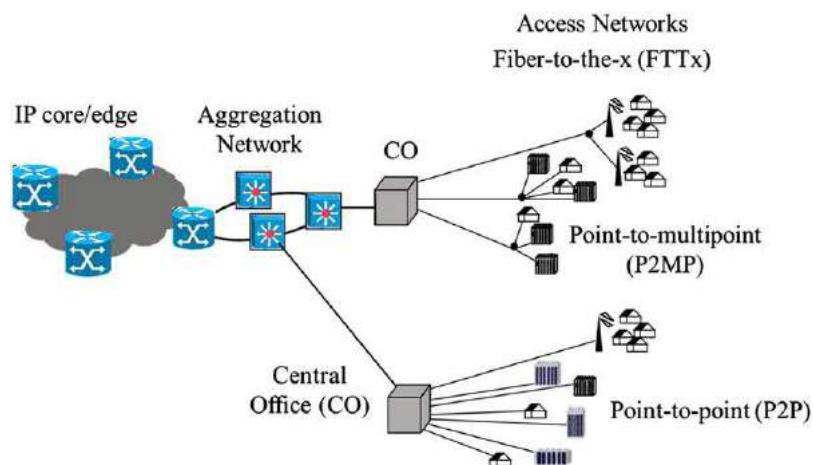


Fig. 1 Schematic of network structure showing the physical topology options of access networks. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. *IEEE/OSA J. Lightw. Technol.* 30(4), 597–608.

is popular in some parts of the world, e.g., Japan, China and Sweden, and has been deployed to connect high revenue fixed access and mobile business customers that require guaranteed bandwidths (Förster, 2006).

New access technologies beyond NG-PON 1, namely 10 Gb/s TDM/TDMA systems, are being explored to support symmetrical average data rates of ~ 1 Gb/s per user, an extended system reach of 60–100 km, a high subscriber count of up to 1000, and heterogeneous service convergence, while meeting the cost constraints of the access market. The standardization of time and wavelength division multiplexed PON (TWDM-PON) as the selected NG-PON2 technology, specifies a 40-Gb/s downstream capacity using an underlying hybrid time and wavelength division multiplexing scheme (ITU, 2013). Each wavelength in a TWDM-PON is shared between multiple ONUs by employing time division multiplexing and multiple access mechanisms.

Beyond the promises of NG-PON2, wavelength division multiplexed (WDM) PONs are being explored as the next-generation solution. The idea of exploiting wavelength division multiplexing in access networks was conceived in the late 1980s (Oakley, 1988) but significant advancement of key enabling technologies such as the athermal arrayed waveguide grating (AWG), WDM filter, reflective semiconductor optical amplifier (RSOA) and Fabry-Perot laser diode (FP-LD) has fuelled the resurgence of WDM-PON research in the last 10 years. A number of commercialized WDM-PON systems have been deployed to service customized business and wireless/wireline backhaul markets, e.g., (see Section Relevant Websites). In a WDM-PON, each end user is given dedicated capacity through the assignment of unique upstream and downstream wavelength channels. Concurrently, extended-reach PONs (ER-PONs) are being explored to consolidate metropolitan area and access networks (Davey *et al.*, 2009). Serving an increased number of clients, accommodating high bandwidth applications, and an increased network span, ER-PON deployments promise lower capital expenditure (CAPEX) and operational expenditure (OPEX) due to reduced number of active network interfaces and elements in the field. More recently, the use of orthogonal frequency division multiplexing (OFDM) in combination with WDM has been vigorously explored for application in the access segment (Cvijetic, 2012). In an OFDM-PON, data is carried by multiple orthogonal subcarriers, and a flexible 3D bandwidth allocation can be realized in time, frequency and modulation formats domain.

This review article is organized as follows. In Section Time-Division Multiplexed (TDM) Passive Optical Networks, we will introduce the early standardized TDM-PONs including BPONs, GPONs and GE-PONs. Then, the standardized 10G-PONs including ITU-T NG-PON1 and IEEE 10 G-EPON will be introduced in Section NG-PON1. In Section NG-PON2, the TWDM-PON structure based NG-PON2 will be introduced from the viewpoint of both physical layer and protocol layer. High capacity TWDM-PONs with advanced techniques to enable 100 Gb/s or beyond capacity will also be discussed. In Section Future PONs, future PONs with potential standardization possibility including WDM-PONs, extended-reach PONs and OFDM-PONs will be discussed. Finally, we will conclude this article in Section Summary and Conclusion.

Time-Division Multiplexed Passive Optical Networks

BPONs

The Broadband Passive Optical Network (BPON) was firstly developed by the Full Service Access Network (FSAN) working group in 1995 and was standardized by the International Telecommunications Union (ITU) in 2001 (ITU-T, 2008a). BPON is based on the asynchronous transfer mode (ATM) protocol and has a maximal downstream rate of 1244.16-Mbit/s and upstream rate of 622.08-Mbit/s. As discussed in Section Introduction, downstream signal is broadcast to all ONUs on the PON. To avoid collisions in the upstream direction, upstream transmission time slots from each ONU are controlled and granted by the OLT in the CO through Time Division Multiple Access (TDMA) (ITU-T, 2008a). The line code employed by BPON in both downstream and upstream transmissions is the Non-Return Zero (NRZ) modulation format. The typical operating wavelength for downstream and upstream is 1550 nm and 1310 nm, respectively.

The BPON MAC frame consists of different numbers of fixed size ATM cells, the actual number depending on the rate of transmission, as shown in Fig. 2. For the downstream MAC frame, a Physical Layer Operation, Administration, and Maintenance (PLOAM) is inserted at the beginning of every 28 ATM cells. PLOAMs are also used to carry grant information relating to an ONU's upstream access. Through PLOAM, an OLT in the CO can allocate and divide the upstream access bandwidth amongst different ONUs. Downstream transmission from the OLT is broadcast to every connected ONU. Each ONU will selectively detect its own cell according to the bandwidth allocation scheduled in OLT. In the upstream direction, transmission occurs in ATM cells with 3 overhead bytes per cell in its own time slot. Each ONU can only transmit its ATM cells during pre-determined time slots. Due to ONUs not being equidistance to the OLT, the OLT will receive burst-mode signals at differing power levels, each from a different ONU.

Though being the first-standardized PON, the uptake of BPON was slow and is currently rarely deployed due to the high cost per unit bandwidth. Further, the use of ATM cells within this MAC frame results in complex processing and large overhead (a 5-byte header to each 48-byte payload), thus limiting large-scale deployment.

Gigabit PONs

The Gigabit PON (GPON) was first standardised by the FSAN/ITU-T under ITU-T G.984 listing different aspects of the GPON such as general characteristics, physical medium, and management and control (ITU-T, 2008a). The standardized GPON supports data

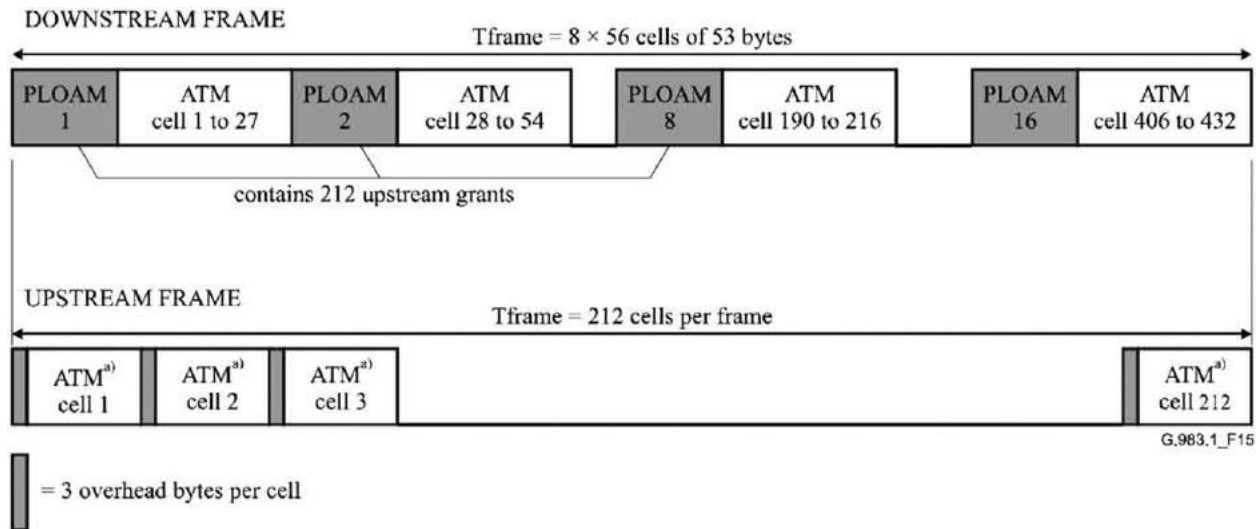


Fig. 2 Frame structure of BPON for 1244.16/622.08-Mbit/s. Adapted from ITU-T, 2008a. G.984 – Series recommendation, Gigabit-capable passive optical networks (G-PON).

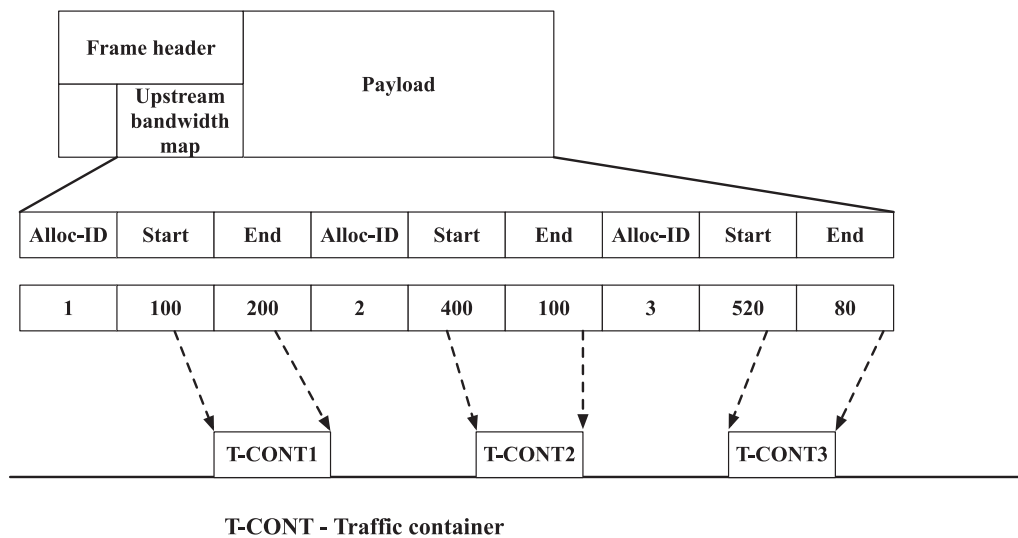


Fig. 3 Downstream GPON packet structure. Adapted from CommScope, 2013. EPON-GPON comparison, CommScope Solutions Marketing.

rates of 2.5 Gb/s and 1.25 Gb/s in the downstream and upstream directions, respectively. Today, GPON is one of the most popular technology of choice in some areas of Europe and North America.

Before entering the physical medium, Ethernet frames are encapsulated into GPON Transmission Control (GTC) layer encapsulation method (GEM) frames. GEM frames are, in turn, encapsulated into GPON Transmission Convergence (GTC) frame (in both upstream and downstream directions) (CommScope, 2013). Fig. 3 illustrates a downstream GTC frame in detail (CommScope, 2013). In GPON, the GTC layer is responsible for aggregating traffic generated from different services, e.g., voice and video, into a common service-independent framework. The GTC header carries an upstream bandwidth map, which includes information on transmission time slots allocated to each ONU. In the GPON, the bandwidth grant is per traffic container (T-CONT). A T-CONT is an ONU object representing a group of logical connections that appear as a single entity for the purpose of upstream bandwidth assignment on the PON. For example, a GPON may have multiple T-CONTs per ONU, each dedicated for a different traffic class, e.g., best-effort and high priority. The allocation ID (Alloc-ID) is used to identify uniquely a T-CONT.

Fig. 4 illustrates the upstream frame structure of GPON (ITU-T, 2008b). The frame length is as same as that of the downstream, 125 μ s and 19,440 bytes long. Each frame contains a number of transmissions from one or multiple ONUs, where the bandwidth map discussed under downstream frame structure schedules the arrangement of these transmissions. In any given upstream transmission, the ONU can transmit one of the four types of overheads and the payload. The four types of overheads are physical

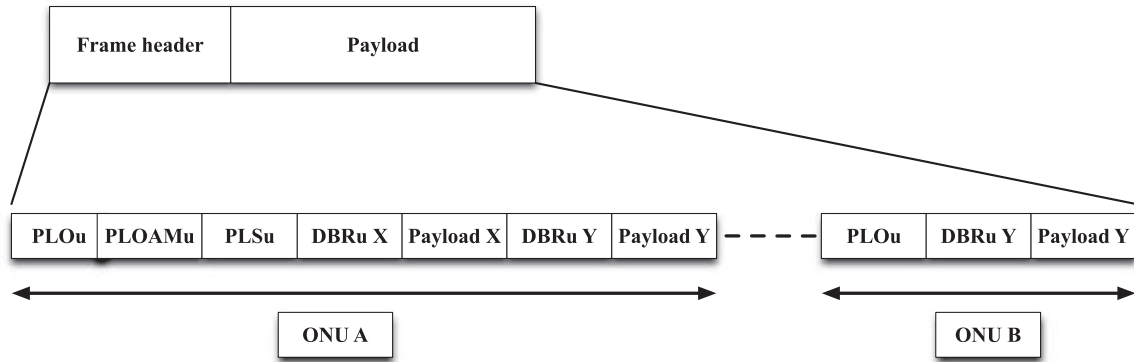


Fig. 4 Upstream GPON packet structure. Adapted from ITU-T, 2008b. G.984 x-series recommendations: Gigabit-capable passive optical networks (G-PON): Transmission convergence layer specification (G.984.3).

layer overhead (PLOu), physical layer operations administrations and management upstream (PLOAMu), power levelling sequence upstream (PLSu), and dynamic bandwidth report upstream (DBRu). The ONUs utilize the dynamic bandwidth report (DBRu) field, located in the header of the upstream GTC frame, to inform the queue length of each T-CONT to the OLT.

Gigabit Ethernet PONs

The IEEE standardized the Gigabit Ethernet PON (GE-PON) under IEEE 802.3 ah (IEEE, 2004). Integrated with conventional Ethernet technology, GE-PON enables seamless integration with IP and Ethernet technologies. The GE-PON supports 1.25 Gb/s link rate in both upstream and downstream directions and a 20 km network span. The deployment of GE-PON is most prevalent in the Asian region, especially in countries such as Japan and Korea.

In GE-PON, Ethernet packets are transmitted natively across the PON. As Ethernet technology is widely deployed in current local area networks (LANs), incorporating Ethernet into PONs requires minimal protocol conversion as far as hardware and technical know-how are concerned. As a TDM-PON, the ONUs in a GE-PON follow a time division multiple access technique, where the upstream bandwidth is divided into distinct time slots. In order to obtain bandwidth requirement information as well as to grant bandwidth from and to each ONU, a GE-PON employs the Multi-Point Control Protocol (MPCP). MPCP relies on control messages to facilitate two modes of operation: auto-discovery and normal operation. The primary MPCP control messages used for each mode of operation are listed below:

- Auto-discovery: GATE, REGISTER REQ, REGISTER, and REGISTER ACK.
- Normal operation: GATE and REPORT.

Auto-discovery

Fig. 5 illustrates the auto-discovery process, an initialization process implemented at both OLT and ONUs. Using auto-discovery, the OLT identifies the MAC addresses and round trip times of ONUs that have recently joined the network (IEEE:802.3ah). For this purpose, MPCP uses four control messages, GATE, REGISTER REQ, REGISTER, and REGISTER ACK. When an OLT discovers new ONUs, it allocates a discovery window, where no other registered ONU can transmit. A discovery GATE message with a local OLT time stamp will be broadcast, which only the uninitialized ONUs will respond to. When an ONU clock reaches the start time of the discovery window, it will wait a random amount of time and send its REGISTER REQ to the OLT. The random delay minimizes potential collisions caused by multiple uninitialized ONUs simultaneously sending REGISTER REQ messages. The REGISTER REQ message contains the MAC address and the local time stamp of the ONU, through which the OLT learns the round trip time to the ONU. Once the ONU is registered, the OLT will assign a unique identification, link layer ID (LLID), to the ONU. The OLT will then send the unique LLID through REGISTER to the ONU. This is then followed by a normal GATE message. Upon receiving the REGISTER and normal GATE messages, the ONU sends the REGISTER ACK to the OLT during the time slot allocated through the normal GATE message.

Normal operation

Once ONUs have established their connection with the OLT, bandwidth requests from and allocations to each ONU are performed under normal operation mode, as illustrated in **Fig. 6** (Kramer, 2005). The GATE message, sent from the OLT to ONUs, is used to grant bandwidth (in the form of transmission time slots) to the ONUs. The GATE specifies the start time and the length of a transmission window and is sent per LLID. As discussed, an LLID which is attached to the preamble of the downstream packets, uniquely identifies an ONU in the network. The REPORT message, sent from each ONU to the OLT, contains bandwidth requirements of the ONUs. It is important to note that the standardized MPCP is not unique to a particular bandwidth allocation algorithm, rather, it is a supporting protocol that encourages the execution of different bandwidth allocation schemes in the GE-PON depending on latency, energy-efficiency, and capacity requirements of the network/service provider.

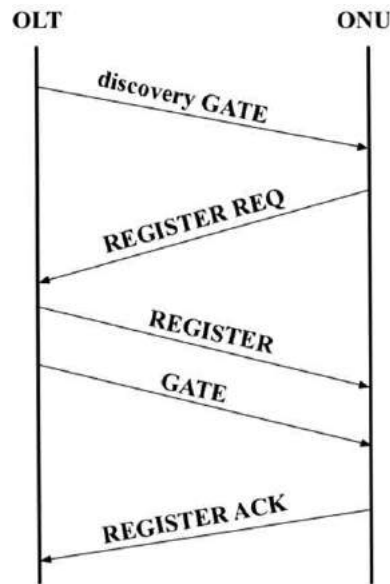


Fig. 5 Auto-discovery process of MPCP. IEEE:802.3ah, Ethernet in the first mile task force.

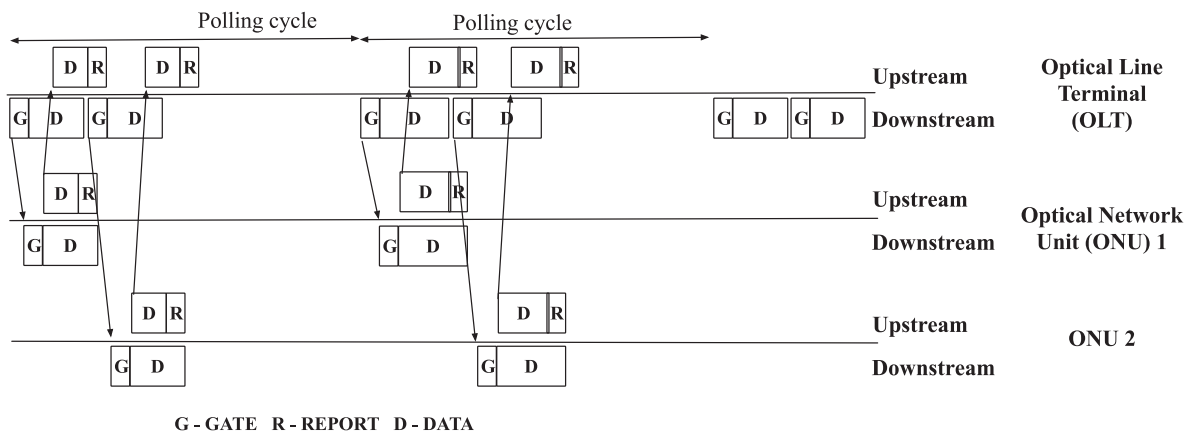


Fig. 6 Normal operation of MPCP. Reproduced from Kramer, G., 2005. Ethernet Passive Optical Networks. New York, NY: McGraw-Hill.

NG-PON1

Overview of 10GE-PON and XG-PON

The IEEE and the ITU-T with the Full Services Access Network (FSAN) group have defined their respective 10 Gb/s solution to address future bandwidth growth over existing ODNs that exploit time-division multiplexing (TDM) in the downstream direction and time division multiple access (TDMA) in the upstream direction. These solutions are specified in the IEEE Std. 802.3av 10 GE-PON (IEEE, 2009) and ITU-T XG-PON (ITU-T, 2010) standards respectively. The 10 GE-PON defines symmetric (10.3 Gb/s) and asymmetric (10.3 Gb/s downstream and 1.25 Gb/s upstream) line-rate operations, the latter to support asymmetric traffic of IP video services. In contrast, XG-PON1 supports an asymmetrical bandwidth capacity of 10 Gb/s downstream and 2.5 Gb/s upstream (ITU-T, 2010). In XG-PON1, new features including enhancing security through authentication of management messages and minimizing energy consumption through powering down parts or all of the ONU are defined. Specific solutions to minimize the energy consumption include (a) powering down user network interfaces that are not actively in use (b) operating in “doze” mode whereby the ONU transmitter is powered down when the user has no real data to send, and (c) operating in “sleep” mode whereby both transmitter and receiver are powered down when the ONU is idle. A symmetric version of XG-PON1, known as XG-PON2 with 10 Gb/s downstream/upstream capacity is also specified (ITU-T, 2010).

Forward error correction (FEC) is defined in both standards to compensate for the reduced receiver sensitivity from using 10 Gb/s optical receivers. NRZ line coding in conjunction with FEC ((RS(248,216) block code) is mandatory in the downstream

direction for XG-PON systems but optional in the upstream direction (ITU-T, 2010). For 10 GE-PON, FEC ((RS(255,223) block code) is mandatory for the symmetric network. As for the asymmetric 10 GE-PON network, the upstream 1 Gb/s links can optionally use the IEEE GE-PON FEC (IEEE, 2009). To enhance coding efficiency in 10 GE-PON, 64b/66b line coding is used instead of the conventional 8b/10b coding in GE-PON.

Coexistence and Wavelength Allocation

Table 1 summarizes the main characteristics of currently standardized TDM-PONs discussed in Sections Time-Division Multiplexed (TDM) Passive Optical Networks and NG-PON1. Coexistence between these different standards over the same ODN can be achieved by observing the wavelength allocation schematically shown in **Fig. 7**. In the figure, the upstream and downstream wavelength bands for XG-PON, 10 GE-PON, GPON, GE-PON, and RF video overlay are indicated. Specifically, GPON and GE-PON use 1480–1500 nm for downstream transmission and 1550–1560 nm for RF video overlay (ITU-T, 2008a; IEEE, 2004). For upstream transmission, GPON uses the 1290–1330 nm waveband whereas GE-PON uses the entire O band. Both XG-PON and 10 GE-PON specify the O-minus band (1260–1280 nm) for upstream transmission and the L-band (1575–1580 nm) for downstream transmission (IEEE, 2009; ITU-T, 2010), thus facilitating coexistence with legacy systems and those implemented with RF video overlay.

NG-PON2

Overview of Time and Wavelength Division Multiplexed-PON

The standardization process of NG-PON2 was implemented by ITU-T study group 15 (SG15) (ITU, 2013). To meet the general requirement of 40-Gb/s downstream capacity, an underlying hybrid time and wavelength division multiplexing scheme is adopted in NG-PON2, giving rise to the time and wavelength division multiplexing PON (TWDM-PON). **Fig. 8** depicts a typical schematic diagram of a TWDM-PON system (Luo *et al.*, 2013). Several wavelength pairs are multiplexed into a single PON system, whereby each wavelength is shared between multiple ONUs by employing time division multiplexing and multiple access mechanisms. NG-PON2 provides at least 40/10-Gb/s downstream/upstream capacity by stacking 4 TDM channel pairs with

Table 1 Comparison of standardized TDM-PONs

	<i>BPON</i>	<i>GPON</i>	<i>GE-PON</i>	<i>XG-PON1</i>	<i>10GE-PON</i>
Standard	ITU-T G.983	ITU-T G.984	IEEE803.2ah	ITU-T XGPON	IEEE 802.3av
Bandwidth					
Down	622 Mb/s	upto 2.5 Gb/s	Symmetric up to 1.25 Gb/s	10.3 Gb/s	10 Gb/s
Up	Up - 155 Mb/s	upto 2.5 Gb/s		10.3 Gb/s	1.25 Gb/s up to 10 Gb/s
Downstream, λ_d	1490 and 1550 nm	1490 nm and 1550 nm	1490 nm	1575–1580 nm	1575–1580 nm
Upstream, λ_u	1310 nm	1310 nm	1310 nm	1270 nm	1270 nm

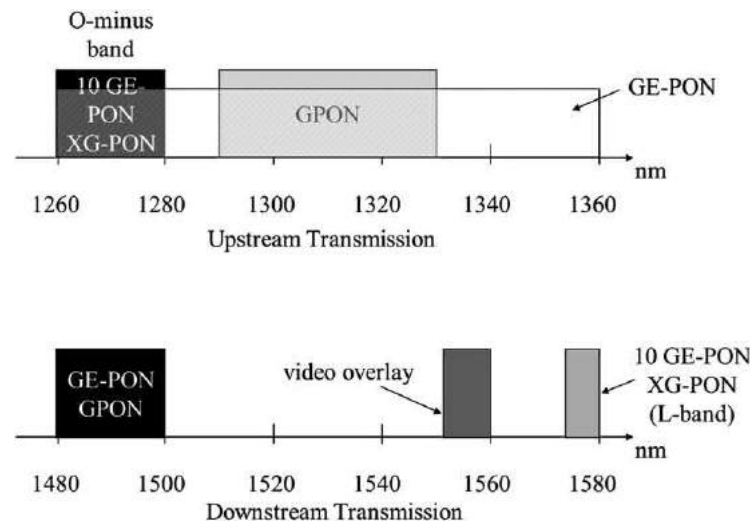


Fig. 7 Spectrum overlay of 10GE-PON and XG-PON showing coexistence with GE-PON and GPON. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. *IEEE/OSA J. Lightw. Technol.* 30(4), 597–608.

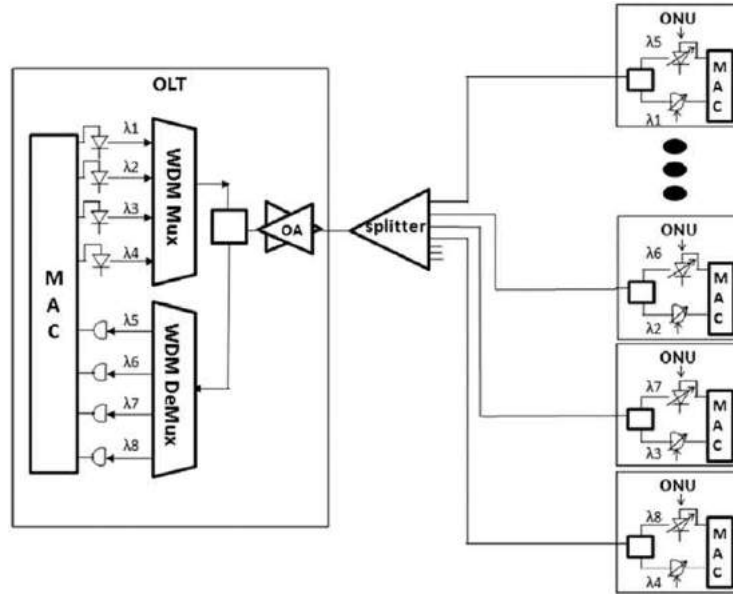


Fig. 8 TWDM-PON system diagram. Adapted from Luo, Y., Sui, M., Effenberger, F., 2012. Wavelength management in time and wavelength division multiplexed passive optical networks (TWDM-PONs). In: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM).

10-Gb/s downstream and 10/2.5-Gb/s upstream capacity per channel. For seamless network upgrade and user migration, the optical technologies specified for NG-PON2 must be compatible to legacy PON systems, including the coexisting wavelength band plan and reusing the power splitting ODNs. As such, L band (1596–1603 nm) and C-band (1524–1544 nm) have been assigned for downstream and upstream transmission, respectively (ITU, 2013).

In addition, PtP-WDM channel overlay is defined in NG-PON2 for supporting services such as fronthaul and backhaul wireless network transmission with an expanded spectrum option of 1524–1625 nm and a shared spectrum option of 1603–1625 nm (ITU, 2013). Regardless, all specified PtP-WDM wavebands must be chosen from unused spectrum by the TWDM-PON and legacy systems. As for link budget, four ODN path loss classes are specified, namely Class N1, N2, E1 and E2, which corresponds to 29 dB, 31 dB, 33 dB and 25 dB respectively. The line code for both downstream and upstream links is scrambled NRZ. To improve receiver sensitivity, FEC support is mandatory at both OLT channel termination and ONU. For 2.5 Gb/s and 10 Gb/s, the FEC codes are RS (248, 332) and RS (255, 223), respectively.

In view of the multi-wavelength operation property of TWDM-PON, additional network and protocol requirements beyond that of legacy PON systems, are critical in TWDM-PONs, e.g., the ability to tune the ONU transmitter and receiver. In the upstream direction, the ONU is tuned to emit at the desired wavelength channel. In the downstream direction, a tunable ONU receiver is required to select the proper wavelength channel. Also, to eliminate the chromatic dispersion induced signal distortion, electrical dispersion compensation (EDC) may be used in the OLT/ONU to achieve the specified loss budget.

Energy-Efficient Dynamic Bandwidth Allocation

Aside from its increased bandwidth capacity, TWDM-PON is also an attractive solution from an energy perspective (Valcarenghi *et al.*, 2014; Dixit *et al.*, 2012). Due to multiple line cards and wavelengths channels, the power consumption of the OLT is comparatively higher in TWDM-PON compared to networks specified under NG-PON1. Nevertheless, the large number of ONUs serviced by the TWDM-PON results in lower energy per user. Considering its expected uptake, future-proofing the TWDM-PON in terms of energy-efficiency is becoming an important deployment factor.

As discussed before, a TWDM-PON consists of multiple wavelengths channels, requiring tunable ONU transceivers (TRXs) that can tune to any of the wavelength channels present in the network. Due to this tunability, at low network loads, the network may be reconfigured by switching off idle wavelengths and thereby saving energy at the OLT (Valcarenghi *et al.*, 2014). In addition, conventional power-saving methods such as sleep and doze can be implemented at the ONUs to improve overall energy-efficiency (Dixit *et al.*, 2012). As a result, recent energy-efficient solutions for TWDM-PON including dynamic wavelength and bandwidth allocation (DWBA) algorithms incorporating wavelength reconfiguration and sleep/doze mode operations, have been proposed.

One of the most important aspects of network reconfiguration through wavelength reallocation is determining how many wavelengths to be switched off without affecting the system performance. Optimizing the number of active wavelengths, subjected to different parameters of the network such as wavelength tuning time (Luo *et al.*, 2012), energy consumption (Yang *et al.*, 2011), or distance from the OLT (Dixit *et al.*, 2013), have been proposed as a result. Energy-saving techniques such as wavelength

reconfiguration and sleep/doze mode operations on the other hand, result in an increase in the average delay experienced by transmission packets. If the tuning time is assumed to be negligible, unlimited wavelength switching could be practised in a network. Although unlimited wavelength switching may theoretically result in significant energy-savings, in reality, the tuning time of an ONU is significant. In fact, the tuning time of an ONU transmitter can range from nanosecond to sub-second order while the ONU receiver tuning time can range from nanosecond to second order based on implementation (Asaka, 2014). It has also been shown that the ratio between the tuning time and the polling cycle time (Kondepu et al., 2014), and the OLT's awareness of the wavelength tuning time (Buttaboni et al., 2014) are strongly correlated to performance matrices. As such, when considering a TWDM-PON with non-zero wavelength tuning time, the improved energy-savings through wavelength reconfiguration will be achieved at the expense of increased average delay of the network. To overcome this shortfall, using the average delay as a constraint in wavelength optimization (Xu et al., 2015) and changing the polling sequences based on the wavelength tuning time (Wang et al., 2015), have been reported in literature.

Introducing sleep/doze mode operations into a wavelength-reconfiguring algorithm increases the average delay further. In the energy-efficient DWBA algorithm reported in Dias et al. (2015), the Quality-of-Service degradation is minimized using Bayesian estimation and prediction techniques. The proposed DWBA algorithm considers a delay-constrained TWDM-PON and determines the maximum polling cycle time, $T_{poll\ max}$ that satisfies a given delay constraint. The $T_{poll\ max}$ is considered to achieve maximum energy-savings at the ONUs through sleep/doze mode operations. Based on this $T_{poll\ max}$, the number of active wavelengths is determined and the remaining idle wavelengths are switched off to achieve energy-savings at the OLT. Under the proposed framework, $T_{poll\ max}$ for both online and offline DWBA algorithms are determined. Under the online DWBA algorithm, the OLT uses average inter-arrival time of packets, estimated using Bayesian estimation at the ONUs, to predict the accumulated bandwidth at each ONU. The proposed traffic estimation and prediction prevents the ONUs from waiting extra cycles, before being allocated any bandwidth for transmission. Fig. 9 illustrates the traffic flow of the proposed DWBA algorithm.

High Capacity TWDM-PONs

To accommodate the anticipated increase in customer bandwidth demand, NG-PON2 systems may support an optional extension for services in excess of 10 Gb/s upstream and/or downstream towards one PON client, as was specified in G989.1 amendment 1 (ITU, 2013). It was defined in the recommendation G.989 series that NG-PON2 system supports multiple wavelength channels and enables flexibility to add capacity as bandwidth demand is expected to grow to 100 Gbit/s and beyond (ITU, 2013). The cost-effective solution for capacity enhancement, such as realizing 25 Gb/s per channel is under hot discussion. Generally, 10 G-class optical devices have been preferred for cost control, thereby necessitating advanced modulation formats with higher spectral efficiency. Four-level pulse amplitude modulation (PAM-4) (Wei et al., 2015) and duobinary (Li et al., 2015) formats are the most promising format candidates, and based on these two modulation formats, many experimental demonstrations of 10 G-class devices enabled 25 Gb/s modulation have been reported. As the demodulation circuit for duobinary and PAM-4 formats are not yet commercially viable, signals are sampled by real-time oscilloscope and processed offline in most of these demonstrations. Several research groups have reported real-time demodulation for duobinary and PAM-4 formats, although further improvements are required for application in practical PON systems. To date, the reported real-time 25-Gb/s PON system demonstrations are mostly based on optical duobinary (ODB) format. Though the ODB format is robust to fiber dispersion, it necessitates costly Mach-Zender Modulators (MZMs) and wide-band receivers for modulation and detection.

Consequently, efforts in increasing the modulation bandwidth as well as decreasing the cost of both optical transmitters and receivers are underway. Related works are being reported, e.g., 30-GHz directly modulation DFB laser (Zhang et al., 2015), 25-Gb/s

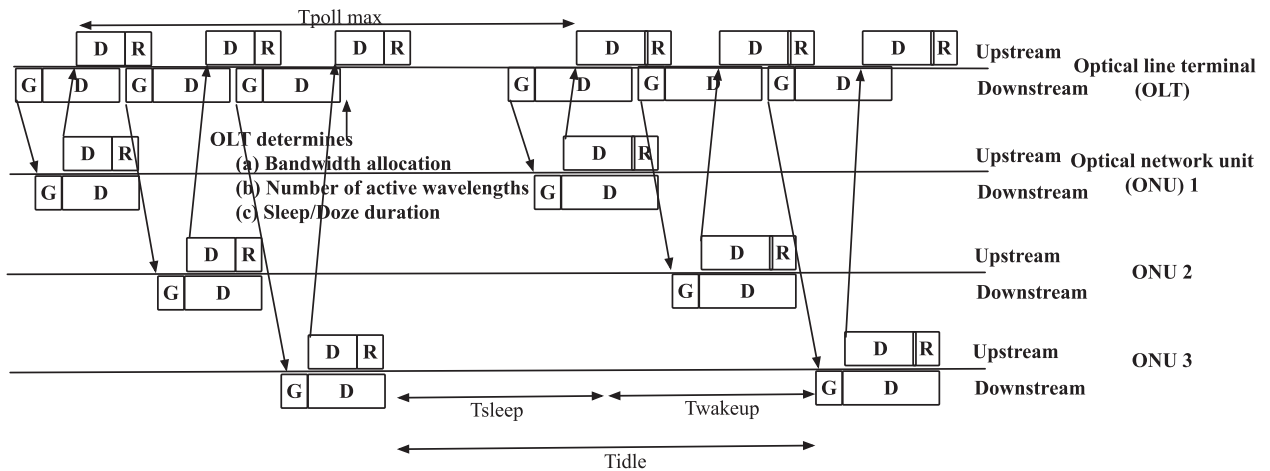


Fig. 9 Traffic flow of the OFF-DWBA algorithm. D-Data, R-REPORT and G-GATE. Dias, M.P.I., Van, D.P., Valcarenghi, L., Wong, E., 2015. An energy-efficient framework for wavelength and bandwidth allocation in TWDM PON. J. Opt. Commun. Netw. 7(6) 496–504.

Si-Ge APD (Huang *et al.*, 2016), and bandwidth improvement of CMOS-APD (Lee *et al.*, 2015), etc. Signal distortion from fiber chromatic dispersion is another issue to be handled, especially for high-speed signal, and some institutes are studying the proper dispersion compensation techniques for high data rate signal. With the development of wide-band transceivers and practical dispersion compensation techniques, NRZ format can still be a potential candidate for high capacity PON applications.

Future PONs

Wavelength Division Multiplexed PONs

The WDM-PON has been touted as a long-term access solution beyond TDM-PONs. The system operation of a WDM-PON is entirely different to the TDM-PON whereby in the former each ONU located in a home/office/building is assigned its own upstream and downstream wavelength channel. This is contrary to more a traditional PON architectures where one optical feed is shared among 32 or more users. In that case, each home operates at the same wavelength, and is allotted a 1/32nd time slot on the main fiber. In a WDM-PON, each home/office/building is assigned its own wavelength and has continual use of the fiber at that wavelength. However, virtually all PON technologies rely on some form of wavelength division multiplexing (WDM) to enable bi-directional communications. For example, in a typical GPON system, the upstream communication runs at 1310 nm wavelength, while the downstream traffic is transmitted at 1490 nm. A third wavelength at 1550 nm is used for video overlay. The utilization of WDM in PON systems is therefore already very commonplace.

Fig. 10 illustrates a typical WDM PON comprising a CO with an OLT, two cyclic AWGs, a trunk or feeder fiber, a series of distributions fibers, and ONUs at the subscriber premises (Wong, 2012). Downstream wavelengths and upstream wavelengths from the ONUs are multiplexed and demultiplexed in the AWG located at the CO. These wavelength channels also under (de) multiplexing at the AWG in the remote node located closer to the ONUs. The trunk fiber carries the multiplexed downstream and upstream wavelengths between the first and second AWG. The distribution fiber carries the demultiplexed wavelengths to and from the remote node to the ONUs. Both AWGs have the same free spectral range (FSR). Note that the downstream and upstream wavelengths allocated to each ONU are intentionally spaced at a multiple of the FSR, allowing both wavelengths to be directed in and out of the same AWG port that is connected to the destination ONU. For example, the downstream wavelength destined for ONU1, denoted λ_1 , and the upstream wavelength from ONU1 denoted $\lambda_{1'}$, are spaced a multiple of the FSR. In a typical WDM PON, wavelength channels are spaced 100 GHz (0.8 nm) apart. In systems classified as dense WDM-PON (DWDM), a channel spacing of 50 GHz or less is deployed.

There are several advantages to the WDM-PON architecture over more traditional TDM-PON systems. Foremost is the dedicated bandwidth available to each ONU. Secondly, WDM-PONs can typically have improved security and scalability features, because each ONU only transmits and receives on assigned wavelength channels. Thirdly, the MAC layer in a WDM-PON is simplified, since the logical topology of a WDM-PON is a PtP topology between the OLT and each ONU, and thus does not require the use of PtMP media access controllers found in conventional TDM-PON networks. Finally, each wavelength in a WDM-PON network is effectively a PtP link, allowing each link to operate at a different speed and if required a different MAC protocol for maximum flexibility and pay-as-you-grow upgrades.

The main challenge with WDM-PON is the cost. Since each ONU is assigned his own wavelength, this suggests that the OLT must transmit on different wavelengths versus one shared wavelength as found in more traditional TDM-PON systems. Likewise, it requires that each of the ONUs on a link operate at a separate wavelength, suggesting that every ONU requires a wavelength-specific laser that can operate at its assigned wavelength. Since each ONU is assigned a unique upstream wavelength channel,

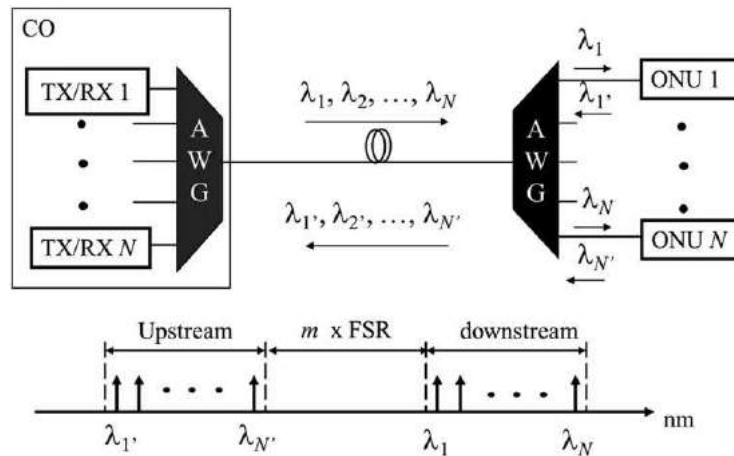


Fig. 10 WDM-PON. Inset: Allocation of upstream/downstream wavelength into two separate wavebands. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30(4), 597–608.

distinct wavelength transmitters must be deployed at the subscriber premises. The simplest solution is to utilize fixed wavelength transmitters with wavelength tunability, e.g., tunable distributed feedback (DFB) laser (Suzuki *et al.*, 2007), tunable vertical cavity surface emitting lasers (VCSELs) (Wong *et al.*, 2006), and distributed Bragg reflector laser diodes (DBR-LDs), such as sampled grating DBR-LDs and super structure grating DBR-LDs (Kuвано *et al.*, 2001; Ishii *et al.*, 1996).

Alternatively, wavelength reuse schemes that eliminate optical sources at the ONUs have been proposed. Aside from carrying downstream signals, the downstream wavelength is used to wavelength seed a reflective semiconductor optical amplifier (RSOA) located at the designated ONU. The RSOA is intentionally operated in the gain saturation region whereby the RSOA amplitude squeezing effect can be exploited to erase the modulation on the seeding downstream wavelength (Katagiri *et al.*, 1999). Wavelength reuse schemes eliminate the need for seeding sources, is less costly than using tunable lasers, and allows direct modulation of the RSOA for upstream transmission. The drawback is severe performance degradation at the CO due to interference between the newly modulated upstream data and residual downstream data. A solution to improve transmission performance is through using orthogonal upstream and downstream modulation formats (Garces *et al.*, 2007) and line coding approaches (Kim *et al.*, 2007; Al-Qazwini and Kim, 2011).

In addressing the potential large inventory and cost of wavelength specific sources, researchers have been concentrating on developing cost-efficient and wavelength independent sources termed “colorless” sources. These sources include directly modulated light-emitting diode (LED) and super-luminescent diode (SLD) (Reeve *et al.*, 1988) which broadband spectrum from each ONU is spectrally sliced by the AWG in the upstream direction. These sources are limited in terms of system reach and modulation bit rate. In yet another category of colorless sources, optical light originating from the CO is fed into the ONUs to injection-lock Fabry-Perot laser diodes (F-P LDs) (Kim *et al.*, 2000) or to wavelength-seed RSOAs (Feuer *et al.*, 1996). The wavelength seeding scheme is identical to the injection-locking scheme except for the use of an RSOA which amplifies and modulates the incoming continuous wave (CW) light. The drawback of these schemes lie in the requirement of additional broadband light source(s), limited transmission bit-rate due to device limitation, and susceptibility to transmission performance impairment from fiber dispersion, Rayleigh backscattering noise, and broadband amplified spontaneously emission (ASE) noise from the broadband light source and the colorless source.

More recently, self-seeding of the RSOA was proposed in which each RSOA is self-seeded by its own spectrally-sliced CW light (Wong *et al.*, 2007). ASE light emitted from each RSOA is spectrally-sliced by the AWG located at the remote node and feedback to the RSOA. Self-seeding removes the need for active temperature tracking between the optical components within the remote node and between the remote node and each RSOA and identical RSOAs can be placed at all ONUs.

Extended-Reach PONs

ER-PONs can serve an increased number of subscribers, accommodate high bandwidth applications, and an increased network span that consolidates metro and access networks. This consolidation lowers capital expenditure (CAPEX) and operational expenditure (OPEX) due to a reduction in the number of active network interfaces and elements in the field. The idea of extending the reach of a PON is not new with many significant demonstrations and field trials carried out in the 1990s (Mestdagh and Martin, 1996). However, advances in optical amplification and WDM technologies in recent years along with the volume manufacturing of optical components such as SOAs, AWGs, and WDM filters, have lowered the cost of ER-PONs to a point that is competitive in the business and access market.

An illustrative example of an ER-PON is shown in Fig. 11 whereby multiple COs which were previously connected to a metropolitan aggregation ring network (dash line) are now consolidated at the main central office (MCO) to support a large

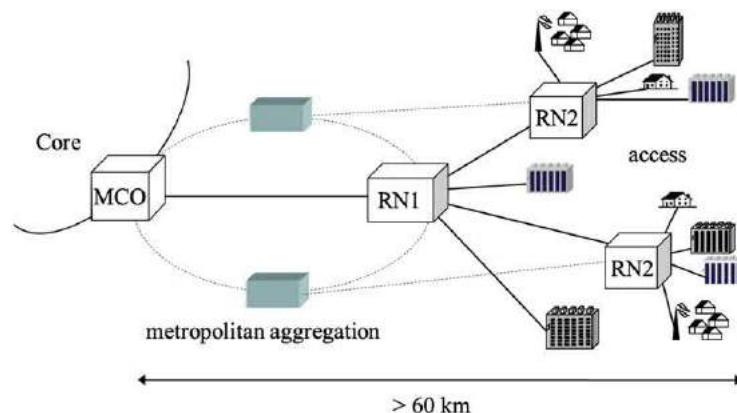


Fig. 11 Schematic diagram of a two stage extended-reach PON (ER-PON) showing consolidation of central offices (COs) previously connected to the metropolitan aggregation ring network at the main central office (MCO). Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. *IEEE/OSA J. Lightw. Technol.* 30(4), 597–608.

number of customers via remote nodes. The ODN of a typical ER-PON comprises an extended-reach FF which connects the MCO to remote node 1 (RN1), DFs which connect RN1 to remote node 2 (RN2), and last mile fibers (LMFs) that connect RN2s to customers. Bandwidth-dedicated business customers and mobile customers can be connected directly to RN1 whereas residential users can be connected to RN2. By exploiting a combination of (a) TDM or WDM in the first stage, and (b) TDM in the second stage, system reach can be extended from the conventional 20 km to 60–100 km while maintaining a 1:32 or higher split ratio. In case where the first stage is based on TDM, a single upstream and downstream wavelength support all users with power splitting carried out in both RN1 and RN2s. When the first stage is based on WDM, wavelengths are multiplexed at the MCO and transmitted simultaneously through a feeder fiber to RN where optical amplification and demultiplexing take place. In turn, in the second stage of the ER-PON, each demultiplexed wavelength is then further power split at RN2 and then transmitted to multiple customers. A reach extender is typically housed in RN1 to compensate for power loss due to the long transmission distance and high split ratio.

A few reach extended technologies have been proposed to compensate for the insertion loss of ER-PONs. These include optical amplification at the MCO and/or RN1, using erbium doped fiber amplifier (EDFA) (Townsend *et al.*, 2008), SOA (Suzuki and Nakagawa, 2005), and distributed Raman amplification (DRA) (Kjaer *et al.*, 2006). EDFAs provide excellent gain power and noise performance in the C and L-bands. However, optical amplification is limited to these wavelength bands and unless gain control through optical gain clamping or pump power variation is implemented, the relatively slow speed in adjusting the EDFA gain makes this option disadvantageous to the bursty nature of upstream TDMA traffic. The SOA, on the other hand, benefits from the fact that it can operate at any wavelength of interest including the O band and with better gain dynamics than the EDFA. Other benefits include compactness and the ability to facilitate additional functionalities such as wavelength conversion and all-optical regeneration. However, the SOA operates on a per-channel basis. The inability to provide simultaneous amplification across multiple channels is its main drawback. The use of DRA in counter and co-propagation configurations enables amplification across a wide wavelength region over a bidirectional fiber link. Raman amplifiers can be tailored to provide a flat optical gain bandwidth that encompasses wavelengths exceeding those of common optical amplifiers. Using simultaneous launching of multiple pump wavelengths, DRA can achieve a flat Raman gain of up to 100 nm of bandwidth. Combining the benefits of EDFA and SOA, DRA allows simultaneous multichannel amplification, fast gain dynamics, and the ability to provide gain at any wavelength contingent on the pump wavelength. Using DRA does however require high pumping power, thereby adding to the power consumption at the RN1 and MCO.

In replacement of optical amplification, electronic repeater(s) can be implemented at RN1 to facilitate 1R or 2R regeneration of both upstream and downstream signals (Davey *et al.*, 2009). In using electronic repeater(s), benefits include the option of wavelength conversion, bit-error-rate monitoring, and optical power equalization of the burst mode signals in the upstream direction. However, a drawback of electronic repeaters is the need for bit-rate specific burst mode receivers that must also be capable of handling a wide dynamic range.

Orthogonal Frequency Division Multiplexing PON

Orthogonal frequency division multiplexing (OFDM) is a modulation technique which is now used in most new and emerging broadband wired and wireless communication systems (Cvijetic, 2012). In OFDM, the spectra of individual subcarriers overlap as depicted in Fig. 12(a), but the signals on each subcarrier are mathematically orthogonal over one OFDM symbol period. As such, the subcarriers can be demodulated without interference. Dispersion-induced signal distortion can be eliminated by adding a cyclic prefix (CP) to the start of each time domain OFDM symbol before transmission, i.e., a number of samples from the end of the symbol is appended to the start of the symbol as shown in Fig. 12(c) (Armstrong, 2009), by which any distortion caused by a linear dispersive channel can be corrected simply using a ‘single-tap’ equalizer. Also, by employing bit-loading techniques, the modulation formats on each subcarrier can be different according to the channel response. Combined with the time- and frequency-domain multiplexing as shown in Fig. 12(b) (Qiu *et al.*, 2011), data can be assigned to end-users by different time slots or subcarriers with different modulation formats, thereby supporting a flexible 3-dimensional (3D) dynamic bandwidth allocation algorithm.

Despite advantages such as high spectral efficiency, high dispersion tolerance, and flexible bandwidth allocation, thus making OFDM a suitable modulation technique for future high-capacity PON systems, its drawbacks mainly due to high peak to average power ratio (PAPR) and high sensitivity to phase noise and frequency offset, easily degrades performance and limits its practical application. Digital signal processing (DSP) algorithms are being proposed to reduce PAPR (Popoola *et al.*, 2014) and to mitigate signal distortion caused by phase noise and frequency offset (Yi *et al.*, 2008). However, as high-speed ADC/DAC and complex DSP algorithms are required for the generation, demodulation and equalization of high-speed OFDM signal, most of the experimental demonstrations are realized off-line. Further, the capacity of the demonstrated real-time systems is generally lower than 10-Gb/s per channel. Another drawback of OFDM modulation is its low sensitivity. With the modulation format of 16-QAM or even higher, the sensitivity of directly-detected signal is much lower than the traditional NRZ format, thereby severely limiting the loss budget of the PON system. To solve this problem, coherent detection is introduced in OFDM-PONs to improve receiver sensitivity. As a result, the loss budget is increased at the cost of a higher system complexity.

So in the short term, limited by the cost of high-speed ADC/DAC and the complex DSP modules, OFDM-PONs are still considered an expensive option. However, with the emergence of novel applications, low spectral-efficiency formats such as NRZ

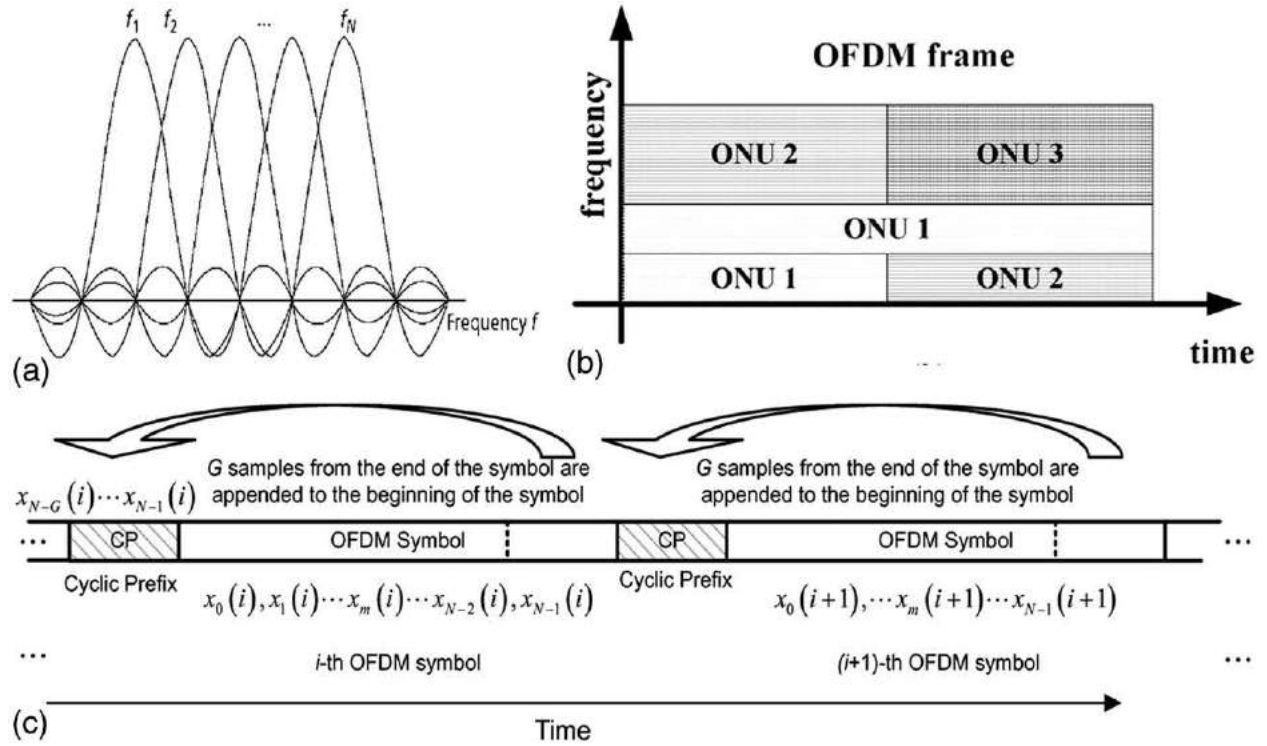


Fig. 12 (a) OFDM spectrum for N orthogonal frequency domain subcarriers; (b) frequency and time domain partitioning of an OFDM frame. (c) Time domain sequence of OFDM symbols showing the cyclic prefix. Adapted from Armstrong, J., 2009. OFDM for optical communications. *J. Lightw. Technol.* 27(3), 189–204 and Qiu, K., Yi, X., Zhang, J., *et al.*, 2011. OFDM-PON optical fiber access technologies. In: Asia Communications and Photonics Conference Optical Society of America, 1–9.

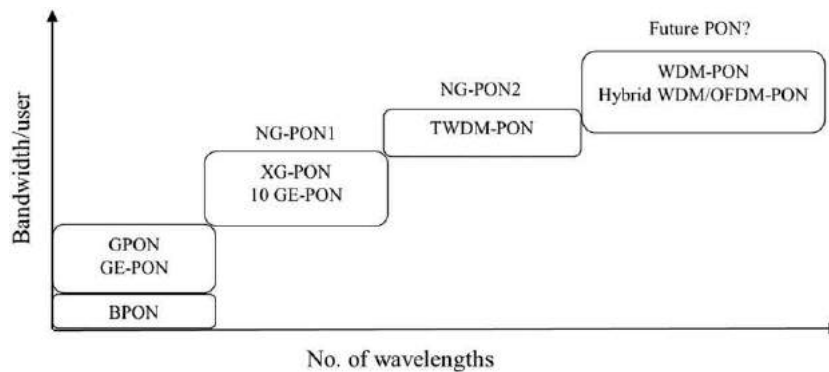


Fig. 13 Evolution of broadband passive optical access networks.

and RZ, will not be able to satisfy the increase in bandwidth demand and ultimately, and the spectral-effective OFDM-PON will be a solution by then.

Summary and Conclusion

A brief overview of the evolution of passive optical networks with particular focus on standardized systems mentioned, and the emerging trends to address future bandwidth needs, was discussed in this article. This summary of the evolution of PONs is illustrated in Fig. 13. Driven by the increasing user bandwidth demand for emerging applications, PONs with increasing capacity and cost-efficiency have been deeply discussed and studied around the world. BPON, GPON, XG-PON have been standardized by ITU, while GE-PON and 10GE-PON have been standardized by IEEE. These technologies are all developed by gradually increasing the data rate. The main reasons behind the push for these TDM/TDMA PON systems are to extend the longevity of existing ODNs and to allow co-existence with the current generation PONs such that the operational impact on existing users will be minimized.

Due to the chromatic dispersion and cost issues, only increasing the data rate per wavelength is more difficult and encounters some bottlenecks especially when the bit rate is beyond 10 Gb/s. In order to increase the system capacity, NG-PON2 employs the technique of stacking wavelength, called as TWDM-PON, which has been standardized by ITU-T in 2015 and is supposed to be commercially available in around 2018. From 2016, IEEE next generation Ethernet passive optical network (NG-EPON) has also been discussed in-depth to provide a capacity of 100-Gb/s with 25-Gb/s per wavelength, which has attracted extensive attention from the industry. In the future, in order to satisfy the bandwidth requirement, more advanced technologies will be proposed to upgrade the system capacity such as OFDM-PON and WDM-PON. High-speed PON with higher data rate per wavelength is also an effective solution to increasing the system capacity. So the combination of stacking more wavelengths and increasing the data rate of per wavelength or using complicated technologies such as coherent detection and digital signal processing will be effective solutions to overcome the problem of system capacity. In order to successfully deploy these technologies, the implementation complexity must be minimized to a level that is comparable to existing commercialized systems and with a cost that is sufficiently low to meet the cost constraints of the access market. In all PON with higher capacity and lower cost will be the ultimate goal from both industry and academia.

See also: Passive Optical Components

References

- CommScope, 2013. EPON-GPON comparison, CommScope Solutions Marketing.
- Al-Qazwini, Z., Kim, H., 2011. Line coding for downlink DML modulation in lambda-shared, RSOA-based asymmetric bidirectional WDM PONs. In: Proceedings Optical Fiber Communication Conference National Fiber Optic Engineers Conference (OFC/NFOEC), paper OMP5.
- Armstrong, J., 2009. OFDM for optical communications. *J. Lightw. Technol.* 27 (3), 189–204.
- Asaka, K., 2014. What will be killer devices and components for NG-PON2. In: Proceedings of the 40th European Conference and Exhibition on Optical Communication (ECOC).
- Buttaboni, A., Andrade, M.D., Tornatore, M., 2014. Dynamic bandwidth and wavelength allocation with coexistence of transmission technologies in TWDM PONs. In: Proceedings of the 16th International Telecommunications Network Strategy and Planning Symposium (Networks).
- Cvijetic, N., 2012. OFDM for next-generation optical access networks. *J. Lightw. Technol.* 30 (4), 384–398.
- Davey, R., Grossman, D.B., Raszlovits-Wiech, M., *et al.*, 2009. Long-reach passive optical networks. *J. Lightw. Technol.* 27, 273–291.
- Dias, M.P.I., Van, D.P., Valcarengi, L., Wong, E., 2015. An energy-efficient framework for wavelength and bandwidth allocation in TWDM PON. *J. Opt. Commun. Netw.* 7 (6), 496–504.
- Dixit, A., Lannoo, B., Colle, D., Pickavet, M., Demeester, P., 2012. ONU power saving modes in next generation optical access networks: Progress, efficiency, and challenges. *Opt. Express* 20, B52–B63.
- Dixit, A., Lannoo, B., Colle, D., Pickavet, M., Demeester, P., 2013. Flexible TDMA/WDMA passive optical network: Energy-efficient next-generation optical access solution. *J. Opt. Switch. Netw.* 10, 491–506.
- Feuer, M.D., Wiesenfeld, J.M., Perino, J.S., *et al.*, 1996. Single-port laser-amplifier modulators for local access. *IEEE Photon. Technol. Lett.* 8, 1175–1177.
- Förster, M., 2006. Worldwide Development of FTTH. Available at: https://www.hft-leipzig.de/fileadmin/image_hft/presse/Institut_HF/FTTH_Foerster.pdf.
- Garces, I., Aguado, J.C., Martinez, J.J., *et al.*, 2007. Analysis of narrow-FSK downstream modulation in colorless WDM PONs. *Electron. Lett.* 43 (8), 471–472.
- Huang, Z., Li, C., Yu, K., *et al.*, 2016. A 25Gbps low-voltage waveguide Si-Ge avalanche photodiode. In: CLEO: Science and innovations, Optical Society of America, STh4E. 6.
- IEEE, 2004. Standard 802.3ah – IEEE Standard for Local and Metropolitan Area Networks – Specific requirements – Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method and physical layer specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks.
- IEEE, 2009. Standard 802.3av – IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment 1: Physical Layer Specifications and Management Parameters for 10 Gb/s Passive Optical Networks.
- Ishii, H., Tanabe, H., Kano, F., *et al.*, 1996. Quasicontinuous wavelength tuning in super-structure-grating (SSG) DBR lasers. *IEEE J. Quantum Electron.* 32 (3), 433–441.
- ITU, 2013. 40-Gigabit-capable passive optical networks (NG-PON2): General requirements, ITU-T Recommendation G.989.1.
- ITU-T, 2005. G.983.1 Recommendation: Broadband optical access systems based on passive optical networks (PON).
- ITU-T, 2008a. G.984 – Series recommendation, Gigabit-capable Passive Optical Networks (G-PON).
- ITU-T, 2008b. G.984 x-series recommendations: Gigabit-capable passive optical networks (G-PON): Transmission convergence layer specification (G.984.3).
- ITU-T, 2010. G.987.x – Series recommendation: 10-Gigabit-capable passive optical network (XG-PON).
- Katagiri, Y., Suzuki, K., Aida, K., 1999. Intensity stabilisation of spectrum-sliced Gaussian radiation based on amplitude squeezing using semiconductor optical amplifiers with gain saturation. *Electron. Lett.* 35 (16), 1362–1364.
- Kim, H.D., Kang, S.-G., Lee, C.-H., 2000. A low-cost WDM source with an ASE injected Fabry-Perot semiconductor laser. *IEEE. Photon. Technol. Lett.* 12, 1067–1069.
- Kim, S.Y., Jun, S.B., Takushima, Y., Son, E.S., Chung, Y.C., 2007. Enhanced performance of RSOA based WDM PON by using Manchester coding. *OSA J. Opt. Netw.* 6, 624.
- Kjaer, R., Monroy, I.T., Oxenlowe, L.K., Jeppesen, P., Palsdottir, B., *et al.*, 2006. Bi-directional 120 km long reach PON link based on distributed Raman amplification. In: Proceedings IEEE LEOS Annual Meeting, paper WEE3, Oct.
- Kondepu, K., Valcarengi, L., Van, D.P., Castoldi, P., 2014. Impact of ONU tuning time in TWDM-PON with dynamic wavelength and bandwidth allocation: An FPGA-based evaluation. In: Proceedings of the 40th European Conference and Exhibition on Optical Communication (ECOC).
- Kramer, G., 2005. Ethernet Passive Optical Networks. New York, NY: McGraw-Hill.
- Kuwano, S., Teshima, M., Uematsu, H., Iwatsuki, K., 2001. WDM optical packet transmission experiment over 235 km of installed fibers. In: Proceedings Optical Fiber Communication Conference National Fiber Optic Engineers Conference (OFC/NFOEC), TuK4, Mar.
- Lee, M.-J., Lee, J.-M., Rucker, H., Choi, W.-Y., 2015. Bandwidth improvement of CMOS-APD with carrier-acceleration technique. *IEEE Photon. Technol. Lett.* 27 (13), 1387–1390.
- Li, Z., Yi, L., Wang, X., *et al.*, 2015. 28 Gb/s duobinary signal transmission over 40 km based on 10 GHz DML and PIN for 100 Gb/s PON. *Opt. Express* 23 (16), 20249–20256.

- Luo, Y., Sui, M., Effenberger, F., 2012. Wavelength management in time and wavelength division multiplexed passive optical networks (TWDM-PONs). In: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM).
- Luo, Y., Zhou, X., Effenberger, F., *et al.*, 2013. Time-and wavelength-division multiplexed passive optical network (TWDM-PON) for next-generation PON stage 2 (NG-PON2). *J. Lightw. Technol.* 31 (4), 587–593.
- Mestdagh, D.G., Martin, C.M., 1996. The super-PON concept and its technical challenges. *Broadband Commun.*
- Oakley, K.A., 1988. An economic way to see in the broadband dawn (passive optical network). In: Proceedings of IEEE Global Telecommunications Conference and Exhibition vol. 3, pp. 1574–1578.
- Popoola, W.O., Ghassemlooy, Z., Stewart, B.G., 2014. Pilot-assisted PAPR reduction technique for optical OFDM communication systems. *J. Lightw. Technol.* 32 (7), 1374–1382.
- Qiu, K., Yi, X., Zhang, J., *et al.*, 2011. OFDM-PON optical fiber access technologies. In: Asia Communications and Photonics Conference Optical Society of America, 1–9.
- Reeve, M.H., Hunwicks, A.R., Zhao, W., *et al.*, 1988. LED spectral slicing for single-mode local loop applications. *Electron. Lett.* 24 (7), 389–390.
- Suzuki, H., Fujiwara, M., Suzuki, T., *et al.*, 2007. Wavelength-tunable DWDM-SFP transceiver with a signal monitoring interface and its application to coexistence-type colorless WDM-PON. In: Proceedings European Conference on Optical Communication, PD3.4.
- Suzuki, N., Nakagawa, J., 2005. First demonstration of full burst optical amplified GE-PON uplink with extended system budget of up to 128 ONU and 58 km reach. In: Proceedings European Conference Optical Communication, Sept., Paper Tu 1.3.3.
- Townsend, P.D., Talli, G., MacHale, E.K., Antony, C., 2008. Long reach PONs. In: Technical Digest of Conference Optical Internet, pp. 1–2.
- Valcarengi, L., Yoshida, Y., Maruta, A., Castodi, P., Kitayama, K., 2014. Energy savings in TWDM(A) PONs: Challenges and opportunities. In: Proceedings of 15th IEEE Transparent Optical Networks (ICTON), p. We.A4.4.
- Wang, H., Su, S., Gu, R., Ji, Y., 2015. A minimum wavelength tuning scheme for dynamic wavelength assignment in TWDM-PON. In: Proceedings of the 14th International Conference on Optical Communications and Networks (ICOON).
- Wei, J., Eiselt, N., Griesser, H., *et al.*, 2015. First demonstration of real-time end-to-end 40 Gb/s PAM-4 system using 10-G transmitter for next generation access applications. In: European Conference on Optical Communication, 32–33.
- Wong, E., 2012. Next-generation broadband access networks and technologies. *IEEE/OSA J. Lightw. Technol.* 30 (4), 597–608.
- Wong, E., Lee, K.L., Anderson, T.B., 2007. Directly modulated self-seeding reflective semiconductor optical amplifiers as colorless transmitters in wavelength division multiplexed passive optical networks. *J. Lightw. Technol.* 25 (1), 67–74.
- Wong, E., Zhao, X., Chang-Hasnain, C.J., Hofmann, W., Amann, M.C., 2006. Optically injection-locked 1.55 μm VCSEL as upstream transmitters in WDM-PONs. *IEEE Photonics Technol. Lett.* 18 (22), 2371–2373.
- Xu, W., Fu, M., Le, Z., 2015. Energy efficiency scheme for delay aware TWDM-PON. In: Proceedings of the 14th International Conference on Optical Communications and Networks (ICOON), pp. 14–16.
- Yang, H., Sun, W., Hu, W., Li, J., 2011. ONU migration in dynamic time and wavelength division multiplexed passive optical network (TWDM-PON). *Opt. Express* 21, 285–295.
- Yi, X., Shieh, W., Ma, Y., 2008. Phase noise effects on high spectral efficiency coherent optical OFDM transmission. *J. Lightw. Technol.* 26 (10), 1309–1316.
- Zhang, Z., Liu, Y., Guo, J., *et al.*, 2015. 30-GHz directly modulation DFB laser with narrow linewidth. In: Asia Communications and Photonics Conference. Optical Society of America, AM1B. 3.

Relevant Websites

<https://www.corecess.com/eng/solution/wdmpn.asp>
Corecess.

<https://www.lg-nortel.com/index.html>
LG-Nortel.

Holographic Recording Media and Devices

Pierre-Alexandre Blanche, The University of Arizona, Tucson, AZ, United States

© 2018 Elsevier Ltd. All rights reserved.

Holography Terminology

Holographic recording materials are able to change their refractive index and/or their absorption according to the intensity pattern created by two interfering laser beams. Then, when the material is re-illuminated by the appropriate light, the material can diffract the incoming beam and display the hologram. Thus, there are two functions that the material must perform: sensitivity to the recording light, and physical transformation of the optical properties.

The distinction between an image recording material and a holographic recording material is their necessary spatial resolution. For imaging, the material only needs to resolve details down to tens of microns, so the image is detailed and sharp to the human eye. For holography, we need to reproduce the modulation obtained by the interference between two laser beams. In this case, the separation between bright and dark regions can be sub-micron scale. This is the reason why digital holography had to wait for the development of high-pitch focal-plane-array sensors to become practical.

The goal of this entry is not to explain holography. However, since it heavily relies on the terminology associated with this method, it is necessary to give a brief introduction on the most important techniques.

When the material records the hologram as an absorption modulation, one talks of an amplitude hologram, since it is the amplitude term of the light wave that is affected when interacting with the media. When it is the index of refraction of the material that is modulated by the recording, one talks of a phase hologram. In this latter case, there is no absorption of the incident light, and the efficiency of the hologram can be larger than for an amplitude hologram.

Holograms can also be recorded by modifying the surface of the material, such as by scribing or by lithography. In this case, the hologram is also a phase modulator, due to the difference in light path.

Another distinction with holograms, is they are either transmission or reflection holograms. With transmission holograms, the incident light goes through the medium and is diffracted on the other side of the material. The Bragg planes in this case are oriented more or less perpendicular to the surface of the material. For reflection holograms, the light is diffracted back toward the same side of the material as the incident beam. In this case, the Bragg planes are roughly parallel to the surface of the material.

There is an important difference between the spectral and angular properties of transmission and reflection holograms. Transmission holograms are angularly selective but spectrally dispersive, which means that when illuminated with a white light (broad band source) they diffract the incident light into a rainbow (the spectrum is dispersed). However, when transmission holograms are illuminated with a monochromatic source, they will only diffract at a very specific angle of incidence, the Bragg angle (angular selection).

The behavior of reflection holograms is quite the opposite: they are spectrally selective and angularly tolerant. Illuminated with a white light, they only diffract back a specific color, acting like a filter (spectral selection). However, reflection holograms can diffract the incoming light over a large angle of incidence.

Beware that a reflection coating can be applied to a transmission hologram to make it look like a reflection hologram: the light will be diffracted back. However, its spectral and angular characteristics will still be that of a transmission hologram.

There exist two regimes of diffraction: Raman-Nath and Bragg. The Raman-Nath regime is characterized by having a significant amount of energy not diffracted by the hologram (strong zero order), and diffracted into orders higher than 1: ± 2 , ± 3 , etc. This happens when there is only a short distance of interaction between the light and the hologram. Holograms operating in that mode are also called "thin", even though the actual thickness of the material is only one of several parameters defining the regime. On the other hand, when in the Bragg regime, holograms diffract most of the energy into the first order. For this to happen, the interaction length between the light and the grating should be relatively long, so these holograms are called "thick". The specific criteria between these two regimes is not just the physical thickness (d) but involves the light wavelength (λ_0), the material index of refraction (n), the grating spacing (Λ), and the angle of incidence (θ) in the Klein and Cook relationship:

$$Q' = \frac{2\pi\lambda_0 d}{n\Lambda^2 \cos\theta} \quad (1)$$

where $Q' < 1$ for the Raman-Nath regime, and $Q' > 1$ for the Bragg regime.

Another criteria for the energy distribution in higher orders is the Moharam and Young equation involving the index modulation (Δn):

$$\rho = \frac{\lambda_0^2}{n\Delta n\Lambda^2 \cos\theta} \quad (2)$$

where $\rho < 1$ for the Raman-Nath regime, and $\rho \geq 1$ for the Bragg regime.

We now need to introduce the different figures of merit characterizing a holographic material:

Sensitivity: sensitivity is the most important criteria defining the amount of energy density in J cm^{-2} units required to achieve a certain amount of efficiency. The sensitivity figure is generally not divided by the efficiency (J cm^{-2}) but given for a specific

efficiency (e.g., J cm⁻² @ 90%). The sensitivity can also be given as the modulation (either index of refraction (Δn) or absorption ($\Delta \alpha$)) according to the energy density. The sensitivity generally depends of the wavelength used to record the hologram, which brings us to the next figure of merit the spectral sensitivity.

Spectral sensitivity: spectral sensitivity is the range of wavelengths to which the material is responsive, and holograms can be recorded. Ideally, it should be a curve of the sensitivity according to the wavelength, but most of the time there is only indication for the main laser lines: krypton red 647 nm, HeNe red 633 nm, argon green 514 nm, doubled YAG green 532 nm, and cadmium blue 441 nm. When a material has a very limited spectral sensitivity confined in only one color, it is called monochromatic. When the material is sensitive to red, green, and blue, even if only at very specific laser lines, it is called panchromatic.

Maximum efficiency ($\eta_{\max}$): maximum efficiency ($\eta_{\max}$) is the maximum diffraction efficiency expressed either as the ratio between the first order diffracted intensity divided by the incident intensity (external efficiency), or the ratio between the first order diffracted intensity and transmitted intensity without hologram (internal efficiency). Note that since the transmitted intensity is already reduced by the absorption, scattering, and Fresnel reflections from the material, the internal efficiency is always larger than the external efficiency.

Maximum modulation: maximum modulation is the maximum index (Δn) or absorption modulation (Δa) of the material. The maximum modulation is important to determine the diffraction efficiency. Recording geometry and wavelength are also important in that regard, but as a rule of thumb, longer wavelengths require larger modulation. Also the maximum modulation is important to consider when multiplexing several holograms into the same material. In this case, each of the multiplexed holograms takes a fraction of the maximum modulation range.

Spatial resolution: spatial resolution is the range of lateral frequencies the material is able to record and reproduce. When plan waves interfere, they produce intensity fringes spaced according to the grating equation:

$$\Lambda = m \frac{\lambda_0 / n}{\sin(\theta_d) - \sin(-\theta_i)} \quad (3)$$

Where Λ is the fringes spacing, m the diffraction order, λ_0 the wavelength of the light, θ_d the diffraction angle, and θ_i the incidence angle, with both angles being defined by the recording geometry. This spacing ranges from a few hundred line pairs per millimeter (lp mm⁻¹) for low angle transmission hologram, to several thousand for reflection holograms at small wavelength (blue). An image hologram can be decomposed as the superposition of a multitude of these plane waves (Fourier decomposition). The material should be able to reproduce these small features in order to reproduce the hologram with maximum angular resolution.

Illumination Dynamic Range

The interference pattern is composed of an intensity modulated structure. So the material must be able to react not only to a certain average level of light (sensitivity), but reproduce an entire range of intensity. The dynamic range is given by the maximum and minimum intensity the material can react to. If the minimum intensity is not zero, the intensity ratio between object and reference beams can be detuned to accommodate the minimum.

Illumination Linearity Response

Within the dynamic range, it is expected that the material reacts linearly to the intensity. Thus, if the fringe pattern is sinusoidal, the modulation is also sinusoidal. If this is not the case, the phase or intensity modulation inside the material introduces higher spatial frequencies which translate into more energy directed into higher orders and a loss of intensity in the ± 1 order.

Stability: stability can be understood as the shelf life of the material before exposure, or the lifetime of the hologram once it has been processed, which are different and has to be specified. In any case, the stability of the material depends on environmental conditions such as temperature, humidity, and UV exposure.

Absorption spectrum: absorption spectrum is the absorption coefficient (cm⁻¹) as a function of wavelength. Unexposed material should have some absorption in order to interact with the recording light (sensitivity spectrum). However, it is important that once processed, the material be as transparent as possible for the waveband of interest. Absorption reduces the external efficiency.

Scattering: scattering in addition to the absorption some loss can be due to scattering. The dispersion is mostly due to inclusions in the material whose size is comparable to the wavelength of light, resulting some Mie scattering. Scattering can be due to micro-bubble formation such as in gelatin emulsion material, or polymer crystal formation such as in photopolymer. These problems will be detailed in the specific material sections.

Thickness change: When the material is processed after exposure, it can experience some thickness change, either shrinkage or swelling. This volume modification reshapes the Bragg's planes by tilting them and imposing a different grating spacing. The tilt changes the diffraction angle and the modification of the spacing affects the diffracted wavelength. Both changes can be computed according to the grating equation (Eq. (3)).

Pulse response: The energy delivered for recording the hologram can be such as a long exposure at relatively low intensity, or as a short pulse with high peak power. Laser pulses can be formed to last several hours (CW), or as short as picosecond, femtosecond, or even attosecond pulses. For a given energy, when the duration of the pulse decreases, the peak intensity increases proportionally. Depending on the mechanism involved to record the hologram in the material, there is a peak power, or a pulse

duration, for which this mechanism is less efficient or changes entirely. At that moment, there is a reciprocity failure in the sensitivity of the material.

Permanent Materials

One can claim that any material can be used to permanently record a hologram; it all depends of the recording energy. In this section we will focus our attention on the most sensitive and widely-used materials today.

Between the recording and display steps, permanent holographic materials need to be processed to enhance the phase or amplitude modulation that have been initiated by the exposure. This post-processing changes the nature of the material and fixes the hologram permanently. This means that new holograms cannot be recorded with further exposure of the material, and the hologram(s) already present cannot be modified.

Permanent holographic recording materials are the first to have been explored, and their development is closely related to the photographic recording medium. They are still developed today due to the large demand for long lasting holographic images, either for the security tag industry (bank notes and anti-counterfeit product), art work, holographic optical elements such as notch filters, dispersion gratings and null testers, or, to a lesser extent, holographic data storage.

Silver Halide

The chemistry of silver halide material is similar to the one used in analog black and white photography. Both of them use a colloidal suspension of silver halide crystals such as AgBr, AgCl, and AgI in gelatin, and require a post-exposure wet processing to make the image/hologram appear. This wet process presented in Fig. 1, uses a developing agent that enhances the latent image by transforming the entire silver halide crystal that has been exposed into an opaque metallic silver particle. The developer bath is followed by a stop bath to halt the reaction, then a fixer bath that removes the remaining unaltered silver halide crystals. At this point the picture/hologram is a gray scale (amplitude modulation) reproduction of the intensity distribution that illuminated the material).

There are two differences between the photographic and holographic material: the size of the crystals, which influences both the spatial resolution and the sensitivity; and the bleaching process, which is used in holography to transform the amplitude modulation into phase modulation and enhance the efficiency.

The size of the features to resolve in the case of a hologram (tens of nanometers) is orders of magnitude smaller than in a case of visual photography (tens of microns). Therefore, the grain size of the silver halide crystal used for hologram recording should be much smaller than for visible photography. In this regard, the holographic plates are similar to the ones used for X-ray medical applications. The sensitivity of silver halide is directly proportional to the size of the grain. So, the holographic emulsions are much less sensitive than the photographic one. For example, an emulsion able to resolve 5000 lp cm^{-1} has a grain size of 10 nm and a sensitivity of a few mJ cm^{-2} .

When the emulsion is processed to generate an amplitude modulation, its diffraction efficiency is limited to a maximum of 3.7% due to the resulting average absorption. It is possible to enhance the efficiency up to 100% by transforming this amplitude modulation into a phase modulation. This transformation is done by another wet post-processing called the bleaching process that is applied in addition to the traditional developer and fixer baths. In this bleaching process, the remaining silver particles are dissolved and leave voids with lower index of refraction than the surrounding gelatin (Fig. 2).

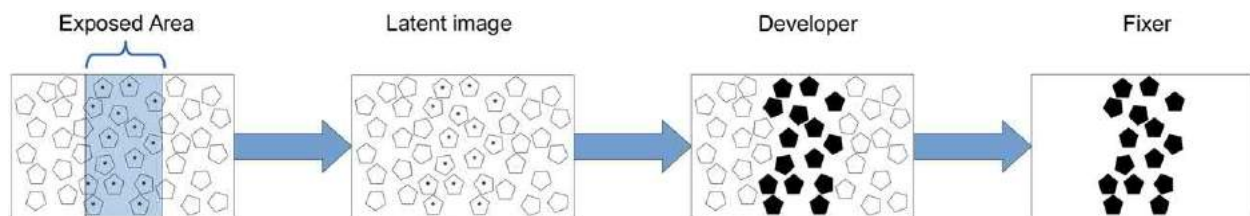


Fig. 1 Exposure and post processing of silver halide emulsion for amplitude modulation.

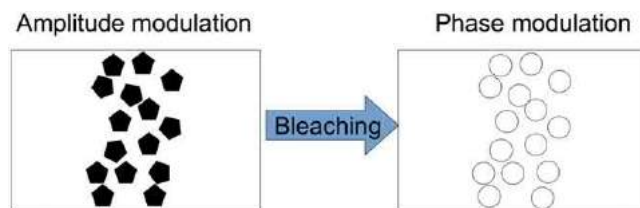


Fig. 2 Transformation of the amplitude modulation (absorption) into phase modulation (refractive index) by bleaching.

The advantage of silver halide emulsion is: a very large spectral sensitivity that covers the entire visible region (panchromatic) and even part of the infrared. This spectral sensitivity allows the recording of full color reflection holograms. Silver halide emulsions also have a very good sensitivity in the order or better than mJ cm^{-2} . This sensitivity is especially valuable when compared to other holographic recording materials that require much higher energy density. The spatial resolution of silver halide could be made such that it allows the recording of both transmission and reflection holograms. However, there is a trade-off between the spatial resolution and sensitivity. So, a careful selection of the emulsion according to the application is advised.

The disadvantages of silver halide material are the relative complexity of manufacturing and processing. Indeed, the chemistry, the application of the emulsion, and the careful handling in dark room, make silver halide a complex material to fully master. Other disadvantages are the limited shelf life, the material shrinkage during processing, and the possible scattering that occurs due to the bleaching process. This scattering is caused by the relatively large size of the voids left when the silver grains are removed, leaving some haze when the holographic image is rendered.

The shrinkage of the emulsion is particularly inconvenient for the production of reflection holograms. Any material thickness change between recording and display induces chromaticity change, and to a lesser extent, optical aberrations. It is possible to control, and even take advantage of the shrinkage by both pre and post processing, but this requires fine-tuning of the procedure.

For further reading about silver halide emulsion for holography see [Bjelkhagen \(1993\)](#).

Dichromated Gelatin

Dichromated gelatin (DCG) is quite similar to silver halide. The medium in both cases is gelatin, which is mainly composed of collagen, a natural protein, and both materials can produce very efficient phase holograms. However, in the case of DCG, the sensitizer is ammonium dichromate $(\text{NH}_4)_2\text{Cr}_2\text{O}_7$ (potassium dichromate is sometime used $\text{K}_2\text{Cr}_2\text{O}_7$), which is easier to obtain than fine silver halide crystals. The trade-off for this easier manufacturing is that both sensitivity and spectral response of the DCG are more limited than silver halide. DCG usually requires exposure in the order of several mJ cm^{-2} and is only sensitive to the blue and green region of the spectrum with a response of about 1–5 for these different wavelengths (1:488 v 5:532 nm). Sensitization of gelatin in the red has been demonstrated using other colorants, such as methyl blue, but this requires even higher energy exposure.

This requirement for large energy means that DCG needs high power lasers for the recording, or longer exposure time. Longer exposure time puts high constraints on the stability of the setup, and eventually a phase stabilization system could be required.

Once the material has been exposed, the dichromate is removed from the emulsion by a fixer bath and rinsed with water. The water also swells the gelatin, which creates larger voids where the gelatin was not exposed. The exposure has cross-linked the collagen molecular chains of the gelatin making it harder. The water is then removed by successive baths of increased concentration of isopropyl alcohol, to end up with a 100% concentration alcohol final bath. After processing, the gelatin film is fully transparent through all of the visible spectrum, as it should no longer contain any dichromate.

The index modulation of the DCG hologram can be finely tuned by:

- The aging of the material before or after exposure, old material is harder and has lower index modulation.
- The exposure energy, very low or high exposure leads to lower index modulation.
- The development process. In this final case, the temperature of the water bath is particularly critical. Higher temperatures produce a higher index modulation.

The modulation can be increased to up to 10% of the average index, which is one of the highest index modulation for any holographic recording material. Past that point, the size of the voids causes Mie scattering and the gelatin film becomes hazy.

In the case when the developing process did not produce the desired index modulation, the gelatin can be reprocessed starting at the water bath, reducing or increasing the modulation by changing the temperature of the water. This is particularly useful when manufacturing dispersing elements where the blaze and superblaze curves (dispersion spectrum) need to be finely tuned.

Yet another advantage of the DCG is its spatial resolution. For silver halide the resolution is determined by the size of the crystals. For DCG, the resolution is due to the photo-reduction of the chromium ions, so resolution up to $10,000 \text{ lp mm}^{-1}$ can be achieved.

To avoid water re-absorption by the gelatin once it has been processed, it can be encapsulated between glass plates using optical glue. The gelatin media is particularly resistant to UV and thermal degradation.

For more information about DCG, see [Stojanoff \(2011\)](#).

Photopolymers

This group of materials are organic synthetic media that undergo a polymerization initiated by the illumination. In addition to this polymerization, the photopolymers also experience a change of refractive index, so the hologram is recorded as a phase modulation. This change of refractive index is not directly due to the polymerization, but rather to the diffusion of the remaining monomers to the polymerized regions. This diffusion by osmosis locally increases the density of the material in these regions and increases the index. When the modulation reaches the desired value the diffusion process can be halted by fully polymerizing the material, which can be achieved by UV illumination or thermal treatment (heating) depending of the material formulation.

The holographic recording process in photopolymer is presented in [Fig. 3](#) where the three steps: photo-polymerization, diffusion, and final polymerization are illustrated.

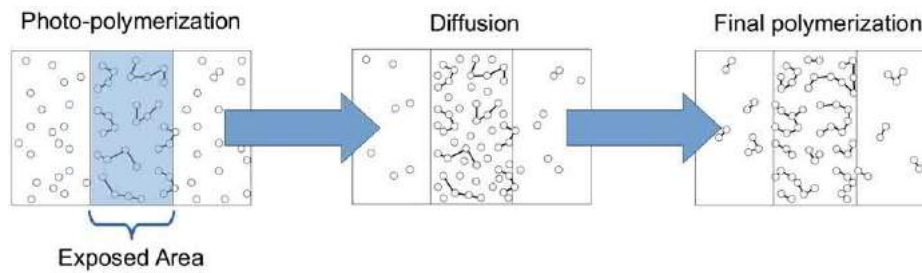


Fig. 3 Holographic recording process in photopolymer.

The sensitivity of photopolymers is on the order of 10 mJ cm^{-2} and they can be sensitized to respond to the entire visible spectrum. The spatial resolution can reach up to 4000 lp mm^{-1} , and the amount of shrinkage is minimal between recording and full polymerization, which is extremely useful for accurate color reproduction. Photopolymer also benefit from a prolonged shelf life and most formulations can be kept in the dark for years without any alteration. Very old samples (5+ years) have shown some crystallization though.

Commercially available photopolymers are the simplest holographic recording material to use. The post processing by UV illumination can be accomplished by direct sunlight exposure, and does not require a dedicated source.

The disadvantage of using commercial photopolymers is the inability to control the thickness of the material, which can be an important factor for dispersive elements. Also the backing medium that encapsulates the material is not under the control of the user, and in some cases has been reported to induce birefringence.

For more information about photopolymers, see [Guo et al. \(2012\)](#), as well as [Berneth et al. \(2011\)](#).

Photoresists and Embossed Holograms

The photoresists used for holography recording are the same materials as the one used by the semiconductor industry for lithography. Once exposed to light, the photoresist changes its ability to be dissolved by a solvent. One can distinguish between the positive resist where the parts illuminated become soluble, and the negative resist where the parts illuminated becomes insoluble. For holography this means that in both cases the modulation recorded is a surface relief.

The surface relief modulation can be used as a holographic pattern without any further modification. In this case, the hologram is used in transmission and the optical path length difference is given by the thickness modulation times the difference between the index modulation of the material and the air. The surface relief can also be coated with a reflective material such as metal. In this case the hologram is used in reflection and the path length difference is twice the thickness modulation.

The most common use for photoresist is to be a master copy for embossed hologram recording. The process is presented in [Fig. 4](#), where the surface relief structure of the photoresist is copied by electroplating. During electroplating, a thick layer of metal is grown on the top of the structure. This metal is used as a stamp to reproduce the relief in a heated thermoplastic material by embossing. The thermoplastic material is then coated with metal and further encapsulated by a protective layer. This process is used for mass producing holograms such as security tags for credit card or as an anti-counterfeiting measure for luxury items.

Because most photoresists have been developed for the photo-lithographic process, they have been optimized for short wavelengths such as UV light where smaller details can be imaged. Photoresist spectral sensitivity decreases dramatically at longer wavelength, such as in the visible. The holographic structure can be recorded by the interference of two UV laser beams, but more frequently it will be transferred to the photoresist by a lithographic process, either UV mask exposure, or by direct beam writing.

For more information about photoresist, see [del Campo and Greiner \(2007\)](#).

Photo-Thermo-Refractive Glasses

Photo-thermo-refractive Glasses (PTRG) are inorganic glasses such as $\text{Na}_2\text{O} - \text{ZnO} - \text{Al}_2\text{O}_3 - \text{SiO}_2$ doped with silver (Ag), cerium (Ce), and/or fluorine (F) atoms. These glasses are highly transparent in the visible and infrared (from 350 nm to 2700 nm), but UV exposure below 350 nm followed by thermal precipitation of crystalline phase produces a decrease of the refractive index. This phenomenon can be used to record phase volume holograms.

The maximum value of the refractive index modulation for PTRG is quite small (10^{-3}) compared to other holographic materials, but this is compensated by the very large material thickness and high transparency. Modulation times thickness ($\Delta n \cdot d$) in PTRG is enough to create very efficient volume hologram with diffraction efficiency exceeding 99%.

The photosensitivity of PTRG for 325 nm irradiation followed by 3 h of development at 520°C is about $1.5 \times 10^{-3} \text{ cm}^2 \text{ J}^{-1}$ which means that a standard exposure for high efficiency hologram recording is in the order of hundreds of mJ cm^{-2} . The spatial resolution of PTRG materials is particularly large ranging from 0 (continuous) up to $10,000 \text{ lp mm}^{-1}$.

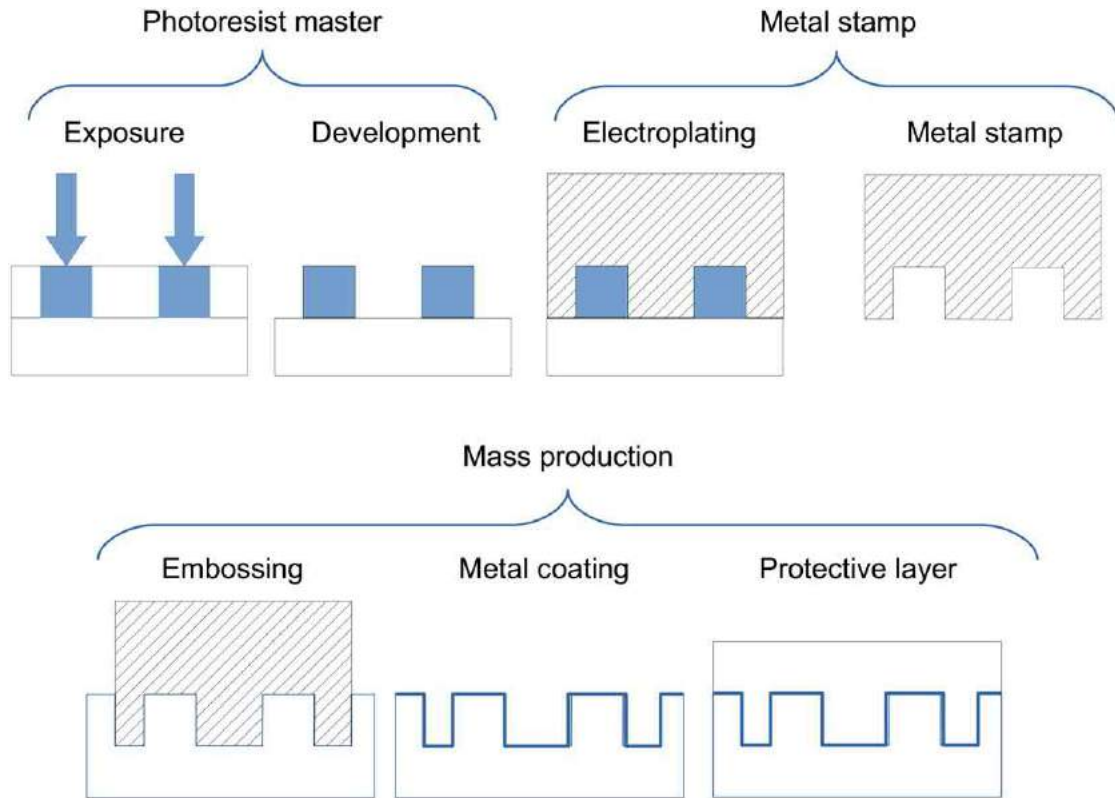


Fig. 4 Production of embossed hologram from a photopolymer surface relief structure.

PTRG materials are particularly important because they are useful for manufacturing the fiber Bragg gratings that can be found in telecommunication fibers and fiber lasers. The fiber Bragg grating is used as a notch filter to select a particular bandwidth and reject all the others. Based on this principle, optical add-drop multiplexers can be built.

Another major advantage of the PTRG materials is their very high damage threshold. Being particularly transparent, they can withstand nanosecond pulses with an energy exceeding tens of mJ cm^{-1} (measured at 1064 nm, and 10 Hz repetition rate), and an average power exceeding 100 kW cm^{-2} of CW irradiation (measured at 1085 nm). This allows for creation of holographic beam combiners for high energy lasers.

For more information about photo-thermo-refractive glasses, see [Glebov \(2002\)](#).

Holographic Sensors

The spectral characteristic of the light diffracted by an hologram is extremely sensitive to the index modulation (Δn), and the spacing between the Bragg planes (Λ). This can be seen from the Bragg's [Eq. \(3\)](#) that can be further differentiated to find the spectral dispersion:

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta n}{n} + \frac{\Delta\Lambda}{\Lambda} + \cot\theta\Delta\theta \quad (4)$$

Therefore, if a recorded hologram is altered by its environment by swelling, or shrinking, the characteristics of the diffracted spectrum will change. This spectral, or color, change can be used as a sensor to detect the component responsible for the thickness change.

Gelatin emulsion, for example, is very well known to absorb the air humidity and swell. Swelling separate the Bragg planes already recorded and makes the hologram color shifts toward the red part of the spectrum, and eventually disappear in the IR. It is possible to then restore the original color of the hologram by dehydrating the gelatin by heating it up.

Based on this example of the gelatin, several other polymer matrices have been used to make the hologram sensitive to a wide variety of components. One can cite, poly(2-hydroxyethyl methacrylate) (pHEMA), poly(acrylamide) (paam), or poly(vinyl alcohol) (PVA), and poly(dimethylsiloxane) (PDMS). These matrices are porous to a selection of solvents, and their absorption will make them swell.

It is also possible to dope the porous polymer matrix with other molecules such as crown ethers that will bind with specific metal ions present in the solution and make the hologram sensitive to that specific ion.

Another example is the addition of 3-(acrylamido)phenylboronic acid (3-APB) into pAAm matrix. The 3-APB molecule can bound with glucose which makes the hologram a potential sensor for monitoring bodily fluid sugar.

This technique can be generalized to any chemical reagent sensitive to specific molecules. Once incorporated into the porous matrix that has been cross linked to form an hologram (see the case for the gelatin in Section Dichromated Gelatin), the spectral dispersion is affected by the presence of that particular molecules, and the hologram can be used as a sensor.

For more information about holographic sensors, see [Yetisen *et al.* \(2014\)](#).

Refreshable Materials

In this section we will introduce the holographic recording materials where the hologram can be erased, and the material reused to record a new diffraction pattern. This class of material is not used to permanently display the hologram, but rather, to present it momentarily, analyze the diffraction, change the recording setup, and write a new interference figure.

These refreshable materials can further be distinguished between those which are dynamic and do not need post processing for the hologram to appear, and those which need post processing. The advantage of the dynamic materials is that they do not need to be moved away from their original position in the recording setup before being read. This allows for a greater stability and reproducibility in the experiment, especially in holographic interferometry. The advantage of the non-dynamic material is that the hologram is more stable and will last longer even during the exposure by the reading beam.

Photochromic Materials

The word photochromism describes a material that changes its absorption coefficient due to illumination. This change of absorption can be used to record amplitude holograms. However, the diffraction efficiency of phase holograms is much larger, and the Kramer–Kronig relationship specifies that a change of absorption always goes along with a refractive index change, eventually at a shifted wavelength. So, it is better to use the photochromic material for its index modulation than its absorption modulation. To do so, the writing wavelength is tuned near the absorption peak, but the reading of the hologram is performed in a more transparent region of the spectrum, where the index modulation is enhanced.

The photochromic effect is a macroscopic, observable modification of the material property, and can be generated by many different microscopic alterations. It can be permanent: such as photobleaching; or reversible: like in photo-isomerization. Photochromism can happen in inorganic material, such as doped glass, or inorganic molecules, such as azoic dyes.

In photochromic glasses, the SiO_2 matrix is doped with a metallic compound such as silver halide crystals. Under light excitation, the transparent silver halide molecules undergo a decomposition into metallic silver particles, which are opaque. Because the halogen molecules are trapped inside the glass and cannot escape, they can recombine with the silver after the illumination is gone, and the glass retrieves its initial absorption profile.

Photochromic glasses are not often used for holography due to their slow response time and low sensitivity.

On the other hand, there exist several types of organic molecules that react to illumination by reversibly changing their absorption spectrum and refractive index. The most commonly used in optics and more specifically for holography recording are spiropyrans, diarylethenes, and azobenzenes. When a photon is absorbed by one of these molecules, it undergoes a conformation change such as isomerisation.

Some molecules such as bacteriorhodopsin have several metastable excitation states that can be addressed with different wavelengths. By taking advantage of this particularity, the hologram can be recorded at one wavelength, read non-destructively with another, and erased by a third. The selection of the different wavelengths is such that the recording is done in a strong absorption region of the molecule, the reading away from it in a transparent part of the spectrum, and the erasing excites the molecule into an unstable conformation that decays back into the stable form. This process is illustrated in [Fig. 5](#).

For more information about photochromic glasses, see [Glebov \(2002\)](#). For more information about photochromic organic molecules, see [Dürr and Bouas-Laurent \(2003\)](#).

Persistent Spectral Hole Burning

When cooled down at cryogenic temperature (liquid helium ≈ 4 K), photochromic materials can experience an inhomogeneous broadening of their absorption spectra. This means that each of the absorption centers (molecules or ions) acquire a narrow spectrum that is shifted in frequency due to the interaction with the host matrix (either glass or polymer). The material spectrum still spans a large bandwidth, but individual centers possess a narrow line as shown in [Fig. 6](#).

The observation of the inhomogeneous broadening requires very low temperature. This is because at room temperature the spectrum broadening of individual centers is due to the interaction with phonons and other excitation. At cryogenic temperature, these interactions are minimized, revealing the inhomogeneous broadening of the material.

If the host matrix was perfect at cryogenic temperature each absorption center would be in the exact same state, and the spectrum would be a single narrow line. In these conditions, no inhomogeneous broadening happens and no spectral hole burning would be possible. There will be only one spectral line.

The hole burning itself happens when the inhomogeneous broadened photochromic material is illuminated with a narrow bandwidth source such as a laser. The source will only excite the centers with a spectrum overlapping the light frequency. Once

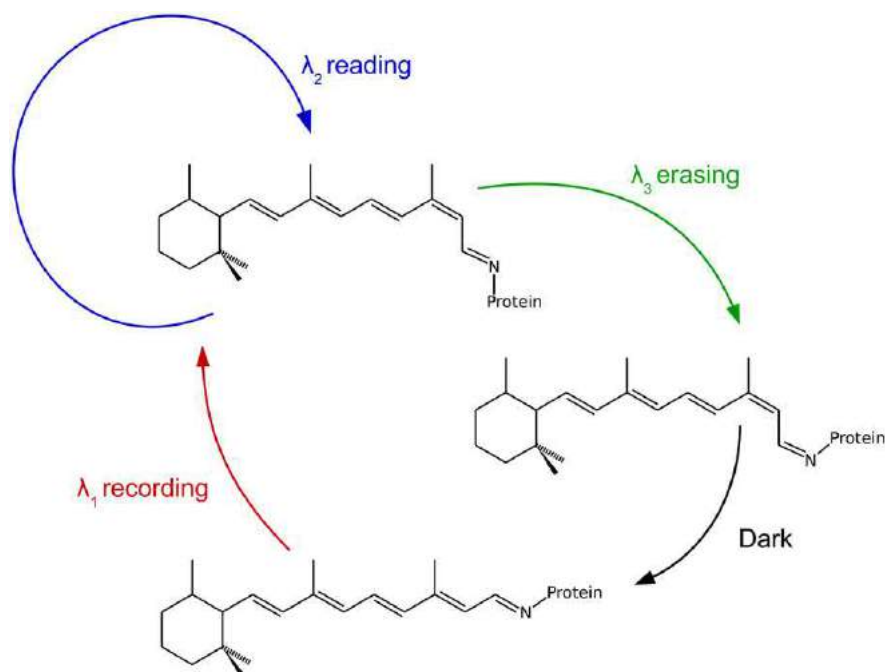


Fig. 5 Molecular structures of the various isomeric forms of bacteriorhodopsin, and the different wavelengths used for writing, reading and erasing an hologram.

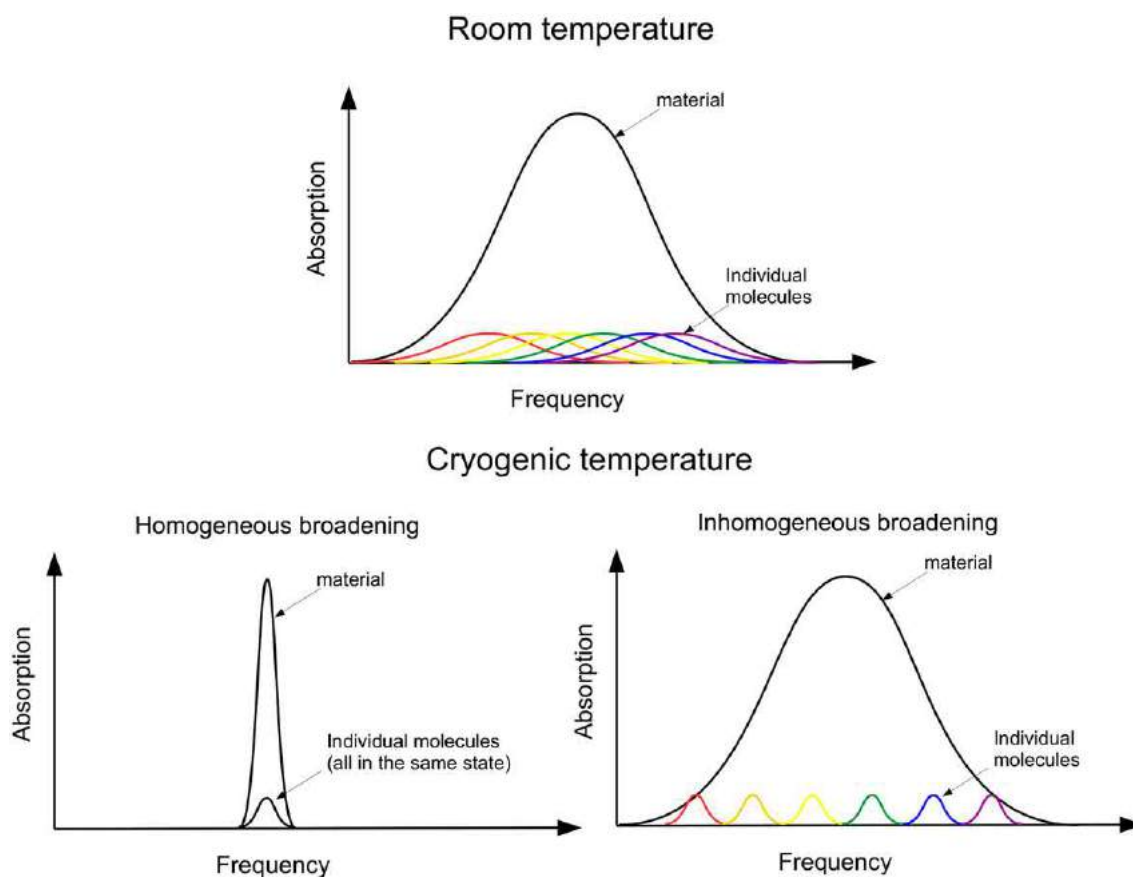


Fig. 6 Spectral broadenings. Top: at room temperature, the broadening is due to phonon interaction and the spectrum of individual center is large. Bottom left: at cryogenic temperature, in a perfect lattice, all the centers are in the same state and the spectrum of the material would be narrow. Bottom right: the spectrum of individual center is narrow but due to the interaction with the lattice each spectrum is shifted in frequency.

photo-excited, the molecule or ion composing the center will have a different absorption spectrum and leave a notch in the original absorption band as presented in Fig. 7.

An important parameter for spectral hole burning material is the ratio between the width of the notch that can be formed, Γ_{ZPL} (or zero phonon line), and the width of the material spectrum due to the inhomogeneous broadening, Γ_{inh} . This ratio indicates the number of individual frequencies that can be addressed in the material to record information, and can be as large as 10^6 .

Spectral hole burning materials can either be permanent or not. It is reversible if the photo-excited state of the absorption centers can decay back to the original state. The temporal behavior is strongly dependent on the material, the temperature, as well as the mechanisms of relaxation that bring back the excited state to its original state. Hours of dark decay time have been demonstrated.

The interest of spectral hole burning for holography comes from the possibility to use massive wavelength multiplexing. Each individual spectral notch that can be created is an opportunity to write a different and independent hologram. The use of the frequency domain to record the hologram is a new dimension for encoding the information in addition of time and space, which increases dramatically the quantity of information that can be stored in the material. Spectral hole burning gives the opportunity to increase the capacity of holographic data storage, and can also be used to design very high density correlation filters.

For more information about persistent spectral hole burning and its application for holography, see Moerner (1988).

Polarization Sensitive Materials

Some of the photochromic materials are sensitive to the light polarization and can be used to record not only amplitude-modulation holograms but also polarization-modulation holograms. This mechanism is also referred to as orientational hole burning in comparison to spectral hole burning.

Polarization holograms are formed when two coherent beams with orthogonal polarization interfere. In that case, the intensity is constant but the polarization vector changes along the period of the grating. For left and right circular polarized recording beams, the modulation is a linear polarized vector whose direction is rotating around the bisector of the beams propagation. For linear s- and p-polarized recording beams, the modulation is elliptical polarizations changing direction. These cases are presented in Fig. 8.

Sensitivity of photochromic molecules to polarization can be explained by the preferential absorption when the electric field is aligned to the axis of delocalized electronic orbital. This is the case of the azobenzene dyes that become oriented perpendicular to the light electric vector after multiple trans-cis photoisomerization and cis-trans relaxation cycles. During each cycle, the molecule

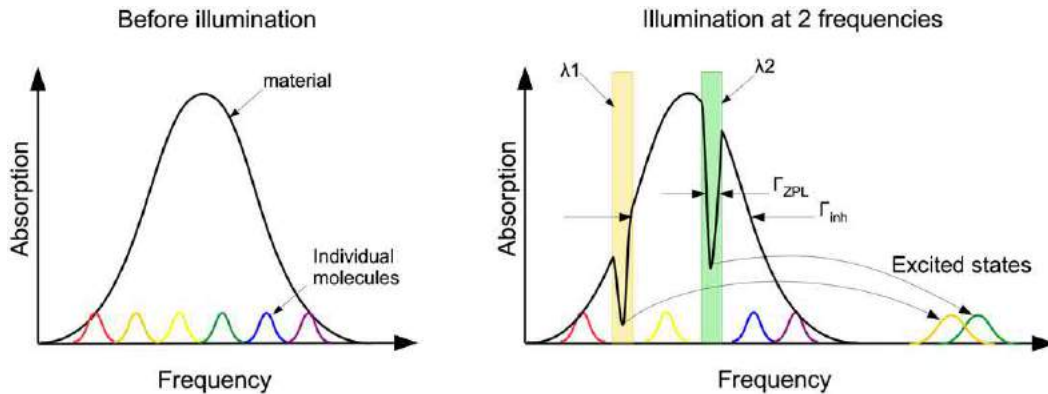


Fig. 7 Spectral hole burning principle: a narrow band source only excites the centers that have a spectrum overlapping the source frequency. The center excited states have a different spectrum which leave a notch in the material initial spectrum at the same frequency as the source.

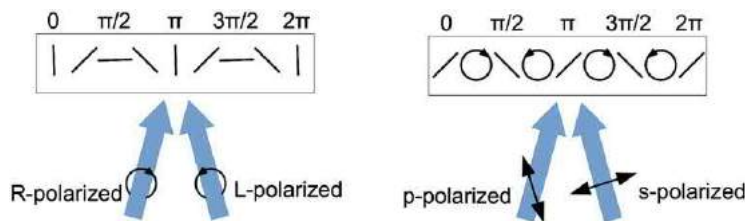


Fig. 8 Polarization gratings formed by the interference of, left: left and right circular polarized coherent beam. Right: s- and p- polarized coherent beams.

can rotate by a finite and random amount. Over a large number of excitation-relaxation cycles, the molecule distribution becomes predominantly organized with the main axes orthogonal to the polarization. The orientation is a statistical process and is not driven by any torque, which means that it is rather inefficient, explaining the slow dynamic time of the material in comparison to the very fast isomerization process.

In some materials, the multiple transitions also induce a molecular migration that can leave a surface relief modulation in addition to the volume phase hologram.

Even though the *trans*- and *cis*- forms of the azobenzene have different absorption spectra, the photo-orientation process leaves the molecules in their relaxed form (*trans*). The hologram formed in this case is a phase modulation due to the birefringence induced by the anisotropic molecular distribution.

Sensitivity of polarization sensitive azobenzene is on the order of mJ cm^{-2} , and the index modulation achievable is rather large, on the order of 10^{-2} .

For more information on polarization holography and polarization sensitive materials, see [Nikolova and Ramanujam \(2009\)](#).

Photorefractive Materials

The literal meaning of a photorefractive material is one that changes its refractive index under illumination. However, the scientific community has come to define the photorefractive process very specifically as the reversible and dynamic change of the index due to an electronic process. So, photorefractive materials should not be confused with photochromic even though the observable macroscopic effect is similar.

The photorefractive effect is a multiple step process that starts with the absorption of photons and the generation of electric charges inside the material. This aspect is similar to the photovoltaic process. Both electron and hole charges are created since there is conservation of the material electrical neutrality. However, in photorefractive materials, the mobility of the charge carrier depends on its sign. One type of charge carriers migrate inside the material while the other type stays in the localization where it was created. There are several charge transport mechanisms such as diffusion, drift (under external electric field), and photovoltaic that explain this migration. The mobile charge carriers are eventually trapped in the dark regions of the material, and the local charge distribution creates a space-charge electric field between the illuminated and dark regions of the material. This space-charge field modulates the index of refraction by either nonlinear electro-optic effect (in inorganic crystals) or molecular orientation (in organic materials).

The different steps leading to the photorefractive process is presented in [Fig. 9](#).

The photorefractive effect is characterized by a phase shift (φ) between the intensity pattern and the index modulation. This phase shift is responsible for the self-diffraction of the recording beams in a two beam coupling experiment.

Photorefractive materials can be organized into two different categories: inorganic crystals and organic compounds. Inorganic crystals are grown at high temperature and their composition is imposed by their stoichiometry and crystalline structure. Some inorganic dopants, such as metallic atoms, can be incorporated to modify the electro-optical properties like absorption band or carrier mobility. Examples of inorganic crystals include the silenites ($\text{Bi}_{12}\text{SiO}_{20}$, $\text{Bi}_{12}\text{GeO}_{20}$), strontium-barium niobate ($\text{Sr}_x\text{Ba}_{1-x}\text{Nb}_2\text{O}_6$), barium titanate (BaTiO_3), lithium niobate (LiNbO_3), as well as semiconductors such as GaAs, InP, and CdTe.

Photorefractive organic compounds are mixtures of several organic molecules exhibiting specific function to achieve the photorefractive effect. They are mainly composed of a photoconductive polymer matrix that allows the charge transport. Such matrices are PVK: poly (N-vinylcarbazole), or PATPD: poly (acrylic tetraphenyldiaminobiphenol). To enhance the sensitivity in the

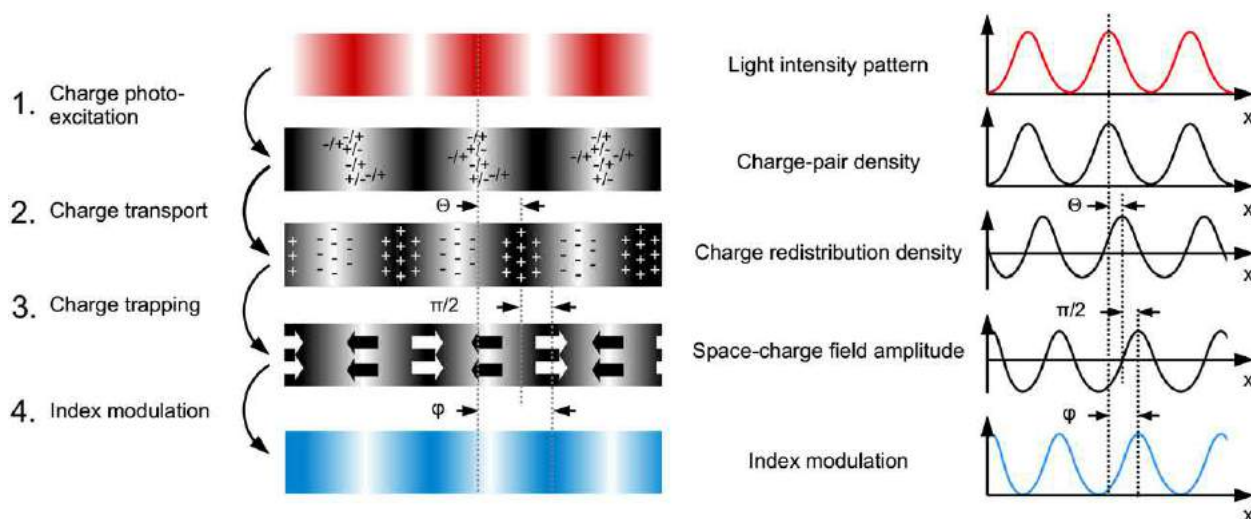


Fig. 9 Photorefractive process starting with the intensity distribution of the interference pattern, charge photo-generation, charge transport and distribution after migration and trapping, space-charge field, and final index modulation.

visible the matrix is doped with a sensitizer that is responsible for the charge photogeneration. The most used sensitizer that covers a large band in the visible region is a derivative of the C_{60} fullerene molecule, PCBM: phenyl-C61-butiric acid methyl ester. The index modulation is provided by chromophores, rod-like molecules with large dipolar moment and polarizability, such as DMNPAA: 2,5-dimethyl-4-nitrophenylazoanisole, or 7-DCST: 4-azacycloheptylbenzylidene-malonitrile. To increase the mobility of the chromophores that need to orient in the space-charge field, the glass transition of the polymer matrix is lowered with plasticizers, low molecular weight molecules such as ECZ: N-ethylcarbazole, or BBP: n-butyl benzyl phthalate.

The optical properties of the photorefractive crystals and organics are strongly dependent on their nature and cannot be generalized. Their sensitivity ranges from $\mu\text{J cm}^{-2}$ to mJ cm^{-2} , and their spectral response varies through all the visible up to the infrared.

Due to their very specific properties, photorefractive materials are used in numerous applications, such as phase conjugation, coherent beam amplification, imaging through turbid or scattering media, dynamic holographic imaging, and holographic interferometry.

For more information on photorefractive inorganic crystals, see [Günter and Huignard \(2007\)](#). For more information on photorefractive organic materials, see [Blanche \(2016\)](#).

Photo Thermoplastic Process

The term Photo Thermoplastic does not describe a material, but a material process. Thermoplastic polymers have their structural rigidity altered by temperature. By heating a sheet of material it becomes supple and can be deformed. This property can be exploited to record a surface relief hologram which is encoded in the deformation of the surface of the sheet. To do so, the thermoplastic material is coated on the top of a photoconductor. The surface of the photoconductor is charged by corona discharge which attracts opposite charge on the top of the thermoplastic. When illuminated, the charges that were on the top of the photoconductor migrate through the material and came in contact with the thermoplastic. After illumination, the photoconductor is charged a second time. Because of this second charging, there is now more charge in the region that was illuminated than where the material was not exposed to light. At that moment, the thermoplastic material is heated up and the electrostatic attraction between the charges squeeze the film, modulating its thickness. To erase the surface relief, the material is evenly charged and heated. These processes are illustrated in [Fig. 10](#).

For more information about the photo thermoplastic process, see [Lin and Beauchamp \(1970\)](#).

Electronic Devices

Twenty years ago, there was a technological transition from analog to digital photography. This transition was possible thanks to the development of the focal plan array detector and the personal printer. Today, we are assisting to the same kind of transformation for holography. More and more often, the analog recording media is substituted for an electronic device. This transformation is made possible by the high resolution of both the detectors and the micro display devices. Once again, when photography requires micrometer resolution, holography, which records the wavefront features, requires nanometer precision. Needless to say, this is much more difficult to achieve.

The advantages of the electronic devices are numerous. They are dynamic and do not require post processing to capture the hologram, they can be used over a very large wavelength bandwidth, they allow computer manipulation of the holographic pattern for direct interpretation or eventual transformation, they are easy to use, and they are reusable with no pre-processing.

The figure of merit for the recording and reproduction of holograms by electronic devices is the space-bandwidth product (SBP), which is a measure of the rendering capacity of an optical system. It is defined as the product of the spatial frequency bandwidth and the spatial extent of the image. In other words, at constant SBP you can either capture/reproduce a small image at high resolution, or a large image at low resolution. The transformation from one to another can simply be performed by a magnifying lens.

Although very useful and convenient, electronic devices are still suffering from a coarse resolution and pixel count compared to the recording materials (small SBP), but this might be solved in the very near future. A more fundamental problem is their inability to record and reproduce volume holograms. Detectors only record the interference along a plane, and the current micro-

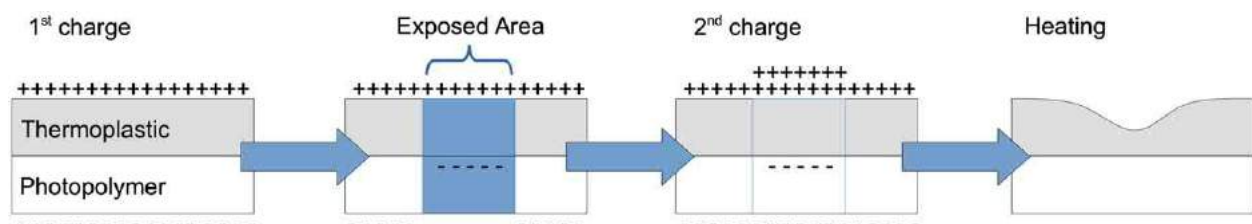


Fig. 10 The photo thermoplastic process: charging the material, exposure drives the charges inside the photopolymer, the second charging increases the electrostatic attraction where the material has been exposed, heating the thermoplastic modulates the surface relief.

displays only produce surface relief modulation. For a volume hologram to be recorded, the phase needs to be sensed over a non-negligible thickness, and reproduced likewise. This excludes both the sensing and reproduction of reflection hologram by electronic devices.

Even with these limitations, electronic devices are used in a large variety of holographic setups, from 3D display to microscopy and adaptive optics. They are promising the same bright future for holography as the one they created for photography.

Focal Plane Array Detector

The function of detecting the intensity modulation generated by the interference of the coherent beams is provided by the focal plane array detector (FPAD). Two technologies can be used: either charge-coupled devices (CCDs), or complementary metal-oxide-semiconductors (CMOSs).

These devices have been popularized by digital photography and are very well known, even outside the scientific community. It is only recently, however, that the pixel pitch and count made them interesting for holography. A commercial grade FPAD can now have 20 million pixels with a 4-micron square size. This size of pixel correspond to a resolution of 125 lp mm^{-1} which is far from the several thousand that analog materials offer, but is enough for detecting interference fringes for in-line holography, or low angular separation beams at long wavelength.

Scientific grade FPAD can have a pixel size down to 2.2 μm for a total of approximately 120 million pixels on a single sensor. The problem with decreasing the pixel size is the same as with the silver halide emulsion: when the size shrink, there are fewer and fewer photons interacting with the pixel (crystal), and the sensitivity decreases dramatically. This is particularly true for the CMOS technology where the pixel surface is shared between the sensing area and the electronic amplification and logic.

Once the interference pattern has been detected by the FPGA, it can be used to reconstruct the 3D light field (phase and amplitude) using a computer, or displayed by another electronic device to reproduce the hologram.

For more information about the use of focal plane array detector for holography, see Schnars and Jüptner (2002).

Acousto-Optic Modulator

A sound wave is a compression wave propagating inside a material. This compression (and dilatation) modulates the index of refraction and if the frequency is carefully selected, can be used to diffract the light. Acousto-optic modulators (AOM), use this principle to generate diffraction gratings. They are composed of a piezo-electric transducer (PZT) that generates the wave, which is coupled to a transparent medium, generally glass, quartz, or other crystal. Two configurations are presented in Fig. 11. The first one is the commonly used Bragg cell, where the sound wave generated by the transducer travels inside the material. The second is one where the transducer is put on top of a waveguide and generates a surface acoustic wave. The wave diffracts the light out of the waveguide so this configuration is called leaky-mode coupling.

There are two modes of operation for the Bragg cell: static and traveling wave. When the acoustic wave is traveling, it shifts the diffracted beam frequency due to the Doppler effect. This shift is about 1 GHz due to the speed of sound in the material. To avoid this shift, a sound reflector can be used at the other end of the material. This reflection generates a standing wave like in musical wind instruments.

By changing the frequency and the amplitude of the vibration at the PZT, it is possible to change the diffraction angle and amplitude. AOM can achieve a diffraction efficiency up to 99%.

The sound wave can only generate diffraction gratings that redirect the incident beam but do not change its wave front. Used simply as is, AOMs would not be able to form a 3D image. However, if instead of using a fixed frequency, the AOM is driven with a sum of many frequencies, it becomes similar to a one-dimensional arbitrary holographic pattern: any waveform can be decomposed as a sum of sinusoids.

The advantage of the AOMs is their very large space-bandwidth product compared to the spatial light modulator. An AOM can produce larger images size, and larger view angle than the other technologies.

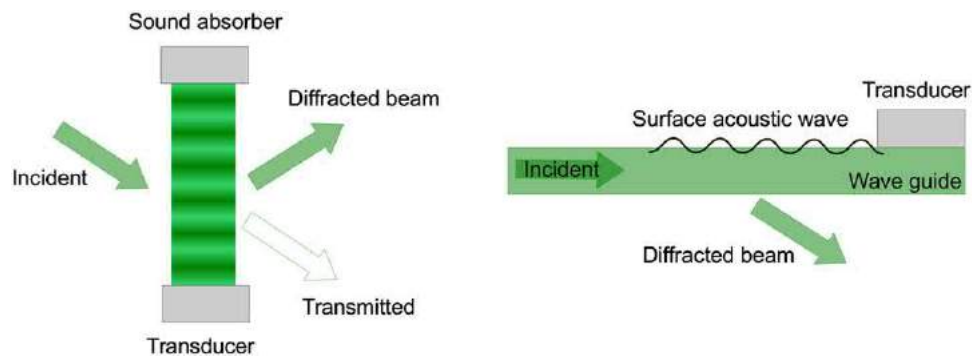


Fig. 11 OAM configurations. Left: Bragg cell where the sound wave generated by the transducer modulates the index inside the material. Right: leaky-mode mode coupling where the transducer produces a surface acoustic wave on top of a waveguide.

Due to the very fast reaction time, AOM are driven by a signal ranging from tens to hundreds of MHz. AOMs are used for Q-switching pulsed lasers, for pulse shaping, in telecommunication to modulate the optical signal so it carries the information, and in spectroscopy for frequency control.

For more information about acousto-optic modulator, see [Efron \(1994\)](#).

Spatial Light Modulator

Spatial Light Modulators (SLM) are dynamic pixelated electronic devices where each pixel can be individually controlled to change the amplitude or the phase of an incoming light beam. Initially developed for the display industry, SLMs are also called micro displays because of the small size of the pixels. Thanks to that small size, they can also be used to display holographic patterns, which diffract the incident visible light over an appreciable angle. Larger pixels reduce the angle of diffraction according to the Bragg equation and need to be used at larger wavelengths.

Micro displays where the light is directly emitted from the pixels such as in light emitting diodes (LEDs) and thin film transistors (TFTs), cannot be used for holography since the phase of the different sources cannot be controlled individually. A micro display system where the phase of the light emitter could be controlled would be the equivalent of a phase array radar for the visible and is currently the object of active research.

Because SLMs are dynamic and refreshable at will, they are extremely attractive for dynamic holography applications such as optical tweezers, optical switching, non-mechanical scanner, wave front correction, and holographic interferometry.

Of course, 3D display is also a potential application, but the space-bandwidth product of commercially available SLMs (number of pixels over the pitch) is still too small to generate a comfortable 3D image, even with a 4 K UHD resolution: 4096×2160 pixels. Systems tiling several SLMs together to increase the SBP have been demonstrated though.

Two types of SLMs can be distinguished according to their mode of operation: liquid crystal on silicon (LCoS) and micro-opto-electro-mechanical systems (MOEMS).

Liquid crystal on silicon

Liquid crystals (LCs) are liquid form materials composed of birefringent molecules that self-align. Due to their large dipole moment, the LC molecules are also sensitive to externally applied electric field. When an electric field is applied the molecules rotate, which changes the optical properties of the material: birefringence amplitude and axis, as well as the refractive index.

In LCoS SLMs, a layer of LC is deposited on top of a complementary metal-oxide-semiconductor (CMOS) structure forming individual cells where the voltage can be applied independently. This is the same technology used in liquid crystal display (LCD) and LC television. When a voltage is applied, the LCs induce phase retardation in the polarized incident light as large as a few wavelengths, which is enough to produce a fully modulated (2π) holographic phase pattern. This phase modulation can take multiple values (usually 8 bits = 256 levels), which allow the LCoS to reproduce continuous phase holograms.

It has to be noted that at maximum spatial frequency, the pattern is only represented by 2 pixels per period, and only binary phase can be displayed. To reproduce multilevel functions, several pixels need to be used and the frequency is reduced.

To increase the reflectivity of the device, a layer of aluminum or dielectric mirror is coated between the LC and the CMOS electronic. The mirror can be ordered according to the wavelength at which the device is supposed to be used (IR, visible, or specific laser line) which give a reflectivity larger than 90%.

Amplitude modulation LCoS displays use a polarizer in front of the SLM so the phase retardation is converted into amplitude modulation following Malus's law: $I = I_0 \cos^2 \rho$, where ρ is the angle between the polarization vector and polarizer direction. This mode of operation is not interesting for holography since amplitude holograms only reach 10% in the first order.

A phase only LCoS is more interesting for holography due to the higher diffraction efficiency (up to 100%). In this case, the device modulates the index of refraction directly, according to the main axes of the birefringence.

Typical LCoS pixel pitch is a few microns, which offers a maximum diffraction angle of a few degrees in the visible (4 mm pitch $\approx 4^\circ$ in the visible). The pitch is limited by the field bleeding from one cell to another and it would be hard to reduce it further in the future. An LCoS benefits from a fill factor (active pixel area over pixel size) as large as 90%. The refresh rate is limited by the viscoelastic relaxation of the molecules but can increase up to a few hundred Hz. Unfortunately, this refresh rate does not allow for the time multiplexing or very fast reconfiguration sought in some holographic applications.

For more information about LCoS SLM and holographic application, see [Osten and Reingand \(2012\)](#).

Micro-opto-electro-mechanical systems

Micro-electro-mechanical systems (MEMSs) are devices where a micron scale mechanical feature can be activated by an applied voltage. When this device is also used to interact with light, it is named MOEMS with the introduction of the term "opto".

The best known of these MOEMSs, due to its commercial success, is the Texas Instruments DMD (digital micromirror device) which is also known as the DLP (digital light processor). The DMD is used in display applications such as televisions and projectors. It is composed of an array of micron size mirrors: $13 \mu\text{m}$ for the $0.7''$ XGA, down to $5.4 \mu\text{m}$ for the 1080p "pico". The size of the DMD follows the resolution standard for display: XGA = 1024×768 , 1080p = 1920×1080 , WQXGA = 2560×1600 .

The DMD is a binary device and the mirrors can only take two orientations: tilted left or right by 12° for the large DMD, or 17° for the pico, according to the surface normal. Accordingly, the DMD can be used to display binary amplitude holograms. The incident light is reflected by the mirrors, and since the pattern is composed of left and right tilted mirrors there are two reflection

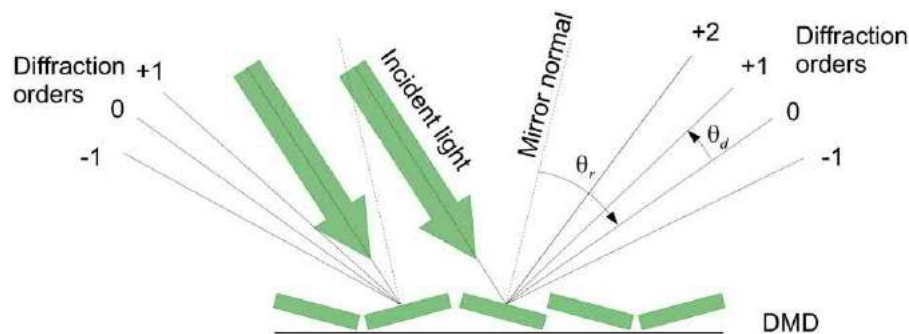


Fig. 12 Geometry of diffraction from the DMD.

directions θ_r (or zero orders). In each of these directions, there are multiple diffraction orders ($\pm 1, \pm 2, \dots$) due to the diffraction by the structure created by the mirror orientation (see Fig. 12). The maximum angle of diffraction (θ_d) around the 0 orders is given by the pixel pitch and is about 2° for $13 \mu\text{m}$.

The maximum diffraction efficiency for the binary amplitude hologram displayed by the DMD is 10.1% as predicted by the Fourier decomposition theory. However, the big advantage of the DMD is its refreshing rate, which can be up to 20 kHz (100 times faster than LCoS). This enables applications that are not accessible with LCoS SLMs.

That fast refresh rate is what allows the DMD to display numerous intensity levels (10 bit gray scale), although it is a binary device that can only display black (light reflected away from viewer) or white (light reflected toward viewer). By oscillating the mirror at different frequencies, the intensity directed to the viewer is modulated. This only works because the eye integration time is rather slow ($\approx 24 \text{ Hz}$) compared to the 20 kHz of the device. DMDs cannot be used for applications requiring smaller integration time.

Another diffractive MOEMS worth mentioning is the grating light valve (GLV) which is composed of parallel thin ribbons (few micron pitch) that can take two positions: up or down. The phase difference generated by these two positions can be used to diffract the incident light with 40% efficiency (binary phase hologram). The GLV has been shown to have a refresh rate of up to 50 GHz (2000 times faster than the DMD), but also has some severe limitations for holography: it is a 1-dimensional array and only 1 or 2 ribbons can be activated at a time.

Other types of MOEMS are under development specifically for holography. One such device is an analog piston MOEMS where the micro-mirrors are moved up or down according to the voltage applied. Like in the GLV, this movement creates a phase difference between the mirrors which diffracts the light. The difference from the GLV is that the mirrors can be positioned at multiple levels instead of just two. This allows users to display a continuous phase hologram which can have a diffraction efficiency as high as 100%. Such a piston MOEMS will combine both the efficiency of the LCoS, and the speed of the DMD. If a piston MOEMS can be manufactured with a very large number of mirrors (hundreds of trillions), it could enable true holographic television along with other diffractive applications.

For more information about MOEMS SLM and holographic application, see [Blanche et al. \(2015\)](#) as well as [Blanche et al. \(2017\)](#).

Acknowledgements

The author would like to thank Mr. Colton Bigler for the careful revision of the manuscript.

See also: Overview: Holography

References

- Berneth, H., Bruder, F.-K., Fäcke, T., et al., 2011. Holographic recording aspects of high-resolution bayfol[®] hx photopolymer. Proceedings of SPIE 7957, 79570H.
- Bjelkhagen, H.I., 1993. Silver-Halide Recording Materials: For Holography and Their Processing. Springer.
- Blanche, P.-A. (Ed.), 2016. Photorefractive Organic Materials and Applications. Springer.
- Blanche, P.-A., Banerjee, P., Moser, C., Kim, M.K., 2015. Special section guest editorial: Special section on the interface of holography and mems. Journal of Micro/Nanolithography, MEMS, and MOEMS 14 (4), 041301.
- Blanche, P.-A., LaComb, L., Wang, Y., Wu, M.C., 2017. Diffraction-based optical switching with MEMS. MDPI Applied Sciences 7 (4), 411.
- del Campo, A., Greiner, C., 2007. Su-8: A photoresist for high-aspect-ratio and 3d submicron lithography. Journal of Micromechanics and Microengineering 17 (6), R81.
- Dürr, H., Bouas-Laurent, H., 2003. Photochromism: Molecules and Systems. Gulf Professional Publishing.
- Efron, U., 1994. Spatial Light Modulator Technology: Materials, Devices, and Applications, 47. CRC Press.
- Glebov, L., 2002. Photochromic and Photo-Thermo-Refractive Glasses. John Wiley & Sons, Inc.

- Günter, P., Huignard, J.-P., 2007. Photorefractive Materials and Their Applications. Springer.
- Guo, J., Gleeson, M.R., Sheridan, J.T., 2012. A review of the optimisation of photopolymer materials for holographic data storage. *Physics Research International* 2012.
- Lin, L.H., Beauchamp, H., 1970. Write-read-erase in situ optical memory using thermoplastic holograms. *Applied Optics* 9 (9), 2088–2092.
- Moerner, W.E., 1988. Persistent Spectral Hole-Burning: Science and Applications. Berlin, Heidelberg: Springer.
- Nikolova, L., Ramanujam, P.S., 2009. Polarization Holography. Cambridge University Press.
- Osten, W., Reingand, N., 2012. Optical Imaging and Metrology: Advanced Technologies. John Wiley & Sons.
- Schnars, U., Jüptner, W.P.O., 2002. Digital recording and numerical reconstruction of holograms. *Measurement Science and Technology* 13 (9), R85.
- Stojanoff, C.G., 2011. A review of selected technological applications of dcg holograms. *Proceedings of SPIE* 7957, 79570L. (-79570L-15).
- Yetisen, A.K., Naydenova, I., Da Cruz Vasconcellos, F., Blyth, J., Lowe, C.R., 2014. Holographic sensors: Three-dimensional analyte-sensitive nanostructures and their applications. *Chemical Reviews* 114 (20), 1065410696.

Colour Holography: Perception Versus Technical Reality

Andrew Pepper, Nottingham Trent University, Nottingham, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

The opportunity for optical holography to record and display three-dimensional objects (and the volume surrounding them), at extremely high fidelity, has prompted a distorted public perception. Almost since we became aware of the process, and its capabilities, there has been an assumption that the remarkable dimensional images would be in full (natural) colour. This is not a surprising supposition due to the clarity and illusionistic impact of holographic images displaying full parallax (Benton, 1977). Our familiarity with looking into display spaces, very much like looking through a window into a room in a building (Pepper, 2000), suggests that what we appear to 'see' must surely be in full colour.

Most of the first holograms produced were monochromatic; a limitation of the technical processes which created them. When Dennis Gabor first published the announcement about his work with optical holograms in 1947 (Gabor, 1948), he used a heavily filtered mercury arc lamp to record and illuminate his photographic plates, which stored the microscopic interference patterns on which holograms rely. This was the only suitable and accessible source of light at the time, which fit the specific requirements of his optical recording configuration, but it did limit the size of the object he could record to a few millimetres.

These early holograms used Gabor's 'in-line' recording process, which placed the 'object' to be recorded directly between his chosen source of light and the photographic plate. Light waves, directly from the illuminating source, and those modulated through diffraction by the object, were combined on the photographic plate to record the resulting, microscopic, interference wave pattern.

The reference to an 'object' here is intentionally placed in inverted commas as Gabor used a flat photographic transparency for his recording. Although the transparency is physically three-dimensional, and millimetres thick, there is a tendency to describe it as flat because of our learned visual, perceptual and verbal framework used to view and describe 3-D. This aspect of these early recordings, and the demonstration of the optical and physical viability of this new method for recording three-dimensional objects caused a slight perceptual and cultural friction. If it is three-dimensional, it is likely to be incredibly accurate, and if it is so accurate, it is reasonable to assume that it is correctly coloured. Feedback from visitors to exhibitions of optical holography includes incorrect assumptions that some of the monochromatic images they have encountered are, when recalled and discussed, full colour. We have a natural tendency to 'fill in the gaps', and this is incredibly easy to do when confronted with the astounding three-dimensional illusion of a holographic image.

The tendency to make such an incorrect assumption is not an isolated phenomenon. Evidence through teaching fine art to undergraduate students suggests that they encounter much of the art they look at (or explore for research) online. They view multiple images of the same work, made up of gallery installation shots, close-up details and peripherally photographed views, which are scattered across the web or in printed catalogues, and often believe they have 'seen' the original. It is only when they then encounter the actual work that they realise they have not seen the work, only its documentation.

A major shift in public perception of holography was directly related to two significant developments of the technical process in the United States and USSR. Emmett Leith and Juris Upatnieks developed an off-axis geometry, which used coherent, monochromatic, laser light, to record and display holograms. This more powerful, coherent, light source, and the optical pathway they developed, allowed the recording of 'actual' three-dimensional objects (rather than Gabor's 'flat' transparencies).

The results of their research were disseminated to the scientific community (Leith and Upatnieks, 1962) but an article in the widely circulated magazine 'Scientific American' (Leith and Upatnieks, 1965) brought it to the attention of a wider audience, interested in broad scientific topics. Culturally this was a positive, future-looking, period with an enthusiastic public awareness of new technologies stimulated by the space race and rapid scientific advances. This publication informed a different demographic beyond the technical papers previously published in field-specific scientific conferences and journals. Unsurprisingly, the 'improvement' on Gabor's original invention, and its implied connection with photography (in the title of the article and through the text), captured the public view and that of popular media. Generalised information about optical holography spread quickly, with many popular articles appearing worldwide (Johnston, 2006). Subsequently, public interest and expectations about holography increased, which fuelled many inaccuracies - one of the main ones being that holograms were full-colour.

This public confusion about the realistic nature of holograms was evident during the late 1970s when two important exhibitions of large-scale pulsed laser holograms, produced by Nick Phillips (1933–2009) at the Royal Academy of Arts, London, proved extremely successful and not only attracted the attention of an 'art' sensitive public, visiting a renowned institute, but also that of the media. Some of the reports and images disseminated around this time blurred the line between what was possible with the technique and high public expectation. It was difficult to be sure whether the images from these exhibitions, reproduced in glossy colour supplements, might be full colour.

Leith and Upatnieks did propose colour holographic recording as early as 1964 (Leith and Upatnieks, 1964), but there was a need for the technical components to become available and accessible, mainly multiple colour, coherent, laser light and high-resolution recording materials.

During the same period, Yuri Denisyuk demonstrated significant developments, which also used an off-axis system of recording and display (Denisyuk, 1962). This incorporated a single beam of coherent laser light which passed through the holographic recording plate, bathed a three-dimensional object in laser light, and the subsequently reflected light from that object was combined with the laser light shining through the plate, to generate the required microscopic interference pattern. The advantage

of this 'single beam' technique (not to be confused with Gabor's in-line method) is that the resulting hologram could be displayed using white light. This significant development played an important part in the adoption of holography as a display (and archival process). All previous systems required expensive, and potentially dangerous, mercury arc (Gabor) or laser (Leith and Upatnieks) light. Here, inexpensive white lights, such as spotlights or direct sunlight, could be used to reconstruct the recorded three-dimensional images.

A drawback to Denisyuk's technique was that objects needed to be relatively shallow, compared to the amount of volume which was possible to record in Leith and Upatniek's technique. It appears the more accessible, English language, publications about Leith and Upatniek's activities placed them prominently in the public eye and that of the US and European media outlets. It is also worth noting that the term 'hologram' was not in common use before the mid-1960s, with terminologies such as wavefront reconstruction, diffraction microscopy, Gaboroscopy, holoscopy, wave photography and lensless photography, appearing in publications (Johnston, 2009).

The various techniques and optical processes were, theoretically, capable of recording full-colour holograms. As early as 1966, Lin and his team successfully produced the first full-colour reflection hologram, which used two different wavelengths of light, recorded and combined onto a single holographic plate (Lin *et al.*, 1966). This was, in fact, a recording of a colour photographic slide, which we might interpret as 'flat' but, like Gabor's original work, it demonstrated the feasibility of using multiple wavelengths (colours) of light which, combined, will produce the impression of full colour.

Red, Green, Blue

Most of the early holograms recorded, irrespective of the optical technique employed, used a monochromatic light source to record and display a three-dimensional object. As previously stated, the optical illusion and impact of the recorded images were so significant that the 'memory' and recall of seeing these holograms often prompted a distorted perception among viewers, thinking they had 'seen' the images in full colour.

An image recorded with red, helium-neon, laser light, for example, will display a tonal red, three-dimensional image, when displayed using the same red laser light. There will be no other colours present. Much of the early research on colour holography used the principle that it would be possible to use multiple colour lasers to record a hologram and then use the same lasers to display the results.

As white light is made up of multiple colours (wavelengths), using the light primary colours of red, green and blue will allow the recording and reconstruction of a full-colour image. This phenomenon of colour reproduction is based on the tristimulus theory of colour vision, which suggests that any colour can be reproduced (or matched) by combining the three light primary colours.

It is possible, therefore, to make a holographic recording using red laser light (helium-neon), then make a second exposure using green laser light (argon) and finally a third using blue laser light (helium-cadmium). The colour of a laser can vary, depending on the gases/materials used within the device to cause the light amplification process. This can include krypton ion (red), neodymium-doped yttrium aluminium garnet (green) and argon ion (blue).

Once processed, the developed hologram can then be illuminated with these three colours of laser light. In effect, each of the three-dimensional recordings in the hologram 'overlap' and then combine, spatially, to reconstruct the colours of the original object in their original position in space.

Three objects are presented, which overlap each other in space and combine to create a single, full-colour object/image. This is a unique property of holography, and one of the few opportunities to place visually solid, three-dimensional images of objects into the same space at the same time. Such 'overlapping' has always been possible with multiple exposure photography and use of multiple camera feeds in video, but only holography allows this in full parallax, full three dimensions.

During the 1970s, there was a considerable amount of research in the field of colour holographic recording, which took advantage of the technical improvement in light sources, recording materials and optical components (Hariharan, 1983).

Although the cost of lasers has dropped and their quality (and variety) improved, they remain complex, and in some cases fragile or dangerous (depending on their power and wavelength). This has limited their use in display. Clearly, the option of recording holograms with multiple colour laser light and then viewing them with white light is a practical solution, and a considerable amount of research has been undertaken, using reflection holography (often employing the Denisyuk process) with a significant pioneering research by Bjelkhagen (2006).

Research suggests that the optical quality and colour rendition of holographic displays can be enhanced using multiple wavelengths of laser light and more than three colour recordings (Mirlis *et al.*, 2005). The reproduction of colour has become accurate enough for it to be viable as an archival system for paintings, which can be captured using a contact recording technique, this not only accurately records the colours within the painting, but also the three-dimensional undulations of the painting surface. Each brush stroke, its thickness, and texture of the paint, is stored and can later be displayed (Bjelkhagen and Vukicevic, 2002). As an archival process, this has considerable potential – extending the visibility of paintings, which now have a restricted display life due to potential light damage. Maintaining their colour is a significant requirement of the preservation process even if, in this case, it is shown in a high-resolution facsimile.

One holographic technique, invented by Stephen Benton (1941–2003) in 1968 (Benton, 1969), impacted on the production of colour holography significantly and was extremely popular within the display, advertising and creative arts industries during the 1970–90s. This two-step process used a laser transmission hologram of a three-dimensional object, which was then 'copied', or rerecorded, onto a second holographic plate through a restricted optical 'slit', with the resulting

hologram being displayed using inexpensive white light. The optical and practical 'trade-off' here was that the display hologram produced a three-dimensional, full-parallax, image when an observer moved from left to right (or right to left) in front of the display. This allowed the observer to look 'round' objects in the hologram offering shifting points of view, as you would expect when looking at a scene through a window. You can, for example, move your point of view and look behind a three-dimensional object in the foreground so that you can see a hidden, three-dimensional object in the background, much like we do in real life. The reason why this type of hologram is of interest in connection with colour holography is that the optical production and reconstruction of this kind of hologram produces a rainbow-colour effect, which 'covers' the displayed image. These are often referred to as rainbow holograms or white-light transmission holograms.

Although parallax exists when moving left to right – when an observer moves up and down, the parallax view remains the same (so the images 'freezes') but also changes colour from red, through green, to blue. Its apparent size can also appear to shrink or expand. Although this limit in parallax viewing and shift in colour is a side effect of the production process, it presented commercial and visual advantages. The technique allowed printed mass production of holograms and prompted a surge in their use on credit cards, publications, textiles, manufacturing plastics, novelty gifts and commercial products.

Pseudo Colour and the Manipulative Artist

Benton's technique is often described as pseudo colour holography because the colours on show have no direct connection to the actual colour of the three-dimensional object which was initially recorded. This term can be applied to any hologram which produces a similar effect and covers both white-light reflection and transmission techniques. A US patent was granted for pseudo colour holography in 1991.

Pseudo colour holographic display was incorporated into a number of different systems, including the separation of two-dimensional graphic images, which could be combined (overlapped) to produce full-colour images. This had considerable impact on mass-produced commercial printing and novelty gift production (McGrew, 1982).

Although this 'inaccurate' or sometimes 'unfixed' colouration within pseudo colour holograms is viewed as a disadvantage to precise reproduction of objects, it is acknowledged by artists and designers as a flexible and creative, spatial, colour technique. Benton understood this and supported many artists in the use of his technique, offering master classes as well as accessible conference publications. A number of key artists adopted this technique (or a variety of it) and tested it to its visual and technical limits, allowing them to 'manufacture' or 'capture' light, change its colour and use it as a sculptural material in space. Artist Rudie Berkhout (1946–2008) is acknowledged as a pioneer in this field, and his colour holograms are in significant public and private collections worldwide. One aspect of the use and development of holographic colour by artists is that they are unbridled from the specific requirements of industry, results-driven, exploration. Where most optical scientists work on a particular problem which requires a quantifiable solution, often set externally by a funding agency or commercial consultation, artists set their own boundaries to develop skills and techniques relevant to their practice. Berkhout used innovative and practical solutions for recording and displaying colours in his work, which might be considered an unlikely requirement, or aberration, in an optics research programme. He was able to create and manipulate three-dimensional objects and images, which were not recorded from a corresponding (recognisable) object, and as his career progressed, this exploration became more abstract and confident.

Berkhout's work has a visual vocabulary developed over many years, and several other artists have used the technical opportunity to display 'pieces' of light and colour within space. A line, shape, mark or word can be presented, unsupported in space, using holographic colour recording techniques. Early works by artist Eduardo Kac displayed 'floating' and overlapped letters and words which made up his series of Holopoems, produced between 1983 and 1993. Unlike Berkhout, they are grounded through the use of text, which can be viewed from multiple angles, as an observer moves around the hologram, offering spatial fluidity and semantic interpolation. There are a large number of pioneering artists who produced exceptional works with colour and their contributions to the visual arts can be found in numerous exhibitions, catalogues, public and private collections.

Another pioneer in the field, artist John Kaufman, used the often unwanted side effect of reflection holography, namely the swelling of the recording emulsion, which results in a colour shift when the hologram is processed and displayed.

The majority of holograms are recorded onto a high-resolution photographic emulsion, which is then wet-processed in chemicals (similar to those used in the development of photographs). Submersion in these chemicals, and wash baths, causes the exposed photographic emulsion to swell. Once dry, the emulsion returns to its original volume/thickness (or very close to it). In white-light reflection holograms, this is an essential element in their production and processing. During display with white light, the processed emulsion absorbs unwanted wavelengths of light and then reflects the wavelength (colour) of light used in the recording. If a red laser is used to record the hologram, the thickness of the processed emulsion should cause only red light to be reflected, and the resulting holographic image would appear red. Occasionally, the processing of the hologram, or inconsistent drying of the emulsion after wet processing, can cause the emulsion to dry to a different thickness. This will change the wavelength of light absorbed and reflected during reconstruction of the hologram and subsequently change its apparent colour.

Artists and display companies have used this to their advantage, and there was a period in the 1980s when reflection holograms for commercial and archival display used emulsion which was intentionally swelled so that the resulting image would appear yellow gold, rather than the original colour of the laser used to record it.

Kaufman developed a technique using chemicals to pre-swell the emulsion on the holographic plate, record a three-dimensional object and, before processing, re-swell the emulsion to a different thickness, record another object, and often repeat

this several times until finally chemically processing the hologram. What resulted was an emulsion which dried to slightly different thicknesses and, when illuminated with white light, would absorb and reflect a variety of wavelengths in the various regions of the holographic plate. The viewer of these works would see fully three-dimensional, full-parallax, images in different colours within the recorded holographic space. This is a controlled system with long production times and a very high degree of uncertainty, as the resulting image is only visible once all the swelling, recording and processing is complete. It is not possible to return to a particular stage of the process and change it – the entire hologram would need to be remade. This limited the use of this process commercially, but enhanced its use in limited, or unique, editions by artists.

Other artists have used a less prescriptive method of emulsion swelling, which subsequently produces a more abstract spatial result. This includes swelling small parts of the emulsion across the surface of the holographic plate without predefining the result. Some artists have used their body surface as a tool for application of swelling agents onto the surface, others incorporate brushes, fine drawing devices or specially cut templates to define and contain the area of swelling. The process functions much like mark-making or the application of paint on a surface but, in this case, the area being manipulated with colour is spatial and volumetric – drawing and colouring in space.

Colour generation and ‘shifting’ is not limited to emulsion swelling. Several artists exploited the rainbow effect within white-light transmission holography, particularly the production and manipulation of holographic optical elements (HOEs). Already a successful process for generating optical elements with extremely specific characteristics for photonic research and development, they are also used by artists as a method of capturing, structuring and manipulating white light to explore the liminal spaces these constructions can produce. Fred Unterseher, one of the pioneers in this technique, and a significant artist/teacher of accessible holography, commented that; “The virtue of a holography class and the holography experience is to transform the way you see, to transform the way you experience; the value of this is to see light and to experience (it) in new ways...” (Johnston, 2006).

As the techniques and technology of holography have developed, there are now a number of commercial companies and research facilities which offer accurate full-colour analogue and digital holographic recording. This has clear commercial impact in military reconnaissance, advertising, architecture, visualisation and museum archiving. Subsequent adoption in this area should stimulate the industry and its competition with illusionistic computer graphics or virtual reality. The undeniable advantage of colour holography continues to be its ability to record and display extremely high-resolution objects in full parallax three dimensions without the need for a viewing device. Where this, perhaps, differentiates itself from developing display technologies is its perceived state of hyper-reality, evidenced by the recent significant advances in ultra-realistic holographic imaging techniques (Bjelkhagen and Brotherton-Ratcliffe, 2016) The ability to specify colours in space can allow significant applications in spatial ‘signalling’ or object differentiation, while artists have a physical and conceptual opportunity to explore and manipulate colour in ‘real’ three dimensions.

What began with Gabor in 1947 as an attempt to improve the resolution of the electron microscope has partly developed into a new technical and creative vocabulary for colouring our spatial environment.

See also: Overview: Holography

References

- Benton, S.A., 1969. Hologram reconstructions with extended incoherent sources. *Journal of the Optical Society of America* 59, 1545–1546.
- Benton, S.A., 1977. White-light transmission/reflection holographic imaging. In: *Applications of Holography and Optical Data Processing: Proceedings of the International Conference, August 23–26, 1976, Jerusalem, Israel*, pp. 401–409.
- Bjelkhagen, H., Brotherton-Ratcliffe, D., 2016. *Ultra-Realistic Imaging: Advanced Techniques in Analogue and Digital Colour Holography*. Boca Raton, FL: CRC Press.
- Bjelkhagen, H.I., Vukicevic, D., 2002. Color holography: A new technique for reproduction of paintings. In: *Practical Holography XVI and Holographic Materials VIII: Proceedings. SPIE*, 4659. doi: 10.1117/12.469251.
- Bjelkhagen, H.I., 2006. Color holography: Its history, state-of-the-art, and future. In: *Holography 2005: International Conference on Holography, Optical Recording, and Processing of Information: Proceedings. SPIE*, 6252. doi: 10.1117/12.677173, ISBN: 9780819463111.
- Denisyuk, Y.N., 1962. On the reflection of optical properties of an object in a wave field of light scattered by it. *USSR Academy of Sciences (Doklady Akademii Nauk)* 7 (144), 1275–1278.
- Gabor, D., 1948. A new microscopic principle. *Nature* 161 (4098), 777–778. doi:10.1038/161777a0.
- Hariharan, P., 1983. IV Colour holography. *Progress in Optics*. 263–324. doi:10.1016/s0079-6638(08)70279-8.
- Johnston, S.F., 2006. *Holographic Visions: A History of New Science*. Oxford: Oxford University Press, [ISBN-13: 978-0-10-857122-3].
- Johnston, S.F., 2009. The parallax view: the military origins of holography. In: Rieger, S., Schröter, J. (Eds.), *Das holographische Wissen*, Dortmund, Diaphane, pp. 33–57; ISBN 978-3-03734-071-4.
- Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. *J. Opt. Soc. Am.* 52 (10), 1123–1130.
- Leith, E.N., Upatnieks, J., 1964. Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54 (11), 1295. doi:10.1364/josa.54.001295.
- Leith, E.N., Upatnieks, J., 1965. Photography by laser. *Scientific American* 212 (6), 24–35. doi:10.1038/scientificamerican0665-24.
- Lin, L.H., Pennington, K.S., Stroke, G.W., Labeyrie, A.E., 1966. Multicolor holographic image reconstruction with white-light illumination. *Bell System Technical Journal* 45 (4), 659–661. doi:10.1002/j.1538-7305.1966.tb01050.x.
- McGrew, S., 1982. A graphical method for calculating pseudocolor hologram recording geometries. In: *Proceedings of the First International Symposium on Display Holography*, Lake Forest, USA (Illinois), Lake Forest College.
- Mirlis, E., Turner, M.J., Bjelkhagen, H.I., 2005. Selection of optimum wavelengths for holography recording. *Practical Holography XIX: Materials and Applications: Proceedings. SPIE* 5742, 113–118. doi:10.1117/12.583224.
- Pepper, A., 2000. Windows with memories: Creative holography in the real world. *This Side Up* 10, 15–18. (Summer).

High-Resolution Underwater Holographic Imaging

John Watson, University of Aberdeen, Aberdeen, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Holography is well-known to most of us: many will have seen white-light display holograms in exhibitions, shops and on credit cards. The word “hologram” has passed into general usage and has come to be associated, somewhat erroneously, with any three-dimensional or pseudo-3D image regardless of how it was formed. The impact and influence of true holography, however, spreads far and wide beyond display and entertainment confines and into the whole spectrum of science, engineering, commerce and everyday life. Holography is now a standard requirement in vibration analysis, image processing, holographic-optical elements, and optical computers. The potential for high-precision imaging and accurate dimensional measurement makes holography an ideal tool for in situ imaging of plankton and other microscopic-sized aquatic organisms and particles. It is this unique ability of holography to recreate, from a flat optical sensor, a three-dimensional image that is optically indistinguishable from the original that sets it apart from other imaging techniques and makes it so useful in underwater aquatic science and engineering for imaging, identification, and measurement of aquatic organisms and particles.

A detailed knowledge of the distribution and dynamics of marine and freshwater organisms and particles, such as plankton, is crucial to our understanding of the planet’s biodiversity and how these organisms affect our environment. Traditional methods of gathering in situ data on aquatic particles, such as collection using nets (which can destroy fragile organisms), electronic counting or photography, are not usually suited to observing precise spatial relationships. The overriding benefit of holography is that it permits non-invasive and non-destructive observation and analysis of organisms in their natural environment, whilst preserving their relative spatial distribution.

Initial applications of holography in underwater measurement were based on recording the holograms on photographic film or plates – so called “classical analogue holography.” This method requires wet-chemical processing of the recorded hologram together with its subsequent replay in the laboratory, on dedicated (and expensive) replay systems, to create a 3D image of the original scene. Furthermore, taking classical holography out of the laboratory and into a field environment, requires the use of short-duration pulsed lasers to reduce the effects of vibration or movement in any of the optical components or in the objects being recorded. However, this does have the advantage of freezing the scene at the recording instant, thereby allowing fast moving particles to be imaged.

With the dramatic improvements in electronic sensor technology and computer performance, digital holography (DH) – sometimes known as digital holographic microscopy (DHM) – has now begun to dominate the field of aquatic holography. In contrast to classical methods, digital holography records the hologram on an electronic sensor array, such as a charge-coupled device (CCD) or complementary metal oxide semiconductor (CMOS), and the resulting interference pattern is stored in fast computer memory. Reconstruction of the digital holographic image is carried out, in a computer, by numerical simulation of the beam propagation through the hologram and the resultant image at a particular plane in space with reference to the hologram, viewed on a monitor.

Regardless of whether or not analogue or digital recording is employed, a hologram captures vastly more data than a photograph; the reconstructed scene retains the parallax of the original and is free from perspective distortion. Image resolution is high (a few micrometres under favourable conditions) over a wide depth-of-field and field-of-view. Sequential holograms can record changes within a recording volume over a defined period of time and the wide recording dynamic range allows images to be captured that would otherwise be lost in noise. Additionally, in digital holography, “holovideos” can be produced that capture the time dimension as well as spatial dimensions.

For a general background to classical analogue holography, *Optical Holography* (Hariharan, 1996) provides a good introduction; and for the basic principles of digital holography, *Digital Holography and Wavefront Sensing* (Schnars *et al.*, 2015) and *Digital Holography and Three-dimensional Television* (Poon, 2006) provide comprehensive overviews of the techniques.

Analogue Holographic Recording and Replay

Although digital holographic recording is now the dominant method of underwater holography, it makes sense to start with an outline of classical analogue holography, since this helps to explain the concepts and bring out the crucial advantages of holography for aquatic studies. Of the possible recording methods of holography (whether classical or digital) two in particular, the “in-line reference beam hologram” (ILH) and the “off-axis reference beam hologram” (OAH), find use for high-resolution in situ imaging underwater.

In an ILH, a single laser beam (Fig. 1) is directed through the sample volume towards the sensor (photographic or electronic). The optical interference which occurs between light diffracted by the object and the undeviated portion of the illuminating beam is recorded on the sensor. A laser is needed in recording to provide the optical coherence properties necessary to produce the optical interference pattern.

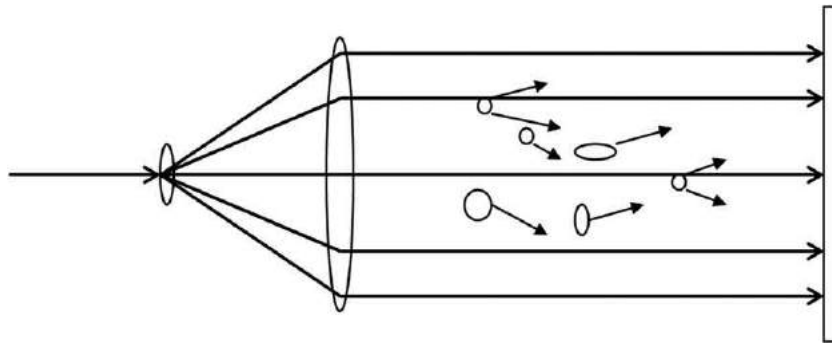


Fig. 1 Recording of an in-line hologram of suspended particles.

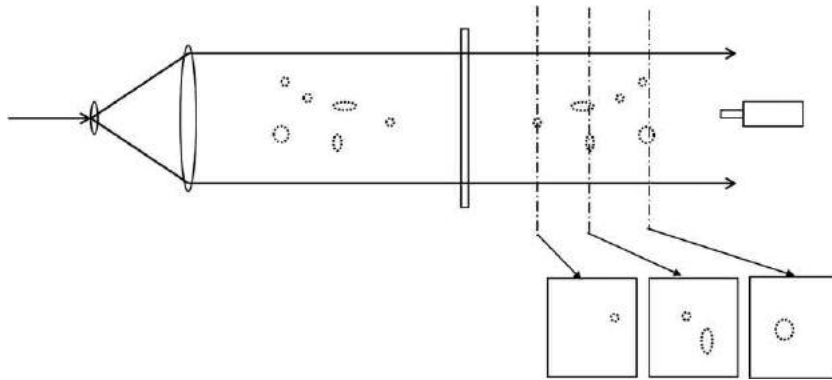


Fig. 2 Replay of an in-line hologram.

The method of reconstructing or replaying the hologram to create a 3D image is where analogue and digital holography differ. In the analogue form, the holographic film must be placed in a replica of the original reference beam (in terms of wavelength, wavefront curvature, phase and direction) and the resultant image viewed with some auxiliary optical system such as a video camera (**Fig. 2**). The replayed hologram simultaneously forms two images, one “virtual,” the other “real,” which are located on the optic axis on opposite sides of the holographic plate. These two images are co-axial, and located (for the typical case of collimated recording and replay beams) at equal distances in front of and behind the hologram plane. The virtual image appears to be located behind the hologram and seen through it (like viewing through a window – although this should not be done since the viewer will also see the straight-through laser beam). The real image is formed in space in front of the hologram, between the hologram and the viewer. Although a true 3D real image is produced, because of the dimensions of the sensor it has almost no parallax, and cannot easily be seen by the unaided eye. This image has inverted depth perspective and is seen in darkfield against the out-of-focus virtual image. Traversing a camera through the projected real image allows optical “sectioning” of the image at any plane in the image space.

In OAH, a two beam recording geometry is utilized, although both must be generated from the same laser: one beam illuminates the scene, and the other directly illuminates the holographic film at an oblique incidence angle (**Fig. 3**). Optical interference occurs between the diffuse light reflected from the scene and the angularly separated reference beam. On replay, the real and virtual images are angularly separated which makes their interrogation easier. OAH is primarily applied to opaque subjects of large volume. To view the “virtual image” in OAH, the processed hologram is replaced in the position in which it was originally recorded and illuminated with an exact duplicate of the original reference wave, in terms of its wavelength, curvature, and beam angle. In this mode, a virtual image is observed as if viewing the scene through a window. Other than colour, this image is almost optically indistinguishable from the original scene.

Once again for data extraction and measurement, the projected “real” image (**Fig. 4**) is most useful. If we illuminate an off-axis hologram from behind, with a wave which is the exact phase conjugate of the original (i.e., one which possesses the same wavelength as the original but with the opposite direction and curvature), we generate an image that appears to float in the real space in front of the observer. Often a collimated reference beam is used in recording, since reconstruction is then a simple case of turning the hologram around. No lens is needed to form this image, and it is optically identical to the original wave save that it is reversed left-to-right and back-to-front (pseudoscopic). Planar sections of the real image can be optically interrogated by directly projecting onto a video camera (often with the lens removed) which can be moved throughout the 3D volume.

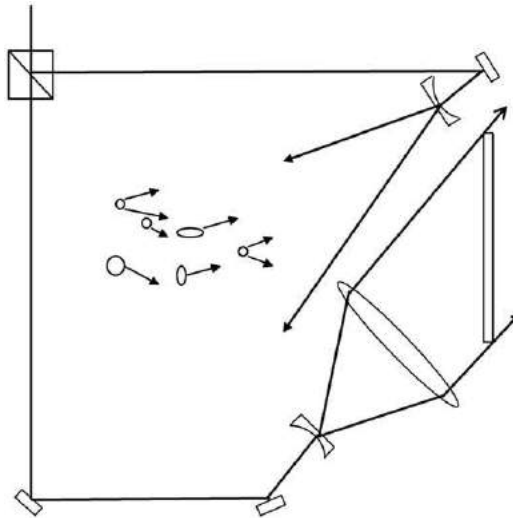


Fig. 3 Recording of an off-axis hologram.

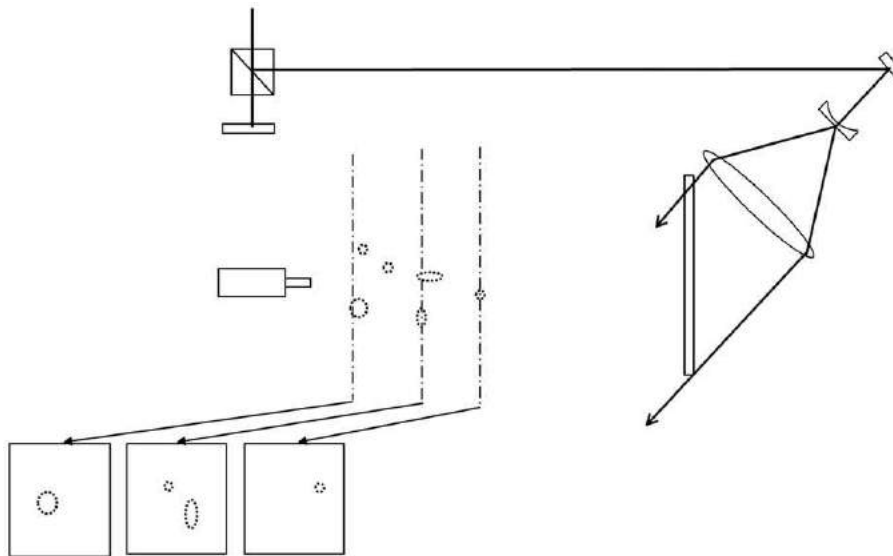


Fig. 4 Real image reconstruction from off-axis hologram.

Digital Holographic Recording and Replay

The concepts of digital holography are broadly similar to those of analogue: the main differences being the recording medium and the method of reconstruction of the holographic image. In digital holography, the interference field is stored directly into computer memory and the hologram is replayed by numerical simulation of the propagation of the optical field through space. In this way planar sections of the image can be recreated at any distance from the hologram plane, at any time, and displayed on a computer monitor, allowing size, location, identification and distribution of particles to be extracted. Computer-based algorithms are used to pre- and post-process the holograms, which are replayed and displayed on a TV monitor at any plane in the recording space. Importantly, the phase as well as the intensity of the wave field is retained on reconstruction. Apart from speed, convenience, and freedom from wet chemical processing, a distinct advantage of electronic recording is the ability to record holographic videos. This allows not only 3D spatial interrogation but allows visualisation of the movement of organisms within a particular scene and tracking the path of individuals.

For digital holographic particle sizing, the ILH mode of recording is generally the favored option by virtue of its geometric simplicity, and consequently its lower cost and potentially higher resolution. Because of the much higher sensitivity of electronic sensors compared with holographic photosensitive film, low power continuous-wave (c.w.) lasers can be used for illumination. In field applications though, pulsed lasers are often a better choice, particularly if the subject is fast-moving or subject to vibration,

or the holocamera cannot be held steady with respect to the target. To record large, opaque or dense aggregates of particles, OAH allows the use of front or side illumination of the subject in conjunction with a separate side or off-axis (angular) reference beam.

One major limitation, though, of digital recording is the lower resolution of electronic sensors compared with that of photographic film. The sensor must be able to resolve the finely spaced interference pattern recorded in the hologram; thus the grain or pixel size has a great bearing on imaging capabilities. The maximum spatial frequency of the resolvable interference pattern depends on the angle between the reference beam and the object beam. The typical grain size of 30 nm or so for holographic film is about two-orders of magnitude greater than that of a typical electronic sensor of about 3 μm pixels. Accordingly the achievable fringe frequency is much greater and reference beam angles up to 45° or so are possible with film, whereas the electronic sensor can only support beam angles up to about 10° . It is this requirement which so far makes in-line recording the dominant method of DH, since the reference beam angle is essentially zero.

Generally the computer algorithms developed for digital recording and numerical replay of a hologram are based upon the Fresnel–Kirchhoff formulation of a complex monochromatic wavefield propagating from a scene through the aperture of an optical sensor. The algorithms employed achieve computational efficiency by applying appropriate approximations and simplifications to make the Fresnel–Kirchhoff equation more suitable for digitisation and computer implementation. In most cases, the integral is implemented as a Fourier transform, thereby allowing the use of existing Fast Fourier Transform (FFT) libraries that can be easily incorporated into software. It is possible to recreate the intensity and phase distribution of the wavefield at any plane in the reconstructed volume of the hologram; thereby simulating the effect of traversing an image sensor through an optically replayed hologram (but with the added advantage that phase information can also be extracted).

Subsea Holography and Holocameras

In general, the usefulness of holography for accurate inspection and measurement is dependent on its ability to reproduce an image of the object which is low in optical aberrations and high in image resolving power. Regardless of whether or not the hologram is recorded underwater, high resolution, high contrast, and low noise are the dominant requirements. All the primary monochromatic aberrations (spherical aberration, astigmatism, coma, distortion, and field curvature) of any optical system may be present in the holographic image. For precise and accurate reconstruction of the real image, the reconstruction reference beam should be an exact copy of the original (where possible), but opposite in curvature (the phase conjugate) of that used in recording. Under these conditions, the lateral, longitudinal, and angular magnifications of the real image with all equal unity and aberrations can be reduced to a minimum.

When holography is applied subsea, the hologram is recorded in water but the image is replayed, in the laboratory (if classical) in air, or by computer (if digital). Because of the refractive index mismatch between recording and replay spaces, optical aberrations in the reconstructed image will be exaggerated, and can impair the potential for high-precision measurement. In ILH, only spherical aberration is significant since both reference and object beam angles are normal to the recording plane. However, in OAH astigmatism and coma dominate; they increase the greater the field angle of the subject in the reconstructed image. These limit resolution and introduce uncertainty in co-ordinate location. These additional aberrations are entirely due to the refractive index mismatch between object and image spaces and are unconnected with the holographic process itself. Furthermore, the water itself may be expected to influence the quality of the images produced. An increase in the overall turbidity of the water will adversely affect both ILH and OAH recording and will create background noise that will reduce image fidelity. In most underwater applications the objects (or the camera) are in motion during the exposure. The effect of this motion is to blur out the finer fringes, and thus reduce resolution and contrast.

A recurring issue in all forms of holographic recording (digital or classical) of microscopic particles, and one which imposes limits on the wider use of the technique, is extracting, processing and analyzing the vast amount of data contained in a hologram. Although analysis of classical holograms, can be performed by manual scanning of the optical reconstructions, this is tedious and time-consuming and requires high levels of concentration; automatic interrogation of the data in a dedicated reconstruction facility is essential for all practical applications of the technique. Since the main thrust of this article is on digital holography the following discussion is restricted to subsea digital methods.

Since a single holovideo may contain as much as several gigabytes of data, a first step in interpretation of digital holograms and holovideos of particles often involves localizing every particle within a frame (in x , y , z , t), and distilling particle shape and positional information from it. This facilitates application-specific processing, with subsequent image recognition, particle tracking, counting and sizing. For each video frame, the hologram is reconstructed in incremental steps and an image of each slice parallel to the sensor plane is sequentially obtained. This is equivalent to discretized simulation of the wave-field projected into real image space by an analogue hologram when reconstructed by an optical beam. When a particle coincides with a reconstruction plane it will appear in-focus, and is characterized by a maximization of image gradients at the particle edges.

Focus detection is typically the first step in hologram analysis and with suitable software, particle identification can be carried out on focused particle images and 3D plots of relative position and distribution produced. Most of these algorithms implement optimal-focus metrics which depend variously on image intensity gradient, variance, correlation, histograms and frequency-domain analysis. These metrics all rely on the premise that focused images have higher information content than blurred (out-of-focus) images due to the existence of larger gradients with higher variance across them. This leads to a greater deviation between maxima and minima in the brightness histogram and the local maximization of power in higher frequency components when the image is transformed to the frequency-domain. Due to “speckle effects” in the hologram an algorithm with good noise immunity

is required. One such algorithm relies on estimating the contour gradient around a particle and has been used to analyze a number of holographic videos recorded in the North Sea.

The first-known use of holography for recording living marine plankton was by Knox in 1966, who recorded classical in-line holograms, using a low coherence ruby laser, of a variety of living marine plankton species in a tank. Following this pioneering work, a host of subsea holographic systems, based on classical holography were developed and deployed up until the late 1990s and early 2000s (e.g., [Carder, 1979](#); [Katz 1999](#)). Although these holocameras successfully demonstrated the power and ability of the technique, they were big, bulky, heavy and not easy to deploy from research vessels or remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs).

With the development of high-resolution electronic sensors such as CCD or CMOS together with increased power and speed of modern PCs and high storage methods, digital holography took over from the early 2000s. Now holography was freed from the constraints of laboratory replay, wet-chemical processing and could be replayed in near real-time. Other advantages included recording with low-power diode lasers (under appropriate conditions) and the introduction of the time dimension allowing holographic videos to be captured.

Digital Subsea Holocameras

The first digital holocamera for underwater applications was developed by [Owens and Zozulya \(2000\)](#). It utilized a 10 mW continuous wave (c.w.) diode laser in an in-line configuration onto a CCD array over a maximum in-water path length of 25 cm.

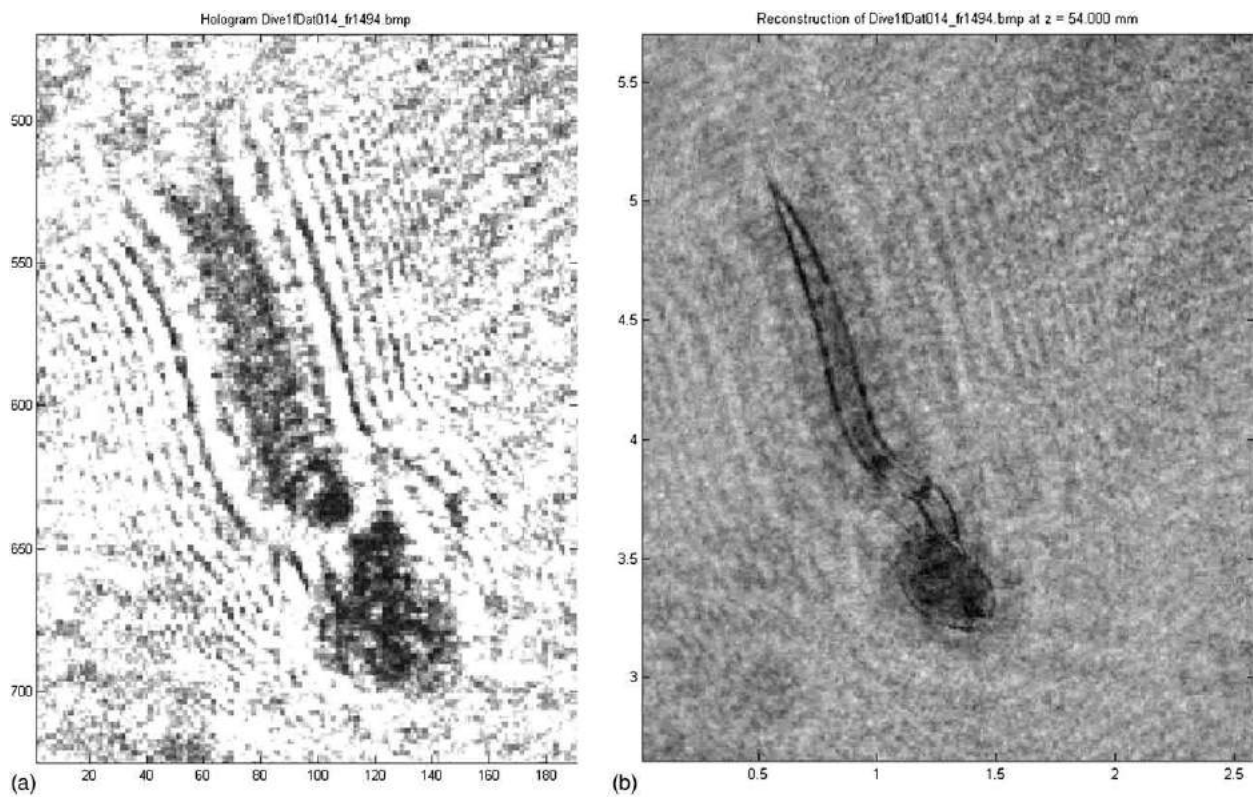


Fig. 5 Appendicularia (a) unreconstructed hologram, (b) reconstructed image (scale in mm).



Fig. 6 Internal layout of *eHolocam* showing the laser, power supply and computer on the right and the smaller CMOS housing on the left.

The use of a c.w. laser can be justified on the grounds that the enhanced sensitivity of electronic sensors over photofilm reduces exposure times to about 100 μ s. However, this still limits application to slowly moving systems. This holocamera was successfully deployed to depths of about 50 m in Tampa Bay, FL.

Since then, several groups world-wide have exploited DH for underwater environmental science (e.g., [Dyomin et al., 2003](#); [Jericho et al., 2006](#); [Nimmo Smith, 2008](#)); and submersible digital holocameras are now becoming commercially available. Although these systems share common elements such as use of the in-line geometry and, usually, charge-coupled device (CCD) arrays, there are subtle differences amongst them relating to their application and deployment method. Many were developed primarily for studies of phytoplankton over sampling volumes of around 1 mm³; their use of continuous lasers limits them to low particle velocities.

One such system is shown in [Fig. 5](#); and known as “*eHoloCam*” ([Sun et al., 2007](#)). It embodied a different approach from many of the others. Although still adopting in-line holography, *eHoloCam* utilized a collimated (parallel) beam from a pulsed frequency-doubled Nd-YAG. For subsea purposes, a laser with a wavelength in the blue-green region of the spectrum will match the peak transmission window of seawater. A frequency-doubled Nd-YAG laser with a wavelength of 532 nm is a good choice, although ruby lasers operating at 694 nm are also acceptable if transmission distance is not an issue. *eHoloCam* was designed with larger planktonic species in mind and featured a large recording volume of 36.5 cm³ over a 453 mm path length to allow the sparser zooplankton species to be captured rather than the more abundant, and smaller, phytoplankton.

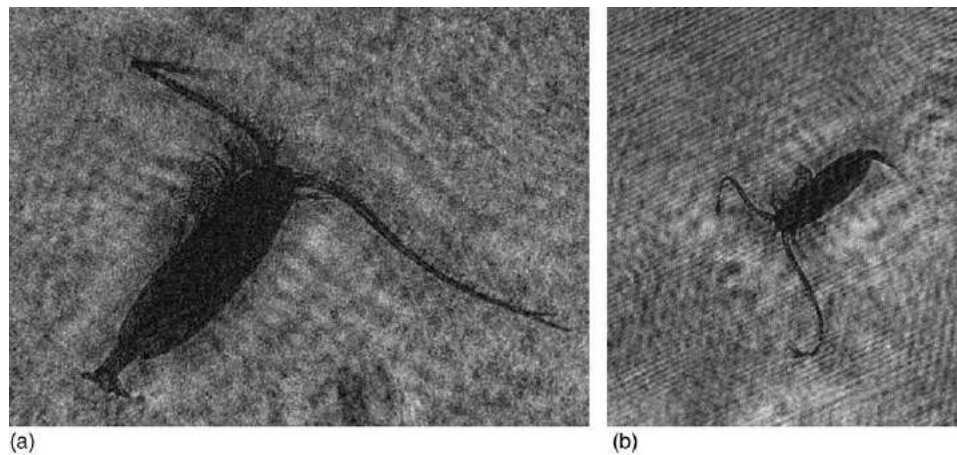


Fig. 7 Reconstruction of a calenoid copepod.

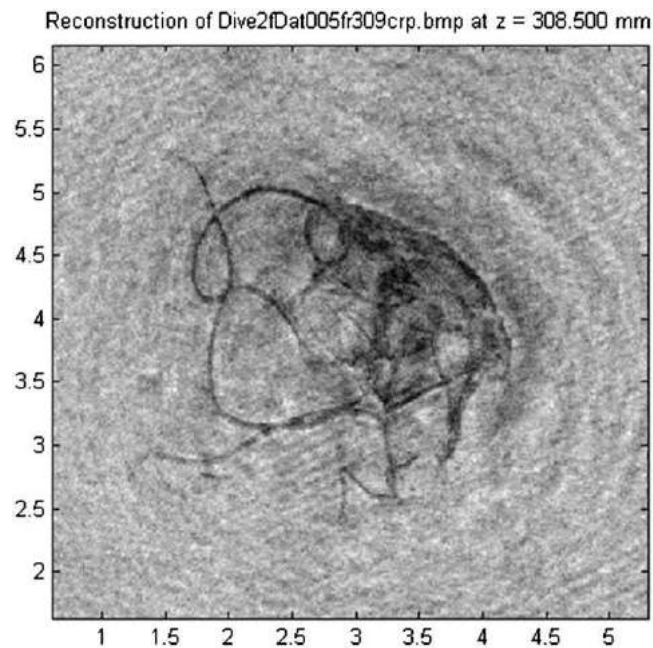


Fig. 8 Reconstruction of a jellyfish larvae (scale is mm).

eHoloCam consists of two water-tight housings: the larger housing contains the laser, a single board computer, storage hard drives and beam forming optics (Fig. 6). The second, smaller, housing contains a CMOS sensor of 2208×3000 , $3.5 \mu\text{m}$ square pixels on a $10.50 \text{ mm} \times 7.73 \text{ mm}$ area. Both housings were designed to an operational pressure of 30 MPa (3 km depth). Sapphire windows allowed beam transmission through the water from laser to sensor. Holograms were recorded at video frame rates up to 25 Hz.

eHoloCam was deployed in the North Sea from the RV Scotia (Marine Scotland Science, Aberdeen, Scotland) on four cruises over 1 year (December 2005 to December 2006). It was operated from a sampler frame Auto-Recording Instrumented Environmental Sampler (ARIES – Marine Scotland Science, Aberdeen, Scotland) and the short pulse duration (4 ns) allowed it to be towed at up to 4 knots (about 2 m s^{-1}) to depths of 450 m. Water flows through the “jaws” of *eHoloCam* laterally and perpendicular to the beam path. The system is self-contained in power and data storage and executes embedded control software. The use of ILH and pulsed laser allowed hologram recording at energies of less than a millijoule.

Hologram frames have been reconstructed using the angular spectrum algorithm and manually extracted from the recorded holovideos. Fig. 6 shows an Appendicularia firstly as an unreconstructed hologram and then shows the reconstructed, focused image. Fig. 7 shows a calenoid copepod and Fig. 8 shows a jellyfish larvae.

Conclusions

Since its inception in the early 1960s, holography has shown itself to be particularly useful for studies of microscopic species in the aquatic environment. Its non-destructive and non-invasive nature lend itself particularly for this environment. We have seen that holography in both its classical analogue form and its digital form provides significant advances and features not present in conventional methods of imaging. These features such as 3D, optical sectioning, freedom from parallax and perspective effects, high depth-of-field all offer important benefits to aquatic biologists in their drive to understand the complexities of our aquatic environment. In particular, digital holography also offers the ability to record holographic videos which allow the preservation of the time dimension as well as particle location. Now with the improvements in electronic sensors and computer technology, in-line digital holography is well on its way to becoming a standard tool for studies of microscopic organisms in the sea. Specifically holography aids biologists in their studies of the behavior of organic species such as plankton and other marine organisms and aquatic particles in the oceans of the world. With further development holography is set to become one of the most valuable tools available for aquatic biologists.

See also: Overview: Holography

References

- Carder, K.L., 1979. Holographic microvelocimeter for use in studying ocean particle dynamics. *Optical Engineering* 18, 524–525.
- Dyomin, V.V., Polovtsev, I.G., Makarov, A.V., *et al.*, 2003. Submersible holocamera for microparticle investigation: Problems and solutions. *Atmos Oceanic Optics* 16, 778–785.
- Hariharan, P., 1996. *Optical Holography*, second ed. Cambridge: Cambridge University Press.
- Jericho, S.K., Garcia-Suserquia, J., Xu, W., Jericho, M.H., Kreuzer, H.J., 2006. Submersible digital in-line holographic microscope. *Review of Scientific Instruments* 77, 043706.
- Katz, J., 1999. Submersible holocamera for detection of particle characteristics and motions in the ocean. *Deep-Sea Research* 46, 1455–1481.
- Nimmo Smith, W.A.M., 2008. A submersible three-dimensional particle tracking velocimetry system for flow visualization in the coastal ocean. *Limnology & Oceanography: Methods* 6, 96–104.
- Owen, R.B., Zozulya, A.A., 2000. In-line digital holographic sensor for monitoring and characterizing marine particulates. *Optical Engineering* 39, 2187–2197.
- Poon, T.C., 2006. *Digital Holography and Three-Dimensional Television*. New York, NY: Springer.
- Schnars, U., Falldorf, C., Watson, J., Jüptner, W., 2015. *Digital Holography and Wavefront Sensing*. Heidelberg: Springer.
- Sun, H., Hendry, D., Player, M., Watson, J., 2007. In situ electronic holographic camera for studies of plankton. *IEEE Journal of Oceanic Engineering* 32, 373–382.

Further Reading

- Burns, N.M., Watson, J., 2011. A study of focus metrics and their application to automated focusing of inline transmission holograms. *Imaging Science Journal* 59 (2), 90–99.
- Burns, N.M., Watson, J., 2014. Robust particle outline extraction and its application to digital in-line holograms of marine organisms. *Optical Engineering* 53 (11), 8. doi:10.1117/1.OE.53.11.112212.
- Gopalan, B., Katz, J., 2015. Turbulent shearing of crude oil mixed with dispersants generates long microthreads and microdroplets. *Physical Review Letters* 104 (054501), 1–4.
- Knox, C., 1966. Holographic microscopy as a technique for recording dynamic microscopic subjects. *Science* 153, 989–990.
- Watson, J., 2013. Subsea imaging and vision: An introduction. In: Watson, J., Zielinski, O. (Eds.), *Subsea Optics and Vision*. Cambridge: Elsevier, pp. 17–34. ISBN: 978-0-85709-341-7.
- Watson, J., Burns, N.M., 2013. Subsea holography and submersible holocameras. In: Watson, J., Zielinski, O. (Eds.), *Subsea Optics and Imaging*. Cambridge: Elsevier, pp. 294–326. ISBN: 978-0-85709-341-7.

Digital Holographic Display

Daping Chu, Jia Jia, and Jhensi Chen, University of Cambridge, Cambridge, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

Overview of a Holographic Display

History of Holography

Holography was invented by Dennis Gabor in 1948 for the correction of electron lens aberration to improve electron microscope resolution (Dennis, 1948). The word of holography is composed of 'holo-' and '-graphy', which stand for 'whole' and 'writing', respectively. The advantage of being able to reconstruct the wavefront in space was appreciated gradually, and the research works on various holographic applications started to develop.

The electron wave based holography was not immediately implemented in optical applications because of the lack of adequate coherent light source at the time. Mercury arc lamps were used commonly as the light source and the image quality of the holographic experiments was poor due to the low coherence of light. The situation changed completely after the laser was invented in 1960, and the first practical optical hologram with recorded 3D objects was realized in 1962 by Yuri Denisyuk in the Soviet Union (Denisyuk, 1962) and Leith and Upatnieks at the University of Michigan (Leith and Upatnieks, 1962).

Although people have been hoping to realize 3D images through holography for decades, as shown frequently in novels and sci-fi movies, there has been almost no actual implementation commercially so far. Instead, holography finds its applications in holographic optical elements (HOE), such as holographic gratings and holographic lenses, and in interferometry as a standard optical measurement technology, known as the non-destructive testing and holographic microscopy. It also found its way back in the aberration correction for optical components, as what Gabor intended to at the beginning. Furthermore, holography is used in anti-counterfeiting and authentication, such as the holograms laminated on credit cards and identification documents. It also has the potential to be used in optical data storage and optical computing. Nevertheless, this review will focus on the development and issues related to holographic displays, in particular digital holographic displays.

Principle of Holography

A light wave, or an electromagnetic (EM) wave more generally, can be described by using its amplitude and phase. Consider a certain area, such as a square or a circle, in a space which the light flux propagates through. The basic concept of holography is to record the complex light field (both amplitude and phase) in this area and then reconstruct it for various applications. This principle can also be extended to non-EM situations, such as electron waves used in the electron holography. For display applications, the light waves to be considered are in the visible range with a wavelength between 400 and 700 nm.

Recording a complex light field was done traditionally by using photosensitive materials, which response to light intensity rather than phase. Using a single coherent light source, the object light scatters from an object and interferes with the reference light in space. The interference pattern carries the information of both the amplitude and phase of the wavefront of the object light. Such a pattern is called a hologram, or to be precise an intensity hologram when recorded on the photosensitive material. To reconstruct the recorded wavefront of the object, a reference light is used to illuminate the intensity hologram from the same angle as it was recorded.

Fig. 1 illustrates the conventional way of recording a hologram and reconstructing a holographic image. Detailed mathematical description of these processes can be found in Benton and Bove (2008). In practice, it requires suitable equipment and finds experimental skills to successfully produce high quality holograms and reconstructed images.

Using intensity holograms produced by interference to record and reconstruct wavefront is not the only method of holography. The main reason to use an intensity hologram was historical because of the difficulty in modulating the phase of light directly in the past. As the technology progresses, it is now possible to use an array of mega pixels on a phase-only spatial light modulator (SLM) to form a phase-only hologram and modulate the phase of light. This allows us to manipulate the wavefront of light directly without the need to manipulate its intensity, hence the maximum light efficiency.

Both intensity and phase-only holograms can be generated by using appropriate SLMs. Digital holographic displays use those SLMs which are pixelated.

Digital Holography and Computer Generated Hologram

Holographic interference patterns, i.e., holograms, carry the information of light propagation, and can be sampled (recorded) discretely. This discrete information can be kept and reconstructed digitally, and this technique is called digital holography and the corresponding holograms and displays called digital holograms and digital holographic displays, respectively. Computer generated hologram (CGH) is a part of digital holography, and it specifically means those digital holograms which are generated by calculations on computers. Calculation of CGHs involves light propagation calculations and complex information approximation (further information see Section Computer Generated Holograms). Note that the wave information which a CGH carries can be

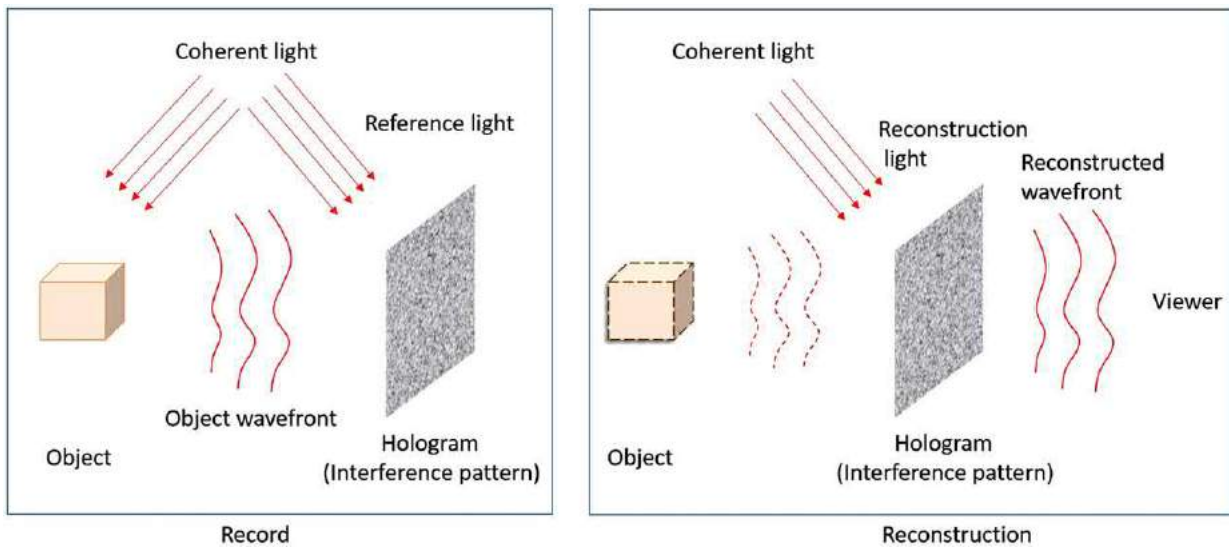


Fig. 1 Illustration of the recording and reconstruction of a 3D scene using holography.

any kind of light propagations, including image and beam propagations. The hardware to load CGH and reconstruct images is mostly of a pixelated structure except few exceptions, and this will be discussed in Section Hardware for a Holographic Display.

Hardware for a Holographic Display

Holograms can be recorded on photosensitive materials such as silver halide films from Geola (Zacharovas *et al.*, 2001), holographic photopolymer from Bayfol (Jurbergs *et al.*, 2009) and DuPont (Gambogi *et al.*, 1994), as described in Section Principle of Holography. These photosensitive materials are one-off use in general and not suitable for dynamic display applications (Berneth *et al.*, 2013; Bjelkhagen *et al.*, 2008; Cody *et al.*, 2012; Stevenson, 1997). Here, we will introduce only the materials and devices suitable for dynamic holographic displays, i.e., for holographic video displays.

Passively Dynamic Holographic Display Devices

One straightforward approach to display holographic images at video rate is to reconstruct sequential static images from sequential static holograms, just as the way which movies are shown using static image films. To implement this approach, updatable mediums, on which holograms can be recorded and removed repeatedly, were proposed and researched, such as a photorefractive polymer (PRP) films (Blanche *et al.*, 2010a,b; Kinashi *et al.*, 2012; Tay *et al.*, 2008) and optical addressed spatial light modulators (OASLMs) (Shrestha *et al.*, 2015; Slinger *et al.*, 2004; Stanley *et al.*, 2004) and azobenzene-based devices (Rasmussen *et al.*, 1999; Shishido, 2010). They can be regarded as passive dynamic holographic display devices (PDHDDs). “Passive” here means that the hologram information is provided from another device, and a PDHDD only carries it passively.

PRP films are simple in configuration and can support nanometer size grains, but they need high voltages ($> \text{kVolts}$) for its operation. OASLMs, on the other hand, can be operated at low voltages ($< 10 \text{ Volts}$), but with much larger reported spot sizes. Researches on both devices have improved their performance over the years, leading to better resolution, higher sensitivity, faster response time and larger area with good uniformity. A paper in 2015 shows that it is able to achieve sub-micro resolution for an OASLM (Shrestha *et al.*, 2015). In addition, updatable hologram materials have shown further potential in applications, such as data storages and futuristic optical displays for optical computing systems.

To use updatable holographic films for dynamic holographic images projection, it requires holograms to be written on the films in real time. This is not normally feasible if the diffraction patterns come from a real object directly. However, it is possible to do so if they are provided by active dynamic holographic display devices (ADHDDs), which will be introduced in the next section. Nowadays, almost all reported demonstrations of updatable holographic films are using ADHDDs to write target diffraction patterns on them.

One advantage of using a PDHDD device is that it can provide a large aperture and hologram size for the holographic image reconstruction. Existing ADHDDs all have limited apertures and hologram sizes, and this limits the minimum image spot size it can reconstruct (Hecht, 1998). The use of a PDHDD allows holograms from an ADHDD to be written and tiled up on it, to provide a better image quality in reconstruction. Nevertheless, as spatial light modulators (SLMs) develop over the time, $4\text{k} \times 2\text{k}$ chips are now commercialized (Jasper Display Corp., 2017; Texas Instruments, 2017) and $8\text{k} \times 4\text{k}$ chips are also available (Senoh *et al.*, 2013), and hence the aperture issue is of a less concern.

Active Dynamic Holographic Display Devices

An ADHDD is the kind of SLMs which modulate light directly by themselves. In holographic display applications, two types of ADHDDs are widely used, pixelated type SLMs and acoustic-optic modulators.

A pixelated SLM consists of a large number usually millions of pixels, each of them can be controlled individually to modulate either the amplitude or the phase of light in normally a digital manner. It is possible to control even both of the amplitude and phase simultaneously through cascade approaches (Hsieh *et al.*, 2007; Jesacher *et al.*, 2008a). The pixel pitch has been brought down over from 15–20 μm a few years ago to 1–5 μm now (Jasper Display Corp., 2017; Senoh *et al.*, 2013; Isomae *et al.*, 2016; Oppenheim and Lim, 1981). It is expected that the pixel pitch will become even smaller in future. Holographic images are reconstructed usually with a plane wave illuminating the holograms uploaded on the pixelated area of an SLM. Approaches to generate holograms from computation will be introduced in Section Computer Generated Holograms.

Liquid crystal on silicon (LCOS) devices are probably the most commonly used pixelated SLMs in a holographic display. LCOS combines the functionality of the cutting-edge CMOS integrated circuit technology with a nonlinear electro-optic material which could be switched by low voltages. The device operates in the reflective mode and its construction is similar to that of a liquid crystal cell apart from that a silicon backplane with a reflection layer is used in place of one of the traditional glass substrates. There are various types of LCOS, including grey-level amplitude-only, grey-level phase-only ones as well as binary amplitude-only and phase-only ones using ferroelectric liquid crystals (ferroelectric LCOSs). A cross section sketch of a typical phase-only LCOS device is shown in Fig. 2 (Zhang *et al.*, 2014). In holographic displays, phase-only LCOSs are normally used because of its high light efficiency and good reconstruction image quality (Oppenheim and Lim, 1981). Other applications of phase-only LCOSs can be found in Collings *et al.* (2011). Ferroelectric LCOSs are also often used because they have very high frame rate. Temporal based methods are used in the image reconstruction by Ferroelectric LCOS SLMs, to compensate the low image quality due to the binary nature. The advantage of LCOS SLM technology includes: benefit on the development of silicon CMOS technology, integration of high performance driving circuitry on the silicon chip, high pixel fill factor, high quality process technology for excellent pixel mirror reflectivity, and scalability to smaller feature size and higher pixel number.

Digital micro-mirror devices (DMDs) are also a widely used in holographic displays. Although a DMD can only support binary amplitude modulation, its high frame rate (up to 30 kHz (Texas Instruments, 2017)) makes it attractive since it can support a large information bandwidth for a holographic video display with large image size and viewing angle.

Apart from LCOS and DMD, there are other SLMs developed previously which are less used for holographic displays nowadays, such as liquid crystal transmissive SLMs which have low fill factors. There are also some devices under development but not yet fully realized, such as diffractive nano-devices (Sun *et al.*, 2013) and flexo-electro LCOS (Chen *et al.*, 2009).

In addition to the pixelated digital devices, an acoustic-optic modulator (AOM) has also been applied in holographic display. An AOM is an analog-type SLM without pixelation structure and it is often used in laser Q-switches (Hecht, 1998). It modulates light using the acoustic-optic effect of certain materials, which refractive index changes when an acoustic wave passes through it. The acoustic wave modulated periodic index changes form an effective grating which modulates the light accordingly. Different to a pixelated SLM, which reconstructs the object light for the whole target image simultaneously, an AOM delivers one light beam at a single time. This means that with an AOM the object light is reconstructed beam by beam. In other words, an image is generated pixel by pixel in time sequence. Well-known holographic displays built on AOM(s), the Mark system, will be introduced in Section Selected Holographic Display Systems.

Bandwidth Requirements for Decent Display Size and Viewing Angle

We define the information bandwidth of a device or display system as the product of its spatial bandwidth (number of pixels to provide at a given time) and temporal bandwidth (refresh rate or frame rate).

To generate a genuine holographic 3D image from diffraction, we can calculate how much spatial information or the number of pixels is required to support the given specifications of an image. The viewing angle, $\Delta\theta$, is:

$$\Delta\theta = \sin^{-1}(W/2d) \quad (1)$$

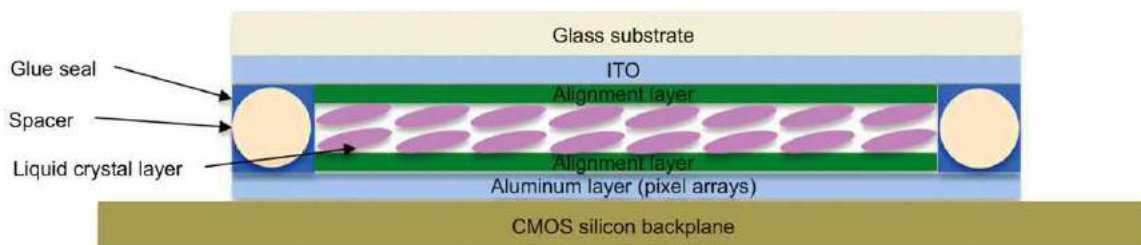


Fig. 2 Cross section of a typical phase-only LCOS device. Reproduced from Zhang, Z., You, Z., Chu, D.P., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. Light Sci. Appl. 3, e213.

and the image size, Δh , is:

$$\Delta h = d \sin \theta_d = d\lambda/p \quad (2)$$

where W is the SLM width in one dimension, d the distance between the SLM and the reconstructed image, θ_d the diffraction angle for the 1st order, λ the wavelength, p the hologram pixel pitch.

In the far field $d \gg W$, $\Delta\theta \sim W/2d$ and the product of Eqs. (1) and (2) gives the number of pixels required, n , as:

$$n \simeq 2\Delta h\Delta\theta/\lambda \quad (3)$$

This result is the same as that obtained from a Fourier holographic display geometry in Stanley *et al.* (2000).

Assuming the targeted viewing angle, θ , is tiled by a number of $\Delta\theta$ and the targeted image size, H , is tiled by several Δh . For a holographic display of two dimensions with the image sizes, H_x and H_y , and the corresponding viewing angles, θ_x and θ_y , respectively, the spatial bandwidth or the overall number of pixels, N_{tot} , needed to reconstruct such an image will be from Eq. (3):

$$N_{\text{tot}} = N_x N_y \simeq 4H_x H_y \theta_x \theta_y / \lambda^2 \quad (4)$$

Multiplying this number of pixels with the required frame rate will give us the necessary information bandwidth.

Considering a holographic video display with a display size of 200 mm $\times$ 200 mm, viewing angles of $30^\circ \times 20^\circ$ and frame rate of 60 Hz, it would require the system to have an information bandwidth capable of delivering at least $\sim 10^{12}$ pixels/s for just the red colour ($\lambda \sim 633$ nm). If we want to support full colour with additional green and blue, the information bandwidth required will be more than tripled.

However, existing display devices described in this section can only manage to deliver at most $\sim 10^{10}$ pixels/s, which is two orders less than that required. Besides, the available computational speed and data bus transfer rate at present are also nowhere near the levels needed for real time videos. Such hardware limitations significantly restrict the development and deployment of holographic 3D displays.

Selected Holographic Display Systems

In the last a few decades, there have been many holographic display prototypes reported. We review here a few representative systems, to illustrate how different holographic displays are developed.

QinetiQ system

QinetiQ developed a holographic 3D display which is known as the so-called Active Tiling (AT) display system (Stanley *et al.*, 2000; Slinger *et al.*, 2001), in which an electrically addressed SLM (EASLM) is used to write the hologram onto an OASLM time-sequentially. The OASLM works as the information keeper. After all the holograms are written onto an OASLM array, the images are read out at once. Because the EASLM can be driven very quickly and the OASLM array can reconstruct a hologram at a larger size, the AT system potentially can deliver several giga-pixels at once.

In 2004, QinetiQ achieved an active tiling system that combines 4 channels as shown in Fig. 3(a) (Stanley *et al.*, 2004), each of which contains one $1024 \times 1024 \times 750$ Hz binary EASLM (ferroelectric LCOS) and a 5×5 OASLM array with the total equivalent resolution of 5120×5120 . As a result, the system displayed a CGH of 20,480 pixels wide $\times$ 5120 pixels high at a pixel pitch of 6.6 μm and a frame rate for the whole hologram up to 30 Hz, which is equal to 3000 Mpixel per second. As a result, full parallax holographic images with the size of 140 mm wide were reconstructed as shown in Fig. 3(b) (Stanley *et al.*, 2004).

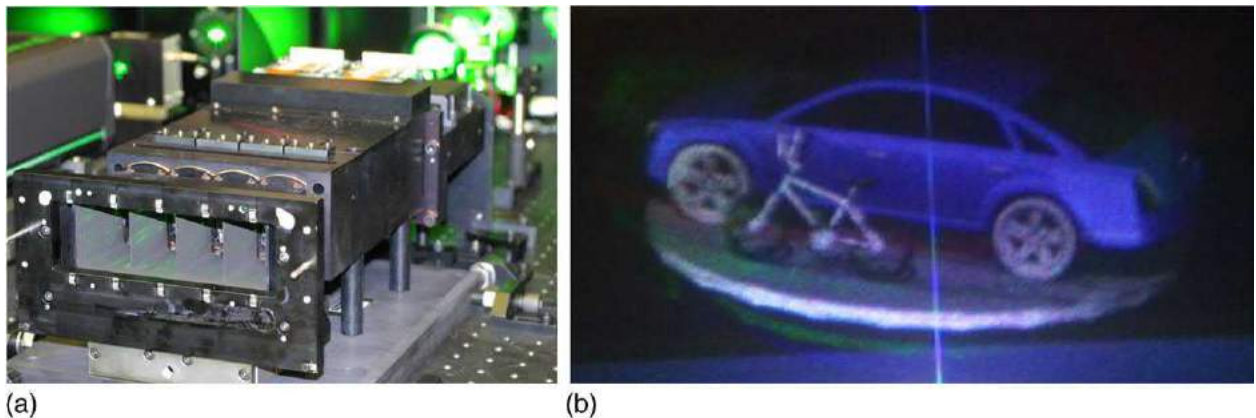


Fig. 3 (a) The QinetiQ 4 channel Active Tiling system; (b) image of an animated movie, showing a full parallax and 140 mm wide holographic image replayed by using a 100 Mpixel Active Tiling system. Reproduced from Stanley, M., Smith, M.A., Smith, A.P., *et al.*, 3D electronic holography display system using a 100-megapixel spatial light modulator. Proc. SPIE 5249, 297.

Mark systems

The Mark holographic systems are developed by Massachusetts Institute of Technology (MIT) Media Lab. There have been a few versions already since the first Mark I system was developed in early 1990s (Lucente, 1994; Smalley, 2006; Smalley *et al.*, 2007). The work in 2013 (Smalley *et al.*, 2013) showed anisotropic waveguide-based modulators (AWM) with more than 40 channels. Each channel can deliver equivalent to 100 million pixels per second (100 Mpixels/s). The total amount of information in pixels reaches the order of Gpixels/s. The diffracted output of light from the modulator was scanned into 2D with horizontal and vertical galvanometric mirrors. The reported holographic video monitor based on AWM is shown in Fig. 4. The holographic stereogram images were generated with a native resolution of 296 pixels $\times$ 156 pixels for one view, then the image resolution was improved to 29,600 pixels $\times$ 156 pixels by scanning, and finally to a composite resolution of 355,200 pixels $\times$ 156 pixels by stitching 12 of these images together. The size of display area is 35 mm $\times$ 20 mm. The system is still under development, and new versions of Mark IV and V are in the pipeline (Gneiting *et al.*, 2016).

National institute of information and communications technology (NICT)

NICT used 16 units of 4K $\times$ 2K-pixel SLMs to project holograms as shown in Fig. 5, with an information delivery rate in the order of 10^9 pixels/s (Sasaki *et al.*, 2014a,b). To tile multiple SLMs together seamlessly, a complex optical system is designed. The experimental device was able to produce full-parallax colourful holographic 3D at video rate of 20 fps with an image size of 74 mm $\times$ 42 mm and a horizontal viewing-zone angle of 5.6 deg $\times$ 2.8 deg. They also developed 8k $\times$ 4k LCOS chips and tilted three of them together to deliver 6 billion pixels/s and create a real time holographic 3D colour display (Senoh *et al.*, 2013, 2014).

Agency for science, technology and research (A*STAR)

In comparison with the NICT system, A*STAR system not only used physical tiling but also employed optical scan-tiling of high-speed SLMs to increase the total pixel count of holograms, as shown in Fig. 6 (Lum *et al.*, 2013). 24 units of high-speed ferroelectric LCOS of 1280 $\times$ 1080 pixels in resolution were physically tiled to form a 3 $\times$ 8 SLM array. The optical output from this array was tiled in space to form a large image equivalent to that of 36 $\times$ 8 SLMs by using a horizontal galvanometric scanning mirror with 12 steps. A large computer-generated hologram for the 3 $\times$ 8 SLM array was calculated and divided to upload onto

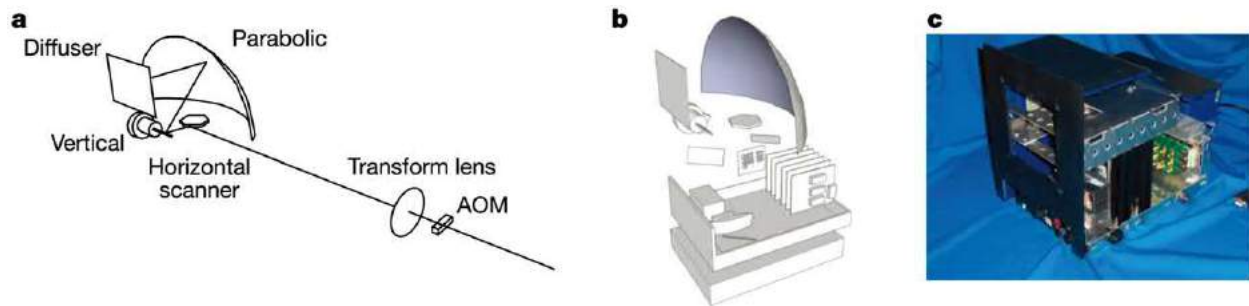


Fig. 4 A holographic video system based on AWMs, (a) optical layout, (b) folding of optical path, and (c) assembled system. Reproduced from Smalley D.E., Smithwick, Q.Y.J., Bove, V.M., Barabas, J., Jolly, S., 2013. Anisotropic leaky-mode modulator for holographic video displays. *Nature* 498, 313–317.

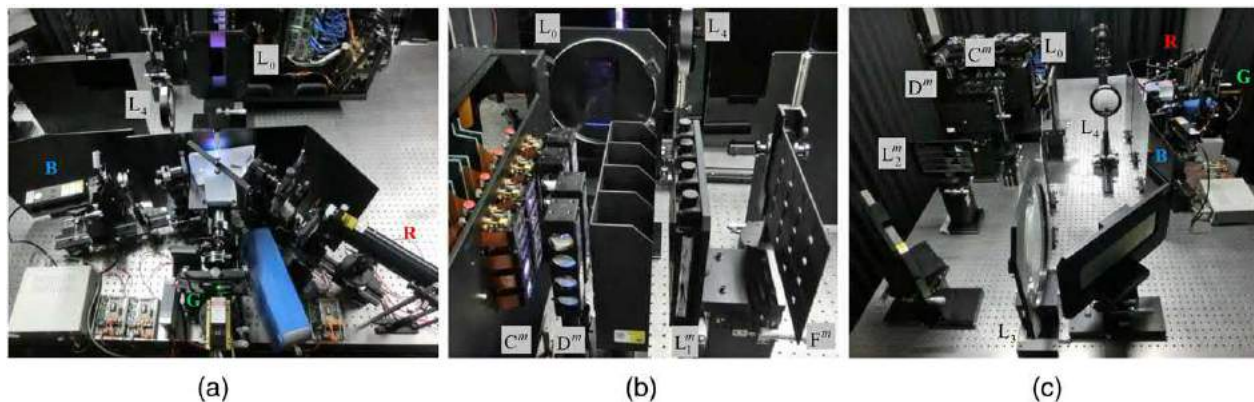


Fig. 5 The NICT (a) light source optical unit, and (b) image reading out optical unit, and (c) an overview of the whole optical system (Sasaki *et al.*, 2014).

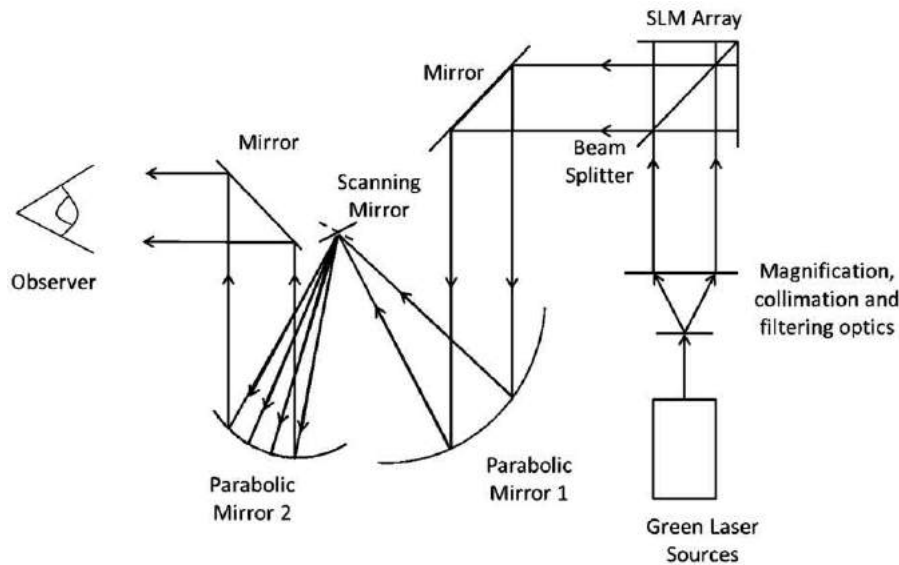


Fig. 6 The optical layout diagram of the A*STAR system. Reproduced from Lum, Z.M.A., Liang, X., Pan, Y., Zheng, R., Xu, X., 2013. Increasing pixel count of holograms for three-dimensional holographic display by optical scan-tiling. *Optical Engineering* 52, 15802–15802.

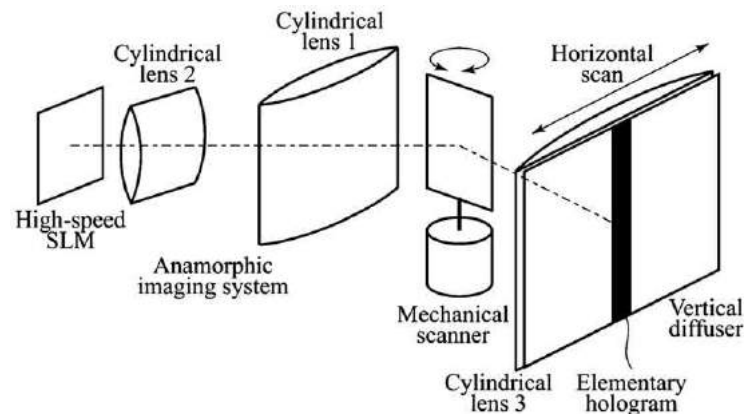


Fig. 7 The TAUT scree-scanning holographic display system. Reproduced from Takaki, Y., Okada, N., 2009. Hologram generation by horizontal scanning of a high-speed spatial light modulator. *Appl. Opt.* 48, 3255–3260.

individual SLMs (Pan *et al.*, 2013). A full-colour and full-parallax three-dimensional image of 10 in. in width was projected at a rate of 60 frames per second.

Tokyo university of agriculture and technology (TAUT)

TAUT developed a one-dimensional (1D) scanning approach called screen-scanning system, which consists of two orthogonally aligned cylindrical lenses (Takaki *et al.*, 2015; Takaki and Okada, 2010, 2009). The sub-holograms generated from a DMD was de-magnified in the horizontal direction and magnified along the vertical direction. The vertically expanded elementary holograms were then scanned horizontally using a galvanometric scanner onto a vertical diffuser screen to achieve a large screen size. As the pixel pitch is de-magnified horizontally, the horizontal viewing angle is enlarged. The vertical parallax is removed by the use of the vertical diffuser, making this display system one of the “horizontal parallax only” (HPO) systems. An illustration of the system is shown in Fig. 7 (Takaki and Okada, 2009).

Besides, TAUT also developed the resolution distribution technique (Takaki and Hayashi, 2008) and the viewing-zone scanning system (Takaki and Fujii, 2014). The resolution redistribution applies multiple shifted light sources to re-shape the effective reconstructed window, which is more suitable for one direction scanning. The viewing-zone scanning system employs a magnification image system to increase the hologram in both horizontal and vertical direction. In this case, the viewing angle is reduced. Then the reduced viewing zone is scanned by the horizontal scanner to enlarge the viewing zone, as shown in Fig. 8 (Takaki and Fujii, 2014).

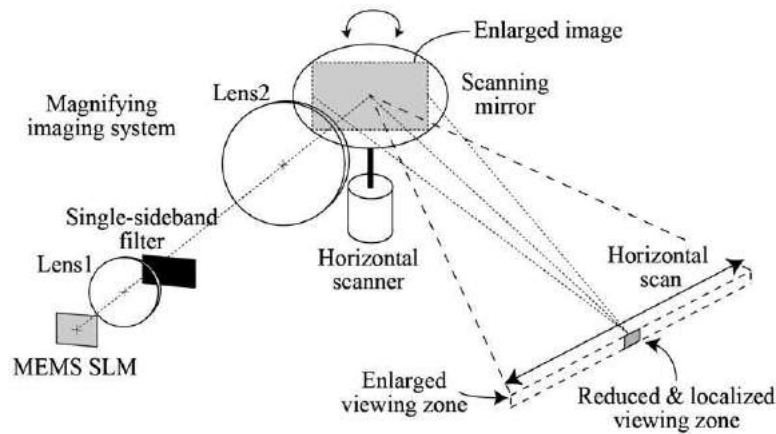


Fig. 8 Viewing zone-scanning holographic display system. Reproduced from Takaki,Y., Fujii,K., 2014. Viewing-zone scanning holographic display using a MEMS spatial light modulator. *Opt. Express* 22, 24713–24721.

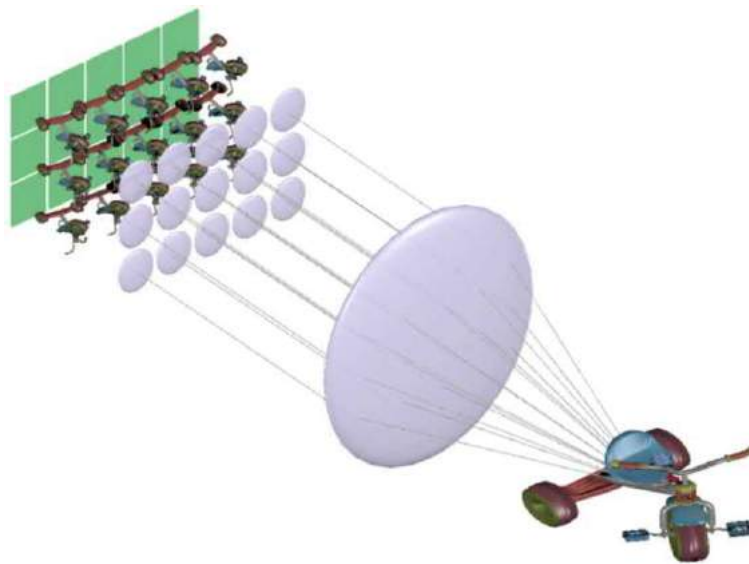


Fig. 9 Course integral holographic display system. Reproduced from Chen, J.S., Smithwick, Q.Y.J., Chu, D.P., 2016. Coarse integral holography approach for real 3D color video displays. *Opt. Express* 24, 6705–6718.

Course integral holographic display (CIH)

CIH display uses coarse integral optics to angularly tile low spatial bandwidth sub-holograms to form 3D images of a decent size with full parallax features and wide viewing angles in both horizontal and vertical directions, as shown in [Fig. 9](#) ([Chen et al., 2016](#)). The result reported in 2015 ([Chen et al., 2016](#)) showed full-colour holographic 3D images consisting of 141.6 Mpixels for each colour with a size of 49.6 mm × 49.6 mm, a frame rate of 23.33 Hz and viewing angles of $12^\circ \times 3.2^\circ$. Compared with integral imaging and holographic display, CIH has a better performance on producing accommodation depth cue than integral imaging, and larger viewing angles than that of a normal holographic display.

A diffraction-based scanning 3D colour video holographic display

A diffraction-based scanning 3D colour video holographic display was developed employing tiled gratings and a vertical diffuser, as shown in [Fig. 10](#) ([Jia et al., 2017](#)). Its main idea is to use a rotational tiled grating, which is slim and lightweight, to replace a galvanometer or polygon mirror, both of which have a limited scanning speed. It can significantly increase the distributing capability of the scanner, and accommodate the information provided from a high-speed SLM, such as a DMD, without any waste. Its result reported in 2017 ([Jia et al., 2017](#)) showed a full colour HPO holographic display projecting 3D images with the size of 60 mm × 30 mm and horizontal viewing angle of 37° at a frame rate of 75 Hz.

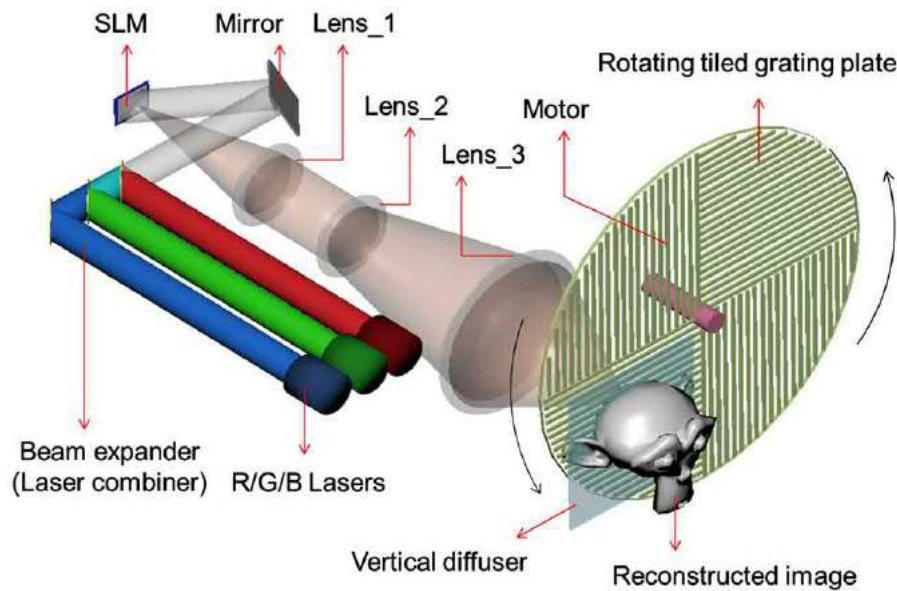


Fig. 10 A diffraction-based scanning 3D colour video holographic display system. Reproduced from Jia, J., Chen, J.S., Yao, J., Chu, D.P., 2017. A scalable diffraction-based scanning 3D colour video display as demonstrated by using tiled gratings and a vertical diffuser. *Scientific Reports* 7, 44656.

360 degrees holographic display

Apart from the most common front-viewing holographic display systems, there are various systems to produce images with a horizontal viewing angle of 360 degrees (full-around) (Butler *et al.*, 2011; Inoue and Takaki, 2015; Jones *et al.*, 2007; Kakue *et al.*, 2015; Miyazaki *et al.*, 2012; Sando *et al.*, 2016; Takaki and Nakamura, 2014; Takaki and Uchida, 2012; Xia *et al.*, 2013; Yoshida, 2016). This requires a huge amount of data and an information distribution system with high capacity. DMD (s), which can provide a high frame rate, are often used in 360 degrees system for prototype demonstrations. For the information distribution system, there are various approaches, such as the uses of array of projectors (Inoue and Takaki, 2015), tilted mirror (Sando *et al.*, 2016) and parabolic mirror (Butler *et al.*, 2011; Kakue *et al.*, 2015), respectively.

Holography can be integrated with the concept of 360 degrees viewing to support accommodation cue for every view (Inoue and Takaki, 2015; Lim *et al.*, 2016; Teng *et al.*, 2012). A rotating tilted mirror was used to distribute holographic images projected from a high speed SLM (Teng *et al.*, 2012). A parabolic mirror was also used to project a floating holographic image on the table (Lim *et al.*, 2016). The largest diameter of the screen was 35.2 mm among these systems. To support a large image size, a large mirror is necessary but it will increase the weight of the rotation component and slow down the scanning speed. The rotating off-axis lens, which can be made light-weight, was used to distribute holographic images to 360 degrees (Inoue and Takaki, 2015).

To summarise, a comparison of the bandwidth property of the selected holographic displays is shown in Fig. 11. The bandwidth improving strategies are spatial tiling, temporal tiling or hybrid tiling methods. No matter which technology is used for a holographic display, the most fundamental principle is to utilize the information bandwidth, in other words, the total amount of information which the device can provide per second to reconstruct holographic images.

Light Sources

Light sources for holographic displays are also developed to improve the quality of holographic images. Holograms, which takes the advantage of interference and diffraction to reconstruct images, require a coherent light source. However, a high level of coherence also produces speckle as a side-effect. To reduce speckle, various approaches are proposed. Some of them are to search a suitable light source (Kozacki and Chlipala, 2016; Mori *et al.*, 2014; Pan *et al.*, 2014; Pang *et al.*, 2015; Zhao *et al.*, 2012), which has an appropriate coherence (low enough to reduce the speckle but high enough to maintain the image quality). Although the image sharpness and speckle are related to both spatial and temporal coherences, studies show that the spatial coherence can be linked directly to the image sharpness and the temporal coherence to the speckle (Deng and Chu, 2017). To provide more natural colour from holograms, lasers with a wide range of wavelengths are researched (Sarakinis *et al.*, 2013) to refine the colour match and the commercial light sources for holography are also developed (Bjelkhagen and Brotherton-Ratcliffe, 2014). Currently, these sources already exist, and the challenge is to reduce its cost (by searching for a cheaper alternative or simplifying its fabrication) and make them affordable.

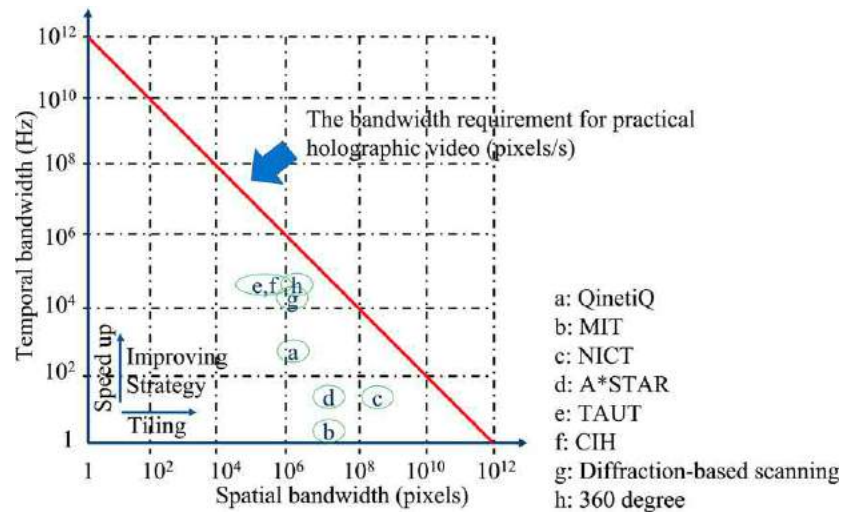


Fig. 11 The bandwidth property of selected holographic display systems.

Computer Generated Holograms

The process of calculating a digital hologram has three steps: (1) define the data of the object to be reconstructed, which can be a virtually created object or the graphic data of a real object; (2) calculate the complex field propagated from the object to the hologram plane according to the wave propagation principles; (3) transfer the calculated hologram onto a film or a SLM, using a phase-only, amplitude-only, binary or any specific format.

In general, the films or SLMs cannot control the amplitude and the phase of a complex field simultaneously, unless using some cascaded approaches mentioned earlier (Hsieh *et al.*, 2007; Jesacher *et al.*, 2008a).

To transfer hologram on films, there are techniques to encode and decode the complex field information of the targeted wave plane: detour holograms and kinoform holograms. The detour-phase hologram, which was invented by Brown and Lohmann, (1969), is a technique transforming the complex information into binary patterns printable by most plotting devices. The principle of making a detour-phase hologram is to divide the hologram into separate cells, each of which contains several pixels. The total number of pixels of the "on" state represents the amplitude and the positions of these pixels encode the phase. Note that although it is named as "phase", detour-phase technique is an amplitude modulation approach. It produces an approximation of the planned modulation and involves with errors, and it has been less used after SLMs being widely used.

The detour-phase technique was later replaced by phase contour interferogram proposed by Lee (1979, 1974). Phase contour interferogram is an encoding approach, which plots the pattern along the hologram phase contour. It can be adapted to plotters to produce a better approximation than that of detour-phase technique. It can also be applied on amplitude-only SLMs which can modulate the intensity in grey levels pixel by pixel.

Unlike pure amplitude modulations, the kinoform hologram supports a pure phase modulation and provides a better efficiency (Lesem *et al.*, 1969). The concept of kinoform holograms is based on an assumption that the phase carries the majority of information about the objects and the loss of amplitude information is acceptable (Oppenheim and Lim, 1981).

Along with the development of SLMs, amplitude-only and phase-only SLMs are commercialized and easily available. The name of detour, contour and kinoform are less used and people simply call amplitude-only or phase-only holograms. Interestingly, while DMD, a popular SLM for holographic displays, is a binary device, it does not apply detour hologram techniques. One reason is that detour-phase hologram has a limited accuracy. Besides, the necessary pixel count per second for a holographic image is high, and it is preferred to keep each pixel as one independent cell, rather than to spend multiple pixels to form an equivalent cell.

Overall, the transformation from a complex hologram to a phase-only, amplitude-only, detour or binary hologram (either phase or amplitude) is only after the complex wave propagation calculation is done. To calculate the complex amplitude and phase, there are mainly two principles. One is to adopt the wave propagation equation on each object element (point, polygon, etc.) and calculate their accumulation on the hologram plane. Another is to simply take the Fourier transformation (FT) of the target 2D image as the hologram, since the reverse FT of this pattern is the target 2D image. This corresponds to the optics fact that two planes on both sides of a lens are FT of each other. Since Fourier hologram is easier to calculate and it becomes the common implementation for 2D image hologram projection. Following we will introduce algorithms about the complex hologram calculation in 2D/2.5D/3D images.

Phase Retrieval Approach for 2D Images

To calculate holograms, we should have both amplitude and phase information of the object light. The intensity of an object is fixed according to the target while the phase of the object light at the object surface can be changed to the arbitrary number since

eyes cannot directly detect the phase of light. Meanwhile, most films and SLMs cannot modulate amplitude and phase of light simultaneously (note that there are few approaches can do this, but they still cannot be widely used). In the phase-only modulation case, the amplitude information of the calculated hologram is lost and this directly degrades the reconstruction quality of the image. Since the phase of object light is flexible, it is possible to find a phase pattern to form a phase-only hologram, which has the least loss of image quality. This action of searching is called “phase retrieval”, which is an optimization problem. The fast Fourier transformation (FFT) is the common mathematics tool used in phase retrieval. However, the FFT assumes that a hologram is an infinitely periodic arrangement without considering the aperture effect, hence the pixels in the Fourier plane may be wider than expected and overlap with each other. The consequent noise is on the image plane. Therefore, a multiple iteration method is necessary to minimise this problem.

Gerchberg-Saxton (GS) algorithm is probably the most famous approach for phase retrieval (Gerchberg and Saxton, 1972). The principle of GS algorithm is to calculate the converging solution through iterative FFT between two planes by fixing the amplitude components in one plane (the 2D image plane) and optimize the phase components in the other plane (the hologram plane). Please refers to Gerchberg and Saxton, (1972) for detailed procedure steps.

Different improvements of the GS algorithm have been proposed. The Fienup algorithm (Fienup, 1978) accelerates the converging speed by feeding the errors between the actual reconstruction image and the target image as a feedback into the iteration process. The Fienup with don't-care (Fidoc) method was proposed in 1986 (Akahori, 1986) and the further improved one in 1990 (Wyrowski, 1990). It is a specific case of the Fienup algorithm by applying the don't-care area on the image plane. The don't-care areas may be placed around or within the image, which is well known as zero padding and zero circumscribing. The Fidoc algorithm performs better than GS and Fienup algorithms in terms of image quality but at the cost of increased computational load and number of pixels because of the additional noise disposal area.

In commercialization, a handheld holographic video projector with 60 Hz frame rate and real-time CGHs was developed in CAPE of Cambridge University and demonstrated by Alps Electric in Japan in 2006, as shown in Fig. 12 (SPIE, 2017).

Holographic Stereogram for 2.5D Images

Between 2D and 3D images, we should also define the 2.5D image. By 2.5D we mean that it provides some but not all depth cues for 3D visual perception. For example, a display using multiple 2D images can provide motion parallax for 3D perception but there is a lack of accommodation cue. Stereogram, which provides 3D perception by binocular cue, is another example of a 2.5D image.

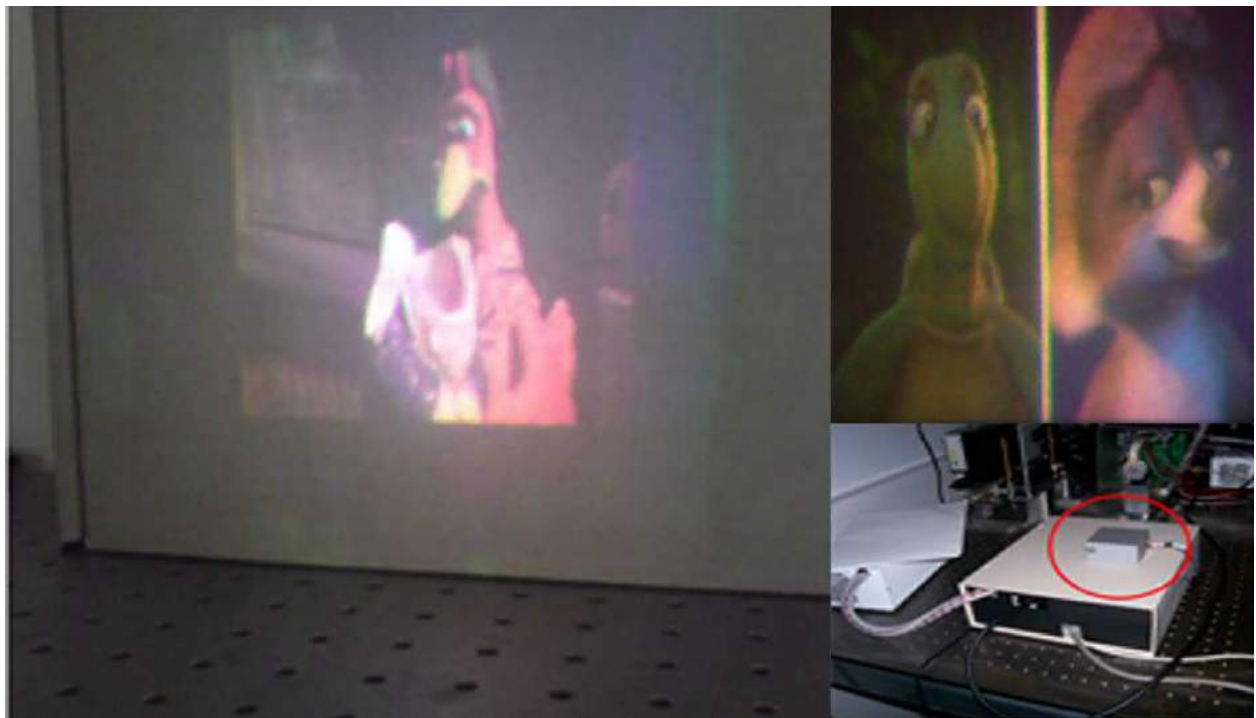


Fig. 12 A handheld holographic video projector with real-time CGHs developed in CAPE of Cambridge University and made by Alps Electric in Japan. Reproduced from SPIE, 2017. Available at: <http://spie.org/newsroom/1015-projection-display-using-computer-generated-phase-screens?highlight=x2408&ArticleID=x20068> (accessed 21.05.17).

The concept of multiple 2D images is also used in holographic display, and it is known as holographic stereogram. In early days, it records 2D pictures of the object taken from different angles onto different parts of the hologram by moving a slit step by step (Benton and Bove, 2008). Later on in 1980s, computer calculation of holographic stereogram was also realized. The principle of implementation is to divide the hologram into multiple segments as the basic elements, get the 2D projection with different angles of the 3D object on the hologram plane to get the ray information, transform the ray information (including intensity and angle) to the diffraction pattern on each segment, and then form the final hologram. The way to transform a ray into a diffraction pattern is to conduct Fourier transformation (FT) on a 2D image representing the ray direction and intensity in the image domain. Since the diffraction calculation can be done by FFT, the calculation load of computer holographic stereograms is much less than that of the conventional CGH techniques for 3D images (which will be introduced in Section 3D Images (Point/Polygon/Layer Based Methods and Improvements)).

Another advantage of holographic stereogram is that it can produce occlusion effect directly because of its use of 2D projection of the 3D object, and it can be integrated with 3D graphics without much additional computation (Barabas *et al.*, 2011; Sando *et al.*, 2013), while occlusion culling methods for conventional 3D image hologram calculation involves with heavy computation (Ichikawa *et al.*, 2013a,b; Matsushima and Kondoh, 2004; Matsushima and Nakahara, 2009; Underkoffler, 1997; Zhang *et al.*, 2011).

The concept of using individual rays is an approximation to the 3D wave propagation. To improve this approximation, phase-added stereogram was proposed (Gao *et al.*, 2015; Kang *et al.*, 2016; Yamaguchi *et al.*, 1993). It calculates the directional information of the light from the object and uses a phase factor to yield a better coherent wavefront. As a result, the depth range of the image is wider while the amount of calculation stays almost the same. In commercialization, Geola and Zebra Imaging print customized 3D holographic stereogram holograms.

3D Images (Point/ Polygon/Layer Based Methods and Improvements)

To calculate the complex wave plane of one object, the information of location and intensity should be considered. The object data can come from a virtual object or the graphics data reconstructed from a real object by a 3D camera. To simplify the calculation procedure, a 3D object is broken down to smaller basic units and the wave propagation of each unit is calculated and added up to form the final hologram.

According to the type of the basic unit in use, 3D image CGH methods can be classified into three categories. One is the point-source method, which samples 3D objects by cloud points (Eldeib and Yabe, 1996; Stein *et al.*, 1992). Another is the polygon-based method (Matsushima, 2010, 2008; Matsushima and Shimobaba, 2009), which discretizes 3D objects into a number of polygons, usually triangles. The third one is image-based approach (Barabas *et al.*, 2011; Chen and Chu, 2015; Jia *et al.*, 2014) which hybrids graphic rendered 2D images with other approaches to generate 3D effects.

The point-based method is also known as the coherent ray tracing (CRT) method. It treats every point on the object as a self-illuminating point source and calculates the emitted spherical wave to all the pixels on a hologram plane using a light propagation function (Matsushima and Takai, 2000; Nishitsuji *et al.*, 2015; Weng *et al.*, 2011). To support a high quality 3D image, the object needs to be sampled with a large number of points and this leads to a high computational load.

To improve the calculation speed, Lucente proposed the look-up-table (LUT) method in 1993 (Lucente, 1993) to store all the possible diffraction patterns for the points offline. The point diffraction patterns are directly read out from the table during the online calculation. However, the necessary size of the LUT is huge, such as hundreds of gigabytes, and it is overwhelming even for a modern computer. Several improved LUT methods (Gao *et al.*, 2015; Dong *et al.*, 2014; Jia *et al.*, 2013; Jiao *et al.*, 2017; Kim and Kim, 2008; Kim *et al.*, 2013, 2012, 2008, 2015; Kwon *et al.*, 2016; Pan *et al.*, 2009) were proposed to reduce the memory size down to the order of kilobytes.

Another method to improve the calculation speed of point-based methods is the wave recording plane (WRP) technique developed by Phan *et al.* (2014) and Shimobaba *et al.* (2009, 2010). A WRP is a virtual plane that is located at a near distance to the objects. Because the short distance and the limited diffraction angle, each object point will only diffract light within a small area onto the WRP. Then the information on the WRP propagates to the hologram plane by Fresnel diffraction. With the WRP technique, a hologram of 2048×2048 pixels in size, representing an image scene composed of 4×10^6 points, can be generated in less than 25 ms (Tsang *et al.*, 2011).

On the other hand, in the polygon-based method, the 3D object is represented by thousands of polygons rather than millions of points. It essentially reduces the number of basic units while the computation for each basic unit becomes more complicated. The computation of a polygon unit involves a rotational transformation in the frequency domain to get the pattern corresponding to the normal angle, and the angular spectrum approach which can calculate the complex information along the propagation axis (Matsushima, 2008). In 2008, a fully analytical method was developed (Ahrenberg *et al.*, 2008). It proposes a new structure to conduct polygon based methods (FFT is no longer used). The method was implemented on GPU for parallel computing in order to achieve a faster speed and better image quality (Liu *et al.*, 2010). When polygon-approach was developed, the overall calculation speed was expected to be faster than that of other approaches. However, point-based and image-based methods were also improved, and now polygon-based approach remains the slowest approach.

The image-based approach was firstly used in holographic stereogram to provide multiple 2D images for different views. This approach, later on, is applied to the calculation of real 3D images. A hybrid of point-based approach and image-based approach was proposed (Barabas *et al.*, 2011; Jia *et al.*, 2014), so that the occlusion cue is provided by graphics while the accommodation cue

is provided by point-based method. The layer-based method was also proposed (Chen and Chu, 2015; Chen *et al.*, 2014). It uses 2D rendered depth map to divide a 3D image into several layers. The hologram of each 2D image is calculated, by adding a lens of a specific focal length to the FT of that layer. Finally, holograms of individual layers are added together to form the hologram of the 3D image. It takes advantage of the factor that eyes have a limited resolution of accommodation cue, so few layers are enough to project a good accommodation cue. Research shows that only 28 depth layers are required to produce a full depth of field from 25 cm to infinite distance for human eyes (Akeley, 2004; Rolland *et al.*, 1999).

Error Reduction in CGH

Existing SLMs, such as LCOS, can only display either the amplitude or the phase component of light. The loss of either amplitude or phase information brought up the phase retrieval issue, as discussed in Section Phase Retrieval Approach for 2D Images. Different phase retrieval approaches, such as GS and Fienup methods, optimize the reconstruction image quality by using an iterative calculation. In contrast, the loss of either amplitude or phase information may lead to heavy errors in the holograms calculated by using wavefront propagation functions discussed in 3D Images (Point/Polygon/Layer Based Methods and Improvements). The random phase was first added to the object points prior to the generation of the digital hologram to minimise this error, but the random phase makes the speckle noise worse (Pan *et al.*, 2014; Makowski *et al.*, 2016; Shimobaba *et al.*, 2016,a,b; Takaki and Taira, 2016; Takaki and Yokouchi, 2011).

One solution to this problem is to cascade two SLMs and reconstruct the complex amplitude hologram (Hsieh *et al.*, 2007; Leportier *et al.*, 2016). In some implementations, an optical element (can be a prism or grating) is used to merge the wave propagations from different regions of the same device in order to achieve the complex modulation (Jesacher *et al.*, 2008a,b; Liu *et al.*, 2011; Song *et al.*, 2012). Another effective approach is to place a moving diffuser at the imaging plane (Kuratomi *et al.*, 2010; Pan and Shih, 2014). It can average the noise, but requires an extra mechanical device.

To deal with this problem through calculation, the error reduction methods are proposed which can be grouped into temporal-based and spatial-based methods. For the temporal-based method, multiple sub-frames, each of which represents the same object scene added with different random phase patterns, are displayed rapidly in sequence to the observers to average the speckle noise (Buckley, 2011; Buckley *et al.*, 2006). It is also known as the random averaging method. The superposition of N holograms with uncorrelated phase patterns can reduce the speckle noise to $1/\sqrt{N}$ at the cost of computational load and the reduction of effective frame rate.

The spatial-based error reduction method is also known as the error diffusion method (Tsang and Poon, 2013; Tsang *et al.*, 2014). It converts a digital complex hologram into a phase-only hologram without using a random phase pattern. The process is accomplished with an error diffusion mechanism. It scans each hologram pixel sequentially, updates it with the values it was past to, then pass the error to the unvisited neighborhood, and then carry on. Compared with the temporal-based method, this method does not consume the frame rate of the SLM, and the reconstructed images are very similar to those obtained with the original complex holograms. The downside of spatial-based error distribution is its additional calculations. It is noted that the error diffusion methods can be applied directly on an existing complex amplitude hologram, without knowing the original data of the object image. This allows it to be applied on holograms calculated by any approach as a post-processing approach.

Holographic Display VR/AR

Virtual reality (VR) has been a technology dream for decades. It is now gradually commercialized after the launches of Oculus Rift, HTC Vive and Sony PlayStation VR between 2015 and 2016. At the same time augmented reality (AR) is also attracting public attention. One of the representative examples is Microsoft HoloLens, released in 2016. It successfully applies a holographic waveguide combiner to be part of the AR headset while the image contents stay as 2D. On the other hand, holography can be used as true 3D image projectors and it attracts efforts on building a prototype of holographic HMD. This section introduces these two holographic applications in VR/AR.

Holographic Waveguide/Lightguide Combiner in VR/AR HMD

In early 1990s, the concept of a holographic waveguide display was patented in which a waveguide hologram was used to couple a collimated image into a glass-based waveguide, and transmit the images by Total Internal Reflection (TIR). The spectral and angular Bragg selectivity is the key functionality in the holographic combiners both in reflective or transmission holograms. Compared with a transmission hologram, a reflective hologram can achieve a narrower spectral bandwidth and a wider angular bandwidth at the same time. A larger angular bandwidth can produce a larger field of view, and narrower spectral bandwidth can reduce the chromatic aberration.

Since the waveguide can be made very thin and light, there are companies such as Sony Ltd, Konica/Minolta and Nokia/Vuzix which have applied this concept to the HMDs. Their works mainly focused on improving the light efficiency and reducing the chromatic aberration in waveguide due to the wavelength dependency of holograms.

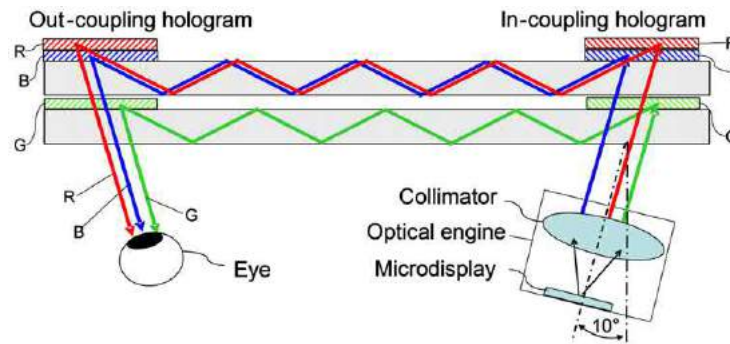


Fig. 13 Sony holographic waveguide combiner. Reproduced from Mukawa, H., Akutsu, K., Matsumura, S., 2008.8.4: distinguished paper: A full color eyewear display using holographic planar waveguides. In: SID Symposium Digest of Technical Papers 39, pp. 89–92.

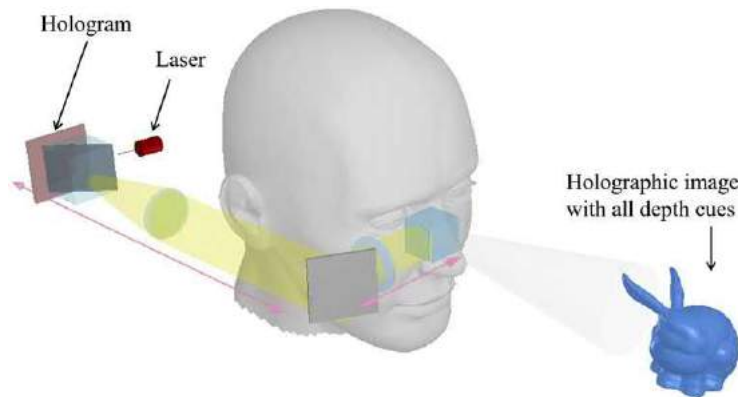


Fig. 14 A holographic 3D head-mounted display. Reproduced from Chen, J.-S., Chu, D.P., 2015. Improved layer-based method for rapid hologram generation and real-time interactive holographic display applications. Opt. Express 23, 18143–18155.

In Sony and Konica/Minolta's holographic waveguide combiners, the traditional reflective holograms are used with high diffraction efficiency because of the Bragg selectivity. The chromatic aberration in waveguide can be compensated by using three superimposed holograms, such as in Sony's approach shown in [Fig. 13](#) ([Mukawa et al., 2008](#)). Each hologram is exposed to a specific wavelength. Otherwise, Konica/Minolta and other researchers exposed one hologram three times to different wavelengths with the corresponding illuminating angle ([Kress and Starner, 2013](#); [Shi et al., 2012](#)). One drawback of using holographic waveguides is its low diffraction efficiency due to the material limitation, and its use in HMDs is replaced by free space ([Zhou et al., 2017](#)), free-form optics ([Hu and Hua, 2014](#); [Hua and Javidi, 2014](#)) or prisms ([Love et al., 2009](#)).

Holographic 3D Display in VR/AR HMD

Conventional stereoscopic HMDs may cause eye fatigue because of the conflict of accommodation and convergence depth cues. Applications of holographic 3D displays in HMD is one solution to provide accommodation cue and remove the fatigue.

A number of prototypes have been built ([Chen and Chu, 2015](#); [Li et al., 2016](#); [Tanjung et al., 2010](#); [Yoneyama et al., 2014, 2013](#)). The concept of a holographic 3D HMD is shown in [Fig. 14](#) ([Chen and Chu, 2015](#)), in which the holographic 3D images generated by the SLM are projected into human pupil directly. The necessary bandwidth (both optical and computational) for a holographic HMD is much less than that of a holographic display because of the small size of a pupil, therefore holographic HMDs are much more feasible in terms of bandwidth. Nevertheless, there are still issues remain to be solved, including speckle noise, small field of view (FOV), chromatic aberration and small eye-box. This field is still relatively new, and its development will take time before any solid impact is produced.

Key Challenges and Conclusions

Although huge efforts and funding have been put into the development of digital holographic displays for several decades, the process of its commercialization is still very slow. The key challenge is that the user experience, including the image size, viewing angle and image quality, is not as good as that of traditional 2D displays. In comparison with 2D displays, a high quality 3D

display requires a pixel size smaller than $1\ \mu\text{m}$ and information bandwidth larger than 10^{12} pixels/s for a decent size display screen. This is difficult to achieve by the existing technology. Such a huge amount of data also leads to the heavy computational load for rendering the 3D data and generate CGHs. Moreover, it is difficult to realize the real-time display because of the limited data transmission bandwidth of the data bus. For example, a DMD can run at a high frame rate but pre-uploading of the images is required, which can take a couple of minutes or even more.

To move holographic 3D display technologies forward into a mass market is not easy. We need to develop suitable technologies and products to overcome all the obstacles. Fortunately, future SLMs have the potential to be made with sub-micron pitches (Shrestha *et al.*, 2015; Isomae *et al.*, 2016) and to be driven at a fast speed (Forth Dimensions Displays, 2017), in order to produce the required amount of data. High speed computers and fast CGH calculation algorithms make it possible to generate holograms in real time (Ichihashi *et al.*, 2012), and customized data board for transmitting holograms may solve the data bus issue.

Finally, to reconstruct only the wavefront at the eye position using eye-tracking technology can significantly reduce the amount of information to be processed. Similarly, applying digital holographic 3D technology for HMDs requires much less amount of information if compared with traditional holographic 3D displays.

See also: Module: Digital Holography

References

- Ahrenberg, L., Benzie, P., Magnor, M., Watson, J., 2008. Computer generated holograms from three dimensional meshes using an analytic light transport model. *Appl. Opt.* 47, 1567–1574.
- Akabori, H., 1986. Spectrum leveling by an iterative algorithm with a dummy area for synthesizing the kinoform. *Appl. Opt.* 25, 802–811.
- Akeley, K., 2004. Achieving Near-Correct Focus Cues Using Multiple Image Planes. PhD Thesis. Stanford University.
- Barabas, J., Jolly, S., Smalley, D.E., Bove, V.M., 2011. Diffraction specific coherent panoramagrams of real scenes. *Proc. SPIE* 7957, 795702.
- Benton, S.A., Bove, J.V.M., 2008. *Holographic Imaging*. Wiley-Interscience.
- Berneth, H., Bruder, F.-K., Fäcke, T., *et al.*, 2013. Holographic recordings high beam ratios Improved Bayfol®HX photopolymer. *Proc. SPIE* 8776, 877603.
- Bjelkhagen, H.I., Brotherton-Ratcliffe, D., 2014. Ultrarealistic imaging the future of display holography. *Optical Engineering* 53, 112310.
- Bjelkhagen, H.I., Crosby, P.G., Green, D.P.M., Mirlis, E., Phillips, N.J., 2008. Fabrication ultra-fine-grain silver halide recording material color holography. *Proc. SPIE* 6912, 691209.
- Blanche, P.-A., Bablumian, A., Voorakaranam, R., 2010a. Future photorefractive based holographic 3D display. *Proc. SPIE* 7619, 76190L.
- Blanche, P.-A., Bablumian, A., Voorakaranam, R., *et al.*, 2010b. Holographic three-dimensional telepresence using large-area photorefractive polymer. *Nature* 468, 80–83.
- Brown, B.R., Lohmann, A.W., 1969. Computer-generated binary holograms. *IBM J. Res. Dev.* 13, 160–168.
- Buckley, E., 2011. Real-time error diffusion for signal-to-noise ratio improvement in a holographic projection system. *Journal of Display Technology* 7, 70–76.
- Buckley, E., Cable, A., Lawrence, N., Wilkinson, T., 2006. Viewing angle enhancement for two- and three-dimensional holographic displays with random superresolution phase masks. *Appl. Opt.* 45, 7334–7341.
- Butler, A., Hilliges, O., Izadi, S., *et al.*, 2011. Vermeer direct interaction with a 360° viewable 3D display. In: *Proceedings of the 24th Annual ACM Symposium User Interface Software Technology, UIST'11 (ACM, 2011)*, pp. 569–576.
- Chen, J., Morris, S.M., Wilkinson, T.D., Freeman, J.P., Coles, H.J., 2009. High speed liquid crystal over silicon display based on the flexoelectro-optic effect. *Opt. Express* 17, 7130–7137.
- Chen, J.-S., Chu, D., Smithwick, Q., 2014. Rapid hologram generation utilizing layer-based approach and graphic rendering for realistic three-dimensional image reconstruction by angular tiling. *Journal of Electronic Imaging* 23, 023016.
- Chen, J.-S., Chu, D.P., 2015. Improved layer-based method for rapid hologram generation and real-time interactive holographic display applications. *Opt. Express* 23, 18143–18155.
- Chen, J.S., Smithwick, Q.Y.J., Chu, D.P., 2016. Coarse integral holography approach for real 3D color video displays. *Opt. Express* 24, 6705–6718.
- Cody, D., Naydenova, I., Mihaylova, E., 2012. New non-toxic holographic photopolymer material. *Journal of Optics* 14, 015601.
- Collings, N., Davey, T., Christmas, J., Chu, D.P., Crossland, B., 2011. The applications and technology of phase-only liquid crystal on silicon devices. *Journal of Display Technology* 7, 112–119.
- Deng, Y., Chu, D., 2017. Coherence properties of different light sources and their effect on the image sharpness and speckle of holographic display. *Sci. Rep.* 7 (1), 5893.
- Denisjuk, Y.N., 1962. On the reflection of optical properties of an object in a wave field of light scattered by it. *Doklady Akademii Nauk SSSR* 144, 1275–1278.
- Dennis, G., 1948. A new microscopic principle. *Nature* 161, 777–778.
- Dong, X.-B., Kim, S.-C., Kim, E.-S., 2014. MPEG-based novel look-up table for rapid generation of video holograms of fast-moving three-dimensional objects. *Opt. Express* 22, 8047–8067.
- Eldeib, H., Yabe, T., 1996. Computer-generated holograms: A fast ray-tracing approach. *Journal of King Saud University - Computer and Information Sciences* 8, 139–151.
- Finup, J.R., 1978. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.* 3, 27–29.
- Forth Dimensions Displays, 2017. Available at: <http://www.forthdd.com/> (accessed 21.05.17).
- Gambogi Jr, W.J., Weber, A.M., Trout, T.J., 1994. Advances and applications of DuPont holographic photopolymers. *Proc. SPIE* 2043, 2.
- Gao, C., Liu, J., Li, X., *et al.*, 2015. Accurate compressed look up table method for CGH in 3D holographic display. *Opt. Express* 23, 33194–33204.
- Gerchberg, R.W., Saxton, W.O., 1972. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik (Jena)* 35, 237.
- Gneiting, S., Smalley, D.E., Qaderi, K., *et al.*, 2016. Optimizations for robust, high-efficiency, waveguide-based holographic video. In: *2016 IEEE 14th International Conference Industrial Informatics (INDIN)*, pp. 576–581.
- Hecht, E., 1998. *Optics*. Addison-Wesley.
- Hsieh, M.-L., Chen, M.-L., Cheng, C.-J., 2007. Improvement of the complex modulated characteristic of cascaded liquid crystal spatial light modulators by using a novel amplitude compensated technique. *Optical Engineering* 46.070501–070501–3.
- Hu, X., Hua, H., 2014. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. *Opt. Express* 22, 13896–13903.
- Hua, H., Javidi, B., 2014. A 3D integral imaging optical see through head-mounted display. *Opt. Express* 22, 13484–13491.
- Ichihashi, Y., Oi, R., Senoh, T., Yamamoto, K., Kurita, T., 2012. Real-time capture and reconstruction system with multiple GPUs for a 3D live scene by a generation from 4K IP images to 8K holograms. *Opt. Express* 20, 21645–21655.

- Ichikawa, T., Yamaguchi, K., Sakamoto, Y., 2013a. Realistic expression for full-parallax computer-generated holograms with the ray-tracing method. *Appl. Opt.* 52, A201–A209.
- Ichikawa, T., Yoneyama, T., Sakamoto, Y., 2013b. CGH calculation with the ray tracing method for the Fourier transform optical system. *Opt. Express* 21, 32019–32031.
- Inoue, T., Takaki, Y., 2015. Table screen 360-degree holographic display using circular viewing-zone scanning. *Opt. Express* 23, 6533–6542.
- Isomae, Y., Shibata, Y., Ishinabe, T., Fujikake, H., 2016. Optical phase modulation properties of 1 μm -pitch LCOS with dielectric walls for wide-viewing-angle holographic displays. In: *SID Symposium Digest of Technical Papers* 47, 1670–1673.
- Jasper Display Corp., 2017. Available at: <http://www.jasperdisplay.com/success-stories/> (accessed 21.05.17).
- Jesacher, A., Maurer, C., Schwaighofer, A., Bernet, S., Ritsch-Marte, M., 2008a. Near perfect hologram reconstruction with a spatial light modulator. *Opt. Express* 16, 2597–2603.
- Jesacher, A., Maurer, C., Schwaighofer, A., Bernet, S., Ritsch-Marte, M., 2008b. Full phase and amplitude control of holographic optical tweezers with high efficiency. *Opt. Express* 16, 4479–4486.
- Jia, J., Chen, J.S., Yao, J., Chu, D.P., 2017. A scalable diffraction-based scanning 3D colour video display as demonstrated by using tiled gratings and a vertical diffuser. *Scientific Reports* 7, 44656.
- Jia, J., Liu, J., Jin, G., Wang, Y., 2014. Fast and effective occlusion culling for 3D holographic displays by inverse orthographic projection with low angular sampling. *Appl. Opt.* 53, 6287–6293.
- Jia, J., Wang, Y., Liu, J., *et al.*, 2013. Reducing the memory usage for effective computer-generated hologram calculation using compressed look-up table in full-color holographic display. *Appl. Opt.* 52, 1404–1412.
- Jiao, S., Zhuang, Z., Zou, W., 2017. Fast computer generated hologram calculation with a mini look-up table incorporated with radial symmetric interpolation. *Opt. Express* 25, 112–123.
- Jones, A., McDowall, I., Yamada, H., Bolas, M., Debevec, P., 2007. Rendering for an interactive 360° light field display. *ACM Trans. Graph.* 26.
- Jurbergs, D., Bruder, F.-K., Deuber, F., *et al.*, 2009. New recording materials for the holographic industry. *Proc. SPIE* 7233, 72330K.
- Kakue, T., Nishitsuji, T., Kawashima, T., *et al.*, 2015. Aerial projection of three-dimensional motion pictures by electro-holography and parabolic mirrors. *Scientific Reports* 5, 11750.
- Kang, H., Stoykova, E., Yoshikawa, H., 2016. Fast phase-added stereogram algorithm for generation of photorealistic 3D content. *Appl. Opt.* 55, A135–A143.
- Kim, S.-C., Dong, X.-B., Kim, E.-S., 2015. Accelerated one-step generation of full-color holographic videos using a color-tunable novel-look-up-table method for holographic three-dimensional television broadcasting. *Scientific Reports* 5, 14056.
- Kim, S.-C., Dong, X.-B., Kwon, M.-W., Kim, E.-S., 2013. Fast generation of video holograms of three-dimensional moving objects using a motion compensation-based novel look-up table. *Opt. Express* 21, 11568–11584.
- Kim, S.-C., Kim, E.-S., 2008. Effective generation of digital holograms of three-dimensional objects using a novel look-up table method. *Appl. Opt.* 47, D55–D62.
- Kim, S.-C., Kim, J.-M., Kim, E.-S., 2012. Effective memory reduction of the novel look-up table with one-dimensional sub-principle fringe patterns in computer-generated holograms. *Opt. Express* 20, 12021–12034.
- Kim, S.-C., Yoon, J.-H., Kim, E.-S., 2008. Fast generation of three-dimensional video holograms by combined use of data compression and lookup table techniques. *Appl. Opt.* 47, 5986–5995.
- Kinashi, K., Wang, Y., Tsujimura, S., Sakai, W., Tsutsumi, N., 2012. Dynamic holographic images using photorefractive composites. In: *Biomedical Optics 3-D Imaging*. Optical Society of America, p. JM3A.58.
- Kozacki, T., Chlipala, M., 2016. Color holographic display with white light LED source and single phase only SLM. *Opt. Express* 24, 2189–2199.
- Kress, B., Starnier, T., 2013. A review of head-mounted displays technologies and applications for consumer electronics. *Proc. SPIE* 8720, 87200A.
- Kuratomi, Y., Sekiya, K., Satoh, H., *et al.*, 2010. Speckle reduction mechanism in laser rear projection displays using a small moving diffuser. *J. Opt. Soc. Am. A* 27, 1812–1817.
- Kwon, M.-W., Kim, S.-C., Kim, E.-S., 2016. Three-directional motion-compensation mask-based novel look-up table on graphics processing units for video-rate generation of digital holographic videos of three-dimensional scenes. *Appl. Opt.* 55, A22–A31.
- Lee, W.-H., 1974. Binary synthetic holograms. *Appl. Opt.* 13, 1677–1682.
- Lee, W.-H., 1979. Binary computer-generated holograms. *Appl. Opt.* 18, 3661–3669.
- Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. *J. Opt. Soc. Am.* 52, 1123–1130.
- Leportier, T., Park, M.C., Kim, T., 2016. Numerical alignment of spatial light modulators for complex modulation in holographic displays. *Journal of Display Technology* 12, 1000–1007.
- Lesem, L.B., Hirsch, P.M., Jordan, J.A., 1969. The kinoform: A new wavefront reconstruction device. *IBM Journal of Research and Development* 13, 150–155.
- Li, G., Lee, D., Jeong, Y., Cho, J., Lee, B., 2016. Holographic display for see-through augmented reality using mirror-lens holographic optical element. *Opt. Lett.* 41, 2486–2489.
- Lim, Y., Hong, K., Kim, H., *et al.*, 2016. 360-degree tabletop electronic holographic display. *Opt. Express* 24, 24999–25009.
- Liu, J.-P., Hsieh, W.-Y., Poon, T.-C., Tsang, P., 2011. Complex Fresnel hologram display using a single SLM. *Appl. Opt.* 50, H128–H135.
- Liu, Y.-Z., Dong, J.-W., Pu, Y.-Y., *et al.*, 2010. High-speed full analytical holographic computations for true-life scenes. *Opt. Express* 18, 3345–3351.
- Love, G.D., Hoffman, D.M., Hands, P.J.W., *et al.*, 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Opt. Express* 17, 15716–15725.
- Lucente, M., 1994. Diffraction-specific fringe computation for electro-holography. PhD, Massachusetts Institute of Technology.
- Lucente, M.E., 1993. Interactive computation of holograms using a look-up table. *Journal of Electronic Imaging* 2, 28–34.
- Lum, Z.M.A., Liang, X., Pan, Y., Zheng, R., Xu, X., 2013. Increasing pixel count of holograms for three-dimensional holographic display by optical scan-tiling. *Optical Engineering* 52, 15802.
- Makowski, M., Shimobaba, T., Ito, T., 2016. Increased depth of focus in random-phase-free holographic projection. *Chin. Opt. Lett.* 14, 120901.
- Matsushima, K., 2008. Formulation of the rotational transformation of wave fields and their application to digital holography. *Appl. Opt.* 47, D110–D116.
- Matsushima, K., 2010. Wave-field rendering in computational holography: The polygon-based method for full-parallax high-definition CGHs. In: *2010 IEEE/ACIS Proceedings of the 9th International Conference Computer Information Science*, pp. 846–851.
- Matsushima, K., Kondoh, A., 2004. A wave-optical algorithm for hidden-surface removal in digitally synthetic full-parallax holograms for three-dimensional objects. *Proc. SPIE* 5290, 90.
- Matsushima, K., Nakahara, S., 2009. Extremely high-definition full-parallax computer-generated hologram created by the polygon-based method. *Appl. Opt.* 48, H54–H63.
- Matsushima, K., Shimobaba, T., 2009. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Opt. Express* 17, 19662–19673.
- Matsushima, K., Takai, M., 2000. Recurrence formulas for fast creation of synthetic three-dimensional holograms. *Appl. Opt.* 39, 6587–6594.
- Miyazaki, D., Akasaka, N., Okoda, K., Maeda, Y., Mukai, T., 2012. Floating three-dimensional display viewable from 360 degrees. *Proc. SPIE* 8288, 82881H.
- Mori, Y., Fukuoka, T., Nomura, T., 2014. Speckle reduction in holographic projection by random pixel separation with time multiplexing. *Appl. Opt.* 53, 8182–8188.
- Mukawa, H., Akutsu, K., Matsumura, I., *et al.*, 2008. 8.4: Distinguished paper: A full color eyewear display using holographic planar waveguides. In: *SID Symposium Digest of Technical Papers* 39, pp. 89–92.
- Nishitsuji, T., Shimobaba, T., Kakue, T., Arai, D., Ito, T., 2015. Simple and fast cosine approximation method for computer-generated hologram calculation. *Opt. Express* 23, 32465–32470.

- Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. *Proceedings of the IEEE* 69, 529–541.
- Pan, J.-W., Shih, C.-H., 2014. Speckle reduction and maintaining contrast in a LASER pico projector using a vibrating symmetric diffuser. *Opt. Express* 22, 6464–6477.
- Pan, Y., Xu, X., Liang, X., *et al.*, 2013. Large-pixel-count hologram data processing for holographic 3D display. *Proc. SPIE* 8644, 86440F.
- Pan, Y., Wang, Y., Liu, J., Li, X., Jia, J., 2014. Improved full analytical polygon-based method using Fourier analysis of the three-dimensional affine transformation. *Appl. Opt.* 53, 1354–1362.
- Pan, Y., Xu, X., Solanki, S., *et al.*, 2009. Fast CGH computation using S-LUT on GPU. *Opt. Express* 17, 18543–18555.
- Pang, X.-N., Chen, D.-C., Ding, Y.-C., *et al.*, 2015. Image quality improvement of polygon computer generated holography. *Opt. Express* 23, 19066–19073.
- Phan, A.-H., Piao, M., Gil, S.-K., Kim, N., 2014. Generation speed and reconstructed image quality enhancement of a long-depth object using double wavefront recording planes and a GPU. *Appl. Opt.* 53, 4817–4824.
- Rasmussen, P.H., Ramanujam, P.S., Hvilsted, S., Berg, R.H., 1999. A remarkably efficient azobenzene peptide for holographic information storage. *Journal of the American Chemical Society* 121, 4738–4743.
- Rolland, J.P., Krueger, M.W., Goon, A.A., 1999. Dynamic focusing in head-mounted displays. *Proc. SPIE* 3639, 463–470.
- Sando, Y., Barada, D., Yatagai, T., 2013. Hidden surface removal of computer-generated holograms for arbitrary diffraction directions. *Appl. Opt.* 52, 4871–4876.
- Sando, Y., Barada, D., Yatagai, T., 2016. Optical rotation compensation for a holographic 3D display with a 360 degree horizontal viewing zone. *Appl. Opt.* 55, 8589–8595.
- Sarakinos, A., Zervos, N., Lembessis, A., 2013. Holofos: An optimized LED illumination system for color reflection holograms display. *Proc. SPIE* 8644, 86440I.
- Sasaki, H., Yamamoto, K., Ichihashi, Y., Senoh, T., 2014a. Image size scalable full-parallax coloured three-dimensional video by electronic holography. *Scientific Reports* 4, 4000.
- Sasaki, H., Yamamoto, K., Wakunami, K., *et al.*, 2014b. Large size three-dimensional video by electronic holography using multiple spatial light modulators. *Scientific Reports* 4, 6177.
- Senoh, T., Ichihashi, Y., Oi, R., Sasaki, H., Yamamoto, K., 2013. Study of a holographic TV system based on multi-view images and depth maps. *Proc. SPIE* 8644, 86440A.
- Senoh, T., Wakunami, K., Ichihashi, Y., *et al.*, 2014. Multiview image and depth map coding for holographic TV system. *Optical Engineering* 53, 112302.
- Shi, R., Liu, J., Zhao, H., *et al.*, 2012. Chromatic dispersion correction in planar waveguide using one-layer volume holograms based on three-step exposure. *Appl. Opt.* 51, 4703–4708.
- Shimobaba, T., Kakue, T., Ito, T., 2016. Random phase-free computer holography and its applications. *Proc. SPIE* 9867, 98670M.
- Shimobaba, T., Makowski, M., Nagahama, Y., *et al.*, 2016. Color computer-generated hologram generation using the random phase-free method and color space conversion. *Appl. Opt.* 55, 4159–4165.
- Shimobaba, T., Masuda, N., Ito, T., 2009. Simple and fast calculation algorithm for computer-generated hologram with wavefront recording plane. *Opt. Lett.* 34, 3133–3135.
- Shimobaba, T., Nakayama, H., Masuda, N., Ito, T., 2010. Rapid calculation algorithm of Fresnel computer-generated-hologram using look-up table and wavefront-recording plane methods for three-dimensional display. *Opt. Express* 18, 19504–19509.
- Shishido, A., 2010. Rewritable holograms based on azobenzene-containing liquid-crystalline polymers. *Polym. J.* 42, 525–533.
- Shrestha, P.K., Chun, Y.T., Chu, D.P., 2015. A high-resolution optically addressed spatial light modulator based on ZnO nanoparticles. *Light Sci. Appl.* 4, e259.
- Slinger, C.W., Bannister, R.W., Cameron, C.D., *et al.*, 2001. Progress and prospects for practical electroholographic display systems. *Proc. SPIE* 4296, 18.
- Slinger, C.W., Cameron, C.D., Coomber, S.D., *et al.*, 2004. Recent developments computer-generated holography: Toward a practical electroholography system interactive 3D visualization. *Proc. SPIE* 27, 5290.
- Smalley, D.E., 2006. Integrated optics for holographic video. M.E.: Massachusetts Institute of Technology.
- Smalley, D.E., Smithwick, Q.Y.J., Bove, V.M., 2007. Holographic video display based on guided-wave acousto-optic devices. *Proc. SPIE* 6488, 64880L.
- Smalley, D.E., Smithwick, Q.Y.J., Bove, V.M., Barabas, J., Jolly, S., 2013. Anisotropic leaky-mode modulator for holographic video displays. *Nature* 498, 313–317.
- Song, H., Sung, G., Choi, S., *et al.*, 2012. Optimal synthesis of double-phase computer generated holograms using a phase-only spatial light modulator with grating filter. *Opt. Express* 20, 29844–29853.
- SPIE, 2017. Available at: <http://spie.org/newsroom/1015-projection-display-using-computer-generated-phase-screens?Highlight=x2408&ArticleID=x20068> (accessed 21.05.17).
- Stanley, M., Conway, P.B., Coomber, S.D., *et al.*, 2000. Novel electro-optic modulator system for the production of dynamic images from giga-pixel computer-generated holograms. *Proc. SPIE* 3956, 13.
- Stanley, M., Smith, M.A., Smith, A.P., *et al.*, 2004. 3D electronic holography display system using a 100-megapixel spatial light modulator. *Proc. SPIE* 5249, 297.
- Stein, A.D., Wang, Z., Leigh Jr., J.S., 1992. Computer-generated holograms: A simplified ray-tracing approach. *Computers in Physics* 6, 389–392.
- Stevenson, S.H., 1997. DuPont multicolor holographic recording films. *Proc. SPIE* 3011, 231.
- Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E.S., Watts, M.R., 2013. Large-scale nanophotonic phased array. *Nature* 493, 195–199.
- Takaki, Y., Fujii, K., 2014. Viewing-zone scanning holographic display using a MEMS spatial light modulator. *Opt. Express* 22, 24713–24721.
- Takaki, Y., Hayashi, Y., 2008. Increased horizontal viewing zone angle of a hologram by resolution redistribution of a spatial light modulator. *Appl. Opt.* 47, D6–D11.
- Takaki, Y., Matsumoto, Y., Nakajima, T., 2015. Color image generation for screen-scanning holographic display. *Opt. Express* 23, 26986–26998.
- Takaki, Y., Nakamura, J., 2014. Generation of 360-degree color three-dimensional images using a small array of high-speed projectors to provide multiple vertical viewpoints. *Opt. Express* 22, 8779–8789.
- Takaki, Y., Okada, N., 2009. Hologram generation by horizontal scanning of a high-speed spatial light modulator. *Appl. Opt.* 48, 3255–3260.
- Takaki, Y., Okada, N., 2010. Reduction of image blurring of horizontally scanning holographic display. *Opt. Express* 18, 11327–11334.
- Takaki, Y., Taira, K., 2016. Speckle regularization and miniaturization of computer-generated holographic stereograms. *Opt. Express* 24, 6328–6340.
- Takaki, Y., Uchida, S., 2012. Table screen 360-degree three-dimensional display using a small array of high-speed projectors. *Opt. Express* 20, 8848–8861.
- Takaki, Y., Yokouchi, M., 2011. Speckle-free and grayscale hologram reconstruction using time-multiplexing technique. *Opt. Express* 19, 7567–7579.
- Tanjung, R.B.A., Xu, X., Liang, X., *et al.*, 2010. Digital holographic three-dimensional display of 50-Mpixel holograms using a two-axis scanning mirror device. *Optical Engineering* 49, 025801. –025801–9.
- Tay, S., Blanche, P.-A., Voorakaranam, R., *et al.*, 2008. An updatable holographic three-dimensional display. *Nature* 451, 694–698.
- Teng, D., Liu, L., Wang, Z., Sun, B., Wang, B., 2012. All-around holographic three-dimensional light field display. *Optics Communications* 285, 4235–4240.
- Texas Instruments, 2017. Available at: <http://www.ti.com/lit/dsp/advanced-light-control/ultraviolet/ultraviolet-products.page#p2439-32.552> (accessed 21.05.17).
- Texas Instruments, 2017. Available at: <http://www.ti.com/lit/dsp/technology/markets/dlp-products-for-4k-ultra-hd.page> (accessed 21.05.17).
- Tsang, P., Cheung, W.-K., Poon, T.-C., Zhou, C., 2011. Holographic video at 40 frames per second for 4-million object points. *Opt. Express* 19, 15205–15211.
- Tsang, P.W.M., Jiao, A.S.M., Poon, T.-C., 2014. Fast conversion of digital Fresnel hologram to phase-only hologram based on localized error diffusion and redistribution. *Opt. Express* 22, 5060–5066.
- Tsang, P.W.M., Poon, T.-C., 2013. Novel method for converting digital Fresnel hologram to phase-only hologram based on bidirectional error diffusion. *Opt. Express* 21, 23680–23686.
- Underkoffler, J.S., 1997. Occlusion processing and smooth surface shading for fully computed synthetic holography. *Proc. SPIE* 19, 3011.
- Weng, J., Shimobaba, T., Oikawa, M., Masuda, N., Ito, T., 2011. Fast recurrence algorithm for computer-generated-hologram. In: *Digital Holography Three-Dimensional Imaging*. Optical Society of America, p. DTuC21.
- Wyrowski, F., 1990. Diffractive optical elements iterative calculation of quantized, blazed phase structures. *J. Opt. Soc. Am. A* 7, 961–969.
- Xia, X., Liu, X., Li, H., *et al.*, 2013. A 360-degree floating 3D display based on light field regeneration. *Opt. Express* 21, 11237–11247.
- Yamaguchi, M., Hoshino, H., Honda, T., Ohya, N., 1993. Phase-added stereogram: Calculation of hologram using computer graphics technique. *Proc. SPIE* 1914, 25.

- Yoneyama, T., Ichikawa, T., Sakamoto, Y., 2014. Semi-portable full-color electro-holographic display with small size. Proc. SPIE 9006, 900617.
- Yoneyama, T., Yang, C., Sakamoto, Y., Okuyama, F., 2013. Eyepiece-type full-color electro-holographic binocular display with see-through vision. In: Digital Holography Three-Dimensional Imaging. Optical Society of America, p. DW2A.11.
- Yoshida, S., 2016. fVisiOn 360-degree viewable glasses-free tabletop 3D display composed of conical screen and modular projector arrays. Opt. Express 24, 13194–13203.
- Zacharovas, S.J., Rodin, A.M., Ratcliffe, D.B., Vergnes, F.R., 2001. Holographic materials available Geola. Proc. SPIE 4296, 206.
- Zhang, H., Collings, N., Chen, J., *et al.*, 2011. Full parallax three-dimensional display with occlusion effect using computer generated hologram. Optical Engineering 50, 074003–074005. [074003–].
- Zhang, Z., You, Z., Chu, D.P., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. Light Sci. Appl. 3, e213.
- Zhao, Y., Cao, L., Zhang, H., He, Q., 2012. Holographic display with LED illumination based on phase-only spatial light modulator. Proc. SPIE 8559, 85590B.
- Zhou, L., Chen, C.P., Wu, Y., *et al.*, 2017. See-through near-eye displays enabling vision correction. Opt. Express 25, 2130–2142.

Holography: Computer Generated Holograms

WJ Dallas, University of Arizona, Tucson, AZ, United States
AW Lohmann[†]

© 2018 Elsevier Ltd. All rights reserved.

From the Classical Hologram to the Computer Generated Hologram: CGH

Holography is a method to form images. The method consists of two steps: recording and reconstruction. In the recording step, some interference fringes are recorded on a photographic plate (Fig. 1(A)). The two interfering waves come from the object and from a reference light source. After being developed by the usual photo-chemical treatment the photographic plate is called a hologram. In the reconstruction step, the hologram acts as a diffracting object that is illuminated by a replica of the reference light. Due to diffraction, the light behind the hologram is split into three parts (Fig. 1(B)). One part proceeds to the observer who perceives a virtual object where the genuine object had been during the recording step.

This was a brief description of “classical – holography”. In “computer holography” the first step, recording, is synthetic. In other words, the propagation of the light from the object to the photographic plate is simulated digitally. The simulation includes the addition of the object wave to the reference wave and the subsequent modulus square process, which describes the transition from the two complex amplitudes to the hologram irradiance:

$$|u_O(x) + u_R(x)|^2 = I_H(x) \quad (1)$$

The amplitude transmittance of the hologram is

$$T_H(x) = c_0 + c_1 I_H(x) = c_1 u_O u_R^* + \dots \quad (2)$$

The coefficients c_0 and c_1 have to do with the photochemical development. In the reconstruction process (Fig. 1(B)) the transmittance T_H is illuminated by the reference beam. Hence, we get

$$u_R(x) T_H(x) = c_1 u_O(x) |u_R(x)|^2 + \dots \quad (3)$$

If the reference intensity $|u_R(x)|^2$ is constant, this term represents a reconstruction of the object wave $u_O(x)$. The omitted parts in Eq. (2) and in Eq. (3) describe those beams that are ignored by the observer (Fig. 1(B)). The hologram transmittance $T_H(x)$ is an analog signal:

$$0 \leq T_H \leq 1 \quad (4)$$

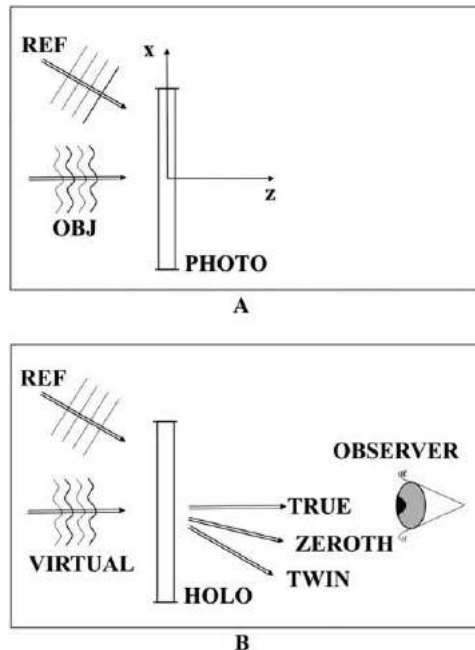


Fig. 1 Holographic image formation in two steps: (A) Recording the hologram. (B) Reconstructing the object wave.

[†]Deceased.

50 years ago, when computer holograms were invented (Brown and Lohmann, 1966; Lohmann and Paris, 1967), the available plotters could produce only binary transmittance patterns. Hence, it was necessary to replace the analog transmission $T_H(x)$ by a binary transmission $B_H(x)$. That is possible indeed, without loss of image quality. In addition, binary computer holograms have better light efficiency and a better signal-to-noise ratio. They are also more robust when being copied or printed.

Another advantage of a computer generated hologram (CGH) is that the object does not have to exist in reality. That is important for some applications. The genuine object may be difficult to manufacture or, if it is three-dimensional, difficult to illuminate. The CGH may be used to implement an algorithm for optical image processing. In that case, the term “object” becomes somewhat fictitious. We will discuss more about applications, in Section Some CGH Applications, towards the end of this entry.

From a Diffraction Grating to a Fourier CGH

Now we will explain, in some detail, the “Fourier type” CGH. Fourier holograms are more popular than two other types: “image-plane” CGH’s and “Fresnel-type” CGH’s. The Fresnel-type will be treated in a section on “software” because of some interest in 3-D hologram’s (Section About some CGH algorithms). Certain hardware issues will appear in Section On Some Hardware Aspects.

The explanation of Fourier CGH’s begins with grating diffraction (Fig. 2). The only uncommon feature in Fig. 2 is the off-axis location of the source. It is arranged such that the plus-first diffraction order will hit the center of the output plane.

We will now convert a simple Ronchi grating (Fig. 3(A)) into a Fourier CGH. The Ronchi grating transmittance can be expressed as the Fourier series:

$$G(x) = \sum_{(m)} C_m \exp\left(\frac{2\pi i m x}{D}\right); \quad C_0 = \frac{1}{2}; \quad C_m = \frac{\sin(\pi m/2)}{\pi m} \quad (5)$$

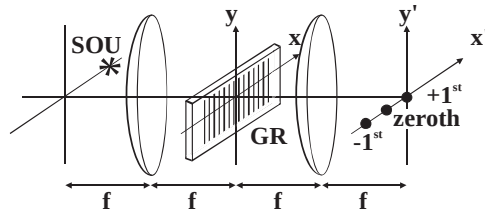


Fig. 2 Grating diffraction, but with an off-axis source.

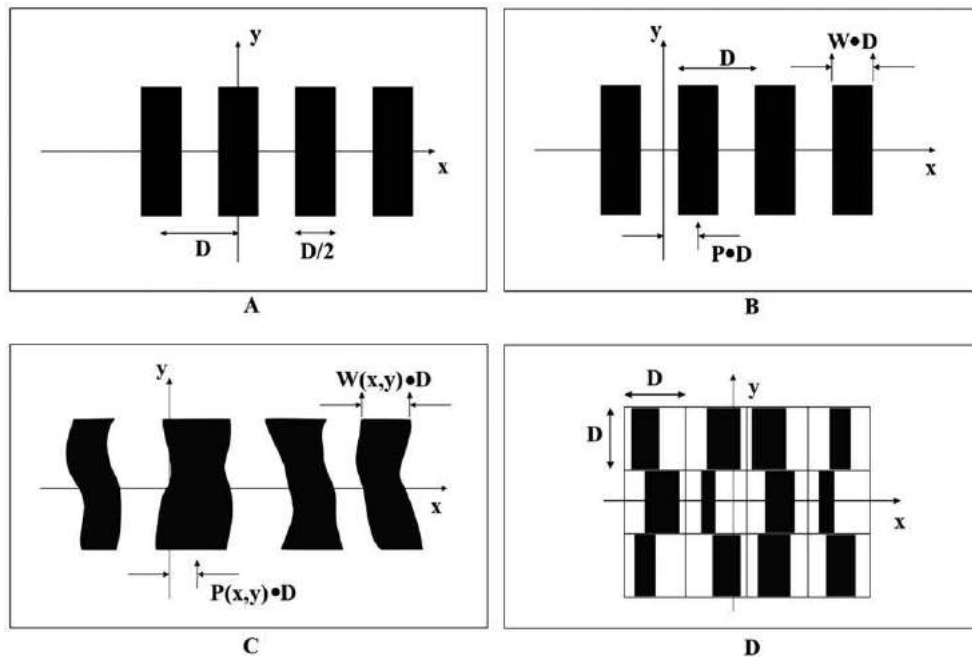


Fig. 3 Modifying a grating into a hologram in 3 steps: (A) Ronchi grating. (B) Modified width and location of the grating bars. (C) Distorted grating bars. (D) As in Fig. 3(C), but now discrete modifications.

Now we shift the grating bars by an amount PD . And, we modify the width from $D/2$ to WD , as shown in Fig. 3(B). The Fourier coefficients are now

$$C_m = \frac{\exp(2\pi i m P) \sin(\pi m W)}{\pi m} \quad (6)$$

The C_{+1} coefficient is responsible for the complex amplitude at the center of the readout plane:

$$C_{+1} = \frac{\exp(2\pi i P) \sin(\pi W)}{\pi} \quad (7)$$

We are able now to control the beam magnitude, $A = (\frac{1}{\pi})\sin(\pi W)$, by the relative bar width W and the phase $\phi = 2\pi P$ by the relative bar shift P . This particular kind of phase shift is known as “detour phase” (Hauk and Lohmann, 1958).

So far, the light that ends up at the center of the exit plane, leaves the grating as a plane wave with a wave vector parallel to the optical axis. Now we distort the grating on purpose (Fig. 3(C)):

$$W \rightarrow W(x, y); P \rightarrow P(x, y) \quad (8)$$

These distortions should be mild enough that the plus-first order light behind the distorted grating can be described by the complex amplitude:

$$\left(\frac{1}{\pi}\right) \sin[\pi W(x, y)] \exp[2\pi i P(x, y)] = u_H(x, y) \quad (9)$$

This is a generalization of Eq. (7). The phase $2\pi P(x, y)$ covers the range from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ if the shift is bounded by $|P| \leq \frac{1}{2}$. That is enough to cover the range of complex amplitudes within the circle of $|u_H| \leq \frac{1}{\pi}$. We will return shortly to the condition “mild-enough distortions” in the context of Fig. 4.

Most plotters move their pens only in x -direction and y -direction and printers place dots on a grid. Therefore it is convenient to replace the continuous (x, y) variations of W and P by piece-wise constant variations as in Fig. 3(D). That restriction turns out to be acceptable, as demonstrated by the first computer holograms. The theoretical proof involves the sampling theorem (Lohmann and Paris, 1967).

Recall that the light propagation through a 2-f system (Fig. 2, between grating and output plane) can be described by a Fourier transformation:

$$\iint u_H(x, y) \exp[-2\pi i (xx' + yy')/\lambda f] dx dy = \tilde{u}_H(x'/\lambda f, y'/\lambda f) \quad (10)$$

Suppose now that we wish to see a particular image: $v(x', y')$. Hence, we request that

$$v(x', y') = \tilde{u}_H(x'/\lambda f, y'/\lambda f) \text{RECT}(x'/\lambda f) \quad (11)$$

The RECT function indicates that only the PLUS first diffraction order is of interest. As a consequence we have to select the grating distortions $W(x, y)$ and $P(x, y)$ such that the hologram $u_H(x, y)$ is the Fourier transform of $v(x', y')$:

$$u_H(x, y) = \iint v(x', y') \exp[2\pi i (xx' + yy')/\lambda f] dx' dy' \quad (12)$$

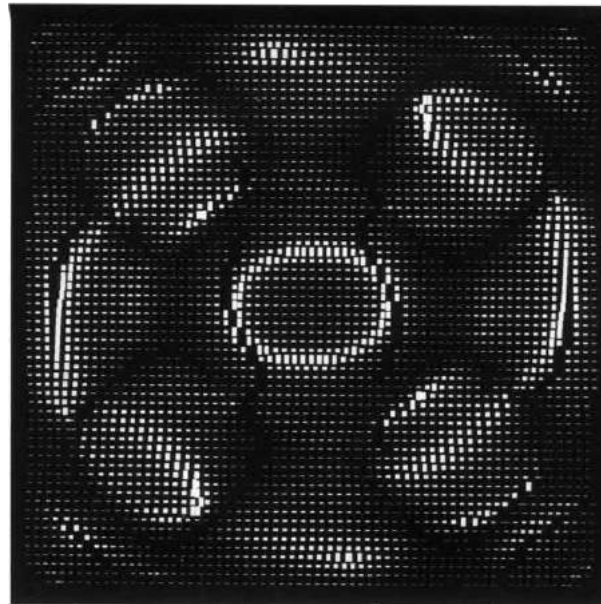


Fig. 4 CGH with 64×64 cells.

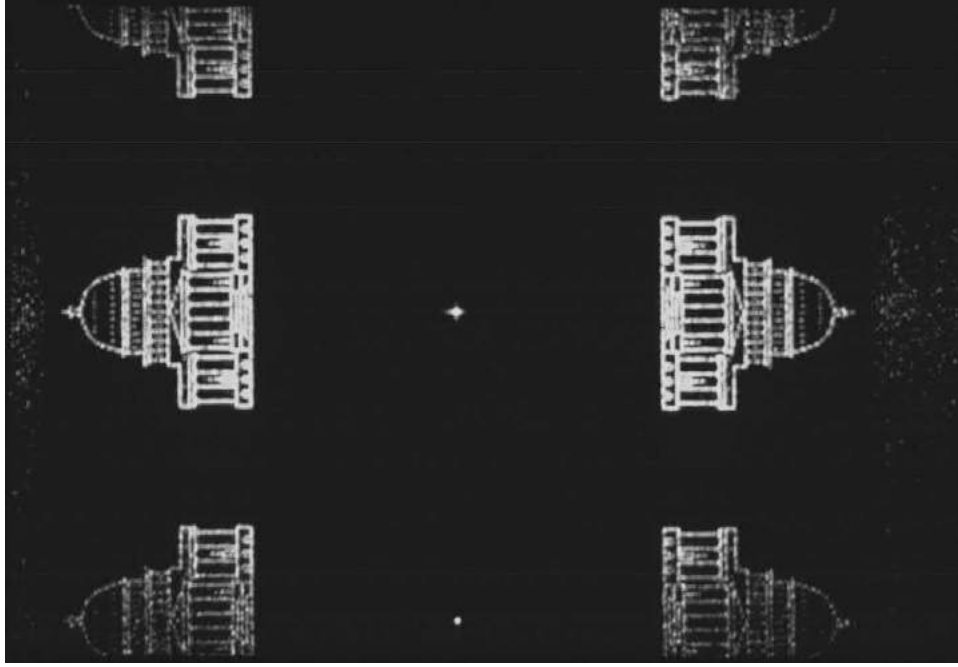


Fig. 5 A reconstruction from a 1024×1024 CGH. The true image and the symmetrical twin image appear. The zeroth-order is blocked. Reproduced from Brown, B.R., Lohmann, A.W., 1969. IBM J. R&D 13, 160.

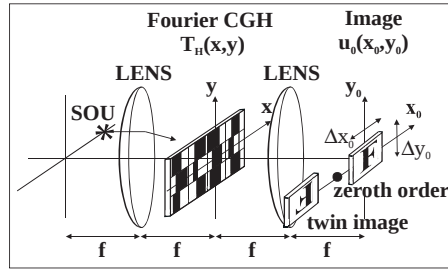


Fig. 6 Reconstruction setup for a Fourier CGH.

The condition “mild enough distortions” has been treated in the literature. There is not enough space here to present that part of the CGH theory. But we show an actual CGH with 64×64 cells (**Fig. 4**). This CGH is no longer a grating. But, it shows enough resemblance to our heuristic derivation. Hence, the CGH theory appears to be trustworthy. A CGH looks less regular if the image $v(x', y')$ contains a random phase that simulates a diffuser. The diffuser spreads light across the CGH and so levels out unacceptably large fluctuations in the amplitude.

With 512×512 or 1024×1024 cells in a CGH one can obtain an image of the quality seen in **Fig. 5** (Brown and Lohmann, 1969). The reconstruction setup is shown in **Fig. 6**, here with the letter F as the image.

About Some CGH Algorithms

The computational effort for getting the desired hologram amplitude of a Fourier hologram u_H is reasonable due to the powerful FFT algorithm. To use it the data must be discrete. In other words the CGH consists of cells, centered at $x_m = mD$, $y_n = nD$ as shown in **Fig. 3(D)**. The questions now are: “What is the proper cell size D ?” and “How many cells are needed?” The answers to those questions depend on the parameters of the image. If the zero order (see **Fig. 6**) should be at the distance Δx_0 , then the quasi-grating period D should obey:

$$\lambda f / D \geq \Delta x_0 \quad (13)$$

And if the size of an image pixel should be as fine as δx_0 the size Δx_H of the hologram should obey:

$$\lambda f / \Delta x_H \leq \delta x_0 \quad (14)$$

This condition is plausible because the finite size Δx_H of the hologram acts like a resolution-limiting aperture. The combination of Eqs. (13) and (14) yields

$$\Delta x_0 / \delta x_0 \leq \Delta x_H / D \quad (15)$$

This means that the number of image pixels $\Delta x_0 / \delta x_0$ is bounded by the number of CGH cells $\Delta x_H / D$. The generalization to two dimensions is straightforward. Again, Fourier CGH's benefit from the fact that the light propagation from the virtual object to the hologram plane can be described by a Fourier transformation. Hence, the FFT can be used.

And now we move to the synthesis of a Fresnel hologram. A Fresnel hologram is recorded at a finite distance z from the object. Hence, we have to simulate the wave propagation in free space from the object (at $z=0$) to the hologram at a distance z . This free space propagation can be described (in the paraxial approximation) as a convolution of object and quasi-spherical wave:

$$u_0(x) \rightarrow \int_{-\infty}^{\infty} u_0(x') \exp \left[\frac{-i\pi(x-x')^2}{\lambda z} \right] dx' = u(x, z) \quad (16)$$

In the Fourier domain the corresponding operation is:

$$\tilde{u}_0(\mu) \rightarrow \tilde{u}_0(\mu) \exp[-i\pi\lambda z \mu^2] = \tilde{u}(\mu, z) \quad (17)$$

The indirect execution of Eq. (16), based on Eq. (17), is easy: a Fourier transformation of the object converts $u_0(x)$ into $\tilde{u}_0(\mu)$, which is then multiplied by the quadratic phase factor, yielding $\tilde{u}(\mu, z)$. An inverse Fourier transformation produces $u(x, z)$. The two Fourier transformations consist typically of $4N \log N$ multiply/add operations. N is the number of pixels: image size Δx_0 , divided by the pixel size δx_0 . In the case of a rectangular object $u_0(x, y)$, the number of pixels is $\Delta x_0 \Delta y_0 / \delta x_0 \delta y_0$.

And now to the algorithmic effort for computing the convolution. The $u_0(x)$ has N pixels. And, the lateral size of the exponential is essentially $M = z / \delta z$, where δz means "depth of focus". The M describes the lateral size of the diffraction cone, measured in pixelsize units δx_0 . The convolution involves MN multiply/adds. A comparison of these two algorithms is dictated by the ratio

$$M / 4 \log N \quad (18)$$

The direct convolution is preferable if the distance z is fairly small. The integration limits of Eq. (16) depend on the oscillating phase $\pi x^2 / \lambda z$. Effectively, nothing is contributed to this integral, when this phase varies by more than 2π over the length, δx_0 , of an object pixel. Note that the pixel size of $u(x, z)$ does not change with the distance z , because the bandwidth $\Delta \mu(z) = \Delta \mu_0$ is z -independent. This is so because the power spectrum is z -invariant:

$$|\tilde{u}(\mu, z)|^2 = |\tilde{u}_0(\mu)|^2 \quad (19)$$

It is easy to proceed from here to the CGH synthesis of a 3-D object. One starts with an object detail $u_1(x)$ at z_1 , for example, a house as in Fig. 7. Then, one propagates to the next object detail at z_2 . This second detail may act as a multiplication (transmission) and as addition (source elements). This process is repeated as the wave moves towards the plane of the hologram. There, a reference wave has to be added just as in the Fourier case.

A very powerful algorithm for the CGH synthesis is iterative Fourier transform algorithm (IFTA) (Lesem *et al.*, 1969; Gerchberg and Saxton, 1972; Fienup, 1981). Suppose, we want to design a Fourier CGH, which looks like a person and whose reconstructed image shows the signature of that person (Fig. 8(A) and (B)). In other words, the two intensities

$$|u_H(x, y)|^2 \text{ and } |\tilde{u}_H(\mu, (v))|^2 \quad (20)$$

are determined. But, the phase distributions of u_H and $\tilde{u}_H$ are still at our disposal. The IFTA algorithm will yield those phases in many cases at the expense of ten, hundred or more Fourier transformations.

The IFTA algorithm is sometimes called "Fourier Ping-Ping" because it bounces back and forth between the (x, y) and the (μ, v) domain. The amplitudes $|u_H|$ and $|u_0|$ are enforced at every opportunity. But the associated phases are free to vary and to converge, eventually. If so, u_H and u_0 are completely known. The IFTA will not always converge, for example not if

$$|u_H(x, y)| = \delta(x, y) \text{ and } |\tilde{u}_H(\mu, (v))| = \delta(\mu, (v)) \quad (21)$$

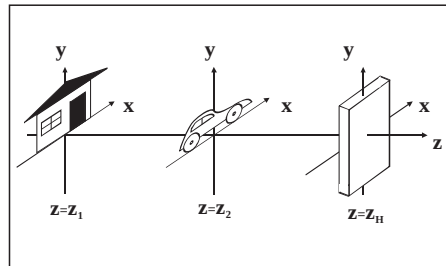


Fig. 7 Synthesis of a hologram for a three-dimensional object.

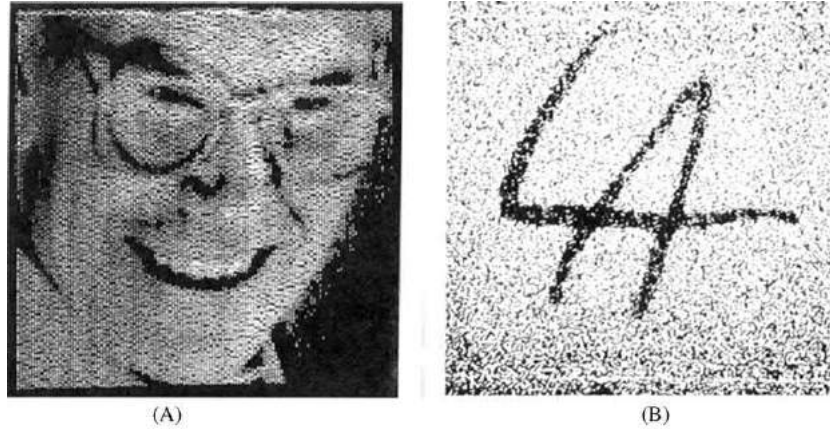


Fig. 8 IFTA-CGH: (A) The Fourier CGH. (B) The optical reconstruction thereof. Courtesy of Bartelt, H.O.

But the chances are fairly good if both amplitudes are very often close to one and seldom near zero.

Finally, a few words about image holograms. An image plane CGH is trivial from an algorithmic point of view. The simulated propagation from object plane to image plane is usually an identity, apart from the finite bandwidth of the image forming system. Nevertheless, image plane holograms can be quite useful, for example for interferometric testing of a non-spherical mirror of an astronomical telescope (Hansler, 1968; McGovern and Wyant, 1971; Ichioka and Lohmann, 1972). The CGH is laid out to provide a wavefront that is suitable for an interferometric “null test”. The CGH serves as a “synthetic prototype”.

A fourth kind of CGH, to be named near-field CGH may emerge, perhaps. It operates in the near-field. And, it employs evanescent waves as they occur if object details are comparable in size with, or even smaller than, the wavelength. Conceivable topics are “super resolution” and “sub-lambda lithography”. Fundamentals and history of classical evanescent holography are reported in Bryngdahl (1973).

On Some Hardware Aspects

And now a few comments about the hardware aspects of the CGH’s. Remembering the CGH cells structure in Fig. 3(D) we saw the amplitude encoded as relative width W (Eq. (6)). Instead of the width one may use the relative height $H \leq 1$ as the amplitude parameter (Lohmann and Sinzinger, 1995). More efficient, but more difficult to manufacture are saw tooth shaped hologram cells (Brown and Lohmann, 1969). The saw tooth prism may be approximated by a phase stair, which is possible with today’s lithographic technology. The corresponding theory is called phase quantization (Goodman and Silvestri, 1970; Dallas, 1971).

Some CGH Applications

Optical Filtering

Some applications had been mentioned already before, for example, 3-D Display (see Fig. 7). The data volume can be very high. Hence, shortcuts such as eliminating the vertical perspective, are called for. The largest CGH’s for visual light are used for TESTING astronomical telescopes by means of interferometry. The IFTA algorithm has been used, among other things, for designing a CGH which acts as an optimal diffuser. Speckles and light efficiency are the critical issues in diffuser design. Another project that also benefits from IFTA is beam shaping. That is a broad topic with aims such as homogenizing the output of laser, beam structuring for welding (Dresel *et al.*, 1996), cutting, and scanning.

A CGH can also be used as a spatial filter for image processing. For example, an input image may be Hilbert-transformed in the setup shown in Fig. 9(A) (Hauk and Lohmann, 1958). The filter (Fig. 9(B)) is a grating, of which one sideband is shifted by one-half of the grating period. Such a geometric shift generates the π -phase shift that is needed for the Hilbert transformation. The same setup can be used also for implementing Zernike’s phase contrast (Lohmann and Paris, 1968). Now the grating is distorted such that the zero-frequency region at the optical axis is reduced in amplitude by (typically) a factor of 0.2. In addition, the phase is shifted by $\pi/2$ due to a fringe shift of a quarter period (Fig. 10(A)). In the image plane one can see an ordinary image on-axis, and two phase-contrast images in the plus-minus first diffraction orders. One of them has positive phase contrast, the other one has negative phase contrast (Fig. 10(B)). Another simple spatial filter (Fig. 11(A)) produces a derivative of the input $u(x,y)$: $\frac{\partial u(x,y)}{\partial x}$ (Fig. 11(B)). Eqs. (22)–(24) explain this differential filtering experiment.

$$u(x, y) \rightarrow \partial u(x, y) / \partial x = v(x, y) \quad (22)$$

$$\tilde{u}(\mu, (v)) \rightarrow 2\pi i \mu \cdot \tilde{u}(\mu, (v)) = \tilde{v}(\mu, (v)) \quad (23)$$

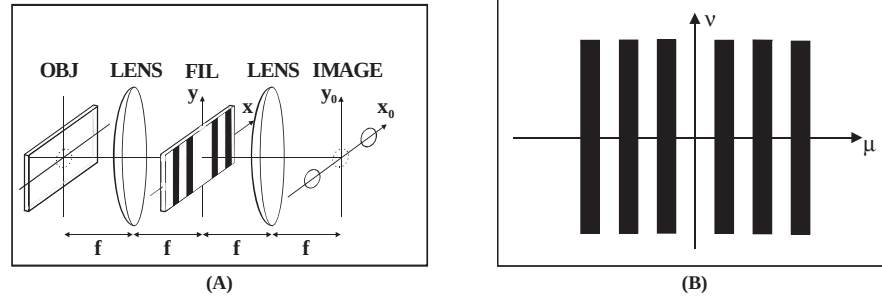


Fig. 9 (A) Hilbert setup. (B) Hilbert filter.

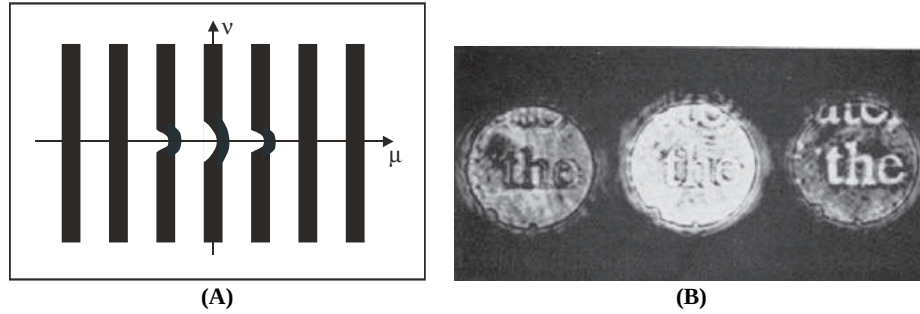


Fig. 10 (A) Phase contrast filter. (B) Phase contrast output.

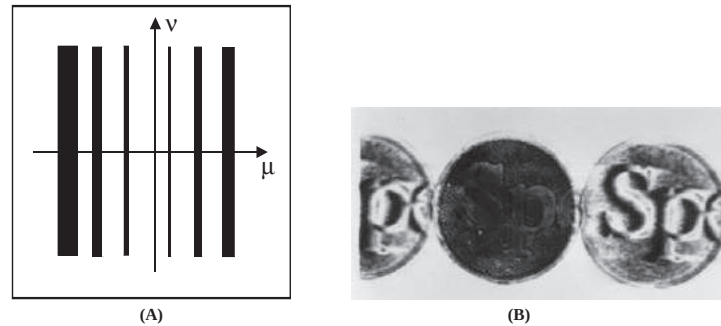


Fig. 11 (A) Differentiation filter. (B) Differentiation output.

$$\tilde{p}(\mu, (v)) = 2\pi|\mu| \cdot \begin{cases} +i(\mu > 0) \\ -i(\mu < 0) \end{cases} \quad (24)$$

The phase portion of the filter (**Fig. 11(A)**) is the same as for the Hilbert filter. But, the filter amplitude now enhances the higher frequencies. The object used for getting the results of **Figs. 10(B)** and **11(B)** was a phase-only object: a bleached photograph.

A very famous spatial filtering experiment is “matched filtering” ([Vander Lugt, 1964](#); [Kozma and Kelly, 1965](#)). It can be used for pattern recognition. The original matched filters were classical Fourier holograms of the target pattern. Computer hologram's can do the same job ([Brown and Lohmann, 1966](#)), even with spatially incoherent objects ([Lohmann and Werlich, 1971](#)). Work on spectrally incoherent spatial filtering was initiated by [Katyl \(1972\)](#).

Fourier holograms are attractive also for data storage due to their robustness against local defects. A local error in the Fourier domain spreads out all over the image domain. The computed hologram has an advantage over classical holograms because of the freedom to incorporate any error detection and correction codes for reducing the bit error rate even further.

One of the most recent advances is in matched filtering with totally incoherent light ([Andrés et al., 1999](#); [Pe'er et al., 1999](#)). This step allows moving spatial filtering out of the well-guarded laboratory into hostile outdoor environments.

Optical Vortices

The mathematical key to understanding CGH generation of optical vortices ([Rozas et al., 1997](#)) is the circular harmonic. It is a function of two variables where θ is the angular member of the polar coordinate pair and n is the order of the harmonic. The

circular harmonic is

$$\exp(in\theta) \quad (25)$$

An optical wave at a plane with this complex amplitude $\exp(in\theta)$ is called an optical vortex with a topological charge of n . In order to design a CGH that converts a plane wave to an optical vortex the circular harmonic decomposition is used. Extending Eq. (2) to polar coordinates, the transmittance of the CGH is

$$T_H(r, \theta) = c_0 + c_1 \exp(in\theta) u_R^* + \dots \quad (26)$$

For a unit-amplitude normally-incident wave, the exiting optical vortex is numerically equal. The corresponding CGH mask is shown in Fig. 12.

These waves can be used, as propeller beams (Zhang *et al.*, 2010) that apply torque to components (cranks) of micro-mechanical devices. They can be used as optical traps (Ng *et al.*, 2010), for example, to trap Bose-Einstein condensates. They can also be employed in solar coronagraphs (Foo *et al.*, 2005).

Polarization CGH

Polarized light can be decomposed into two orthogonal components. For example, if the light is propagating in the z -direction, light in a plane can be decomposed into linearly polarized components: x -polarized and y -polarized. An arbitrarily polarized field can be composed by superimosing a field of x -polarized and a field of y -polarized light. This process is known as polarization synthesis (Lohmann, 1967; Noble *et al.*, 2010). An idealized synthesis process is illustrated in (Fig. 13) where two Fourier CGHs are inserted into a Mach-Zehnder interferometer. It combines the waves exiting both holograms into the arbitrarily-polarized reconstructed image.

Determining the strength of each component is known as polarization analysis. The generation steps for calculating the encoded waves of the CGHs are illustrated in the flow diagram of Fig. 14.

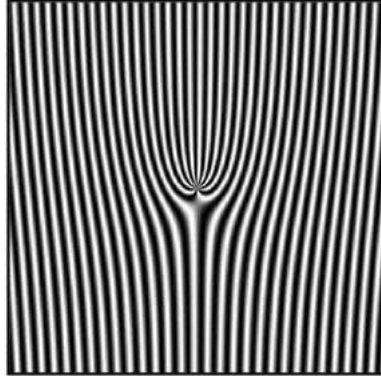


Fig. 12 Ninth-order vortex CGH.

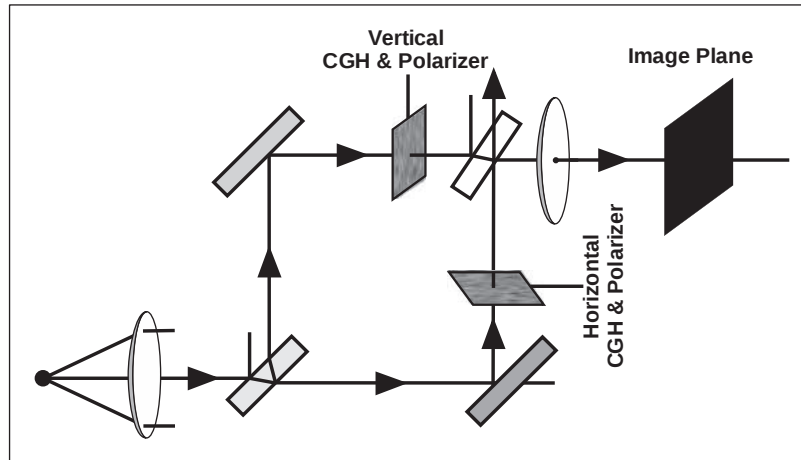


Fig. 13 Polarized reconstruction.

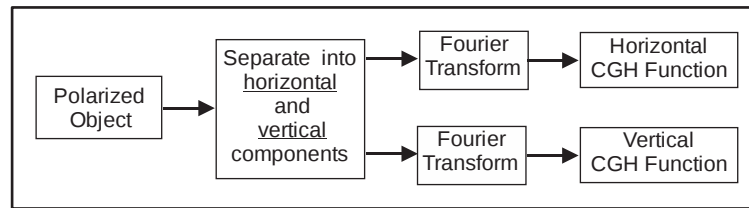


Fig. 14 Flow diagram for straight-forward designing of a polarization CGH.

Terminology

Several terms mean more or less the same as CGH: synthetic hologram, digital hologram, holographic optical element (HOE), diffractive optical element (DOE) or digital optical element.

CGH's have been made also for other waves, such as acoustical waves, ocean waves, and microwaves. Digital antenna steering can be understood as a case of computer holography.

See also: Module: Digital Holography

References

- Andrés, P., *et al.*, 1999. *Opt. Lett.* 24, 1331.
 Brown, B.R., Lohmann, A.W., 1966. *Appl. Opt.* 5, 967.
 Brown, B.R., Lohmann, A.W., 1969. *IBM J. R&D* 13, 160.
 Bryngdahl, O., 1973. *Prog. Opt.* 11, 168.
 Dallas, W.J., 1971. *Appl. Opt.* 10, 673.
 Dresel, T., Beyerlein, M., Schwider, J., 1996. *Appl. Opt.* 35, 4615.
 Fienup, J.R., 1981. *Proc. SPIE* 373, 147.
 Foo, G., Palacios, D.M., Swartzlander, G.A., 2005. *Opt. Lett.* 30 (24), 3308.
 Gerchberg, R.W., Saxton, W.O., 1972. *Optik* 35, 237.
 Goodman, J.W., Silvestri, A.M., 1970. *IBM J. R&D* 14, 478.
 Hansler, R.L., 1968. *Appl. Opt.* 7, 1863.
 Hauk, D., Lohmann, A.W., 1958. *Optik* 15, 275.
 Ichioka, Y., Lohmann, A.W., 1972. *Appl. Opt.* 11, 2597.
 Katyl, R.H., 1972. *Appl. Opt.* 11, 1255.
 Kozma, A., Kelly, D.L., 1965. *Appl. Opt.* 4, 387.
 Lesem, L.B., Hirsch, P.M., Jordan, J.A., 1969. *IBM J. R&D* 13, 150.
 Lohmann, A.W., 1967. Reconstruction of vectorial wavefronts. *Appl. Opt.* 4, 1965.
 Lohmann, A.W., Paris, D.P., 1967. *Appl. Opt.* 6, 1746.
 Lohmann, A.W., Paris, D.P., 1968. *Appl. Opt.* 7, 651.
 Lohmann, A.W., Sinzinger, S., 1995. *Appl. Opt.* 34, 3172.
 Lohmann, A.W., Werlich, H.W., 1971. *Appl. Opt.* 10, 670.
 McGovern, A.J., Wyant, J.C., 1971. *Appl. Opt.* 10, 619.
 Ng, J., Lin, Z., Chan, C.T., 2010. *PhysRevLett.* 104, 103601.
 Noble, H., Ford, E., Dallas, W., *et al.*, 2010. *Opt. Lett.* 35 (20), 3423.
 Pe'er, A., Wang, D., Lohmann, A.W., Friesem, A.A., 1999. *Opt. Lett.* 24, 1469. (2000; vol. 25, p. 776).
 Rozas, D., Law, C.T., Swartzlander Jr., G.A., 1997. *J. Opt. Soc. Am. B* 14, 3054.
 Vander Lugt, A.B., 1964. *IEEE IT* 10, 139.
 Zhang, P., Huang, S., Hu, Y., Hernandez, D., Chen, Z., 2010. *Opt. Lett.* 35, 3129.

Further Reading

- Bryngdahl, O., Wyrowski, F., 1990. *Prog. Opt.* 28, 1–86.
 Dallas, W.J., 1973. *Appl. Opt.* 12, 1179.
 Dallas, W.J., 1980. *Top. Appl. Phys.*, 41. Heidelberg: Springer, pp. 291–366.
 Dallas, W.J., 1999–2003, University of Arizona, Tucson, AZ. Available at: <http://www.radiology.arizona.edu/dallas/CGH.htm>.
 Fienup, J.R., 1982. *Appl. Opt.* 21, 2758.
 Lee, W.H., 1978. *Prog. Opt.* 16, 119–232.
 Sinzinger, S., Jahns, J., 1999. *Micro Optics*. New York: Wiley.
 Wyrowski, F., Bryngdahl, O., 1991. *Rep. Prog. Phys.* 1481–1571.
 Yaroslavskii, L.P., Merzlyakov, N.S., 1980. *Methods in Digital Holography*. New York: Plenum Press.

Module: Digital Holography

Wolfgang Osten, University of Stuttgart, Stuttgart, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

One of the most fascinating properties of light is the fact that it encounters us both as a beam, as a wave, and as a particle. In the context of this article, however, we will only deal with the wave character of light and its sinusoidal nature. The accompanying ability of waves to interfere makes it possible to store spatial information on a two-dimensional target and to reconstruct the complete light field again. This is known as holography (Gabor, 1949). Even more, the superposition of waves and the precise knowledge of the wavelength guide us to a series of interesting methods for the high-resolution measurement of various physical quantities, such as three-dimensional distances, displacements, strains and refractive index distributions. The primary physical quantity of the wave field, which represents its three-dimensional properties, is the phase. Unfortunately, due to the high frequency of light, all media for the storage of light, whether photo-chemical (films, photoplates) or photo-electronic (CCD and CMOS detectors) can only register intensities. But the ability to interfere allows us to apply a trick: the interferometric principle. Simply by means of interferometry, phase changes are converted into measurable intensity changes. This is done in an interferometer where a primary wave front is divided into at least two wave trains, often called as object and reference wave, respectively, which are superimposed again after traveling along common or various paths. Finally, interferometry and holography enabled numerous outstanding applications such as the 3D-imaging of natural objects and scenes (Leith and Upatnieks, 1961; Denisjuk, 1962), the fabrication of diffractive optical elements (Lee, 1978), and the high-resolution inspection of static and dynamic objects having both precisely machined as well as rough surfaces (Steel, 1983; Powell and Stetson, 1965).

As already mentioned, the basic principle of holography consists in the transformation of phase changes into recordable intensity changes. Because of the high spatial frequency of these intensity fluctuations the registration of a hologram requires a light sensitive medium with adequate spatial resolution. Therefore, special photographic emulsions have dominated holographic technologies for a long period. However, the recording of holograms on electronic sensors and their numerical reconstruction is almost as old as holography itself. The first ideas and implementations were already conceived in the 1960s and 1970s (Goodman and Lawrence, 1967; Huang, 1971; Kronrod *et al.*, 1972; Demetrakopoulos and Mitra, 1974). But the most crucial step towards a practicable technology could only be made as digital cameras with reasonable space-bandwidth product and powerful processor technology became widely accessible. At the beginning of the 1990s these technical prerequisites were available and it was only a matter of time before they would be applied to the discrete recording, digital storage and numerical reconstruction of optical wave fronts, which is called digital holography. The article of Schnars and Jüptner from 1994 (Schnars and Jüptner, 1994) can be considered as a kind of initialization for a continuously growing number of implementations and applications. Meanwhile, digital holography has grown into a separate, distinct category in modern optics and has obvious implications in many fields such as microscopy (Yu *et al.*, 2014), 3D imaging (Frauel *et al.*, 2006), metrology (Kreis, 2005), display technology (Gang, 2013), material processing (Hasegawa *et al.*, 2006), data storage (Coufal *et al.*, 2000), and information processing (Yaroslavsky, 2004). A further remarkable extension of the application range of digital holography could be achieved by the application of sophisticated spatial light modulator (SLM) technology (Sutkowski and Kujawinska, 2000; Kohler *et al.*, 2006; Haist and Osten, 2015a,b), complementing the digital recording process with a matching digital optical reconstruction process. Since the late 1990s such devices have been offered by different companies (HOLOEYE Photonic AG; Lazarev *et al.*, 2012; HAMAMTSU Photonics K.K.) with the result that a bunch of new applications such as holographic micro-manipulation (Zwick *et al.*, 2010; Reichert *et al.*, 1999; Daneshpanah *et al.*, 2010), aberration compensation (Ferraro *et al.*, 2003; Liesener *et al.*, 2004), computational microscopy (Maurer *et al.*, 2010; Haist *et al.*, 2014; Marquet and Depeursinge, 2014; Kim, 2011), holographic displays (Onural *et al.*, 2011; Lee and Kim, 2012), comparative digital holography (Osten *et al.*, 2002; Baumbach *et al.*, 2006), and remote holographic laboratories (Osten *et al.*, 2013) could be implemented. Further recent application fields are described in (Osten *et al.*, 2014).

Meanwhile, modern digital holography is more than 25 years old. But it is still an equally young and dynamic discipline. Some emerging fields where digital holography in combination with modern photonic technologies shows their great potential for new applications in technical and living sciences are:

- the implementation of sophisticated illumination, imaging and reconstruction principles for the enhancement of the resolution of holographically stored images and for the reduction of the demands on the space-bandwidth product of holographic sensors,
- the combination of digital-holographic microscopy with SLM-based holographic 3D-micromanipulation for the in vivo investigation of living cells and tissues,
- the evaluation of the spatial coherence function measured by self-referenced interferometry with the objective to reduce the degree of coherence of the light needed for holographic storage and to derive spatial and spectral information about the objects under test,
- the application of modern light sources such as supercontinuum lasers and frequency combs for the high-precision depth mapping and 3D-reconstruction of objects and scenes,

- the iterative evaluation of the propagating wavefront for avoiding sensitive interferometric detection principles,
- the evaluation of light fields transmitted or reflected from strongly scattering or hidden objects, and
- the enabling of holographic procedures and setups for remote access and comparative inspection technologies.

Direct Phase Reconstruction by Digital Holography

In digital speckle pattern interferometry DSPI (Butters and Leendertz, 1971) the object is focused onto the target of an electronic sensor. Thus an image plane hologram is formed as result of the interference with an inline reference wave. In contrast with DSPI, a digital hologram can be recorded without imaging the object onto a target. The target records the superposition of the reference and the object wave in the near-field region – a so-called Fresnel hologram (Schnars and Jüptner, 1994). The basic optical setup in digital holography for recording holograms is the same as in conventional holography, Fig. 1(a). A laser beam is divided into 2 coherent beams. One beam illuminates the object and generates the object beam $u(x,y)$. The other enters the target directly and forms the reference wave $r(x,y)$. On this basis very compact solutions are possible. Fig. 1(b) shows a holographic camera where the interferometer, containing a beam splitter and some wave front shaping components, is mounted in front of the camera target. The object, drafted as a ball, is illuminated from four directions. This scheme has some advantages for the measurement of the shape and displacement of the object under test (see Section “The Fresnel approximation”).

For the description of the principle of digital Fresnel holography we use a simplified version of the optical setup, Fig. 2(a). The object is modeled by a plane rough surface that is located in the (x,y) -plane and illuminated by laser light. The scattered wave field forms the object wave $u(x,y)$. The target of an electronic sensor (e.g. a CCD or a CMOS) used for recording the hologram is located in the (ξ,η) -plane at the distance d from the object. Following the basic principles of holography (Gabor, 1949; Hariharan, 1984) the hologram $h(\xi,\eta)$ originates from the interference of the object wave $u(\xi,\eta)$ and the reference wave $r(\xi,\eta)$ in the (ξ,η) -plane:

$$h(\xi,\eta) = |u(\xi,\eta) + r(\xi,\eta)|^2 = r \cdot r^* + r \cdot u^* + u \cdot r^* + u \cdot u^* \quad (1)$$

The transformation of the intensity distribution $h(\xi,\eta)$ into a discrete and digital gray value distribution $T(m,n)$ that is stored in the image memory of the computer is considered by a characteristic function t of the sensor. This function is in general only approximately linear:

$$T(m,n) = t[h(\xi,\eta)] \quad (2)$$

Because the sensor has a limited spatial resolution the spatial frequencies of the interference fringes in the hologram plane – the so-called micro-interferences – have to be considered. The fringe spacing g in, for example, x -direction and the corresponding spatial frequency f_x , respectively, is determined by the angle β between the object and the reference wave, Fig. 2(b):

$$g = \frac{1}{f_x} = \frac{\lambda}{2\sin(\beta/2)} \quad (3)$$

with λ as the wavelength. If we assume that the discrete sensor has a pixel pitch (distance between two adjacent pixels) $\Delta\xi$ the sampling theorem requires at least 2 pixels per fringe for a correct reconstruction of the periodic function:

$$2\Delta\xi < \frac{1}{f_x} \quad (4)$$

Consequently, we obtain for small angles β :

$$\beta < \frac{\lambda}{2\Delta\xi} \quad (5)$$

Modern high-resolution CCD or CMOS chips have a pitch $\Delta\xi$ of about $4\mu\text{m}$. In that case a maximum angle between the reference and the object wave of only 4° is acceptable. The practical consequence of the restricted angle resolution in digital holography is a limitation of the effective object size that can be stored holographically by an electronic sensor. However, this is only a technical handicap. Larger objects can be placed in a sufficient distance from the hologram, or reduced optically by imaging with a negative lens.

The reconstruction is done by illumination the hologram with a so-called reconstruction wave $c(\xi,\eta)$:

$$u'(x',y') = t[h(\xi,\eta)] \cdot c(\xi,\eta) \quad (6)$$

The transmission function of the hologram $t[h(\xi,\eta)]$ (see Eq. (2)) diffracts the wave $c(\xi,\eta)$ in such a way, that images of the object wave are reconstructed. In general, four terms are reconstructed if the wave $u'(x',y')$ propagates in space. In case of a linear characteristic function $t(h) = \alpha h + t_0$ we can write:

$$u' = T \cdot c = t(h) \cdot c \quad (7)$$

and

$$u' = \alpha [cu^2 + cr^2 + cur^* + cru^*] + ct_0 \quad (8)$$

with two relevant image terms $[cur^*]$ and $[cru^*]$, containing the object wave u and its conjugated version u^* , respectively. The appearance of the image terms depends on the concrete shape of the reconstruction wave. In general, the original reference wave

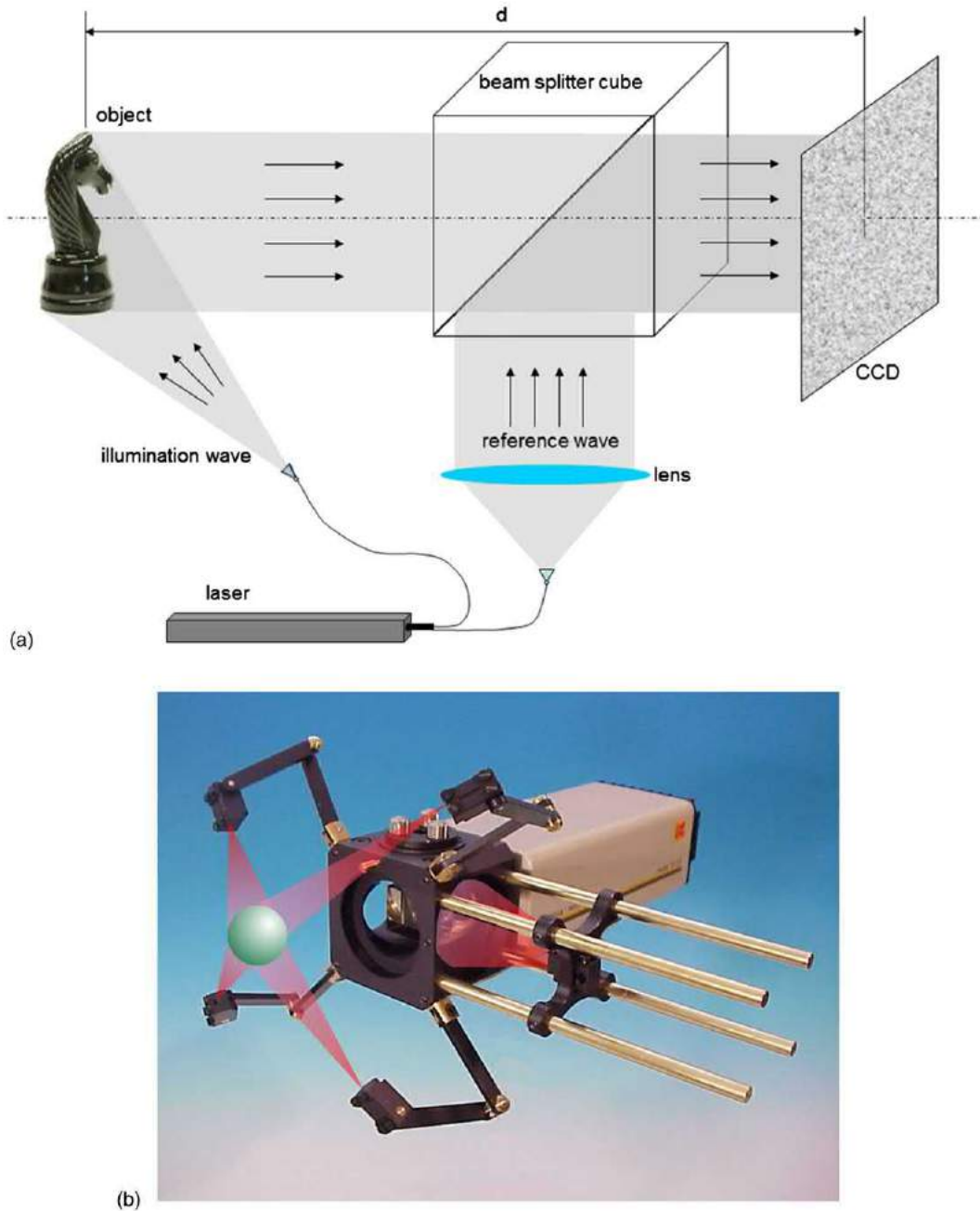


Fig. 1 Setup for recording digital holograms. (a) Schematic setup for recording a digital hologram onto a CCD target, (b) Implementation of a holographic camera by mounting a compact holographic interferometer in front of a CCD-target (four illumination directions are used). Osten, W., Seebacher, S., Jüptner, W. 2001. The application of digital holography for the inspection of micro-components. Proc. SPIE 4400, 1–15.

$c=r$ or its conjugated version $c=r^*$ is applied. In case of the conjugated reference wave a direct or real image will be reconstructed due to a converging image wave that can be imaged on a screen at the place of the original object.

However, in digital holography the reconstruction of the object wave in the image plane $u'(x',y')$ is done by numerical reproduction of the physical propagation process as shown in Fig. 3(a). The reconstruction wave, with a well-known shape equal to the reference wave $r(\xi,\eta)$, propagates through the hologram $h(\xi,\eta)$. Following the Huygens' principle each point $P(\xi,\eta)$ on the hologram acts as the origin of a spherical elementary wave. The intensity of these elementary waves is modulated by the transparency $h(\xi,\eta)$. In a given distance $d'=d$ from the hologram, a sharp real image of the object can be reconstructed as the superposition of all elementary waves. For the reconstruction of a virtual image, $d'=-d$ is used.

Consequently, the calculation of the wave field $u'(x',y')$ in the image plane starts with the pointwise multiplication of the stored and transformed intensity values $t[h(\xi,\eta)]$ with a numerical model of the reference wave $r(\xi,\eta)$. For a normally incident and

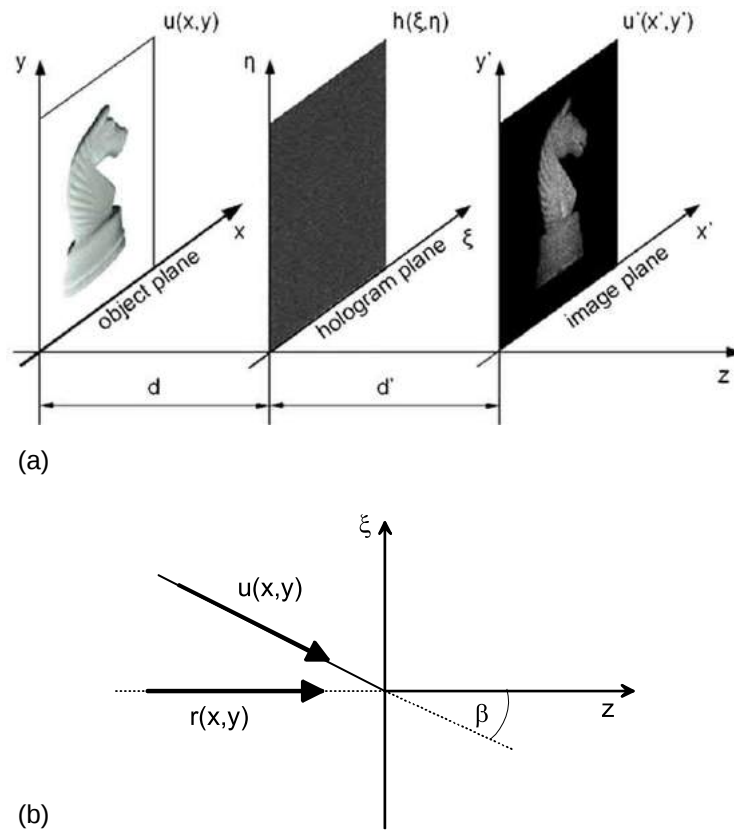


Fig. 2 Geometry for recording and reconstruction of digital holograms. (a) Schematic setup for Fresnel holography, (b) Interference between the object and reference wave in the hologram plane.

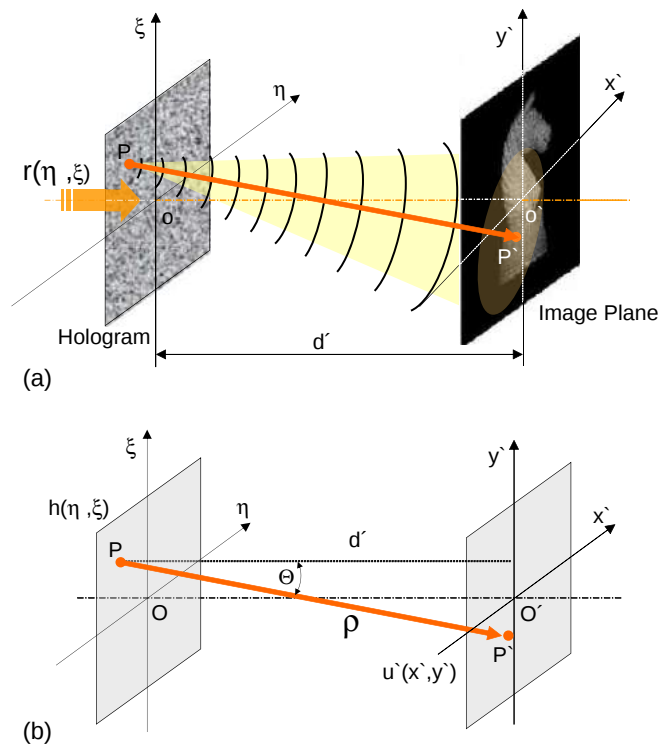


Fig. 3 Reconstruction of digital holograms. (a) Principle of wave front reconstruction, (b) light propagation by diffraction (Huygens-Fresnel principle).

monochromatic wave with unit amplitude, the reference wave can be modeled by $r(\xi, \eta)$. After the multiplication, the resulting field is propagated in free space with reference to the laws of e.g. nearfield diffraction. In the distance d' the diffracted field $u'(x', y')$ can be found by solving the well-known Rayleigh–Sommerfeld diffraction formula, which is also known as the Huygens–Fresnel principle (Goodman, 1996):

$$u'(x', y') = \frac{1}{i\lambda} \iint_{-\infty \dots \infty} t[h(\xi, \eta)] \cdot r(\xi, \eta) \frac{\exp(ik\rho)}{\rho} \cos\theta d\xi d\eta \quad (9)$$

with

$$\rho(\xi - x', \eta - y') = \sqrt{d'^2 + (\xi - x')^2 + (\eta - y')^2} \quad (10)$$

as the distance between a point $O'(x', y', z=d')$ in the image plane and a point $P(\xi, \eta, z=0)$ in the hologram plane and

$$k = \frac{2\pi}{\lambda} \quad (11)$$

as the wave number. The obliquity factor $\cos\theta$ represents the cosine of the angle between the outward normal and the vector joining P to O' . This term is given exactly by

$$\cos\theta = \frac{d'}{\rho} \quad (12)$$

and therefore Eq. (9) can be rewritten

$$u'(x', y') = \frac{d'}{i\lambda} \iint_{-\infty \dots \infty} t[h(\xi, \eta)] \cdot r(\xi, \eta) \frac{\exp(ik\rho)}{\rho^2} d\xi d\eta \quad (13)$$

The numerical reconstruction provides the complex amplitude of the wave front. Consequently, the phase distribution $\phi(x', y')$ and the intensity $I(x', y')$ can be calculated directly from the reconstructed complex function $u'(x', y')$:

$$\phi(x', y') = \arctan \frac{\text{Im}[u'(x', y')]}{\text{Re}[u'(x', y')]} \quad [-\pi, \pi] \quad (14)$$

$$I(x', y') = u'(x', y') \cdot u'^*(x', y') \quad (15)$$

The direct approach to the phase yields several advantages for imaging and metrology applications that are discussed later.

Reconstruction Principles in Digital Holography

The theoretical background for the reconstruction of digital holograms is given by the scalar diffraction theory. Based on this, different numerical reconstruction principles are applied: the Fresnel approximation (Schnars, 1994), the convolution approach (Demetrakopoulos and Mittra, 1974), the angular spectrum method (Goodman, 1996), the lens-less Fourier approach (Takeda *et al.*, 1996; Wagner *et al.*, 1999), the phase shifting approach (Yamaguchi and Zhang, 1979), and the phase-retrieval approach (Zhang *et al.*, 2003). In this section the main techniques are briefly described.

The Fresnel Approximation

If the distance d between the object and hologram plane, and equivalently $d' = d$ between the hologram and image plane, is large compared to $(\xi - x')$ and $(\eta - y')$, so that the Fresnel approximation is fulfilled, then the distance ρ^2 in the denominator of Eq. (13) can be replaced by d'^2 and the parameter ρ in the numerator can be approximated by a binomial expansion for the square root in Eq. (10) where only the first two terms are considered (Goodman, 1996):

$$\rho \approx d' \left[1 + \frac{(\xi - x')^2}{2d'^2} + \frac{(\eta - y')^2}{2d'^2} \right] \quad (16)$$

The resulting expression for the field at (x', y') becomes

$$u'(x', y') = \frac{\exp(ikd')}{i\lambda d'} \iint_{-\infty \dots \infty} t[h(\xi, \eta)] \cdot r(\xi, \eta) \exp \left[\frac{ik}{2d'} \{ (\xi - x')^2 + (\eta - y')^2 \} \right] d\xi d\eta \quad (17)$$

Eq. (17) is a convolution integral that can be expressed as

$$u'(x', y') = \iint_{-\infty \dots \infty} t[h(\xi, \eta)] \cdot r(\xi, \eta) H(\xi - x', \eta - y') d\xi d\eta \quad (18)$$

with the convolution kernel

$$H(x', y') = \frac{\exp(ikd')}{i\lambda d} \exp \left[\frac{ik}{2d} (x'^2 + y'^2) \right] \quad (19)$$

Another notation of Eq. (17) is found if the term $\exp\left[\frac{ik}{2d}(x'^2 + y'^2)\right]$ is taken out of the integral:

$$u'(x', y') = \frac{\exp(ikd')}{i\lambda d'} e^{i\frac{k}{2d'}(x'^2 + y'^2)} \iint_{-\infty \dots \infty} \left\{ t[h(\xi, \eta)] \cdot r(\xi, \eta) \cdot e^{i\frac{k}{2d'}(\xi^2 + \eta^2)} \right\} \exp\left[-i\frac{2\pi}{\lambda d'}(\xi x' + \eta y')\right] d\xi d\eta \quad (20)$$

or

$$u'(x', y') = \frac{\exp(ikd')}{i\lambda d'} e^{i\frac{k}{2d'}(x'^2 + y'^2)} \text{FT}_{\lambda d'} \left\{ t[h(\xi, \eta)] \cdot r(\xi, \eta) \cdot e^{i\frac{k}{2d'}(\xi^2 + \eta^2)} \right\} \quad (21)$$

where $\text{FT}_{\lambda d'}$ indicates the 2D-Fourier transform that has been modified by a factor $1/(\lambda d')$. Eq. (21) makes clear that the diffracted wave front u' results from the Fourier transform FT of the digitally stored hologram $t[h(\xi, \eta)]$ multiplied by the reference wave, $r(\xi, \eta)$, and a quadratic phase factor, the so-called the chirp function $\exp\left\{i\frac{k}{2d'}(\xi^2 + \eta^2)\right\}$. This Fourier transform is scaled by the constant factor $1/(i\lambda d')$ and a phase factor that is independent of the processed hologram.

The discrete sampling in digital holography with the discrete pixel coordinates (m, n) requires the transformation of the infinite continuous integral in Eq. (20) into a finite discrete sum. This results in the finite Fresnel transform:

$$u'(m, n) = \sum_{j=0}^{M-1} \sum_{l=0}^{N-1} t[h(j \cdot \Delta\xi, l \cdot \Delta\eta)] \cdot r(j \cdot \Delta\xi, l \cdot \Delta\eta) \exp\left[\frac{i\pi}{d'\lambda} (j^2 \Delta\xi^2 + l^2 \Delta\eta^2)\right] \times \exp\left\{-i2\pi\left(\frac{j \cdot m}{M} + \frac{l \cdot n}{N}\right)\right\} \quad (22)$$

where constants and pure phase factors preceding the sums have been omitted. The main parameters are the pixel number $M \times N$ and the pixel pitches $\Delta\xi$ and $\Delta\eta$ in the two directions which are defined by the used sensor chip.

The discrete phase distribution $\phi(m, n)$ of the wave front and the discrete intensity distribution $I(m, n)$ on the rectangular grid of $M \times N$ sample points can be calculated from the reconstructed complex function $u'(m, n)$ by using Eqs. (14) and (15), respectively. The wrapped phase distribution is computed directly without the need of additional temporal or spatial phase modulation like phase shifting or spatial heterodyning (Creath, 1993; Kujawinska, 1993). Fig. 4 shows the result of the numerical reconstruction of the intensity and the phase of a tin soldier using a setup of the kind shown in Fig. 1(a).

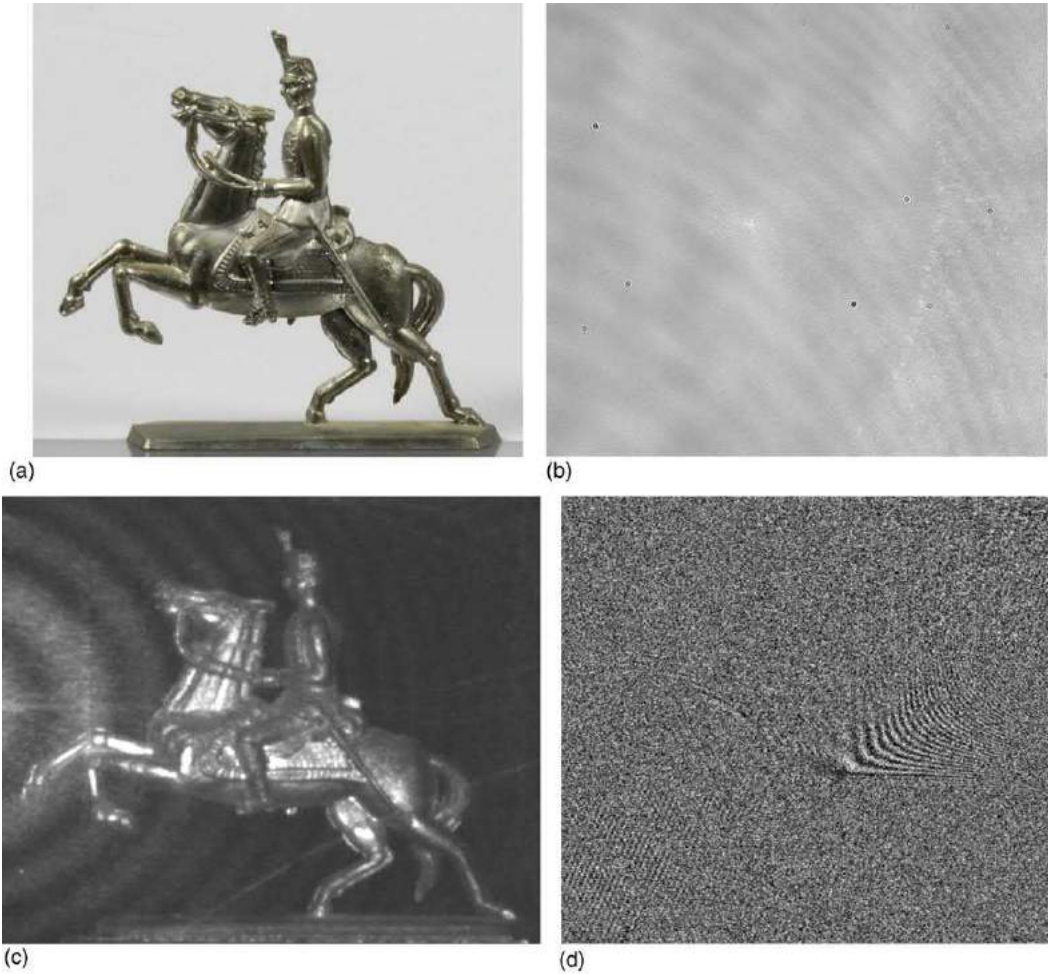


Fig. 4 (a) Tin soldier, (b) digital hologram, (c) Reconstructed Intensity $I(m, n)$, (d) Reconstructed phase distribution $\phi(m, n)$.

For interferometric applications with respect to displacement and shape measurement (Kreis, 2005; Osten and Ferraro, 2007) the phase difference $\delta(m, n)$ of the two reconstructed wave fields $u_1'(m, n)$ and $u_2'(m, n)$ corresponding to the two states of the object needs to be calculated. The algorithm can be reduced to the following equation:

$$\delta(m, n) = \phi_1(m, n) - \phi_2(m, n) \quad (23)$$

Fig. 5 shows the principle of displacement measurement by digital holography. Besides the investigation of surface displacements in the region of the wavelength, digital holography can be applied advantageously for the measurement of the shape of complex objects. In Fig. 6 a turbine blade was investigated with the 2-wavelength contouring method (Osten *et al.*, 1998). Both, the digitally reconstructed mod 2π -phase and its demodulated version after phase unwrapping are presented (Wagner *et al.*, 2000). Phase unwrapping is necessary in digital holography since the fringe-counting problem still remains. However, there are some efficient approaches to overcome the difficulties of that process (Wagner *et al.*, 2000; Osten and Kujawinska, 2000).

Following Eq. (22), the pixel sizes in the reconstructed image along the two directions are

$$\Delta m = \frac{d'\lambda}{M\Delta\xi} \text{ and } \Delta n = \frac{d'\lambda}{N\Delta\eta} \quad (24)$$

In addition to the real image, a blurred virtual image and a bright dc-term, the zero-order diffracted field, are reconstructed. The dc-term can effectively be eliminated by preprocessing the stored hologram (Kreis and Jüptner, 1997; Seebacher *et al.*, 1997), and the different terms can be separated by using the off-axis instead of the in-line scheme. However, the spatial separation between the object and the reference field requires a sufficient space-bandwidth product of the used CCD-chip, as discussed in the section on direct phase reconstruction above.

Numerical Reconstruction by the Convolution Approach

In Section "The Fresnel approximation" we have already mentioned that the connection between the image term $u'(x', y')$ and the product $t[h(\xi, \eta)] \cdot r(\xi, \eta)$ can be described by a linear space-invariant system. The diffraction formula (17) is a superposition integral with the impulse response

$$H(x' - \xi, y' - \eta) = \frac{\exp(ikd')}{i\lambda d'} \exp\left[\frac{ik}{2d'} \left\{(\xi - x')^2 + (\eta - y')^2\right\}\right] \quad (25)$$

A linear space-invariant system is characterized by a transfer function G that can be calculated as the Fourier-transform of the impulse response:

$$G(f_x, f_y) = \text{FT}\{H(\xi, \eta, x', y')\} \quad (26)$$

with the spatial frequencies f_x and f_y . Consequently, the convolution theorem can be applied, which states that the Fourier transform of the convolution $t[h(\xi, \eta)] \cdot r(\xi, \eta)$ with H is the product of the individual Fourier transforms $\text{FT}\{t[h(\xi, \eta)] \cdot r(\xi, \eta)\}$ and $\text{FT}\{H\}$. Thus $u'(x', y')$ can be calculated by the inverse Fourier transform of the product of the Fourier-transformed convolution partners:

$$u'(x', y') = \text{FT}^{-1}\{\text{FT}[t(h) \cdot r] \cdot \text{FT}[H]\} \quad (27)$$

$$u'(x', y') = [t(h) \cdot r] \otimes H \quad (28)$$

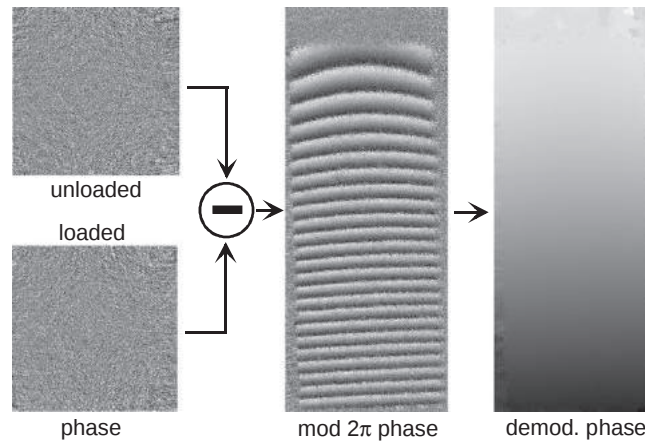


Fig. 5 The principle of digital holographic interferometry demonstrated on example of a loaded small beam. The interference phase is the result of the subtraction of the two phases which correspond to the digital holograms of the both object states.

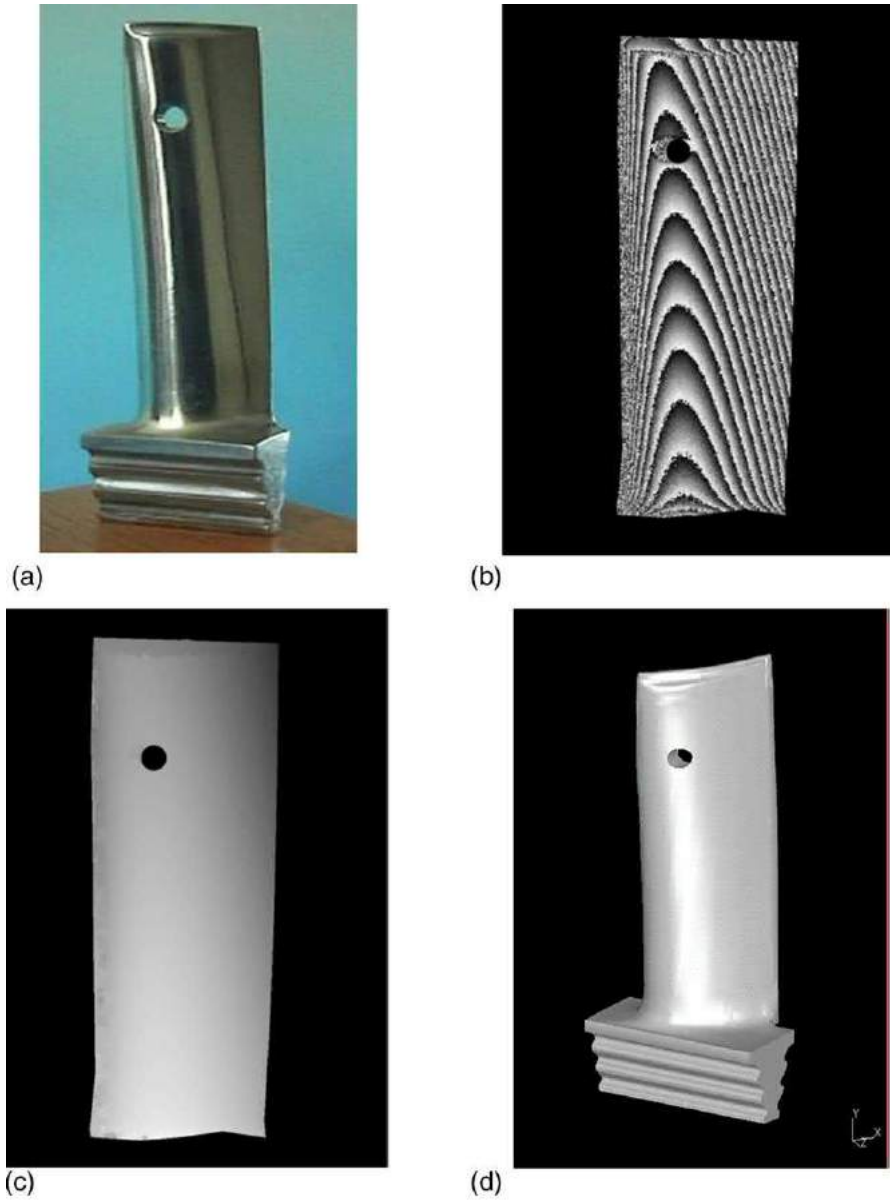


Fig. 6 Phase reconstruction by digital holographic interferometry on example of 2-wavelength contouring of a turbine blade. (a) image of the turbine blade, (b) mod-2 π phase distribution, (c) unwrapped phase distribution, (d) CAD reconstruction of the turbine blade.

with $\otimes$ as the convolution symbol. The computing effort is comparatively high: two complex multiplications and three Fourier transforms. The essential difference compared to the Fresnel approximation is the different pixel size in the reconstructed field (Kreis, 2005). The pixel size is constant independent from the reconstruction distance d' , wavelength λ and pixel number m :

$$\Delta x' = \Delta \xi \text{ and } \Delta y' = \Delta \eta \quad (29)$$

Numerical Reconstruction by the Lensless Fourier Approach

We have already mentioned that the limited spatial resolution of the sensor restricts the angular resolution of the digital hologram. In fact, the sampling theorem requires that the angle between the object beam and the reference beam at any point of the electronic sensor be limited in such a way that the microinterference spacing is larger than double the pixel size. In general, the angle between the reference beam and the object beam varies over the sensor's surface, and so does the maximum spatial frequency. Thus for most holographic setups the full spatial bandwidth of the sensor cannot be used. However, it is very important to use the entire spatial bandwidth of the sensor because the lateral resolution of the reconstructed image depends on a complete evaluation of all the information one can get from the sensor.

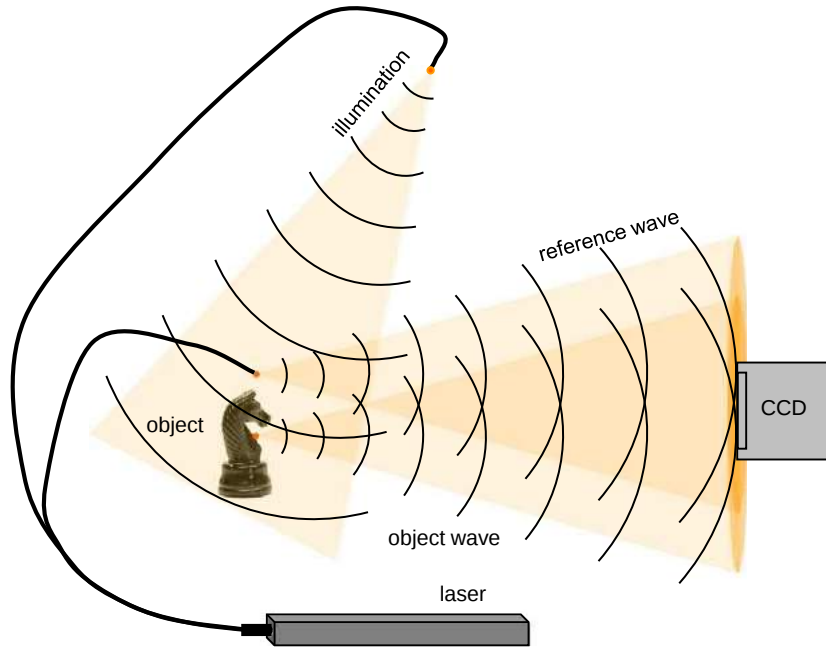


Fig. 7 Scheme of a setup for digital lensless Fourier holography of diffusely reflecting objects.

Even if the speed of digital signal processing is increasing rapidly, algorithms should be as simple and as fast to compute as possible. For the Fresnel approach and the convolution approach several fast Fourier transforms and complex multiplications are necessary. Therefore a more effective approach such as the subsequently described algorithm seems promising.

The lensless Fourier approach (Takeda *et al.*, 1996; Wagner *et al.*, 1999) is the fastest and most suitable algorithm for small objects. The corresponding setup is shown in Fig. 7. It allows us to choose the lateral resolution in a range from a few microns to hundreds of microns without any additional optics. Each point (ξ, η) on the hologram is again considered as a source point of a spherical elementary wave front (Huygen's principle). The intensity of these elementary waves is modulated by $t[h(\xi, \eta)]$ – the amplitude transmission of the hologram. The reconstruction algorithm for lensless Fourier holography is based on the Fresnel reconstruction. Here, again $u(x, y)$ is the object wave in the object plane, $h(\xi, \eta)$ the hologram, $r(\xi, \eta)$ the reference wave in the hologram plane and $u'(x', y')$ the reconstructed wave field.

For the specific setup of lensless Fourier holography a spherical reference wave is used with an origin in the same distance from the sensor as the object itself. In the case that $d' = -d$, $x = x'$, and $y = y'$ both virtual images are reconstructed and the reconstruction algorithm is then

$$u'(x', y') = \frac{\exp(ikd)}{i\lambda d} e^{-i\frac{k}{2d}(x^2 + y^2)} \text{FT}_{\lambda d} \left\{ t[h(\xi, \eta)] \cdot r(\xi, \eta) \cdot e^{-i\frac{k}{2d}(\xi^2 + \eta^2)} \right\} \quad (30)$$

where $\text{FT}_{\lambda d}$ is the two-dimensional Fourier transformation which has been scaled by a factor $1/(\lambda d)$. In this recording configuration, the effect of the spherical phase factor associated with the Fresnel diffraction pattern of the object is eliminated by the use of a spherical reference wave $r(\xi, \eta)$ with the same average curvature:

$$r(x, y) = \text{const} \cdot \exp\left(i\frac{\pi}{\lambda d}(\xi^2 + \eta^2)\right) \quad (32)$$

This results in a more simple reconstruction algorithm which can be described by

$$u'(x', y') = \text{const} \cdot \exp\left(-i\frac{\pi}{\lambda d}(x^2 + y^2)\right) \text{FT}_{\lambda d} \{h(\xi, \eta)\} \quad (33)$$

Besides the faster reconstruction algorithm (only one Fourier transform has to be computed), the Fourier algorithm uses the full space-bandwidth product (SBP) of the sensor chip because it adapts the curvature of the reference wave to the curvature of the object wave.

Influences of Discretization

For the estimation of the lateral resolution in digital holography three different effects related to discretization have to be considered (Wagner *et al.*, 1999): averaging, sampling and the limited sensor size. We assume a quadratic sensor with $N \times M$ pixels of size $\Delta\xi \times \Delta\eta$. Each pixel has a light sensitive region with a side length $\gamma\Delta\xi = \Delta\xi_{\text{eff}}$. The quantity γ^2 is the so-called fill factor ($0 \leq \gamma \leq 1$) that indicates the active area of the pixel. The sensor averages the incident light which has to be considered because of

possible intensity fluctuations over this area. The continuous expression for the intensity $I(\xi, \eta) = t[h(\xi, \eta)]$ registered by the sensor has to be integrated over the light sensitive area. This can be expressed mathematically by the convolution of the incident intensity $I(\xi, \eta)$ with a rectangle function (Goodman, 1996)

$$I^1(\xi, \eta) \propto I \otimes \text{rect}_{\Delta\xi_{\text{eff}}, \Delta\eta_{\text{eff}}} \quad (34)$$

The discrete sampling of the light field is modeled by the multiplication of the continuous assumed hologram with the two-dimensional comb-function:

$$I^2(\xi, \eta) \propto I^1(\xi, \eta) \cdot \text{comb}_{\Delta\xi, \Delta\eta} \quad (35)$$

Finally, the limited sensor size requires that the comb-function has to be truncated at the borders of the sensor. This is considered by the multiplication with a two-dimensional rect-function of the size $N_x \Delta\xi$:

$$I^3(\xi, \eta) \propto I^2 \cdot \text{rect}_{N\Delta\xi, N\Delta\eta} \quad (36)$$

The consequences are amplitude distortion, aliasing and speckling that have to be considered in the reconstruction procedure (Wagner *et al.*, 1999; Seebacher, 2001).

Advantages of Digital Holography

Besides the electronic processing and the direct access to the phase some further advantages recommend digital holography for several applications such as metrology and microscopy (Osten, 2003).

- The availability of the independently reconstructed phase distributions of each individual state of the object and interferometer, respectively, offers the possibility to record a series of digital holograms with increased load amplitude. In the evaluation process the convenient states can be compared interferometrically. Furthermore, a series of digital holograms with increasing load can be applied to unwrap the $\text{mod } 2\pi$ -phase temporally (Wagner *et al.*, 2000). In this method the total object deformation is subdivided in many measurement steps in which the phase differences are smaller than 2π . By adding up those intermediate results, the total phase change can be obtained without any further unwrapping. This is an important feature, since it is essential to have an unwrapped phase to be able to calculate the real deformation data from the phase map. Fig. 8 shows an example of such a measurement. The left image shows the wrapped deformation phase for a clamped coin which was loaded with heat. The right image shows the temporal unwrapped phase which has been obtained by dividing the total deformation into 85 sub-measurements. Thus the displacement of the object can be observed almost in real time during the loading process.
- The independent recording and reconstruction of all states gives also a new degree of freedom for optical shape measurement. In case of multi-wavelength contouring each hologram can be stored and reconstructed independently with its corresponding wavelength. This results in a significant decrease of aberrations and makes it possible to use larger wavelength differences for the generation of shorter synthetic wavelengths (Wagner *et al.*, 2000).
- Because all states of the inspected object can be stored and evaluated independently, only seven digital holograms are necessary to measure the 3D-displacement field and – shape of an object under test: one holograms for each illumination direction before and after the loading, respectively, and one hologram with a different wavelength (or a different source-point of illumination) which can interfere with one of the other holograms to do two-wavelength-contouring (or two-source-point-contouring) for shape measurement. If four illumination directions are used nine holograms are necessary (Seebacher *et al.*, 2001).
- The direct approach to the phase enables various new microscopic principles for the investigation of phase objects. They are categorized with the term Quantitative Phase Contrast Microscopy (Cuche *et al.*, 1999; Kim, 2011).

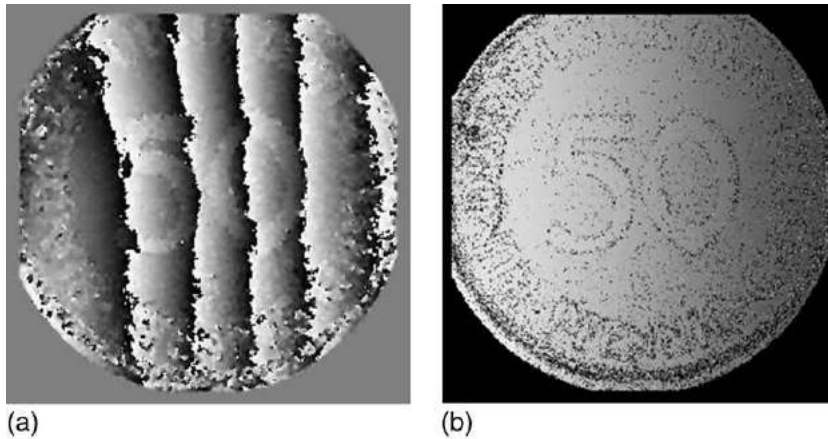


Fig. 8 Temporal unwrapping by digital holography. (a) wrapped phase of a thermal loaded coin, (b) temporally unwrapped phase.

- Digital holography also allows new approaches for the exploration and exploitation of scattering media (Singh *et al.*, 2017).
- Last but not least, the complex holographic sensor setups can be miniaturized drastically (Kolenovic *et al.*, 2003) and methods for remote comparative interferometry (Baumbach *et al.*, 2006) make digital holography to a versatile tool for the solution of numerous future inspection and measurement problems.

See also: Holography: Computer Generated Holograms

References

- Baumbach, T., Osten, W., Kopylow, C., Jueptner, W., 2006. Remote metrology by comparative digital holography. *Appl. Opt.* 45, 925–934.
- Butters, J.N., Leendertz, J.A., 1971. Holographic and video techniques applied to engineering measurement. *J. Meas. Control* 4, 349–354.
- Coufal, H.J., Psaltis, D., Sincerbox, G.T., Glass, A.M., Cardillo, M.J. (Eds.), 2000. *Holographic Data Storage*. Berlin: Springer.
- Creath, K., 1993. Temporal phase measurement methods. In: Robinson, D.W., Reid, G.T. (Eds.), *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.
- Cuche, E., Marquet, P., Depeursinge, C., 1999. Simultaneous amplitude-contrast and quantitative phase-contrast microscopy by numerical reconstruction of Fresnel off-axis holograms. *Appl. Opt.* 38 (34), 6994–7001.
- Daneshpanah, M., Zwick, S., Schaal, F., *et al.*, 2010. 3D Holographic imaging and trapping for non-invasive cell identification and tracking. *J. Display Technol.* 6, 400–409.
- Demetrakopoulos, T.H., Mittra, R., 1974. Digital and optical reconstruction of images from suboptical patterns. *Appl. Opt.* 13, 665–670.
- Demetrakopoulos, T., Mittra, T., 1974. Digital and optical reconstruction of images from suboptical diffraction patterns. *Appl. Opt.* 13, 665–670.
- Denisjuk, Yu.N., 1962. Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Dokl. Akad. Nauk SSSR* 144, 1275–1279.
- Ferraro, P., De Nicola, S., Finizio, A., *et al.*, 2003. Compensation of the inherent wave front curvature in digital holographic coherent microscopy for quantitative phase-contrast imaging. *Appl. Opt.* 42, 1938–1946.
- Frauel, Y., Naughton, T.J., Matoba, O., Tajahuerce, E., Javidi, B., 2006. Three-dimensional imaging and processing using computational holographic imaging. *Proc. IEEE* 94 (3), 636–653.
- Gabor, D., 1949. Microscopy by reconstructed wave-fronts. *Proc. Royal Soc., A* 197, 454–487.
- Gang, J., 2013. Three-dimensional display technologies. *Adv. Opt. Photon.* 5 (4), 456–535.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*. New York: McGraw Hill Comp. Inc..
- Goodman, J.W., Lawrence, R.W., 1967. Digital image formation from electronically detected holograms. *Appl. Phys. Lett.* 11, 77–79.
- Haist, T., Hasler, M., Osten, W., Baranek, M., 2014. Programmable Microscopy. In: Javidi, B., Tajahuerce, E., Andrés, P. (Eds.), *Multidimensional Imaging*. New York: John Wiley & Sons, pp. 153–174.
- Haist, T., Osten, W., 2015a. Holography using pixelated spatial light modulators – Part 1: Theory and basic considerations. *J. Micro/Nanolith MEMS MOEMS* 14 (4), 041310.
- Haist, T., Osten, W., 2015b. Holography using pixelated spatial light modulators – Part 2: Applications. *J. Micro/Nanolith MEMS MOEMS* 14 (4), 041311.
- HAMAMTSU Photonics K.K., <http://www.hamamatsu.com/jp/en/technology/innovation/lcos-slm/index.html>.
- Hariharan, P., 1984. *Optical Holography*. Cambridge: Cambridge University Press.
- Hasegawa, S., Hayasaki, Y., Nishida, N., 2006. Holographic femtosecond laser processing with multiplexed phase Fresnel lenses. *Opt. Lett.* 31, 1705–1707.
- HOLOEYE Photonic AG, <http://holoeeye.com/>.
- Huang, T., 1971. Digital Holography. *Proc. IEEE* 59, 1335–1346.
- Kim, M.K., 2011. *Digital Holographic Microscopy*. Berlin: Springer.
- Kohler, C., Schwab, X., Osten, W., 2006. Optimally tuned spatial light modulators for digital holography. *Appl. Opt.* 45, 960–967.
- Kolenovic, E., Osten, W., Klattenhoff, R., *et al.*, 2003. Miniaturized digital holography sensor for distal three-dimensional endoscopy. *Appl. Opt.* 42 (25), 5167–5172.
- Kreis, T., 2005. *Handbook of Holographic Interferometry – Optical and Digital Methods*. Weinheim: Wiley-VCH.
- Kreis, Th., Jüptner, W., 1997. The suppression of the dc-term in digital holography. *Opt. Eng.* 36, 2357–2360.
- Kronrod, M.A., Merzlyakov, N.S., Yaroslavsky, L.P., 1972. Reconstruction of holograms with a computer. *Sov. Phys. Tech. Phys.* 17, 333–334.
- Kujawinska, M., 1993. Spatial phase measurement methods. In: Robinson, D.W., Reid, G.T. (Eds.), *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.
- Lazarev, G., Hermerschmidt, A., Krüger, S., Osten, S., 2012. LCOS Spatial Light Modulators: Trends and Applications. In: Osten, W., Reingand, N. (Eds.), *Optical Imaging and Metrology: Advanced Technologies*. Weinheim: Wiley-VCH, pp. 1–29.
- Lee, W.-H., 1978. Computer generated holograms: Techniques and applications. In: Wolf, E. (Ed.) *Progress in Optics*, XVI. Amsterdam: North-Holland, pp. 121–230.
- Lee, B., Kim, Y., 2012. Three-dimensional display and imaging: Status and prospects. In: Osten, W., Reingand, N. (Eds.), *Optical Imaging and Metrology: Advanced Technologies*. Weinheim: Wiley-VCH, pp. 31–56.
- Leith, E.N., Upatnieks, J., 1961. New techniques in wavefront reconstruction. *JOSA* 51, 1469–1473.
- Liesener, J., Reicherter, M., Tiziani, H.J., 2004. Determination and compensation of aberrations using SLMs. *Opt. Commun.* 233, 161–166.
- Marquet, P., Depeursinge, C., 2014. Digital Holographic Microscopy: a New Imaging Technique to Quantitatively Explore Cell Dynamics with Nanometer Sensitivity. In: Javidi, B., Tajahuerce, E., Andrés, P. (Eds.), *Multidimensional Imaging*. New York: John Wiley & Sons, pp. 197–212.
- Maurer, C., Jesacher, A., Bernet, S., Ritsch-Marte, M., 2010. What spatial light modulators can do for optical microscopy. *Laser Photonics Rev.* 5, 81–101.
- Onural, L., Yaras, F., Kang, H., 2011. Digital holographic three-dimensional video displays. *Proc. IEEE* 99 (4), 576–589.
- Osten, W., 2003. Active metrology by digital holography. *Proc. SPIE* 4933, 96–110.
- Osten, W., Baumbach, Th., Jueptner, W., 2002. Comparative Digital Holography. *Opt. Lett.* 27, 1764–1766.
- Osten, W., Faridian, A., Gao, P., *et al.*, 2014. Recent advances in digital holography. *Appl. Opt.* 53 (27), G44–G63. doi:10.1364/AO.53.000G44.
- Osten, W., Ferraro, P., 2007. Digital Holography and IST application in MEMS/MOEMS inspection. In: Osten, W. (Ed.), *Optical Inspection of Microsystems*. Boca Raton, FL: Taylor and Francis, pp. 351–425.
- Osten, W., Jüptner, W., Seebacher, S., 1998. The qualification of large scale approved measurement techniques for the testing of microcomponents. *Proceedings 18th Symposium on Seebacher, Experimental Mechanics of Solids*, Jachranka, pp. 43–55.
- Osten, W., Kujawinska, M., 2000. Active phase measuring metrology. In: Rastogi, P.K., Inaudi, D. (Eds.), *Trends in Optical Nondestructive Testing and Inspection*. Elsevier Science B.V., pp. 45–69.
- Osten, W., Wilke, M., Pedrini, G., 2013. Remote laboratories for optical metrology: From the lab to the cloud. *Opt. Engn.* 52 (10), 101914. [Article Number: 101914].
- Powell, R.L., Stetson, K., 1965. Interferometric vibration analysis of three-dimensional objects by wavefront reconstruction. *JOSA* 55, 1593–1598.
- Reicherter, M., Liesener, J., Haist, T., Tiziani, H.J., 1999. Optical particle trapping with computer-generated holograms written in a liquid crystal display. *Opt. Lett.* 9, 508–510.
- Schnars, U., Jüptner, W., 1994. Direct recording of holograms by a CCD target and numerical reconstruction. *Appl. Opt.* 33 (2), 179–181.
- Schnars, U., 1994. Direct phase determination in hologram interferometry with use of digitally recorded holograms. *J. Opt. Soc. Am. A* 11, 2011–2015.

- Seebacher, S., Osten, W., Baumbach, Th., Jüptner, W., 2001. The determination of material parameters of microcomponents using digital holography. *Opt. Las Eng.* 36 (2), 103–126.
- Seebacher, S., Osten, W., Jüptner, W., 1997. 3-D deformation analysis of micro-components using digital holography. *Proc. SPIE* 3098, 382–391.
- Seebacher, S., 2001. Application of digital holography for 3d-shape and deformation measurement of micro components. PhD Thesis, University of Bremen.
- Singh, A., Pedrini, G., Takeda, M., Osten, W., 2017. Exploiting scattering media for exploring 3D objects. *Light: Science & Applications* 6, e16219. doi:10.1038/lsa.2016.219.
- Steel, W.H., 1983. *Interferometry*. Cambridge: Cambridge University Press.
- Sutkowski, M., Kujawinska, M., 2000. Application of liquid crystal (LC) devices for optoelectronic reconstruction of digitally stored holograms. *Opt. Laser Eng.* 33, 191–201.
- Takeda, M., Taniguchi, K., Hirayama, T., Kogho, H., 1996. Single transform Fourier-Hartley fringe analysis for holographic interferometry. In: Füzessy, Z., Jüptner, W., Osten, W. (Eds.), *Simulation and Experiment in Laser Metrology*. Berlin: Akademie Verlag, pp. 67–73.
- Wagner, C., Osten, W., Seebacher, S., 2000. Direct shape measurement by digital wavefront reconstruction and multi-wavelength contouring. *Opt. Eng.* 39 (1), 79–85.
- Wagner, C., Seebacher, S., Osten, W., Jüptner, W., 1999. Digital recording and numerical reconstruction of lensless Fourier holograms in optical metrology. *Appl. Opt.* 38 (22), 4812–4820.
- Yamaguchi, I., Zhang, T., 1979. Phase-shifting digital holography. *Opt. Lett.* 22, 1268–1270.
- Yaroslavsky, L., 2004. *Digital Holography and Digital Image Processing: Principles, Methods, Algorithms*. Dordrecht: Kluwer Academic Publishers.
- Yu, X., Hong, J., Liu, C., Kim, M.K., 2014. Review of digital holographic microscopy for three-dimensional profiling and tracking. *Opt. Eng.* 53, 112306.
- Zhang, Y., Pedrini, G., Osten, W., Tiziani, H., 2003. Image reconstruction for in-line holography with the Yang-Gu algorithm. *Appl. Opt.* 42 (32), 6452–6457.
- Zwick, S., Haist, T., Warber, M., Osten, W., 2010. Dynamic holography using pixelated light modulators. *Appl. Opt.* 49, F47–F58.

Overview: Holography

C Shakher and AK Ghatak, Indian Institute of Technology, New Delhi, India

© 2005 Elsevier Ltd. All rights reserved.

Major Milestones

- 1948: Essential concept for holographic recording by Dennis Gabor.
- 1960: The first successful operation of a laser device by Theodore Maiman.
- 1962: Off-axis technique of holography by Leith and Upatnieks.
- 1962: Yu N Denisjuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.
- 1964: Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.
- 1969: S A Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications.

Holography is the science of recording an entire optical wavefront, both amplitude and phase information, on appropriate recording material. The record is called a hologram. Unlike conventional photography, which records a three-dimensional scene in a two-dimensional format, holography records true three-dimensional information about the scene.

Holography was invented by Gabor in 1948 and his first paper introduced the essential concept for holographic recording – the reference beam. Holography is based on the interference between waves and, it provides us with a way of storing all the light information arriving at the film in such a way that it can be regenerated later.

The use of the reference beam is utilized because the physical detectors and recorders are sensitive only to light intensity. The phase is not recorded but is manifest only when two coherent waves of the same frequency are simultaneously present at the same location. In that case, the waves combine to form a single wave whose intensity depends not only on intensities of the two individual waves, but also on the phase difference between them. This is key to holography. The film record, or hologram, can be considered as a complicated diffraction grating. Holograms bear no resemblance to conventional photographs in that an image is not actually recorded. In fact, the interferometric fringes which are recorded on the recording material are not visible to an unaided eye because of extremely fine interfringe spacing (~ 0.5 micrometer). The fringes which are visible on the recording material are the result of dust particles in the optical system used to produce the hologram.

Gabor's original technique is now known as in-line holography. In this arrangement, the coherent light source as well as the object, which is a transparency containing small opaque details on a clear background, is located along the axis normal to the photographic plate. With such a system, an observer focusing on one image observes it superposed on the out-of-focus twin image as well as a strong coherent background. This constitutes the most serious problem of Gabor's original technique.

The first successful technique for separating the twin images was developed by Leith and Upatnieks in 1962. This is called the off-axis or side band technique of holography. We shall consider mainly the off-axis technique here.

Basic Holography Principle

Light from a laser or any light beam is characterized by spatial coherence (l_c) and temporal coherence (τ_c), which is discussed in all textbooks on optics. For typical laboratory hologram recording, the required degree of coherence of laser radiation is determined by the type of object and geometrical arrangement being used for the recording. The following condition must hold:

$$L \ll c\tau_c \quad (1)$$

where τ_c is the coherence time, and L the maximum path length difference between the two waves chosen to record the hologram. To begin recording, two wavefronts are derived from the same laser source. One wavefront is used to illuminate the object and another is used as a reference wavefront. The wavefront derived by illuminating the object is superimposed with the reference wave and the interference pattern is recorded. These object and reference waves must:

- be coherent (derived from the same laser); and
- have a fixed relative phase at each point on the recording medium.

If the above conditions are not met, the fringes will move during the exposure and the holographic record will be smeared. Therefore, to record the hologram one should avoid air currents and vibrations. The recording medium must have sufficient resolution to record the fine details in the fringe pattern. This will be typically of the order of the wavelength. To create a three-dimensional image from the holographic process one then has to record and reconstruct the hologram. The object and reference waves are derived from the same laser.

Holograms record an interference pattern formed by interference of object and reference wavefronts, as explained above. We may mention here that for the two plane waves propagating at an angle θ between them, the spacing of the interference fringes is given by

$$\alpha = \frac{\lambda_0}{2 \sin(\theta/2)} \quad (2)$$

where λ_0 is the free space wavelength.

Processing of the hologram (developing, fixing, and washing of the recording material) yields a plate with alternating transparent and opaque parts, variation of refractive index, or variation of height corresponding to intensity variation in the fringe pattern. Such a plate can be regarded as a complicated diffraction grating. In this process, the hologram is illuminated with monochromatic, coherent light. The hologram diffracts this light into wavefronts which are essentially indistinguishable from the original waves which were diffracted from the object. These diffracted waves produce all the optical phenomena that can be produced by the original waves. They can be collected by a lens and brought into focus, thereby forming an image of the original object, even though the object has since been removed. If the reconstructed waves are intercepted by the eye of an observer, the effect is exactly as if the original waves were being observed; the observer sees what appears to be the original object in true three-dimensional form. As the observer changes his viewing position, the perspective of the image scene changes; parallax effects are evident and the observer must refocus when the observation point is changed from a near to a distant object in the scene. Assuming that both the construction and reconstruction of the hologram are made with the same monochromatic light source, there is no visual test which can be made to distinguish between the real object and the reconstructed image of the object. It is as if the hologram were a window through which the apparent object is viewed.

Hologram of a Point Object

Consider a hologram recorded with a collimated reference wave normal to the recording plate and a point object inclined at a certain angle. If the hologram is illuminated once again with the same collimated reference wave, it reconstructs two images, one virtual true image and the other real image. However, the two images differ in one very important respect.

While the virtual image is located in the same position as the object and exhibits the same parallax properties, the real image is formed at the same distance from the hologram but in front of it. Corresponding points on the real and virtual images are located at equal distances from the plane of the hologram; the real image has the curious property that its depth is inverted. Such an image is not formed with a normal optical system; it is therefore called a pseudo image as opposed to a normal or orthoscopic image.

This depth inversion results in conflicting visual clues, which make viewing of the real image psychologically unsatisfactory. Thus, if O_1 and O_2 are two elements in the object field, and if O_1 blocks the light scattered by O_2 at a certain angle, the hologram records information only on the element O_1 at this angle and records no information about this part of O_2 . An observer viewing the real image from the corresponding direction then cannot see this part of O_2 , which, contrary to normal experience, is obscured by O_1 , even though O_2 is in front of O_1 .

Production of an Orthoscopic Real Image

An orthoscopic real image of an object can be produced by recording two holograms in succession. In the first step, a hologram is recorded of the object with a collimated reference beam. When this hologram is illuminated once again with the collimated reference beam that was used to record it, it reconstructs two images of the object at unit magnification, one of them being an orthoscopic virtual image, while the other is a pseudoscopic real image. A second hologram is then recorded of this real image with a second collimated reference beam.

When the second hologram is illuminated with a collimated beam it reconstructs a pseudoscopic virtual image located in the same position as the real image formed by the second hologram is an orthoscopic image. Since a collimated reference beam is used throughout, the final real image is the same size as the original object and free from aberrations.

In addition to these characteristic linked intensities with the three-dimensional nature of the reconstruction, the holographic recording has several other properties. Each portion of the hologram can reproduce the entire image scene. If a smaller and smaller portion of the hologram is used for reconstruction, there is loss of image intensity and reconstruction. When a hologram is reversed, such as in contact printing processes, it will still reconstruct a positive image indistinguishable from the image produced by the original hologram.

Simple Mathematical Description of Holography

Let the object and reference waves be given by

$$U_o = O(x, y)e^{i\phi(x, y)} \quad (3)$$

and

$$U_r = R e^{iky \sin \theta} \quad (4)$$

U_r describes the reference (plane wave) propagating at angle θ to the z -axis.

The intensity at the hologram plane is

$$\begin{aligned} I &= |U_r + U_o|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \end{aligned} \quad (5)$$

As can be seen from the above equation, the amplitude and phase of the wave are encoded as the amplitude and phase modulation of a set of interference fringes. The material used to record the patterns of interference fringes described above is assumed to provide linear mapping of the intensity incident during the reconstruction process into amplitude transmitted by or reflected from the recorded material. Usually both light detection and wavefront modulation are performed by photographic plate/film. We assume the amplitude transmission properties of the plate/film after processing to be described by

$$T = T_o - b(E - E_o) \quad (6)$$

where the exposure E , at the film is $E = I\tau$; here τ is the exposure time. T_o is the transmittance of the unexposed plate.

If the hologram is reilluminated by the reference wave, the transmitted wave amplitude will be

$$U_t = U_r(T_o - b\tau(|U_o|^2 + U_r^* U_o + U_r U_o^*)) \quad (7)$$

The first term is reference wave times a constant. The second term is the reference wave modulated by $|U_o|^2 = O(x, y)^2$; it gives small-angle scattering about the reference wave direction. The third term is proportional to U_o . It is the same as the original object wave (note this is only so if the reconstruction wave is identical to the reference wave). The fourth term is

$$-b\tau U_r^2 U_o^* = -b\tau R^2 e^{iky 2 \sin \theta} O(x, y) e^{-i\phi(x, y)} \quad (8)$$

This is essentially a wave traveling in a direction $\sin^{-1}(2 \sin \theta)$ to the z -axis, with the correct object amplitude modulation but its phase reversed in sign, producing a conjugate wave.

Types of Holograms

Primarily, the holograms are classified as thin and thick, based on the thickness of the recording medium. When the thickness of the recording medium is small compared with the average spacing of the interference fringes, the hologram can be treated as a thin hologram. Such holograms are characterized by spatially varying complex amplitude transmittance:

$$t(x, y) = [t(x, y)] \exp[-i\phi(x, y)] \quad (9)$$

Thin holograms can be further categorized as thin amplitude holograms or thin phase holograms. If amplitude transmittance of the hologram is such that $\phi(x, y)$ is constant while $t(x, y)$ varies over the hologram, the hologram is termed an amplitude hologram. For a lossless phase hologram, $|t(x, y)| = 1$, so that the complex amplitude transmittance is caused by variation in phase.

When the thickness of the hologram recording material is large compared to the fringe spacing of the interference fringes, the holograms may be considered as volume holograms. These may be treated as a three-dimensional system of layers corresponding to a periodic variation of absorption coefficient or refractive index, and the diffracted amplitude is at a maximum when Bragg's condition is satisfied. In general, the behavior of thin and thick holograms corresponds to Raman Nath and Bragg diffraction regimes. The distinctions between two regimes is made on the basis of a parameter Q , which is defined by the relation:

$$Q = \frac{2\pi\lambda_o d}{n_o \Lambda^2} \quad (10)$$

where

Λ = spacing of fringe on hologram, measured normal to the surface;

d = thickness of recording medium;

n_o = mean refractive index of the medium; and

λ_o = wavelength of light.

Small values of Q ($Q < 1$) correspond to thin gratings, while large values of Q ($Q > 10$) correspond to volume gratings. For values of Q between 1 and 10, intermediate properties are exhibited. However, this criterion is not always adequate. The boundaries between the thin and thick holograms are given by the value of

$$P = \frac{\lambda_o^2}{\Lambda^2 n_o n_1} \quad (11)$$

where P^{-2} is the relative power diffracted into higher orders, and $P < 1$ for thin holograms and $P > 10$ for thick holograms. For P having values between 1 and 10, the hologram may be thin or thick, depending on the other parameters involved.

Holograms can also be classified based on whether they are reconstructed in transmitted light or reflected light. For transmission holograms, during the recording stage, two interfering wavefronts make equal but opposite angles to the surfaces of

recording medium and are incident on it from the same side. Reflection holograms reconstruct images in reflected light. They are recorded such that the interfering wavefronts are symmetrical with respect to the surface of the recording medium but are incident on it from opposite sides. When the angle between the interfering wavefronts is maximum (180°), the spacing between the fringe planes of the recording medium is minimum. Under such conditions, reflection holograms may have wavelength sensitivity high enough to be reconstructed, even with white light.

An important development in the field of holography was the invention of rainbow holograms by Benton in 1969. This provides a method for utilizing white light for illumination when viewing the holograms. The technique does so by minimizing the blur introduced by color dispersion in transmission holograms, at the price of giving up parallax information in one dimension.

Recording Materials

An ideal recording medium for holography should have a well matching spectral sensitivity corresponding to available laser wavelengths, linear transfer characteristics, high resolution, and low noise. It should also be either inexpensive or indefinitely recyclable. Toward achieving the above-mentioned properties, several materials have been studied, but none has been found so far that meets all the requirements. Materials investigated include: silver halide photographic emulsion; dichromated gelatin plates/films; silver halide sensitized gelatin plates/films; photo-resists; photo polymer systems; photochromics; photo thermoplastics; and ferro-electric crystals. Recently, the use of storage and processing capabilities of computers, together with CCD cameras, has been used for recording holograms.

Silver halide photographic emulsions are a commonly used recording material for holography, mainly because of relatively high sensitivity and easy availability. Manufacture, use, and processing of these emulsions have been well standardized. The need for wet processing and drying may constitute a major drawback. Dichromated gelatin can be considered an almost ideal recording material for volume phase holograms, as it has a large refractive index modulation capability, high resolution, low absorption and scattering.

Because of these features, dichromated gelatin has been extensively investigated. The most significant disadvantage is its comparatively low sensitivity. Pennington *et al.* developed an alternative technique, which combines the advantage of silver halide emulsions (high sensitivity) with that of dichromated gelatin (low absorption scattering and high stability). It involves exposing silver halide photographic emulsion and then processing it, so as to obtain a volume phase hologram consisting solely of hardened gelatin.

Thin phase holograms can be recorded on photoresists, which are light-sensitive organic films yielding relief image after exposure and development. They offer advantages of easy replication using thermoplastics but are slow in response and undergo nonlinear effects at diffraction efficiency greater than ~ 0.05 . Shipley AZ-1350 is a widely used photoresist, with maximum sensitivity in the ultraviolet, dropping rapidly for longer wavelengths towards the blue.

Photopolymers are being keenly investigated because they offer advantages such as ease of handling, low cost, and real-time recording for the application of holography and nonlinear optics. However, they have low sensitivity and short shelf life.

Thin phase surface relief holograms can be recorded in a thin layer of thermoplastics. These materials have a reasonably high sensitivity over the whole visible spectrum, fairly high diffraction efficiency, and do not require wet processing. Their application in holographic interferometry, optical information processing, and in making compact holographic devices has been widely reported.

Photorefractive crystals such as $\text{Bi}_{12}\text{SiO}_{20}$, LiNbO_3 , $\text{Bi}_{12}\text{GeO}_{20}$, BaTiO_3 , LiTaO_3 , etc., as recording materials, offer high-angular sensitivity and provide capability to read and write volume holographic data in real time. Besides the materials discussed here, several other materials have been investigated for recording holograms; these include photocromics, elastomere devices, magneto-optic materials, etc.

Application of Holography

Holography can be constructed not only with the light waves of lasers, but also with sound waves, microwaves, and other waves in the electromagnetic spectrum of radiation. Holograms made with ultraviolet light or X-rays can record images of objects/particles smaller than the wavelength of visible light, e.g., atoms or molecules. Acoustical holography uses sound waves to see through solid objects. Holography has a vast scope of practical applications, which have been classified into two major categories:

1. applications requiring three-dimensional images for visual perception; and
2. applications in which holography is used as a measuring tool.

The unique ability of holography – to record and reconstruct both electromagnetic and sound waves – makes it a valuable tool for education, science, industry, and business. Below are some of the important applications:

1. Holographic interferometry (HI) is one of the most powerful measuring tools. In HI, two states of an object, i.e., initial and deformed state are recorded on the same photographic plate. After reconstruction of the light wave corresponding to two states of an object, interferences and deformations are displayed in terms of the interference pattern. The change of distance of one tenth of a micron, or lower, can be resolved. HI provides scientists/engineers with crucial data for design of critical machine

parts of power-generating equipment, in the aircraft industry, automobile industry, and nuclear installations (say, for example, in the design of containers used to transport nuclear materials, improve the design of aircraft wings and turbine blades, etc.). Presently, HI is being widely used in mechanical engineering, acoustics, and aerodynamics, for nondestructive testing, to investigate oscillation in diaphragms and flow around various objects, respectively.

2. Microwave holography can detect objects deep within spaces, by recording the radio waves they emit.
3. Another important application of holography is the design of optical elements, which possess special properties. A holographic recording of a concave mirror behaves in much the same way as the mirror itself, i.e., it can focus the light. In some cases chromatism can be introduced in the design of elements so that a location of the point, where the beams are focused, depends on the wavelength. This can be achieved by accurately choosing the recording arrangement of the focusing elements, these elements are, in fact, diffraction gratings. These can have low noise levels, freedom from astigmatism, and have other useful properties. Holographic optical elements have found applications in supermarket scanners to read barcodes, headup displays in fighter aircraft to observe critical cockpit instruments, etc.
4. A telephone credit card used in Europe and other developed countries has embossed surface holograms which carry a monetary value. When the card is inserted into the telephone, a card reader discerns the amount due and deducts the appropriate amount to cover the cost of the call.
5. Holography is having applications in analog and digital computers, offering remarkable opportunities to realize various logical operators, devices for identifying images based on matched filtering, and in computer memory units. The basic advantage of holographic memory is that a relatively large volume of information can be stored and that there are a limited number of ways to change the record. The arrival of the first prototype of optical computers, which use holograms as storage material for data, could have a dramatic impact on the holographic market. The yet-to-be-unveiled optical computers will be able to deliver trillions of bits of information faster than the current generation of computers.
6. Optical tweezers are going to become an important tool for the study of micro-organisms/bacteria, etc.
7. Holograms can be used to locate and retrieve information without knowing its location in the storage medium, only needing to know some of its content.
8. The links between computer science and holography are now well-established and are developing, with at least two aspects making computer-generated holograms extremely interesting. First, such holograms enable us to obtain visual 3-dimensional reconstructions of imagined objects. For example, one can reconstruct three dimensions of a model of an object still in the design stage. Second, computer-generated holograms can be used to reconstruct lightwaves with specified wave fronts. This means specially computed and manufactured holograms may function as optical elements that transform the incident light wave into desired wavefronts.
9. Another important applications of holography is its utilization to compensate for the distortion that occurs when viewing objects through optically heterogeneous mediums. It can be achieved based on the principle of beam reversal by the hologram.
10. Finally, let us consider the application of holography to art. The development of holography gives very effective ways of creating qualitative three-dimensional images. Thus, a new independent area of holographic creative work representational/artistic holography has appeared. The art of holographic depiction has developed along two major routes. The first is creation of the view hologram, used as holograms of natural objects that are to be displayed in exhibitions and museums; these are also known as artistic holograms. Portrait holography is also classified under this category. The progress in portrait holography is hampered partly because of imperfection of pulsed lasers and partly because of the deterioration of photographic material when exposed to pulsed electromagnetic radiation.

The principle behind the creation of elusion by using composite holograms is also very convincing for the display of objects. To synthesize the composite holograms, the photographs of various aspects of a scene are printed onto the photographic plate. The synthesis techniques for preparing the composite hologram are very complicated, while the images created by holograms are still far from perfect. However, there is no doubt that composite holography opens holography up as an artistic technique. Recently rainbow holography has been very popular to display such objects.

See also: Fraunhofer Diffraction. Fresnel Diffraction

Further Reading

- Benton, S.A., 1969. On a method for reducing the information content of holograms. *Journal of the Optical Society of America* 59, 1545.
- Denisyuk, Y.N., 1962. Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Sov. Phys.-Dokl.* 7, 543.
- Denisyuk, Y.N., 1963. On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt. I. *Optical Spectroscopy* 15, 279.
- Denisyuk, Y.N., 1965. On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt II. *Optical Spectroscopy* 18, 152.
- Gabor, D., 1949. Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society A* 197, 454.
- Gabor, D., 1951. Microscopy by reconstructed wavefronts: II. *Proceedings of the Physical Society B* 64, 449.
- Gabor, S.D., 1948. A new microscopic principle. *Nature* 161, 777.
- Hariharan, P., 1983. *Optical Holography: Principle, Techniques and Applications*. Cambridge, UK: Cambridge University Press.
- Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. *Journal Optical of the Society of America* 52, 1123.

- Leith, E.N., Upatnieks, J., 1963. Wavefront reconstruction with continuous-tone objects. *Journal of the Optical Society of America* 53, 1377.
- Leith, E.N., Upatnieks, J., 1964. Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54, 1295.
- Pennington, K.S., Lin, L.H., 1965. Multicolor wavefront reconstruction. *Applied Physics Letters* 7, 56.
- Stroke, G.W., Labeyrie, A.E., 1966. White-light reconstruction of holographic images using the Lippmann-Bragg diffraction effect. *Physics Letters* 20, 368.

The Fractional Order Fourier Transform and Fresnel Diffraction

Pierre Pellat-Finet, University of Southern Brittany, Lorient, France

Yezid Torres Moreno, Industrial University of Santander, Bucaramanga, Colombia

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The fractional order Fourier transform is an extension of the standard Fourier transform. Let α be a complex number and let f be a function of two real variables. The 2-dimensional fractional Fourier transform of order α of f is defined by

$$\mathcal{F}_\alpha[f](\vec{\sigma}) = \frac{ie^{-i\alpha}}{\sin \alpha} \exp(-i\pi\sigma^2 \cot \alpha) \int_{\mathbb{R}^2} \exp(-i\pi\rho^2 \cot \alpha) \exp\left(\frac{2i\pi\vec{\sigma} \cdot \vec{\rho}}{\sin \alpha}\right) f(\vec{\rho}) d\vec{\rho} \quad (1)$$

where $\vec{\rho}$ and $\vec{\sigma}$ are 2-dimensional real vectors. The standard Fourier transform is $\mathcal{F}_{\pi/2}$, and $\mathcal{F}_0[f] = f$ for every function f . Fractional order Fourier transforms compose according to $\mathcal{F}_\beta \circ \mathcal{F}_\alpha = \mathcal{F}_{\beta+\alpha}$.

The integral expression of $\mathcal{F}_\alpha$ in Eq. (1) is similar to the integral expression of Fresnel diffraction, so that it is possible to express Fresnel diffraction phenomena in the form of fractional order Fourier transforms and to then develop a method for analyzing and solving some problems of diffraction and light propagation.

Diffraction-Propagation From a Spherical Emitter to a Spherical Receiver

If $\mathcal{A}$ is a spherical emitter or receiver (a spherical segment) with vertex Ω_A and curvature center C_A , its radius of curvature is $R_A = \overline{\Omega_A C_A}$ (Fig. 1). We choose Cartesian coordinates x, y on $\mathcal{A}$ with origin at Ω_A , and use the spatial vector $\vec{r} = (x, y)$ to localize a point on $\mathcal{A}$. We denote $r = \|\vec{r}\| = (x^2 + y^2)^{1/2}$ and $d\vec{r} = dx dy$. We consider monochromatic waves whose wavelength is λ in the propagation medium, which is assumed to be isotropic and homogeneous.

In the framework of a scalar theory of diffraction, the field transfer from a spherical emitter $\mathcal{A}$ (radius R_A , coordinates $\vec{r}$) to a spherical receiver $\mathcal{B}$ (radius R_B , coordinates $\vec{s}$) at a distance D (taken from vertex to vertex, Fig. 1) is expressed by

$$U_B(\vec{s}) = \frac{i}{\lambda D} \exp\left[-\frac{i\pi}{\lambda} \left(\frac{1}{R_B} + \frac{1}{D}\right) s^2\right] \int_{\mathbb{R}^2} \exp\left[-\frac{i\pi}{\lambda} \left(\frac{1}{D} - \frac{1}{R_A}\right) r^2\right] \exp\left(\frac{2i\pi}{\lambda D} \vec{r} \cdot \vec{s}\right) U_A(\vec{r}) d\vec{r} \quad (2)$$

where U_A is the complex (electric) field amplitude on $\mathcal{A}$, and U_B on $\mathcal{B}$. A constant multiplicative phase factor, of the form $\exp(-2i\pi D/\lambda)$, has been omitted in Eq. (2).

Eq. (2) generally corresponds to a Fresnel diffraction phenomenon. If $D = R_A$, we obtain a Fraunhofer phenomenon; moreover if $R_B = -R_A$, the field amplitude U_B is the spatial standard Fourier transform of U_A and the sphere $\mathcal{B}$ is called the Fourier sphere of $\mathcal{A}$, denoted by $\mathcal{F}$ (see Fig. 2).

Mathematical Expression of Diffraction Through a Fractional Order Fourier Transform

The similarity between Eqs. (1) and (2) should be clear. To express Eq. (2) by using a fractional order Fourier transform we proceed as follows. Let K be such that

$$K = \left(1 - \frac{D}{R_A}\right) \left(1 + \frac{D}{R_B}\right) \quad (3)$$

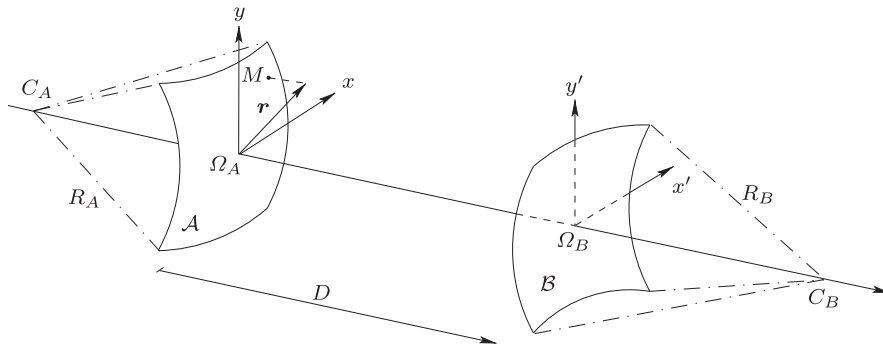


Fig. 1 Describing the field transfer by diffraction-propagation from the spherical emitter $\mathcal{A}$ to the spherical receiver $\mathcal{B}$ at a distance $D = \overline{\Omega_A \Omega_B}$. Generally we have a Fresnel diffraction phenomenon.

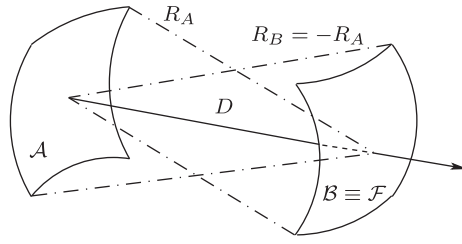


Fig. 2 If $D=R_A=-R_B$, light propagates from $\mathcal{A}$ to its Fourier sphere $\mathcal{F}$ according to a Fraunhofer diffraction phenomenon.

For the sake of simplicity we assume $0 \leq K \leq 1$, and we define α (a real number) by

$$\cos^2 \alpha = K, \quad -\pi \leq \alpha \leq \pi, \quad \alpha D \geq 0 \quad (4)$$

An equivalent definition, that is useful in some derivations, is

$$\cot^2 \alpha = \frac{(D+R_B)(R_A-D)}{D(D-R_A+R_B)}, \quad -\pi \leq \alpha \leq \pi, \quad \alpha D \geq 0 \quad (5)$$

We then define auxiliary parameters ε_A and ε_B by

$$\varepsilon_A = \frac{D}{R_A - D} \cot \alpha, \quad \varepsilon_A R_A > 0 \quad (6)$$

$$\varepsilon_B = \frac{D}{R_B + D} \cot \alpha \quad (7)$$

(Then, necessarily $\varepsilon_B R_B > 0$.)

Eventually we choose scaled spatial variables $\vec{\rho}$ on $\mathcal{A}$ and $\vec{\sigma}$ on $\mathcal{B}$ according to

$$\vec{\rho} = \frac{\vec{r}}{\sqrt{\lambda \varepsilon_A R_A}}, \quad \vec{\sigma} = \frac{\vec{s}}{\sqrt{\lambda \varepsilon_B R_B}} \quad (8)$$

We use scaled field amplitudes V_A on $\mathcal{A}$ and V_B on $\mathcal{B}$ such that

$$V_A(\vec{\rho}) = \sqrt{\frac{\varepsilon_A R_A}{\lambda}} U_A(\sqrt{\lambda \varepsilon_A R_A} \vec{\rho}), \quad V_B(\vec{\sigma}) = \sqrt{\frac{\varepsilon_B R_B}{\lambda}} U_B(\sqrt{\lambda \varepsilon_B R_B} \vec{\sigma}) \quad (9)$$

Then, Eq. (2) becomes

$$V_B(\vec{\sigma}) = e^{i\alpha} \mathcal{F}_\alpha[V_A](\vec{\sigma}) \quad (10)$$

Eq. (10) expresses a Fresnel diffraction phenomenon in the form of a fractional Fourier transform whose order depends on the propagation distance and the radii of the emitter and the receiver. Fraunhofer diffraction is obtained for $\alpha = \pi/2$.

If $K > 1$ or $K < 0$, the order α becomes a complex number and also corresponds to some diffraction phenomenon, depending on propagation distance and radii of curvature.

In order to not repeat definitions of parameters everytime, we will write

$$(R_A, R_B, D, \vec{r}, \vec{s}, U_A, U_B, \lambda) \leftrightarrow (\alpha, \varepsilon_A, \varepsilon_B, \vec{\rho}, \vec{\sigma}, V_A, V_B) \quad (11)$$

to indicate the correspondence between physical parameters and variables (on the left side) and fractional parameters, scaled variables and scaled field amplitudes (on the right side).

Representing diffraction through a fractional order Fourier transform is in accordance with two basic features of light propagation:

1. Continuity of light propagation. Fractional order Fourier transforms are such that $\mathcal{F}_\alpha[f] \rightarrow \mathcal{F}_\beta[f]$ if $\alpha \rightarrow \beta$, for every function f . Near the emitter, we have $\alpha=0$ and $\mathcal{F}[V_A] = V_A$; there is no diffraction. Near the Fourier sphere ($D=R_A$) we have a Fourier transform $\mathcal{F}_{\pi/2}$ and a Fraunhofer diffraction. Physically there is a continuous succession of diffraction phenomenae, from no diffraction (near the emitter) to a Fraunhofer diffraction (near the Fourier sphere), passing by intermediate Fresnel type phenomenae (see Fig. 3).
2. The composition law of fractional order Fourier transforms is in accordance with the Huygens-Fresnel principle that establishes that diffraction from $\mathcal{A}_1$ to $\mathcal{A}_2$ (represented by $\mathcal{F}_\alpha$) can be seen as the succession of two diffraction phenomenae : from $\mathcal{A}_1$ to $\mathcal{A}_3$ ($\mathcal{F}_{\alpha_1}$) and from $\mathcal{A}_3$ to $\mathcal{A}_2$ ($\mathcal{F}_{\alpha_2}$), where $\mathcal{A}_3$ is an intermediate spherical segment ; then $\mathcal{F}_\alpha = \mathcal{F}_{\alpha_2} \circ \mathcal{F}_{\alpha_1}$ and $\alpha = \alpha_1 + \alpha_2$ (see Fig. 4).

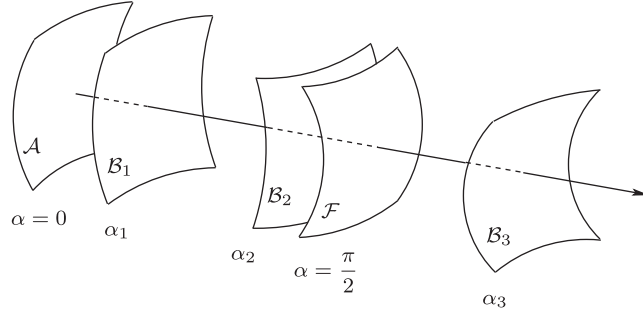


Fig. 3 Various values of α for various receivers: $0 < \alpha_1 < \alpha_2 < \pi/2 < \alpha_3$. The parameter α varies continuously when the receiver moves along the propagation axis.

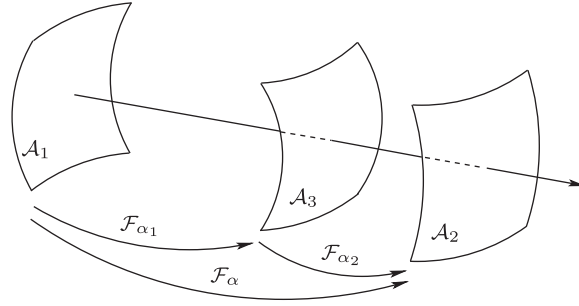


Fig. 4 Illustrating the accordance between the composition law of fractional order Fourier transforms and the Huygens-Fresnel principle: $\alpha = \alpha_1 + \alpha_2$, so that $\mathcal{F}_\alpha = \mathcal{F}_{\alpha_2} \circ \mathcal{F}_{\alpha_1}$.

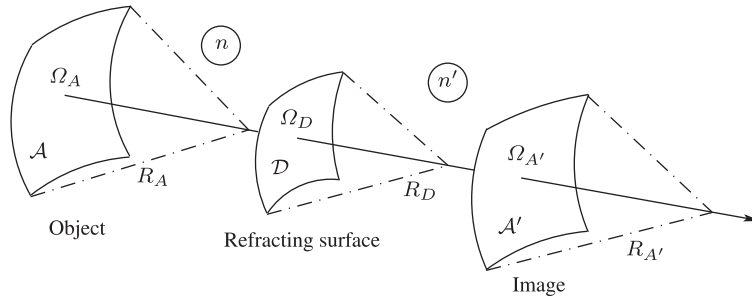


Fig. 5 The refracting surface, separating two propagation media with refractive indices n and n' .

Working With Fractional Orders

The basic idea is that problems in the scalar theory of diffraction can be solved by only manipulating fractional orders without explicitly writing corresponding integral expressions. We illustrate the method by considering imaging through a refracting spherical surface (vertex Ω_D , radius R_D , see Fig. 5). Let $\mathcal{A}$ be the object (spherical emitter), in the object space (with refractive index n), and let $\mathcal{A}'$ be a receiver in the image space (refractive index n'). We denote $d = \overline{\Omega_D \Omega_A}$ and $d' = \overline{\Omega_D \Omega_{A'}}$, as usual in geometrical optics (d is opposite to the distance of propagation from $\mathcal{A}$ to $\mathcal{D}$).

According to the Huygens-Fresnel principle, the field transfer from $\mathcal{A}$ to $\mathcal{A}'$ is the composition of two transfers: from $\mathcal{A}$ to $\mathcal{D}$ and from $\mathcal{D}$ to $\mathcal{A}'$. The corresponding mathematical expression could be obtained by the composition of two integrals that are like the one in Eq. (2). This is a cumbersome method that can be greatly simplified if we use the fractional Fourier transforms associated with light propagation.

The field transfer from $\mathcal{A}$ to $\mathcal{D}$ is described by

$$\left(R_A, R_D, -d, \vec{r}, \vec{s}, U_A, U_D, \lambda \right) \leftrightarrow \left(\alpha, \varepsilon_A, \varepsilon_D, \vec{\rho}, \vec{\sigma}, V_A, V_D \right) \quad (12)$$

and leads to

$$V_D = e^{i\alpha} \mathcal{F}_\alpha[V_A] \quad (13)$$

The field transfer from $\mathcal{D}$ to $\mathcal{A}'$ corresponds to

$$\left(R_D, R_{A'}, d', \vec{s}, \vec{r}', U_D, U_{A'}, \lambda'\right) \leftrightarrow \left(\alpha', \varepsilon'_D, \varepsilon'_A, \vec{\sigma}', \vec{\rho}', V'_D, V'_{A'}\right) \quad (14)$$

so that

$$V'_{A'} = e^{i\alpha'} \mathcal{F}_{\alpha'}[V'_D] \quad (15)$$

The composition of fractional transforms in Eqs. (13) and (15) is possible only if scaled variables on $\mathcal{D}$, $\vec{\sigma}$ and $\vec{\sigma}'$, are identical for both transfers, in order to obtain $V'_D = V_D$. This happens if $\lambda \varepsilon_D = \lambda' \varepsilon'_D$, that is, if

$$\frac{-\lambda d}{R_D - d} \cot \alpha = \frac{\lambda' d'}{R_D - d'} \cot \alpha' \quad (16)$$

For every function f we have $\mathcal{F}_0[f] = f$ and $\mathcal{F}_\pi[f](\vec{\sigma}) = f(-\vec{\sigma})$, so that the field amplitude on $\mathcal{A}'$ is the image of the field amplitude on $\mathcal{A}$ if $\alpha + \alpha' = 0$ or if $\alpha' + \alpha = \pi$ (inverted image). Then $\cot \alpha = -\cot \alpha'$, and since $n\lambda = n'\lambda'$, we deduce from Eq. (16)

$$\frac{n'}{d'} = \frac{n}{d} + \frac{n' - n}{R_D} \quad (17)$$

Eq. (17) is the relation of conjugation of the refracting surface. It has been deduced from a scalar theory of diffraction without writing any integral, but adequately managing parameters of the involved fractional order Fourier transforms.

The imaging between $\mathcal{A}$ and $\mathcal{A}'$ is expressed by $V'_{A'}(\vec{\sigma}') = V_A(\vec{\sigma}')$, if $\alpha + \alpha' = 0$, and by $V'_{A'}(\vec{\sigma}') = V_A(-\vec{\sigma}')$, if $\alpha + \alpha' = \pi$.

We now examine the curvature of the receiver. We denote $q = \Omega_A C_A = d + R_A$ and $q' = \Omega_{A'} C_{A'} = d' + R_{A'}$. Since $\cot \alpha = -\cot \alpha'$, we use Eq. (5) and write

$$\frac{(R_D - d)(R_A + d)}{d(d + R_A - R_D)} = \frac{(d' + R_{A'})(R_D - d')}{d'(d' - R_D + R_{A'})} \quad (18)$$

We use Eq. (16) with $\cot \alpha' = -\cot \alpha$, and deduce from Eq. (18) the relation

$$\frac{q}{n(q - R_D)} = \frac{q'}{n'(q' - R_D)} \quad (19)$$

that is

$$\frac{n'}{q'} = \frac{n}{q} + \frac{n' - n}{R_D} \quad (20)$$

Eq. (20) is the conjugaison relationship written for the curvature centers of $\mathcal{A}$ y $\mathcal{A}'$. We conclude that $\mathcal{A}'$ is the image of $\mathcal{A}$ if, and only if, the vertices of $\mathcal{A}$ and $\mathcal{A}'$ are conjugated and their centers are also conjugated.

Image Formation by a Lens

The use of fractional Fourier transforms in diffraction theory leads us to generalize one of the most important results in modern optics: the field amplitude of an image formed by a lens is the convolution of the object field amplitude with the Fourier transform of the pupil function of the lens. More precisely, we refer to Fig. 6 where a lens images the spherical emitter $\mathcal{A}$ (vertex Ω_A) on the spherical receiver $\mathcal{A}'$ (vertex $\Omega_{A'}$). If g_v denotes the transversal magnification at points Ω_A and $\Omega_{A'}$ we consider first the geometrical image field amplitude U_{GA} on $\mathcal{A}'$, defined by

$$U_{GA}(\vec{r}') = \frac{1}{g_v} U_A\left(\frac{\vec{r}'}{g_v}\right) \quad (21)$$

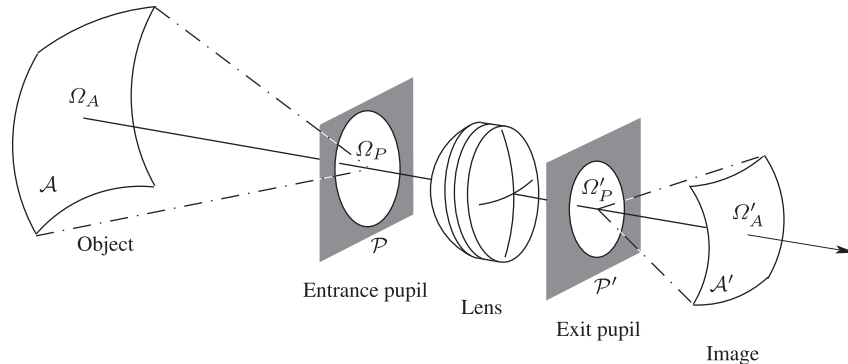


Fig. 6 Image formation by a lens, corresponding to Eq. (23).

We introduce the entrance pupil of the lens, say $\mathcal{P}$, located at Ω_p , and the exit pupil $\mathcal{P}'$ at $\Omega_{p'}$, which is the image of the entrance pupil through the lens. The pupil function of the lens is p such that $p(\vec{s}) = 1$, if the point $\vec{s}$ lays inside the pupil, and $p(\vec{s}) = 0$, if not. Let $d = \overline{\Omega_p \Omega_A}$ and let h be the function such that

$$h(\vec{r}) = \frac{1}{\lambda g_v^2 d^2} \mathcal{F}_{\pi/2}[p]\left(\frac{\vec{r}}{\lambda g_v d}\right) \quad (22)$$

Then the field amplitude on $\mathcal{A}'$ is given by

$$U_{A'} = h * U_{GA} \quad (23)$$

Although very important, Eq. (23) rigorously holds, indeed, only if $\mathcal{A}$ is centered on the entrance pupil $\mathcal{P}$, as shown in Fig. 6 (It can be shown that $\mathcal{A}'$ is centered on the exit pupil.). It does not hold for an arbitrary emitter, not centered on the pupil; nor if the object is the entrance pupil of the lens (whose image is the exit pupil).

By using the fractional Fourier transform representation of diffraction, the previous result can be extended to emitters that are not centered on the entrance pupil. For that purpose we first define the fractional convolution product of functions f and g by

$$f *_{\alpha} g = \mathcal{F}_{-\alpha}[\mathcal{F}_{\alpha}[f] \mathcal{F}_{\alpha}[g]] \quad (24)$$

We have $f *_{\alpha} g = fg$ and $f *_{\pi/2} g = f * g$, where $*$ denotes the standard convolution product.

Let α be the order of the fractional Fourier transform associated with the field transfer from $\mathcal{A}$ to $\mathcal{P}$. We introduce scaled variables and amplitudes V_A on $\mathcal{A}$ and $V_{A'}$ on $\mathcal{A}'$. If p_s is the scaled pupil function and $h_s = \mathcal{F}_{\alpha}[p_s]$ then

$$V_{A'} = h_s *_{\alpha} V_{GA} \quad (25)$$

Eq. (25) is a general expression for image formation by a lens, taking into account the effect of the lens aperture. It holds whatever the position of the object and of its curvature center. If $\mathcal{P}$ is located on the Fourier sphere of $\mathcal{A}$, since $\alpha = \pi/2$, we obtain Eq. (23) once more. If $\mathcal{A}$ is located on the pupil, we obtain $\alpha = 0$ and the result is compatible with the fact that the field on the exit pupil is identical to the field on the entrance pupil (since pupils are conjugated).

Optical Resonator Stability

An optical resonator is made up of two spherical mirrors, say $\mathcal{M}_1$ (radius R_1) and $\mathcal{M}_2$ (radius R_2). Expressing propagation from a mirror to the other by using a fractional order Fourier transform leads us to a complete theory of optical resonators, based on a scalar diffraction theory. Once more, the main interest of the method is deduce results without explicitly writing integrals, but suitably manipulating fractional orders of involved fractional Fourier transforms.

The field transfer from a mirror to the other can be expressed by a fractional Fourier transform whose order α depends on the distance between the mirrors and their radii. This holds true for the reverse transfer, so that a round trip is expressed by the composition of two fractional order Fourier transforms. It can be shown that the orders are the same for both transfers, so that a round trip from $\mathcal{M}_1$ to $\mathcal{M}_1$ is expressed in the form

$$V_1 = e^{-4i\pi D/\lambda} e^{2i\alpha} \mathcal{F}_{2\alpha}[V_1] \quad (26)$$

where V_1 is the scaled field amplitude on $\mathcal{M}_1$ and where the phase factor that was omitted in Eq. (2) has been reintroduced.

A distinction should be made according to whether α is a real or a complex number. If α is real, there is a phase shift in the round trip, and the resonator is said to be stable. If α is a complex number, there is an attenuation of the field amplitude, which is due to diffraction losses; the resonator is said to be unstable.

We state: an optical resonator is stable if, and only if, the field transfer from a mirror to the other is expressed by a fractional Fourier transform whose order is a real number. According to Eqs. (3) and (4) we conclude that an optical resonator is stable if, and only if,

$$0 \leq \left(1 - \frac{D}{R_1}\right) \left(1 - \frac{D}{R_2}\right) \leq 1 \quad (27)$$

(The sign of R_2 in Eq. (27) is opposite to the sign of R_B in Eq. (3), because light propagation direction is changed after reflexion on the corresponding mirror).

Eq. (27) is a very well known condition for a resonator to be stable.

Another consequence of Eq. (26) is that field amplitudes of waves that propagate inside the resonator are eigenfunctions of fractional Fourier transforms (whatever the order). These eigenfunctions are Hermite-Gauss functions, of the form

$$\varphi_{m,n}(\xi, \eta) = H_m(\sqrt{2\pi}\xi) H_n(\sqrt{2\pi}\eta) \exp[-\pi(\xi^2 + \eta^2)] \quad (28)$$

where H_m is the Hermite polynomial of integer order m . Hence the notion of Hermite-Gauss modes of an optical resonator.

Eq. (26) also leads to find longitudinal modes of a resonator. For every integer m , we have

$$\mathcal{F}_{\alpha}[H_m] = e^{2im\alpha} H_m \quad (29)$$

so that for $V_1 = \varphi_{m,n}$ we deduce from Eq. (26)

$$\varphi_{m,n} = e^{-4i\pi D/\lambda} e^{2i\alpha} e^{2i(m+n)\alpha} \varphi_{m,n} \quad (30)$$

and necessarily

$$\frac{2\pi D}{\lambda} - (1 + m + n)\alpha = k\pi \quad (31)$$

where k is an integer. Only waves whose wavelengths satisfy Eq. (31) can propagate in the resonator. They are the longitudinal modes of the resonator.

Gaussian Beams and Gouy Phase

A laser with a stable optical cavity (or resonator) generates Gaussian beams, which are modes of the laser resonator and are described by Hermite-Gauss functions. It can be shown that a Gaussian beam admits a minimal area, called the beam waist and located on a plane (while other wave surfaces are spherical segments). At a distance d from the waist, the Gaussian beam wave surface is a sphere whose curvature radius is R and the field propagation from the beam waist to the previously described surface is expressed by a fractional Fourier transform whose order is α_d with

$$\tan \alpha_d = \frac{\lambda d}{\pi w_0^2} \quad (32)$$

If the phase origine is taken on the waist plane, the field amplitude at a distance d is

$$U_{m,n}(x, y, d) = U_0 \frac{w_0}{w_d} H_m\left(\frac{\sqrt{2}}{w_d} x\right) H_n\left(\frac{\sqrt{2}}{w_d} y\right) \exp\left[-\frac{x^2 + y^2}{w_d^2} - \frac{2i\pi d}{\lambda} + i(m + n + 1)\alpha_d\right] \quad (33)$$

and $(m + n + 1)\alpha_d$ is no more than the Gouy phase.

Uncoupling Variables

A 2-dimensional fractional Fourier transform can be seen like the composition of two 1-dimensional transforms, a fact that can be adequately applied in optics.

We denote by $\mathcal{F}_\alpha^{[1]}$ the 1-dimensional Fourier transform of order α (a real number), defined by

$$\mathcal{F}_\alpha^{[1]}[f](\eta) = \frac{\exp\left[i\left(\frac{\pi}{4}s(\alpha) - \frac{\alpha}{2}\right)\right]}{\sqrt{|\sin \alpha|}} \exp(-i\pi\eta^2 \cot \alpha) \int_{\mathbb{R}} \exp(-i\pi\xi^2 \cot \alpha) \exp\left(\frac{2i\pi\xi\eta}{\sin \alpha}\right) f(\xi) d\xi \quad (34)$$

where $s(\alpha)$ denotes the sign of α .

The Fubini theorem leads us to compose two 1-dimensional fractional order Fourier transforms according to the following diagram, where $\mathcal{F}_{\alpha_1}^{[1]}$ operates on the variable ξ_1 , and $\mathcal{F}_{\alpha_2}^{[1]}$ on ξ_2 ,

$$\begin{array}{ccc} f(\xi_1, \xi_2) & \xrightarrow{\mathcal{F}_{\alpha_1}^{[1]}} & \mathcal{F}_{\alpha_1}^{[1]}[f](\eta_1, \xi_2) \\ \downarrow \mathcal{F}_{\alpha_2}^{[1]} & & \downarrow \mathcal{F}_{\alpha_2}^{[1]} \\ \mathcal{F}_{\alpha_2}^{[1]}[f](\xi_1, \eta_2) & \xrightarrow{\mathcal{F}_{\alpha_1}^{[1]}} & \mathcal{F}_{\alpha_1, \alpha_2}[f](\eta_1, \eta_2) \end{array} \quad (35)$$

where $\mathcal{F}_{\alpha_1, \alpha_2}$ is a 2-dimensional integral. We remark that $\mathcal{F}_{\alpha, \alpha}$ is no more than the 2-dimensional fractional order Fourier transform defined in Eq. (1).

The previous decomposition is useful when dealing with optical emitters and receivers that are segments of tori, that is, have two radii of curvature (see Fig. 7). We consider a toric emitter $\mathcal{A}$ (coordinates x_1 and x_2 , taken along principal sections) and a toric receiver $\mathcal{B}$ (coordinates x'_1 and x'_2) respectively at a distance D . We assume that principal sections of $\mathcal{A}$ and $\mathcal{B}$ are along parallel axes: $x'_1 \parallel x_1$ and $x'_2 \parallel x_2$ (see Fig. 7). Let R_1 and R_2 be the principal radii of $\mathcal{A}$ and R'_1 and R'_2 those of $\mathcal{B}$.

We consider a first 1-dimensional fractional Fourier transform for the transfer from $\mathcal{A}$ to $\mathcal{B}$ but only for the x_1 and x'_1 variables. We define α_1 by

$$\cos^2 \alpha_1 = \left(1 - \frac{D}{R_1}\right) \left(1 + \frac{D}{R'_1}\right) \quad (36)$$

and then

$$\varepsilon_1 = \frac{D}{R_1 - D} \cot \alpha_1, \quad \varepsilon'_1 = \frac{D}{R'_1 + D} \cot \alpha_1 \quad (37)$$

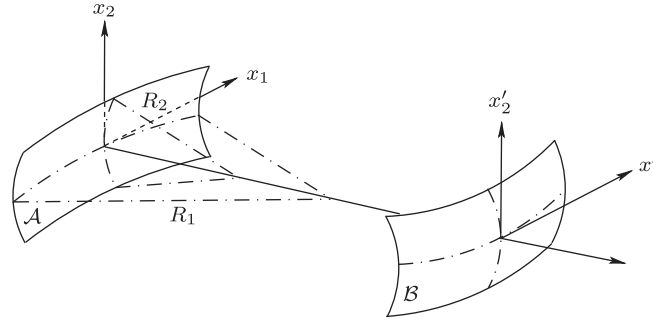


Fig. 7 Describing the field transfer by diffraction-propagation from a toric emitter A to a toric receiver B .

Scaled variables are

$$\xi_1 = \frac{x_1}{\sqrt{\lambda \varepsilon_1 R_1}}, \quad \xi'_1 = \frac{x'_1}{\sqrt{\lambda \varepsilon'_1 R'_1}} \quad (38)$$

and scaled functions

$$V_1(\xi_1, x_2) = \sqrt[4]{\frac{\varepsilon_1 R_1}{\lambda}} U_A\left(\sqrt{\lambda \varepsilon_1 R_1} \xi_1, x_2\right), \quad V'_1(\xi'_1, x_2) = \sqrt[4]{\frac{\varepsilon'_1 R'_1}{\lambda}} U_B\left(\sqrt{\lambda \varepsilon_1 R_1} \xi'_1, x'_2\right) \quad (39)$$

We then obtain

$$V'_1(\xi'_1, x_2) = e^{i x_1/2} \mathcal{F}_{\alpha_1}^{[1]}[V_1](\xi'_1, x_2) \quad (40)$$

Then, we consider a second 1-dimensional Fourier transform for the transfer from A to B but for the x_2 and x'_2 variables. We define $\alpha_2, \varepsilon_2, \varepsilon'_2, \xi_2$ and ξ'_2 as $\alpha_1, \varepsilon_1, \varepsilon'_1, \xi_1$ and ξ'_1 , after changing R_1 into R_2 and R'_1 into R'_2 in Eqs. (36)–(38). Scaled field amplitudes on A and B are

$$V_A(\xi_1, \xi_2) = \sqrt[4]{\frac{\varepsilon_2 R_2}{\lambda}} V_1\left(\xi_1, \sqrt{\lambda \varepsilon_2 R_2} \xi_2\right) \quad (41)$$

$$V_B(\xi'_1, \xi'_2) = \sqrt[4]{\frac{\varepsilon'_2 R'_2}{\lambda}} V'_1\left(\xi'_1, \sqrt{\lambda \varepsilon_2 R_2} \xi'_2\right) \quad (42)$$

so that

$$V_B(\xi'_1, \xi'_2) = e^{i x_2/2} \mathcal{F}_{\alpha_2}^{[1]}[V'_1](\xi'_1, \xi'_2) = e^{i(x_1+x_2)/2} \mathcal{F}_{\alpha_2}^{[1]}\left[\mathcal{F}_{\alpha_1}^{[1]}[V_1]\right](\xi'_1, \xi'_2) = e^{i(x_1+x_2)/2} \mathcal{F}_{\alpha_1, \alpha_2}[V_A](\xi'_1, \xi'_2) \quad (43)$$

Eq. (43) is a generalization of Eq. (10).

The method can be applied to deal with toric (or astigmatic) refractive surfaces and astigmatic lenses. It also provides a method for analyzing propagation and imaging of Gaussian beams with elliptical waists, like those produced by some laser diodes.

See also: Fraunhofer Diffraction. Fresnel Diffraction

Further Reading

- Goodman, J.W., 2005. Introduction to Fourier Optics. Englewood: Robert & Company.
- Namias, V., 1980. The fractional order Fourier transform and its application to quantum mechanics. *Journal of the Institute of Mathematics and its Applications* 25, 241–265.
- Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. The Fractional Fourier Transform, with Applications in Optics and Signal Processing. Chichester: John Wiley & Sons.
- Pellat-Finet, P., 2009. *Optique de Fourier, Théorie métaxiale et fractionnaire*. Paris: Springer.
- Pellat-Finet, P., Fogret, É., 2011. A fractional Fourier transform theory of optical resonators. In: Emersone, P.S. (Ed.), *Progress in Optical Fibers*. New York: Nova Science Publisher, pp. 299–351.
- Pellat-Finet, P., Torres, Y., 1997. Image formation with coherent light: The fractional fourier transform approach. *Journal of Modern Optics* 44, 1581–1594.

Ambiguity Function in Optics

JP Guigay, European Synchrotron Radiation Facility (ESRF), Grenoble, France

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The ambiguity function (AF) associated to a time signal $s(t)$ is defined as the following two-dimensional time-frequency function

$$A(\tau, f) = \int dt s^*(t) s(t - \tau) \exp(-i2\pi ft) \quad (1)$$

It has been introduced by Woodward (1953) in the theory of signal processing of radar or sonar measurements, τ being the delay time related to the distance of the target and f the frequency shift due to its velocity. It bears in its name the idea that it is impossible to perform arbitrarily accurate measurements of both the distance and the velocity of the target. The well-known reason is that the width of a signal in the time domain is inversely proportional to its width in the frequency domain (a narrower signal has a wider spectrum and inversely). From the mathematical point of view, this is just a property of the Fourier transformation; from the physical point of view, this is a common property for waves of any kind and is the basis of the well-known uncertainty relations in quantum physics.

It was shown by Papoulis (1974) that the AF concept can be used in Fourier optics, considering the propagation of a spatially coherent quasi-monochromatic optical field $\psi_z(x)$ transmitted through slit apertures and thin lenses, in the conditions of paraxial propagation in the z -direction; the considered AF is then defined as the following Fourier transform with respect to x :

$$A_z(a, f) = \int dx \psi_z^*(x - a/2) \psi_z(x + a/2) \exp(-i2\pi fx) \quad (2)$$

In the present paper, it will be shown that the concept of AF in optics can be considered as an extension of the spectral analysis of images; the images are two-dimensional intensity patterns $I(x, y)$ which can often be analyzed in a convenient way by considering their Fourier transform (intensity spectrum):

$$\tilde{I}(u, v) = \iint dx dy I(x, y) \exp[-i2\pi(ux + vy)] \quad \text{or} \quad \tilde{I}(\vec{f}) = \int d\vec{x} I(\vec{x}) \exp[-i2\pi\vec{x} \cdot \vec{f}] \quad (3)$$

where two-dimensional vectors are used; the variables (u, v) are the spatial frequencies conjugate to the coordinates (x, y) .

Definition in Terms of the Mutual Intensity

Besides the intensity distribution, the phase correlation between an arbitrary pair of points must be included for a complete description of the optical field. For this purpose, it seems natural to generalize $I(\vec{x})$ by the mutual intensity $\rho(\vec{x}, \vec{x}') = \sqrt{I(\vec{x})I(\vec{x}')} \gamma(\vec{x}, \vec{x}')$, where $\gamma(\vec{x}, \vec{x}')$ is the complex degree of coherence associated to the pair of points $(\vec{x}, \vec{x}')$ (see Born and Wolf (1999)). The mutual intensity is reduced to the usual intensity $I(\vec{x})$ when these two points coincide. The ambiguity is defined (Guigay, 1978), in the frame of Phase-Space Optics (PSO), as the Fourier transform with respect to $\vec{x}$ of the mutual-intensity written as $\rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2)$:

$$A(\vec{f}, \vec{a}) = \int d\vec{x} \exp(-i2\pi\vec{x} \cdot \vec{f}) \rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) \quad (4)$$

$A(\vec{f}, \vec{a})$ is defined in the phase-space which is a four-dimensional space, since $\vec{f}$ and $\vec{a}$ are two-dimensional vectors. It represents a complete description of the partially coherent field optical field. The equivalent representation

$$A(\vec{f}, \vec{a}) = \int d\vec{m} \exp(i2\pi\vec{m} \cdot \vec{a}) \tilde{\rho}(\vec{m} + \vec{f}/2, \vec{m} - \vec{f}/2) \quad (5)$$

is obtained by replacing the mutual-intensity by its Fourier expansion with respect to $\vec{x}$

$$\rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) = \iint d\vec{g} d\vec{h} \exp\{i2\pi[\vec{g} \cdot (\vec{x} + \vec{a}/2) - \vec{h} \cdot (\vec{x} - \vec{a}/2)]\} \tilde{\rho}(\vec{g}, \vec{h}) \quad (6)$$

Formula (4) shows that $A(\vec{f}, 0)$ represents the intensity spectrum defined in (3), which, as well as $I(\vec{x})$, represents the experimental data recorded by a digital detector. Formula (5) shows that $A(0, \vec{a})$ is the inverse Fourier transform of the intensity distribution $\tilde{\rho}(\vec{m}, \vec{m})$ in Fourier space.

A particularly important phase-space function is the Wigner Distribution Function (WDF) which was introduced in optics by Bastiaans (1978) (see Testorf *et al.* (2010) for a precise account of its properties). The WDF is defined as the Fourier transform of $\rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2)$ with respect to $\vec{a}$, instead of $\vec{x}$:

$$W(\vec{x}, \vec{g}) = \int d\vec{a} \exp(-i2\pi\vec{a} \cdot \vec{g}) \rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) \quad (7)$$

It can also be expressed as

$$W(\vec{x}, \vec{g}) = \int d\vec{m} \exp(i2\pi\vec{m} \cdot \vec{x}) \tilde{\rho}(\vec{g} + \vec{m}/2, \vec{g} - \vec{m}/2) \quad (8)$$

The mutual intensity can be obtained from the AF or from the WDF as

$$\rho\left(\vec{x} + \frac{\vec{a}}{2}, \vec{x} - \frac{\vec{a}}{2}\right) = \int d\vec{f} \exp(i2\pi\vec{f} \cdot \vec{x}) A(\vec{f}, \vec{a}) = \int d\vec{g} \exp(i2\pi\vec{g} \cdot \vec{a}) W(\vec{x}, \vec{g}) \quad (9)$$

The AF and the WDF are related to each-other by double Fourier transformation over the position and frequency variables:

$$A(\vec{f}, \vec{a}) = \int d\vec{x} \int d\vec{g} \exp[i2\pi(\vec{a} \cdot \vec{g} - \vec{f} \cdot \vec{x})] W(\vec{x}, \vec{g}) \quad (10)$$

$$W(\vec{x}, \vec{g}) = \int d\vec{a} \int d\vec{f} \exp[i2\pi(\vec{f} \cdot \vec{x} - \vec{a} \cdot \vec{g})] A(\vec{f}, \vec{a}) \quad (11)$$

The property $\rho(\vec{x}, \vec{x}) = (\rho(\vec{x}, \vec{x}))^*$ of the mutual intensity shows that the AF, in general complex, satisfies the relation $A(-\vec{f}, -\vec{a}) = [A(\vec{f}, \vec{a})]^*$. It also shows that the WDF is real (but can be negative as well as positive).

In the exit plane of an object of transmittance $T(\vec{x})$, under uniform coherent illumination, the mutual intensity is equal to the product $T(\vec{x} + \vec{a}/2)T^*(\vec{x} - \vec{a}/2)$ which is often referred to as the “product-space-representation” of the signal $T(\vec{x})$; The corresponding AF and WDF, which are referred to as the AF and the WDF associated to $T(\vec{x})$, are redundant representations of $T(\vec{x})$.

PSO representations are useful tools to characterize the performances of optical systems. They provide elegant approaches to the description and processing of optical signals or images. It has been shown by Nugent (2007) that the concept of AF can be used to unify the various noninterferometric approaches to phase retrieval.

In order to simplify the formulation of the following sections, we shall often consider one-dimensional fields, in which case the two-dimensional vectors are replaced by scalars, without a real lost of generality, because the extension to the general case is usually straightforward.

Intensity Spectrum of a Fresnel Diffraction Pattern Under Coherent Illumination

General Formulation

For simplicity, let us consider a plane wave, of wavelength λ , incident along the z -direction on a thin object of transmittance $T(\vec{x})$ in the plane $z=0$. In the conditions of Fresnel diffraction, the wave-function in a plane $z=D$ is

$$\psi_D(\vec{x}) = |\lambda D|^{-1/2} \exp\left(-i\frac{\pi}{4}\right) \int d\vec{\eta} \exp\left[i\pi \frac{(\vec{\eta} - \vec{x})^2}{\lambda D}\right] T(\vec{\eta}) \quad (12)$$

The corresponding intensity spectrum can be expressed by the multiple integral

$$\tilde{I}_D(\vec{f}) = \int d\vec{x} \exp(-i2\pi\vec{x} \cdot \vec{f}) \iint \frac{d\vec{\eta} d\vec{\eta}'}{\lambda D} \exp\left[i\pi \frac{(\vec{\eta} - \vec{x})^2 - (\vec{\eta}' - \vec{x})^2}{\lambda D}\right] T(\vec{\eta}) T^*(\vec{\eta}')$$

The integration over $\vec{x}$ results in the appearance of a two-dimensional delta-function

$$\int d\vec{x} \exp\left[-i2\pi\vec{x} \cdot \left(\vec{f} + \frac{\vec{\eta} - \vec{\eta}'}{\lambda D}\right)\right] = \lambda D \delta\left(\lambda D \vec{f} + \vec{\eta} - \vec{\eta}'\right) \quad (13)$$

and the intensity spectrum can therefore be reduced to a single integration (Guigay, 1977a,b):

$$\tilde{I}_D(\vec{f}) = \exp(-i\pi\lambda D \vec{f}^2) \int d\vec{\eta} \exp(-i2\pi\vec{f} \cdot \vec{\eta}) T(\vec{\eta}) T^*(\vec{\eta} + \lambda D \vec{f}) \quad (14)$$

This last formula can be expressed in symmetrical form:

$$\tilde{I}_D(\vec{f}) = \int d\vec{x} \exp(-i2\pi\vec{f} \cdot \vec{x}) T\left(\vec{x} - \frac{\lambda D \vec{f}}{2}\right) T^*\left(\vec{x} + \frac{\lambda D \vec{f}}{2}\right) \quad (15)$$

Similar expressions also exist in terms of $\tilde{T}(\vec{f})$:

$$\begin{aligned} \tilde{I}_D(\vec{f}) &= \exp(-i\pi\lambda D \vec{f}^2) \int d\vec{h} \exp(-i2\pi\lambda D \vec{h} \cdot \vec{f}) \tilde{T}(\vec{h} + \vec{f}) \tilde{T}^*(\vec{h}) \\ &= \int d\vec{h} \exp(-i2\pi\lambda D \vec{h} \cdot \vec{f}) \tilde{T}(\vec{h} + \vec{f}/2) \tilde{T}^*(\vec{h} - \vec{f}/2) \end{aligned} \quad (16)$$

It is interesting to note that the AF associated with $T(x)$ is apparent in this formulation if the intensity spectrum is formally considered as a function of f and $a = \lambda Df$.

Derivation of the Transport of Intensity Equation

By linearization of $T\left(\vec{x} - \frac{\lambda D\vec{f}}{2}\right)T^*\left(\vec{x} + \frac{\lambda D\vec{f}}{2}\right)$ as $T(\vec{x})T^*(\vec{x}) + \lambda D\vec{f} \cdot [T\vec{\partial}T^* - T^*\vec{\partial}T]$, where $\vec{\partial}T$ denotes the two-dimensional gradient of the function $T(\vec{x}) = \sqrt{I_0(\vec{x})}\exp[i\phi(\vec{x})]$ written in terms of its modulus and its phase, we obtain the following approximation of formula (15) in the case of small values of $|\lambda D\vec{f}|$:

$$\tilde{I}_D(\vec{f}) = \tilde{I}_0(\vec{f}) - i\lambda D\vec{f} \cdot \int d\vec{x} \exp(-i2\pi\vec{f} \cdot \vec{x}) I_0(\vec{x}) \vec{\partial}(\vec{x})$$

By inverse Fourier transformation of this last formula, we obtain the well-known transport of intensity equation (Teague, 1983):

$$I_D(\vec{x}) = I_0(\vec{x}) - \frac{\lambda D}{2\pi} \vec{\partial} \cdot [I_0(\vec{x}) \vec{\partial}(\vec{x})] \quad (17)$$

which is widely used for phase retrieval.

Application to Simple Objects

This formulation can provide interesting results for some typical Fresnel diffraction patterns; for instance, in the one-dimensional case of a slit of full-width w , we obtain (Guigay, 1977b):

$$\tilde{I}_D(f) = \int_{-(w-|\lambda Df|)/2}^{(w-|\lambda Df|)/2} dx e^{-i2\pi fx} = \frac{\sin[\pi f(w-|\lambda Df|)]}{\pi f} \text{ for } |f| \leq \frac{w}{\lambda D}, \quad 0 \text{ otherwise} \quad (18)$$

Note that this expression is analytically much simpler than the intensity distribution $I(x)$ expressed in terms of Fresnel integrals represented geometrically by the Cornu spiral (see Born and Wolf (1999)).

Contrast Transfer Functions

Considering $T(\vec{x}) = \exp[-B(\vec{x}) + i\phi(\vec{x})]$, where $\exp[-B(\vec{x})]$ and $\phi(\vec{x})$ are the absorption and the phase modulations of the object, we can introduce in formula (11) the approximation

$$T^*\left(\vec{x} + \frac{\lambda D\vec{f}}{2}\right)T\left(\vec{x} - \frac{\lambda D\vec{f}}{2}\right) = 1 - B\left(\vec{x} + \frac{\lambda D\vec{f}}{2}\right) - B\left(\vec{x} - \frac{\lambda D\vec{f}}{2}\right) + i\left[\left(\vec{x} + \frac{\lambda D\vec{f}}{2}\right) - \left(\vec{x} - \frac{\lambda D\vec{f}}{2}\right)\right] \quad (19)$$

which should be valid under the conditions $0B(\vec{x})1$ (weak absorption) and $|\phi(\vec{x}) - \phi(\vec{x} - \lambda D\vec{f})|1$. This last condition is the slowly-varying phase condition (Testorf et al., 2010) which is less restrictive than the weak-phase condition $|\phi(\vec{x})|1$. Under such conditions, the intensity spectrum takes a simple linear form (Guigay, 1977a)

$$\tilde{I}_D(\vec{f}) = \delta(\vec{f}) - 2\cos\left(\pi\lambda D\vec{f}^2\right)\tilde{B}(\vec{f}) + 2\sin\left(\pi\lambda D\vec{f}^2\right)\tilde{\phi}(\vec{f}) \quad (20)$$

The factors of $\tilde{B}(\vec{f})$ and $\tilde{\phi}(\vec{f})$ are named as the absorption-transfer function (ATF) and the phase-transfer function (PTF) respectively.

Formula (20) can be generalized to the case of an imaging system with aberrations others than defocusing. In electron microscopy, for which primary spherical aberration characterized by the coefficient C_s is unavoidable, the following formula is to be used (Hanszen, 1972; Wade, 1974):

$$\tilde{I}_D(\vec{f}) = \delta(\vec{f}) - 2\cos[\omega(\vec{f})]\tilde{B}(\vec{f}) + 2\sin[\omega(\vec{f})]\tilde{\phi}(\vec{f}) \quad (21)$$

$$\text{with } \omega(\vec{f}) = \pi\left(\lambda D\vec{f}^2 + C_s\lambda^3\vec{f}^4/2\right) \quad (22)$$

Partial Talbot Effect

According to the Talbot effect, in the case of a periodic object of period a , a Fresnel diffraction amplitude identical to the object amplitude $T(x)$ is obtained at the so-called Talbot distance $D_T = a^2/\lambda$. It was shown in Guigay (1971) and Arrizón and Ojeda-Castañeda (1992) that the amplitude distribution at any intermediate distance equal to a rational fraction $(n'/n)D_T$ of the Talbot distance is a linear combination of n laterally shifted copies of $T(x)$; this is called the Fractional Talbot effect.

The intensity of any Fresnel diffraction pattern is itself a periodic function of $\vec{x}$ (the letter), also with periodicity a ; according to (15), its Fourier coefficient of order m (m integer) is

$$\tilde{I}_D\left(\frac{m}{a}\right) = \frac{1}{a} \int_{-a/2}^{a/2} dx \exp(-i2\pi x \frac{m}{a}) T^*\left(x + \frac{\lambda Dm}{2a}\right) T\left(x - \frac{\lambda Dm}{2a}\right) \quad (23)$$

This last formula shows that $\tilde{I}_D(pm/a) = \tilde{I}_0(m/a)$, with p being an integer, if the distance D is equal to D_T/m . This result may be named as the “the partial Talbot effect,” because it represents the conservation of a part of the intensity spectrum of the periodic object.

Propagation Through a Paraxial Optical System in Terms of AF

Propagation in Free Space

Let us consider the propagation in free space, with mean direction along the z -axis, of a partially coherent beam. The mutual intensity in the $z=D$ plane is given in terms of the mutual intensity in the $z=0$ plane as:

$$\begin{aligned} \rho_D\left(\vec{x} + \frac{\vec{a}}{2}, \vec{x} - \frac{\vec{a}}{2}\right) &= \frac{1}{\lambda D} \int d\vec{\eta} \exp\left[i\pi \frac{(\vec{\eta} - \vec{x})^2}{\lambda D}\right] \int d\vec{\xi} \exp\left[-i\pi \frac{(\vec{\xi} - \vec{x})^2}{\lambda D}\right] \rho_0\left(\vec{\eta} + \frac{\vec{a}}{2}, \vec{\xi} - \frac{\vec{a}}{2}\right) \\ &= \frac{1}{\lambda D} \int d\vec{\eta} \int d\vec{\xi} \exp\left[i\pi \frac{\eta^2 - \xi^2}{\lambda D} - 2i\pi \vec{x} \cdot \frac{\vec{\eta} - \vec{\xi}}{\lambda D}\right] \rho_0\left(\vec{\eta} + \frac{\vec{a}}{2}, \vec{\xi} - \frac{\vec{a}}{2}\right) \end{aligned}$$

After insertion of this expression in (4), it can be seen that the integration over x results in the delta-function $\lambda D \delta(\vec{\eta} - \vec{\xi} + \lambda D \vec{f})$. The multiple integral is thus reduced to a single integral.

$$\begin{aligned} A_D(\vec{f}, \vec{a}) &= \int d\vec{\eta} \exp(-i2\pi \vec{f} \cdot \vec{\eta}) \rho_0\left(\vec{\eta} + \frac{\vec{a} - \lambda D \vec{f}}{2}, \vec{\eta} - \frac{\vec{a} - \lambda D \vec{f}}{2}\right) \\ \text{Therefore } A_D(\vec{f}, \vec{a}) &= A_0(\vec{f}, \vec{a} - \lambda D \vec{f}) \end{aligned} \quad (24)$$

This represents a translation of amplitude $\lambda D \vec{f}$ of $A_0(\vec{f}, \vec{a})$ on the second variable.

This result can be more readily obtained by recalling that the Fresnel diffraction integral is the convolution of the input function $T(\vec{x})$ by the propagator $G(\vec{x}) = \exp(i\pi \vec{x}^2 / \lambda D - i\pi/4) / \sqrt{\lambda D}$; the output spectrum is therefore the product of the input spectrum $\tilde{T}(f)$ by the spectrum of $G(x)$ which is $\tilde{G}(f) = \exp(-i\pi \lambda D f^2)$; this is translated in terms of mutual intensity as

$$\begin{aligned} \tilde{\rho}_D(\vec{m} + \vec{f}/2, \vec{m} - \vec{f}/2) &= \exp\left\{-i\pi \lambda D \left[(\vec{m} + \vec{f}/2)^2 - (\vec{m} - \vec{f}/2)^2\right]\right\} \tilde{\rho}_0(\vec{m} + \vec{f}/2, \vec{m} - \vec{f}/2) \\ &= \exp(-i2\pi \lambda D \vec{m} \cdot \vec{f}) \tilde{\rho}_0(\vec{m} + \vec{f}/2, \vec{m} - \vec{f}/2) \end{aligned} \quad (25)$$

Inserting this expression in (5), we indeed obtain directly:

$$A_D(\vec{f}, \vec{a}) = \int d\vec{m} \exp[i2\pi \vec{m} \cdot (\vec{a} - \lambda D \vec{f})] \tilde{\rho}_0\left(\vec{m} + \frac{\vec{f}}{2}, \vec{m} - \frac{\vec{f}}{2}\right) = A_0(\vec{f}, \vec{a} - \lambda D \vec{f})$$

As shown in Bastiaans (1978), a similar formula also exists for the WDF:

$$W_D(\vec{x}, \vec{g}) = W_0(\vec{x} - \lambda D \vec{g}, \vec{g}) \quad (26)$$

According to formulas (24) and (26), the AF and the WDF propagate in a uniform medium without a change of their functional forms; only the variables are linearly transformed. This is an elegant representation of Fresnel diffraction phenomena.

Transmission Through a Thin Object

In this case, the incident mutual-intensity $\rho_{inc}(\vec{x}, \vec{x}')$ is multiplied by $T(\vec{x})T^*(\vec{x}')$, where $T(\vec{x})$ is the object transmittance; The AF of the incident beam is then to be convoluted with the object-AF $A_T(\vec{f}, \vec{a})$, as follows

$$A(\vec{f}, \vec{a}) = \int d\vec{h} A_{inc}(\vec{h}, \vec{a}) A_T(\vec{f} - \vec{h}, \vec{a}) \quad (27)$$

where $A_T(\vec{f}, \vec{a}) = \int d\vec{x} \exp(-i2\pi \vec{x} \cdot \vec{f}) T(\vec{x} + \vec{a}/2) T^*(\vec{x} - \vec{a}/2)$. The transmission by a thin lens of focal length F is of special interest; this lens behaves as an object of transmittance $\exp(-i\pi x^2 / \lambda F)$. With $\rho_{inc}(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2)$ being the mutual intensity in the lens entrance surface, the AF in the lens exit surface is calculated as

$$\begin{aligned} A_{exit}(\vec{f}, \vec{a}) &= \int d\vec{x} \exp\left[-i2\pi \vec{x} \cdot \vec{f} - i\pi \frac{(\vec{x} + \vec{a}/2)^2 - (\vec{x} - \vec{a}/2)^2}{\lambda F}\right] \rho_{inc}(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) \\ &= \int d\vec{x} \exp\left[-i2\pi \vec{x} \cdot \left(\vec{f} + \frac{\vec{a}}{\lambda F}\right)\right] \rho_{inc}(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) \\ \text{Therefore } A_{exit}(\vec{f}, \vec{a}) &= A_{inc}\left(\vec{f} + \vec{a}/\lambda F, \vec{a}\right) \end{aligned} \quad (28)$$

This represents a translation of amplitude $(-\vec{a}/\lambda F)$ of $A_0(\vec{f}, \vec{a})$ on the first variable.

The corresponding formula for the WDF is easily found as:

$$W(\vec{x}, \vec{g}) = W_{inc}(\vec{x}, \vec{g} + \vec{x}/\lambda F) \quad (29)$$

Propagation in a Paraxial Optical System

Consider the following process: propagation from an input plane to a thin lens over a distance D_1 , then transmission by this lens of focal length F and finally propagation to the output plane over a distance D_2 . Performing the corresponding transformations successively, according to (24, 28), it is easy to obtain the output AF in terms of the input AF as:

$$A_{out}(\vec{f}, \vec{a}) = A_{in}(\vec{f} - \vec{f}D_2/F + \vec{a}/\lambda F, \vec{a} - \vec{a}D_1/F - \vec{f}\lambda(D_1 + D_2 - D_1D_2/F)) \quad (30)$$

This represents a translation of amplitude $\vec{f}D_2/F - \vec{a}/\lambda F$ on the first variable and a translation of amplitude $\vec{a}D_1/F + \vec{f}\lambda(D_1 + D_2 - D_1D_2/F)$ on the second variable.

This phase-space method is thus a convenient and elegant tool to describe the propagation of a coherent or partially coherent beam in any system comprising coaxial lenses and it is possible to use the WDF instead of the AF. The phase-space method is much simpler than the method based on the propagation of mutual-intensity which would involve convolutions integrals or Fourier transformations.

The AF in Space-Invariant (Isoplanatic) Imaging

The mutual intensity $\rho_{im}(\vec{x}, \vec{x}')$ in the image plane is given in terms of the mutual intensity $\rho_{ob}(\vec{x}, \vec{x}')$ in the object plane (for convenience, the magnification is set equal to 1) as:

$$\rho_{im}(\vec{x}, \vec{x}') = \iint d\vec{\eta} d\vec{\eta}' G(\vec{x} - \vec{\eta}) G^*(\vec{x}' - \vec{\eta}') \rho_{ob}(\vec{\eta}, \vec{\eta}') \quad (31)$$

where $G(\vec{x})$ is the coherent point-spread-function (PSF) of the imaging system. The image-AF is therefore

$$A_{im}(\vec{f}, \vec{a}) = \int d\vec{x} \exp(-i2\pi\vec{x}\cdot\vec{f}) \int d\vec{\eta} G(\vec{x} + \vec{a}/2 - \vec{\eta}) \int d\vec{\eta}' G^*(\vec{x} - \vec{a}/2 - \vec{\eta}') \rho_{ob}(\vec{\eta}, \vec{\eta}') \quad (32)$$

By introducing new variables $\vec{\sigma} = (\vec{\eta} + \vec{\eta}')/2$, $\vec{\tau} = \vec{\eta} - \vec{\eta}'$, $\vec{t} = \vec{x} - \vec{\sigma}$ in this integral expression, we obtain directly (Guigay, 1978; Dutta and Goodman, 1977; Ojeda-Castañeda and Sicre, 1984)

$$\begin{aligned} A_{im}(\vec{f}, \vec{a}) &= \iint d\vec{t} d\vec{\sigma} \exp[-i2\pi\vec{f}\cdot(\vec{t} + \vec{\sigma})] \int d\vec{\tau} G\left(\vec{t} + \frac{\vec{a} - \vec{\tau}}{2}\right) G^*\left(\vec{t} - \frac{\vec{a} - \vec{\tau}}{2}\right) \rho_{ob}\left(\vec{\sigma} + \frac{\vec{\tau}}{2}, \vec{\sigma} - \frac{\vec{\tau}}{2}\right) \\ &= \int d\vec{\tau} A_G(\vec{f}, \vec{a} - \vec{\tau}) A_{ob}(\vec{f}, \vec{\tau}) \end{aligned} \quad (33)$$

This is the convolution integral, with respect to the variable $\vec{a}$, of the AF $A_{ob}(\vec{f}, \vec{\tau})$ in the object plane with the function

$$A_G(\vec{f}, \vec{a}) = \int d\vec{x} \exp(-i2\pi\vec{x}\cdot\vec{f}) G\left(\vec{x} + \frac{\vec{a}}{2}\right) G^*\left(\vec{x} - \frac{\vec{a}}{2}\right) = \int d\vec{g} \exp(i2\pi\vec{g}\cdot\vec{a}) \tilde{G}\left(\vec{g} + \frac{\vec{f}}{2}\right) \tilde{G}^*\left(\vec{g} - \frac{\vec{f}}{2}\right) \quad (34)$$

$\tilde{G}(\vec{g})$ is known as the pupil-function of the imaging system. $A_G(\vec{f}, \vec{a})$ is to be considered equivalently as the AF of the coherent PSF or as the pupil-AF.

The image intensity spectrum, which is a quantity of special interest, is obtained by setting $\vec{a}$ equal to 0 in formula (33):

$$\tilde{I}_{im}(\vec{f}) = A_{im}(\vec{f}, \vec{0}) = \int d\vec{\tau} A_G(\vec{f}, -\vec{\tau}) A_{ob}(\vec{f}, \vec{\tau}) \quad (35)$$

Formula (34) shows that, in the particular case of free-space propagation for which $\tilde{G}(\vec{f}) = \exp(-i\pi\lambda D f^2)$, the pupil-AF is a delta-function $\delta(\vec{a} - \lambda D \vec{f})$. The pupil-AF of an ideal (stigmatic) system is equal to $\delta(\vec{a})$.

The AF of the Image of an Incoherent Source

The AF and the WDF of a point-source represented by the Dirac-function $\delta(\vec{x} - \vec{x}_0)$ are respectively

$$A(\vec{f}, \vec{a}) = \int d\vec{x} \exp(-i2\pi\vec{x}\cdot\vec{f}) \delta(\vec{x} - \vec{x}_0 + \vec{a}/2) \delta(\vec{x} - \vec{x}_0 - \vec{a}/2) = \delta(\vec{a}) \exp(-i2\pi\vec{x}_0\cdot\vec{f}) \quad (36)$$

$$W(\vec{x}, \vec{g}) = \int d\vec{a} \exp(-i2\pi\vec{a}\cdot\vec{g}) \delta(\vec{x} - \vec{x}_0 + \vec{a}/2) \delta(\vec{x} - \vec{x}_0 - \vec{a}/2) = \delta(\vec{x} - \vec{x}_0) \quad (37)$$

The AF and the WDF of an extended incoherent source, considered as a continuous distribution $S(\vec{x}_0)$ of mutually incoherent point-sources, are obtained by summing the AF and the WDF of these mutually incoherent point-sources. They are respectively:

$$A(\vec{f}, \vec{a}) = \int d\vec{x}_0 S(\vec{x}_0) \delta(\vec{a}) \exp(-i2\pi\vec{x}_0 \cdot \vec{f}) = \tilde{S}(\vec{f}) \delta(\vec{a}) \quad (38)$$

$$W(\vec{x}, \vec{g}) = \int d\vec{x}_0 S(\vec{x}_0) \delta(\vec{x} - \vec{x}_0) = S(\vec{x}) \quad (39)$$

Derivation of the Zernike-Van Cittert Theorem from the Propagation of the AF

According to formula (24) and (38), at the distance L from the incoherent source, the AF is therefore $\tilde{S}(\vec{f}) \delta(\vec{a} - \lambda L \vec{f})$ and the mutual-intensity can be obtained from formula (9) as

$$\rho(\vec{x} + \vec{a}/2, \vec{x} - \vec{a}/2) = \int d\vec{f} \exp(i2\pi\vec{x} \cdot \vec{f}) \tilde{S}(\vec{f}) \delta(\vec{a} - \lambda L \vec{f}) = \tilde{S}\left(\frac{\vec{a}}{\lambda L}\right) \exp\left(\frac{i2\pi\vec{x} \cdot \vec{a}}{\lambda L}\right) (\lambda L)^{-1} \quad (40)$$

This result, which is the expression of the Van Cittert-Zernike theorem (Born and Wolf, 1999), can also be obtained from the WDF which is equal to $S(\vec{x} - \lambda L \vec{f})$, since the WDF is easily found to be equal to $S(\vec{x})$ in the plane of the incoherent source.

Partial Coherence Properties in the Image of an Incoherent Source (Guigay, 1978)

The image of an incoherent source (or equivalently in the image of an object under incoherent illumination) by a non-ideal optical system shows some degree of coherence because the light from each point of the primary source is spread over a finite area in the image; it is convenient to introduce the image-AF, which characterizes this image completely, including its coherence properties, and has a simple expression:

$$A_{im}(\vec{f}, \vec{a}) = \int d\vec{\tau} A_G(\vec{f}, \vec{a} - \vec{\tau}) \tilde{S}(\vec{f}) \delta(\vec{\tau}) = A_G(\vec{f}, \vec{a}) \tilde{S}(\vec{f}) \quad (41)$$

The Pupil-AF as a Generalization of the OTF

In the case $\vec{a} = 0$, formula (41) gives the image intensity spectrum as

$$\tilde{I}_{im}(\vec{f}) = A_{im}(\vec{f}, \vec{0}) = A_G(\vec{f}, \vec{0}) \tilde{S}(\vec{f}) \quad (42)$$

This formula shows that $A_G(\vec{f}, \vec{0})$ is identical to the well-known Optical Transfer Function (OTF); a detailed account of the OTF can be found in Born and Wolf (1999). Since formula (41) is obviously a generalization of formula (42), the pupil-AF is to be considered as a generalization of the OTF.

The Pupil-AF and the OTF with Defocusing

According to (24), if the image is recorded at a distance D from its normal position (defocusing), $\vec{a}$ is to be replaced by $\vec{a} - \lambda D \vec{f}$ in (41). This means that the defocused pupil-AF is $A_G(\vec{f}, \vec{a} - \lambda D \vec{f})$. The defocused OTF is consequently $A_G(\vec{f}, -\lambda D \vec{f})$. Therefore, the pupil-AF contains the values of the OTF for any value of the defocusing distance. More precisely, as first pointed out in Brenner *et al.* (1983), the pupil-AF can be seen as a polar display of the OTF for variable defocusing distance: the OTF is displayed, as represented schematically in Fig. 1, along lines going through the origin of coordinates in the $(\vec{f}, \vec{a})$ representation (see chapter 5 of Testorf *et al.* (2010) for details).

This connection between the OTF and the pupil-AF has been used (Ojeda-Castañeda and Berriel-Valdos, 1988; Dowski and Cathey, 1995; Fitzgerald *et al.*, 1997; Castro and Ojeda-Castañeda, 2004; Ojeda-Castañeda *et al.*, 2005; Castro *et al.*, 2006; Ojeda-Castañeda and Gómez-Sarabia, 2015) for designing pupil phase-masks (phase apodizers) which increase the depth of focus without losing lateral resolution and light-gathering power. Furthermore, various effects such as the behaviour of the Strehl ratio and the sensitivity to spherical aberration (Ojeda-Castañeda *et al.*, 1987), or the focal shift (Sheppard and Larkin, 2000), have been studied by considerations based on the pupil-AF. These applications are detailed in chapter 10 of Testorf *et al.* (2010).

Phase-Space Tomography

The idea of phase-space tomography (Raymer *et al.*, 1994; Tu and Tamura, 1998; Dragoman *et al.*, 2002; Tran *et al.*, 2005; Liu and Brenner, 2003; Liu *et al.*, 2008) is to reconstruct the AF (or the WDF) in the plane $z=0$ from a set of intensity measurements $I_D(\vec{x})$ in planes at different distances $z=D_n$. The mutual intensity can then be derived from the reconstructed AF (or WDF) by using formula (9). This is of interest for the characterization of the optical field in the plane $z=0$. If an object of transmittance $T(\vec{x})$ is placed in this plane, we obtain

$$\rho_{inc}\left(\vec{x} + \frac{\vec{a}}{2}, \vec{x} - \frac{\vec{a}}{2}\right) T\left(\vec{x} + \frac{\vec{a}}{2}\right) T^*\left(\vec{x} - \frac{\vec{a}}{2}\right) = \int d\vec{f} \exp(i2\pi\vec{f} \cdot \vec{x}) A(\vec{f}, \vec{a}) \quad (43)$$

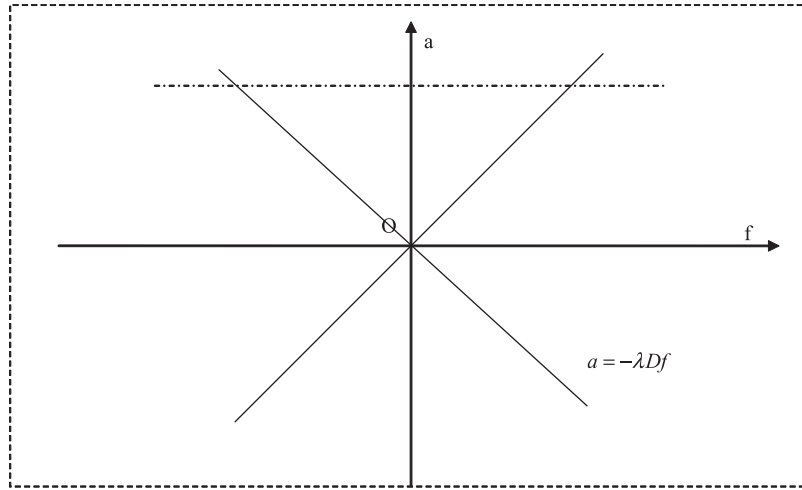


Fig. 1 Schematic PSO representation. The AF along the line $a = -\lambda Df$ correspond to the intensity spectra $\tilde{I}_D(f)$ at distance D from the reference plane. The integration of formula (44) is to be performed along lines parallel to the f -axis.

where ρ_{inc} is the mutual intensity of the incident beam; with $x=a/2$, this is reduced to

$$\rho_{inc}(\vec{a}, \vec{0}) T(\vec{a}) T^*(\vec{0}) = \int d\vec{f} \exp(i\pi \vec{f} \cdot \vec{a}) A(\vec{f}, \vec{a}) \quad (44)$$

As $\rho_{inc}(\vec{a}, \vec{0})$ and the modulus of $T(\vec{0})$ can be measured independently, this last formula allows the determination (up to a constant phase factor) of the complex function $T(\vec{a})$ from the AF.

The WDF tomographic reconstruction is based on the formula

$$I_D(\vec{x}) = \int d\vec{f} W(\vec{x} - \lambda D \vec{f}, \vec{f}) \quad (45)$$

which shows that $I_D(\vec{x})$ is the projection of the WDF in the $(\vec{x}, \vec{f})$ space along a direction which can be varied by the position $z=D$ of the recording plane. The operation which allows the WDF reconstruction is an inverse Radon transform (Testorf *et al.*, 2010; Tu and Tamura, 1998). The feasibility of the tomographic WDF reconstruction has been discussed in Raymer *et al.* (1994), Tu and Tamura (1998), Dragoman *et al.* (2002), and Tran *et al.* (2005).

The AF reconstruction is considered (Tu and Tamura, 1998; Dragoman *et al.*, 2002; Tran *et al.*, 2005; Liu and Brenner, 2003; Liu *et al.*, 2008) to be simpler, because there is no need of inverse Radon transform (the term phase-space tomography may be considered as inappropriate in this case); we only need to perform the Fourier transformation of the measured $I_D(\vec{x})$; the relation $\tilde{I}_D(\vec{f}) = A(\vec{f}, -\lambda D \vec{f})$ shows that the intensity spectra represent the variations of the AF along the radial lines $\vec{a} = -\lambda D \vec{f}$ in the $(\vec{f}, \vec{a})$ space, as depicted schematically in Fig. 1. In order to sample the AF over the complete $(\vec{f}, \vec{a})$ space, it is necessary to use negative, as well as positive, values of the distance D ; this was not possible previously in X-ray optics because appropriate lenses were not available, but this situation may be changing now, because of the development of the so-called compound refractive X-ray lenses (see Simons *et al.* (2017), Terentyev *et al.* (2017) and references therein).

The process of AF reconstruction was first considered in the case of one-dimensional (1-dim) structures (Tu and Tamura, 1998). A 2-dim structure can be considered as an ensemble of 1-dim y -structures $T(x_0, y)$; the corresponding AF are $A(x_0, f_y, a_y)$, from which $T(x_0, y)$ can be derived as a function of y according to (44). For this purpose, a convenient set-up, actually a 1-Dim propagator system, has been proposed by Liu and Brenner (2003) (see also Liu *et al.* (2008)): a cylindrical lens of focal length F produces an exact image (corresponding to no effective propagation) in the x -direction, while there is propagation over the object-image distance in the y -direction; this distance can be varied by using several cylindrical lenses.

A Possible Approach to AF Reconstruction

Let us consider here the case of a one-dimensional object. The AF in the exit plane of an object illuminated by a tilted plane wave of tilting angle α is

$$\begin{aligned} \int dx \exp(-i2\pi f \cdot x) T\left(x + \frac{a}{2}\right) \exp\left[i2\pi \frac{\alpha}{\lambda} \left(x + \frac{a}{2}\right)\right] T^*\left(x - \frac{a}{2}\right) \exp\left[-i2\pi \frac{\alpha}{\lambda} \left(x - \frac{a}{2}\right)\right] \\ = A_{ob}(f, a) \exp\left(i2\pi \frac{\alpha a}{\lambda}\right) \end{aligned} \quad (46)$$

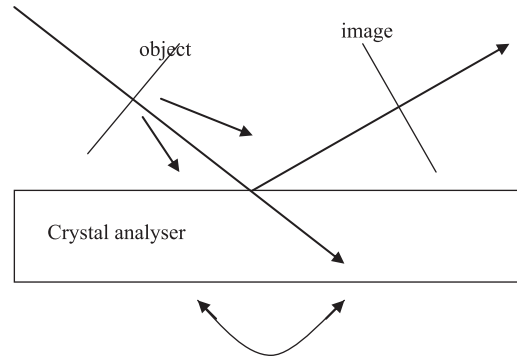


Fig. 2 Principle of the X-ray analyzer-based imaging system. A quasi-parallel and quasi-monochromatic beam is diffracted, after transmission through the object, by a plate of perfect silicon crystal. The arrangement is such that Bragg diffraction occurs, corresponding to reflecting planes parallel to the crystal surface. The Bragg-diffraction process is highly sensitive to the direction of the rays. Images are recorded across the diffracted beam for different angular settings of the crystal.

where $A_{ob}(f, a)$ is the AF associated to the transmittance $T(x)$. Setting $\omega = \alpha/\lambda$, the intensity spectrum of the image delivered by an imaging system with pupil-AF $A_G(f, \tau)$ is

$$\tilde{I}_{im}(f, \omega) = \int d\tau A_G(f, -\tau) A_{ob}(f, \tau) \exp(i2\pi\omega\tau) \quad (47)$$

which is a Fourier transform. Consequently:

$$A_{ob}(f, \tau) A_G(f, -\tau) = \int d\omega \exp(-i2\pi\omega\tau) \tilde{I}_{im}(f, \omega) \quad (48)$$

Supposing $A_G(f, -\tau)$ to be known and $\tilde{I}_{im}(f, \omega)$ to be measured as a function of ω , this last formula provides the possibility to obtain the object-AF.

This approach has been suggested (Guigay *et al.*, 2007b) in the context of X-ray analyzer-based imaging (ABI) which is a Schlieren-type technique (see Fig. 2) based on X-ray Bragg reflection by a perfect crystal plate (analyzer) placed downstream of the object and acting as an angle selector, therefore as a filter in Fourier space; the Fourier transform of the image amplitude is $\tilde{T}(f)\tilde{G}(f)$, where $\tilde{T}(f)$ is the Fourier transform of the object transmittance and $\tilde{G}(f)$ is the complex reflectivity of the crystal for an incident plane wave having angular shift $\delta\theta = \theta - \theta_B = \lambda f$ with respect to the exact Bragg angle θ_B . $\tilde{G}(f)$ is the pupil-function of the imaging system.

The angular range of Bragg reflection is typically a few microradians. For practical reasons, the beam is fixed and the crystal is rotated. When the crystal is rotated by an angle $\delta\theta$ from its peak position in the incident beam, this pupil-function is changed into $\tilde{G}(f - \delta\theta/\lambda)$. It is easily shown from formula (34) that the pupil-AF $A_G(f, a)$ is then changed into $A_G(f, a)\exp(i2\pi a\delta\theta/\lambda)$; we then obtain for the image intensity spectrum the same formula as (47), with $\omega = \delta\theta/\lambda$.

This technique is sensitive to the object structure in one dimension (sensitive to angular deviations in the Bragg diffraction plane). In order to overcome this limitation, it should be just necessary, for each angular position of the crystal-analyzer, to perform a 90-rotation of the object in its plane, in order to obtain finally two-dim information.

Propagation-Based Holographic Phase Retrieval From Several Images

For the sake of simplicity, let us consider the case of a one-dimensional object in the present section.

Fresnel Diffraction Images as In-Line Holograms

The holographic features of Fresnel diffraction images are clearly shown by considering an object transmittance in the form $T(\eta) = 1 + \psi(\eta)$, which is $\tilde{T}(f) = \delta(f) + \tilde{\psi}(f)$ in Fourier space. Formula (16) can be written as:

$$\begin{aligned} \exp(i\pi\lambda D f^2) \tilde{I}_D(f) &= \int dh \exp(-i2\pi\lambda D h f) [\delta(h+f) + \tilde{\psi}(h+f)] [\delta(h) + \tilde{\psi}^*(h)] \\ \exp(i\pi\lambda D f^2) \tilde{I}_D(f) &= \delta(f) + \tilde{\psi}(f) + \exp(2i\pi\lambda D f^2) \tilde{\psi}^*(-f) + \int dh \exp(-i2\pi\lambda D h f) \tilde{\psi}(h+f) \tilde{\psi}^*(h) \end{aligned} \quad (49)$$

This describes the process of holographic reconstruction. The two first terms corresponds to the reconstructed object; the next term corresponds to the out-of-focus image (at distance $2D$) of the conjugate object and the integral term is negligible if $|\psi(x)| \ll 1$ (weak object). The importance of these perturbation terms can be strongly reduced by performing the following summation based on N images recorded at different distances D_n :

$$\frac{1}{N} \sum_{n=1}^N \tilde{I}_{D_n}(f) \exp(i\pi\lambda D_n f^2) = \delta(f) + \tilde{\psi}(f) + \frac{\tilde{\psi}^*(-f)}{N} \sum_{n=1}^N \exp(2i\pi\lambda D_n f^2) + N^{-1} \int dh \tilde{\psi}(h+f) \tilde{\psi}^*(h) \sum_{n=1}^N \exp(-i2\pi\lambda D_n h f) \quad (50)$$

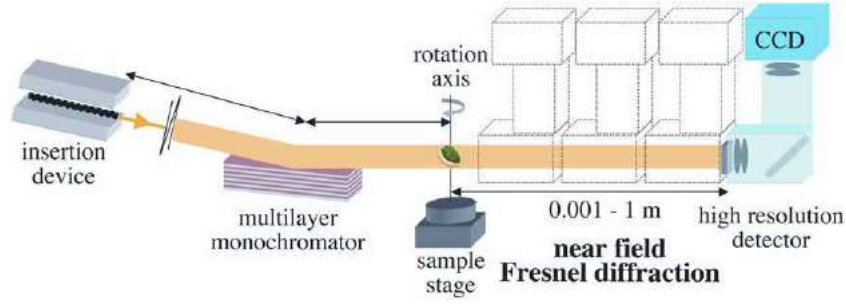


Fig. 3 Scheme of the set-up for X-ray holotomography operated on the ID19 beamline of the ESRF. The Synchrotron X-ray beam is monochromatised by a crystal monochromator or a multilayer, the energy used being typically around 15 keV (wavelength around 0.08 nm). The sample is mounted on a rotating table. The detector ensemble is a scintillator screen coupled by light optics to a CCD camera. This detector can be moved in order to record images close to the object or at different distances from it.

The quantity on the left-hand side may be calculated from digitally recorded images; this allows a good reconstruction of the object if the perturbation terms are nearly cancelled by this summation procedure.

Application to Phase Retrieval and X-ray Holotomography

In the case of a phase object, such that $T(\eta) = \exp[i(\eta)]$, we obtain

$$\tilde{I}_{D_n}(f) = \delta(f) + 2\sin(\pi\lambda D_n f^2)\tilde{\sim}(f) + (NLD)_n \quad (51)$$

Neglecting the nonlinear term (the integral term) denoted as $(NLD)_n$ in the left-hand side of this equation, the following estimation of the phase spectrum (Cloetens *et al.*, 1999) can be obtained, from the experimentally known $\tilde{I}_{D_n}(f)$ by a least-squares fitting as

$$\tilde{\sim}(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2) \tilde{I}_{D_n}(f)}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \quad (52)$$

From this result, it is possible to calculate the nonlinear terms in order to check whether they could indeed be neglected. If necessary, it is possible to take them into account recursively: the calculated $(NLD)_n$ can be subtracted from the experimentally known $\tilde{I}_{D_n}(f)$ to obtain a new estimate of the phase spectrum

$$\tilde{\sim}(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2) [\tilde{I}_{D_n}(f) - NLT_{D_n}]}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \quad (53)$$

and this process can be continued recursively.

If a single image ($N=1$) was used in formula (43), the phase spectrum could not be obtained for the spatial frequencies f such that transfer function $\sin(\pi\lambda D f^2)$ is close to 0. Using several images (typically 4 or 5 images) allows to eliminate this defect and to reduce the influence of the nonlinear terms.

This phase retrieval approach, which has some similarity with the focus variation method used in electron microscopy (Op de Beeck *et al.*, 1996), has been implemented in Synchrotron X-ray optics (see Fig. 3) to provide two-dimensional phase maps, with micrometer resolution, of objects showing a nearly uniform absorption but introducing an important phase modulation; advantage is taken from the high degree of spatial coherence (due to the small lateral size of the source and the long source-specimen-distance) and the good monochromaticity available on modern Synchrotron beamlines using crystal or multilayer monochromators. The phase maps obtained for different orientations of the object are used as input for a tomographic reconstruction of the three-dimensional distribution of the electron density in the sample. This technique named holotomography (Guigay *et al.*, 2007a; Cloetens *et al.*, 1999; Zabler *et al.*, 2005) is of particular interest in the case of objects opaque to visible light. It has been successfully applied to a large variety of objects of interest in biological or material sciences. For a particular recent application, see Mocella *et al.* (2015).

Conclusion

The AF shares with the WDF the ability to describe the propagation of a partially coherent beam in free space and through a paraxial optical system in a simple and elegant way. The AF is a generalization of the intensity spectra which are the basis of important phase retrieval methods. The images given by a space-invariant system are conveniently described in terms of AF of the object and of the pupil-AF which is a generalization of the OTF and has important applications in the design of phase apodizers.

Phase-space tomography is a growing field of research, in which the AF tomographic reconstruction may be more practical than the WDF tomographic reconstruction.

References

- Arrizón, V., Ojeda-Castañeda, J., 1992. Irradiance at Fresnel planes of a phase grating. *J. Opt. Soc. Am. A* 9, 1801–1806.
- Bastiaans, M.J., 1978. The Wigner distribution function applied to optical signals and systems. *Opt. Commun.* 25 (1), 274–278.
- Born, M., Wolf, E., 1999. *Principles of Optics*, seventh ed. Cambridge university press. [Chapter XI].
- Brenner, K.-H., Lohmann, A., Ojeda-Castañeda, J., 1983. The ambiguity function as a display of the OTF. *Opt. Commun.* 44 (5), 323–326.
- Castro, A., Ojeda-Castañeda, J., 2004. Asymmetric phase masks for extended depth of field. *Appl. Opt.* 43 (17), 3474–3479.
- Castro, A., Ojeda-Castañeda, J., Lohmann, A., 2006. Bow-tie effect: Differential operator. *Appl. Opt.* 45 (30), 7878–7884.
- Cloetens, P., Ludwig, W., Baruchel, J., *et al.*, 1999. Holotomography: Quantitative phase tomography with micrometer resolution using hard synchrotron radiation x rays. *Appl. Phys. Lett.* 75, 2912–2914.
- Dowski, E.R., Cathey, W.T., 1995. Extended depth of field through wave-front coding. *Appl. Opt.* 34 (11), 1865.
- Dragoman, D., Dragoman, M., Brenner, K.H., 2002. Tomographic amplitude and phase recovery of vertical-cavity surface-emitting lasers by use of the ambiguity function. *Opt. Lett.* 27 (17), 1519–1521.
- Dutta, K., Goodman, J.W., 1977. Reconstruction of images of partially coherent objects from samples of mutual intensity. *J. Opt. Soc. Am.* 67 (6), 796–803.
- Fitzgerald, A.R., Dowski, E.R., Cathey, W.T., 1997. Defocus transfer function for circularly symmetric pupils. *Appl. Opt.* 36 (23), 5796–5804.
- Guigay, J.P., 1971. On Fresnel diffraction by one-dimensional periodic objects, with application to structure determination of phase objects. *Opt. Acta* 18 (9), 677–682.
- Guigay, J.P., 1977a. Fourier transform analysis of Fresnel diffraction patterns and in-line holograms. *Optik* 49, 121–125.
- Guigay, J.P., 1977b. Analyse spectrale (fréquences spatiales) d'une image de diffraction de Fresnel. *C. R. Acad. Sc. Paris* 284 B, 193–196.
- Guigay, J.P., 1978. The ambiguity function in diffraction and isoplanatic imaging by partially coherent beams. *Opt. Commun.* 26, 136–138.
- Guigay, J.P., Langer, M., Boistel, R., Cloetens, P., 2007a. Mixed transfer function and transport of intensity approach for phase retrieval in the Fresnel region. *Opt. Lett.* 32 (12), 1617–1619.
- Guigay, J.P., Pagot, E., Cloetens, P., 2007b. Fourier Optics approach to X-ray analyser-based imaging. *Opt. Commun.* 270, 180–188.
- Hansen, K.J., 1972. In-line-holographische Erfahrungen mit Radialgittern als Testobjekten in lichtoptischen Modellarrangements für das Elektronenmikroskop. *Optik* 36, 41–54.
- Liu, X., Brenner, K.H., 2003. Reconstruction of two-dimensional complex amplitudes from intensity measurements. *Opt. Commun.* 225, 19–30.
- Liu, X., Hruscha, C., Brenner, K.H., 2008. Efficient reconstruction of two-dimensional complex amplitudes utilizing the redundancy of the ambiguity function. *Appl. Opt.* 47 (22), E1–E7. (Feature issue on Phase-space representations in Optics).
- Mocella, V., Brun, E., Ferrero, C., Delattre, D., 2015. Revealing letters in rolled Herculaneum papyri by X-ray phase contrast imaging. *Nat. Commun.* 5895.
- Nugent, K.A., 2007. X-ray noninterferometric phase imaging: A unified picture. *J. Opt. Soc. Am. A* 24, 536–547.
- Ojeda-Castañeda, J., Andrés, P., Montes, E., 1987. Phase-space representation of the Strehl ratio: Ambiguity function. *J. Opt. Soc. Am. A* 4 (2), 313–317.
- Ojeda-Castañeda, J., Berriel-Valdos, L.R., 1988. Ambiguity function as a design tool for high focal depth. *Appl. Opt.* 27, 790–795.
- Ojeda-Castañeda, J., Gómez-Sarabia, C.M., 2015. Tuning field depth at high resolution by pupil engineering. *Adv. Opt. Photonics* 7 (4).
- Ojeda-Castañeda, J., Landgrave, J.E.A., Escamilla, H.M., 2005. Annular phase-only mask for high focal depth. *Opt. Lett.* 30, 1647–1649.
- Ojeda-Castañeda, J., Sicre, E., 1984. Bilinear systems: Wigner distribution function and ambiguity function representations. *Opt. Acta* 31 (3), 255–260.
- Op de Beeck, M., Van Dyck, D., Coene, W., 1996. Wave-function reconstruction in HRTEM: The parabola method. *Ultramicroscopy* 64, 167–183.
- Papoulis, A., 1974. Ambiguity function in Fourier optics. *J. Opt. Soc. Am.* 64, 779–788.
- Raymer, M.G., Beck, M., McAlister, D.F., 1994. Complex Wave-Field Reconstruction using Phase-Space Tomography. *Phys. Rev. Lett.* 72 (8), 1137–1140.
- Sheppard, C.J.R., Larkin, K.G., 2000. Focal shift, optical transfer function and phase-space representations. *J. Opt. Soc. Am. A* 17 (4), 772–779.
- Simons, H., Ahl, Sonja Rosenlund, Poulsen, H.F., Detlefs, C., 2017. Simulating and optimizing compound refracted-based X-ray microscopes. *J. Synchrotron Radiat.* 24, 392–401.
- Teague, M.R., 1983. Deterministic phase retrieval: A Green's function solution. *J. Opt. Soc. Am.* 73, 1434.
- Terentyev, S., Polikarpov, M., Snigireva, I., *et al.*, 2017. Linear parabolic single-crystal diamond refractive lenses for synchrotron X-ray sources. *J. Synchrotron Radiat.* 24, 103–109.
- Testorf, M., Hennelly, B., Ojeda-Castañeda, J., 2010. *Phase-Space Optics, Fundamentals and Applications*. New York, NY: Mc Graw-Hill.
- Tran, C.Q., Peele, A.G., Roberts, A., *et al.*, 2005. X-ray imaging: A generalized approach using phase-space tomography. *J. Opt. Soc. Am.* 22 (8), 1691–1700.
- Tu, J., Tamura, S., 1997. Wave field determination using tomography of the ambiguity function. *J. Opt. Soc. Am.* 55 (2), 1946–1949. (1998. Analytic relation for recovering the mutual intensity by means of intensity information. *J. Opt. Soc. Am.* 15 (1), 202–206).
- Wade, R.W., 1974. Spectral analysis of holograms and reconstructed images. *Optik* 40, 201.
- Woodward, P.M., 1953. *Probability and Information Theory With Application to Radar*. New York, NY: Pergamon.
- Zabier, S., Cloetens, P., Guigay, J.P., Baruchel, J., Schlenker, M., 2005. Optimization of phase contrast imaging using hard x rays. *Rev. Sci. Instrum.* 76, 073705.

Further Reading

- Brenner, K.H., Ojeda-Castañeda, J., 1984. Ambiguity function and Wigner distribution function applied to partially coherent imagery. *Opt. Acta* 31 (2), 213–223.
- Semichaevsky, A., Testorf, M., 2004. Phase-space interpretation of deterministic phase retrieval. *J. Opt. Soc. Am. A* 21, 2173–2179.

Phase Space Tomography in Optics

Tatiana Alieva and José A Rodrigo, Complutense University of Madrid, Madrid, Spain
Antonio Picón, University of Salamanca, Salamanca, Spain

© 2018 Elsevier Inc. All rights reserved.

Introduction

Tomographic techniques in a wide sense consist in the reconstruction of a n – dimensional (nD) object from a set of data of lower, usually $(n - 1)$, dimension. They are widely applied in medicine and technology for noninvasive inspection of internal structure of macro- and microorganisms and other 3D samples. There exist different approaches developed in optics to obtain this valuable information: computed tomography (CT), optical diffraction tomography, optical coherence tomography, to name the most known ones. Phase space tomography (PST) is intrinsically different from the mentioned tomographic modalities, because it does not directly serve for the tomographic analysis of the sample structure. Meanwhile it is focused on the tomographic study of the light beam itself. From mathematical point of view the PST is closer to the CT technique since both of them are based on the rotation of the object under inspection and measuring a set of its projections. However, the object of reconstruction in the CT and the PST are completely different. In the case of the CT it is the distribution of the absorption coefficient of a physically existed 3D object. While in the case of the PST the object is the Wigner distribution function (WD) which describes coherence properties of classical light (Bastiaans, 2008) or defines a quantum state (density matrix) of light source in quantum systems (D'Ariano *et al.*, 2003). Neither the light coherence characteristics nor density matrix can be measured directly. This information, required for the development of imaging techniques, telecommunication protocols, quantum computing methods, etc., is possible to recover applying the PST tools. The WD is a real function but it can take negative values that impedes its direct measurements. Nevertheless, the projections of the WD are always non-negative and are associated with measurable quantities. This fact together with rotation of the WD after light propagation through certain optical systems described by canonical integral transforms form the basis of the PST (Raymer *et al.*, 1994). Measuring a proper set of the WD projections and using the inverse Radon transform, or other tomography reconstruction algorithm, the WD, and therefore the related beam characteristics are obtained. Below the application of the PST for the analysis of spatial structure of optical beams, short pulses and quantum states are considered.

Light Beam Characterization

Let us consider a linearly polarized optical beam propagated in z direction. The only quantity which can be measured directly is its time-averaged intensity distribution, because the oscillations of the electromagnetic field are too fast for any currently available detectors. However, from the knowledge of the intensity distribution $I(\mathbf{r})$ in a certain plane $z=z_1$ (here we have defined a 2D position vector $\mathbf{r}=[x,y]^t$ at the xy plane transverse to the propagation direction z , where the superscript t denotes the transposition operation) it is impossible to predict the value of $I(\mathbf{r})$ in other plane $z=z_2$ or to describe the beam interaction with a sample under study. In order to do this in the case of quasi-monochromatic coherent beam we need to know its complex field amplitude $f(\mathbf{r})=a(\mathbf{r})\exp[i\varphi(\mathbf{r})]$, where $a(\mathbf{r})$ and $\varphi(\mathbf{r})$ referred to as amplitude and phase are real valued. In this case $I(\mathbf{r})=|f(\mathbf{r})|^2=a^2(\mathbf{r})$ and therefore the amplitude $a(\mathbf{r})$ can be easily measured $a(\mathbf{r})=\sqrt{I(\mathbf{r})}$ using, for example, a digital camera while the determination of the phase $\varphi(\mathbf{r})$ requires some special interferometric or iterative techniques. However, in the most cases light is not fully coherent and its phase is stochastically varying. The description of such partially coherent light requires the application of statistical methods. Often it is assumed that light follows Gaussian statistics and therefore is described by two-point correlation function $\Gamma(\mathbf{r}_1, \mathbf{r}_2)=\langle f(\mathbf{r}_1)f^*(\mathbf{r}_2) \rangle$ known as mutual intensity (MI) (Goodman, 2000). Here $\langle \cdot \rangle$ stands for ensemble averaging which for ergodic process can be substituted by time averaging. The MI is a non-negative definite Hermitian function of four variables $\Gamma(\mathbf{r}_1, \mathbf{r}_2)=\Gamma^*(\mathbf{r}_2, \mathbf{r}_1)$. Note that only the values $\Gamma(\mathbf{r}, \mathbf{r})=\langle |f(\mathbf{r})|^2 \rangle$, corresponding to the intensity distribution, can be directly measured. For the polychromatic light, similar expression known as cross-spectral density, corresponding to the temporal Fourier transform of the mutual coherence function is used for beam description. The MI gauges the field correlation at the points $\mathbf{r}_1$ and $\mathbf{r}_2$ via complex coherence factor (Goodman, 2000): $\gamma(\mathbf{r}_1, \mathbf{r}_2)=\Gamma(\mathbf{r}_1, \mathbf{r}_2)/\sqrt{\Gamma(\mathbf{r}_1, \mathbf{r}_1)\Gamma(\mathbf{r}_2, \mathbf{r}_2)}$.

Instead of the MI, its Fourier transform (FT) with respect to the position difference $\mathbf{r}'=\mathbf{r}_1-\mathbf{r}_2$, known as Wigner distribution (WD) (Bastiaans, 2008)

$$W(\mathbf{r}, \mathbf{p}) = \frac{1}{\sigma^2} \int d\mathbf{r}' \Gamma\left(\mathbf{r} - \frac{\mathbf{r}'}{2}, \mathbf{r} + \frac{\mathbf{r}'}{2}\right) \exp\left(-\frac{i2\pi}{\sigma^2} \mathbf{p}' \mathbf{r}'\right) \quad (1)$$

or the FT with respect to $\mathbf{r}=(\mathbf{r}_1+\mathbf{r}_2)/2$ known as Ambiguity function (AF)

$$A(\mathbf{r}', \mathbf{p}) = \frac{1}{\sigma^2} \int d\mathbf{r} \Gamma\left(\mathbf{r} - \frac{\mathbf{r}'}{2}, \mathbf{r} + \frac{\mathbf{r}'}{2}\right) \exp\left(-\frac{i2\pi}{\sigma^2} \mathbf{p}' \mathbf{r}\right) \quad (2)$$

can be used for beam description. Here $\mathbf{p}=[u,v]^t=\sigma^2\mathbf{k}_\perp$ is a vector proportional to the transverse projection of the spatial frequency vector, $\mathbf{k}_\perp$, and σ is a convenient constant with units of length that is commonly used in Fourier optics. The WD of 2D field is a real valued function of four Cartesian coordinates (x, y, u, v) , which form so called phase space. It follows from the Eqs. (1–2) that the MI can be recovered from the WD (or the AF) applying the inverse FT

$$\begin{aligned}\Gamma(\mathbf{r}_1, \mathbf{r}_2) &= \frac{1}{\sigma^2} \int d\mathbf{p} W\left(\frac{\mathbf{r}_1 + \mathbf{r}_2}{2}, \mathbf{p}\right) \exp\left(\frac{i2\pi}{\sigma^2} \mathbf{p}^t (\mathbf{r}_1 - \mathbf{r}_2)\right) \\ &= \frac{1}{\sigma^2} \int d\mathbf{p} A(\mathbf{r}_1 - \mathbf{r}_2, \mathbf{p}) \exp\left(\frac{i2\pi}{\sigma^2} \mathbf{p}^t (\mathbf{r}_1 + \mathbf{r}_2)\right)\end{aligned}\quad (3)$$

Moreover, if the light is completely coherent, $I(\mathbf{r}_1, \mathbf{r}_2)=1$ and $\Gamma(\mathbf{r}_1, \mathbf{r}_2)=f(\mathbf{r}_1)f^*(\mathbf{r}_2)$, then the phase of the complex field amplitude up to the constant φ_0 can be obtained from the MI

$$\varphi(\mathbf{r}) = \varphi_0 + \arg[\Gamma(\mathbf{r}, \mathbf{r}_0)] \quad (4)$$

where $\mathbf{r}_0$ is a point in which the field intensity is not zero. Note that the phase information is very demanded, for example, in optical metrology and quantitative microscopy.

PST Fundamentals

Actually developed detectors allow measuring only beam intensity distribution and therefore no one of the mentioned functions, MI, WD and AF, can be measured directly: the MI and AF are complex valued functions, while the WD is real but takes non-negative values only in the case of coherent Gaussian beam. Nevertheless, the projection of the WD on the (x, y) – plane corresponds to the beam intensity distribution

$$I(\mathbf{r}) = \frac{1}{\sigma^2} \int d\mathbf{p} W(\mathbf{r}, \mathbf{p}) \quad (5)$$

and therefore is measurable. On the other side the intensity distribution is the FT of the AF slice

$$I(\mathbf{r}) = \frac{1}{\sigma^2} \int d\mathbf{p} A(0, \mathbf{p}) \exp\left(\frac{i2\pi}{\sigma^2} \mathbf{p}^t \mathbf{r}\right) \quad (6)$$

The PST method based on the recovery of the AF is considered, for example, in Cámara *et al.* (2014), Tran *et al.* (2005), and Tu and Tamura (1997).

In order to apply the PST method for beam characterization the WD (or the AF) has to be rotated and a set of corresponding projections have to be measured. In paraxial approximation the beam propagation through a first-order optical system – an optical system composed by centered spherical and cylindrical lenses and/or mirrors separated by homogeneous medium – is described by the 2D linear canonical integral transform $f_T(\mathbf{r}_o) = \int K_T(\mathbf{r}_i, \mathbf{r}_o) f(\mathbf{r}_i) d\mathbf{r}_i$ where $f(\mathbf{r}_i)$ and $f_T(\mathbf{r}_o)$ are the complex field amplitude in the input and output planes of the system and

$$K_T(\mathbf{r}_i, \mathbf{r}_o) = \frac{1}{\sigma^2 \sqrt{\det(i\mathbf{B})}} \exp\left[\frac{i\pi}{\sigma^2} (\mathbf{r}_o^t \mathbf{D} \mathbf{B}^{-1} \mathbf{r}_o - 2\mathbf{r}_i^t \mathbf{B}^{-1} \mathbf{r}_o + \mathbf{r}_i^t \mathbf{B}^{-1} \mathbf{A} \mathbf{r}_i)\right] \quad (7)$$

for $\det \mathbf{B} \neq 0$ and

$$K_T(\mathbf{r}_i, \mathbf{r}_o) = \frac{1}{\sqrt{|\det \mathbf{A}|}} \exp\left[\frac{i\pi}{\sigma^2} \mathbf{r}_o^t \mathbf{C} \mathbf{A}^{-1} \mathbf{r}_o\right] \delta(\mathbf{r}_i - \mathbf{A}^{-1} \mathbf{r}_o) \quad (8)$$

for $\mathbf{B}=\mathbf{0}$. Here $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are 2×2 sub-matrices of the 4×4 dimensionless ray transformation matrix $\mathbf{T}$, which provides the relation between the position $\mathbf{r}$ and direction (proportional to $\mathbf{p}$) of the ray in the input ($\mathbf{r}_i$ and $\mathbf{p}_i$) and output ($\mathbf{r}_o$ and $\mathbf{p}_o$) planes of an optical system:

$$\begin{bmatrix} \mathbf{r}_o \\ \mathbf{p}_o \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{bmatrix} = \mathbf{T} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{bmatrix} \quad (9)$$

Note, that for the matrix describing the linear ray propagation in free space

$$\mathbf{T}_z = \begin{bmatrix} \mathbf{I} & \frac{z}{\lambda} \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (10)$$

where $\mathbf{I}$ is a unity matrix, λ is the wavelength and z is the distance between the input and output planes, the expression Eq. (7) with $\sigma=\lambda$ is reduced to the kernel of the Fresnel transform. Meanwhile the matrix

$$\mathbf{T}_L = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{C} & \mathbf{I} \end{bmatrix} \quad (11)$$

corresponds to spherical/cylindrical thin lens or mirror action yielding to quadratic phase modulation of the wavefield according to the Eq. (8) with $\sigma=\lambda$. For example, the sub-matrix $\mathbf{C} = \begin{bmatrix} -\lambda f_x^{-1} & 0 & 0 & -\lambda f_y^{-1} \end{bmatrix}$ describes the orthogonal superposition of two

convergent cylindrical lenses with front focal distances f_x and f_y , that for $f_x = f_y$ corresponds to the spherical lens. Two assembled cylindrical lenses with transverse axes forming any other angle, called *generalized lens*, is associated with symmetric sub-matrix $C = [c_{xx}, c_{xy}; c_{xy}, c_{yy}]$. Using a set of cylindrical lenses and choosing appropriate distances between them optical system performing any 2D canonical integral transform can be constructed.

The WD, $W_T(\mathbf{r}, \mathbf{p})$, of the beam at the output plane of a first-order system described by the ray transformation matrix $T = [A, B; C, D]$, is related to the WD at the input plane, $W(\mathbf{r}, \mathbf{p})$, by an affine transformation (Bastiaans, 2008)

$$W_T(\mathbf{r}, \mathbf{p}) = W(D^t \mathbf{r} - B^t \mathbf{p}, -C^t \mathbf{r} + A^t \mathbf{p}) \quad (12)$$

which, in general, includes rotation, shearing and scaling in phase space. Similar expression is valid for the AF

$$A_T(\mathbf{r}, \mathbf{p}) = A(D^t \mathbf{r} - B^t \mathbf{p}, -C^t \mathbf{r} + A^t \mathbf{p}) \quad (13)$$

In particular, the rotations in phase space, which are described by the ray transformation matrix T_R with parameters $A = D$ and $B = -C$, are needed to apply the PST method. There are two principle kind phase space rotators: the fractional Fourier transform (FrFT) (Ozaktas et al., 2001) associated with the matrix $T_{FrFT}(\alpha_x, \alpha_y)$ defined by

$$\begin{aligned} A_{FrFT} &= D_{FrFT} = \begin{bmatrix} \cos \alpha_x & 0 \\ 0 & \cos \alpha_y \end{bmatrix} \\ B_{FrFT} &= -C_{FrFT} = \begin{bmatrix} \sin \alpha_x & 0 \\ 0 & \sin \alpha_y \end{bmatrix} \end{aligned} \quad (14)$$

which describes the WD rotation in the planes xu and yv for angles $\alpha_{x,y}$ respectively, and image rotator (IR) associated with the matrix $T_{IR}(\beta)$ defined by

$$A_{IR} = D_{IR} = \begin{bmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{bmatrix} \text{ and } B_{IR} = C_{IR} = 0 \quad (15)$$

which corresponds to the simultaneous WD rotation in xy and uv for angle β . Their combination defines any other phase space rotator associated with $T_R(\gamma, \alpha_x, \alpha_y, \beta) = T_{IR}(\gamma) T_{FrFT}(\alpha_x, \alpha_y) T_{IR}(\beta)$. Note that for the PST applications the parameter γ is irrelevant since it describes the rotation of the coordinate system in the output xy plane. The intensity distribution in the output plane of a phase space rotator corresponds to the WD projection given by

$$I_{T_R}(\mathbf{r}) = \langle |f_{T_R}(\mathbf{r})|^2 \rangle = \frac{1}{\sigma^2} \int d\mathbf{p} W_{T_R}(\mathbf{r}, \mathbf{p}) = \frac{1}{\sigma^2} \int d\mathbf{p} W(D_R^t \mathbf{r} - B_R^t \mathbf{p}, -C_R^t \mathbf{r} + A_R^t \mathbf{p}) \quad (16)$$

For complete WD recovery using the inverse Radon transform, or other tomography reconstruction algorithm a set of 2D projections $\{I_{T_R}(\mathbf{r})\}$ for two independent angles (both covering a π -interval) associated with the rotation in two orthogonal planes of the phase space is needed. Initially the PST method was established for the projection set corresponding to the FRFT, $T_R(0, \alpha_x, \alpha_y, 0) = T_{FrFT}(\alpha_x, \alpha_y)$ (Raymer et al., 1994) while in Cámara (2015) and Cámara et al. (2013) it has been shown that the phase space rotator described by $T_R(0, 0, \alpha, \beta)$ or by $T_R(0, \alpha, 0, \beta)$ with

$$\begin{aligned} A_R(0, \alpha, 0, \beta) &= \begin{bmatrix} \cos \alpha \cos \beta & \cos \alpha \sin \beta \\ -\sin \beta & \cos \beta \end{bmatrix} \\ B_R(0, \alpha, 0, \beta) &= \begin{bmatrix} \sin \alpha \cos \beta & \sin \alpha \sin \beta \\ -\sin \beta & \cos \beta \end{bmatrix} \end{aligned} \quad (17)$$

can be applied.

While the theoretical fundamentals of the PST method seem straightforward its experimental realization is rather difficult since real-world applications require fast acquisition and processing of the experimental data together with comprehensive analysis of the resulting four-dimensional functions associated to the light coherence information. Any a priori information about a studied beam can simplify the task and decrease the number of projections required for WD reconstruction. Some methods assume certain hypothesis about field model (Tian et al., 2012), its symmetry (Agarwal and Simon, 2000; Alieva et al., 2011; Cámara, 2015; Cámara et al., 2014) or coherence homogeneity. Below two schemes for practical realization of the PST analysis of the beam coherence state in the absence of a priori information are considered.

PST for 1D Beam Characterization

The application of the PST for beam analysis meets inherent difficulty caused by the high dimension of the WD and related functions. It yields to the excessive computation effort to process the measured data (WD projections) into the requested result (the WD, AF or MI). One of solutions of this complex problem is to decompose a 2D beam into 1D signals and recover their WDs. In order to select a 1D signal a thin slit is usually used. This mask has to be moved and rotated in the xy input plane in order to

explore consequently the beam coherence properties along the corresponding lines. The WD of a selected 1D signal described by the complex field amplitude $f(x)$ is a 2D function, $W(x, u)$, that can be easily presented graphically and analyzed. In the associated 2D phase space xu there is only one phase space rotator: the 1D FrFT whose the kernel is written as

$$K_\alpha(x_i, x) = \frac{1}{\sigma \sqrt{i \sin \alpha}} \exp \left[\frac{i\pi (x^2 + x_i^2) \cos \alpha - 2xx_i}{\sin \alpha} \right] \quad (18)$$

The 1D FrFT produces the rotation of the WD

$$W_\alpha(x, u) = W(x \cos \alpha - u \sin \alpha, x \sin \alpha + u \cos \alpha) \quad (19)$$

Measuring the WD projections

$$P_\alpha(x) = \frac{1}{\sigma} \int du W(x \cos \alpha - u \sin \alpha, x \sin \alpha + u \cos \alpha) \quad (20)$$

for angles covered a π -interval $\alpha \in [0, \pi]$ which corresponds to the Radon transform of the WD (also called as Radon-Wigner transform) the $W(x, u)$ can be reconstructed applying, for example, the filtered back-propagation algorithm

$$W(x, u) = \frac{1}{\sigma^2} \int_0^\pi \int dx d\rho |P_\alpha(\rho)| \exp \left[-\frac{2\pi}{\sigma^2} \rho (x \cos \alpha - u \sin \alpha) \right] \quad (21)$$

If the projections are available for the interval $\alpha \in [\alpha_0, \alpha_0 + \pi]$ then the WD, $W(x \cos \alpha_0 - u \sin \alpha_0, x \sin \alpha_0 + u \cos \alpha_0)$ reconstructed by this expression (the integration limits in this case are from α_0 to $\alpha_0 + \pi$) is rotated at angle α_0 with respect to one, $W(x, u)$, described the input signal. The MI is obtained correspondingly as

$$\Gamma(x_1, x_2) = \frac{1}{\sigma} \int du W\left(\frac{x_1 + x_2}{2}, u\right) \exp\left(\frac{i2\pi}{\sigma^2} u(x_1 - x_2)\right) \quad (22)$$

The principal advantage of this method is a possibility to registered in one shot all required projections of the $W(x, u)$ in the digital camera display. For this purpose a 1D signal is converted into a 2D one, which is its multiple-copy and varifocal lenses or Fresnel zone plates are applied to perform the FrFT of every copy for different angle. Several designs for such system, known as Radon-Wigner display (RWD), have been proposed (Cámara *et al.*, 2011; Granieri *et al.*, 1997; Mendlovic *et al.*, 1996). One of them is discussed below. While this method is suitable for analysis of 1D optical signals it has important drawbacks when it is applied for the 2D beams. The slit mask application introduces artifacts. The conversion of 1D signal into a 2D one produces significant power reduction of the analyzed field and therefore decreases the signal to noise ratio. For analysis of the beam coherence properties along another line of xy plane the displacement and/or rotation of the slit mask and the RWD are required.

PST for 2D Beam Characterization

Another approach for practical application of the PST is based on an appropriate choice of the projections set and the order of their acquisition. As it was mentioned above several projection sets can be used for the reconstruction of the 4D-WD. As well as for the tomographic analysis of horizontal slices of the 3D object it is beneficial rotate the object along the vertical axes, there exists the 4D-WD projection set which is more appropriate for the beam coherence analysis. This set, hereinafter referred as $\{P_{\alpha, \beta}(\mathbf{r})\}$, is obtained by rotating the WD in the xy and xu planes for angles β and α , respectively (phase space rotator described by the ray transformation matrix $T_R(0, \alpha, 0, \beta)$) (Cámara, 2015; Cámara *et al.*, 2013).

The projection set $\{P_{\alpha, \beta}(\mathbf{r})\}$, comprises several subsets, $\{P_{\alpha, \beta_0}(\mathbf{r})\}$, defined by fixing $\beta = \beta_0$, which are acquired consecutively. Each subset provides the information about field correlation at whichever points $\mathbf{r}_1$ and $\mathbf{r}_2$ that are contained in a line which forms an angle β_0 with axes x . The MI at such a line is defined as $\Gamma_{\beta_0}(\mathbf{r}_0, s) = \Gamma(\mathbf{r}_0, \mathbf{r}_0 + s\mathbf{n})$, where $\mathbf{r}_0$ is the reference point, $\mathbf{n} = [\cos \beta_0, \sin \beta_0]^T$ and s are the direction and running coordinate of the line, respectively. For example, from one slice of the projection subset $\{P_{\alpha, 0}([x, y_0]^T)\}$ for any value of y_0 (see Fig. 1) the 2D-WD, $W(x, u; y_0)$, of the 1D signal $f(x; y_0)$ is obtained using the Eq. (21), where

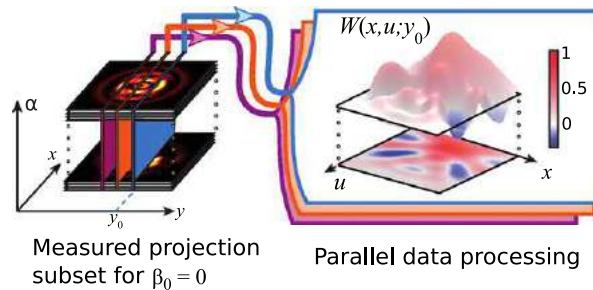


Fig. 1 Scheme explaining the parallel data processing for WD reconstruction. The Wigner distribution $W(x, u; y_0)$ of the field along a chosen line $y = y_0$ is reconstructed for each projection subset slice defined by both β_0 and y_0 (see, e.g., the blue slice). Note that the Wigner distribution takes positive and negative values (white color indicates zero level).

$P_\alpha(\rho) = P_{\alpha,0}([\rho, \gamma_0]^t)$. The correlation function for the field at reference point $\mathbf{r}_0 = [x_0, \gamma_0]^t = \mathbf{r}_1$ and any other point $\mathbf{r}_2 = \mathbf{r}_0 + s[1, 0]^t$ contained in the horizontal line $\gamma = \gamma_0$ is then calculated as

$$\Gamma_0(\mathbf{r}_0, s) = \frac{1}{\sigma} \int du W(x_0 + s/2, u; \gamma_0) \exp\left(\frac{i2\pi}{\sigma^2} su\right) \quad (23)$$

Note that from the same subset $\{P_{\alpha,0}([x, \gamma]^t)\}$ the MI for any two points contained in the same horizontal line ($\gamma = \text{constant}$) is obtained. To find the coherence relations for two points with different γ coordinates, one has to perform the rotation in the xy plane for the corresponding angle β_0 and to repeat the procedure described above using the projection subset $\{P_{\alpha,\beta_0}(\mathbf{r})\}$. Doing this, we recover the desired information about field correlation avoiding the reconstruction of the entire 4D-WD and its posterior processing as it needed, for example, if the projection set corresponding to 2D-FrFT ($T_R(0, \alpha_x, \alpha_y, 0) = T_{FrFT}(\alpha_x, \alpha_y)$) is used. Although the amount of data required for full beam characterization is the same as in other realizations of the PST, this choice of projection set, $\{P_{\alpha,\beta}(\mathbf{r})\}$, brings powerful benefits: (1) The data acquisition and processing tasks are performed simultaneously since every projection subset is an independent entity. (2) As one projection subset slice is independent of the rest, parallel computing can be used for the data processing. (3) Physically meaningful information is obtained from a reduced number of projections that allows starting the beam analysis before the complete projection set is acquired.

Optical Setups for PST

Several experimental phase space rotators described by ray transformation matrix T_R have been proposed. All of them are based on a combination of lens (or other optical elements producing quadratic phase modulation of wavefield) separated by the free space intervals. Since only the intensity distribution is measured then the phase in the output plane of the optical system applied for the PST can be arbitrary. This allows increasing a number of suitable setups and often simplifying their design. Indeed the intensity distributions in the output of the systems described by the ray transformation matrices $T_L T_R$ and T_R coincide and depend only on T_R . The scanning of the rotation angles is achieved by modification of the lens powers or distances between them. However, the insert of a lens or change of the distance usually require the system alignment that is incompatible for a fast projection acquisition required for the practical application of the PST. Moreover, for accurate WD reconstruction it is recommended to avoid the projection re-scaling. All these conditions can be fulfilled if digital lenses implemented by spatial light modulator (SLM) are used. Indeed the phase-only SLM easily performs a quadratic phase wavefront modulation that is an action of a thin lens. This allows designing an optical setup with all elements located in fixed position and a programmable control of the projection parameters. In this case the speed of the projection acquisition is limited by the SLM and digital camera (usually video rate).

Different projections sets including ones discussed above (corresponding to $T_R(0, \alpha_x, \alpha_y, 0)$ and $T_R(0, \alpha, 0, \beta)$) can be automatically acquired using the programmable optical setup (Cámara *et al.*, 2013; Rodrigo *et al.*, 2009) shown in Fig. 2. The setup comprises two generalized lenses implemented by two SLMs and a digital camera. The distance z between every two consecutive elements is fixed. Specifically, for acquisition the projection set $\{P_{\alpha,\beta}(\mathbf{r})\}$ these lenses L_j ($j=1,2$) have the following transmission functions

$$L_j(x, y) = \exp\left[-i\pi \frac{(x \cos \beta + y \sin \beta)^2}{\lambda f_j}\right] \exp\left[-i\pi \frac{(-x \sin \beta + y \cos \beta)^2}{\lambda g_j}\right] \quad (24)$$

where the focal lengths are $f_1 = 2z/(2 - \cot(\alpha/2))$, $f_2 = z/(2 - 2 \sin \alpha)$ and $g_1 = z$, $g_2 = z/2$, are given as a function of the transformation angle $\alpha \in [\pi/2, 3\pi/2]$. This setup is described by the matrix $T_L T_R(-\beta, \alpha, 0, \beta)$. Then the intensity distribution in its output plane is equal to one corresponding to the desired transformation $T_R(0, \alpha, 0, \beta)$ except for a rotation at an angle $-\beta$. To compensate this effect and obtain the corresponding WD projection the image acquired by the digital camera has to be rotated an angle β that can be easily performed digitally. The number of the acquired projections which is one of the principle parameters defining the resolution of the WD reconstruction can easily reach a value 180 per one subset $\{P_{\alpha,\beta_0}(\mathbf{r})\}$.

The varifocal lens needed for RWD implementation can also be realized by the SLM (Cámara, 2015; Cámara *et al.*, 2011). An illustration of such system is displayed in Fig. 3. It consists in two SLMs separated a fixed distance z , and two cylindrical lenses of

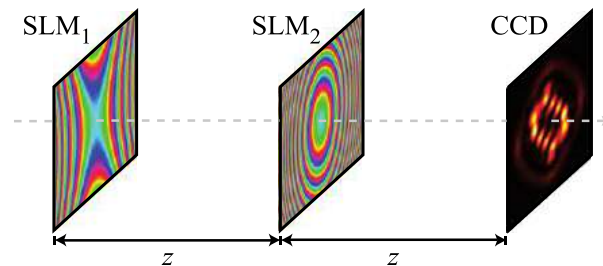


Fig. 2 Scheme of a programmable setup for complete WD projection set acquisition, required for the WD reconstruction. For every β and α the SLM₁ and SLM₂ implement generalized lenses L_1 and L_2 , respectively, given by Eq. (24).

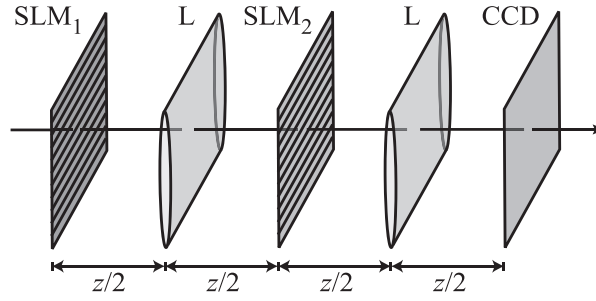


Fig. 3 Scheme of a programmable Radon Wigner display using spatial light modulators. SLMs implement N 1D lenses described by Eqs. (25)–(26).

focal length $z/4$ with phase modulation in the y direction also separated a fixed distance z . N copies of the input 1D signal $f(x)$ arranged in set of N vertical channels are projected into input plane of the RWD, which coincides with the position of the first SLM. The SLMs implement the multichannel phase masks. Each channel of the phase mask is programmed to perform the 1D-FrFT operation for an arbitrary angle in the range $\alpha \in [\pi/2, 3\pi/2]$. In order to acquire the full-range WD projections for N equidistant angles, the n th channel of the j th SLM implements the transmission function $L_j(x, n)$ given by:

$$L_1(x, n) = \exp \left[-i\pi \frac{x^2}{\lambda z} \left(1 - \frac{1}{2} \cot \frac{\alpha_n}{2} \right) \right] \quad (25)$$

$$L_2(x, n) = \exp \left[-i2\pi \frac{x^2}{\lambda z} (1 - \sin \alpha_{N-n+1}) \right] \quad (26)$$

where $n = 1, \dots, N$ and $\alpha_n = \pi/2 + (n-1)/N\pi$ is the FrFT angle associated with the j th channel. While the SLMs perform the FrFT in the x direction, the glass cylindrical lenses compensate the beam propagation in the y direction by imaging the first SLM into the second, and the second into the output plane. Since the first cylindrical lens inverts the channel position, the n th channel of the second SLM implements the angle labeled by $N - n + 1$ in Eq. (26). At the output plane, placed at a distance $z/2$ from the last cylindrical lens, a digital camera registers in one shot all required WD projections of the input 1D signal. The number of channels and therefore the WD projections depend on the size of the SLM and camera displays and can easily reach $N=300$ if the 2D expansion of the 1D signal is energetically resolved.

Chronocyclic Tomography for Optical Pulse Characterization

We have considered the application of the phase space tomography for characterization of the spatial light coherence, meanwhile it also can be used for the analysis of short optical pulses (Beck *et al.*, 1993; Dorrer and Kang, 2003). In this case the time t and temporal frequency ν are coordinates of the phase space and the method of the corresponding WD recovery is called chronocyclic tomography. The electrical field of an optical pulse is described by

$$E(t) = |E(t)| \exp[i\phi_t(t)] \exp(-i2\pi\nu_0 t) \quad (27)$$

where $|E(t)|$ and $\phi_t(t)$ are the time-dependent pulse envelop and phase correspondingly and ν_0 is the carrier frequency. The instantaneous power $|E(t)|^2$ can be measured only if the pulse duration is sufficiently larger than the response time of the photodetector. Alternatively the pulse can be described in frequency domain as $\tilde{E}(\nu) = |\tilde{E}(\nu)| \exp[i\phi_\nu(\nu)]$, where $\tilde{E}(\nu)$ is a FT of the $E(t)$, $\phi_\nu(\nu)$ is a spectral phase. The spectral density $|\tilde{E}(\nu)|^2$ can be easily measured by a spectrometer. Usually an ensemble of pulses are involved in the experiments and its statistical properties, in particular, two-time correlation function $\Gamma(t_1, t_2) = \langle E(t_1) E^*(t_2) \rangle$ or the related WD

$$W(t, \nu) = \eta \int dt' \Gamma(t - t'/2, t + t'/2) \exp(-i2\pi t' \nu) \quad (28)$$

are of interest. Here η is a convenient constant with units of frequency. If all pulses are identical then the amplitude and the phase of $E(t)$ up to the constant factor can be retrieved from $\Gamma(t_1, t_2)$.

As well as in the spatial PST case the devices which produce the rotation of the WD, $W(t, \nu)$, are needed for chronocyclic tomography. A quadratic temporal phase modulator and the propagation in dispersive medium or two-grating compressor (quadratic spectral phase modulation) are the analogues of lens and Fresnel propagator, respectively. After pulse propagation through a temporal fractional FT system constructed by such devices, whose parameters control the rotation angle α , the required number of the WD projections $P_\alpha(\nu) = \eta \int W(t\eta^2 \cos \alpha + \nu \sin \alpha, \nu \cos \alpha - t\eta^2 \sin \alpha) dt$ can be registered by a spectrometer. Then the WD (and therefore the two-point correlation function $\Gamma(t_1, t_2)$) is reconstructed applying, for example, the inverse Radon transform.

PST for Quantum Light Characterization

In quantum optics, light is described by the concept of *photons* which are light particles with a particular energy (frequency), polarization, and wave vector (direction of propagation). In the classical textbooks where the second quantization is derived (D'Ariano *et al.*, 2003), photons have a plane-wave spatial structure. However, it is easy to imagine photons with a general transverse structure just by assuming a superposition of plane-wave photons (Calvo *et al.*, 2006). Hence, these quantum entities can be well localized both in space and time.

Many quantum information and quantum computation protocols exploit high-dimensional Hilbert spaces. Photons, which constitute the main carrier of information between nodes of quantum networks, can store high-dimensional quantum bits in their spatial degrees of freedom. These degrees of freedom can be tailored by resorting to the symplectic invariant approach based on lossless linear canonical transformations introduced before in the context of classical optics, Eqs. (7–8). These transformations enable one to manipulate the transverse structure of a single photon prepared in superpositions of paraxial modes. Before describing those transformations acting on photons, it is suitable at this stage to introduce in a nutshell the concept of a photon in the paraxial approximation.

In quantum electrodynamics the photon is interpreted as an excitation of the quantized mode field (D'Ariano *et al.*, 2003). Within this quantization, we can define a set of creation and annihilation operators: those that create a photon from the vacuum with a certain paraxial transverse profile, for example, with a well-defined Laguerre-Gaussian mode or a Hermite-Gaussian mode. The most general (paraxial) single-photon pure state can be described as (Calvo *et al.*, 2006)

$$|\psi\rangle = \sum_{\sigma, n_x, n_y} \int_0^\infty d\omega C_{\sigma, n_x, n_y}(\omega) \hat{b}_{\sigma, n_x, n_y}^\dagger(\omega) |vac\rangle \quad (29)$$

Here, $\hat{b}_{\sigma, n_x, n_y}^\dagger(\omega)$ denotes the bosonic creation operator of a photon with a Hermite-Gaussian mode (n_x, n_y) , a linear polarization σ , and a frequency ω acting on the vacuum state $|vac\rangle$. The commutation relations read as $[\hat{b}_{\sigma, n_x, n_y}(\omega), \hat{b}_{\sigma', n_x', n_y'}^\dagger(\omega')] = \delta_{\sigma\sigma'} \delta_{n_x n_x'} \delta_{n_y n_y'} \delta(\omega - \omega')$. The complex coefficients $C_{\sigma, n_x, n_y}(\omega)$ can be interpreted as the probability amplitudes for finding a photon in the state $\hat{b}_{\sigma, n_x, n_y}^\dagger(\omega) |vac\rangle = |\sigma\rangle \otimes |n_x, n_y\rangle \otimes |\omega\rangle$. Here we will focus on the spatial part of Eq. (29).

In order to introduce the link between the Wigner formalism used in previous sections and the single-photon description, let us consider the connection between the cross-spectral density $w_c(\mathbf{r}, \mathbf{r}')$ and the density matrix operator. The density matrix operator for a paraxial wave state $|\psi\rangle$ from Eq. (29) is $\hat{\rho} = |\psi\rangle\langle\psi|$. Here, we deal with coherent waves, but this formalism is more general, and can be extended to partially coherent waves. For coherent waves, the amplitude of the spatial distribution of the photon is defined by $\psi(\mathbf{r}) = \langle \mathbf{r} | \psi \rangle$. In contrast, partially coherent waves do not have a deterministic amplitude, but instead they are described by an ensemble of amplitudes with an associated probability, for example, consider the physical scenario of a phase that is stochastically varying. Within the density operator formalism, partially coherent waves can be described as $\hat{\rho} = \sum_k p_k \hat{\rho}_k = \sum_k p_k |\psi_k\rangle\langle\psi_k|$, where $\hat{\rho}_k$ is a coherent wave with an associated probability p_k . As it is expected for a probability distribution, $\sum_k p_k = 1$ must be satisfied. When $p_k = 1$ for a certain k we return to the coherent wave. Therefore, for both coherent and partially coherent waves, the cross-spectral density is defined as

$$w_c(\mathbf{r}, \mathbf{r}') = \langle \mathbf{r} | \hat{\rho} | \mathbf{r}' \rangle \quad (30)$$

where $\hat{\rho}$ is the density operator of the wave. In the particular case of coherent waves, the cross-spectral density reads as $w_c(\mathbf{r}, \mathbf{r}') = \psi(\mathbf{r})\psi^*(\mathbf{r}')$. As well as in the classical case (see Eq. (1)) the Wigner distribution related by the FT to the cross-spectral density is a faithful representation of quantum state.

Identically to previous sections in which light is considered to be a classical electromagnetic wave, we are interested in tomographic techniques to retrieve the WD and therefore the spatial structure of photons. The main difference between the quantum and the classical picture is the concept of measurement. When the square modulus of the amplitude of the photon is measured in a particular position, the quantum state collapses after the measurement and the information of the initial quantum state that we are interested in is lost. Due to this limitation, it is impossible to retrieve the quantum state of a single photon in a single shot. In general, all methods for characterizing quantum states of photons are based on an ensemble of photons, i.e., a set of many photons with the same quantum state. Then the strategy consists in measuring separately all photons and inferring the quantum state from the information obtained from these measurements.

A spatial structure of a photon propagating through a first-order optical system can be modified as explained in Eq. (12). Assuming that we have N identical photons, we send the photons one at the time through the optical system that performs the fractional FT and the image rotator. In each measurement, we measure a *click* in a particular transverse position of our detector. After repeating the experiment N times, we will obtain a distribution of *clicks* in our detector, related to the modulus square of the amplitude (at the position of the detector) of the quantum state. Using the same strategy than in the classical case, we implement the required transformations to retrieve the WD corresponding to a single photon by changing accordingly the optical system. Hence, the discussed PST techniques can be easily extended to the single-photon regime.

Note that the photon can also be described by a superposition of paraxial modes with different frequencies ω (Eq. (29)) and be localized in time. In that case, we can also apply the chronocyclic tomography strategies discussed for short pulses.

So far, the simple case of a single-photon state has been considered. However, there exist more complicated quantum states of light composed of different photons, and the number of photons not necessarily have to be well-defined (D'Ariano *et al.*, 2003).

Then the Eq. (29) given for single-photon state has to be modified in order to account for more than one photon. If there is well-defined number N of photons in the same mode, the quantum state corresponds to the Fock state given by

$$|\psi_F\rangle = \frac{[\hat{b}_{\sigma,n_x,n_y}^\dagger(\omega)]^N}{\sqrt{N!}} |\text{vac}\rangle \quad (31)$$

Other common states are the coherent and squeezed states that do not have a well-defined number of photons. For example, the coherent states can be described via the displacement operator as

$$|\psi_C\rangle = e^{\alpha \hat{b}_{\sigma,n_x,n_y}^\dagger(\omega) - \alpha^* \hat{b}_{\sigma,n_x,n_y}(\omega)} |\text{vac}\rangle \quad (32)$$

where α is related to the average photon number and might take a complex value. Note that by expanding the exponential a superposition of Fock states is obtained and then the photon number is not well-defined. In most applications of quantum tomography, there is some prior information about the system. For example, it might be the knowledge that the light is described by a coherent state and the problem is then reduced to the determination of the value of α .

Note that in the previous examples for the Fock and coherent states all photons are described by the same mode, which is the common situation in quantum optics experiments. However, there exist more complicated scenarios in which the quantum states of photons with different modes, including the spatial transverse structure, have to be retrieved. In this case the optical system used for quantum state characterization depends on the mode representation or encoding. For example, a mode representation could be a superposition of photons with different transverse spatial profile, or it could be a superposition of photons propagating in different directions, both being possible mode representations. In the latter, which is well-known in the context of linear optical quantum computing, the quantum state is represented by photons propagating in different directions in an optical circuit constituted by beam splitters, phase shifters, and mirror. In particular, it has been proven that any unitary transform in the Fock space for a specific mode can be implemented by using a set of beam splitters and phase shifters (Knill *et al.*, 2001). These transforms can be expressed by a similar mathematical formalism to the one introduced above for the PST characterization of the spatial structure of classical light.

The PST allows establishing a series of powerful techniques to retrieve both quantum and classical states of lights which are of relevance for a wide range of fields such as optical metrology, microscopy, astronomy, classical and quantum optics communication, to name a few.

Acknowledgements

T.A and J.A.R acknowledge the Spanish Ministerio de Economía y Competitividad for the grant TEC2014-57394-P, A.P. acknowledges the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 702565.

References

- Agarwal, G.S., Simon, R., 2000. Reconstruction of the Wigner transform of a rotationally symmetric two-dimensional beam from the Wigner transform of the beam's one-dimensional sample. *Opt. Lett.* 25 (18), 1379–1381.
- Alieva, T., Cámara, A., Rodrigo, J.A., Calvo, M.L., 2011. Phase-space tomography of optical beams. In *Optical and Digital Image Processing*. Wiley-VCH.
- Bastiaans, M.J., 2008. Applications of the Wigner distribution to partially coherent light beams. In *Advances in Information Optics and Photonics*. SPIE.
- Beck, M., Raymer, M.G., Walmsley, I.A., Wong, V., 1993. Chronocyclic tomography for measuring the amplitude and phase structure of optical pulses. *Opt. Lett.* 18, 2041–2043.
- Calvo, G.F., Picón, A., Bagan, E., 2006. Quantum field theory of photons with orbital angular momentum. *Phys. Rev. A* 73, 013805.
- Cámara, A., 2015. *Optical beam characterization via phase-space tomography*. Springer.
- Cámara, A., Alieva, T., Castro, I., Rodrigo, J.A., 2014. Phase-space tomography for characterization of rotationally symmetric beams. *J. Opt.* 16 (1), 015705.
- Cámara, A., Alieva, T., Rodrigo, J.A., Calvo, M.L., 2011. Phase-space tomography with a programmable Radon–Wigner display. *Opt. Lett.* 36 (13), 2441–2443.
- Cámara, A., Rodrigo, J.A., Alieva, T., 2013. Optical coherenscopy based on phase-space tomography. *Opt. Express* 21 (11), 13169–13183.
- D'Ariano, G.M., Paris, M.G.A., Sacchi, M.F., 2003. Quantum tomography. *Adv. Imag. Elect. Phys.* 128, 205–308.
- Dorrer, C., Kang, I., 2003. Complete temporal characterization on short optical pulses by simplified chronocyclic tomography. *Opt. Lett.* 28, 1481–1483.
- Goodman, J.W., 2000. *Statistical Optics*, first ed. Wiley-Interscience.
- Granieri, S., Furlan, W.D., Saavedra, G., Andres, P., 1997. Radon–Wigner display: A compact optical implementation with a single varifocal lens. *Appl. Opt.* 36 (32), 8363–8369.
- Knill, E., Laflamme, R., Milburn, G.J., 2001. A scheme for efficient quantum computation with linear optics. *Nature* 409, 46–52.
- Mendlovic, D., Dorsch, R.G., Lohmann, A.W., Zalevsky, Z., Ferreira, C., 1996. Optical illustration of a varied fractional Fourier-transform order and the Radon–Wigner display. *Appl. Opt.* 35 (20), 3925–3929.
- Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. *The fractional Fourier transform with applications in optics and signal processing*. New York, NY: Wiley.
- Raymer, M.G., Beck, M., McAlister, D.F., 1994. Complex wave-field reconstruction using phase-space tomography. *Phys. Rev. Lett.* 72 (8), 1137–1140.
- Rodrigo, J.A., Alieva, T., Calvo, M.L., 2009. Programmable two-dimensional optical fractional Fourier processor. *Opt. Express* 17 (7), 4976–4983.
- Tian, L., Lee, J., Baek Oh, S., Barbastathis, G., 2012. Experimental compressive phase space tomography. *Opt. Express* 20 (8), 8296–8308.
- Tran, C.Q., Peele, A.S., Roberts, A., *et al.*, 2005. X-ray imaging: A generalized approach using phase-space tomography. *J. Opt. Soc. Am. A* 22 (8), 1691–1700.
- Tu, J., Tamura, S., 1997. Wave field determination using tomography of the ambiguity function. *Phys. Rev. E* 55 (2), 1946–1949.

Coordinate Transformations and the Hough Transform

Filippus S Roux, University of Witwatersrand, Johannesburg, South Africa and National Metrology Institute of South Africa, Pretoria, South Africa

© 2018 Elsevier Ltd. All rights reserved.

Introduction

An optical coordinate transformation is a linear optical implementation of a class of point transforms that is applied to the intensity distribution of an optical beam. Such a coordinate transformation can be implemented with the aid of a thin optical element having a specially designed phase-only transmission function. The optical element is placed in the input plane where it modulates the input optical field in such a way that the subsequent linear optical system produces the required transformed optical field in the output plane. Another optical element with a phase-only transmission function is sometimes placed in the output plane to remove the residual phase, induced by the transformation, so that the transformed amplitude distribution would remain intact upon subsequent propagation (apart from normal diffraction).

The initial concept for the implementation of optical coordinate transformations was proposed in [Bryngdahl \(1974\)](#). However, for optical elements with continuous phase functions, this procedure can only implement geometrical transformations that obey a continuity condition ([Cederquist and Tai, 1984](#)). Subsequently, a number of techniques have been proposed to overcome this restriction. One way is to produce a multifaceted optical element with a discontinuous phase function ([Case et al., 1981](#)). Another approach is to separate the geometrical transformation into two cascaded geometrical transformations, each of which satisfies the continuity condition ([Stuff and Cederquist, 1990](#)). A third method ([Roux, 1994](#)) is to allow the phase function to contain phase singularities, which would produce optical vortices ([Nye and Berry, 1974](#)) in the propagating optical field.

Optical coordinate transformations have been used to implement, among others: a log-polar transform for rotation and scale invariant feature extraction ([Saito et al., 1983](#)); a rotationally symmetric beam-shaping system ([Han et al., 1983](#)); a ring-to-point transform ([Cederquist and Tai, 1984](#)); a rotation transformation ([Roux, 1993a](#)); and a polar formatting transform to compensate for distortion in synthetic-aperture radar data ([Roux, 1995](#)). Recently, a refractive optical implementation of the log-polar transform was proposed for the efficient measurement of the azimuthal index of Laguerre-Gauss modes ([Berkhout et al., 2010](#)).

A different kind of point transform is where the input and output coordinates are related by an implicit transformation equation. The Hough transform is an example of such an implicit transformation. It maps lines in the input plane to individual points in the output plane. After considering the optical implementation of coordinate transformation, we'll discuss the optical implementation of the Hough transform.

Coordinate Transformations

A coordinate transformation that can be implemented optically maps the points on a two-dimensional input plane $\mathbf{x}=(x,y)$ onto the points on a two-dimensional output plane $\mathbf{u}=(u,v)$. Being continuous and one-to-one, such a coordinate transformation is defined by a pair of equations, expressing the output coordinates in terms of functions of the input coordinates $u(\mathbf{x})$ and $v(\mathbf{x})$. These relationships are called the transformation equations, which can be represented as a vector field

$$\mathbf{u}(\mathbf{x}) = u(\mathbf{x})\hat{x} + v(\mathbf{x})\hat{y} \quad (1)$$

The optical implementation employs a phase-only transmission function to direct the light from each point in the input plane to the required location on the output plane. Such a phase-only transmission function can be physically realized as a thin refractive optical element or as a diffractive optical element. It can also be implemented with a programmable spatial light modulator.

Bryngdahl Method

First, we'll consider an optical system comprising a 2-f system with a phase-only transmission function located in the front focal plane, which serves as the input plane. Such a system is shown diagrammatically in [Fig. 1](#). The output amplitude distribution in the back focal plane (the output plane) of a 2-f system is given by the Fourier transform of the amplitude distribution in the input plane. For the case of the system in [Fig. 1](#), the amplitude distribution in the input plane is the input optical field times the phase-only transmission function. To implement the required geometrical transformation, the phase-only transmission function modulates the light at each point in the input plane with an appropriate spatial frequency so that the Fourier transforming action of the lens will place the light from that input point at the required location in the output plane.

Bryngdahl used the stationary-phase approximation to derive a set of differential equations that relates the transformation equations to the gradient of the phase of the transmission function in the input plane of the 2-f system. In vector notation, this

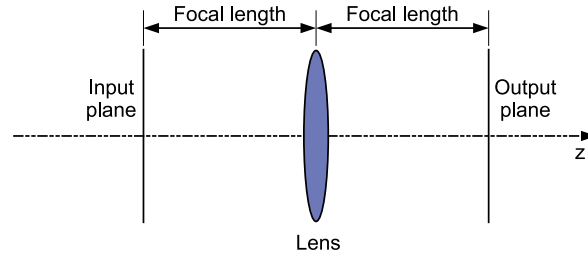


Fig. 1 Optical configuration of a 2-f system.

differential equation is given by

$$\nabla\theta(\mathbf{x}) = \frac{k}{f}\mathbf{u}(\mathbf{x}) \quad (2)$$

where $\theta(\mathbf{x})$ is the phase of the transmission function, k is the wave number and f is the focal length of the lens. Since we'll only work on the two-dimensional input and output planes, all vectors are assumed to be two-dimensional and the gradient operator only operates on the two-dimensional transverse plane $\nabla \equiv \hat{x}\partial_x + \hat{y}\partial_y$.

Note that, Eq. (2) only applies for the case of a 2-f system. One can also use an optical system without a lens, where the input and output planes are separated by a distance L that is large enough to allow the output to be given by Fresnel diffraction. In such a case, the differential equation is

$$\nabla\theta(\mathbf{x}) = \frac{k}{L}[\mathbf{u}(\mathbf{x}) - \mathbf{x}] \quad (3)$$

By solving the differential equation in either Eq. (2) or Eq. (3), depending on the system under consideration, one can obtain the required phase function for the transmission function in the input plane. The transmission function with this phase function can then be implemented with the aid of a refractive optical element or a diffractive optical element or simply a spatial light modulator to realize an optical system that would perform the required coordinate transform.

Continuity Condition

If the required phase function $\theta(\mathbf{x})$ in Eq. (2) is a continuous function, then $\nabla \times \nabla\theta = 0$. It would then imply that $\nabla \times \mathbf{u} = 0$ for both the 2-f system and the lens-less system. In other words, the transformation equations, expressed as a vector field, must be non-rotational. One can also express this requirement as a continuity equation, given by

$$\partial_y u(\mathbf{x}) = \partial_x v(\mathbf{x}) \quad (4)$$

By implication, the two transformation equations are not independent.

There are many useful transformations that do not obey the continuity condition in Eq. (4). A simple example is where the distribution is rotated by an angle α around the optical axis. The transformation equations for such a rotation transformation are

$$\begin{aligned} u(\mathbf{x}) &= \cos(\alpha)x - \sin(\alpha)y \\ v(\mathbf{x}) &= \sin(\alpha)x + \cos(\alpha)y \end{aligned} \quad (5)$$

Here, we have $\partial_y u(\mathbf{x}) = -\sin(\alpha)$, while $\partial_x v(\mathbf{x}) = \sin(\alpha)$. Since these transformation equations do not satisfy the continuity condition, the transformation cannot be implemented with an optical element having a continuous phase function.

There are a number of different approaches that one can follow to overcome the limitations imposed by the continuity condition in Eq. (4). We'll consider three such approaches: representing the transformation as two separate cascaded transformations, as discussed in Section Cascaded Transformations; implementing the transformation with a multifaceted transmission function, as discussed in Section Multifaceted Transmission Function; or incorporating phase singularities in the phase function of the transmission function, which is discussed in Section Transmission Function With Phase Singularities.

Cascaded Transformations

It was shown (Stuff and Cederquist, 1990) that coordinate transformations, which are bijective¹ geometrical transformations, can be implemented with either one or two transmission functions, depending on whether they obey the continuity condition or not. However, there does not seem to be a general approach to formulate the two separate geometrical transformations that would implement a geometrical transformation not satisfying the continuity condition.

A fairly simple example of such a cascaded implementation is where two mirror transformations are used to implement a rotation transformation. Unlike the rotation transformation, mirror transformations do obey the continuity condition. For a

¹A bijective transform is one-to-one and covers the whole domain.

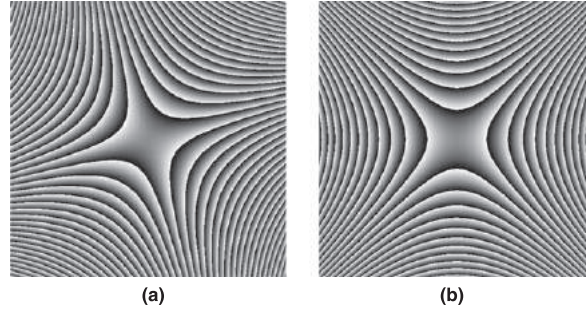


Fig. 2 Two phase functions for the two cascaded mirror transformations, are shown in (a) and (b), respectively. Together they implement a 60 degree rotation. The phase is represented in terms of gray scales, with black and white representing zero and 2π , respectively.

mirror transformation with respect to an arbitrary angle α , the transformation equations are given by

$$u(\mathbf{x}) = \cos(\alpha)x + \sin(\alpha)y, v(\mathbf{x}) = \sin(\alpha)x - \cos(\alpha)y \quad (6)$$

Hence, $\partial_y u(\mathbf{x}) = \partial_x v(\mathbf{x}) = \sin(\alpha)$. If this mirror transformation is followed by another mirror transformation with $\alpha = 0$, the result is the rotation transformation, given in Eq. (5).

Using the differential equation for the 2-f system, given in Eq. (2), together with the mirror transformation equations in Eq. (6), we obtain a phase function given by

$$\theta(\mathbf{x}) = \frac{k}{2f} \cos(\alpha)(x^2 - y^2) + \frac{k}{f} \sin(\alpha)xy \quad (7)$$

This phase function defines the phase-only transmission function in the input plane of the first 2-f system. The output plane of the first 2-f system serves as the input plane for another 2-f system having a similar phase-only transmission function in its input plane, but with $\alpha = 0$. The output of the second 2-f system is the required transformed amplitude distribution, which in this case is a rotated version of the amplitude distribution in the input plane of the first 2-f system. Examples of the two phase functions for the two cascaded mirror transformations that together would implement a 60 degree rotation, are shown in Fig. 2.

Multifaceted Transmission Function

A general transformation can be implemented optically in a brute force approach, by discretizing the input and output planes. The input plane is divided into a grid of small areas. Each of these areas is then treated as a single point in the input plane, which is to be transformed to a single point in the output plane. A diffraction grating with the appropriate spatial frequency is then placed in each of the small areas. Each of these small areas acts like a small aperture that selects a part of the input field and modulates it with the particular diffraction grating function in that aperture. The effect is that the input field is divided into these input areas and each wavelet then gets diffracted in such a way that the light is placed at the correct location in the output plane.

The phase of the diffraction grating for a particular input area, labelled by the indices m and n for the x - and y -directions, respectively, is given by

$$\theta_{m,n}(\mathbf{x}) = \frac{k}{f} [u(x_m, y_n)x + v(x_m, y_n)y] \quad (8)$$

where (x_m, y_n) are the input coordinates at the center of the small input area and $u(x_m, y_n)$ and $v(x_m, y_n)$ denote the output coordinates that are associated with the input point at (x_m, y_n) . The transmission function for each area is then given by

$$t_{m,n}(\mathbf{x}) = \exp[i\theta_{m,n}(\mathbf{x})] \quad (9)$$

and the total transmission function $t_{\text{tot}}(\mathbf{x})$ is formed as the combination of the transmission functions of all the areas

$$t_{\text{tot}}(\mathbf{x}) = \sum_{m,n} t_{m,n}(\mathbf{x}) \prod \left(\frac{x}{\Delta x} - m, \frac{y}{\Delta y} - n \right) \quad (10)$$

where $\prod(\cdot)$ is a rectangular aperture function defining the regions of the small input areas.

Although this method can implement any coordinate transformation, the output field would have the appearance of a discrete array of spots, rather than a smooth optical field. By making the input regions smaller, the output spots would become bigger and be located closer to each other, so that they may eventually overlap. However, depending on the particular coordinate transformation that is being implemented, the spots may be closer to each other in one part of the output plane than in another. Moreover, the overlapping spots in the output plane may cause undesirable interference fringes.

Another result of the multifaceted transmission function is that the sharp boundaries between adjacent input regions cause additional diffraction effects that may be undesirable. One can reduce this effect by allowing the gratings in the different small regions to merge into those of adjacent regions by providing each with a smooth envelope function that extends into the adjacent region and merges with the envelope function of that region. In that case, one would replace the rectangular aperture function

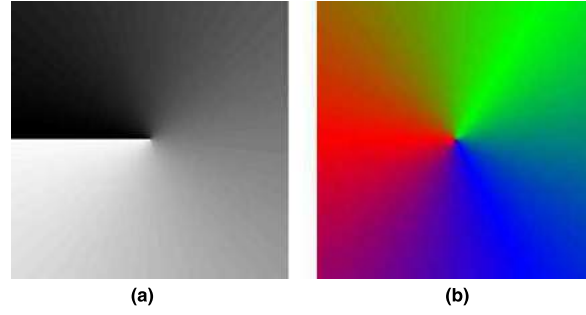


Fig. 3 The phase function of a single isotropic phase singularity. In (a) the phase value is represented in terms of a gray scale, with black and white representing zero and 2π , respectively. Note that the discontinuity along the black-to-white transition is only an artifact of the representation, since 0 and 2π represents the same phase. In (b) a cyclic range of rainbow colors is used to represent the phase values, showing the fact that there is no discontinuity in the phase around the defect in the center.

$\Pi(\cdot)$, defined in Eq. (10), by a suitably smooth envelope function. The result would be a varying diffraction grating over the entire input plane. However, forked grating lines would appear at the boundaries where the grating periods don't match up exactly. These forked grating lines represent phase singularities in the phase of the transmission function, which is the topic of the next section.

Phase Singularities

To discuss the method based on the inclusion of phase singularities in the transmission function, we first need to consider phase singularities in general. Their role in those coordinate transformations that do not obey the continuity condition and the process to compute the required phase function are explained in Section Transmission Function With Phase Singularities.

Phase Functions

One can represent phase as a point on a unit circle around the origin on the complex plane. Each point on the circle represents a unique phase value and all possible phase values are represented by points on the circle. There is no 'largest value' or 'smallest value' for phase. A phase value of 2π represents the same phase as zero. The value of the phase always increases (decreases) for anticlockwise (clockwise) motion round the circle. The unit circle on the complex plane is an example of a compact domain, where the values remain finite even though one never reaches a boundary. This also implies that the circle has a nontrivial topological structure; it is not simply connected.

The topological properties of the configuration space of phase give phase functions special characteristics that make them different from ordinary functions.² A phase function $\theta(\mathbf{x})$ is a continuous mapping from the two-dimensional (x,y) -plane to the circle of phase values. The continuity of the mapping requires that a connected line of points on the (x,y) -plane maps to a connected line of points on the phase circle. The nontrivial topological structure of the circle now allows for the appearance of topological defects. Consider, for instance, all the possible ways that one can map a closed contour from the (x,y) -plane onto or into the circle of phase values. This could give connected phase values that are all located on one side of the circle, so that if the contour shrinks to a point on the (x,y) -plane, the phase values would also shrink to a point on the phase circle. However, if the mapping of the closed contour produces a wrapping around the phase circle, the phase values on one side of the phase circle would have to 'jump' to the other side when the closed contour shrinks to a point on the (x,y) -plane. The 'jumping' process represents a tearing of the topological space, which is not allowed for any process that conserves the topology of the space. This indicates the presence of a topological defect inside the closed contour on the (x,y) -plane.

For a phase function, such topological defects are called phase singularities, because at such a point the phase is undefined and around such a point the phase takes on all the possible phase values. It is therefore classified as an essential singularity. These phase singularities have topological charges, which may be positive or negative depending on the handedness of the wrapping. The topological charges can also be larger than 1 (or smaller than -1) depending on the number of times that the mapping wraps around the phase circle.

The phase function of a single isotropic phase singularity is shown in Fig. 3. Two different ways are used to represent the phase. Henceforth, phase functions are only presented in terms of gray scales.

Mathematically, the presence of a phase singularity inside a close contour can be determined with the aid of an index integral, which integrates the gradient of the phase along the closed contour. The result is an integer multiple of 2π :

$$\oint_C \nabla \theta(\mathbf{x}) \cdot d\hat{s} = 2\pi n \quad (11)$$

where n is an integer representing the net enclosed topological charge.

²In this context, an 'ordinary' function is one with a non-compact range that has no non-trivial homotopy groups.

As an example, consider a phase function that is given by the azimuthal angle $\phi(\mathbf{x})$, expressed as a function of Cartesian coordinates. In this case the index integral can be expressed as

$$\int_0^{2\pi} \partial_\phi \phi \, d\phi = 2\pi \quad (12)$$

As expected, one finds a phase singularity at the origin with a topological charge of $+1$. This result confirms that phase functions are not ordinary functions.

Derivatives of a Phase Function

Applying the Stokes theorem to Eq. (11), one finds

$$\int_A [\nabla \times \nabla \theta(\mathbf{x})] \cdot \hat{z} \, d^2x = 2\pi n \quad (13)$$

where A is the area enclosed by the contour. The expression in Eq. (13) implies that $\nabla \times \nabla \theta(\mathbf{x}) \neq 0$. In fact, the curl of the gradient of a phase function produces a Dirac delta function at the location of each phase singularity, multiplied by its topological charge (Roux, 2006)

$$[\nabla \times \nabla \theta(\mathbf{x})] \cdot \hat{z} = 2\pi \sum_p (\nu_p) \delta(\mathbf{x} - \mathbf{x}_p) \quad (14)$$

where ν_p represents the topological charge of each phase singularity and $\mathbf{x}_p$ is the location of the phase singularity. For example, in the case of a single phase singularity, as represented by a shifted function of the azimuthal angle, one gets

$$\nabla \times \nabla \phi(\mathbf{x} - \mathbf{x}_0) = 2\pi \delta(\mathbf{x} - \mathbf{x}_0) \hat{z} \quad (15)$$

Here the azimuthal angle represents the generic phase function for a single isotropic phase singularity, as shown in Fig. 3.

In more general terms, one can regard the right-hand side of Eq. (14) as a topological charge distribution $T(\mathbf{x})$. So then

$$\nabla \times \nabla \theta(\mathbf{x}) = 2\pi T(\mathbf{x}) \hat{z} \quad (16)$$

By integrating both sides of Eq. (16) over the area A , one recovers Eq. (13), with the understanding that the net topological charge is given by the integral of the topological charge distribution

$$n = \int_A T(\mathbf{x}) \, d^2x \quad (17)$$

General Phase Function

We see now that a phase function in general consists of a continuous part and a singular part. The continuous part is an ordinary function. The singular part can be represented as a sum of shifted versions of the azimuthal angle, multiplied by their topological charges. Hence, a general phase function can be expressed by

$$\theta(\mathbf{x}) = \theta_{\text{cont}}(\mathbf{x}) + \sum_p (\nu_p) \phi(\mathbf{x} - \mathbf{x}_p) \quad (18)$$

where the first term on the right-hand side represents the continuous part and the second term represents the singular part.

Transmission Function With Phase Singularities

Required Topological Charge Distribution

With the new understanding of phase functions, we can take another look at the continuity condition. First we note that the requirement for a continuous phase function is the reason why this restriction exists, because it implies that $\nabla \times \nabla \theta_{\text{cont}}(\mathbf{x}) = 0$. Therefore, if we have a set of transformation equations for which $\nabla \times \mathbf{u} \neq 0$, then it simply means that we need to include phase singularities in the phase function.

Let's apply once again the curl on both sides of Eq. (2), this time using Eq. (16). Then we find that

$$\nabla \times \nabla \theta(\mathbf{x}) = 2\pi T(\mathbf{x}) \hat{z} = \frac{k}{f} \nabla \times \mathbf{u}(\mathbf{x}) \quad (19)$$

The rotational part of the vector field $\mathbf{u}(\mathbf{x})$ precisely gives the topological charge distribution that is required to implement the transform

$$T(\mathbf{x}) = \frac{1}{\lambda f} [\nabla \times \mathbf{u}(\mathbf{x})] \cdot \hat{z} \quad (20)$$

Note, however, that the topological charge distribution thus obtained is a continuous function and not a summation of discrete topological charges associated with distinct phase singularities. Since phase singularities are by their very nature always discrete

points, the required topological charge distribution can at best be approximated by an arrangement of phase singularities over the (x,y) -plane.

The example of the rotation transformation, defined in Eq. (5), requires a constant topological charge distribution given by

$$T(\mathbf{x}) = \frac{1}{\lambda f} [\partial_x v(\mathbf{x}) - \partial_y u(\mathbf{x})] = \frac{2\sin(\alpha)}{\lambda f} \quad (21)$$

It should be noted that such a constant topological charge density cannot be implemented over an arbitrary large region. Due to limitations on the density of phase singularities (Roux, 2003), the maximum radius over which one can maintain a constant topological charge density, as given in Eq. (21), is

$$R_{\max} = \frac{f}{\sin(\alpha)} \quad (22)$$

For larger rotation angles α , one would need to increase the focal length.

Continuous Part of the Phase Function

While the required topological charge distribution is readily computed with the aid of Eq. (20), the continuous part of the phase function requires more attention. One needs to obtain a set of modified transformation equations that excludes the contribution of the phase singularities. Such a modified set of transformation equations would then obey the continuity condition (representing a non-rotational vector field) that could be used to compute a continuous phase function.

The total phase function, given in Eq. (18), consist of a continuous part and the singular part. Computing the gradient of the total phase function, one obtains

$$\nabla\theta(\mathbf{x}) = \nabla\theta_{\text{cont}}(\mathbf{x}) + \sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \quad (23)$$

The left-hand side is given by the transformation equations according to Eq. (2). The required transformation equations for the continuous part of the phase function is thus given by

$$\nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f} \mathbf{u}(\mathbf{x}) - \sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \quad (24)$$

It now remains to determine the last term (the singular term) on the right-hand side of Eq. (24). One can reconstruct the singular term from the topological charge distribution that is required to implement the transform. Although we don't know the locations and topological charges of the individual phase singularities, as required in the last term of Eq. (24), we do know the required topological charge distribution. So one can convert the summation into an integral, where we replace the individual topological charges with the known topological charge distribution

$$\sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \rightarrow \int T(\mathbf{x}') \nabla\phi(\mathbf{x} - \mathbf{x}') d^2x' \quad (25)$$

The result is a convolution integral, which convolves the topological charge distribution with the gradient of a single phase singularity function. The resulting differential equation for the continuous part of the phase function now reads

$$\nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f} \mathbf{u}(\mathbf{x}) - 2\pi \mathbf{F}_s \quad (26)$$

where

$$2\pi \mathbf{F}_s = \int T(\mathbf{x}') \nabla\phi(\mathbf{x} - \mathbf{x}') d^2x' \quad (27)$$

is the singular phase gradient.

As a test of consistency, we apply a curl operation on both sides of Eq. (26). The result gives

$$\nabla \times \nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f} \nabla \times \mathbf{u}(\mathbf{x}) - 2\pi \nabla \times \mathbf{F}_s \quad (28)$$

where

$$2\pi \nabla \times \mathbf{F}_s = \int T(\mathbf{x}') \nabla \times \nabla\phi(\mathbf{x} - \mathbf{x}') d^2x' \quad (29)$$

Since $\theta_{\text{cont}}(\mathbf{x})$ is an ordinary function, the left-hand size of Eq. (28) becomes zero. One can use Eq. (15) to evaluate the integral in Eq. (29). In the end, Eq. (28) becomes

$$0 = \frac{k}{f} \nabla \times \mathbf{u}(\mathbf{x}) - 2\pi T(\mathbf{x}) \quad (30)$$

which is in agreement with Eq. (19). Hence, we have confidence that Eq. (26) represents the required differential equation for the continuous part of the required phase function.

Gradient of the Singular Phase Function

The evaluation of the convolution integral in Eq. (27) requires the expression for the gradient of the generic function for a single phase singularity. One can express the azimuthal angle in terms of the transverse Cartesian coordinate by³

$$\phi(\mathbf{x}) = \frac{1}{i2} \ln \left(\frac{x + iy}{x - iy} \right) \quad (31)$$

The gradient of this phase function is

$$\nabla \phi(\mathbf{x}) = \frac{x\hat{y} - y\hat{x}}{x^2 + y^2} = \frac{\hat{z} \times \mathbf{x}}{|\mathbf{x}|^2} \quad (32)$$

The convolution integral in Eq. (27) thus becomes

$$2\pi F_s = \int T(\mathbf{x}') \frac{\hat{z} \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^2} d^2 x' \quad (33)$$

or, by redefining the integration variable,

$$2\pi F_s = \int T(\mathbf{x} - \mathbf{x}') \frac{\hat{z} \times \mathbf{x}'}{|\mathbf{x}'|^2} d^2 x' \quad (34)$$

The expressions in Eqs. (33) and (34) are real-valued. One can simplify the expressions by combining the x - and y -components as the real and imaginary parts of the expression. The results are

$$2\pi F_s = \int \frac{iT(\mathbf{x} - \mathbf{x}')}{x' - iy'} d^2 x' \quad (35)$$

or

$$2\pi F_s = \int \frac{iT(\mathbf{x})}{(x - x') - i(y - y')} d^2 x' \quad (36)$$

whichever is more convenient to use. The real part of the resulting integral is the x -component and the imaginary part is the y -component.

Application: Rotation Transform

For a constant topological charge density, such as the one required for the rotation transformation, the integral for the singular part of the phase gradient can be readily evaluated with the aid of contour integration. For this purpose we use the expression in Eq. (36) and obtain the result

$$2\pi F_s = \int \frac{iT_0}{(x - x') - i(y - y')} d^2 x' = i\pi(x + iy)T_0 \quad (37)$$

where T_0 represents a constant topological charge density.

For the rotation transformation, the singular part of the phase gradient is therefore given by

$$2\pi F_s = \frac{k}{f} (ix - y) \sin(\alpha) \quad (38)$$

which means that

$$2\pi F_s = \frac{k}{f} (x\hat{y} - y\hat{x}) \sin(\alpha) \quad (39)$$

The differential equation for the continuous phase function is obtain by substituting the transformation equations and the singular phase gradient into Eq. (26). For the rotation transformation, the differential equation becomes

$$\nabla \theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f} \cos(\alpha) \mathbf{x} \quad (40)$$

This differential equation can now be solved to give

$$\theta_{\text{cont}}(\mathbf{x}) = \frac{k \cos(\alpha)}{2f} (x^2 + y^2) \quad (41)$$

The continuous phase function is a lens function with a focal length that depends on the rotation angle. Together with the required topological charge distribution given in Eq. (21), one can now construct the complete phase function that is required to implement the rotation transformation, using a 2-f system. A part of the phase function is shown in Fig. 4. It contains the phase singularities placed on an equilateral triangular grid.

³This definition in terms of the $\ln$ -function is preferred over that with the $\arctan$ -function, because the latter only produces phase values between $-\pi/2$ and $\pi/2$.

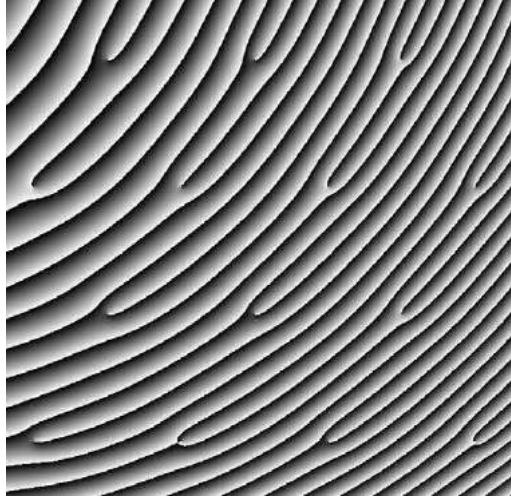


Fig. 4 Part of the phase function with phase singularities of the transmission function for a 45 degree rotation.

Hough Transform

Implicit Transformation

The coordinate transformations, which are discussed above, are defined in terms of a pair of explicit equations that give the output coordinates in terms of the input coordinates and are therefore called explicit transformations. Such transformations are one-to-one, meaning that each output point is associated with only one unique input point.

There are also other transformations that produce one-to-many mappings between input and output points; one input point goes to a set of different output points. Such transformations are defined in terms of implicit equations – a single expression that contains the input coordinates and the output coordinates implicitly – and are therefore referred to as implicit transformations.

Definition of the Hough Transform

An example of an implicit transformation is the conventional Hough transform (Duda and Hart, 1972), which maps lines in the input plane to points in the output plane. The implicit equation for the Hough transform is given by

$$u = x \cos(Kv) + y \sin(Kv) \quad (42)$$

where K is a constant parameter that converts the output coordinate v into an angle. One can define $K = 2\pi/L$, where L is the width of the output region in the v -direction.

For a fixed output point (u, v) , the equation in Eq. (42) describes a line in the input plane. Alternatively, for a fixed input point (x, y) the equation in Eq. (42) describes a sinusoidal curve in the output plane.

Optical Implementation

The optical implementation of the Hough transform can be done with the aid of a single transmission function (Roux, 1993b). Such a transmission function needs to be constructed as a superposition of several transmission functions, each implementing an explicit transformations. Here, we'll describe the implementation with the aid of a multifaceted transmission function, as discussed in Section Multifaceted Transmission Function.

The transmission function of a general optical transform (explicit or implicit), implemented with a multifaceted transmission function, can be expressed as

$$t(\mathbf{x}) = \sum_{m,n,p,q} \alpha_{mnpq} A(x - m\Delta x, y - n\Delta y) \times \Phi^*(p\Delta u, q\Delta v; x, y) \quad (43)$$

where α_{mnpq} denotes the binary coefficients (either equal to 1 or 0) that indicate whether or not there is a link between particular input and output points; $A(\mathbf{x})$ is a suitable envelope function; and $\Phi^*(\mathbf{u}; \mathbf{x})$ is the complex conjugate of the appropriate kernel function (Fourier or Fresnel), depending on the type of optical system in which the transform is implemented. The grid spacings in the input and output planes along the two orthogonal directions are denoted by Δx , Δy , Δu and Δv , respectively. The envelope function could be $\prod(\cdot)$, which is used in Eq. (10), or it could be a smoothing window function to produce a more continuous transmission function.

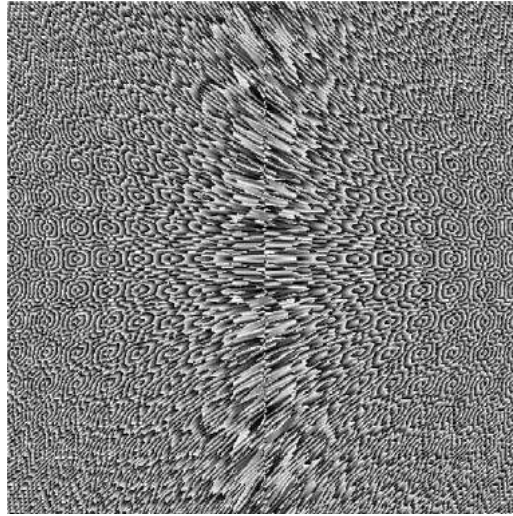


Fig. 5 The phase function of the transmission function for the Hough transform.

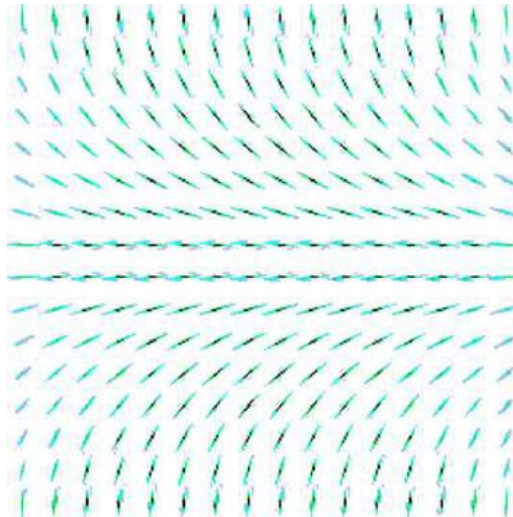


Fig. 6 Output diffraction pattern produced by a fully illuminated transmission function for the Hough transform.

The phase function of the transmission function for an optically implemented Hough transform is shown in [Fig. 5](#). It was computed with [Eq. \(43\)](#) for an implementation in a 2-f system. The Hanning window function was used as envelope function. Using the resulting transmission function, we calculated the diffraction pattern in the output plane when the transmission function is fully illuminated. It is shown in [Fig. 6](#). The diffraction pattern contains an array of discrete output points as produced due to the discrete nature of the implementation. Each output point is elongated in a direction perpendicular to the orientation of the line in the input plane that is mapped to that output point.

General Comments

There are a number of general considerations that one needs to keep in mind when designing optical systems to implement coordinate transformations or more general implicit transformations, such as the Hough transform. These are:

- The stationary phase approximation that is used to derive the differential equations for the phase of the transmission function ignores the amplitude distribution of the incident optical field. As a result, the effect of the phase of the incident optical field is not taken into account in the design of the optical element. If the incident optical field contains a severely tilted phase front at the location of a particular input point, the location of the associated output point would be noticeably shifted away from the

location defined by the transformation. For instance, a helical phase that is unwrapped by the log-polar transform would give a distorted output line in the case of high azimuthal index in the incident optical field.

- The size of an input area, restricted by an aperture, gives rise to a finite area in the output plane with a minimum size. For a smaller (larger) input area, the size of the output area becomes larger (smaller). This relationship is well-known for the 2-f system, where the input and output distributions are related by a Fourier transform,

$$d_{\text{out}} = \frac{\lambda f}{d_{\text{in}}} \quad (44)$$

where d_{in} and d_{out} represent the sizes of the input and output areas, respectively. A similar relationship applies for other systems such as the lens-less system.

The relationship between the sizes of these areas in the input and output can be used to compute a quantity called the space-bandwidth-product (SBP), which is determined as follows. The size of the total usable output area defines a smallest feature size in the input plane via the Fourier relationship in Eq. (44). Anything smaller would scatter light beyond the boundary of the usable output area. At the same time, the input plane also have a finite size for its total usable area that produces a small spot in the output, which thus defines the resolution in the output plane. The SBP is now given by the ratio of the total input area divided by the smallest feature size and it is also equal to the ratio of the total output area divided by the smallest resolution cell in the output plane.

A coordinate transformation, defines a mapping from separate points on the input plane to distinguishable points on the output plane. In practice, these points would be represented by finite size areas both on the input and the output planes. What do the sizes of these areas need to be and how many can there be? One cannot use the number given by the SBP of the system as an indication of the number of points that are mapped between input and output, because if we use the smallest input feature size in the input, then the output points would be as large as the entire output area. Moreover, if one wants the output points to be the size of the resolution cells, the input points would need to cover the entire input area. Hence, the optimal number of points that can be used (in the case of the rotation transform, for instance) would be the square root of the SBP. This gives as many distinguishable output points as separated input points. For this reason, a coordinate transform is not an effective way to implement beam shaping.

- Depending on the type of implementation method that is used, one may have to accept a limited diffraction efficiency. The result is that there could be undiffracted light that also propagates together with the diffracted light, after passing through the device that implements the transmission function. To separate the diffracted light from the undiffracted light, one can add an overall phase grating (linear phase tilt) to the phase of the transmission function. The magnitude of the phase tilt needs to be larger than the largest spatial frequency in the original phase function to ensure that the desired first diffraction order would be completely separated from the undiffracted zeroth order. The desired first diffraction order then appears next to the undiffracted zeroth order in the output plane and can be separated from it by spatial filtering.
- In the case of an implicit transformation where one input point is mapped to several output points, such as with the Hough transform, the transmission function consists of the superposition of several transmission functions for the individual output points. The superposition causes interference, which modulates the amplitude (magnitude) of the resulting transmission function. As a result the final transmission function for the total implicit transform is not a pure phase function (it is not 'phase-only'), such as the transmission functions for the individual output points.

This gives rise to a number of consequences. To implement the complex amplitude of the total transmission function in a phase-only device, such as a spatial light modulator or diffractive optical elements, special encoding techniques need to be used. Alternatively, one can use older technologies that incorporate absorption. The modulated amplitude of the total transmission function implies that some of the incident optical power would be lost due to absorption or scattering, depending on how the complex amplitude of the transmission function is implemented. The result is a poor power efficiency for the implementation. One can try to use only the phase of the total transmission function to implement a phase-only optical transform, but in general the resulting transformation would then suffer from low fidelity.

Conclusions

Point transforms that are implemented in linear optical systems can be divided into explicit transformations, such as coordinate transformations, and implicit transformations, such as the Hough transform. Explicit transformations include those that only require a continuous phase function in the transmission function and therefore satisfy the continuity condition and those that also require phase singularities in the phase function of the transmission function and therefore do not satisfy the continuity condition.

Those explicit transformations that do satisfy the continuity condition can be implemented by a direct solution of the Bryngdahl differential equations. Explicit transformations that do not satisfy the continuity condition require additional attention. These explicit transformations either need to be separated into two cascaded explicit transformations, each satisfying the continuity condition or they need to be implemented as a multifaceted transmission function or they need the inclusion of phase singularities in the phase function of the transmission function.

The Hough transform is an implicit transformation that maps lines in the input plane to separate points in the output plane. It can be implemented as a superposition of several discrete transmission functions that map particular input points to particular output points.

The optical implementation of geometrical transformations provide a fast effective way to process optical information. However, care must be taken in the design of these systems to ensure that the implementation provide an accurate result.

References

- Berkhout, G.C.G., Lavery, M.P.J., Courtial, J., Beijersbergen, M.W., Padgett, M.J., 2010. Efficient sorting of orbital angular momentum states of light. *Phys. Rev. Lett.* 105, 153601.
- Bryngdahl, O., 1974. Geometrical transformations in optics. *J. Opt. Soc. Am.* 64, 1092–1099.
- Case, S.K., Haugen, P.R., Lokberg, O.J., 1981. Multifacet holographic optical elements for wave front transformations. *Appl. Opt.* 20, 2670–2675.
- Cederquist, J.N., Tai, A.M., 1984. Computer-generated holograms for geometrical transformations. *Appl. Opt.* 23, 3099–3104.
- Duda, R.O., Hart, P.E., 1972. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM* 15, 11–15.
- Han, C.-Y., Ishii, Y., Murata, K., 1983. Reshaping collimated laser beams with gaussian profile to uniform profiles. *Appl. Opt.* 22, 3644–3647.
- Nye, J.F., Berry, M.V., 1974. Dislocations in wave trains. *Proc. R. Soc. Lond. A* 336, 165–190.
- Roux, F.S., 1993a. Diffractive optical implementation of rotation transform performed by using phase singularities. *Appl. Opt.* 32, 3715–3719.
- Roux, F.S., 1993b. Implementation of general point transforms with diffractive optics. *Appl. Opt.* 32, 4972–4978.
- Roux, F.S., 1994. Branch-point diffractive optics. *J. Opt. Soc. Am. A* 11, 2236–2243.
- Roux, F.S., 1995. Single-element diffractive optical system for realtime processing of synthetic aperture radar data. *Appl. Opt.* 34, 5045–5052.
- Roux, F.S., 2003. Optical vortex density limitation. *Opt. Commun.* 223, 31–37.
- Roux, F.S., 2006. Fluid dynamical enstrophy and the number of optical vortices in a paraxial beam. *Opt. Commun.* 268, 15–22.
- Saito, Y., Komatshu, S., Ohzu, H., 1983. Scale and rotation invariant real time optical correlator using computer generated hologram. *Opt. Commun.* 47, 8–11.
- Stuff, M.A., Cederquist, J.N., 1990. Coordinate transformations realizable with multiple holographic optical elements. *J. Opt. Soc. Am. A* 7, 977–981.

Single-Pixel Imaging Using the Hadamard Transform

Fernando Soldevila, Pere Clemente, Enrique Tajahuerce, and Jesús Lancis, Jaume I University, Castelló, Spain

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Sometimes great revolutions occur in science. All the different branches of physics had short periods of time where new ideas flooded their communities and sparked new technologies. Perhaps, Optics has been the most affected by those kinds of revolutions. We can clearly see at least three periods of time that were crucial to the development of light sciences as we know them today: the improvement in optical lenses fabrication, the generation of laser based light sources, and the development of electronic light sensors. The first one occurred in the 19th century, thanks to the works of Zeiss, Schott, and Abbe. They provided not only technical developments in the fabrication of lenses, but also the theoretical framework of modern optics. In the 1950s, the first laser sources appeared, and proved to be exceptional light sources that rapidly replaced white light sources in almost every experiment. With their use, novel techniques emerged, such as lithography, range measuring, and laser cutting. Furthermore, new devices were developed around them, such as optical disk drives. Lastly, at the end of the 1960s, charge-coupled devices (CCD) were invented. Now, optical fields could be digitally stored, and soon digital imaging techniques provided information that would have never been possible by working with conventional photographic films.

Even though this last revolution happened half a century ago, it slowly started a trend that continues today: the symbiosis of Electronics, Computer Science and Optics. Now, new optical techniques not only need to build optical systems considering lenses and light sources, but also the electronics present in their detectors and the computer algorithms they use to recover useful information. Moreover, new electronical devices have been built to control light, such as spatial light modulators, optical filters, etc. This blending of disciplines has proved to be an exceptional tool in a plethora of industries, life sciences, telecommunications, and astronomy.

Although CCD and complementary metal-oxide-semiconductor (CMOS) sensors have established as the way-to-go technology in digital imaging, several limitations remain unsolved. Obtaining images in low light level scenarios, in exotic ranges of the electromagnetic spectrum and at extremely high speeds still entails very expensive devices or with great limitations regarding spatial resolution and response time. Due to this, scientists have never stopped their search for new imaging modalities using different detectors. One of the modern solutions to these problems consists of using very specialized sensors without spatial resolution. One can think of those detectors as a digital camera with only one pixel (thus they are usually called single-pixel detectors) (Duarte *et al.*, 2008). Several questions should arise now: why using just one pixel? and, how do I obtain an image (with hundreds or thousands of pixels) with just a single-pixel detector?

The first question has a simple explanation. Building just one sensor with fast response time, sensibility or spectral response is usually easy, and it is always easier than building an array of thousands or millions of sensors with the same characteristics. Then, these cameras using single-pixel detectors present enhanced capabilities in challenging scenarios for traditional digital cameras. Now, the second question needs a little longer explanation, which we will see in the following section.

Single-Pixel Imaging Basics

In order to understand the operation of a single-pixel camera, we can start with a conventional digital camera, as shown in Fig. 1. In a conventional setup, the scene to be acquired is imaged with the aid of an optical system (think of the photographic objective of a digital camera, for example) onto the sensor. Due to image formation laws, each region of the scene is imaged onto a different

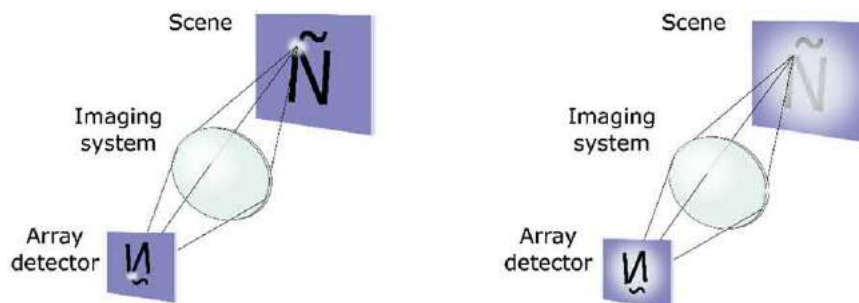


Fig. 1 Conventional digital camera setup. Left: using an imaging system, different regions of the scene are imaged onto different regions of the sensor, that measures the light irradiance. Using this information, a digital image is acquired. Right: as an array detector is used, this process is made in parallel, and each pixel of the detector obtains the information of a different region of the scene at the same time.

region of the sensor, so we have a 1:1 map between regions of the scene and regions of the sensor. Now, each pixel of the sensor measures the light irradiance coming from the scene. In a conventional CCD sensor, every pixel measures at the same time, so the image is acquired in a single shot. However, there are other ways of obtaining the same information.

One can illuminate only a small region of the scene (either by using a galvanometric mirror and a laser beam, or with a spatial light modulator), and measure the light irradiance coming from that region with a detector. This detector does not need to have millions of pixels any more, as light is collected with the optics onto a small spot, and the only relevant quantity to measure is not where the light goes to, but the amount of light coming from the scene. This is usually known in the imaging community as raster scanning, and it's the common operation mode for techniques such as confocal microscopy. The process is depicted in Fig. 2. Now, as the detector does not need spatial resolution, instead of a CCD sensor, different sensors can be used. For example, if one needs to work in low light level scenarios, photomultiplier tubes or photon counting detectors can be used. Even beam spectrometers and polarimeters have been used, providing polarimetric and multispectral images. This measurement process can be depicted by using simple mathematical operations. We consider a bidimensional object, $O(x, y)$. In order to obtain an image, we make the superposition of the object and a function that belongs to an orthogonal basis. In the raster scanning example, those functions are Dirac deltas, each one centered at a different spatial position. Then, each measurement, I_p , can be expressed as

$$I_p = O(x, y) \cdot \delta_p(x - x_p, y - y_p) \quad (1)$$

For each delta function, the single-pixel detector measures the amount of light coming from the object. After all the measurements are done, we can recover the object with the equation

$$O = \sum_p I_p \cdot \delta_p \quad (2)$$

It must be stated that, in order to recover an image with a number N of pixels, N delta functions need to be generated and sequentially measured with the detector, whereas in a conventional camera all the measurements are done in one single shot.

After the introduction of the raster scanning approach, it is easy to understand that there is no need to limit oneself to the use of delta functions. The same mathematical problem can be solved in any base of orthogonal functions, and this knowledge can be very useful in practical scenarios. As shown in Fig. 3, now instead of scanning the scene, different light patterns (functions of the basis) are sequentially generated onto the object, performing what is usually known as a basis scan. The main advantage of using different functions is that, instead of just illuminating a small region of the scene, several parts of the object are lighted at the same time. This increases the signal level at the detector, which ultimately improves the signal-to-noise ratio of the measurements. This is usually known as the Fellgett or multiplex advantage. Here, a plethora of functions have been discussed in the references, such as Fourier functions, DCT, and wavelets. Even non-orthogonal functions, such as random functions, can be used to obtain images, as

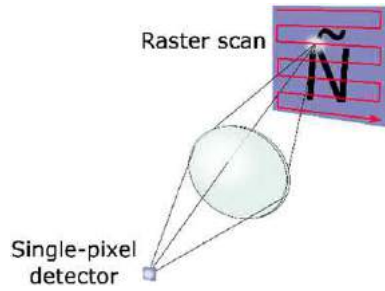


Fig. 2 Raster scanning imaging. The scene under study is scanned with a light beam, and the reflected irradiance is collected by an optical system and measured with a single-pixel detector. By scanning the full scene, the image can be recovered.

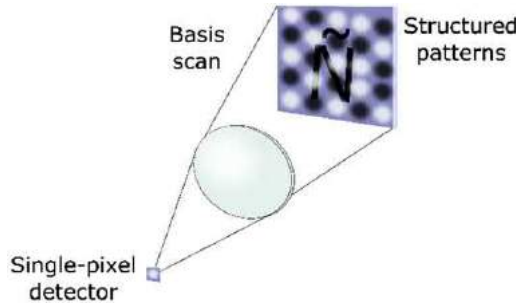


Fig. 3 Basis scanning imaging. The scene under study is lighted with a complete set of functions from an orthogonal basis, and the reflected irradiance is collected by an optical system and measured with a single-pixel detector. After all the functions have been projected, the set of measurements can be used to recover the image of the scene.

in the case of computational ghost imaging techniques. Each one of the basis presents its benefits and characteristic experimental implementations given its mathematical properties, and usually the choice between one on another can be made depending on the specifications of the device that one wants to build. Here, we are going to discuss the use of Walsh-Hadamard functions, its properties, and the requirements needed in order to build a single-pixel camera using them.

Walsh-Hadamard Single-Pixel Imaging

Walsh functions can be derived from Hadamard matrices, which are square matrices with dimensions of power of 2, entries either $+1$ or -1 , and the property that the product of two different rows or columns is zero (Pratt *et al.*, 1969) (see Fig. 4). Then, each row (column) can be understood as a square binary function which is orthogonal to all the other rows (columns). Hadamard matrices can be calculated for arbitrary big dimensions, and are easily stored and transmitted. Furthermore, due to being binary matrices, its technical implementation using spatial light modulators tends to be straightforward. Given those reasons, they are an excellent candidate to operate in a single-pixel architecture. Using vector form, now the measurement process can be described by

$$y = H \cdot x \quad (3)$$

where the object under study is denoted by a vector, x , containing the N pixels of the scene arranged in column form. The measurement process is represented by a Hadamard matrix, H , with dimension $N \times N$. Each row of this matrix contains a Walsh function. The measurements are arranged in the column vector y . In each element of the measurement vector the superposition of a Walsh function and the whole object is stored. This process of projecting the Walsh functions and measuring the resulting irradiance can be understood as an optical way of calculating the Walsh-Hadamard transform of the object (see Fig. 5). This is similar to the measurement of the Fourier Transform of an object by using a digital camera in the focal plane of a lens. Finally, after the measurements are done, one can recover the object by inverting the problem in Eq. (3). This can also be understood as performing the inverse Walsh-Hadamard transform of the measurement vector, y .

However, there can be a problem regarding the implementation of Walsh functions in spatial light modulators. As their entries consist of either $+1$ or -1 , amplitude modulators cannot directly implement Walsh functions (there is no available -1 state). A usual workaround entails the use of two different functions to codify a function. Instead of working with the true functions, a Hadamard matrix can be expressed as $H = H^+ - H^-$. The first matrix is built by substituting the -1 entries by 0, and the second one by replacing the $+1$ entries by zero, and the -1 entries by 1. Then, the measurement process can be expressed by

$$y = H \cdot x = H^+ \cdot x - H^- \cdot x \quad (4)$$

which matrices can be easily codified on amplitude spatial light modulators. It has to be noted that, if one uses phase spatial light modulators, this problem does not occur. However, due to the extremely fast operation modes of amplitude spatial light modulators and the sequential nature of single-pixel cameras, those are the ones usually used.

Once the fundamentals of Walsh-Hadamard imaging have been established, here we show an example of a single-pixel camera using a digital micromirror device (DMD) as an amplitude spatial light modulator. A scheme of the camera setup is shown in Fig. 6. Single-pixel imaging can be divided into two main stages: illumination (or codification) and detection. In the illumination

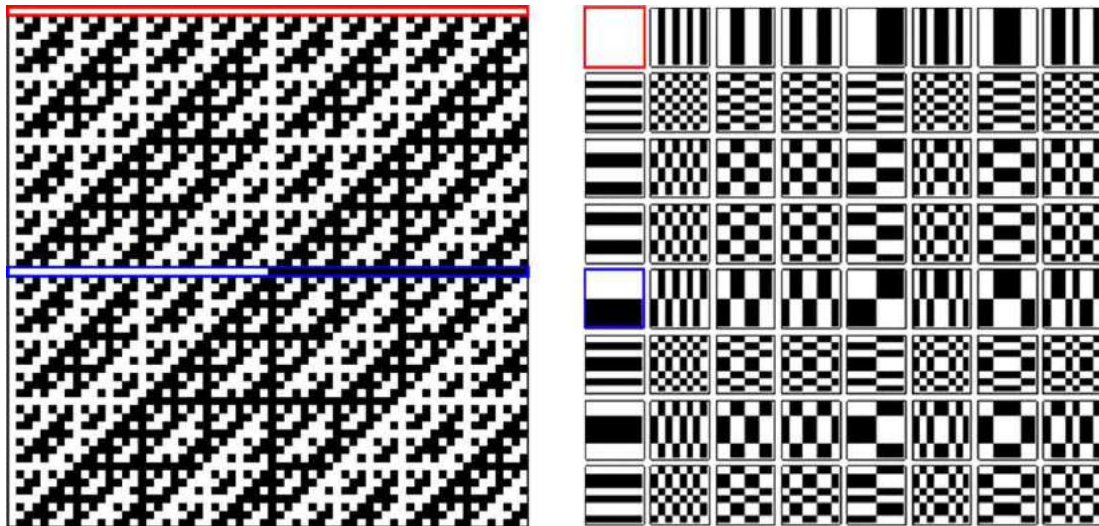


Fig. 4 Hadamard matrix and patterns example. Left: Hadamard matrix of dimension 64. White regions represent $+1$ entries and black regions represent -1 entries. Right: Hadamard patterns derived from the Hadamard matrix. In order to build each one of the 64 patterns, one row of the matrix is selected and arranged into a 8×8 pattern. Two example patterns, highlighted red and blue, are marked with their correspondent rows of the Hadamard matrix.

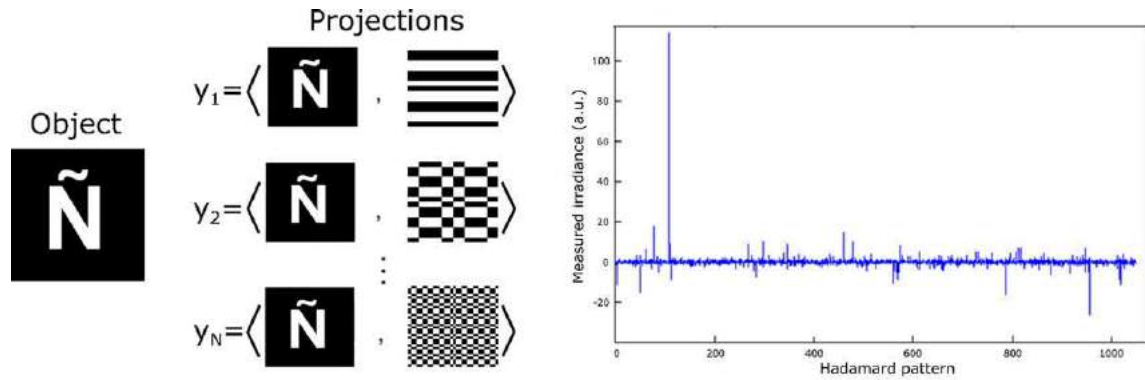


Fig. 5 Single-pixel Hadamard measuring fundamentals. The object is measured by making the projections onto the basis of Hadamard patterns. Those projections consist on overlapping both the object and each Hadamard pattern, and then measuring the total irradiance coming to the detector. By doing this, one gets the decomposition of the object in the Hadamard basis (this graph is the 1D Walsh-Hadamard transform of the object expressed in vector form). Once this decomposition is measured, the recovery can be obtained by performing the inverse Walsh-Hadamard transform of the coefficient vector.

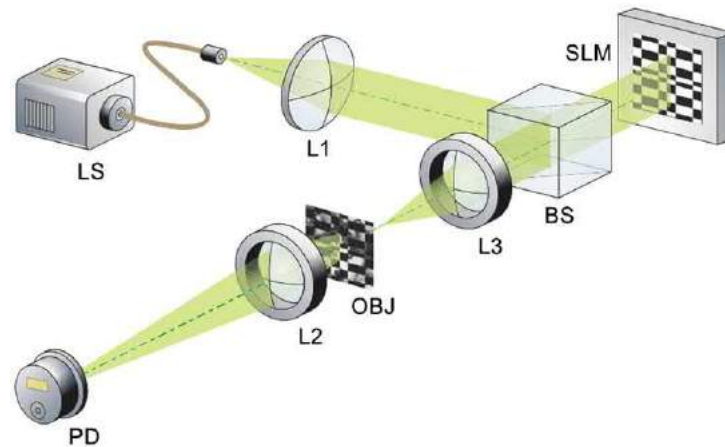


Fig. 6 Single-pixel camera setup. Light coming from a light source (LS) is structured with a spatial light modulator (SLM). An image of the SLM is projected onto the object (OBJ) with several lenses and a beam-splitter (BS). After the superposition of both the Walsh-Hadamard functions and the object, the resulting irradiance is measured with a single-pixel detector (a photodiode in this case, PD).

stage, a light source illuminates a spatial light modulator, where the Walsh functions are generated. In this case, each one of the rows of the Hadamard matrix is rearranged in a bidimensional mask, usually called Hadamard pattern. Those patterns are projected into the scene that one wants to measure with the aid of an optical projection system. Then, the object and the Hadamard patterns overlap in one plane. In the detection stage, the resulting light distribution is collected with the aid of another optical system. A single-pixel detector measures this irradiance value, which is the projection of the object into one of the functions of the Walsh basis. Then, the spatial light modulator codifies another Walsh function and the detector measures again. After the full set of Hadamard patterns have been projected, the image of the object can be recovered. Two examples of images obtained by using this approach are shown in Fig. 7.

It must be stated that the spatial codification of information is realized when both the Hadamard patterns and the object overlap. Here we have shown one way of doing this, by projecting the patterns generated by the spatial light modulator onto the object. However, this projection can be done in different ways. For example, one can image the object onto the spatial light modulator with an optical system, and then make the overlapping in the spatial light modulator plane. After that, light is collected and measured with a single-pixel detector in the same way as in the first approach. This flexibility can be exploited in different scenarios, given that sometimes it is easy to work with an active illumination (sending light patterns to an object) and sometimes it is better to work in a passive configuration (imaging the scene onto the spatial light modulator).

Walsh-Hadamard Single-Pixel Applications

Given the benefits mentioned before, examples of single-pixel imaging techniques using Walsh-Hadamard functions can be extensively found in the references. They can be divided in four main groups. First, given the ease to build single-pixel detectors

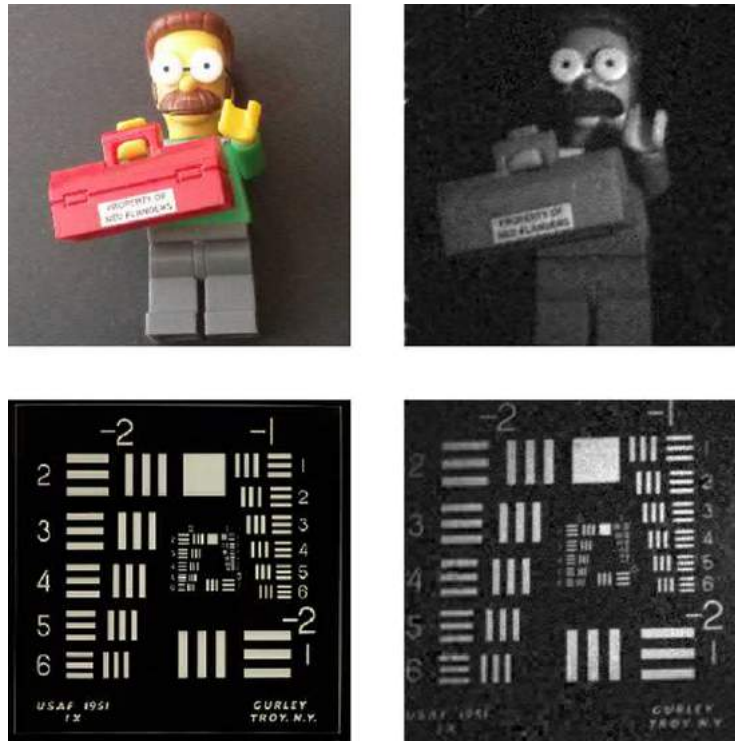


Fig. 7 Single-pixel imaging using Hadamard patterns. In the top row, we show a 256×256 pixels LEGO® Ned Flanders picture (left) and its single-pixel reconstruction. In the bottom row, we show a 512×512 pixels USAF1951 test image (left) and its single-pixel reconstruction (right). Images extracted from Soldevila, F., Salvador-Balaguer, E., Clemente, P., Tajahuerce, E., Lancis, J., 2015. High-resolution adaptive imaging with a single photodiode. *Scientific Reports* (Nature Publishing Group) 5, 7 (Article No. 14300). Licensed under CC BY 4.0.

that work in exotic ranges of the spectral domain, multiple single-pixel cameras have been built in the IR and THz spectral regions (Radwell *et al.*, 2014; Watts *et al.*, 2014). Second, using simple detectors in combination with other optical and mechanical elements has allowed to perform multidimensional imaging. Here we can name multi and hyper spectral imaging (Soldevila *et al.*, 2013; Li *et al.*, 2017; Studer *et al.*, 2012), polarimetric imaging (Durán *et al.*, 2012; Welsh *et al.*, 2015), phase imaging (Clemente *et al.*, 2013), and 3D imaging (Zhang *et al.*, 2016), among others. Third, given the simplicity of the detection scheme, dedicated sensors can also be used to work in challenging scenarios, such as low-light level regimes, by using photomultiplier tubes or photon counting detectors. This has enabled single-pixel cameras to provide images, for example, in presence of highly scattering media, where distinguishing between non-scattered photons and diffuse light is of paramount interest (Tajahuerce *et al.*, 2014; Durán *et al.*, 2015). Lastly, single-pixel ideas have also been applied to well established optical techniques, improving their performance in demanding situations. Here we can name several applications, such as improving the SNR in two-photon microscopy (Ducros *et al.*, 2013) and photoacoustic imaging (Wang *et al.*, 2012), and enhancing resolution in traditional microscopy setups (Rodríguez *et al.*, 2014).

Challenges, Future Work

The main drawback of single-pixel cameras is the need to perform a sequential codification of information: whereas a traditional digital camera images the scene in a single shot, single-pixel cameras need to project a set of patterns onto the scene to recover an image. Furthermore, the higher the spatial resolution one wants to obtain, the higher the number of projections that need to be measured. Even though fast spatial light modulators are used, such as digital micromirror devices, that have repetition rates above 20 kHz, images higher than 64×64 pixels are rarely seen in applications where real-time operation is needed. For several applications, this resolution is high enough, but there are situations where high spatial resolution is needed, such as medical imaging.

In order to tackle this problem, multiple approaches have been proposed. Using Compressive Sensing techniques enable single-pixel cameras to reduce the number of required pattern projections (Duarte *et al.*, 2008; Candès, 2006; Romberg, 2008). Even though they usually entail high post-processing times, novel approaches have also been used to provide real-time video (Sankaranarayanan *et al.*, 2012). Other techniques, based on adaptive algorithms, have also been proposed as a way to reduce the measurement time with no need of complex post-processing stages (Soldevila *et al.*, 2015; Phillips *et al.*, 2017). However, none of those solutions have been able to provide results comparable to traditional digital cameras in terms of spatial resolution and frame

rate as of today. Further development in both computational algorithms and technical improvements in the fabrication of spatial light modulators (i.e., the development of faster SLMs) will be key to solve the present limitations of single-pixel cameras.

References

- Candès, E., 2006. Compressive sampling. In: Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006. Zurich, Switzerland: European Mathematical Society Publishing House, pp. 1433–1452.
- Clemente, P., Durán, V., Tajahuerce, E., *et al.*, 2013. Compressive holography with a single-pixel detector. *Opt. Lett.* 38 (14), 2524–2527.
- Duarte, M.F., Davenport, M.A., Takhar, D., *et al.*, 2008. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* 25 (2), 83–91.
- Ducros, M., Goulam Houssen, Y., Bradley, J., de Sars, V., Charpak, S., 2013. Encoded multisite two-photon microscopy. *Proc. Natl. Acad. Sci. USA* 110 (32), 13138–13143.
- Durán, V., Clemente, P., Fernández-Alonso, M., Tajahuerce, E., Lancis, J., 2012. Single-pixel polarimetric imaging. *Opt. Lett.* 37 (5), 824–826.
- Durán, V., Soldevila, F., Irlés, E., *et al.*, 2015. Compressive imaging in scattering media. *Opt. Express* 23 (11), 14424.
- Li, Z., Suo, J., Hu, X., *et al.*, 2017. Efficient single-pixel multispectral imaging via non-mechanical spatio-spectral modulation. *Sci. Rep.* 7, 41435.
- Phillips, D.B., Sun, M.-J., Taylor, J.M., *et al.*, 2017. Adaptive foveated single-pixel imaging with dynamic supersampling. *Sci. Adv.* 3 (4), e1601782.
- Pratt, W., Kane, J., Andrews, H., 1969. Hadamard transform image coding. *Proc. IEEE* 57 (1), 58–68.
- Radwell, N., Mitchell, K.J., Gibson, G., *et al.*, 2014. Single-pixel infrared and visible microscope. *Optica* 1 (5), 285–289.
- Rodríguez, A.D., Clemente, P., Irlés, E., Tajahuerce, E., Lancis, J., 2014. Resolution analysis in computational imaging with patterned illumination and bucket detection. *Opt. Lett.* 39 (13), 3888–3891.
- Romberg, Justin, 2008. Imaging via compressive sampling. *IEEE Signal Process. Mag.* 25 (2), 14–20.
- Sankaranarayanan, A.C., Studer, C., Baraniuk, R.G., 2012. CS-MUVI: Video compressive sensing for spatial-multiplexing cameras. In: 2012 IEEE International Conference on Computational Photography (ICCP), pp. 1–10.
- Soldevila, F., Irlés, E., Durán, V., *et al.*, 2013. Single-pixel polarimetric imaging spectrometer by compressive sensing. *Appl. Phys. B Lasers Opt.* 113 (4), 551–558.
- Soldevila, F., Salvador-Balaguer, E., Clemente, P., Tajahuerce, E., Lancis, J., 2015. High-resolution adaptive imaging with a single photodiode. *Sci. Rep.* 5, 14300.
- Studer, V., Bobin, J., Chahid, M., *et al.*, 2012. Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proc. Natl. Acad. Sci. USA* 109 (26), E1679–E1687.
- Tajahuerce, E., Durán, V., Clemente, P., *et al.*, 2014. Image transmission through dynamic scattering media by single-pixel photodetection. *Opt. Express* 22 (14), 16945–16955.
- Wang, Y., Maslov, K., Wang, L.V., 2012. Spectrally encoded photoacoustic microscopy using a digital mirror device. *J. Biomed. Opt.* 17 (6), 66020.
- Watts, C.M., Shrekenhamer, D., Montoya, J., *et al.*, 2014. Terahertz compressive imaging with metamaterial spatial light modulators. *Nat. Photonics* 8, 605–609.
- Welsh, S.S., Edgar, M.P., Bowman, R., Sun, B., Padgett, M.J., 2015. Near video-rate linear Stokes imaging with single-pixel detectors. *J. Opt.* 17 (2), 25705.
- Zhang, Y., Edgar, M.P., Sun, B., *et al.*, 2016. 3D single-pixel video. *J. Opt.* 18 (3), 035203.

Linear Canonical Transforms

Kurt B Wolf, National Autonomous University of Mexico, Cuernavaca, Morelos, Mexico

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Canonical transformations preserve the *structure* of Hamiltonian systems, be they *classical* as geometric optics or classical mechanics, or *wave* as monochromatic optics or quantum mechanics. When these transformations are *linear*, they can be represented by matrices, a most useful technique to keep account of their composition, which facilitates finding the end product of optical elements in a setup.

One can broadly define a D -dimensional system to be *Hamiltonian* when it contains variables or observables of *position*, $\mathbf{Q} = \{Q_i\}_{i=1}^D$, and an equal number of *momentum* variables or observables $\mathbf{P} = \{P_i\}_{i=1}^D$, which constitute its *phase space*; and a Hamiltonian function $H(\mathbf{Q}, \mathbf{P})$ that determines its *canonical evolution* along an optical axis z , or time t as in mechanics. If the transformation is to be *linear*, one must assume that phase space is the full real $2D$ -dimensional space $\mathbb{R}^{2D}$. This proviso allows only *paraxial* optical models to be addressed directly, as well as non-relativistic classical and quantum mechanics, since their treatment is mathematically similar to the corresponding classical and wave models in paraxial optics.

Conservation of Hamiltonian Structure

An important property of Hamiltonian systems is their *structure*; this boils down to a set of relations that must be conserved between the phase space coordinates under any transformation: *Poisson brackets* in the classical models and *commutators* in the wave models, namely

$$\{q_i, p_j\} = \delta_{ij}, \quad \{q_i, q_j\} = 0, \quad \{p_i, p_j\} = 0, \quad q_i \equiv Q_i \in \mathbb{R} \quad (1)$$

$$[\hat{Q}_i, \hat{P}_j] = i\hbar \delta_{ij}, \quad [\hat{Q}_i, \hat{Q}_j] = 0, \quad [\hat{P}_i, \hat{P}_j] = 0, \quad \hat{Q}_i \equiv Q_i \text{ on } \mathcal{L}^2(\mathbb{R}^D) \quad (2)$$

These operations are bilinear: $\{[c_1 A_1 + c_2 A_2, B]\} = c_1 \{[A_1, B]\} + c_2 \{[A_2, B]\}$, skew-symmetric: $\{[A, B]\} = -\{[B, A]\}$, satisfy the Jacobi identity and the Leibniz rule. They generate thus the Lie algebra with derivation referred by Heisenberg and Weyl. The wave models (of reduced wavelength $\tilde{\lambda} = \lambda/2\pi$, or $\hbar = h/2\pi$ in quantum mechanics) further require the Hilbert space $\mathcal{L}^2(\mathbb{R}^D)$ of Lebesgue square-integrable 'wave'-functions, where the operators $\hat{Q}_i$ and $\hat{P}_j$ are essentially self-adjoint.

The 4×4 matrix of Lie brackets (1)–(2) with the $2D$ column vector $\mathbf{W} := \begin{pmatrix} \mathbf{Q} \\ \mathbf{P} \end{pmatrix}$ and its transpose $\mathbf{W}^T$, written as

$$\{[\mathbf{W}^T, \mathbf{W}]\} := \left\{ \left[\begin{pmatrix} \mathbf{Q} & \mathbf{P} \end{pmatrix}, \begin{pmatrix} \mathbf{Q} \\ \mathbf{P} \end{pmatrix} \right] \right\} := \begin{pmatrix} \{[Q_i, Q_j]\} & \{[P_i, Q_j]\} \\ \{[Q_i, P_j]\} & \{[P_i, P_j]\} \end{pmatrix} = \begin{pmatrix} 0 & -\delta_{ij} \\ \delta_{ij} & 0 \end{pmatrix} \quad (3)$$

requires that, under linear transformation by a $2D \times 2D$ matrix $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ with four $D \times D$ blocks, which maps $\mathbf{W} \mapsto \mathbf{MW}$, should continue to satisfy

$$\{[\mathbf{W}^T, \mathbf{W}]\} = \{[(\mathbf{MW})^T, \mathbf{MW}]\} = \mathbf{M} \mathbf{\Omega} \mathbf{M}^T = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} =: \mathbf{\Omega} \quad (4)$$

In 2×2 block form,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -\mathbf{AB}^T + \mathbf{BA}^T - \mathbf{AD}^T + \mathbf{BC}^T \\ -\mathbf{CB}^T + \mathbf{DA}^T - \mathbf{CD}^T + \mathbf{DC}^T \end{pmatrix} \quad (5)$$

this entails the $D(2D - 1)$ *symplectic conditions*:

$$\mathbf{AB}^T, \mathbf{CD}^T \text{ are symmetric (and also } \mathbf{A}^T \mathbf{C}, \mathbf{B}^T \mathbf{D}), \quad \mathbf{AD}^T - \mathbf{BC}^T = 1 \quad (6)$$

$$\Rightarrow \mathbf{M}^{-1} = \mathbf{\Omega} \mathbf{M}^T \mathbf{\Omega}^{-1} = \begin{pmatrix} \mathbf{D}^T & -\mathbf{B}^T \\ -\mathbf{C}^T & \mathbf{A}^T \end{pmatrix} \quad (7)$$

Matrices that satisfy (6) are called *symplectic* and provide a minimal matrix representation of linear canonical transformations. The product of two such matrices is also symplectic, and so is the unit $\mathbf{1}$ and the inverse (7). They are thus a $D(2D + 1)$ -dimensional manifold that forms a *group*, denoted $\text{SP}(2D, \mathbb{R})$, where $\mathbf{\Omega}$ is called the symplectic metric matrix.

Geometric Canonical Transforms

In plane paraxial optics (with line $D=1$ screens), canonical transformations are represented by the 3-parameter manifold of 2×2 matrices of unit determinant $\text{Sp}(2, \mathbb{R})$ acting on $\begin{pmatrix} q \\ p \end{pmatrix}$. In the common setups with plane $D=2$ screens, the group of canonical transformations is the 10-parameter symplectic group $\text{Sp}(4, \mathbb{R})$ represented by 4×4 matrices acting on $(\mathbf{q}, \mathbf{p})^T = (q_x, q_y, p_x, p_y)^T$.

The correspondence between quadratic functions of the phase space coordinates, matrices and optical elements is established when the former are expressed in terms of the *Poisson operators* of differentiable functions $f(\mathbf{q}, \mathbf{p})$, given by

$$\{f, \circ\} := \sum_{i=1}^D \left(\frac{\partial f(\mathbf{q}, \mathbf{p})}{\partial q_i} \frac{\partial}{\partial p_i} - \frac{\partial f(\mathbf{q}, \mathbf{p})}{\partial p_i} \frac{\partial}{\partial q_i} \right) \quad (8)$$

acting on phase space functions with the Poisson bracket: $\{f, \circ\}g(\mathbf{q}, \mathbf{p}) = \{f, g\}(\mathbf{q}, \mathbf{p})$. When exponentiated with Taylor series in powers $\{f, \circ\}^n := \{\{f, \circ\}^{n-1}, \circ\}$, one finds in the $D=2$ group $\text{Sp}(4, \mathbb{R})$

3 anamorphic lens parameters:

$$\exp \left\{ - \sum_{i \leq j} c_{ij} q_i q_j, \circ \right\} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ -\mathbf{c} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_{x,x} & c_{x,y} \\ c_{x,y} & c_{y,y} \end{pmatrix} \quad (9)$$

3 anisotropic free space displacements:

$$\exp \left\{ - \sum_{i \leq j} b_{ij} p_i p_j, \circ \right\} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{b} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_{x,x} & b_{x,y} \\ b_{x,y} & b_{y,y} \end{pmatrix} \quad (10)$$

4 anisotropic magnifiers:

$$\exp \left\{ - \sum_{i,j} a_{ij} q_i p_j, \circ \right\} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} e^a & \mathbf{0} \\ \mathbf{0} & e^{-a^T} \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_{x,x} & a_{x,y} \\ a_{y,x} & a_{y,y} \end{pmatrix} \quad (11)$$

When applying matrices on phase space vectors to represent canonical transformations in an optical setup, one should remember that (if light goes from left to right) the leftmost optical element will correspond with the rightmost matrix factor (to act first), and successively in inverse order.

The group of $D=2$ canonical transformations $\text{Sp}(4, \mathbb{R})$ contains 4 *rotations* of phase space which close into the *Fourier group*, which is isomorphic to the unitary group of 2×2 (complex) matrices $\text{U}(2)_{\mathbb{F}}$. This contains x - and y - fractional Fourier transforms (FrFT), gyrations and screen rotations, which are also *generated* by exponentials of Poisson operators of quadratic functions of phase space given below. Writing for brevity $c := \cos \theta$ and $s := \sin \theta$, and identifying their generator functions, they are

$$\begin{array}{cccc} \text{isotropic FrFT} & \text{anisotropic FrFT} & \text{gyrations} & \text{rotations} \\ \begin{pmatrix} c & 0 & -s & 0 \\ 0 & c & 0 & -s \\ s & 0 & c & 0 \\ 0 & -s & 0 & c \end{pmatrix}; & \begin{pmatrix} c & 0 & -s & 0 \\ 0 & c & 0 & s \\ s & 0 & c & 0 \\ 0 & -s & 0 & c \end{pmatrix}, & \begin{pmatrix} c & 0 & 0 & -s \\ 0 & c & -s & 0 \\ 0 & s & c & 0 \\ s & 0 & 0 & c \end{pmatrix}, & \begin{pmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & c & -s \\ 0 & 0 & s & c \end{pmatrix} \\ \frac{1}{2}(p_x^2 + p_y^2 + q_x^2 + q_y^2) & \frac{1}{2}(p_x^2 - p_y^2 + q_x^2 - q_y^2) & p_x p_y + q_x q_y & q_x p_y - q_y p_x \end{array} \quad (12)$$

These matrices are both symplectic and *orthogonal* (called ortho-symplectic); the isotropic FrFT commutes with the other three and is generated by the isotropic harmonic oscillator Hamiltonian. In general D dimensions, $\text{Sp}(2D, \mathbb{R})$ contains a $\text{U}(D)_{\mathbb{F}}$ subgroup of such phase space rotations represented by $2D \times 2D$ ortho-symplectic matrices.

Canonical Integral Transforms

The transformation between input and output wavefields due to the passage through an ideal paraxial optical setup are canonical transforms. Their kernel was found by Stuart A. Collins in optics and, independently, by Marcos Moshinsky and Christiane Quesne as a fundamental question in quantum mechanics, in 1970 and 1971.

Wave optics and quantum mechanics provide a Hilbert space of square-integrable wavefunctions $f(\mathbf{q})$ of $\mathbf{q} \in \mathbb{R}^D$ on which the phase space operators, satisfying (2) act as $\hat{Q}_i f(\mathbf{q}) = q_i f(\mathbf{q})$ and $\hat{P}_i f(\mathbf{q}) = -i \partial f(\mathbf{q}) / \partial q_i$ (for $\hbar = 1$ or $\hbar = 1$). Canonical transformations $\mathcal{C}_M$ characterized by generic symplectic matrices $\mathbf{M}$ will act on these operators through,

$$\mathcal{C}_M \begin{pmatrix} \hat{\mathbf{Q}} \\ \hat{\mathbf{P}} \end{pmatrix} \mathcal{C}_M^{-1} = \mathbf{M}^{-1} \begin{pmatrix} \hat{\mathbf{Q}} \\ \hat{\mathbf{P}} \end{pmatrix} = \begin{pmatrix} \mathbf{D}^T \hat{\mathbf{Q}} & -\mathbf{B}^T \hat{\mathbf{P}} \\ -\mathbf{C}^T \hat{\mathbf{Q}} & +\mathbf{A}^T \hat{\mathbf{P}} \end{pmatrix} \quad (13)$$

The *inverse* matrix (7) is required, as can be easily ascertained verifying that $\mathcal{C}_{M_1} \mathcal{C}_{M_2} = \sigma \mathcal{C}_{M_1 M_2}$, but for a sign σ that is not detectable in (13).

Regarding the action of $\mathcal{C}_M$ on wavefunctions, since the Fresnel transform of free flight is an integral transform, and so is the fractional Fourier transform, canonical transformations $\mathcal{C}_M$ will be *integral transform* operators that act as

$$f_M(\mathbf{q}) \equiv (\mathcal{C}_M f)(\mathbf{q}) := \int_{\mathbb{R}^D} d^D \mathbf{q}' C_M(\mathbf{q}, \mathbf{q}') f(\mathbf{q}') \quad (14)$$

where the integral kernel $C_M(\mathbf{q}, \mathbf{q}')$ is found applying $\mathcal{C}_M$ to $\hat{Q}_i f$ and $\hat{P}_i f$,

$$\mathcal{C}_M(\hat{Q}_i f) = (\mathcal{C}_M \hat{Q}_i \mathcal{C}_M^{-1}) \mathcal{C}_M f = \sum_{j=1}^D (D_{j,i} \hat{Q}_j - B_{j,i} \hat{P}_j) f_M \quad (15)$$

$$\mathcal{C}_M(\hat{P}_i f) = (\mathcal{C}_M \hat{P}_i \mathcal{C}_M^{-1}) \mathcal{C}_M f = \sum_{j=1}^D (-C_{j,i} \hat{Q}_j + A_{j,i} \hat{P}_j) f_M \quad (16)$$

Replacing the integral (14) on the left, $\hat{Q}_i$ and $\hat{P}_i$ will act on the $\mathbf{q}'$ argument of f , while on the right they act on the exterior argument $\mathbf{q}$ of the kernel $C_M(\mathbf{q}, \mathbf{q}')$. The operator $-i\partial/\partial q'_i$ in the left can be integrated by parts and applied on the $\mathbf{q}'$ argument of the kernel; this leads to 2D simultaneous differential equations:

$$q'_i C_M(\mathbf{q}, \mathbf{q}') = \sum_{j=1}^D \left(D_{j,i} q_i + i B_{j,i} \frac{\partial}{\partial q'_j} \right) C_M(\mathbf{q}, \mathbf{q}') \quad (17)$$

$$-i \frac{\partial}{\partial q'_i} C_M(\mathbf{q}, \mathbf{q}') = \sum_{j=1}^D \left(C_{j,i} q_i + i A_{j,i} \frac{\partial}{\partial q'_j} \right) C_M(\mathbf{q}, \mathbf{q}') \quad (18)$$

Multiplying these vector equations by $\mathbf{B}^{-1}$ and proposing a quadratic exponential solution, with appropriate normalization one finds

$$C_M(\mathbf{q}, \mathbf{q}') = K_M \exp i \left(\frac{1}{2} \mathbf{q}^T \mathbf{B}^{-1} \mathbf{D} \mathbf{q} - \mathbf{q}^T \mathbf{B}^{-1} \mathbf{q}' + \frac{1}{2} \mathbf{q}'^T \mathbf{A} \mathbf{B}^{-1} \mathbf{q}' \right) \quad (19)$$

$$K_M := \frac{1}{\sqrt{(2\pi i)^D \det \mathbf{B}}} \equiv \frac{e^{-i\pi D/4} \exp i(-\frac{1}{2} \arg \det \mathbf{B})}{\sqrt{(2\pi)^D |\det \mathbf{B}|}} \quad (20)$$

where $\arg \det \mathbf{B} \in \{0, \pm\pi\}$. When $M_0 = \begin{pmatrix} \mathbf{A} & 0 \\ \mathbf{C} & \mathbf{A}^{-1} \end{pmatrix}$, the kernel collapses to a Dirac δ^D and the transformation is no longer integral, but *point-to-point*:

$$(\mathcal{C}_{M_0} f)(\mathbf{q}) = \frac{\exp i(\frac{1}{2} \mathbf{q}^T \mathbf{C} \mathbf{A}^{-1} \mathbf{q})}{\sqrt{\det \mathbf{A}}} f(\mathbf{A}^{-1} \mathbf{q}) \quad (21)$$

The normalization and phase are determined by the requirement that, at the unit transformation, $\lim_{M \rightarrow 1} C_M(\mathbf{q}, \mathbf{q}') = \delta^D(\mathbf{q} - \mathbf{q}')$. The canonical integral transforms are *unitary* in $\mathcal{L}^2(\mathbb{R}^D)$:

$$(f, \mathcal{C}_M g) = (\mathcal{C}_M^\dagger f, g) = (\mathcal{C}_M^{-1} f, g) = (\mathcal{C}_{M^{-1}} f, g) \quad (22)$$

because

$$C_M(\mathbf{q}', \mathbf{q})^* = C_{M^{-1}}(\mathbf{q}, \mathbf{q}') \quad (23)$$

As in the geometric case, integral canonical transforms are fully invertible and no information is lost between input and output images.

Because it is emblematic, we should write the $D=1$ case of (19)–(21) explicitly; it is

$$C_M(q, q') = \frac{1}{\sqrt{2\pi i B}} \exp \frac{i}{2B} (Dq^2 - 2qq' + Aq'^2) \quad (24)$$

where $1/\sqrt{2\pi i B} \equiv e^{-i\pi/4} \exp(-\frac{i}{2} \arg B) / \sqrt{2\pi |B|}$, and the lower-triangular case (21) follows directly.

Issues Pertaining Integral Transforms

Canonical integral transforms include the Fresnel and fractional Fourier transforms as particular one-parameter subgroups of translation and a phase space rotation (of a spin-like nature – see below). They also include the time evolution generated by all quadratic wave/quantum operators and allow for complex extensions of some transformation parameters, thereby adding coherent states to the functions subject to group theoretical treatment.

The Metaplectic Sign

The square root in (20) and (24) indicates a double valuation: indeed, when for symplectic $M_1 M_2 = M_3$ we check carefully the composition of the integral kernels we find that

$$\int_{\mathbb{R}^D} d^D \mathbf{q}' C_{M_1}(\mathbf{q}, \mathbf{q}') C_{M_2}(\mathbf{q}', \mathbf{q}'') = \sigma_{1,2,3} C_{M_3}(\mathbf{q}, \mathbf{q}'') \quad (25)$$

with the *metaplectic* sign $\sigma_{1,2,3} = \text{sign}(\det B_3 / \det B_1 B_2)$.

Although overall signs are not commonly detected in wavefields, this sign ‘ambiguity’ is real and mathematically unavoidable. It is due to the fact that the set of canonical *integral* transforms covers the geometric $\text{Sp}(2D, \mathbb{R})$ group *twice*. The issue reminds us of the double cover that the spin group affords over the group of space rotations. A handle on this matter is provided by the kernel of the 1D Fourier transform $\mathcal{F}$, namely $F(q, q') = e^{iqq'} / \sqrt{2\pi}$, versus the *canonical* Fourier transform kernel for $F = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \Omega$, which is $C_F(q, q') = e^{i\pi/4} e^{iqq'} / \sqrt{2\pi}$. Thus $C_F = e^{i\pi/4} \mathcal{F}$ and, while $\mathcal{F}^4 = 1$, one has $C_F^4 = -1$, and only $C_F^8 = 1$. The manifold of the symplectic group has the topology of a one-sheeted hyperboloid, whose *waist* is the ‘circle’ subgroup of isotropic fractional Fourier transforms, which may be infinitely covered. The double cover is called the *metaplectic* group $\text{Mp}(2D, \mathbb{R})$, of which linear canonical transforms are a faithful representation.

Exponentials of Second-Order Differential Operators

There is an important connection between canonical integral transforms and exponentials of *second-order* differential operators which follows the generation of the geometric $\text{Sp}(2D, \mathbb{R})$ presentation in (9)–(11) and in (12), replacing q_i, p_j with $\hat{Q}_i, \hat{P}_j$, Poisson brackets with commutators, and $\exp(\circ)$ with $\exp(i\circ)$.

There follow properties of self-reproduction of certain functions under a subgroup of canonical transforms: for example, writing $C_M \equiv C(M)$, the Hermite-Gauss wavefunctions $\Psi_n(q) = e^{-q^2/2H_n(q)/\sqrt{2^n n!} \sqrt{\pi}}$ in a planar waveguide satisfy

$$\exp\left(it \frac{1}{2} (\hat{P}^2 + \hat{Q}^2)\right) \Psi_n(q) = C\left(\exp t \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\right) \Psi_n(q) = \left(C \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \Psi_n\right)(q) = e^{i(n+\frac{1}{2})t} \Psi_n(q) \quad (26)$$

Upon decomposing $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ \gamma & \alpha^{-1} \end{pmatrix} \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$, and using (26) and (21), one can then find the general linear canonical transform of these Hermite-Gauss functions as

$$\left(C \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Psi_n\right)(q) = \frac{\exp\left(-\frac{d-ic}{a+ib} \frac{q^2}{2}\right)}{\sqrt{2^n n! \sqrt{\pi} \left(\frac{a+ib}{a-ib}\right)^n}} \frac{H_n(q/\sqrt{a^2+b^2})}{\sqrt{a+ib}} \quad (27)$$

Similar manipulations can be performed with the eigenfunctions of all quadratic operators: the repulsive oscillator $\frac{1}{2}(\hat{P}^2 - \hat{Q}^2)$, free waves $\exp(ikq)/\sqrt{2\pi}$ of $\frac{1}{2}\hat{P}^2$, eigenfunctions $\sim q^{i\lambda-1/2}$ of the dilatation generator $\frac{1}{2}(\hat{Q}\hat{P} + \hat{P}\hat{Q})$, or $\delta(q-\lambda)$ of $\hat{Q}^2$. Finally, Airy eigenfunctions of the linear potential $\frac{1}{2}\hat{P}^2 + \hat{Q}$ can be brought into the picture with the addition of linear terms of $\hat{Q}$, $\hat{P}$, and $\hat{1}$ to generate the six-parameter *Weyl-symplectic* group $\text{WSp}(2, \mathbb{R})$.

Complex Extension of Canonical Transforms

The parameters of the symplectic groups can be extended into a *complex* domain, characterizing thus $\text{Sp}(2D, \mathbb{C})$. The integral kernel (19) remains valid as long as it is an oscillatory and/or *decreasing* Gaussian in $\mathbf{q}' \in \mathbb{R}^D$. This condition requires that the term $i \mathbf{q}'^T \mathbf{A} \mathbf{B}^{-1} \mathbf{q}'$ must have a zero or negative real part. In the 1D case of (24) this means that $\text{Im}(b/a) \leq 0$.

Out of a Dirac delta at c , $\delta_c(q) \equiv \delta(q-c)$, one can produce a Gaussian of width $\omega > 0$ through the complex canonical transform

$$G_\omega(q-c) := \left(C \begin{pmatrix} 1 & -i\omega \\ 0 & 1 \end{pmatrix} \delta_c\right)(q) = \frac{1}{\sqrt{2\pi\omega}} \exp\left(-\frac{1}{2\omega}(q-c)^2\right) \quad (28)$$

Thus $c \begin{pmatrix} 1 & -i\tau \\ 0 & 1 \end{pmatrix}$ is the complex canonical transform for *diffusion* in time $\tau \geq 0$. It is no longer unitary in $\mathcal{L}^2(\mathbb{R})$, but within certain conditions can be made unitary in Bargmann-type Hilbert spaces over the complex domain.

The evolution of Gaussian wavefunctions in a harmonic waveguide, under fractional Fourier transforms, or under a quantum harmonic oscillator potential in time τ , can be computed using matrix algebra,

$$G_\omega(q-c; \tau) = \left(C \begin{pmatrix} \cos \tau & -\sin \tau \\ \sin \tau & \cos \tau \end{pmatrix} C \begin{pmatrix} 1 & -i\omega \\ 0 & 1 \end{pmatrix} \delta_c\right)(q) = \left(C \begin{pmatrix} \cos \tau & -\sin \tau & -i\omega \cos \tau \\ \sin \tau & \cos \tau & -i\omega \sin \tau \end{pmatrix} \delta_c\right)(q) \quad (29)$$

Coherent states are also complex canonical transforms of a Dirac δ_c . These are $\Upsilon_c(q) := \exp(c\hat{A}^\dagger) \Psi_0(q)$, with the raising operator $\hat{A}^\dagger := \frac{1}{\sqrt{2}}(\hat{Q} - i\hat{P})$, the ground oscillator state $\Psi_0(q) = \pi^{-1/4} e^{-q^2/2}$, and of displaced center $c\sqrt{2}$,

$$\Upsilon_c(q) = \pi^{-1/4} e^{\frac{1}{2}c^2} \exp\left(-\frac{1}{2}(q-c\sqrt{2})^2\right) = (2\pi)^{1/4} \left(C \begin{pmatrix} 1/\sqrt{2} & -i/\sqrt{2} \\ -i/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \delta_c\right)(q) \quad (30)$$

This the *Bargmann* transform that also maps the Hermite-Gauss wavefunctions $\Psi_n(q)$ on power functions $\sim z^n$ of a variable z on the complex plane. The time evolution of $\mathcal{Y}_c(q)$ under a harmonic waveguide can be found as in (29), and results in the center $c(\tau) = ce^{i\tau}$ oscillating with position $\sqrt{2\text{Re } c(\tau)}$ and momentum $\sqrt{2\text{Im } c(\tau)}$, drawing out circles on phase space.

Radial Canonical Transforms

When an $\text{Sp}(2D, \mathbb{R})$ matrix has diagonal blocks $\begin{pmatrix} a1 & b1 \\ c1 & d1 \end{pmatrix}$, canonical transforms of functions with definite angular momentum are reduced to 1D *radial* canonical transforms that conserve the angular momenta. The 2D kernel can be separated in polar coordinates as

$$q_1 = r \cos \theta, \quad q_2 = r \sin \theta, \quad r \in \mathbb{R}^+ = [0, \infty), \quad \theta \in \mathbb{R} \bmod 2\pi, \quad (31)$$

resulting in the measure $d^2q = r dr d\theta$. The generator of rotations, $\hat{L} = \hat{Q}_1 \hat{P}_2 - \hat{Q}_2 \hat{P}_1 = -i\partial_\theta$ is invariant under all $c \begin{pmatrix} a1 & b1 \\ c1 & d1 \end{pmatrix}$'s due to (6) and so are its eigenspaces of functions $f(r)e^{im\theta}$, for all integers $m \in \mathbb{Z}$. When ϕ is the angle between $\mathbf{q} \cdot \mathbf{q}' = rr' \cos \phi$, in each of these spaces we can perform the integral over ϕ to obtain

$$e^{im\theta} C_M^{(m)}(r, r') := \int_{-\pi}^{\pi} d\phi C_M(\mathbf{q}, \mathbf{q}') e^{im(\theta - \phi)} \quad (32)$$

This provides kernel for the m -radial canonical transforms,

$$f_M^{(m)}(r) \equiv (C_M^{(m)} f)(r) = \int_{\mathbb{R}^+} r dr' C_M^{(m)}(r, r') f(r') \quad (33)$$

$$C_M^{(m)}(r, r') = e^{i\pi(m-1)/2} \frac{1}{b} \exp\left(\frac{i}{2b}(dr^2 + ar'^2)\right) J_m\left(\frac{rr'}{b}\right) \quad (34)$$

where $J_m(z)$ is the Bessel function of the first kind. These transforms are unitary in the Hilbert space $\mathcal{L}^2(\mathbb{R}^+)$, with the inner product

$$(f, g)_{\mathcal{L}^2(\mathbb{R}^+)} := \int_0^\infty r dr f(r)^* g(r) \quad (35)$$

Among the 1D quadratic generators, squared momentum now displays a 'centrifugal' term:

$$\hat{P}_{(m)}^2 = -\left(\frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} - \frac{m^2}{r^2}\right) \quad (36)$$

while $\frac{1}{2}(\hat{Q}_{(m)} \hat{P}_{(m)} + \hat{P}_{(m)} \hat{Q}_{(m)}) = r d/dr$ and $\hat{Q}_{(m)}^2 = r^2$. The eigenfunctions will thus be of the corresponding quantum potentials, such as Laguerre-Gauss functions for the harmonic waveguide Hamiltonian $\frac{1}{2}(\hat{P}_{(m)}^2 + \hat{Q}_{(m)}^2)$.

Of course, the set $C_M^{(m)}$ with $M \in \text{Sp}(2, \mathbb{R})$ also represents that group, and belongs to *Bargmann's discrete* representation series D_+^k for $k = \frac{1}{2}(m+1)$. The well-known *oscillator* representation of $\text{Sp}(2, \mathbb{R})$ that occupied the previous two sections belongs to $D_+^{1/4} \oplus D_+^{3/4}$ for even and odd states on $\mathbb{R}$. There, $m = \mp \frac{1}{2}$ corresponds to kernels with $J\left\{\mp \frac{1}{2}(z) = \sqrt{2/\pi z} \begin{cases} \cos z \\ \sin z \end{cases}\right.$ on $\mathbb{R}^+$, which on the full real line reconstitute the oscillator representation that is double valued because it corresponds to half-integer m . For integer m the representations are single-valued so the composition sign σ in (25) is unity.

Acknowledgements

Support for this research has been provided by the *Óptica Matemática* projects of UNAM, presently PAPIIT IN101115.

See also: The Fractional Order Fourier Transform and Fresnel Diffraction

Further Reading

- Bargmann, V., 1947. Irreducible unitary representations of the Lorentz group. *Ann. Math.* 48, 568–642.
 Bargmann, V., 1961. On a Hilbert space of analytic functions and an associated integral transform, Part I. *Commun. Pure Appl. Math.* 20, 187–214.
 Bargmann, V., 1970. Group representation in Hilbert spaces of analytic functions. In: Gilbert, P., Newton, R.G. (Eds.), *Analytical Methods in Mathematical Physics*. New York, NY: Gordon & Breach, pp. 27–63.
 Collins Jr, S.A., 1970. Lens-system diffraction integral written in terms of matrix optics. *J. Opt. Soc. Am.* 60, 1168–1177.
 Condon, E.U., 1937. Immersion of the Fourier transform in a continuous group of functional transformations. *Proc. Natl. Acad. Sci.* 23, 158–163.
 Forbes, G.W., Man'ko, V.I., Ozaktas, H.M., Simon, R., Wolf, K.B. (Eds.), 2000. Feature issue on wigner distributions and phase space in optics. *J. Opt. Soc. Am. A* 17 (12), 2440–2463.
 Healy, J.J., Alper Kutay, M., Ozaktas, H.M., Sheridan, J.T. (Eds.), 2016. *Linear Canonical Transforms. Theory and Applications*. New York: Springer.
 Kauderer, M., 1994. *Symplectic Matrices. First Order Systems and Special Relativity*. Singapore: World Scientific.

- Liberman, S., Wolf, K.B., 2015. Independent simultaneous discoveries visualized through network analysis: The case of linear canonical transforms. *Scientometrics* 104, 715–735.
- Moshinsky, M., Quesne, C., 1971. Linear canonical transformations and their unitary representation. *J. Math. Phys.* 12, 1772–1780.
- Namias, V., 1980. The fractional order Fourier transform and its applications in quantum mechanics. *IMA J. Appl. Math.* 25, 241–265.
- Ozaktas, H.M., Zalevsky, Z., Alper Kutay, M., 2001. *The Fractional Fourier Transform with Applications in Optics and Signal Processing*. Chichester: Wiley.
- Quesne, C., Moshinsky, M., 1971. Linear canonical transformations and matrix elements. *J. Math. Phys.* 12, 1780–1783.
- Rodrigo, J.A., Alieva, T., Bastiaans, T.J., 2011. Phase space rotators and their applications in optics. In: Cristóbal, G., Schelkens, P., Thienpont, H. (Eds.), *Optical and Digital Image Processing: Fundamentals and Applications*. Weinheim: Wiley-VCH Verlag, pp. 251–271.
- Sánchez-Mondragón, J., Wolf, K.B. (Eds.), 1986. *Lie Methods in Optics*. Lecture Notes in Physics 250. Heidelberg: Springer.
- Simon, R., Wolf, K.B., 2000. Fractional Fourier transforms in two dimensions. *J. Opt. Soc. Am. A* 17, 2368–2381.
- Wolf, K.B., 1974. Canonical transforms. I. Complex linear transforms. *J. Math. Phys.* 15, 1295–1301.
- Wolf, K.B., 1979. *Integral Transforms in Science and Engineering*. New York: Plenum.
- Wolf, K.B. (Ed.), 1989. *Lie Methods in Optics. II*. Lecture Notes in Physics 352. Heidelberg: Springer.
- Wolf, K.B., 2004. *Geometric Optics on Phase Space*. Heidelberg: Springer.

Phase-Space Representations of Freeform Optical Systems

Alois M Herkommer, University of Stuttgart, Stuttgart, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Since the early days of optics the classical shape of an optical surface used to be spherical or at least rotational symmetric, mainly because for this shape accurate manufacturing methods have been available. Accordingly the majority of optical systems was, and still is, rotational symmetric. Based on that symmetry assumption the well-known aberration theories, such as Seidel aberration theory (Gross, 2005), or Aldis calculus (Cox, 1964; Brewer, 1976) have mathematically been developed and intensely used for system analysis. However modern manufacturing methods, such as high precision diamond turning or 3D-printing, today allow for the accurate generation of optical surfaces which are free of any symmetry. Such surfaces are attractive to the optical designer, since they offer additional degrees of freedom to correct for astigmatic or other aberrations in complex folded systems (Thompson and Rolland, 2012). However, as freeform systems per definition do not allow any assumptions on symmetry, the well-developed aberration theories cannot be applied and more sophisticated analysis, such as nodal aberration theory (Fuerschbach *et al.*, 2014), is required. However these approaches are quite complex and require in-depth mathematical analysis of the system. Therefore it is desirable to find alternate simple methods to calculate and visualize aberration contributions within freeform systems, in order to allow the designer to analyze the optical system in a similar simple way as for rotational symmetric systems.

In this article we will apply the method of phase space in optics (Testorf *et al.*, 2010; Torre, 2005) for this purpose. The method allows an alternate access to the analysis of optical systems by employing an illustration of positions and angles of limiting rays. This has proven to be especially helpful for paraxial system analysis, but as we will see also aberrations can be very intuitively discussed in this context. Since symmetry assumptions are not required, this method proves to be a general tool to analyze and visualize freeform systems.

Within this article we will first introduce the general principles of phase space in geometrical optics. In a next step we will employ the method to visualize aberrations in symmetric and non-symmetric systems for the simplified two-dimensional case. Extending the analysis to all dimensions is however possible and finally allows an exact mathematical treatment and extraction of individual aberration contributions.

The Concept of Phase Space in Geometrical Optics

The concept of phase space in geometrical optics is based on an illustration of ray angles versus ray positions. This is most easily explained if we for the moment only consider meridional rays in the xz -plane of an optical system, as illustrated in Fig. 1. Any meridional ray at a given z -position is completely defined by two quantities, namely the distance to the optical axis x and the angle θ to the optical axis. If we furthermore weight the ray direction with the refractive index n of the propagation medium, we can define

$$u = n \cdot \tan \theta \quad (1)$$

Thus the vector $\mathbf{r} = (x, u)^T$ completely defines the ray-tracing behavior of any ray in an optical system. The vector components (x, u) of the ray can be illustrated as one point in a xu -diagram, which is called the phase space diagram. Propagation of the ray in an optical system in consequence corresponds to trajectories, respectively transformations in phase space. In Fig. 1 the relationship between standard ray propagation versus the corresponding trajectory in phase space is illustrated.

In the paraxial regime the ray-propagation and the corresponding trajectory in phase space is closely related to the so-called ABCD matrix-optics, respectively linear optical systems theory (Kloos, 2007). In this approximation the real optical system can mathematically be replaced by a sequence of linear transformations. This is illustrated in Fig. 2 where a sequence of reference surfaces D_i and D'_i , located at the vertex of the refractive surface, is introduced. The paraxial ray propagation then corresponds to

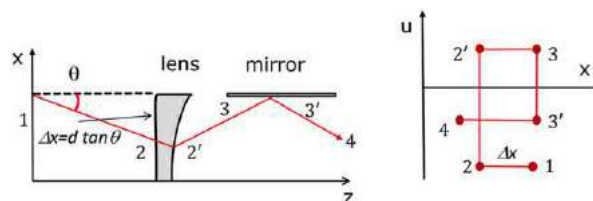


Fig. 1 Relationship between ray-propagation (left) and phase space trajectory (right) in an optical system. Reproduced from Rausch, D., Herkommer, A. 2012. Phase space approach to the use of integrator rods and optical arrays in illumination systems. Advanced Optical Technologies. Available at: <https://doi.org/10.1515/aot-2011-0002>.

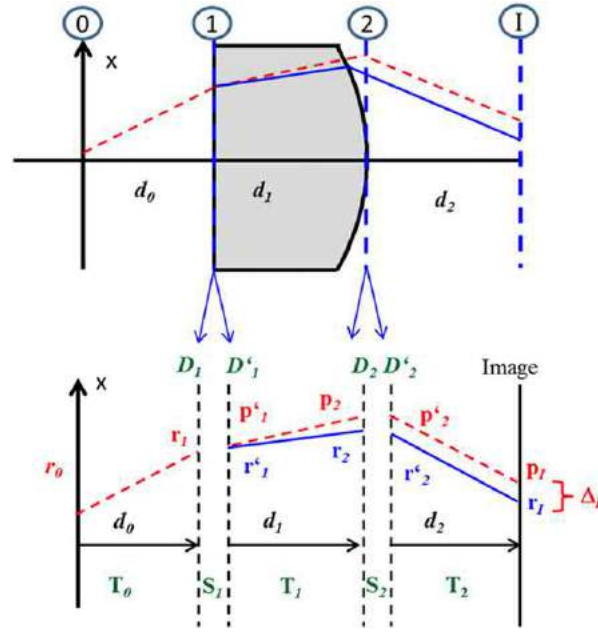


Fig. 2 Real geometry (top) and paraxial representation (bottom) of a refractive lens system. The dashed line represents the paraxial ray-tracing, the solid line real ray-tracing.

the application of propagation steps T_i in between the surfaces D_i and D_{i+1} , and paraxial refraction S_i from D_i to D'_i at the spherical surfaces. If we denote the paraxial ray vector by $\mathbf{p}$, the matrix-optics rules define the following set of linear transformations as the ray propagates:

$$\mathbf{p}_{i+1} = T_i \cdot \mathbf{p}_i \text{ where } T_i = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix} \quad (2)$$

$$\mathbf{p}'_i = S_i \cdot \mathbf{p}_i \text{ where } S_i = \begin{pmatrix} 1 & 0 \\ -\phi & 1 \end{pmatrix} \quad (3)$$

Here d denotes the distance between surfaces along the optic axis and ϕ is the optical power of the surface. Thus any paraxial ray, corresponding to an input ray $\mathbf{r}_0$ can paraxial be traced to surface D_i by application of a sequence of linear operations. Mathematically this is:

$$\mathbf{p}_i = \underbrace{T_{i,i-1} * S_{i-1} \cdots S_1 * T_{1,0}}_{M_{i,0}} * \mathbf{r}_0 \quad (4)$$

We can now translate the paraxial ray-tracing behavior into transformations within phase space. A visualization of appropriate reference rays in phase space, for example, the marginal ray and the chief ray, reveals the basic transformation properties and parameters of an optical system. This is illustrated in **Fig. 3** for a 2f-lens system, representing the propagation from the front focal plane to the back focal plane of a single lens. According to Eq. (2) the linear matrix T for free propagation corresponds to a shear into the x -direction, whereas a paraxial lens S , Eq. (2), corresponds to a shear in the angular u -direction. Thus the total propagation from the front focal plane to the back focal plane corresponds to a 90° rotation in phase space, which is expected since the 2f lens system translates angles into position and vice versa. It is also worth to note that area in phase space has an important interpretation. If the chief ray $(\bar{x}, \bar{u})$ and the marginal ray (x, u) of an optical system are employed as reference points, then the product $L = \bar{u} \cdot x - \bar{x} \cdot u$ is equivalent to the Lagrange invariant (Gross, 2005), which is a conserved quantity in conventional optical systems. In the phase space diagram this product corresponds to the shaded area, as illustrated in **Fig. 3**. In this sense we can generally state that area in phase space is conserved and paraxial propagation corresponds to a sequence of affine transformations to this area.

Visualization of Aberrations in Phase Space

Already analyzing the linear behavior in phase space reveals a lot of information about the system, especially the primary quantities like magnification or focal length (Kloos, 2007). However the paraxial equations do not contain any aberration effects, since each paraxial ray will behave in the same linear way and lead to the same transformation behavior. For aberration analysis it is therefore necessary to record and visualize the exact ray-tracing behavior and compare it to the above linear behavior. To do so,

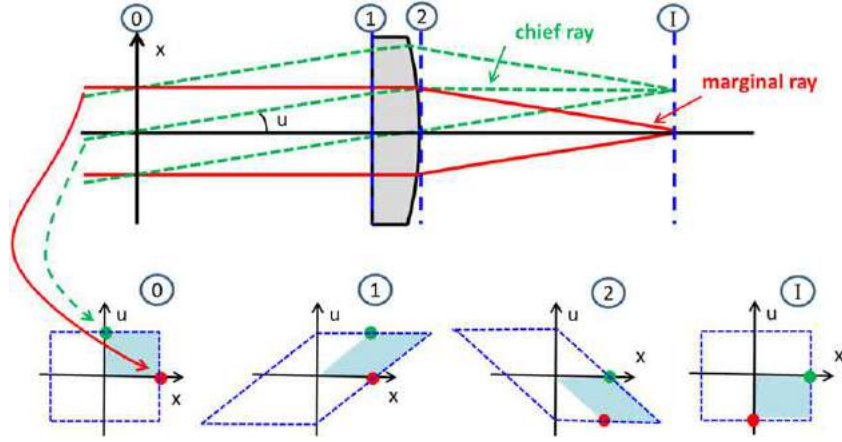


Fig. 3 Illustration of chief and marginal reference rays in phase space and the paraxial transformation representing the system. The enclosed area corresponds to the Lagrange invariant.

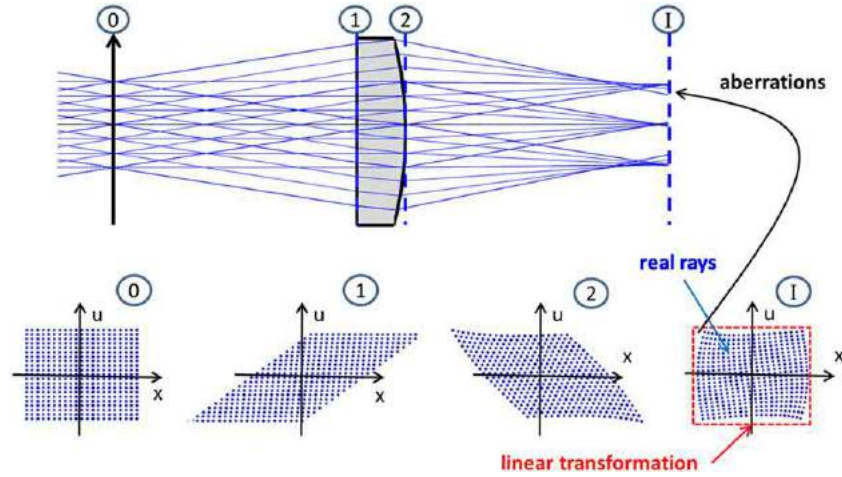


Fig. 4 Illustration of real ray behavior in phase space, revealing nonlinear distortions corresponding to aberrations.

we can introduce a series of planar dummy surfaces D_i and D'_i at the vertex of each refractive surface, as illustrated in Fig. 2. In a sequential ray-tracer the surface D_i is placed directly before the real surface, and records the input ray, whereas the surface D'_i is arranged at the same position but sequentially arranged after the real surface and thus monitors the output ray. By performing raytracing the real ray intersection data $\mathbf{r}_i$ at each “input”-dummy surface D_i , and $\mathbf{r}'_i$ on the corresponding “output”-dummy surfaces D'_i can be recorded. Since the dummy surfaces are at the same position as the paraxial power of the linear system model, we can directly compare the paraxial behavior and real raytracing behavior. To illustrate this, a grid of rays, sampling the object (x) and aperture (u) space, is traced through the system and each ray is plotted in the phase space diagram. The result is presented in Fig. 4 and reveals the phase space transformation of the real system.

If we compare this result to the paraxial transformation as was shown in the previous Fig. 3 we notice, that non-linear deformations of the ray-pattern, respectively the phase space area, become apparent. At the image plane these deviations have a clear meaning and are illustrated in detail in Fig. 5. Since one ray fan at the image plane, corresponding to one column of the ray-grid, represents a set of rays that arrive at the same image point but under different angles, the lateral shift relative to the central ray simply corresponds to the transverse aberrations in x -direction. This is depicted in Fig. 5(b). In other words at the image plane we can define a general aberration vector, describing the difference between the linear ray $\mathbf{p}_I$ and the corresponding real ray $\mathbf{r}_I$

$$\Delta_I = \mathbf{r}_I - \mathbf{p}_I = \begin{pmatrix} \Delta x_I \\ \Delta u_I \end{pmatrix} \quad (5)$$

Obviously the spatial component Δx_I of this vector then represents the transverse aberration of this ray. The angular component can also be interpreted as the pupil aberration of the system. Following along these lines, we can define an aberration vector at any surface i within the system in the same way.

$$\Delta_i = \mathbf{r}_i - \mathbf{p}_i \quad (6)$$

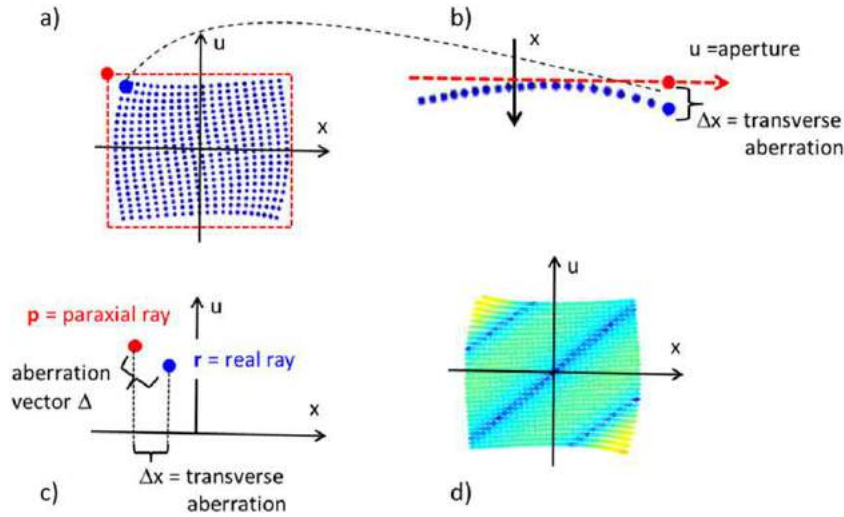


Fig. 5 Difference in ray positions at the image corresponding to aberrations: (a) real ray grid at the image plane and linear behavior, (b) one ray fan corresponding to transverse aberrations (c) aberration vector in phase space, (d) color coded ray grid, where the color indicates the length of the aberration vector.

This vector represents the local deviation of the real ray from the linear ray in phase space and is shown in [Fig. 5\(c\)](#). The length of this aberration vector may thus be interpreted as a general measure of the current magnitude of the aberration of this ray. In consequence we can visualize the aberration of any ray in phase space by illustrating the length of the corresponding aberration vector $|\Delta|$, for example, as a color code. In [Fig. 5\(d\)](#) we show the color coded aberration field at the image.

It is worth to note, that since we have defined $u = n \tan \theta$ via [Eq. \(1\)](#), the free propagation between parallel reference surfaces D_i and D_{i+1} is of exactly the same linear form as the paraxial matrix in [Eq. \(2\)](#). Thus real raytracing in between parallel dummy surfaces exactly corresponds to the paraxial raytracing transformation.

$$\mathbf{r}_{i+1} = \mathbf{p}_{i+1} = \mathbf{T}_i \cdot \mathbf{r}_i \quad (7)$$

In other words free propagation does not introduce differences between paraxial and real ray propagation, i.e., aberrations. Therefore as expected the aberrations within the system are only due to the spherical surfaces and correspond to the difference between the real raytracing via the spherical interface according to Snell's law versus the linear behavior according to the paraxial refraction ([Herkommer, 2013](#)).

Visualizing Aberrations in Freeform Systems

The above described method is quite general, since all that is required to calculate the aberration vector of [Eq. \(6\)](#) are real ray-tracing data and linear ray-propagation data. In consequence this analysis can be transferred to more general optical systems, such as freeform systems ([Chen and Herkommer, 2016](#)). To illustrate this we use the following freeform prism from literature ([Takahashi, 1997](#)), which is often employed as a test and reference system for freeform designs. The design is illustrated in [Fig. 6\(a\)](#) and the detailed design data are listed in [Table 1](#). The illustrated prism geometry is, for example, employed in head-mounted displays to image a display into the pupil of the eye. Note that the design is usually obtained backwards, so in [Fig. 6\(a\)](#) the eye-pupil corresponds to the entrance pupil of the prism and the display corresponds to the image.

To illustrate aberrations in this freeform system the paraxial and real ray propagation in phase space can be compared, just as before for the rotational symmetric case. Again we denote the real rays by $\mathbf{r}$ and the paraxial rays by $\mathbf{p}$. However in a freeform system moreover a reference ray, serving as the reference axis and origin of the coordinate system, has to be defined, as we propagate through the system. In the rotational symmetric case this was not required, since the rotation axis naturally and uniquely defines the optical axis, however in freeform systems the reference has to be fixed. Here we choose the center ray on the pupil starting with an angle of 0° as a reference axis, as illustrated in [Fig. 6\(b\)](#). This reference ray is traced through the freeform system, defining surface intersection points at each surface $S1$ to $S4$. At each intersection we again insert a pair of dummy surfaces D_i and D'_i similar to the rotational symmetric system. The position of both dummy surfaces is identical to the reference ray intersection point, however again the first dummy is mathematically before, whereas the second dummy is after the surface in a sequential ray-tracer. Moreover all dummy surfaces have to be tilted to be exactly orthogonal to the reference ray in order to properly define the ray angles relative to the reference ray. This is illustrated in [Fig. 6\(b\)](#) for the example of surface 2 only. The positions of all dummy surfaces are listed in [Table 1](#). Note that positioning of these dummy surfaces can be performed in any ray-tracer after recording the global ray-tracing data and intersections of the reference ray.

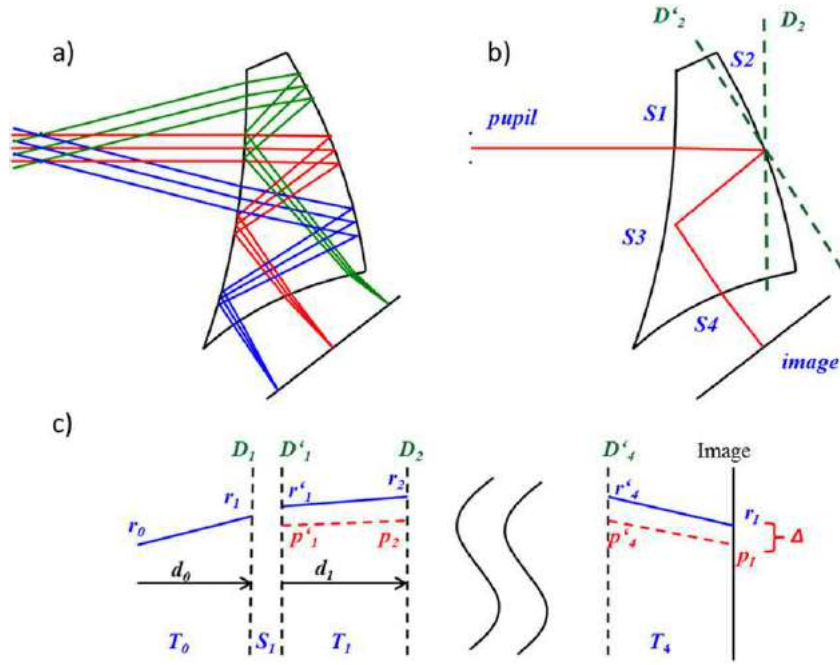


Fig. 6 Illustration of a patented freeform prism (a) and corresponding location of auxiliary reference planes (b), leading to an unfolded auxiliary system of a sequence of dummy surfaces (c).

What should be mentioned is that the tilt and decenter of the surfaces in [Table 1](#) are all in global coordinates and referred to the entrance pupil (surface 0). The horizontal and the vertical field angle of this lens as we analyze here is $\pm 15^\circ$ in each dimension. Besides, the pupil diameter is 4 mm. The surfaces 1, 2 and 3 are anamorphic freeform surfaces, where K_x and K_y represent the conic constant, AR and BR describe the 4th and 6th order symmetric coefficients, AP and BP illustrate the 4th and 6th order asymmetric coefficients, Y and Z correspond to the decenter of the surface in Y and Z direction, and α shows the surface tilt.

After the above design preparations we can perform raytracing and record the real raytracing data r_i on each “input”-dummy surface D_i , and r'_i on the corresponding “output”-dummy surfaces D'_i with the help of any ray-tracer. Just in the same way as we did in the rotational symmetric case, we can thus define an unfolded auxiliary system of sequential dummy surfaces to record the ray patterns in phase space, as illustrated in [Fig. 6\(c\)](#). In order to compare the real rays to the paraxial rays we additionally have to determine the linear transformation matrices S_i and T_i . For the propagation matrices T_i all that is required are the distances d_i between consecutive parallel dummy surfaces D'_i and D_{i+1} . These data are available from any ray-tracer. Determination of the linear transformation S_i of the tilted and decentered real surfaces is more challenging and several methods can be implemented: The real ray grid data at the dummy surfaces D_i and D'_i can be used to determine the best fit linear transformation (S_i) from a sufficiently large set of rays and the corresponding set of equations ([Chen and Herkommer, 2016](#)). Alternatively we can use the predefined ABCD-function for tilted and decentered systems in [Synopsis CodeV10.8](#) to calculate the paraxial transformation matrix S_i . Both methods are appropriate to visualize aberrations fields, since the average linear transformation can be subtracted, however the later method will be used later in this paper for exact surface aberration calculations. In [Fig. 7](#) we illustrate the phase space behavior of the freeform prism. The color scale corresponds to difference between the real ray and the corresponding best fit linear ray and represents the length of the aberration vector $|\Delta|$, as defined in [Eq. \(6\)](#). [Fig. 7](#) thus illustrates the aberration field, which is built up as the rays propagate through the freeform system. The figure nicely illustrates the nodes of the aberration field, i.e., the areas where the aberration vector is small, moreover the missing symmetry leads to non-symmetric distributions.

Extension to 4d Phase Space

To explain the principles and allow visualization the above description was limited to the 2d phase space, since we only considered rays in the xz -plane. However this allows describing meridional rays within an optical system only. Already for rotational symmetric systems with finite field we are missing the saggital ray fans. Moreover in freeform systems rotational symmetry cannot be assumed. Therefore for the general case we have to consider a 4d phase space, which means that the position and direction of each ray in a certain reference plane has to be described by four quantities, namely $\mathbf{r} = (x, y, u, v)^T$. Here x, y are the

Table 1 Lens data of the freeform prism of reference including additional dummy surfaces

Surface	Surface shape/radius (mm)	Surface position	Inclination angle(α)	Glass(n,v)
0	Infinity (pupil)	Origin		
D1	Infinity	Z 30.002		
1	$R_y - 108.187$ $R_x - 73.105$ $K_y 0$ $K_x 0$	AR 5.542E-7 BR 8.176E-11 AP - 0.080 BP - 1.379	- 14.700°	1.492, 57.471
D'1	Infinity	Z 30.002	- 1.066°	1.492, 57.471
D2	Infinity	Y - 0.251 Z 43.485	- 1.066°	1.492, 57.471
2	$R_y - 69.871$ $R_x - 60.374$ $K_y - 0.1368$ $K_x - 0.123$	AR-7.233E-11 BR-4.529E-12 AP29.075 BP - 2.085	36.660°	1.492, 57.471
D'2	Infinity	Y - 0.251 Z 43.485	38.376°	1.492, 57.471
D3	Infinity	Y - 11.858 Z 28.827	38.376°	1.492, 57.471
3	$R_y - 108.187$ $R_x - 73.105$ $K_y 0$ $K_x 0$	AR5.542E-7 BR8.176E-11 AP - 0.080 BP - 1.379	- 14.700°	1.492, 57.471
D'3	Infinity	Y - 11.858 Y - 11.858	- 55.019°	1.492, 57.471
D4	Infinity	Y - 23.067 Z 36.667	- 55.019°	1.492, 57.471
4	77.772	Y - 35.215 Z 18.817	- 47.770°	
D'4	Infinity	Y - 23.067 Z 36.667	- 50.668°	
5 (Image)	Infinity	Y - 30.892 Z 43.083	- 50.668°	

Source: Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. Optics Express 24 (23), 26999–27008.

intersection points of the ray with the reference surfaces and u, v are the index-weighted direction tangents, which can easily be computed from the normalized direction vector (L, M, N) and the refractive index n by using

$$u = n \cdot \frac{L}{N}, \quad v = n \cdot \frac{M}{N} \quad (8)$$

In consequence the above employed matrix calculus needs to be extended to 4d, which however is straightforward. If we for example consider free ray-propagation between two consecutive parallel reference planes D'_i and D_{i+1} , as illustrated in Fig. 8, the corresponding exact ray-transfer matrix is quite similar to the 2d case.

$$\mathbf{r}_{i+1} = \mathbf{T}_i \cdot \mathbf{r}'_i \text{ where } \mathbf{T}_i = \begin{pmatrix} \mathbf{I} & d \cdot \mathbf{I} \\ 0 & \mathbf{I} \end{pmatrix} \text{ and } \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (9)$$

Thus only the introduction of the two-dimensional unit matrix $\mathbf{I}$ is required to extend the matrix $\mathbf{T}$ to the general case. Here it is again important to note, that since we employ the exact ray direction tangents in Eq. (8) and not any “paraxial” angles the above propagation matrix is exact, linear and identical to the paraxial transformation matrix also for larger angles. In other words again free propagation does not introduce any deviation between paraxial and exact ray-tracing.

Thus aberrations are only induced by the action of the real surface in between each dummy surface pair D_i and D'_i . The linear/paraxial transformation $\mathbf{S}_i$ of any 4d-ray input ray $\mathbf{r}_i$ into a paraxial output ray $\mathbf{p}'_i$ can be described by a linear matrix, which however now is 4-dimensional:

$$\mathbf{p}'_i = \mathbf{S}_i \cdot \mathbf{r}_i \text{ where } \mathbf{S}_i = \begin{pmatrix} A_{xx} & A_{xy} & B_{xx} & B_{xy} \\ A_{yx} & A_{yy} & B_{yx} & B_{yy} \\ C_{xx} & C_{xy} & D_{xx} & D_{xy} \\ C_{yx} & C_{yy} & D_{yx} & D_{yy} \end{pmatrix} \quad (10)$$

For the case of rotational symmetric lenses the matrix simplifies to a similar form as Eq. (9) and can be calculated from the surface power, in the same way as discussed in Eq. (2). In the general case we can determine the matrix either from a set of rays and

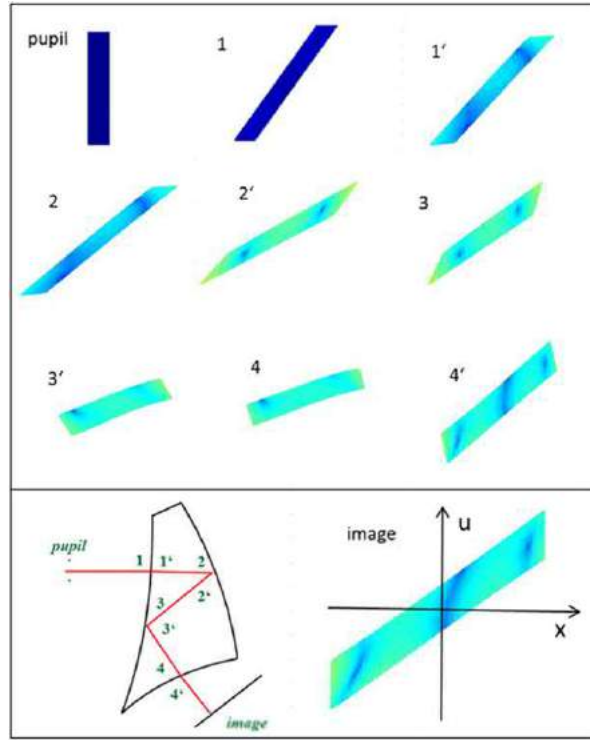


Fig. 7 Illustration of the phase space transformations corresponding to propagation through a freeform prism. The color code represents the normalized length of the aberration vector.

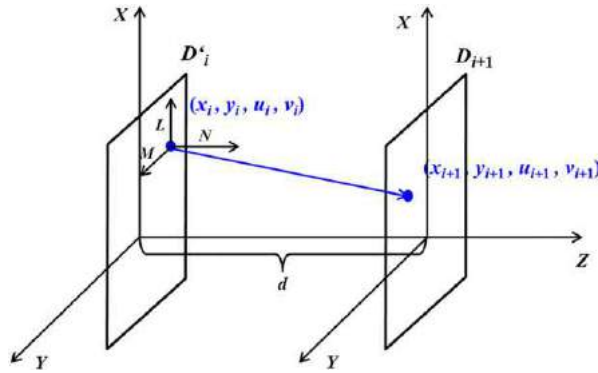


Fig. 8 Ray definition and propagation in between parallel reference planes. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. *Optics Express* 24 (23), 26999–27008.

then solving the equations to find the best-fit linear four-dimensional matrix. Another option is to employ the ABCD-function in Synopsis CodeV ([Synopsis CodeV10.8 News](#)), which is valid for decentered and tilted surfaces and defined for the general 4d-case. Application of this function in between D_i and D'_i will return the above 16-element linear transformation matrix. Note that in the general freeform case the sub-matrices A, B, C and D will not be symmetric, since especially a freeform surface will contain different curvatures in x and y direction and also prismatic structures.

If we, in comparison to the above linear analysis, now employ any ray-tracer to find the real surface intersection coordinates $\mathbf{r}_i$ at all of the reference surfaces we again can define, according to Eq. (6), an aberration vector Δ , which however is 4-dimensional. Therefore in summary the complete method as described in 2d is mathematically straightforward to be extended to the general case. However illustration of a 4d-vector and aberration field is challenging and in consequence visualization of the full information is difficult. Thus typically a reduction (projection) of the data will be required for visualization. For example the length of the aberration vector $|\Delta|$ can again be understood of as an aberration measure and illustrated as a color code, leading to similar pictures as shown in Fig. 7 in any 2d-visualization of the 4d-parameter field.

Even though visualization of the general situation is difficult, calculation is not. We will show below that the 4d-calculus is able to quantify aberrations and aberration contributions within freeform systems.

Calculation of Aberration Contributions at the Image

For rotational symmetric systems a long established theory of aberrations exists, and different classes of aberrations are defined, such as spherical, coma and astigmatism. The corresponding third order aberrations can be derived from paraxial raytracing only, and individual surface contributions can be calculated (Gross, 2005). Also an exact calculation of higher order surface contributions to lateral aberrations at the image exists: Within Aldis theory (Cox, 1964; Brewer, 1976) the aberration calculation is based on real ray tracing and paraxial (linear) ray-tracing of a given input ray. Application of this theorem allows exact calculation of the surface induced contribution to lateral aberrations at the image plane including all orders. Unfortunately for freeform systems this theorem can, due to the lack of symmetry, not be applied.

The phase space method described here does not require symmetry and is, same as Aldis theorem, based only on paraxial and exact rays. We will now mathematical prove that it can be applied to exactly quantify surface aberration contributions in freeform systems. To do this, the aberration generation and propagation in phase space has to be understood. Per definition the general aberration vector at a reference surface is a 4d-vector describing the deviation between the real ray coordinates and the paraxial ray coordinates. To better explain the relationships and propagation laws for aberrations let us switch back to the 2d-case for visualization. However the mathematical treatment is applicable to both, 2d and 4d cases.

In Fig. 9 the aberration generation and propagation is illustrated for the simple system of Fig. 2, where only two surfaces are involved. We will now mathematically and graphically follow the ray and aberration vectors through phase space.

During the analysis we will consider the aberration vector in front and after of each surface until we reach the image plane. We start with a general input ray r_0 at the object plane, as for example illustrated in Fig. 2. The free propagation to the first system surface D_1 is exact, via Eq. (7), and thus the real ray and the corresponding paraxial ray are equivalent:

$$r_1 = p_1 = T_0 \cdot r_0$$

Therefore the aberration vector in front of the real surface S_1 is zero, as

$$\Delta_1 = r_1 - p_1 = 0$$

As the ray passes the first surface (propagation from D_1 to D'_1) the ray picks up some aberration, since the input ray splits into a paraxial ray p'_1 (resulting from the linear surface matrix S_1) and the real ray r'_1 (refracted at the real surface), as illustrated in Fig. 9(a). Therefore we find the local aberration vector after surface S_1 to be:

$$\Delta'_1 = r'_1 - p'_1 = \underbrace{r'_1 - S_1 r_1}_{\delta_1}$$

Here and for later use we define the surface contribution δ_i to be the difference between the real traced ray r'_i and the paraxial traced ray $S_i r_i$ at surface i as:

$$\delta_i = r'_i - S_i r_i \quad (11)$$

After free propagation from surface S_1 to the front of surface S_2 we find the transformed, respectively sheared aberration vector, as illustrated in Fig. 9(b). In mathematical terms this corresponds to the application of the propagation matrix:

$$\Delta_2 = r_2 - p_2 = T_1 \delta_1$$

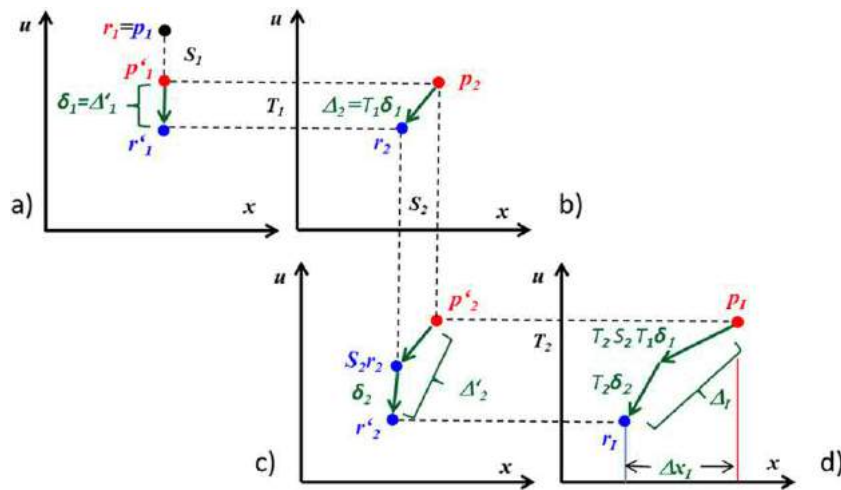


Fig. 9 Illustration of aberration propagation in phase space, where (a) to (b) represents the free propagation to surface S_2 , (b) to (c) illustrates the action of surface S_2 , and (c) to (d) resembles propagation to the image.

Similar as before the action of the subsequent surface S2 introduces some additional aberration contribution, which due to the above relations between the rays can be calculated as:

$$\Delta'_2 = \mathbf{r}'_2 - \mathbf{p}'_2 = \mathbf{r}'_2 - \mathbf{S}_2 \mathbf{p}_2 = \mathbf{r}'_2 - \mathbf{S}_2 \mathbf{T}_1 \mathbf{p}'_1 = \mathbf{r}'_2 - \mathbf{S}_2 \mathbf{T}_1 (\mathbf{r}'_1 - \delta_1) = \mathbf{S}_2 \mathbf{T}_1 \delta_1 + \underbrace{(\mathbf{r}'_2 - \mathbf{S}_2 \mathbf{r}'_1)}_{\delta_2}$$

The result can be interpreted in such a way that the surface S2 introduces an additional aberration term δ_2 which is added to the existing transformed (linear propagated) aberration vector of surface S1. In Fig. 9(c) we have again illustrated this graphically. Further propagation of the complete aberration vector to the front of surface S3 yields:

$$\Delta_3 = \mathbf{T}_2 \mathbf{S}_2 \mathbf{T}_1 \delta_1 + \mathbf{T}_2 \delta_2$$

In our graphical illustration of Fig. 9(d) this already corresponds to the image plane. At this point it has to be remembered that at the image plane the spatial component directly represents the lateral aberration of the particular ray. The equation above already suggests, that a recurrence relation can be established, since the aberration vector consists of terms corresponding to a sum of linear transformed aberrations δ_i from all proceeding surfaces. Indeed if we have n surfaces in the system we can continue this calculation to the image plane to find the following general result.

$$\Delta_I = \mathbf{r}_I - \mathbf{p}_I = \sum_{i=1}^n \mathbf{M}_{I,i} \delta_i = \sum_{i=1}^n \Delta_I^i \quad (12)$$

Here $\mathbf{M}_{I,i}$ represents the linear propagation matrix from the exiting dummy surface D'_i to the image plane I and is given by:

$$\mathbf{M}_{I,i} = \mathbf{T}_n \mathbf{S}_n \dots \mathbf{S}_{i+1} \mathbf{T}_i \quad (13)$$

Here the index n represents the last surface in the system, and $\mathbf{T}_n$ is the propagation from this last surface n to the image plane.

To summarize the above mathematical treatment: With Eq. (12) we have found a mathematical exact prove, that the total aberration of any ray at the image plane of an optical system consisting of n surfaces, can be expressed as the sum of individual surface contributions. Those individual surface contributions result from the differential rays at each surface i (i.e., the difference between linear and real ray tracing across surface i), which are linearly propagated to the image. In other words we have found a mathematical relation between surface aberration contributions and the exact ray aberrations at the image plane, including all

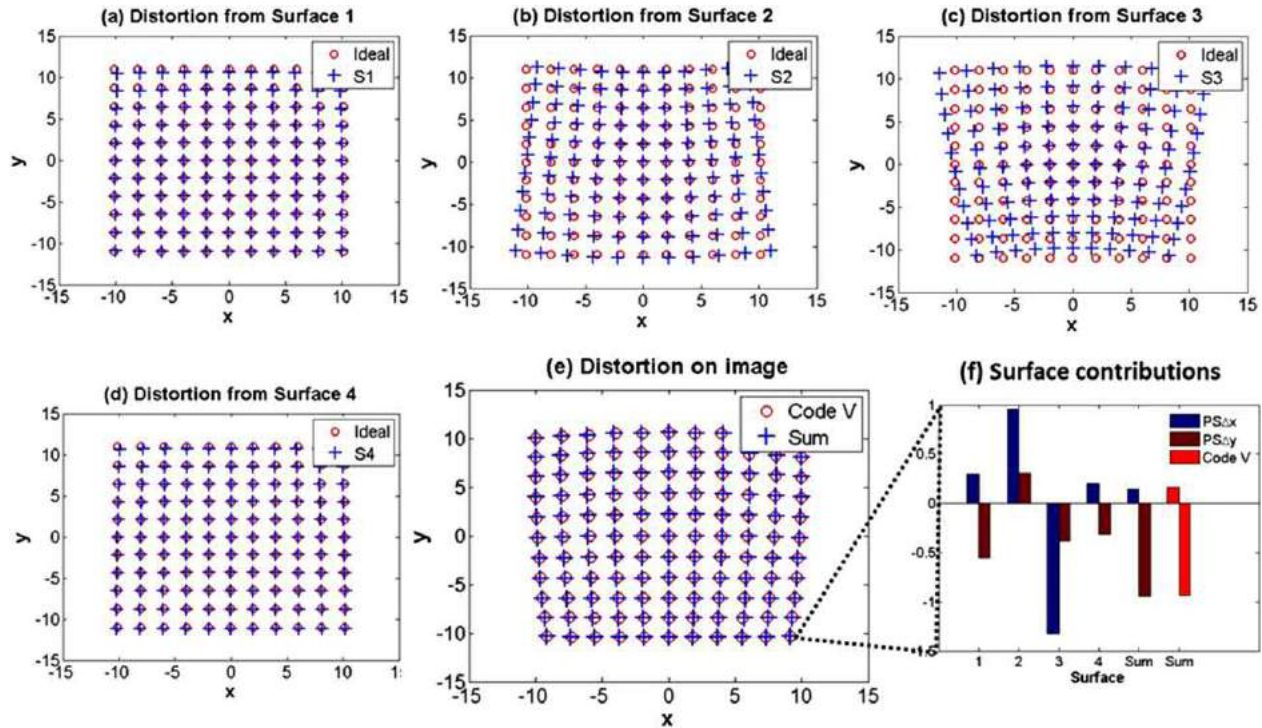


Fig. 10 Distortion analysis: In (a)–(d) the individual surface contributions are shown, where the red circles are the paraxial positions of the rays on image, and the blue cross represents the lateral distortion contribution of the surface, (e) illustrates the sum of the distortion contributions and is compared to the exact ray-tracing result at the image, (f) reveals the individual surface contributions for a single chief ray. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. *Optics Express* 24 (23), 26999–27008.

orders, quite similar to Aldis theory, however valid for freeform systems. The individual surface aberration contributions at the image plane, given in Eq. (12), are:

$$\Delta_I^i = M_{I,i} \delta_i = M_{I,i} (\mathbf{r}'_i - S_i \mathbf{r}_i) \quad (14)$$

This expression contains only real ray-tracing data before and after each surface and the linear System matrix $M_{I,i}$. Therefore this contribution vector can be obtained from the paraxial system representation and ray-tracing data only. No assumptions are made about symmetry or surface shape.

Aberration Contributions in a Freeform Prism

In order to demonstrate the application of the above mathematical treatment we apply it to the already introduced freeform prism. Following the above preconditioning, we have inserted dummy surfaces, as given in Table 1, and moreover determined all necessary transformation matrices S_i , T_i and $M_{I,i}$ by using CodeVs ABCD-linear system analysis (Synopsis CodeV10.8 News). Therefore calculation of the aberration contribution Δ_I^i of every ray at the image is now possible via Eq. (14), if in a prior real ray-tracing the ray-coordinates $\mathbf{r}_i$ and $\mathbf{r}'_i$ are recorded.

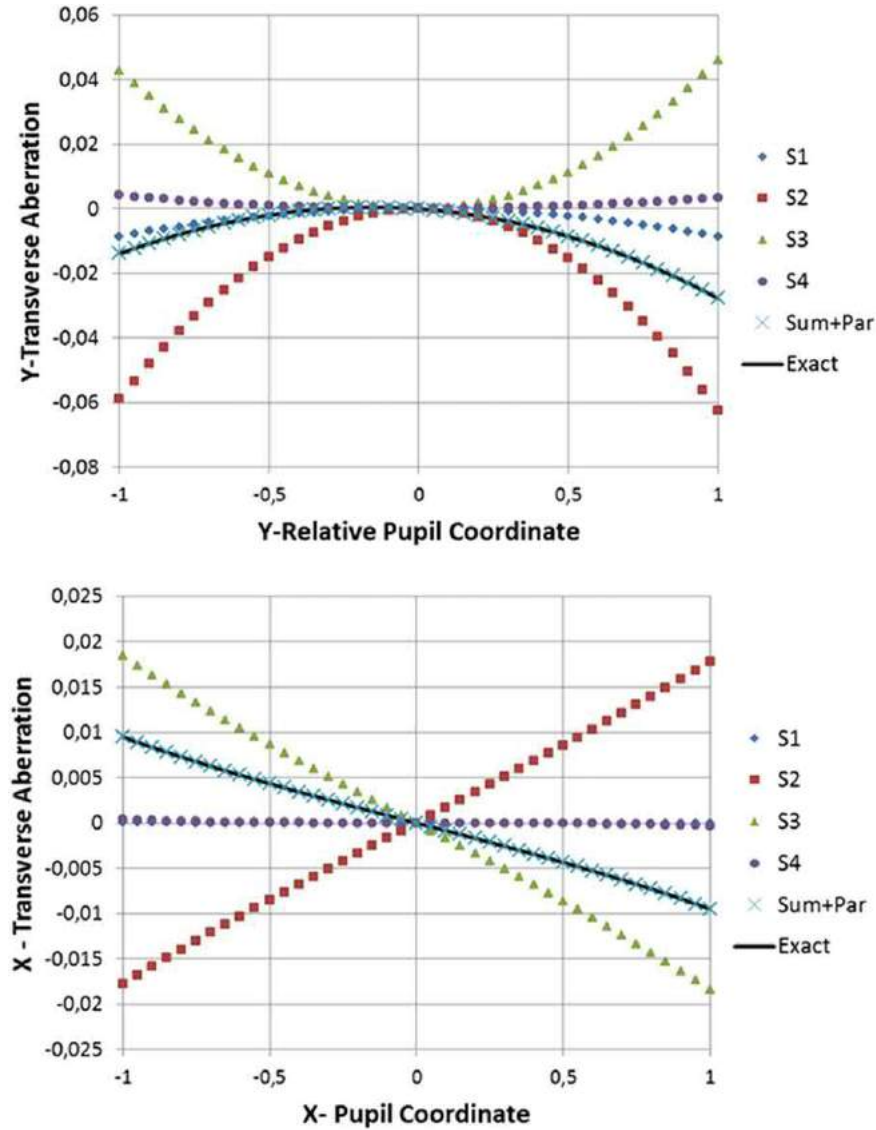


Fig. 11 Transverse aberrations for the (0°,0°) field, resolved for individual surface contributions. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. *Optics Express* 24 (23), 26999–27008.

As a first simple application the aberration contribution of a set of single rays can be studied. The chief-rays, representing different field angles, allow for a first useful application of the method. Since lateral aberrations of the chief-rays at the image plane correspond to the distortion of the system, we can simply apply our method to study distortion contributions in the freeform prism. To do so, we trace a set of chief-rays (i.e., rays starting at the center of the pupil) through the system. We use a set of 11×11 rays, scanning the field of view of $\pm 15^\circ$ in the x and y-direction. For every ray the ray-coordinates r_i and r'_i are recorded at all dummy surfaces. By applying Eq. (14) we can calculate the aberration contribution vector Δ_i^j on the image plane. The spatial (x,y) components of this aberration vector correspond to lateral distortion, i.e., the distortion, which is induced by each of the surfaces. This is illustrated in Fig. 10.

In Fig. 10(a)–(d), red circles represent the ideal positions of the rays on the image, calculated by the linear system matrixes, whereas the blue cross represents the ray aberration contributions of surfaces 1–4 on the image. Therefore the deviations of the blue cross from the red dots are the distortion contributions of each surface. The summation of all distortion contributions corresponds to the overall distortion at the image plane and are drawn as blue crosses in Fig. 10(e) and compared to the red circles directly obtained from real ray-tracing of the full system to the image in CodeV. As expected from Eq. (12) they exactly agree. The benefit of the phase space method is illustrated in Fig. 10(f), where the individual surface contributions are shown for a single ray of this grid. Obviously surface S2 and S3 are strongly contributing to distortion.

As a further example of the application of this method we can calculate and illustrate transverse ray-aberrations for a particular spot in the image plane. In Fig. 11 the meridional and saggital ray aberration contributions for each surface are illustrated for the central spot ($0^\circ, 0^\circ$) of the freeform system. The figure reveals that here surface S2 and surface S3 are strongly compensating each other. The sum of all surfaces, plus the paraxial contribution, perfectly coincides with the results achieved by exact tracing of these rays to the image surface, which again proves the validity of Eq. (12).

Conclusions

This paper presents a matrix based method to visualize aberrations via an analysis of phase space transformations. The difference between real and paraxial rays define a general measure which can be employed to illustrate the aberration generation and propagation in freeform systems.

Moreover mathematical exact aberration contributions in freeform systems could be derived. It was mathematical proven, that the exact ray aberrations at the image plane is identical to the sum of individual surface contributions. Those individual surface contributions can be calculated from a ray-tracing analysis of the system and paraxial propagation of the real and paraxial differential vector induced from an individual surface. All calculations can be performed with standard ray-tracing, provided additional dummy surfaces are installed in the system before. No assumptions about symmetry were required. In other words the method resembles a kind of Aldis theory for freeform systems, allowing the exact calculation of surface aberration contributions.

We have illustrated the method at the example of a freeform prism and shown that distortion and transverse ray-fans can be calculated and agree with the result of a standard ray-tracer.

This general work thus allows an “easy to implement” visualization and analysis of the aberration contributions within any freeform system. This will be beneficial in comparing freeform designs, performing tolerance analysis and improved optimization strategies.

References

- Brewer, S.H., 1976. Surface contribution algorithms for analysis and optimization. *Journal of the Optical Society of America* 66 (1), 8–13.
- Chen, B., Herkommer, A.M., 2016. High order surface aberration contributions from phase space analysis of differential rays. *Optics Express* 24 (6), 5934–5945.
- Cox, A., 1964. *A System of Optical Design*. Focal Press.
- Fuerschbach, K., Jannick, P.R., Kevin, P.T., 2014. Theory of aberration fields for general optical systems with freeform surfaces. *Optics Express* 22, 26585–26606.
- Gross, H. (Ed.), 2005. *Handbook of Optical Systems*, vols. 2 and 3. Wiley VCH.
- Herkommer, A.M., 2013. Phase space optics: An alternate approach to freeform optical systems. *Optical Engineering* 53 (3), 031304.
- Kloos, G., 2007. *Matrix Methods for Optical Layout*, vol. 77. SPIE Press.
- Synopsis CodeV10.8 News. Available at: <https://optics.synopsys.com/codev/codev-whatsnew.html>
- Takahashi, K., 1997. Head or face mounted image display apparatus. U.S. Patent No. 5,701,202.
- Testorf, M., Hennelly, B., Ojeda-Castaneda, J., 2010. *Phase-Space Optics*. McGraw-Hill.
- Thompson, K.P., Rolland, J.P., 2012. Freeform optical surfaces: A revolution in imaging optical design. *Optics and Photonics News* 23 (6), 30–35.
- Torre, A., 2005. *Linear Ray and Wave Optics in Phase Space*. Amsterdam: Elsevier.

Silicon Photonics; Ring Modulator Transmitters

M. Ashkan Seyed and Marco Fiorentino, Hewlett Packard Labs, Palo Alto, CA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

A silicon-based photonics integrated circuit (PIC) is a device that integrates multiple photonics components in a single planar structure. Silicon PICs lag behind compound semiconductor PICs in maturity. However the prospect of building large numbers of complex PICs for data communication applications at low cost by leveraging the CMOS industrial base has attracted much attention in academia and industry. One of the key components of a data communication PIC is a modulator. An integrated optical modulator is used to vary the properties of a light beam propagating in a PIC. Most commonly, modulators change the beam intensity but phase modulators are also used. An electro-optical modulator is used to translate an electrical signal into an optical signal that can be then transmitted and received through a photonic circuit. A ring optical modulator consist of a ring-shaped optical cavity that is optically coupled with a bus waveguide. Schematic top views of ring resonators are shown in Fig. 1. In a silicon PIC, the waveguides have a rectangular cross section with dimensions ranging between hundreds of nanometers to a few microns and the ring diameters for modulators are of the order of ten microns, whereas larger rings with diameters up to hundreds of microns can be used as filters.

Fig. 1(a) shows the waveguide arrangement for an “all pass” resonator while Fig. 1(b) shows the arrangement for an “add-drop filter” resonator. In the case of the “all pass” arrangement, all light incident in the device is passed through except for the resonant wavelengths that are absorbed or scattered in the resonator. In the “add-drop filter” configuration, resonant light is dropped through the “drop” port while non resonant light is transmitted to the “pass” port. The transfer function of both configurations from the input to the pass port is characterized by a Lorentzian notch centered at the cavity resonant frequency. Changing the effective index of the ring cavity shifts the resonant frequency and this mechanism can be used to modulate the amplitude of a laser in a wavelength-selective way. This is the main attraction of ring modulators: rings of different resonant wavelengths can be cascaded on a bus waveguide to modulate a number of different laser wavelengths. Similarly resonant “add-drop” filters can be used to de-multiplex a multi-wavelength signal at the receiver. In a silicon-based modulator, the change in the effective index can be achieved in a number of ways. Most commonly one uses plasma dispersion effects that link changes in the carrier concentration to changes in the real and imaginary parts of the index of refraction of silicon. Changes in the refractive index Δn as well as in the absorption coefficient $\Delta \alpha$ at a wavelength of 1.55 μm are given by (Soref and Bennet, 1987).

$$\Delta n = - \left[8.8 \times 10^{-22} \times \Delta n_e + 8.5 \times 10^{-18} \times (\Delta n_h)^{0.8} \right] \quad (1)$$

$$\Delta \alpha = 8.5 \times 10^{-18} \times \Delta n_e + 6.0 \times 10^{-18} \times \Delta n_h \quad (2)$$

where Δn_e and Δn_h indicate changes in the free-electron and free-hole carrier concentrations, respectively. The number of free electrons and free holes in the silicon can be changed using charge depletion, charge injection, or charge accumulation.

Various modulator schemes are shown in Fig. 2. A horizontal depletion modulator is shown in Fig. 2(a). The modulator consists of a light waveguide with a PN junction in the middle. By reverse-biasing the junction, one can expand the depletion region and therefore change the charge concentration. This modulation scheme is very fast and energy efficient but, because of the intrinsic doping of the light-carrying waveguide, it adds considerable optical losses to the device. It should also be noted that the maximum amount of index change is determined by the initial doping of the waveguide. A vertical depletion modulator (Fig. 2(b)) uses a similar modulation mechanism but the PN junction runs parallel to the plane of the circuit. This reduces the requirements for overlay control between the various doping steps. A current injection modulator consists of a PIN junction where the intrinsic region is the light carrying waveguide (Fig. 2(c)). This modulation scheme is relatively slow as it is limited by the majority carrier lifetime and has a relatively high power consumption due to the need to continuously

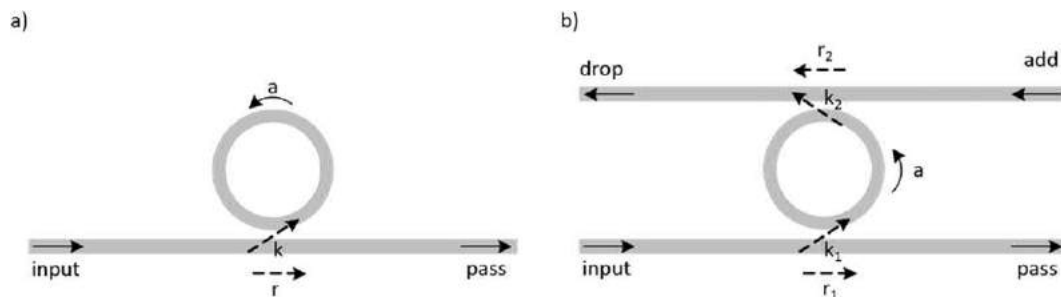


Fig. 1 Schematic top view of ring resonators in two configurations: (a) “all Pass” and (b) “add-drop filter”.

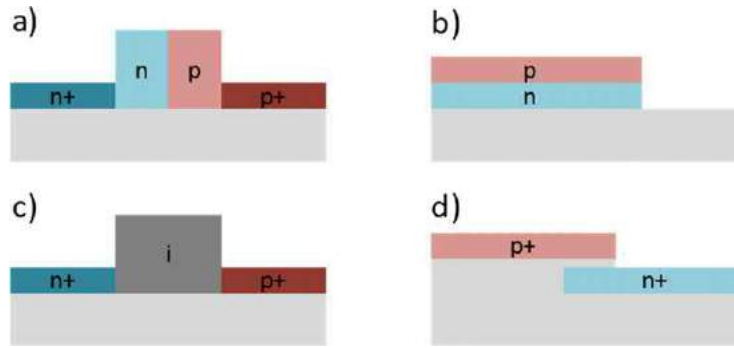


Fig. 2 Waveguide cross sections for various types of modulators: (a) horizontal depletion, (b) vertical depletion, (c) injection, and (d) accumulation.

replace carriers in the waveguide by injecting current in the junction. Balancing this disadvantages is the fact that large changes of the index are possible and the fact that the modulator can be built with low-loss intrinsic silicon. An accumulation modulator is shown in Fig. 2(d). The modulator consist of a waveguide with a MOS capacitor in the middle (running parallel to the plane of the PIC in this case). The capacitor can be charged and discharged through the p and n electrodes and the charge accumulating around the oxide barrier creates the index change. This scheme has the advantage of being fast and can in principle be built with relatively low intrinsic losses. It should be noted, however, that while it is possible to build a ring or a disk modulator based on this modulation mechanism, none has been demonstrated so far. There are other possible modulation schemes. One can change the index of the waveguide by changing the temperature of the silicon. This method is slow and it is not used for data encoding but given that it allows a large tuning range, it is usually used to tune ring resonance wavelengths to a desired value. Other methods use an electro-optical effect in the material that clads the silicon waveguide, for example there have been several demonstration of polymer-based ring modulators.

Fabrication

Photonic devices and circuits require an optical isolation layer and thus are based on a silicon-on-insulator (SOI) wafer. In order to create such a wafer, a layer of silicon with a thickness of a few hundred nanometers is placed atop a layer of silicon dioxide that is a few microns thick. Below this buried oxide is a bulk silicon substrate. The index contrast provided by the oxide layer below and air or a deposited oxide above enables optical confinement in the upper, thin, silicon layer in the vertical direction (Bruel, 1995).

The fabrication of ring-based modulators uses basic CMOS foundry process steps such as implantation of ions to form junctions, vias for ohmic contact, and a multi-layer back end of line. However, to enable lateral optical confinement, waveguides are formed by a partial etch of the silicon layer using plasma gas dry etch processes, a key step that enabled adoption of photonics into CMOS processes. In the most advanced foundries, 193 nm immersion lithography is used to define the optical structures in a resist, but lower resolution lithography has also been used successfully. For ring-based fabrication the critical dimension is the gap between the ring and the bus waveguide that can be as small as 100nm and requires tight control to avoid large dispersion in the optical properties. The lithographic pattern is then transferred into an oxide-based hard mask and finally into the SOI layer using a plasma etch. The optical propagation loss is a critical metric of performance for PICs and is highly dependent on the quality of the lithography and ensuing etch steps to ensure smooth, straight sidewalls. The depth of the partial etch (or alternatively, the height of the waveguide side wall) determines this loss value as it exposes more roughness to the optical signal. Therefore, this depth and its uniformity across a wafer is a tightly-controlled process parameter.

Often, there are two partial etches of the SOI wafer: one to define optical grating structures used for input/output of the optical signal onto the chip and the second for waveguides, often referred to as a shallow waveguide etch. The advantage of this shallow etch, which is usually 50% of the top silicon layer of the SOI wafer, is that it has low propagation loss on the order a few dB per centimeter for a straight waveguide. However, as mentioned, rings are formed by tight bends of the waveguide where the performance of shallow etched waveguides is very poor. This is primarily due to low optical confinement that causes the optical field to leak out of the waveguide through radiative modes. Bend losses can be alleviated by a deeper partial etch, resulting in a higher waveguide sidewall. Of course, this enhanced optical confinement comes at the expense of increased optical propagation loss. Therefore, these competing mechanisms present an optimization problem between process controls, etch depth and ring resonator bend radius in order to achieve optimized optical performance. A cross section of a partially etched waveguide is shown in Fig. 3 where important dimensions are denoted. A further complication of ring modulators that use a lateral junction is the resistance of the remaining portion of the un-etched SOI layer, referred to as the rib. A deeper etch results in better confinement which enables a tighter bend radius. However, increased optical loss and higher resistance of the rib layer present a design trade-off that must be properly accounted for to optimize device performance.

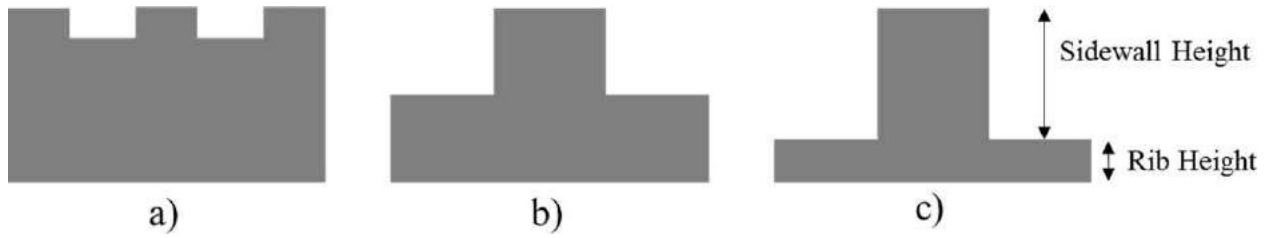


Fig. 3 Cross section of various etch depths for (a) TE-polarized gratings, (b) shallow rib waveguide, and (c) deep rib waveguide.

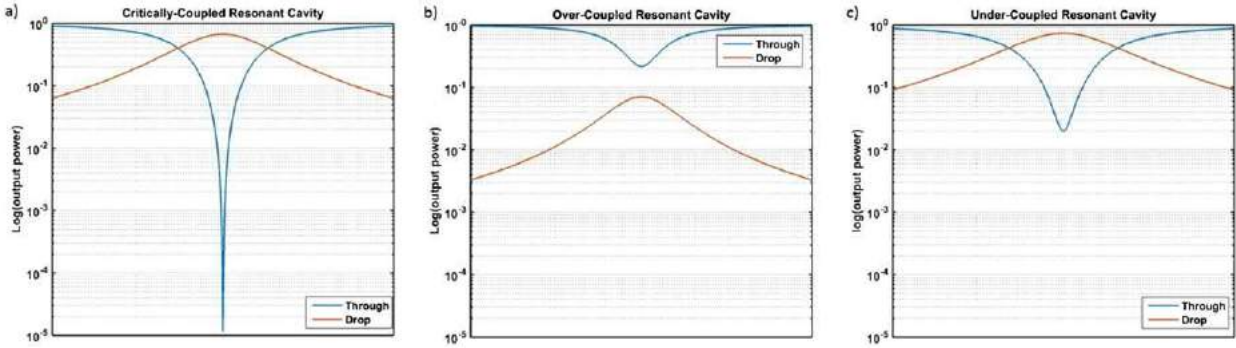


Fig. 4 Through and Drop port spectra for ring resonator in (a) critically coupled, (b) over-coupled, and (c) under coupled conditions.

Optoelectronic Properties

As mentioned before, the transfer function of a ring resonator structure is a Lorentzian, characterized by its extinction ratio (ER) and full-width half-maximum (FWHM), a parameter which is often used to calculate the cavity quality factor (Q) by an inverse relation. A resonant cavity can be under, critically, or over-coupled which is easily determined by the trends of the ER and cavity Q. Following the treatment in [Bogaerts et al., \(2012\)](#), the transmitted optical spectrum for an all-pass configuration, as shown in [Fig. 1\(a\)](#) is given by the following equation:

$$T_n = \frac{I_{pass}}{I_{input}} = \frac{a^2 - 2ra \cos\phi + r^2}{1 - 2ar \cos\phi + (ra)^2} \quad (3)$$

where $a^2 = \exp(-\alpha L)$ is the internal loss mechanisms and comprises radiative losses due to surface roughness, bend loss, and free-carrier absorption from junction doping. These factors are largely process-dependent and can be considered static for a ring resonator design. However, careful consideration must be given to the coupler region design to ensure a proper balance between these three factors to achieve the desired coupling state. The coefficient r quantifies the coupling between the ring and the bus waveguide. Similarly, the transmitted optical power for the through and drop port for a ring resonator in the add/drop configuration shown in [Fig. 1\(b\)](#) is given by

$$T_p = \frac{I_{pass}}{I_{input}} = \frac{r_2^2 a^2 - 2r_1 r_2 a \cos\phi + r_1^2}{1 - 2r_1 r_2 a \cos\phi + (r_1 r_2 a)^2} \quad (4)$$

$$T_d = \frac{I_{drop}}{I_{input}} = \frac{(1 - r_1^2)(1 - r_2^2)a}{1 - 2r_1 r_2 a \cos\phi + (r_1 r_2 a)^2} \quad (5)$$

where r_1 and r_2 are the coupling coefficients between the ring and the add and drop waveguide, respectively. The state of coupling to the ring resonator is a relation between the input coupling, internal loss, and output coupling. Referring to the add/drop configuration in [Fig. 1\(b\)](#), the relationship

$$r_1 = r_2 a \quad (6)$$

defines the critical coupling condition. When $r_1 > r_2 a$ the cavity is under-coupled whereas when $r_1 < r_2 a$ the cavity is over-coupled.

In the critically-coupled regime the amount of light coupled to the cavity is equal to the amount coupled out and radiated due to losses. In this condition the extinction ratio (ER) and quality factor (Q), as can be seen from the spectra of the through and drop port plotted in [Fig. 4\(a\)](#), are very high. For SOI-based ring resonator cavities, a cavity Q in excess of 15,000 is routinely possible with greater than 15 dB ER. As the resonator cavity is moved into the under-coupled regime, the ER drops drastically as Q increases still, as can be seen from the plotted spectra in [Fig. 4\(b\)](#). An over-coupled resonator has both low ER and low Q. the spectra of which are plotted in [Fig. 4\(c\)](#). Therefore, ring resonator amplitude modulators that are critically coupled are desired as this

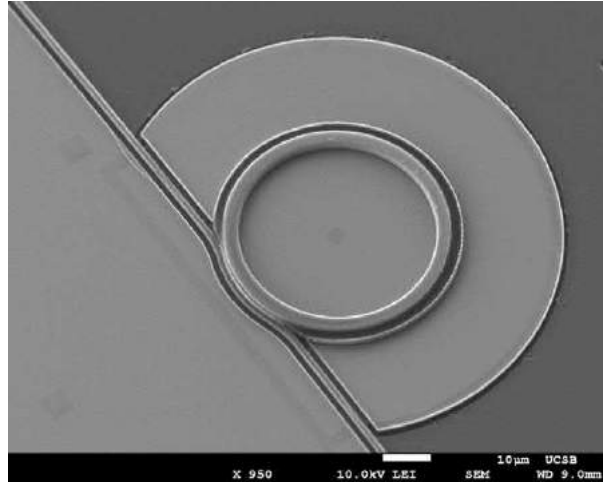


Fig. 5 Curved bus/ring resonator cavity coupler region.

maximizes the ER of the optical signal while keeping the cavity Q within a reasonable range. The cavity Q presents yet another design optimization as higher Q allows for denser channel spacing due to higher spectral isolation. However, higher Q means long photon lifetime within the optical cavity, increasing inter-symbol interference (ISI) as the modulation data rate increases.

There are many design types for coupler regions. In one example two parallel waveguides of a desired length are used to transfer power between the bus waveguide and ring cavity. This design allows for an analytical solutions for coupling strength using coupled mode theory (Marcuse, 2013). However, as is often desired for critical coupling, low input/output coupling ($\sim 5\%$ or less) are difficult to achieve due to the required short interaction length. Another design type of the coupler region is a co-curved bus/ring interaction region, shown in Fig. 5 where the bus waveguide follows the curve of the ring cavity with a constant gap separation.

This design offers a great degree of flexibility and allows for low coupling percentages between the bus and ring waveguides and is often preferred for processes where the minimum gap possible between two waveguides is large (on the order of 250 nm or greater). This design allows for shorter effective interaction lengths between the two waveguides as compared to the parallel waveguide design, it allows the resonant cavity to maintain a fully circular shape (which is often required for maintaining a constant loss and mode profile inside the cavity). However, it is more prone to fabrication variation in the gap between the ring/bus waveguide. The third design technique for the coupler region uses a straight bus waveguide and circular ring cavity as shown in Fig. 1, but with a narrower waveguide width for the coupler (W_{cp}) versus the ring cavity. In using a narrower waveguide width for the bus, the effective indices for the fundamental states of the two waveguides are increased, thereby enhancing the coupling between the two fundamental waveguide modes for a fixed effective interaction length. Another design factor for the add/drop configuration of the ring resonator cavity are the gaps between the ring cavity and the through and drop gap, referred to as G_{thru} and G_{drop} , respectively. For a fixed W_{cp} and ring width, one can control precisely the coupling percentage between the two waveguides by narrow variations in the gap between the two waveguides. This allows for accurate control of the input/output coupling to achieve critical coupling, as defined above.

A ring resonant cavity can be used as an amplitude modulator when its wavelength of resonance is modulated by electro-optic interactions. However, it is difficult, if not impossible, to control the absolute value of the resonance wavelength by fabrication techniques alone. The resonance wavelength can be controlled by changing the effective index of the material by using either the electro-optic effect or localized heating using integrated heaters. As mentioned, injecting carriers into the ring cavity causes a blue shift in the resonance wavelength and this is the main scheme for modulation. However, by maintaining a small DC bias on the diode, the resonance wavelength can be controlled with the same diode that is performing the modulation. This scheme is challenging to implement along with high-speed modulation in the driver design. By using a localized heater, the effective index of silicon can be finely tuned causing a red shift in the resonance wavelength.

The two main designs for an integrated heater are using the lowest metal layer directly above the ring cavity or with a resistive heater in the silicon rib material. To maximize the thermal profile overlap, the metal layer traces over the ring rib waveguide for most of the circumference of the ring cavity. This design is more robust to fabrication errors and burn-out from high current. However, due to the low resistivity of the typical metal layers available in CMOS-compatible processes, this design requires higher voltages as compared to other designs. Furthermore, due to CMOS back-end-of-line (BEOL) processes, the metal layer is several hundred microns away, further decreasing its thermal impact on the underlying silicon waveguide. Decreasing this length is not directly an advantage as interaction of the optical field with metal can lead to very high losses. These challenges can be somewhat mitigated by using highly resistive compounds like TiN instead of copper or aluminum, however, this material is not compatible with CMOS processes.

An integrated resistive heater in the silicon rib is created by forming a conductive channel in the otherwise intrinsic silicon material. Ohmic contacts are formed to this region by forming regions of high doping and conventional BEOL steps. By using an integrated resistor in the silicon rib, the local temperature within a few microns of the ring rib can be controlled, which in turn controls the material effective index, allowing for accurate resonance wavelength control. This design offers very efficient wavelength control and consumes power on the order of tens of microwatts per gigahertz in the wavelength range of interest. Careful consideration must be given to the resistor design in doping type and concentration of the current channel, its length and width to avoid leakage/fringing fields, and to optimize the resistance to achieve efficient wavelength control. A drawback of this design is the minimum proximity of the current channel to the waveguide due to foundry design rules. By using local temperature change, often near the coupling regions of the cavity and bus waveguides, a relatively small footprint of a few microns squared is required to efficiently control the wavelength of resonance, furthering the advantage of this heater design.

As mentioned previously, injection-type ring modulators use a P-i-N junction inside the waveguide to modulate the optical signal. The intrinsic bandwidth of this diode, however, is typically ~ 1 GHz. This is due to long lifetime of carriers in the intrinsic region. To overcome this bandwidth limit, equalization, or pre-emphasis, of the driving signal is required. The general principle of this approach is to decrease the amplitude of the lower frequency content of the driving signal, thereby boosting the amplitude of the high frequency content of the driving signal, relatively. The convolution of this driving signal frequency content and the intrinsic electrical frequency response of the diode results in increased bandwidth. This technique has achieved open eye diagrams up to 25 Gb/s using carrier-injection ring modulators with an intrinsic bandwidth of 1 GHz (Wu, *et al.*, 2016). As mentioned previously, there exists also an intrinsic bandwidth of the modulator from the photon lifetime due to cavity Q. So while pre-emphasis may boost the electrical bandwidth, the photon lifetime is a consequence of optical cavity design and is not affected by a pre-emphasized electrical signal.

In order to first understand the intrinsic electrical bandwidth of a ring modulator, an accurate circuit model is needed. The BEOL metal contribute small but known parasitic effects in the form of resistance and capacitance due to metal vias, substrate and metal layers, and are independent from the type of ring modulator, namely carrier depletion or injection. To develop a large signal model, the P-i-N junction may be modeled as a current source and a capacitor in parallel which are placed in series with a resistor, as shown in Fig. 6(a) from Wu *et al.* (2016). Similarly, a carrier-depletion ring modulator can be analyzed by the effective circuit shown in Fig. 6(b) from Rhim *et al.* (2015).

By performing S_{11} electrical measurements, the parasitics from these circuits can be derived through fitting. Subsequent DC current-voltage tests will further define the diode's current behavior. Note that the optical bandwidth due to the photon lifetime effect is also captured in the carrier-injection model. It is critical to note that the capacitor in this model, represented by C_{diff} is not a representation of the junction capacitance but that it is a mathematical consequence of the diode's small-signal behavior. Namely, it captures the transit time of the carriers across the junction and represents the change in current as a function of a bias change in time, which is the definition of capacitance. Therefore, C_{diff} must be carefully calculated using physical parameters and fitting models to accurately predict the time-dependent behavior of the modulator. Once the frequency response of the diode current is known, it can be mapped directly to a change in material index and thus modulation of the optical signal. This allows for an opto-electrical co-simulation to understand the high-speed modulation response of the modulator and allows for optimization of the pre-emphasized driving scheme often required to overcome the bandwidth limitations of a carrier-injection modulator.

The pre-emphasized signal can be generated by a few approaches and two will be briefly discussed here. By using a differential PRBS pattern, one can include a tunable delay and DC bias on the complementary output of the differential pair and combine the two outputs. This forms then the over/under-shoot components of the driving signal. The amount of delay relative to the data rate UI determines the frequency boost seen by the higher frequency components of the driving signal. Analogously, using a multi-tap FIR filter can produce a similar scheme. Often 2 or 3 taps are sufficient to produce the correct frequency response necessary for this driving mechanism. By using external hardware, one can accurately understand the required frequency response, making the design of a CMOS driver ASIC more direct. Once this frequency response is known, a tunable FIR filter integrated into the CMOS driver design can be easily implemented. Both approaches have been demonstrated (Li *et al.*, 2014) successfully to drive carrier-injection ring modulators.

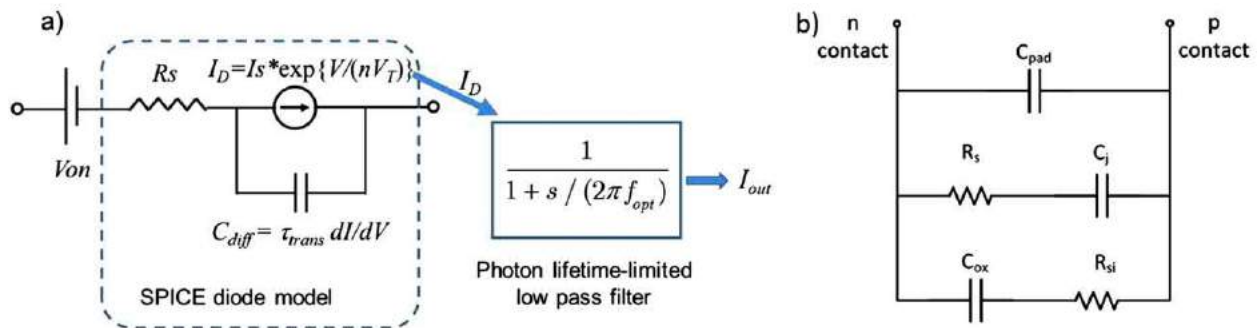


Fig. 6 Equivalent circuit for a (a) carrier-injection and (b) carrier-depletion ring modulator.

Nonlinear Effects

Light that is tightly confined in a silicon photonic waveguide will experience optical nonlinearities. Whereas silicon is a centrosymmetric material and therefore does not exhibit any nonlinear effect of the lowest order ($\chi^{(2)}$) third-order ($\chi^{(3)}$) effects are going to be present. The resonant nature of the ring resonators further reduces the threshold power necessary to observe nonlinear effects. Some researchers have strived to demonstrate and exploit these effects for a range of applications including Raman amplification and lasing, self-phase modulation, all-optical processing, and even photon-pair generation. In the context of modulation, however, nonlinear effects are detrimental. In particular the most commonly observed nonlinear effect in ring resonators derives from thermal effects. Light absorption in the waveguide (either from two-photon absorption or from absorption by silicon defects) creates carriers that thermalize thus increasing the waveguide temperature. Because the index of refraction depends on temperature this effect will change the resonator resonant wavelength and therefore its transfer functions. The effect of this nonlinear behavior is shown in Fig. 7. As the power increases, the transfer function of the drop port of an “add-drop filter” changes (Fig. 7(a)). The system also shows a bi-stable behavior as evidenced by the difference in the transfer function measured in the same resonator by scanning the wavelength in the two directions. Initial modeling of this effects for high-speed depletion modulators has been published (Shin *et al.*, 2016).

Applications: Ring-Based DWDM Transceivers

The most important application of silicon microring modulators is in dense wavelength multiplexing (DWDM) transceivers. Because of their wavelength-selective nature microrings naturally lend themselves to DWDM applications. A schematic view of a DWDM transceiver is shown in Fig. 8. In this example a multi-wavelength laser source is injected in the transmitter circuit comprising a number of modulator rings each resonant with one wavelength of the incoming source. The modulators are used to encode a data stream on the individual laser lines. At the receiver side add-drop resonators are used de-multiplex the incoming data streams and direct them on individual detectors. Waveguides and possibly optical fibers connect the transmitter and the receiver. Transceivers like the one described here have been studied but no commercial implementation of this scheme is available at the time of writing. A key issue to warrant widespread deployment of ring-based DWDM transceivers is the implementation of

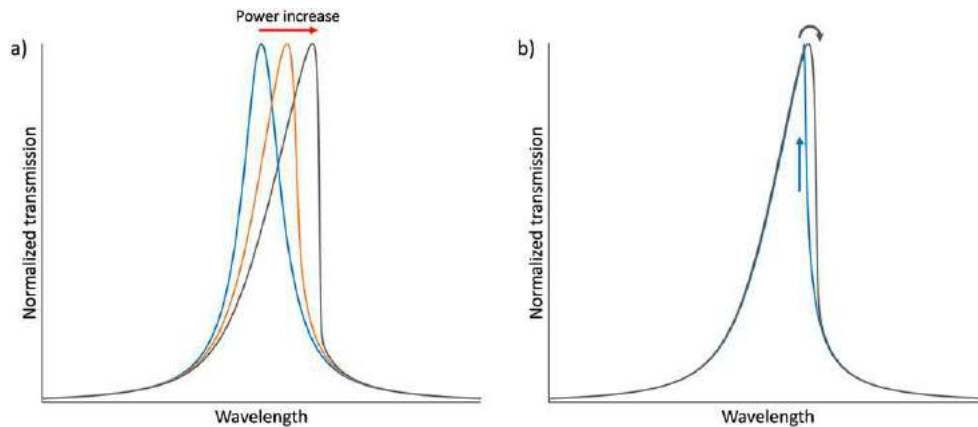


Fig. 7 Nonlinear bistability in ring resonators. (a) Change in the normalized transfer function of a drop port as a function of optical power. (b) Bistability, the transfer function changes depending on the direction of the wavelength sweep.

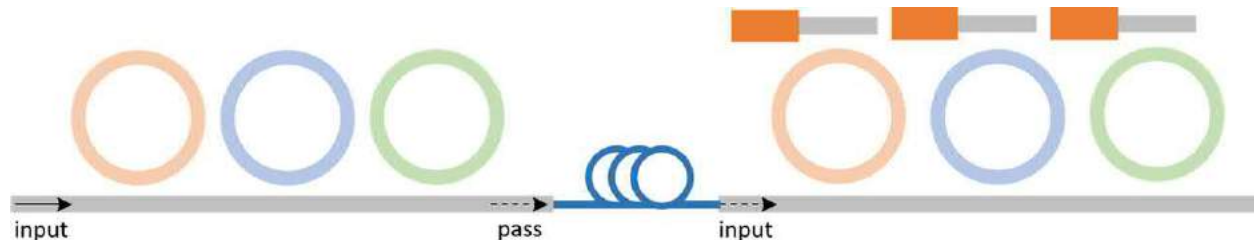


Fig. 8 Dense wavelength division multiplexing link. A multi-frequency source is injected in the transmitter and the various components are separately modulated. After transmission involving on-chip waveguides and off-chip fibers the modulated signals are coupled into a receiver chip where the various components are demultiplexed using drop filters and detected with on-chip photodiodes.

appropriate electronic circuits that drive the modulators, amplify and discriminate the received signal, and tune the rings to be in resonance with the appropriate laser line. Much work has gone into the design of special purpose circuits to achieve these goals. Most of these circuits have been designed in standard CMOS processes to take advantage of the large industrial base that supports these processes and given the relatively low analog bandwidth (10–20 GHz) required for these circuits.

Transmitter Circuits

Transmitter circuits differ considerably depending on the nature of the ring modulator. Depletion rings behave as capacitive loads and have bandwidths in excess of 20 GHz. For this reason relatively simple drivers comprising a chain of inverters and a final driver can be used. With such drivers bit rates in excess of 50 Gbps have been demonstrated. The bandwidths of injection rings on the other hand is severely limited by the carrier recombination time and typically are of the order of 1 GHz or less. For this reason injection-ring drivers require a significant amount of pre-emphasis and relatively large driving voltages. To achieve this pre-emphasis or finite impulse response filter stages need to be added to the driver circuitry. These additional circuits increase the complexity and power consumption of the drivers. With these caveats injection-ring modulator drives have been demonstrated for bit rates in excess of 20 Gbps.

Receiver Circuits

The receiver circuits for microring links do not present any peculiarity compared to standard photonic receivers. The main blocks of a receiver circuit include a trans-impedance amplifier, optional additional amplification stages, discrimination stages, and clock data recovery. The fact that the Silicon Photonic circuit can be closely integrated with the receiver circuit either through 3D integration or monolithic integration simplifies the design of the receiver first stage. The receiver sometimes is integrated with the control circuitry necessary to keep the drop filter in resonance with the incoming modulated laser.

Controllers

Controller circuitry is used to keep the transmitter modulator and drop-off filter in the receiver on resonance with the laser line. These circuits require a sensor and an actuator. For the sensing element a photodiode is often used to verify that the peak level of the modulated light is optimized. This can be accomplished using a peak detector in the receiver circuit. On the transmitter side a similar approach can be used if an add-drop configuration is used for the modulator ring. Two actuator strategies are commonly used with microrings. Increasing a ring temperature allows one to red-shift the resonant frequency. Charge injection allows one to blue-shift the resonance. Charge injection is more energy efficient than thermal tuning but the tuning range is limited. Ideally a combination of the two methods should be used in order to optimize power consumption.

Laser Sources

A multi-wavelength laser source is a key element of the DWDM link described earlier. Such source can be realized by combining multiple lasers through an AWG (array waveguide grating) or similar multiplexing component. Intrinsically multi-wavelength sources can also be used. Comb lasers based on quantum dot gain materials provide multiple independent low-noise lines that can be used for modulation. This solution is much simpler and less expensive than using multiplexed individual lasers.

Link-Level Considerations

Not much can be said about link-level tradeoffs here. In general a system-level analysis of the link is necessary to achieve optimal performance whereas optimizing a single component or figure of merit (such as the line toggle rate) will most likely result in sub-optimal link performance. It is also important to consider the compatibility and availability of different technologies in a single platform.

References

- Bogaerts, W., *et al.*, 2012. Silicon microring resonators. *Laser & Photonics Reviews* 6 (1), 47–73.
- Bruel, M., 1995. Silicon on insulator material technology. *Electronics Letters* 31 (14), 1201–1202.
- Li, C., *et al.*, 2014. Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65 nm CMOS. *IEEE Journal of Solid-State Circuits* 49 (6), 1419–1436.

- Marcuse, D., 2013. Theory of dielectric optical waveguides, 2nd ed London: Elsevier.
- Rhim, J., *et al.*, 2015. Verilog-A behavioral model for resonance-modulated silicon micro-ring modulator. *Opt. Express* 23, 8762–8772.
- Shin, M., *et al.*, 2016. Parametric characterization of self-heating in depletion-type Si micro-ring modulators. *IEEE Journal of Selected Topics in Quantum Electronics* 22 (6), 116–122.
- Soref, R., Bennet, B., 1987. Electrooptical effects in silicon. *IEEE J. Quant. Electron.* 123–129.
- Wu, R., *et al.*, 2016. Large-signal model for small-size high-speed carrier-injection silicon microring modulator. New Orleans, LA: OSA Integrated Photonics Research, p. IW1B.4.

Optical Switches

Dritan Celo, Dominic J Goodwill, and Eric Bernier, Huawei Technologies Canada, Kanata, ON, Canada

© 2018 Elsevier Ltd. All rights reserved.

Introduction

An optical switch is a networking element that directs light from optical input fiber(s) to optical output fiber(s) without converting the light to an electronic signal. An optical switch is controlled by an electronic controller. Applications for optical switches include DWDM network reconfiguration, test automation and packet networking.

Evolution From Static Links to Optical Switching

An optical fiber link carries information from a transmitter to a receiver. Early optical fiber network deployments comprised static optical links, of three general types. Wavelength add-drop networks are used for metro and long-haul deployments, and comprise wavelength add-drop rings and multi-wavelength point-to-point links. Passive optical networks (PON) have a root and leaf topology with time-division multiplexing and, in some cases, wavelength filtering. Intra-datacenter, LAN and super-computer deployments comprise single-point to single-point optical links. In all three environments, all of the optical paths were fixed when the equipment was installed. Switching of circuits, flows or packets happens in electronic elements positioned between an optical receiver and an optical transmitter, in an architecture known as optical-to-electrical-to-optical (OEO), Fig. 1. Even today this remains the conventional way to switch information. However, such OEO conversion uses transceivers at the switch, which are expensive, bulky, energy-intensive, and whose design usually supports a small set of bit-rates and protocols.

It is also possible to reconfigure or switch the optical links. This capability is known generically as optical switching. In optical switching, an electronic controller actuates a photonic element to change the connection of light between the optical transmitters and optical receivers. The architecture is known as optical-optical-optical (OOO), where the center “O” represents the optical switch itself. As will be described below, in some cases the photonic element is a switch that is positioned part-way along the optical path, and in other cases some aspect of the transmitter and/or receiver is actuated (e.g., the wavelength is tuned) which results in a change in the optical path.

An optical switch is transparent to signaling format and protocol (within limitations of spectral bandwidth, loss budget and other physical optical effects) and therefore simplifies the hardware requirements in the data plane. Because there is no electronic conversion of the signals, it is believed that this method of switching will consume less energy, an important requirement to maintain growth of data centers and super computers. Therefore, upgrades of bit-rate or protocol can be accommodated without the need to replace the switch.

Terminology of Optical Switching

The term used in this article is “optical switching”. The term “photonic switching” is also used for this capability, as it is the destination of the photons which is changed.

The term “all-optical switching” is sometimes used, but this term is ambiguous. For many years, in research labs, there have been elements where a control optical beam directly affects the destination of an optical signal, without the intervention of any transistors – this is truly an “all-optical” switch. It is out of scope for this article, as its practical use in communications systems is rare to date.

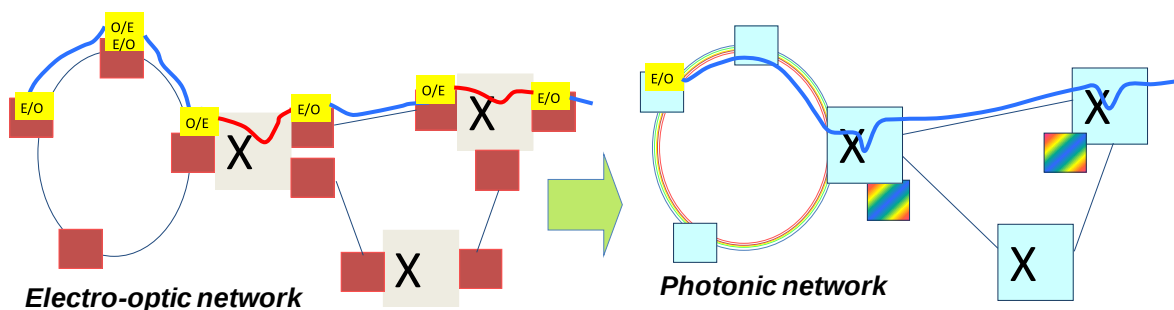


Fig. 1 A conceptual electrically switched optical-to-electrical-to-optical (OEO) network, compared to an optically switched optical-optical-optical (OOO) network.

Existing and Proposed Deployments of Optical Switches

Protection and provisioning

Optical switching is already widely deployed in metro and long-haul networks. The switching granularity is either a fiber, typically using 1×2 and 2×1 switches for protection switching when a fiber fails, or an individual wavelength, typically for adding new customers or adding additional bandwidth to a network, using reconfigurable optical add/drop (ROADM) and wavelength selective switch (WSS) technology. In current deployments of wavelength-granularity systems, the actual switching happens very rarely, although dynamic reconfiguration is frequently proposed as an important future opportunity. Switching time of milliseconds is typical within the switch itself, but the overall optical link takes 10's of seconds to several minutes to achieve stability in existing production networks.

Patch panels

Optical switching with fiber granularity is used in lawful intercept (fiber taps with selector switch), video production and test automation applications. Here, switches act as automated patch panels. Example technologies are: free-space MEMS to redirect beams from an input fiber collimator to a focusing lens of an output fiber, piezoelectric mechanical actuators to move a fiber relative to collimating and focusing lenses, and robot arms that move fiber patch cords. Switching time of milliseconds to seconds is typical.

Reconfiguration within data center

Optical switching with fiber granularity is currently being promoted for data center use where it can provide reconfiguration as workloads change, and can mitigate fiber wiring errors by craft people. Switching time of milliseconds is typical. However, there is a lack of published examples of real deployments.

Packet switch capacity growth

Optical switching with very fine granularity is a subject of much ongoing research, with the intent of capping or displacing OEO flow switches and packet switches, although there are currently no published commercial deployments. Switching time of nanoseconds to microseconds is typical. In this application, the energy-per-bit is a key metric, and photonic switching can in principle be more energy efficient than electrical switching. For clarity, optical switching applications are summarized in [Table 1](#).

Switch Characteristics

Granularity

The granularity of a switch is the smallest change that it can make. Example granularities of optical switches are fiber, wavelength, flow, and packet (ranked from coarsest granularity to finest granularity). As a rule of thumb, optical switches with coarser granularity are simpler to implement, easier to integrate with network operations, have better optical performance, are cheaper, and are more widely deployed than optical switches with finer granularity. Nonetheless, optics is penetrating further and further into electronic switches, and the opportunity for fine granularity optical switches is growing.

Scalability

The scalability of a switch is a conceptual term that indicates whether the technology and architecture of a switch can implement a large number of ports, without a catastrophic reduction in the quantitative parameters, compared to a switch with a small number of ports.

Quantitative Performance

Optical switches are commonly characterized by the following quantitative performance parameters.

Blocking probability is the fraction of legal connection maps that cannot be met because there is no available path inside the switch. Illegal connection maps are not included – for example, a map where two inputs are supposed to go to the same output is illegal, because this can never be serviced in a bufferless switch.

Table 1 Applications of optical switches

<i>Application</i>	<i>Switching time of the switch element itself</i>	<i>Commercial status</i>
DWDM network provisioning	Milliseconds to seconds	Widely deployed
Protection	Milliseconds	Deployed
Patch panel	Milliseconds to seconds	Deployed
Reconfiguration within data center	Microseconds to milliseconds	Emerging
Packet switch capacity growth	Nanoseconds	Development

Reconfigurability is related to blocking – it indicates whether or not existing connections have to be temporarily broken and reconnected when a new connection request arrives.

Switching time is the time between the switch receiving a request for a new connection, and the new lightpath is established and meets all the technical parameters.

Insertion loss (IL) is the fraction of optical signal power lost from the input to the output of the switch, typically expressed in dB.

Path dependent loss is the uniformity of insertion loss over different paths, typically expressed in dB. Optical receivers have finite dynamic range, and the noise penalty from optical amplifiers depends on the optical power level into the amplifier. Hence it is desirable if the path dependent loss is small.

Optical crosstalk is (in its simplest form) the ratio of the power at a specific output port from the desired input port, to the total power leaking from all the other inputs into that output port. A more realistic optical crosstalk calculation has a complicated dependence on the relative wavelengths of the input signals, coherence length and other link engineering parameters.

Multipath interference (MPI) is the fraction of optical power which takes a different path through the switch than the desired path. It arises from multiple back-reflections and leakages within the switch.

Extinction ratio (ER) is the ratio of the output power in the on-state to the output power in the off-state. Typical ER values are 40–50 dB for opto-mechanical switches, while the ER for photonic integrated circuit switches varies widely depending on the technology and architecture with reported values from 10 dB to 60 dB.

Polarization-dependent loss (PDL) is the range of insertion loss for different polarizations of the input light.

Differential group delay (DGD) is the range of time delay through the switch, for different polarizations of the input light. DGD causes deterioration of the optical signal quality, especially when large DGD is combined with large PDL as this is a non-linear effect that cannot be compensated using digital signal processing at the receiver.

Spectral flatness is the variation in the other technical parameters over the overall operating wavelength range of the switch, while ripple is the variation within the pass-band of a given channel. In the case of switches that are not wavelength selective (e.g., Mach-Zehnder switches, MEMS space switches), ripple and spectral flatness are equivalent terms. In the case of switches that are wavelength selective (e.g., WSS or ring-resonator switches), ripple indicates how much an individual optical signal will be corrupted, whereas spectral flatness indicates how the major parameters (insertion loss, PDL, DGD and so on) vary from one wavelength channel to another wavelength channel.

Engineering considerations of size, power consumption, cost, reliability and the ability to monitor the switch are also important parameters.

As a general rule, switches with small optical penalty are preferred, although to an extent the insertion loss and other penalties can be compensated using optical amplifiers and improved transmitters and receivers. The performance requirements vary greatly depending on what services are passing through the switch and the optical link budget from transmitters to receivers. Hence the quantitative requirements on an optical switch are specific to the network in which it is deployed.

Families of Optical Switch Technologies

Table 2 lists the families of optical switching technology that are most commonly used today. Key technologies are described in more detail later in this article.

Table 2 Optical switch technologies

<i>Family</i>	<i>Propagation medium</i>	<i>Mechanism</i>	<i>Types of actuator</i>
Free-space optical	Gas	Angle of the beam is steered by actuators	Tilting MEMS mirror Diffractive beam steerer comprising MEMS or liquid crystal Piezoelectric motor moves fiber relative to lens
Fiber-optic	Fiber	Fiber is physically moved by actuators	Mechanical motor moves fiber
Planar lightwave circuit	Optical waveguides on a chip	Actuators on the chip determine which waveguides the light passes through	Variable interferometer using thermo-optic, carrier injection, or electro-optic effects. Typical materials: silicon, silicon oxide/nitride, III-V semiconductor. Mach-Zehnder and microring interferometers are common. Variable gain in III-V semiconductor amplifier.
Tunable laser	Static network of DWDM filters	Actuator in the laser determines wavelength, which determines the route through the filters	Waveguide coupler, moved by MEMS. Thermo-optic, carrier injection, or movable mirror, in III-V semiconductor laser

Key Application Today: Reconfigurable Optical Add Drop Multiplexer, Using Wavelength Selective Switch

A reconfigurable optical add-drop multiplexer (ROADM) is an optical switch system with wavelength granularity, often used in access ring networks, and in rings and mesh networks that cover cities or larger regions. ROADMs can provide add/drop functions on rings, and cross-connection at hub locations. Initially, ROADMs were controlled by network provisioning software. Currently, many ROADM vendors offer the capability to integrate the ROADM controller into a software defined network.

At the heart of most ROADMs that are deployed today is one or more wavelength selective switch (WSS). A WSS has optical fiber inputs and outputs that each carries multiple wavelengths. The WSS selects which wavelengths from a given input go to a given output. Each output fiber can only carry one signal of a given wavelength. The WSS cannot change the color of the light. The WSS may also provide balancing of optical power across the wavelengths on an output fiber, and monitoring of optical power in each wavelength.

The fibers connected to a WSS are intended to carry multiple wavelengths, and thus network fibers connect directly to the WSS (typically via an optical amplifier to compensate for insertion loss and fiber loss). However, WSS are also used to connect from transceivers to/from the network fibers. A transceiver performs electro-optic conversion on only one wavelength at a time. Therefore between the WSS and optical transceivers there is an optical de/MUX, optical splitter/combiner, and/or an optical cross-connect to separate and direct the individual wavelengths. If the ROADM is used with coherent transceivers, then an external deMUX is not needed, because a coherent optical receiver contains a built-in wavelength filtering function achieved by its local oscillator laser. However, if the ROADM is used with direct detect transceivers, then an external deMUX is needed, because the receiver cannot distinguish light of different wavelengths.

Fig. 2 shows architectural differences between three types of ROADM: classic ROADM, colorless directionless (CD) ROADM, and colorless-directionless-contentionless (CDC) ROADM. The CDC ROADMs offer greater flexibility, but are more complex. For illustration purposes, the classic ROADM (which is colored, has direction, has contention) is shown with direct detect transceivers and therefore has external deMUX, whereas the CDC ROADM is shown with coherent transceivers and thus does not need an external deMUX.

ROADMs have the following attributes:

Degree

The degree is the number of network fiber pairs that connect to a ROADM.

Colorless

The ROADM is colorless if the wavelength that can be added at a given add fiber, or dropped at a given drop fiber, can be changed. Early ROADMs had a fixed wavelength at each add fiber and each drop fiber, and thus were “colored”.

Directionless

Each add fiber/drop fiber can connect to/from any network fiber port. If a given add fiber and drop fiber can only connect to a given network fiber, then the ROADM is not directionless.

Contentionless

In a contentionless ROADM, the ROADM can add and drop as many instances of a given wavelength as it has network fiber pairs. For example, consider a ROADM that has 3 network fiber pairs, as well as add and drop. Imagine that each fiber is carrying signals on the 1549.32 nm wavelength, as well as on other wavelengths. An operator may wish to add and drop all three of the 1549.32 nm signals. If the ROADM can achieve this, it is contentionless. In ROADMs that have contention, instances of the same wavelength would collide at some internal element, and hence cannot be simultaneously added/dropped.

WSS can be implemented using two types of wavelength plan:

Fixed Grid

The wavelengths are on a predetermined grid, typically 100 GHz or 50 GHz, and have a predetermined and static optical pass-band.

Flex-Grid

The separation of the wavelengths and the pass-band shape of each wavelength can be varied.

The most adaptable type of ROADM is a flex-grid CDC (colorless-directionless-contentionless).

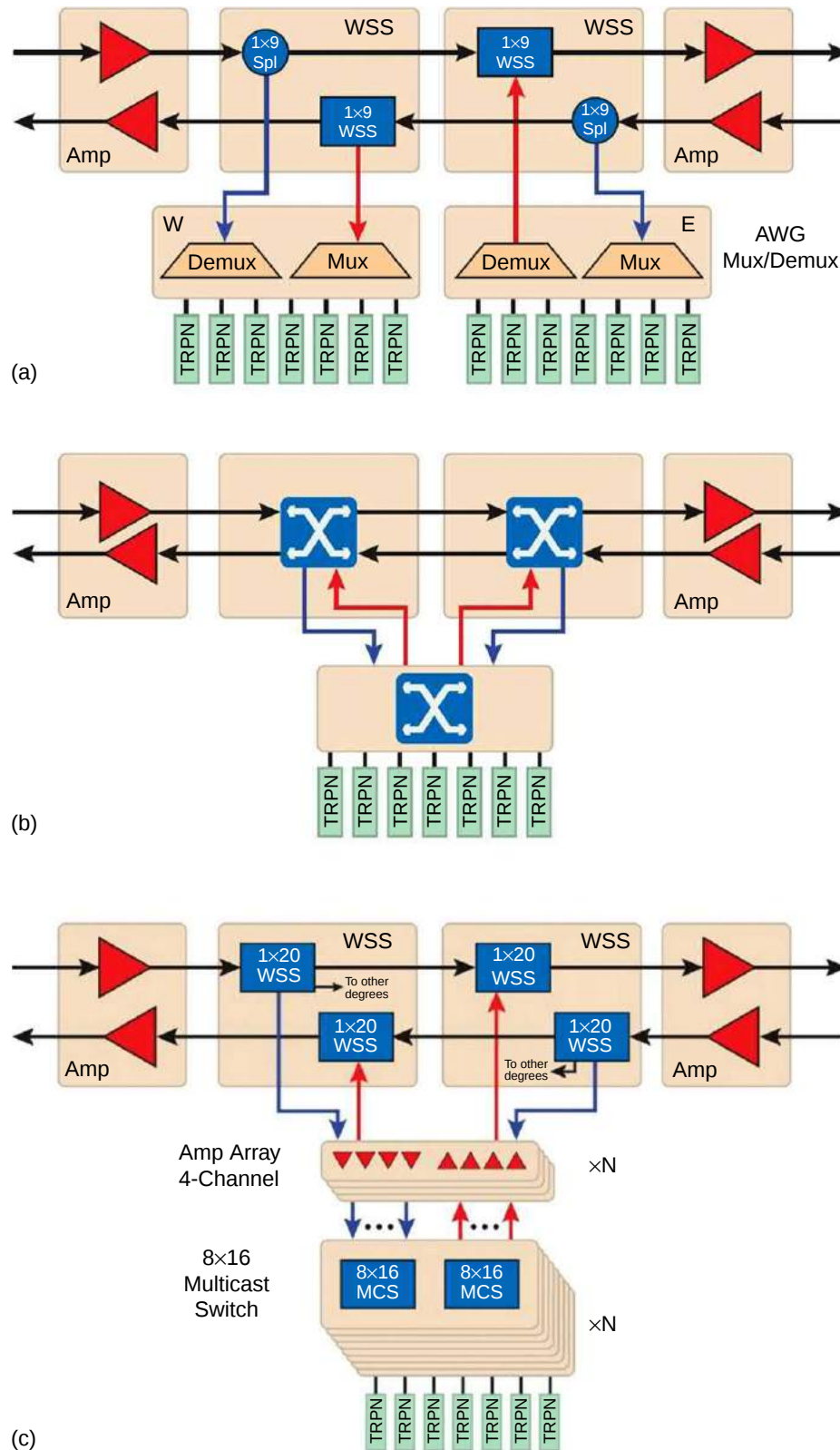


Fig. 2 Types of ROADM (a) classic, (b) CD, and (c) CDC. The CDC ROADMs offer greater flexibility, but are more complex. For illustration purposes, the classic ROADM (colored, has direction, has contention) is shown with direct detect transceivers and therefore has external deMUX, whereas the CDC ROADM is shown with coherent transceivers and thus does not need an external deMUX. (After “CD ROADM – Fact and Fiction” Part 1: Applications, Fujitsu).

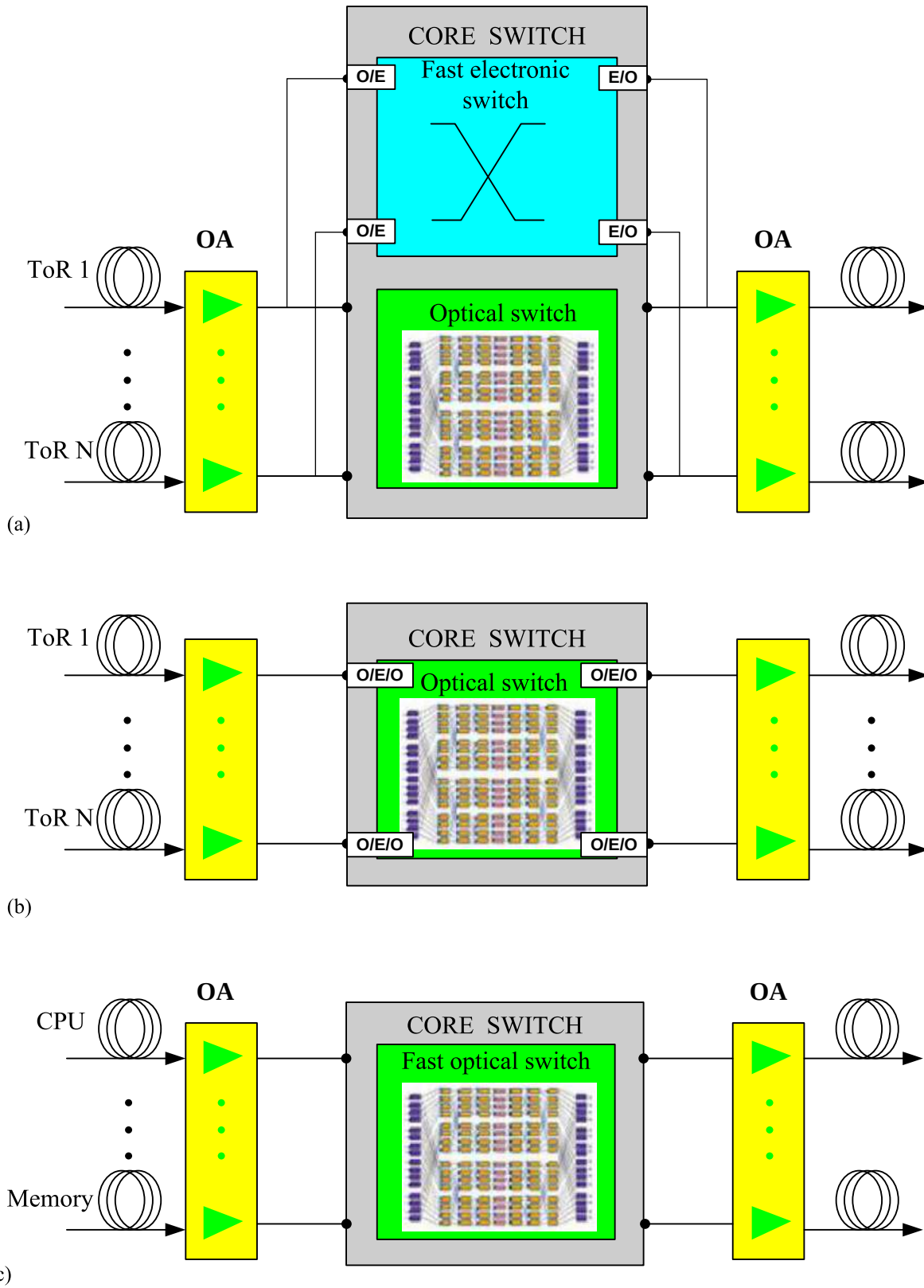


Fig. 3 Applications of optical packet switches: (a) Core router growth, (b) switch core growth, (c) high performance computing switched intraconnect.

Key Proposed Application: Optical Packet Switching

Optical packet switches are the subject of ongoing research, although they are not yet significantly deployed. For the sake of brevity, optical packet switches here include switches capable of switching individual packets, but also optical switches capable of switching custom-length frames or cells, or microsecond-duration dataflow.

Three applications are identified and shown in Fig. 3. Bandwidth requirements continue to increase in these applications. Electronic packet switches are ubiquitous today, but increasing their capacity faces increasing technical challenges related to heat dissipation, power consumption, and signal integrity.

Core Router Growth

At the core routers of a datacenter, an optical packet switch is added alongside an existing electrical router, to offload it. This may be particularly effective to offload longer packets and flows onto the optical packet switch, while leaving short packets in the electrical packet switch, because it is not practical to switch short packets optically.

Switch Core Growth

Conventional OEO switches contain OE and EO interfaces from/to other network elements, and these interfaces are connected through electrical switch chips and an electrical backplane inside the switch. Instead, the internal circuit can be replaced with additional EO and OE interfaces inside the switch, connected through an optical packet switch and an optical backplane.

HPC Switched Intraconnect

Dynamic connection between processing, memory, storage and I/O in a high performance computer (HPC).

In the first application, the optical packet switch switches light that has come from elsewhere in the network – for example, the optical transceivers may be on the other side of a datacenter from the optical packet switch. Hence the optical transceivers are also providing data transport. In the second and third applications, the additional optical transceivers are part of the switch, and are not providing data transport. The second and third applications become increasingly attractive when electrical chips have direct optical I/O, which has been proposed as a solution to the worsening I/O interconnect bottleneck of electronic chips.

Fast optical switch technologies have enormous data throughput, and can be highly efficient in terms of energy per bit. Their switch times are still relatively slow (several nanoseconds at best, for technologies that have good scalability), compared to electronic switches. Optical switches operate strictly as time-of-flight switches, and have no scalable capability for random-access-memory buffers. On the other hand, electrical packet switches switch at sub-nanosecond speed and most such switches today use a memory-based switching architecture, including random-access-memory buffering to handle contention. Hence a network using an optical packet switch cannot simply copy a conventional electronic packet network. Instead, the most practical solutions use a scheduler and ingress buffering to solve time-of-flight contention, and may also retain a conventional electrical switch for some classes of traffic (Fig. 4). Ingress buffering means that the source of the data is buffered at the network element before the optical

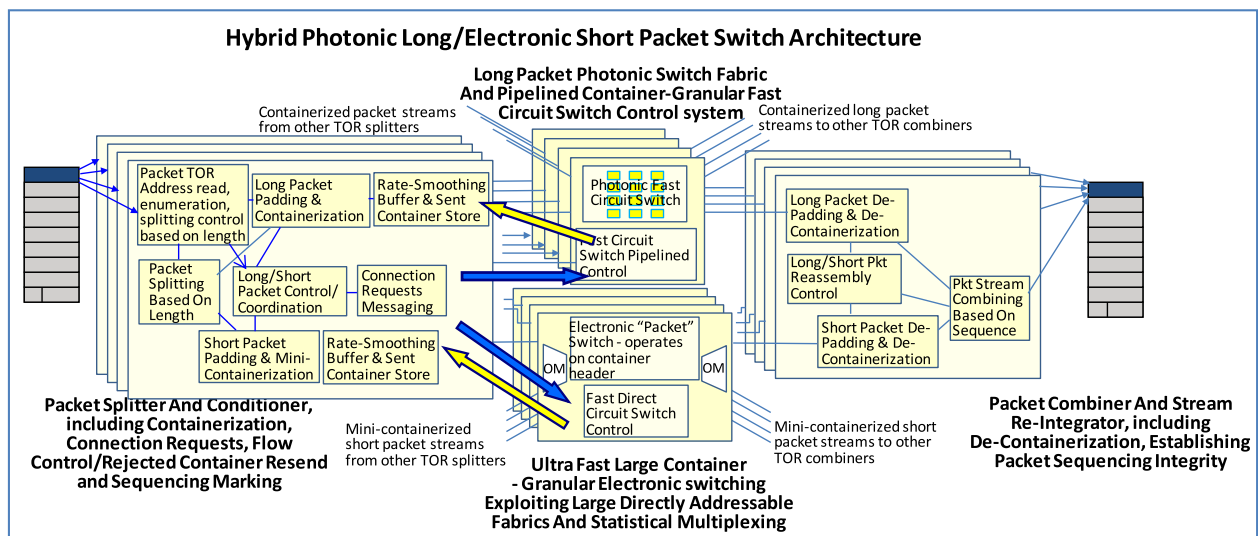


Fig. 4 Hybrid optical packet switch and electrical packet switch, illustrating ingress buffering for data that is handled by the optical packet switch.

packet switch, until the scheduler indicates that it is time to send. This architecture is illustrated in the figure. Thus, it is likely that electrical and optical packet switches will coexist and cooperate.

As will be discussed below in the section on planar lightwave technology, a realistic limit today for a fast photonic switch chip is around 32×32 to 64×64 ports. By a careful choice of parallel architecture, a switch implemented using these chips can implement 1000–4000 I/O ports, each having a throughput of 400 Gb/s or higher. Thus, the overall capacity of an optical packet switch can reach Petabits per second.

Free-Space Optical Switch Technologies

This section describes optical switches that operate by steering beams of light in air (or some other suitable gas). These switches are suitable for circuit switching as they switch relatively slowly (milliseconds), although some LCoS optical switches can achieve microsecond switching time and have been used for experiments with switching short-duration data flows. Generally, free-space optical switches have excellent optical performance. They are precision opto-mechanical assemblies, and so their manufacturing is relatively time-consuming. Careful engineering of vibration control and actuator feedback is used to achieve good alignment.

The 2D-MEMS, 3D-MEMS and piezoelectric optical switches described in this section have fiber switching granularity. If the application calls for switching individual wavelengths, then fiber-coupled wavelength deMUX and MUX components are needed before and after the optical switch. Single-mode and multi-mode versions have been built.

The WSS described in this section have wavelength switching granularity. They contain built-in diffraction gratings that perform wavelength deMUX and MUX. They always use single-mode fibers.

Liquid-Crystal on Silicon Wavelength-Selective Switch

Wavelength selective switches have a collimating lens after each input fiber, and collimate the incoming light onto a diffraction grating. The diffraction grating disperses each signal wavelength into a beam at a different angle. The beams then fall upon an actuator, which reflects them to a desired output fiber focusing lens. Generally, a diffraction grating is used to recombine signal wavelengths (which may be from different inputs) into the output fiber.

The form of the actuator has evolved over time. Early implementations used a MEMS-actuated mirror to tilt each beam, or used a small number of liquid crystal pixels to reflect the beam. By applying a controllable voltage to each liquid crystal pixel, its refractive index is changed. A set of liquid crystal pixels forms a phased-array reflector, which steers the reflected beam. These implementations are limited in their spectral resolution, and typically have fixed grid wavelength granularity at 50 GHz or more channel separation.

Liquid-crystal on silicon (LCoS) WSS were introduced commercially in the mid-2000s and are particularly intended for flexible grid operation. A concept demonstration of a WSS with LCoS optical sensor is shown in Fig. 5. The liquid crystal medium is directly on top of a silicon CMOS control chip. The size of the pixels in an LCoS is very small compared to the size of the beam. The number of liquid crystal pixels hit by the beam depends on the spectral width of the signal in the beam. The control system chooses the appropriate set of pixels depending on the present set of wavelength channels in each fiber, which is learned from network operating software. LCoS granularity down to 6 GHz channel separation is possible. Commercial LCoS WSS with up to 20 fibers are commonly available today.

3D Micro-Electro-Mechanical Optical Switch

3D-MEMS optical switches became well-known around the year 2000. They are currently deployed mostly in patch panel and reconfiguration applications.

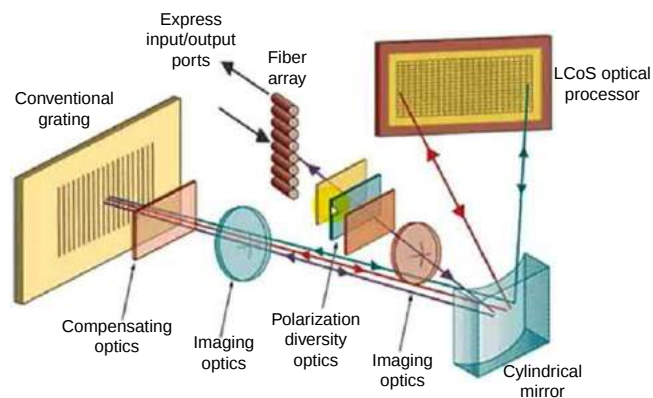


Fig. 5 Wavelength selective switch using liquid crystal on silicon.

A 3D-MEMS, as schematically demonstrated in Fig. 6, has a collimating lens on each input fiber that produces a collimated beam in a gas. The beam reflects off a mirror in a first plane of mirrors, which sends the beam to a mirror in a second plane of mirrors that steers the beam into the acceptance angle of a focusing lens of an output fiber. The beam travels a distance of centimeters through the gas. Each mirror array is a silicon substrate, with tiltable micro-mirrors that are driven by MEMS actuators. A controller tilts each mirror by a variable amount to achieve the desired beam alignment. Typically, an optical monitor in the output fiber provides feedback to the controller.

Most 3D-MEMS optical switches have from 64 to a few hundred ports, and the largest built had more than 1000 ports. Insertion loss is typically less than 3 dB, return loss and crosstalk are low, and switching time is in the order of milliseconds to 10's of milliseconds. The optical switch is non-blocking, operates over a wide wavelength range, and may be bi-directional (depending on the control system). The controller requires careful engineering to achieve and maintain good alignment, and to compensate for environmental vibration.

For a 3D-MEMS optical switch with N inputs and N outputs, there are $2N$ mirrors. For example, a 300×300 switch needs 600 mirrors.

The figures illustrate the 3D-MEMS concept. In practice, the optical path may be folded using static mirrors, to reduce physical volume and to allow the input and output fibers to be assembled in the same plane.

2D Micro-Electro-Mechanical Optical Switch

2D-MEMS optical switches were fashionable in the late 1990s, attracting significant research and industry investment. However, in more recent years they have attracted less attention than 3D-MEMS.

In a 2D-MEMS device, as conceptually illustrated in Fig. 7, each optical input fiber has a lens which creates a collimated beam, and each output fiber has a focusing lens and is at 90° to the input. There is a 2D array of mirrors built on a silicon substrate by MEMS manufacturing processes. To connect an input to an output, one mirror is inserted into the beam, and the other mirrors are moved out of the way. The mirror has two very well defined states; it is either "not inserted" or is "inserted". Several insertion

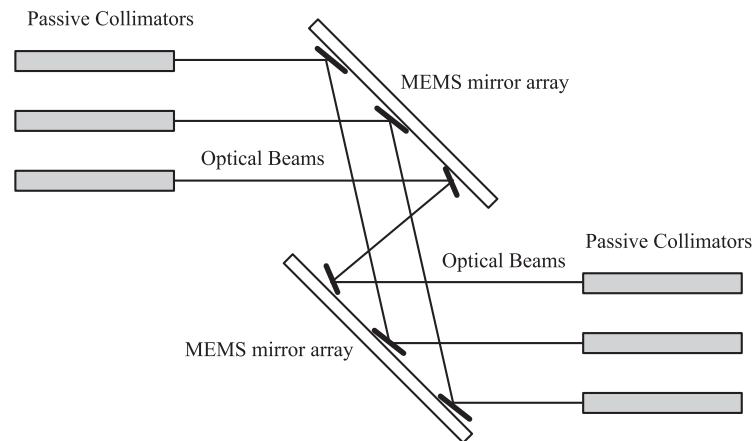


Fig. 6 Schematic of a 3D-MEMS optical switch.

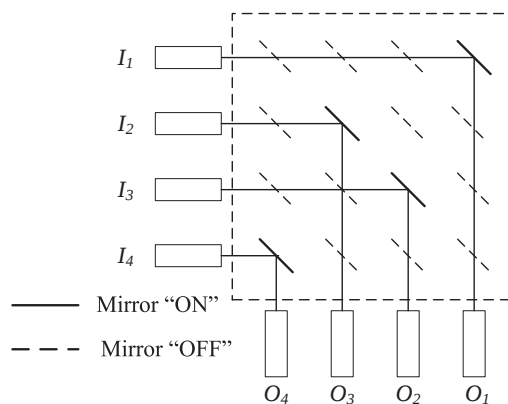


Fig. 7 Schematic of a 4×4 2D-MEMS optical switch.

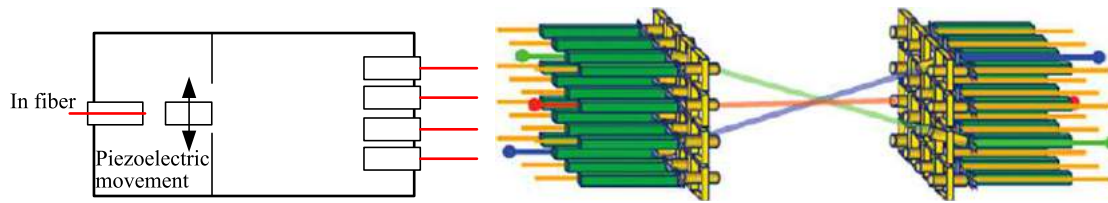


Fig. 8 Schematic of a piezoelectric optical switch. Reproduced from Jencotech, Available at: <http://www.jencotech.com/documents/PolatisOverviewRev1.pdf>.

mechanisms have been used: the mirror is stood up/laid down, the mirror is slid sideways into the beam, or the mirror is lowered into the beam. Actuators are electrostatic or micro-motors. The mirror is typically silicon, with a reflective coating. Generally, the angle of the mirror is defined by the manufacturing and is not adjustable. For an optical switch with N inputs and N outputs, there are N^2 mirrors.

Switching time is typically in the order of sub-milliseconds, which is 5–10 times faster than the 3D-MEMS devices. The advantage is due to simple direct digital control operation of the mirror movement from a rest position to an active position.

A 2D-MEMS has excellent optical properties because the optical path contains just 2 lenses and one mirror. However, it becomes difficult to achieve good optical beam control at large port count, as the distance between the lenses is larger and thus angular alignment during manufacturing becomes more critical. Typical larger elements are around 16×16 ports.

Piezoelectric Moving-Fiber Optical Switch

A piezoelectric optical switch (Fig. 8) is used in similar applications as a 3D-MEMS switch. It has similar capabilities, performance, physical size, and number of ports as 3D-MEMS, but it uses a different optomechanical mechanism.

In a piezoelectric optical switch, each input fiber has a collimating lens, and each output fiber has a focusing lens and there is a gas region between the lenses in the middle, much like the 3D-MEMS. The position of each fiber is moved slightly with respect to its lens, using piezoelectric actuators, so that the beam from the collimating lens is steered toward a desired focusing lens, and the acceptance angle of the focusing lens is tilted so as to couple the beam into the output fiber.

Planar Lightwave Circuit Optical Switch Technologies

Planar lightwave circuit (PLC) optical switches comprise optical waveguides which are manufactured on wafers using lithographic processes, then diced into chips and packaged with optical fibers for optical I/O. Each connection from an input to an output on the chip is called a lightpath. To set up the lightpaths, electrical controllers connect to the planar lightwave circuit. The electrical controllers are usually off-chip. In almost all cases, PLC optical switches are single mode and therefore are compatible only with single mode fiber.

Materials and Physical Principles

A wide variety of materials have been used for PLC optical switches, including silicon-on-insulator, silicon dioxide, silicon oxynitride, GaAs, InP and related III-V semiconductors, polymers, electro-optic polymers, and electro-optic crystals of which lithium niobate and PLZT are the most popular.

PLC optical switches operate using one of the following physical principles:

Interferometer

The electrical signal creates a change in refractive index of the optical waveguide. Each optical switch cell is arranged as an interferometer with 2 inputs and 2 outputs. Changing the refractive index changes which input is connected to which output. The most common forms of interferometer are Mach-Zehnder and microring resonator. To change the refractive index, the electrical signal changes the temperature of the waveguide (thermo-optic effect in dielectric, semiconductor or polymer), changes the density of electrical carriers (carrier injection in semiconductors), or changes the electric field (in materials with an electro-optic coefficient).

Moving waveguide coupler

The electrical signal causes a waveguide to move physically. The moving waveguide is part of a directional coupler with 2 inputs and 2 outputs. Moving the waveguide changes which input is connected to which output. See the section on waveguide-MEMS.

Blocker

The electrical signal changes the optical waveguide from absorbing to having gain. The optical switch cell is a gate with 1 input and 1 output. The gate blocks or transmits light. See the section on SOA switches.

All these PLC optical switches are spectrally broadband and therefore switch all wavelengths within a 10's-nm wide spectrum, except the microring resonator which is narrowband and switches a single WDM wavelength.

Scaling up a Planar Lightwave Optical Switch From Small Building Blocks

Many planar lightwave optical switch technologies can only implement cells with 1 or 2 inputs and outputs. To achieve scalability, these cells are connected together in a switch matrix, preferably with as many cells as possible on one die. In addition to the external characteristics of the overall optical switch, the internal topology of the switch matrix can be described using the following characteristics:

Symmetric

In a symmetric topology, every lightpath uses the same number of cells. In an asymmetric topology, the number of cells in a lightpath depends on the lightpath.

Deterministic

In a deterministic topology, there is exactly one path from a given input to a given output. In a non-deterministic topology, there is more than one possible path, and a routing algorithm considers all the requested connections.

Stages

In a single stage switch, one cell is actuated per lightpath. In a multi-stage switch, multiple cells are actuated per lightpath.

Some important topologies are listed in the [Table 3](#) and illustrated in [Fig. 9](#). For simplicity, the table shows the case where the number of input ports N is equal to the number of output ports, which is known as an $N \times N$ switch. Multi-stage switches are most efficient when the number of ports is a power of 2. Any of these switch matrices can, in principle, be cascaded to make a Clos architecture, although currently actual deployments of a Clos optical switch are rare to nonexistent.

The **crossbar** fabric is a type of switching technology where each node is connected to every other node. The interconnection between the inputs and the outputs is achieved by appropriately setting the states of the 2×2 switches. For a $N \times N$ switch, the total number of 2×2 switching cells is N^2 . [Fig. 9\(a\)](#) demonstrates connectivity in a 4×4 switch matrix. To connect input i to output j , the light traverses the 2×2 switches in row i until it reaches the column j and then traverses the switches in column j until it reaches output. The shortest light path length is 1, and the longest is $2N - 1$, resulting in large variation in path length. The crossbar architecture is non-blocking, has a simple layout, and small cell count at small N . Drawback of this topology is the large number of crossings and large cell count at large N .

The **switch and select** architecture consists of an input switch array, an output switch array, and a passive crossover network. The example of [Fig. 9\(b\)](#) demonstrates a 4×4 switch matrix. The number of switches is uniform in all optical paths and follows logarithm scaling instead of the linear scaling. The passive crossover network connects the N^2 outputs from the input switch array to the N^2 inputs of the output switch array. The large cell count at large N and the large number of on-chip waveguide crossings limit the potential for scalability and introduces different insertion loss on the different paths.

Table 3 Switch topologies used with planar lightwave circuit switch cells

Topology	# Cells per path	# Cells total	Type	Blocking	Strength	Weakness
Crossbar	1 to $2N - 1$	N^2	Single stage, symmetric deterministic	None	Simple layout Small cell count at small N No blocking	Many crossings Large variation in path length Large cell count at large N
Switch & select	$2 \log_2 N$	$2N(N - 1)$	Multi-stage symmetric deterministic	None	One plane of crossings	Large cell count at large N
Path independent loss	$2 \log_2 N + 2$	$2N^2$	Multi-stage symmetric non-deterministic	None	Equal loss on all paths	Large cell count at large N
Banyan	$\log_2 N$	$(N \log_2 N)/2$	Multi-stage symmetric non-deterministic	Blocking	Small number of cells	High crosstalk Blocking
Benes	$2 \log_2 N - 1$	$N \log_2 N - N/2$	Multi-stage symmetric non-deterministic	Almost none	Small number of cells	Very high crosstalk
Dilated Benes	$2 \log_2 N + 2$	$2N(\log_2 N + 2)$	Multi-stage symmetric non-deterministic	Almost none	Low crosstalk Efficient routing algorithm	Larger number of cells in path

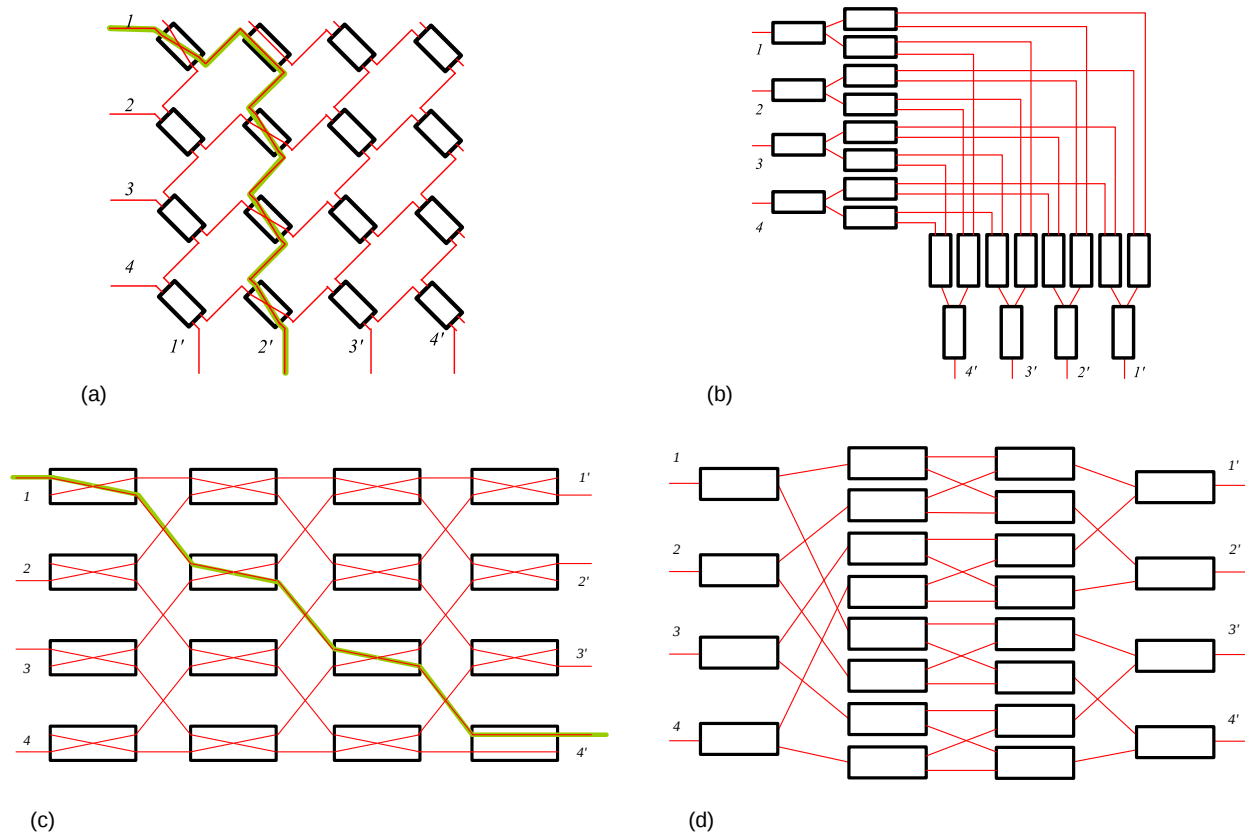


Fig. 9 Schematic of four topologies constructed from small switches: (a) crossbar switch, (b) switch and select switch, (c) path-independent loss (PILOSS) switch, and (d) dilated Benes switch.

PILOSS is another standard switch topology for building $N \times N$ switches. It consists of a matrix of 2×2 element switches as shown in Fig. 9(c). It usually is designed to have the element switches 'normally' OFF (cross-state), therefore reducing the total power consumption of the switch matrix. In this topology, light always passes through N element switches and $N - 1$ intersections, thus experiences the equal attenuation for all the $N^2 -$ ways of paths. The maximum number of ON state element is N , making the PILOSS topology energy-efficient. For this reason, it is believed to be one of the most promising towards larger N .

A hybrid dilated Benes with crosstalk suppression is shown in Fig. 9(d), comprised of an ingress column of 1×2 cells, two vertically replicated 4×4 Benes in the middle, and an egress column of 2×1 cells. By replacing the central column of 2×2 cells each with a 2×2 dilated Banyan, crosstalk noise can be suppressed. In addition to crosstalk suppression, use of wavelength constraint routing (routing of lightpaths within the switch matrix to minimize how many time different lightpaths meet within the switch matrix) reduces output crosstalk.

Materials for Planar Lightwave Optical Switches

Silica-on-silicon is a dielectric that is widely used in passive planar lightwave circuits. The core and cladding are composed of silicon dioxide with different dopants that modify the refractive index slightly. The fabrication process is inexpensive, connection to single-mode optical fiber is relatively easy and low loss, and optical propagation loss in the waveguide is low. Silica has a thermo-optic effect which can be used as a switch actuator. However, bends are very large, due to the small refractive index step between the core and cladding of the waveguide. Therefore it has very poor scalability for an optical switch.

Silicon oxynitride (SiON) is similar to silica-on-silicon, except that the waveguide is made of different compositions of SiON. The refractive index of the SiON can be adjusted in the range 1.45–2.01 depending on the composition, which is wider than the refractive index variation in silica. This allows smaller bend radius. SiON is increasingly used for passive planar lightwave circuits, but its use for optical switches is relatively uncommon.

Perovskite crystals include lithium niobate and lead lanthanum zirconium titanate (PLZT). Waveguides are formed by doping the material, typically by diffusing metal ions into the material. Perovskites are used for very fast switches with up to 8 ports. They have an electro-optic effect, whose switch time is limited only by the electronic controller. The electro-optic effect is much weaker per unit length than thermo-optic and carrier injection effects, so a perovskite optical switch cell is large. Therefore, perovskites do not scale up to very large port count.

III-V semiconductors that are epitaxially grown on a GaAs or InP wafer can create thermo-optic interferometer optical switches, carrier-injection interferometer optical switches and SOA optical switches. The waveguide is formed using III-V materials of differing composition. Further, when using III-V semiconductor interferometric optical switches, non-switching SOAs can be monolithically integrated to provide optical gain to compensate for optical losses in the switch circuit. III-V semiconductors are ubiquitous for optical transmitters and receivers, but activity on III-V semiconductor optical switches has declined in the 2010s.

Silicon photonics (SiPh) is fabricated in the silicon layer of a silicon-on-insulator substrate, with silicon dioxide as the waveguide cladding. SiPh can create thermo-optic interferometer optical switches, carrier-injection interferometer optical switches and waveguide-MEMS optical switches. Activity on SiPh semiconductor optical switches has grown in the 2010s as SiPh manufacturing technology becomes more widespread. Silicon is transparent over the optical telecommunication wavelength band (1300 nm–1600 nm). It has high refractive index contrast $n[\text{Si}]/n[\text{SiO}_2] = 3.48/1.45$ which allows very dense circuit layouts. It is cheap and can be manufactured using legacy silicon electronics processes that have exceptionally good process control. The small waveguide size and large index step causes large propagation loss, which can be overcome using very wide waveguides for long-distance routes on the chip.

Interferometric Optical Switch

Mach-Zehnder interferometer optical switch

A Mach-Zehnder interferometer optical switch cell (Fig. 10) comprises a 50/50 input splitter, two internal optical waveguides and a 50/50 output combiner. Commonly used 50/50 couplers are multi-mode interferometers (MMI) and directional couplers. The two internal optical waveguides are nominally the same length. An electrically-controlled phase shifter on one or both arms changes the relative phase of the light on the two arms as it enters the output combiner, which causes the output light to move from one output waveguide to the other. When the phase difference between the arms is $2N\pi$, the switch is in cross state, which means that input 1 is connected to output 2 and input 2 is connected to output 1. When the phase difference between the MZ arms is $(2N+1)\pi$, the switch is in bar state, which means that input 1 is connected to output 1 and input 2 is connected to output 2.

This figure shows only the phase shifter(s) that perform the switching function. However, it is impossible to manufacture the two waveguide arms to have equal optical path length to within a small fraction of a wavelength. Therefore, carrier injection switches also have a static thermo-optic phase shifter to bias the initial phase. If care is not taken, the power consumption of this bias thermo-optic phase shifter dominates the power consumption of the whole switch.

Microring resonator optical switch

Microring optical switches use the thermo-optic effect or the carrier injection effect, and switch a single wavelength of light. They have been proposed for add/drop bus switches, but not to-date been commercially deployed for this purpose.

A microring resonator (MRR) couples light from an input waveguide to an output waveguide when the wavelength of the light resonates with the round-trip of the microring. Microring resonators are usually made in silicon photonics. The diameter of the ring is from several micrometers to tens of micrometers. A single microring has a sharp resonance peak with a Lorentzian lineshape. This lineshape is not compatible with most WDM transmission formats. Multiple microrings can be coupled to produce a flatter-top lineshape that is compatible with a conventional WDM signal.

A significant challenge is that MRR are highly temperature dependent, very sensitive to manufacturing variations, and difficult to maintain alignment with the desired wavelength. In addition, a single-MRR with a sharp Lorentzian transmission is not favorable in a high-speed data transmission system where flat-top response is preferred to minimize the signal waveform distortion and thus the data transmission errors.

Fig. 11 shows a schematic view of an optical add/drop bus, formed using microring optical switch cells. Each microring couples one wavelength from the through waveguide to the add/drop waveguides. Each microring contains a phase shifter. Varying the phase shifter using the thermo-optic effect or the carrier injection effect determines which wavelength is added/dropped.

Microring optical switches have also been used to create switch matrices, such as the crossbar illustrated in Fig. 12. Each microring couples light from a west-to-east input waveguide to a north-to-south output waveguide. The wavelength that is coupled is controlled by the phase shifter in the microring.

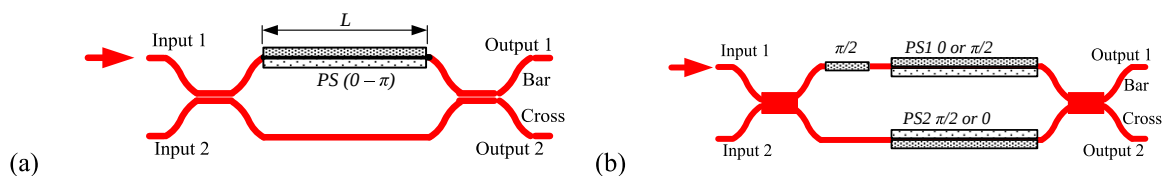


Fig. 10 Mach-Zehnder interferometer optical switch cell: (a) single-arm drive, (b) dual arm drive (push-pull).

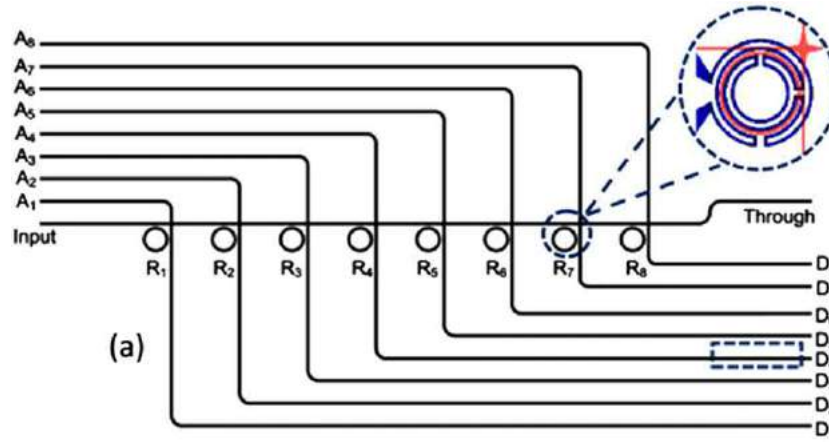


Fig. 11 Add/drop bus of an optical switch with MRRs. Reproduced from Danning Wu, Yuanda Wu, Yue Wang, Junming An, Xiongwei Hu, 2016. Reconfigurable optical add-drop multiplexer based on thermally tunable micro-ring resonators. *Optics Communications* 367, 44–49 (Elsevier).

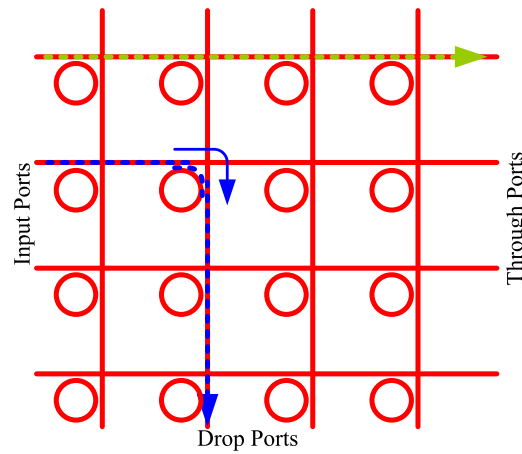


Fig. 12 Microring optical switch, arranged as a crossbar.

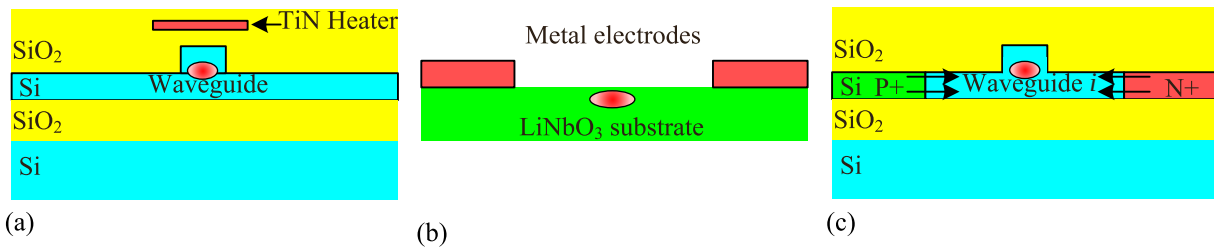


Fig. 13 Cross-section of a (a) thermo-optic phase shifter in silicon photonics (b) electro-optic phase shifter in lithium niobate substrate, and (c) carrier-injection phase shifter in silicon photonics or III-V semiconductor.

Optical phase shifters

Cross sections of typical phase shifters are shown in [Fig. 13](#). The phase change induced through the length of the phase shifter is $\Delta\phi = (2\pi\Delta n)/\lambda L$ where Δn is the change in the effective index of refraction of the guided mode due the electrical control, λ is the free-space wavelength of the propagating light, and L is the length of the phase shifter.

A goal of optical switching is to reduce the energy consumed per bit, compared to an electrical switch. However, a simple thermo-optic phase shifter in silicon photonics consumes around 30 mW of electrical power for π -phase shift (P_π), which is much larger than 1 pJ/bit for a complex circuit such as a 32-port matrix at a line rate of 100 Gb/s per port. A thermal undercut technique greatly improves the power consumption of a thermo-optic switch. [Fig. 14](#) demonstrates a thermo-optic switch cell in silicon

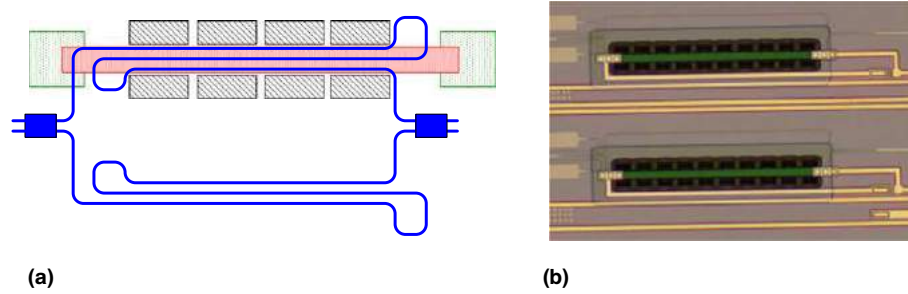


Fig. 14 Thermo-optic Mach Zehnder silicon photonic optical switch cell with thermal undercut: (a) schematic layout where a resistive heater (pink) located above one of the waveguide arms (blue) is used to change the index of refraction causing a phase shift in the propagating light wave, (b) microscope image of the fabricated switch.

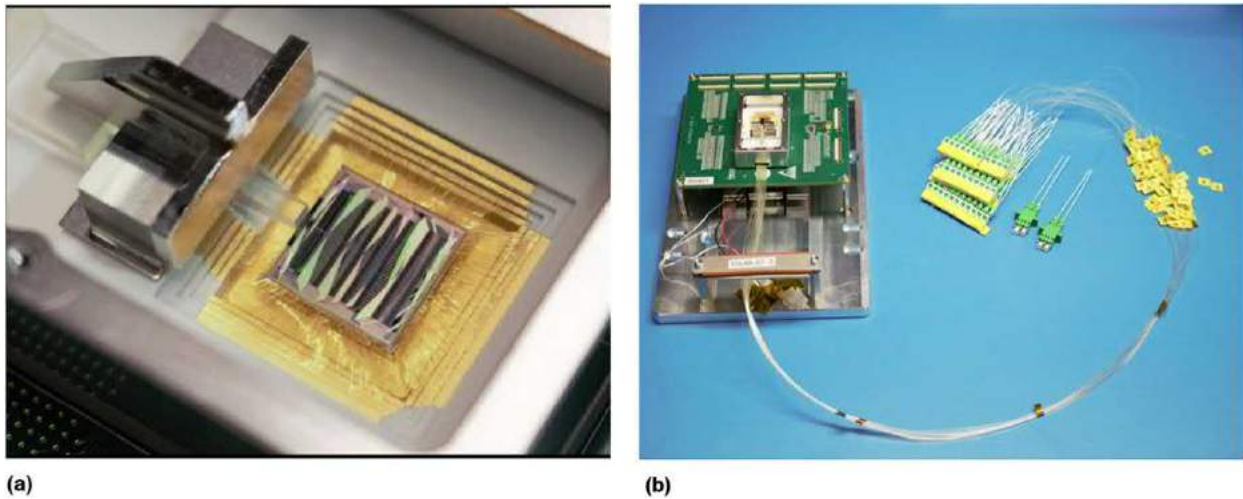


Fig. 15 32×32 SiPh switch (a) packaged die containing 448 cells with 900 photodiodes, wire-bonded in CBGA and silica waveguide mode adapter couples 64-fiber ribbon to SiPh, and (b) system assembly comprising optical package on breakout board, fiber ribbon, detector A/D boards (on back of breakout board), and thermal test plate. Note: the FPGA and heater driver board are not shown. After Dumais, P., Goodwill, D., Kiaei, M., *et al.*, 2017. Silicon Photonic Switch Subsystem With 900 Monolithically Integrated Calibration Photodiodes and 64-Fiber Package, OFC.

photonics with thermal undercut. Deep trench windows are etched vertically on each side of the thermo-optic phase shifter, followed by anisotropic wet chemical etching of the silicon substrate locally under the heated region. This creates a suspended heater and waveguide, which has low thermal conductivity to the silicon substrate, reducing the power consumption by a factor of 20. In addition, folded waveguide arms under the heater increases the interaction length between the heater and waveguide, thus reducing the power consumption by a factor of the number of folds, which is typically 3. The energy consumption then falls to 100 fJ/bit for the entire optical switch matrix, which is competitive with electrical switches.

A demonstration of a fully-assembled, fully-functional thermo-optic 32×32 silicon photonic optical switch matrix with thermal undercut for reduced power consumption is shown in **Fig. 15**.

A better-isolated heater region consumes less power, but needs more time to tune. **Fig. 16** shows the time-response, when the switch is driven by an electrical square wave with amplitude $P\pi$. The 10%–90% switching time increases from 70 μ s without undercut, to 1.34 ms with undercut. Nonetheless, as described earlier, millisecond response time is appropriate for many applications.

In a carrier injection phase-shifter, the mechanism to actuate the switch is the free-carrier plasma dispersion effect in a lateral p - i - n junction, which is centered on the optical waveguide. A forward-biased current is passed through the junction, which increases the carrier density. The lifetime of the carrier density is 1–2 nanoseconds, and hence the switch time of the phase shifter is several nanoseconds. Therefore, this mechanism is appropriate for optical packet switches. The electrical consumption is typically 1 to 3 mA at about 1 V. The increase in carrier density creates a decrease in refractive index, which saturates as the carrier density increases. However, it also creates an increase in absorption which increases monotonically with carrier density. In silicon photonics, the refractive index change before saturation is large enough to create a push-pull switch with only 0.5 dB absorption loss per switch cell and extinction ratio of 25 dB, which is sufficient for a scalable optical switch. However, a push-only switch has

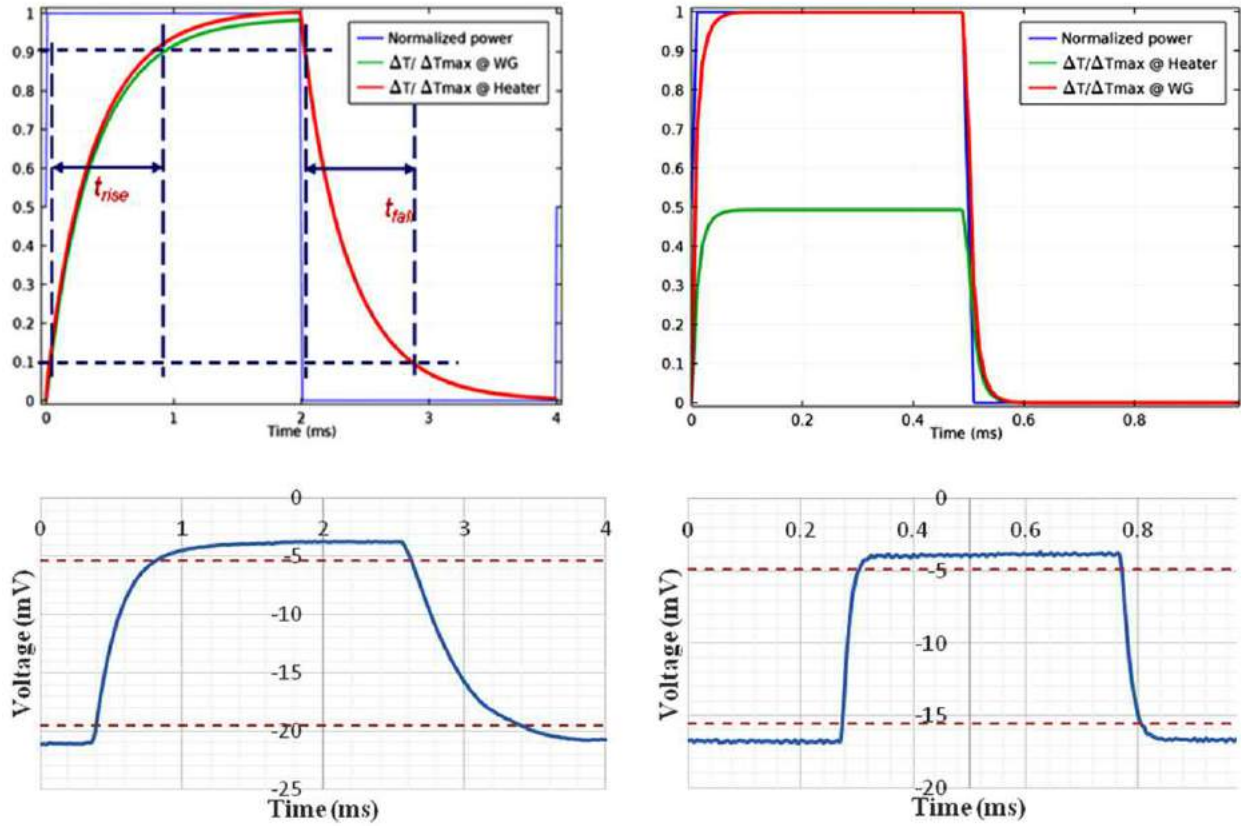


Fig. 16 Switching time of thermo-optic Mach Zehnder silicon photonic optical switch cell: (a) modeled and (b) measured switching time response, with (left) and without (right) thermal undercut. After Celo, D., Goodwill, D.J., Jiang, J., *et.al.*, 2016. Thermo-optic silicon photonics with low power and extreme resilience to over-drive. IEEE Optical Interconnects Conference (OI), pp.26–27.

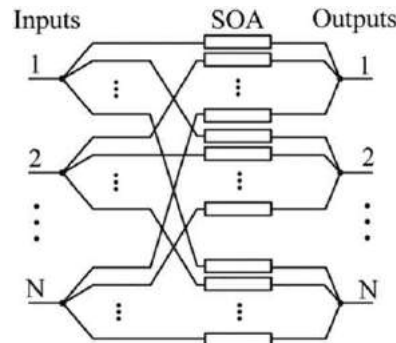


Fig. 17 Semiconductor optical amplifiers (SOA) used in a split-and-select optical switch.

to be driven too hard into saturation, and the absorption loss is 2 dB per switch cell with extinction ratio of 18 dB or less, and therefore push-only carrier injection switches are not effective for scalable optical switches.

Semiconductor Optical Amplifier Switch

Semiconductor optical amplifiers (SOAs) are used in optical switches as an *ON-OFF* switch by varying the bias voltage. SOAs provide nanosecond switching times and broadband operation, have relatively high gain (15–28 dB) with saturation power in the range of +8 to +15 dBm, are small (100's micrometers to few millimeters in length), and can be integrated with other semiconductor and optical components. The typical architecture of an SOA optical switch is a split-and-block topology, integrating a bank of passive splitters, a bank of SOA, and a bank of passive combiners, as shown in Fig. 17. Larger port counts may use several stages. SOA optical switches demonstrated to date have up to 32×32 ports.

An SOA gate can be switched from a high-gain state to a high-loss state within a nanosecond, by changing the drive current from an electronic driver. The difference between large amplification in the *ON*-state and large absorption in the *OFF*-state provides very high extinction ratio, which is needed for the split-and-select architecture.

SOA optical switches have some performance weaknesses. They have relatively high power consumption, of 10's to 100's mW per SOA in the *ON* state, although the *OFF* state power consumption may be small or zero. Many SOAs are polarization dependent, thus requiring polarization diversity lightpaths or single-polarization networks. SOA are best used to carry a single wavelength, because they have large nonlinearity that causes mixing between signals on different wavelengths.

In an alternative use, SOAs serve as optical power amplifiers to compensate for optical losses in other types of planar lightwave optical switches. III-V semiconductor gain elements can be heterogeneously integrated onto a silicon photonic circuit, with the Si and InP chips independently optimized for high performance, reliability and yield.

Waveguide-MEMS Optical Switch

A waveguide-MEMS switch is a planar lightwave optical switch, using waveguides that are moved by micro-mechanical actuators. Each moving waveguide is one half of an adiabatic waveguide coupler. Each switch cell comprises a pair of adiabatic couplers joined by a 90° bend (Fig. 18(a)). In the *OFF* state, the coupler waveguides are moved away from the bus waveguides, and there is no light coupled between the bus waveguides. Upon application of a voltage to the actuator electrodes, the two movable waveguide couplers are brought in close vicinity of the bus waveguide, creating optical coupling between the fixed and the movable waveguides. In the *ON* state, the optical signal is thus routed from the input port to the desired drop port. The switching time is 10's of microseconds.

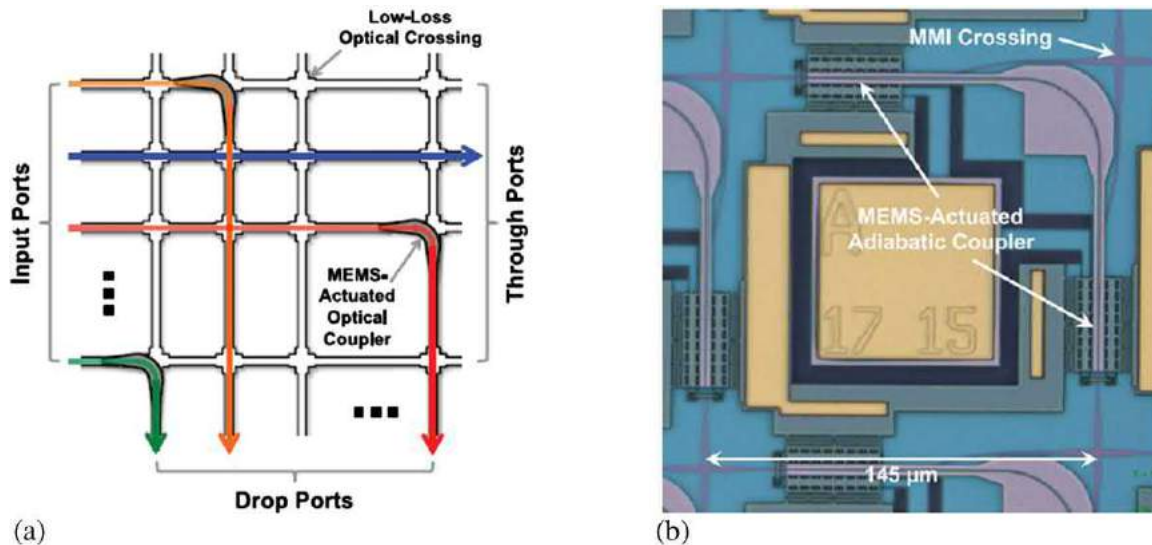


Fig. 18 Optical switch with MEMS-actuated adiabatic waveguide couplers (a) schematic, and (b) microscope image of a representative unit cell. Reproduced from Seok, T.J., Quack, N., Han, S., Muller, R.S., Wu, M.C., 2016. Highly scalable digital silicon photonic MEMS switches. *Journal of Lightwave Technology* 34(2), 365–371.

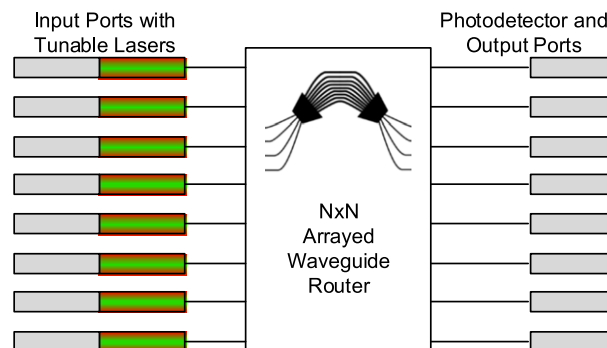


Fig. 19 Tunable laser optical switch – schematic representation of the system including tunable lasers and $N \times N$ AWGR (for $N = 4$).

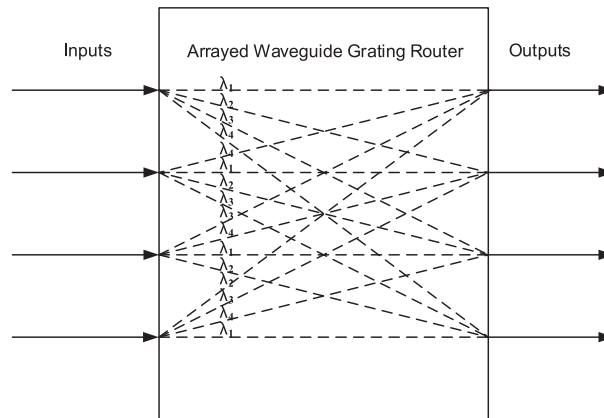


Fig. 20 Details of arrayed waveguide grating router (AWGR) behavior, as used in a tunable laser optical switch. An AWGR is a passive, static planar lightwave circuit, typically made in silica.

A 50×50 optical switch has been demonstrated, using an optical crossbar. The insertion loss compares favorably to other technologies.

Tunable Laser Optical Switch Technologies

A tunable laser optical switch comprises a rapidly-tuned laser at the source, and passive routing using an arrayed waveguide grating routing (AWGR). The switch is illustrated in Fig. 19 and the behavior of the AWGR is illustrated in Fig. 20. To select a destination, the wavelength of a laser is tuned so that the light comes out from the desired output port of the AWGR.

The actuator is in the transmitter – in this manner, the tunable laser optical switch is fundamentally different to all other types of optical switch that are described in this article, as in all other cases the actuator is in the core switch.

An AWGR is similar to the more widely-used optical demultiplexing AWG, where an optical input fiber carries light of multiple wavelengths. The light power is split, passes through an array of curved waveguides of unequal length, and then recombines in a star coupler. The star coupler has multiple output waveguides, and the phase relationship of the curved waveguides is constructed so that light from each wavelength interferes constructively at only one of the output waveguides. An AWGR is similar, except that the input coupler has multiple input fibers, and the geometry is organized so that the mapping of input wavelength at output ports is cyclical. A typical AWGR has -25 dB adjacent channel crosstalk, 7 dB insertion loss, and 3 dB uniformity.

Tunable laser optical circuit switches have been used experimentally for core router growth and for switch core growth (see earlier section for definition of these applications). In both cases, the switching time is limited by the time that it takes to accurately switch the wavelength of the laser. Microsecond switching time is possible, but control is very challenging at nanosecond switch time. Scalability is limited to 40 ports, and AWGRs cannot be cascaded to form large port count systems. Also, AWGRs require DWDM transmitters, which are not commonly used within data centers or high performance computers as they are expensive and consume a large power.

Further Reading

- Ben Yoo, S.J., 2006. Optical packet and burst switching technologies for the future photonic internet. *Journal of Lightwave Technology* 24 (12), pp. 26–27.
- Celo, D., Goodwill, D.J., Jiang, J., *et al.*, 2016. Thermo-optic silicon photonics with low power and extreme resilience to over-drive. *IEEE Optical Interconnects Conference (OI)*, pp. 26–27.
- Dharmaweera, M.N., Parthiban, R., Sekercioglu, Y.A., 2015. Toward a power-efficient backbone network: The state of research. *IEEE Communication Surveys & Tutorials* 17 (1), 198–227.
- Dumais, P., Goodwill, D., Kiaei, M., *et al.*, 2017. Silicon Photonic Switch Subsystem With 900 Monolithically Integrated Calibration Photodiodes and 64-Fiber Package, *OFC*.
- Dupuis, N., Ryljakov, A.V., Schow, C.L., *et al.*, 2016. Nanosecond-scale Mach-Zehnder-based CMOS photonic switch fabrics. *Journal of Lightwave Technology* 35 (4), 615–623.
- Han, S., Seok, T.J., Quack, N., 2015. Large-scale silicon photonic switches with movable directional couplers. *Optica* 2 (4), 370.
- Hinton, H.S., 1993. An introduction to photonic switching fabrics. In: Lucky, R.W. (Ed.), *Applications of Communications Theory*. New Jersey: Bellcore, Red Bank.
- Kachris, C., Tomkos, I., Bergman, K., *et al.*, 2013. *Optical Interconnects for Future Data Center Networks*. New York, NY: Springer-Verlag.
- Stabile, R., Albores-Mejia, A., Rohit, A., Williams, K.A., 2016. Integrated optical switch matrices for packet data networks. *Microsystems & Nano-Engineering* 2, 15042.
- Strasser, T.A., Wagener, J.L., 2010. Wavelength-selective switches for ROADMs applications. *IEEE Journal of Selected Topics in Quantum Electronics* 16 (5), 1150–1157.
- Yeow, T.W., Law, K.L.E., Goldenberg, A., 2001. MEMS optical switches. *IEEE Communications Magazine* 39 (11), 158–163.
- Zhang, Z., You, Z., Chu, D., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. *Light: Science & Applications* 3, 1–10.

Indium Phosphide Photonic Integrated Circuits

Yuliya Akulova, Lumentum, Milpitas, CA, United States

© 2018 Elsevier Ltd. All rights reserved.

Why Do We Need Photonic Integration?

To overcome fiber transmission impairments at higher modulation rates – and to overcome speed limitations of optical and electrical components – optical transceiver architectures increasingly rely on transmission components with multiple functional elements (lasers, modulators, power splitters/combiners and wavelength multiplexers). For example, 100, 200, and 400 Gb/s optical transceiver architectures for the local area network applications combine 25 Gbaud on-off keying (OOK) or four-level pulse amplitude modulation (PAM4) with multiplexing four or even eight wavelengths into one single-mode fiber (LAN/MAN Standards Committee of the IEEE Computer Society, 2010). In the metro, and long-haul space, demand for increased line-card density and reduced power dissipation is being supported by 10 Gb/s SFP + transceivers, either operating at a fixed wavelength or tunable over the C-band. This application utilizes traditional OOK modulation format and requires an externally modulated laser to meet the requirements for dispersion tolerance. In addition, starting from 40 Gb/s, long-haul and metropolitan area telecom applications adopted advanced modulation formats, including polarization multiplexing and multilevel quadrature amplitude modulation (Winzer, 2012). A polarization-multiplexed advanced-modulation-format transmitter requires a laser and two quadrature modulators, typically realized using a nested Mach-Zehnder (MZ) architecture. Widespread deployment of the coherent technology in the datacenter interconnects and metro space requires development of new photonic components that are able to support cost-effective, compact, high efficiency pluggable coherent transceivers.

The requirements for increased optical transmission speed and complexity combined with stringent constraints on cost, size, and power dissipation can be met only by implementing photonic integration to replace several packaged components with one.

What Is Photonic Integration?

Generally speaking, photonic integration refers to any technology that combines two or more functional optical elements to replace several separately packaged components with one. Photonic integration can be categorized into several types: Monolithic integration combines multiple active and passive optical elements fabricated on a single semiconductor chip to form a Photonic Integrated Circuit (PIC). Heterogeneous integration refers to a technology that uses dissimilar materials to provide full set of optical functions (e.g., heterogeneous integration of InP-based epitaxial material with Si photonics enables optical gain function). Finally, hybrid integration co-packages several active optical components that are optically coupled using micro-optics or planar lightwave circuits. In practice, a combination of hybrid and monolithic integration is frequently used. As the title of this article suggests, we will focus on monolithic PICs technology in InP.

Monolithic integration provides a drastic reduction in system footprint, inter-element optical coupling losses, power dissipation, and assembly and test cost. To illustrate the reduction in size, Fig. 1 compares optical components that are required to realize transmission function using discrete laser and modulator with the same functions integrated on an InP PIC. In addition, monolithic integration frequently results in improved reliability since no relative shifts of different optical elements are possible.

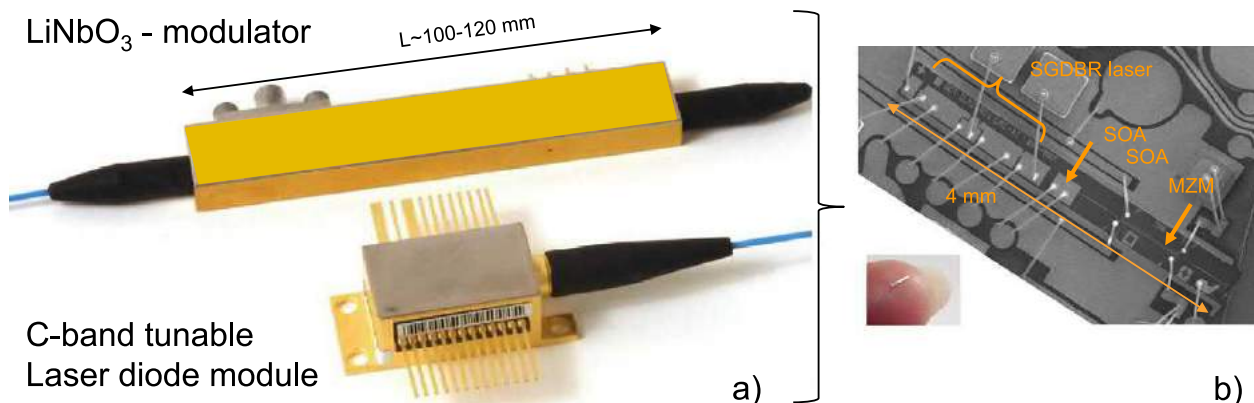


Fig. 1 Optical components required to realize 10 Gb/s C-band tunable transmitter: (a) C-band tunable laser diode module and Lithium Niobate modulator; (b) InP photonic integrated circuit with the same functionality – C-band tunable laser integrated with Semiconductor Optical Amplifier and Mach-Zehnder modulator.

Functional Requirements

The primary functions required to implement high-functionality PICs include (1) light-generation and amplification, (2) high-speed phase and/or amplitude modulation, (3) photodetection, and (4) optical routing. Optical routing functions may include only passive elements such as splitters, combiners, etc. and variety of functional elements that rely on the active refractive index tuning to adjust the optical phase. Fortunately, fundamental material properties of indium phosphide (InP) and related quaternary indium gallium arsenide phosphide (InGaAsP) and indium aluminum gallium arsenide (InAlGaAs) compounds enable all required functionalities.

In most cases, all functional sections of an InP PIC are *p-i-n* heterostructure diodes, with i-waveguide core containing bulk quaternary material or multiquantum well (MQW) layers. In general, different materials have to be employed to optimize the performance for each of the functional elements as schematically illustrated in Fig. 2. Specifically, active regions of a high performance laser typically require several relatively narrow compressively strained QWs with the fundamental transition energy close to the target photon energy. The laser MQW active waveguide is typically surrounded by a separate confinement heterostructure to maximize overlap between the optical mode and injected carriers. The upper and lower InP cladding regions are relatively heavily doped to minimize resistive heating. Electroabsorption (EA) and electro-optic MZ modulators in InP material systems fundamentally rely on the quantum-confined Stark effect, which produces red shift of the MQW absorption edge and change of the refractive index when reverse bias voltage is applied to p-i-n hetero-structure. In contrast to the laser, active regions of high efficiency, high speed EA and MZ modulators are blue shifted relative to the operating wavelength and typically require wider QWs to increase the efficiency of the quantum confined Stark effect. For EA modulators the difference between operating wavelength and modulator absorption edge wavelength, known as detuning, is chosen such that the absorption is negligible in the absence of reverse bias but increases rapidly as reverse bias is applied. MZ modulators are designed with larger detuning to achieve efficient index tuning without introducing excess loss. It is always desirable to be able to select the number of the QWs in a modulator active material for a specific performance targets: modulation bandwidth and drive voltage. The optical routing sections of the PIC require material with a bandgap larger than operating photon energy and low doping to minimize propagation loss. They can employ bulk or MQW material. The refractive index tuning can be accomplished through carrier injection mechanism by forward biasing the p-i-n diodes, voltage induced refractive index change under the reverse bias as outlined above, or thermal index tuning mechanism using local resistive heaters. Finally, light detection components require materials with a bandgap smaller than the photon energy to ensure efficient optical absorption.

To ensure high performance of the PIC based devices an integration technology has to provide flexibility in design optimization for different functional elements, efficient optical coupling between them and absence of parasitic reflections that can detrimentally affect laser wavelength stability, noise, and chirp. It is also desirable to minimize the transition regions between different functional sections. Ideally the spacing between different functional sections is selected based on the required electrical and thermal isolation and not constrained by the dimensions of transitions.

InP Integration Technologies

Active-Passive Axial Waveguide Integration

The compelling advantages and possibility of monolithic photonic integration in InP have been recognized from the early 1980s, and multiple approaches to integrate regions with different absorption/gain properties together along a single waveguide have been developed and are widely used in InP PIC fabrication (Coldren *et al.*, 2013).

Fig. 3 illustrates five most common techniques employed for monolithic integration of regions with different absorption/gain properties together along a single waveguide. For Offset Quantum wells Fig. 3(a) and twin waveguide Fig. 3(b) integration technologies different optical functions are separated into vertically displaced waveguides and the integration is realized by selective removal of the unwanted upper waveguides. Selective Area Growth (SAG) technology Fig. 3(c) relies on variation of the local material growth rates and hence the transition energy of the MQWs by changing the width of the pre-patterned dielectric masks before material growth. Quantum well intermixing technology Fig. 3(d) blue shifts the MQW region absorption edge by

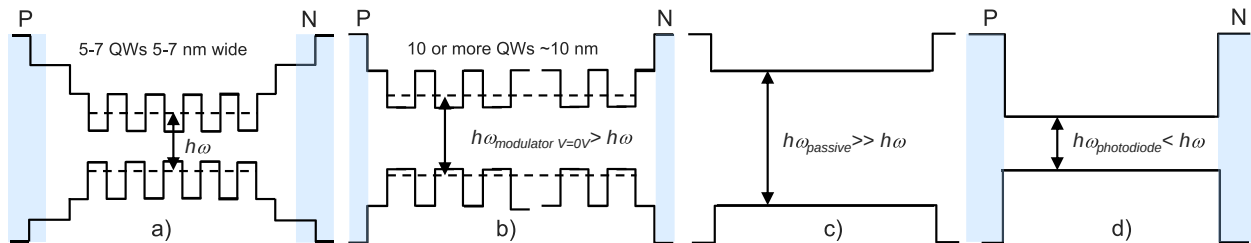


Fig. 2 Schematic band diagrams of the different functional sections of an InP PIC: (a) laser, (b) electro-absorption or Mach-Zehnder modulator, (c) passive waveguide, (d) photodiode.

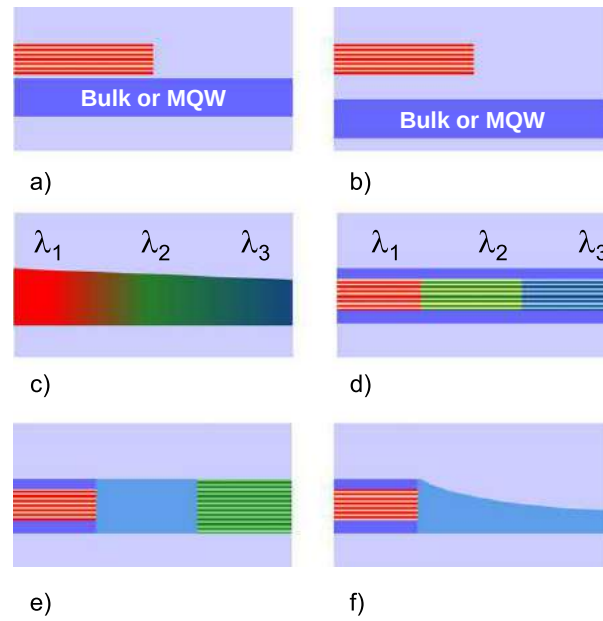


Fig. 3 Active/passive waveguide integration approaches: (a) Offset Quantum wells, (b) twin waveguide, (c) Selective Area Growth, (d) Quantum well intermixing, (e) Material regrowth, (f) Material regrowth combined with selective area growth.

impurity or point defect assisted atomic inter-diffusion between the wells and barriers at a high annealing temperature. In the Material regrowth technology [Fig. 3\(e\)](#), different optical functions are created by sequential material removal and growth in different regions of the wafer. Several of the integration technologies can be further combined to achieve desired functionality. For example, as illustrated in [Fig. 3\(f\)](#) material regrowth can be combined with SAG by tailoring the dielectric mask shape and thus local waveguide thickness to match laser or modulator waveguide on the left and thin core waveguide for a spot-size converter on the right.

The ‘offset quantum-well’ technology and its variance, the ‘dual-quantum-well’ structure is the most simple integration approach. In this technology the waveguide is subdivided into a wider band-gap section at the bottom and a narrower bandgap MQW stack to be used as a gain section for a laser or an SOA at the top. The bottom section of the waveguide can be made of bulk quaternary alloy or contain a MQW stack embedded in the bulk waveguide. These closely spaced sections are grown in one epitaxial run and form single waveguide that supports only fundamental mode. The gain MQW material is selectively removed in the regions intended for modulators and passive waveguides and, if required, grating is etched into the passive waveguides. A single blanket regrowth of the InP cladding and InGaAs contact layer completes the manufacturing process. Offset quantum-well technology results in relatively low gain confinement factor that limits performance of lasers and SOAs with exception of booster amplifiers where low confinement factor is desirable for achieving high saturation power. However, this technology is very attractive due to its simplicity and manufacturability and has been successfully employed to demonstrate a variety of high performance PICs including widely tunable lasers integrated with SOAs, EA and MZ modulators (see reference [Coldren et al. \(2011\)](#) and references therein).

Another approach for monolithic photonic integration was developed based on twin-guide (TG) technology. As illustrated in [Fig. 3\(b\)](#), a TG structure consists of active and/or passive devices formed in separate, vertically displaced waveguides. In difference from the offset quantum well structures, the waveguide layers in TG structures are spaced further apart. Such waveguide structure can support two optical modes (even and odd). Enhanced version of TG integration technology uses asymmetric twin guides (ATG) structure designed with different effective indexes of the active and passive waveguides. Light is transferred between the waveguides via lateral, adiabatically tapered mode transformers, allowing different optical functions to be realized in the different waveguides. The design of ATG structures and adiabatic tapers were described in details in [Xia et al. \(2005\)](#). The optimum design of the taper (shape and length) depends on the asymmetry of the ATG and the authors’ analysis shows that using the optimal taper shape, a transfer efficiency of 90% can be achieved with the taper length of 290 μm . However, regardless of the fabrication simplicity, AGT technology carries significant drawbacks. These include long tapers, ~ 0.5 dB coupling loss between neighboring sections, and absorption in the gain tapers if left unpumped. Pumping can be extended to include the tapers but even then the gain and saturation power of the SOAs can be significantly compromised in the output taper. Alternative approach is to use QWI to blue-shift the absorption edge of the active material in the taper. Overview of the PICs that have been realized using ATG technology is provided in reference ([Menon et al., 2005](#)). To list a few, these include EMLs, SOA/p-i-n detectors, integrated TG lasers, and integrated arrayed waveguide grating/p-i-n detector.

Quantum well intermixing (QWI) is an effective post-growth technique for increasing the bandgap energy of QWs in the desired locations on the wafer thus integrating laser gain with modulator and passive sections. QWI techniques has been used for

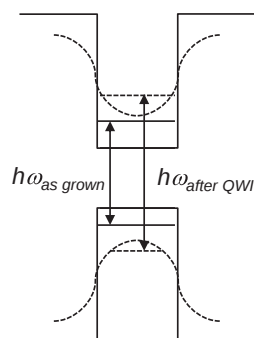


Fig. 4 Schematic representation of a QW potential profile as grown (solid) and after inter-diffusion of atomic species at QW/barrier heterointerfaces (dash).

the fabrication of various PICs in InGaAsP and InAlGaAs MQW materials and relies on inter-diffusion of atomic species at QW/barrier hetero-interfaces at elevated temperatures. Intermixing of the as-grown rectangular wells and barriers results in a graded potential profile and blue shift of the transition energy for the fundamental confined state for electrons and holes as schematically illustrated in Fig. 4. The inter-diffusion process is enhanced dramatically in the presence of point defects or vacancies in the interface region. This enhancement of the intermixing is used to provide spatial selectivity of the band gap shift that is required for monolithic integration. Several intermixing techniques including impurity induced disordering, impurity-free vacancy enhanced intermixing, and ion-implantation enhanced intermixing have been developed.

We will describe the latter two techniques in more details. In the vacancy enhanced QWI process the areas of the wafer that require wider bandgap are capped with a dielectric mask and the wafer is subjected to Rapid Thermal Annealing (RTA) typically at the temperature range of 700–800°C for approximately 30 s. The out-diffusion of group III elements at the semiconductor/dielectric interface create vacancies that rapidly propagate into QW region enhancing the rate of QW intermixing. Vacancy induced QW intermixing has been employed to achieve large bandgap shifts (> 100 nm) in InGaAsP and AlGaAs MQW structures.

The most complex and high performance PICs were realized using implant-enhanced QWI process. This method has been employed for InGaAsP and InAlGaAs materials and produces good spatial resolution and controllable band-gap shifts using anneal time, temperature, and implant dose. Furthermore, any number of QW band edges can be achieved on the same wafer using selective removal of the catalyst.

We will describe implant-enhanced QWI process in more details using widely tunable sampled-grating DBR (SG-DBR) laser-EA modulator PIC reported in Skogen *et al.* (2005) as an example. In the first step, passive and modulator device regions are defined by selectively implanting phosphorous (P+) to introduce point defects into a sacrificial InP implant buffer layer grown above the MQW active region. Next, an RTA step is performed to shift the MQW band edge to that desired in the EAM regions (~ 35 nm). Once the EAM band edge is reached, the point defects are removed in the EAM region by selective removal of the InP implant buffer layer. The annealing is then continued until the desired band gap shift is reached for the tuning sections of the laser (~ 95 nm). Finally, blanket regrowth of the upper cladding and contact layer completes the formation of the waveguide in vertical direction.

Similarly to offset QW integration technology, implementation of PICs based on QWI requires only two epitaxial growth steps. However, QWI technology offers several significant advantages that include ability to place active QWs in the center of the waveguide and therefore maximize gain confinement factor, absence of multiple hetero-barriers and low doped material under the gain sections, ability to use band-gap shifted QWs in the carrier injection tuned phase sections therefor achieving higher index tuning efficiency due to band-filling effect in the step-function density of states. Finally, since QWI does not change the average material composition in the optical waveguide but only slightly changes the compositional profile, the index discontinuity and mode shape miss match at the interface between adjacent sections are small. This results in very low optical coupling loss and eliminates parasitic reflections between adjacent functional sections.

The drawbacks of the QWI techniques include shared number of QWs throughout the PIC and requirement for careful calibration of the intermediate band-gap shifts. The intermixing must also be done in the absence of other rapidly diffusing species in the wafer, such as zinc. Hence, regrowth of the p-doped cladding has to be done after the intermixing step resulting in somewhat poor controllability of the doping profile just above the active regions of the laser and modulator.

SAG has been employed in photonic integration using Metal-Organic Chemical Vapor Deposition (MOCVD) for InGaAsP and AlGaInAs bulk and MQW materials. With SAG technology different band-gap materials can be defined simultaneously using single epitaxial growth with patterned dielectric mask. In the absence of deposition on the mask, extra material is available over the masked zone resulting in the lateral concentration gradient. The gradient leads to extra lateral diffusion of species around the mask and to a local enhancement of the growth rate in the vicinity of the mask. In addition, since the diffusion coefficient in the vapor phase and sticking rate constant to the wafer surface are different for different species, the growth enhancement is accompanied by composition, band-gap, and strain variation as shown schematically in Fig. 5. The region between the mask stripes is used for the laser section and has thicker MQW stack, narrower band-gap, and slight compressive strain. Note, that in the case of MQWs both the composition and QW width are contributing to the red shift of the effective band-gap. Band-gap shifts up to 150 nm for MQW

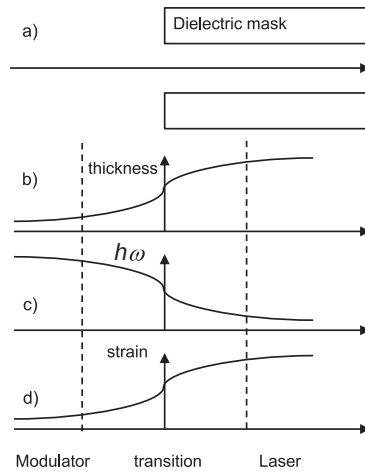


Fig. 5 Selective area growth: (a) top view of the mask layout and illustration of variation of waveguide thickness (b), material band-gap (c), and strain (d).

material and growth enhancement factor of ~ 2.5 can be achieved with appropriate design of the mask pattern. Examples of PICs fabricated using SAG include DFB-EA, tunable DBR-EA, wavelength selectable DFB arrays, and spot-size converters.

Three dimensional vapor phase diffusion models have been successfully developed and can accurately describe spatial variation of the thickness, composition, bandgap, and biaxial strain of the materials grown with different dielectric mask shapes. The models were validated by employing optical interferometer microscopy, high-resolution micro-photoluminescence, and micro X-ray. The models allow for constructive engineering of the waveguide thickness and composition through the manipulation of dielectric mask shape and also can predict undesired spatial variation that can occur due to neighboring mask cells in the case of high mask density.

The best applications for SAG technology include multi-wavelength laser arrays and spot size converters. Multi-wavelength laser arrays require adjustment of the MQW band-gap of 40 and 60 nm to cover C-band and CWDM O-band grid, respectively. Such band-gap engineering is well within SAG capability. In the case of a spot size converter both waveguide thickness and refractive index can be tapered to produce lossless interface with the rest of the PIC and the desired mode field diameter at the SSC output. In the case of a laser-electroabsorption modulator integration SAG can produce sufficient band-gap shift but limits freedom to optimize the material for each functional section. The design has to rely on the same number of QWs and doping throughout the PIC. In addition, SAG produces wider QWs for the laser and narrower QWs for the modulator imposing severe trade-off between laser and modulator efficiency and adds relatively long ($\sim 100 \mu\text{m}$) transition regions. Such design trade-offs have proven to be acceptable for 2.5 and 10 Gb/s components. However, as the industry moves to 100 and 400 Gb/s transmission, more stringent requirements imposed on the modulation speed and PIC output power will require independent optimization of each functional section and therefore favor regrowth integration.

Among various integration technologies, regrowth integration is the only technology that enables flexibility in the design optimization of each PIC section with minimum transition length. In the regrowth integration technique, the laser gain layer stack is typically grown first on a planar substrate and then protected with a dielectric mask where it is required in the PIC. The exposed gain layers are then etched away, and the layers for the next functional section are grown. This process can be repeated several times to produce optimized layer stack for each PIC section. The primary advantage of the regrowth integration technology is that the layer composition (bulk or MQW) and doping can be changed abruptly, and therefore optimum design can be independently selected for each of the functional sections. The coupling loss between the neighboring sections are primarily determined by vertical alignment of the waveguides which can be controlled to very high accuracy by using selective etch. Even with the perfect vertical waveguide alignment some band-gap engineering is usually required to improve mode overlap and index matching at the transition. Well-designed regrowth transition can consistently produce less than 0.1 dB loss.

InP Waveguides

After the active-passive waveguides are defined along the light propagation direction lateral waveguide geometry has to be chosen based on the desired functionality of the PIC elements. **Fig. 6** illustrates several choices that are available in InP technology. They are arranged from highest to lowest index contrast as indicated in the figure. The deep- and shallow-ridge waveguides are formed by etching after all epitaxial steps are completed. The deep-ridge is fabricated using dry etch technique and due to high index contrast is suitable for bends with small radius of curvature. This type of waveguide is most frequently used as a routing guide for large scale PICs when multiple 90 degree bends, U-turns, and waveguide crossings are required to minimize the chip footprint. The deep-ridge waveguide is also the technology of choice for high-speed Mach-Zehnder modulators since in this geometry the active

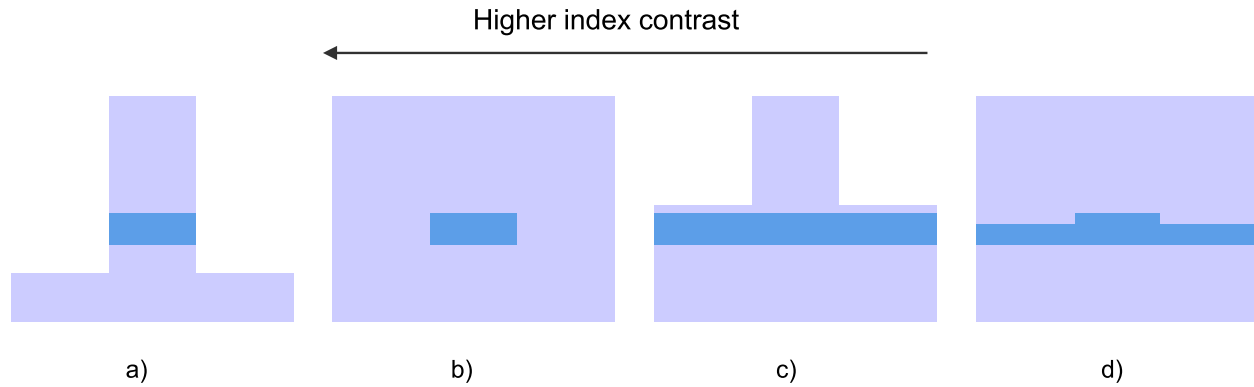


Fig. 6 Lateral waveguide cross-sections available in InP technology: (a) deep ridge, (b) buried channel, (c) shallow or surface ridge, (d) buried rib.

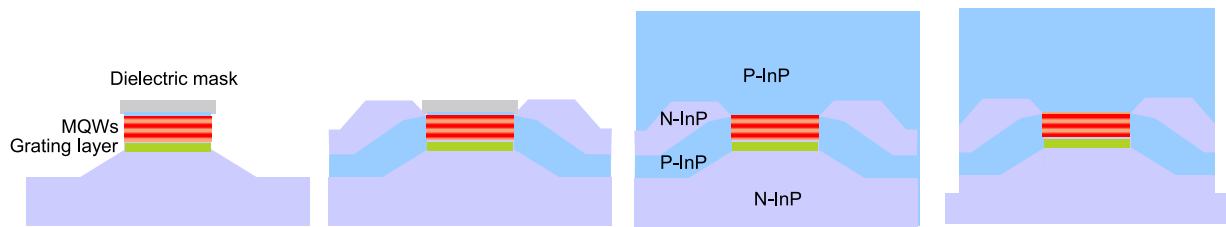


Fig. 7 Fabrication flow for planar buried hetero-structure laser.

region capacitance is determined only by the area of the waveguide without parasitic contribution of the electric field fringing effect in the surrounding material.

The shallow ridge waveguide (also called surface ridge) is fabricated through a combination of dry and selective wet etching. Selective wet etch produces very smooth sidewalls with the ridge profile depended on the crystallographic orientation. Due to the selectivity of the wet etch, the depth of the shallow ridge waveguide is controlled precisely. This type of lateral waveguide is a popular choice for many functional section of InP PICs. It is especially well suited for fabrication of CW lasers (FP, DFB, and DBRs), SOAs, and electro-absorption modulators. In contrast to MZ modulators, EAMs are usually short and very high modulation bandwidth can be attained even with larger junction capacitance of the shallow ridge structure.

The buried channel and buried rib waveguides must be defined prior to epitaxial regrowth. If such waveguides are intended for low loss routing they can be capped with thick undoped layer of InP. Buried channel waveguide is also used for manufacturing of high performance CW and directly modulated lasers (DMLs) and laser arrays by using PN or PNIN regrowth to block carrier flow around the active waveguide. **Fig. 7** illustrates the fabrication sequence for a planar buried heterostructure laser. The process starts with the growth of the grating layer followed by the grating etch and regrowth of a spacer and MQW laser active. The lateral waveguide is formed using a combination of dry and wet etch and several layers of InP with different doping are re-grown with the mask still in place. Finally, the mask is removed and the top cladding and contact layers are grown. PN blocking structure is effective for CW lasers but produces very high parasitic capacitance, so Fe-doped InP i-region sandwiched between n-doped InP layers forming PNIN blocking structure is used as an alternative for manufacturing of high-speed DMLs or electro-absorption modulated lasers (EMLs). The advantages of this technology include low threshold current and high differential gain for the lasers due to reduced active region volume and the fact that carriers are confined to the center of the optical mode, relatively symmetric output beam, and good thermal properties. However, this technology adds critical regrowth step with regrowth interfaces positioned along the whole cavity length between MQW active and the blocking. Since the carriers can diffuse rapidly in the QW layers, non-radiative recombination at these interfaces can be detrimental to laser performance and reliability. Excessive parasitic capacitance can limit modulation bandwidth even if PNIN blocking structure is used due to high dielectric constant of InP. Parasitic capacitance can be somewhat reduced by etching trenches to narrow the width of the blocking structure as shown in **Fig. 7**. Finally, it is difficult to engineer low loss, low reflection couplers between a buried heterostructure laser and deep-ridge routing waveguides and high-speed MZ modulators if such laser is used a part of a PIC. This is due to the significant mismatch in the optical mode sizes and effective refractive indexes and the fact that such transitions have to rely on the precise lithography since the mask used to fabricate buried waveguide section is removed prior to the regrowth of the cladding.

Transitions between different waveguide types are essential elements of the integration technology. Transition between deep- and shallow ridge waveguide is the most frequently used transition in large scale PICs that incorporate lasers and SOAs together with high speed MZ modulators and high functionality routing sections. Fortunately, since both types of waveguides are formed after regrowth of the top cladding, this transition can be fabricated using the same dielectric mask to define the waveguide along

the light propagation direction. In this self-aligned process, the coupling efficiency is not impacted by the accuracy of the lithographic alignment and, as described in reference (Coldren *et al.*, 2011), the waveguide mask can be tapered to ensure better overlap between the modes in the shallow and deep ridge sections. Transitions between buried ridge and other ridge types have to rely on an accurate alignment of the surface ridge waveguide mask to the underlying buried waveguide. However, even in this case coupling efficiency in excess of 95% with large manufacturing tolerance ($\pm 0.3 \mu\text{m}$) can be obtained by tapering both waveguides at the junction.

Low loss InP waveguides

Losses in InP photonic waveguides are dominated by scattering loss and absorption provided that the material composition is chosen such that the contribution of the direct band-gap absorption is negligible. Smooth sidewalls are essential for achieving low loss single mode waveguides. As process control improves, the p-doped cladding becomes the most important contribution to loss through inter-valence band absorption. The n-type free carrier absorption is dominated by intra-valley transitions and requires interaction with phonon or ionized impurity scattering to provide conservation of momentum and thus much less significant than p-type absorption. This analysis implies that low loss waveguides on an InP PIC have to be fabricated with undoped cladding or p-doping has to be setback far enough to have negligible overlap with the optical mode. Optical propagation loss below 0.3 dB/cm has been demonstrated for shallow and deep ridge waveguides with doping setback. Excess loss below 0.1 dB per 90 degree turn with 100 μm radius of curvature can be readily achievable in deep ridge waveguides.

Since InP PICs typically include laser and SOA functions, special care has to be taken to suppress parasitic reflections that can occur due to abrupt changes in optical cross-sectional areas or/and effective index of the different components. Such reflections can detrimentally affect noise properties of the laser, introduce adiabatic chirp if reflection occurs after a modulator, and create gain ripples in the SOAs. Tilting the transitions with refractive index discontinuities and introducing tapered features at the transitions from shallow to deep waveguides and between single and multimode regions enables to maintain sufficiently low reflection level without introducing excess loss. Detailed description of splitters, polarization devices, microbends, and de/multiplexers developed for high-functionality InP PICs can be found in Williams *et al.* (2015).

Finally, ion implantation can be employed for electrical isolation between different functional sections of the chip. Hydrogen implant is typically used for passivation of p-doped material. Helium implant can be used for electrical isolation of n-doped material. In both cases the implantation dose and energy need to be tailored to provide the desired isolation without introducing excess optical loss.

We encourage the readers to follow up on the references provided above to learn more about high performance InP PICs developed in various integration platforms. In the last part of this article we will discuss design and performance for several InP PICs fabricated using regrowth integration. However, it is important to understand that the choices that PIC designers have to make are not limited to the integration technology but also include appropriate selection of the architecture (e.g., EA vs. MZ modulator) and additional functional elements that can be included in the PIC to further alleviate design trade-offs and increase fabrication tolerances, enhance functionality, and achieve sufficient margin relative to all performance specifications to enable high manufacturing yield and ultimately low cost. As we describe the PICs, we will point out why certain integration technology, device architecture, and features were adopted for each specific application.

Integrated (C-Band Tunable) Laser Mach-Zehnder Modulator (ILMZ) PIC

Development of ILMZ PIC drastically reduced footprint and power consumption of widely tunable optical transmitters and enabled transition of 10 Gb/s DWDM metro and LH transmission architecture from discrete implementations on a linecard or in 300 pin transponders into the realm of small form factor pluggable modules such as T-XFP and SFP+.

As illustrated in Fig. 8(a), ILMZ PIC consists of a four-section SG-DBR laser, an SOA, and a MZ modulator. The SG-DBR laser is an efficient widely tunable laser source that can be easily integrated with other optical components. This laser was pioneered at UCSB and was used in a variety of high functionality PICs. An SG-DBR laser is a PIC in itself and consists of a MQW gain section

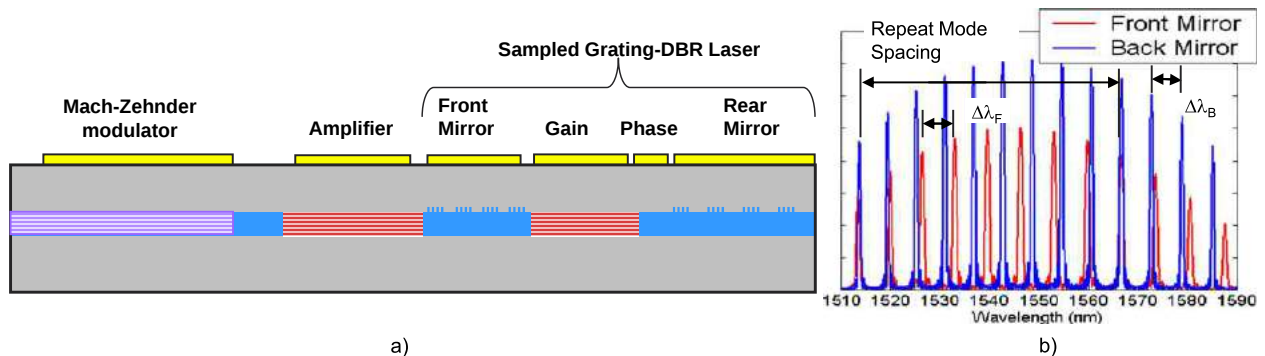


Fig. 8 (a) Schematic cross-section of ILMZ PIC. (b) reflectivity spectra of an SGDBR laser.

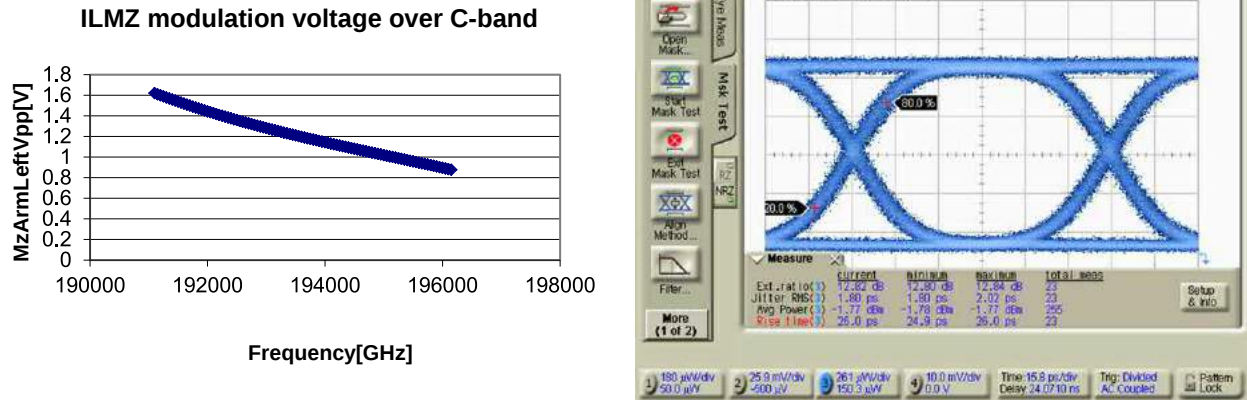


Fig. 9 Performance characteristics of ILMZ PIC based transmitter: (a) modulation voltage vs wavelength across C-band, (b) eye diagram at 10.7 Gb/s.

positioned between sampled-grating DBR mirrors. SG-DBR laser utilizes Vernier effect to achieve wavelength tuning in excess of 40 nm with only moderate index tuning in the mirrors. The detailed description of an SG-DBR laser is given in Coldren *et al.* (2004). In brief, by imposing additional periodicity on the holographically defined grating the reflectivity spectra of the mirrors are transformed into combs of reflectivity peaks as shown in Fig. 8(b). The spacing between adjacent peaks is inversely proportional to the grating sampling period. Selecting different sampling periods for the front and back mirrors ensure that only one set of reflectivity peaks can coincide for a particular set of mirror tuning currents resulting in lasing at the specified wavelength. The laser is tuned by adjusting currents in the mirrors to bring the desired reflectivity peaks into alignment. The phase section is used for alignment of the cavity mode to the peak of mirrors' reflectivity. Continuous tuning can be achieved by tuning both mirrors simultaneously. The specific design of SG-DBR laser implemented in ILMZ PIC uses carrier injection index tuning mechanism. Carrier injection is very efficient and fast refractive index tuning mechanism but it introduces cavity losses caused by free carrier absorption. Therefore it is highly desirable to integrate an SOA as an independent power adjustment knob. The integrated SOA compensates modulator loss and cavity losses caused by free carrier absorption in the tuning sections and allows wavelength independent power leveling, beam blanking during wavelength switching, and variable optical attenuator functionality.

Finally, MZ modulator is selected since the device has to operate over full C-band with high extinction ratio and good control of the chirp. Chirp of a differentially driven InP MZ is determined by the build in phase shift and split ratio and thus can be optimized (negative or zero) for the target application.

The PIC is fabricated using regrowth technology for lossless integration of the laser and SOA MQW active regions with the bulk quaternary mirror and phase sections, and independently optimized MQW MZ modulator section. Shallow ridge waveguide is used throughout the PIC. As described above, shallow ridge waveguide is the most simple and robust choice among the possible waveguide structures. From manufacturability point of view, shallow ridge is optimum technology choice for applications that require a single MZ with a moderate bandwidth. Fig. 9 shows some of the key performance characteristics of ILMZ based optical transmitter. Notably, the differentially driven lumped MZ modulator requires less than 1.6 Vpp of modulation voltage per arm. Such low drive voltage is compatible with low power dissipation modulator drivers and is another very important factor that enabled full C-band tunable pluggable modules.

Narrow Linewidth High Power SG-DBR Laser

As described in the introduction, the LH, metro, and datacenter interconnects adopted coherent transmission format and continue to evolve towards higher bit rates by increasing the baud rate and the level of QAM modulation. This trend results in more stringent requirements imposed on the output power of the tunable lasers (> 16 dBm) and, even more importantly, on the laser linewidth (< 100 kHz). As discussed above, performance of widely tunable SG-DBR lasers with carrier injection tuning was adequate to meet full system requirements for 10 Gb/s OOK applications but additional tuning-induced loss is a challenge preventing injection tuned widely-tunable lasers from achieving narrow linewidth across the C-band. To overcome this limitation, Larson *et al.* (2015) developed thermally tuned SG-DBR laser and employed thermal engineering tailoring the impedance of the tuning sections to minimize power consumption and maintain millisecond-timescale tuning speed.

In addition, flexible-grid and colorless/directionless/contention-less network architectures employing power-combined launch require transmission lasers with high spectral purity in order to minimize noise interference on other channels. In particular, side mode suppression ratios (SMSR) greater than 47–50 dB are needed. SMSR of monolithic widely tunable DBR lasers integrated with SOAs is typically limited to 40–45 dB due to gain tilt and amplification of the spontaneous emission generated by the SOA as it is reflected from the back mirror. Addition of the thermally tuned broad-band filter into the PIC enabled to overcome this limitation, thereby simultaneously achieving > 50 dB SMSR and < 70 kHz linewidth at $+ 17$ dBm fiber coupled output power across C-band.

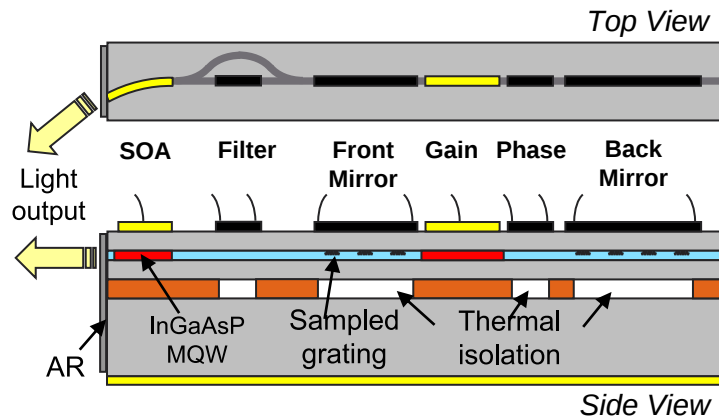


Fig. 10 (a) Schematic layout and cross-section of thermally tuned SG-DBR laser with integrated filter and SOA, (b) Instantaneous linewidth measure for 104 ITU channels at 16.5 and 17.5 dBm fiber coupled power.

The thermally tuned SG-DBR laser PIC comprising of the laser, delay interferometer tunable filter, and SOA is schematically shown in Fig. 10. The device uses regrowth integration technology similar to the injection tuned SG-DBR described above with the micro-heaters deposited above the shallow ridge structure in all tuning sections. The thermal impedance engineering is accomplished by placing sacrificial InGaAs layer underneath the entire PIC. As illustrated in Fig. 10, the sacrificial layer is selectively removed through lateral undercut etching in the thermally tuned sections. These sections become thermally isolated from the substrate and the heat generated in the micro heaters must flow laterally before reaching the substrate through unetched InGaAs. For all other sections the InGaAs remains as a spacer layer through which heat can flow vertically to the substrate/heat sink.

The thermally-tuned SGDBR laser described above is high performance optical source for high bit rate coherent transmitters and local oscillators and is widely adopted in coherent links as a discrete component utilized in combination with stand-alone modulators and coherent receivers. In addition, this all-monolithic laser design lends itself to integration with high-speed IQ modulators to produce high performance fully monolithic dual polarization coherent transmitter PIC.

Dual Polarization IQ Transmitter PIC

A polarization-multiplexed multilevel quadrature amplitude modulation transmitter requires a narrow linewidth laser and two IQ modulators. Compact IQ modulators with nested Mach-Zehnder (MZ) architecture have been successfully demonstrated using InP and Si modulator technologies. Power consumption of such transmitter is directly proportional to the active load (laser and SOA drive currents, modulator photocurrent and termination) and the required modulation voltage. Since the dual-polarization IQ modulator requires four drivers the power efficiency of the transceiver can be significantly improved by lowering modulation voltage.

Electro-optic MZ modulators in InP material systems fundamentally rely on the quantum-confined Stark effect, which produces red shift of the multi-quantum well (MQW) absorption edge and change of the refractive index when reverse bias voltage is applied to p-i-n hetero-structure. The modulation efficiency can be enhanced by optimizing the material band-gap, width of the QWs and using the material with larger conduction band discontinuity such as InAlGaAs. Modulator design optimization also includes the device length, depletion region width and electrode configuration to achieve desired modulation bandwidth. For a 32Gbaud MZ modulator the V_π voltage is typically below 2 V across C-band making InP modulator technology extremely attractive for coherent applications.

Another important consideration for minimizing transmitter power consumption and enabling high optical output power is low loss optical coupling between the laser and modulator. This can be accomplished only by monolithic integration of a narrow linewidth laser with an IQ modulator in a single PIC. Monolithic InP coherent PICs have been developed in multi-channel (Evans *et al.*, 2011) and tunable configurations (Binetti *et al.*, 2012; Akulova, 2016). In this article we will review the design, integration technology, and performance of the PIC reported in Akulova (2016).

The monolithic InP PIC is shown schematically in Fig. 11. The chip integrates C-band thermally tuned narrow linewidth SG-DBR laser with two IQ MZ modulators and three semiconductor optical amplifiers (SOAs). Light exiting the front mirror of the laser is split into three tributaries for X, Y, and Local Oscillator (LO) optical ports. Each optical path integrates an SOA for independent optical power leveling across C-band and variable optical attenuation functionality. The IQ modulators are realized using conventional nested MZ architecture. The signal electrodes of the IQ modulators are implemented using traveling wave design. Differential dc phase tuning sections are integrated into each of the IQ modulators and into individual MZs. The integration of multiple active and passive functional elements in one PIC has been accomplished using regrowth integration technology which is a platform technology for ILMZ and narrow linewidth SG-DBR laser PICs described above. The nested MZ modulator are fabricated using deep ridge waveguides. The performance characteristics of the monolithic DP-IQ transmitter PIC include V_π voltage of less than 2 V and the bandwidth in excess of 38 GHz (Fig. 12).

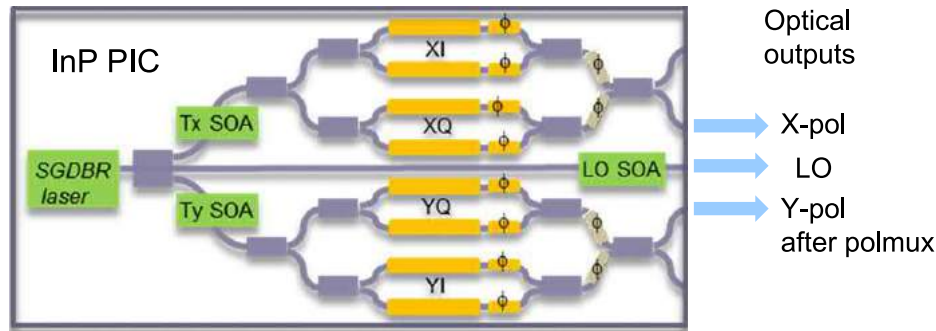


Fig. 11 Schematic layout of monolithic InP PIC consisting of thermally tuned narrow linewidth C-band tunable SGDBR laser, dual-polarization IQ modulators, and three SOAs.

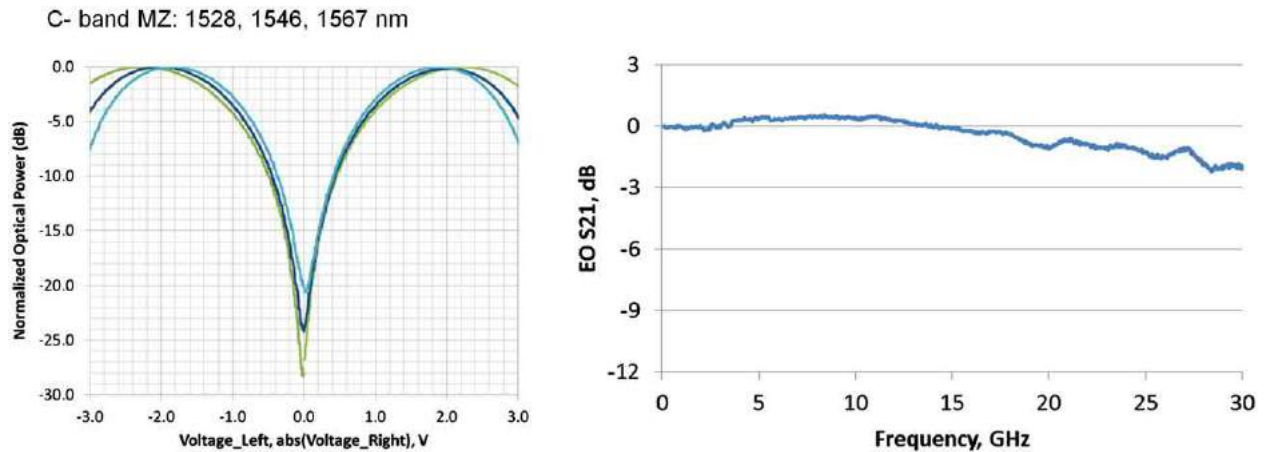


Fig. 12 Performance characteristics of monolithic widely tunable coherent transmitter PIC: (a) dc transfer function of MZ modulator measured at several wavelengths across C-band; (b) small signal modulation bandwidth.

High-efficiency of InP MZs modulators combined with the capability of lossless monolithic integration with tunable lasers and SOAs positions InP PIC technology as a primary candidate for realization of cost-effective, compact, high-efficiency optical transmitter engines for pluggable coherent modules for 100 and 200 Gb/s and thus enabling wide deployment of coherent transmission in metro and datacenter interconnects links. With the potential for higher bandwidth, high modulation efficiency, and increased scale of photonic integration, the InP PIC technology will continue to scale enabling compact, cost effective solutions for 400 Gb/s and beyond.

Electro-Absorption Modulated Lasers

An EML – the smallest scale InP PIC- consists of a DFB laser and electro-absorption modulator. EMLs have been extensively used in DWDM and ZR metro links 2.5 and 10 Gb/s for decades. They have been commercialized in 100 Gb/s transceivers starting from 2014 and currently being designed into 200 and 400 Gb/s small form factor pluggable transceivers. Wide adoption of the EMLs for these high bit rate application is driven by the transceiver architectures and link budget requirements.

Transmission component architectures adopted by IEEE for the inter-datacenter optical links (500 m to 2 km) and local area networks (up to 40 km) at 100 Gb/s combines 25 Gb/s OOK modulation with multiplexing four wavelengths into one single-mode fiber. These links utilize O-band wavelength range in the minimum fiber dispersion window. Similarly, standards that are currently in development for 200 and 400 Gb/s transmission rely on four and eight wavelengths, respectively, with 50 Gb/s pulsed amplitude modulation (PAM4) format which does not require increase in modulation bandwidth relative to 100 Gb/s architecture. Alternatively, 400 Gb/s transmission can be implemented with only four wavelength using PAM4 modulation at 100 Gb/s per wavelength. In all the cases implementation of components has to fit into power dissipation budget and footprint of small form factor pluggable modules such as QSFP28 and QSFP-DD. These links are un-amplified and have to support 6–18 dB link loss budget (including fiber and connectors loss and transmitter and dispersion penalty) resulting in stringent requirements imposed on transmitter output power and extinction ratio.

Links in the minimum fiber dispersion window relax the requirement on transmitter chirp and can be implemented using directly modulated lasers (DMLs) and EMLs. Several high-performance DMLs and DML array integrated with multiplexer PICs

have been reported (see, e.g., Kanazawa *et al.* (2016) and Matsui *et al.* (2016) and references therein). However, it is advantageous to separate light generation and modulation functions to enable uncooled operation with large margin relative to all performance specifications and therefore high manufacturing yield. EML based transmitters enable independent optimization of the laser for the target output power and low relative intensity noise while EA modulator can easily meet the bandwidth and ER requirements and offers superior dispersion tolerance as compared to DMLs, and therefore can be used up to 40 km applications.

Fig. 13 presents schematic x-section of an integrated DFB laser with a high speed EA modulator. The EML is fabricated using regrowth integration and shallow ridge technology. With exception of the modulator MQW regrowth step, the fabrication process for an EML is similar to the flow for the fabrication of a stand alone DFB laser. Interestingly, the fabrication process for 25 Gb/s DMLs is frequently more complex than for EMLs. In the effort to meet the requirements for modulation bandwidth, DMLs have to be designed with very small active volume and the fabrication might include buried heterostructure process and integration of passive sections to maintain manufacturability as the laser cavity length is reduced to below 150 μm .

EMLs are compact and can be co-packaged or monolithically integrated with multiplexers to realize small form factor transmitter optical sub-assemblies (TOSAs) compatible with QSFP28 transceivers. Fig. 14 shows typical transfer functions of an EA

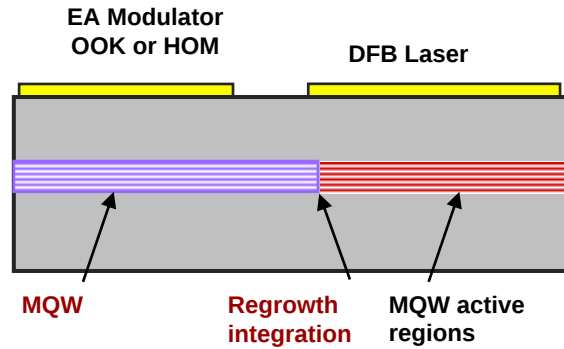


Fig. 13 Schematic cross-section of an EML chip.

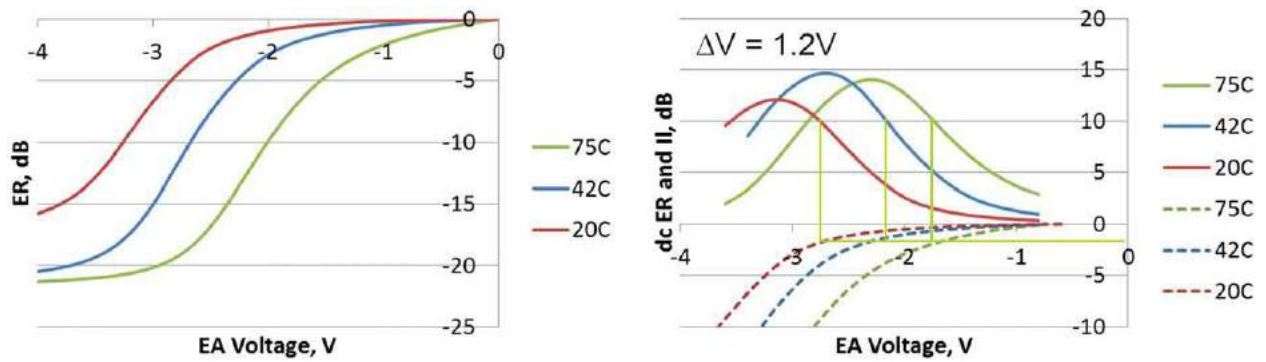


Fig. 14 (a) Transfer function of uncooled EML; (b) Extinction ratio vs insertion loss for fixed modulation voltage ($V_{pp} = 1.2$ V).

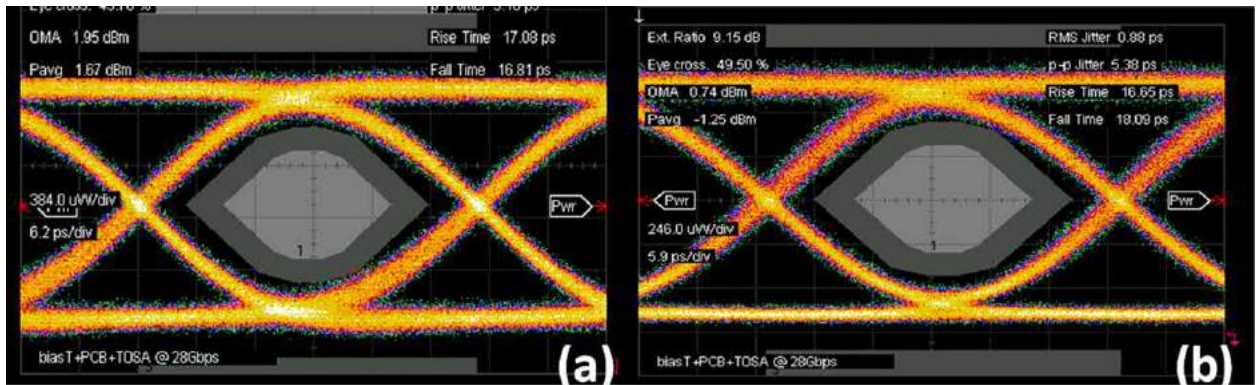


Fig. 15 Eye diagrams measured at 28 Gb/s: (a) $ER > 5$ dB, $MM > 40\%$, $V_{pp} = 1.2$ V; (b) $ER > 9$ dB, $MM > 40\%$, $V_{pp} = 1.5$ V.

modulator measured at several temperatures and ER versus insertion loss calculated from the measured transfer function by changing dc bias with fixed modulation voltage. The ER in excess of 10 dB with insertion loss below 2 dB is demonstrated with 1.2 V modulation voltage across wide temperature range. The appropriate bias and modulation voltage is selected based on the target extinction ratio during transceiver calibration. Fig. 15 presents eye diagrams measured at 28 Gb/s. Extinction ratio in excess of 5 and 9 dB is demonstrated with the modulation voltage of 1.2 and 1.5 V, respectively. Such low requirements on the drive voltage ensure compatibility with low power dissipation modulator drivers and improve the power budget of the transceivers.

In summary, we have reviewed available InP integration technologies and discussed how optical link performance requirements drive the selection of PIC functional sections and integration approaches. We presented design, integration technology, and performance for several InP PICs that have been commercialized in Datacom, Metro, and LH transmission systems. InP PICs performance continues to scale to deliver higher transmission capacity at lower cost and power dissipation. In addition, InP technology enables integration of wide range of photonic functions and impacts broad range of market segments including photonic solutions for sensing, imaging, and high-speed signal analysis.

References

- Akulova, Y., 2016. Advances in integrated widely tunable coherent transmitters. In: Proceedings of Optical Fiber Communications Conference W4H.1, IEEE.
- Binetti, P.R.A., Lu, M., Norberg, E.J., *et al.*, 2012. Indium phosphide photonic integrated circuits for coherent optical links. *Journal of Quantum Electronics* 48, 279–291.
- Coldren, L.A., Corzine, S.W., Mashanovitch, M.L., 2013. Diode lasers and photonic integrated circuits, second ed. John Wiley & Sons.
- Coldren, L.A., Fish, G.A., Akulova, Y., *et al.*, 2004. Tunable semiconductor lasers: A tutorial. *Journal of Lightwave Technology* 22, 193–202.
- Coldren, L.A., Nicholes, S.C., Johansson, L., *et al.*, 2011. High performance InP based photonics ICs – A tutorial. *Journal of Lightwave Technology* 29, 554–570.
- Evans, P., Fisher, M., Malendevich, R., *et al.*, 2011. Multi-channel coherent PM-QPSK InP transmitter photonic integrated circuit (PIC) operating at 112 Gb/s per wavelength. In: Proceedings of Optical Fiber Communications Conference PDPC7, IEEE.
- Kanazawa, S., Kobayashi, W., Ueda, Y., *et al.*, 2016. 30-km Error-free transmission of directly modulated DFB laser array transmitter optical sub-assembly for 100-Gb application. *Journal of Lightwave Technology* 34, 3646–3652.
- Larson, M.C., Bhardwaj, A., Xiong, W., *et al.*, 2015. Narrow linewidth sampled-grating distributed bragg reflector laser with enhanced side-mode suppression. In: Proceedings of Optical Fiber Communications Conference M2D.1, IEEE.
- LAN/MAN Standards Committee of the IEEE Computer Society, 2010. IEEE Std 802.3ba-2010 and IEEE Std 802.3bs- draft. Available at: www.ieee802.org.
- Matsui, Y., Pham, T., Sudo, T., *et al.*, 2016. 28-Gbaud PAM4 and 56-Gb/s NRZ performance comparison using 1310-nm Al-BH DFB laser. *Journal of Lightwave Technology* 34, 2677–2683.
- Menon, V.M., Xia, F., Forrest, S.R., 2005. Photonic integration using asymmetric twin-waveguide (ATG) technology: Part II – Devices. *IEEE Journal of Selected Topics in Quantum Electronics* 11, 30–42.
- Skogen, E.J., Raring, J.W., Morrison, G.B., *et al.*, 2005. Monolithically integrated active components: a quantum-well intermixing approach. *IEEE Journal of Selected Topics in Quantum Electronics* 11, 343–355.
- Williams, K.A., Bente, E.A.J.M., Heiss, D., *et al.*, 2015. InP photonic circuits using generic integration. *Photonics Research* 3, B60–B68.
- Winzer, P.J., 2012. High-spectral-efficiency optical modulation formats. *Journal of Lightwave Technology* 30, 3824–3835.
- Xia, F., Menon, V.M., Forrest, S.R., 2005. Photonic integration using asymmetric twin-waveguide (ATG) technology: Part I – Concepts and theory. *IEEE Journal of Selected Topics in Quantum Electronics* 11, 17–29.

CMOS Transceiver Circuits for Optical Interconnects

Samuel Palermo, Texas A&M University, College Station, TX, United States

© 2018 Elsevier Ltd. All rights reserved.

Optical Interconnects

Optical interconnects for computing systems use electronic driver circuitry to transmit digitally-modulated light from a source laser over an optical channel, most commonly an optical fiber, to a photodetector where the signal is converted back into the electrical domain with a receiver circuit. Relative to electrical interconnects, these optical communication systems overcome key interconnect bottlenecks and greatly improve data transfer efficiency due to the optical fiber's flat channel loss over a wide frequency range and also relatively small crosstalk and electromagnetic noise. Another important feature of optical interconnects is the ability to combine multiple data channels on a single waveguide via wavelength-division-multiplexing (WDM) and greatly improve bandwidth density. These key features have motivated the application of optical interconnect systems in high-performance computing and data center systems where transmission distances can exceed 1 km.

The dominant optical interconnect transmitter technology for these applications has been directly-modulated multi-mode vertical-cavity surface-emitting lasers (VCSELs) due to these devices being inexpensive and the relaxed alignment tolerances offered by the multi-mode fiber system. However, modal dispersion limits performance as distances reach the km-range and data rates climb above 25 Gb/s. This motivates single-mode solutions that allow for increases in both transmission distances and bandwidth density via WDM. These single-mode systems often utilize external modulation of a continuous-wave (CW) laser operating at a constant power level in order to overcome laser bandwidth limitations and optical spectrum broadening due to modulated laser chirp. Common external modulator devices include electroabsorption modulators (EAMs) and refractive modulators, such as the Mach-Zehnder and ring resonator modulators.

At the receive side, a photodiode is typically utilized to sense the high-speed optical power and produce an input current. This photocurrent is then converted to a voltage and amplified sufficiently for data resolution. In order to support high data rates, sensitive high-bandwidth photodiodes are necessary. High-speed p-i-n photodiodes are typically used in optical receivers due to their high responsivity and low capacitance. In these device structures, the light is absorbed in the intrinsic region and the generated carriers are collected at the reverse bias terminals, thereby causing an effective photocurrent to flow. The amount of current generated for a given input optical power is set by the detector's responsivity. For devices where the light is normally incident, there is an inherent tradeoff between responsivity and bandwidth set by the intrinsic layer width. A wider intrinsic region will result in an increased amount of absorbed photons and higher responsivity, but will also cause increased carrier transit times that results in reduced detector bandwidth. This responsivity-bandwidth tradeoff is broken in waveguide p-i-n structures where the light travels horizontally down the intrinsic region and the electric field is formed orthogonal. These structures allow for both a thin intrinsic region for short carrier transit times and a sufficiently long region for high responsivity.

Energy efficient optical interconnect systems require low-power and area circuits that interface to these optical sources and detectors. This article focuses on CMOS driver and receiver circuits, as this technology allows for low-voltage and low-power operation and the potential to integrate the optical front-ends with the main data processing integrated circuits (ICs). While high-performance silicon-germanium (SiGe) or other compound semiconductor technologies may offer superior performance relative to CMOS implementations, these processes are often more expensive and/or not competitive in terms of energy efficiency for short distance optical interconnect applications.

Transmitter Circuitry

Directly modulated lasers and optical modulators, both electroabsorption and refractive, have been proposed as high bandwidth sources for optical interconnects, with these different sources displaying tradeoffs in both device and circuit driver efficiency. Vertical-cavity surface-emitting lasers are an attractive candidate due to their ability to directly emit light with low threshold currents and reasonable slope efficiencies; however their speed is limited by both electrical parasitics and carrier-photon interactions. A device which doesn't display this carrier speed limitation is the electroabsorption modulator, based on either the quantum-confined Stark effect (QCSE) or the Franz-Keldysh effect, which is capable of achieving acceptable contrast ratios at low drive voltages over tens of nm optical bandwidth. Ring resonator modulators are refractive devices that display very high resonant quality factors and can achieve high contrast ratios with small dimensions and low capacitance, however their optical bandwidth is typically less than 1 nm. Another refractive device capable of wide optical bandwidth (>100 nm) is the Mach-Zehnder modulator, however this comes at the cost of a large device and high voltage swings. All of the optical modulators also require an external source laser and incur additional coupling losses relative to a VCSEL-based link. The following describes CMOS circuitry optimized for these particular optical transmitter devices.

VCSEL Transmitters

A VCSEL, shown in Fig. 1(a), is a semiconductor laser diode which emits light perpendicular from its top surface. These surface emitting lasers offers several manufacturing advantages over conventional edge-emitting lasers, including wafer-scale testing ability and dense 2D array production. The most common VCSELs are GaAs-based operating at 850 nm, with longer wavelength devices manufactured in various material structures also available that operate near 1000, 1310, and 1550 nm. Current-mode drivers are often used to modulate VCSELs due to the device's linear optical power-current relationship. A typical CMOS VCSEL output driver is shown in Fig. 1(b), with a differential stage steering current between the optical device and a dummy load, and an additional static current source used to bias the VCSEL sufficiently above the threshold current, I_{TH} , in order to ensure adequate bandwidth.

Total VCSEL bandwidth is limited by a combination of electrical parasitics and the electron-photon interaction dynamics. The laser diode's dominant electrical time constant comes from the bias-dependent junction RC. In addition to the bias-dependent junction resistance, there is also significant series resistance due to the large number of distributed Bragg reflector (DBR) mirrors used for high reflectivity. VCSEL optical bandwidth is regulated by two coupled differential equations which describe the electron-photon interaction. Derived from these rate equations, the VCSEL relaxation oscillation frequency ω_R , which is proportional to the effective bandwidth, is directly proportional to the square-root of the injected current above the threshold current I_{TH} .

$$\omega_R \propto \sqrt{I - I_{TH}} \quad (1)$$

Output power saturation due to self-heating and also device lifetime concerns restrict excessive increase of VCSEL average current levels to achieve higher bandwidth. VCSEL reliability potentially poses a series impediment to very high speed modulation, as the mean time to failure (MTTF) is inversely proportional to the current density squared. The conflicting dependencies of VCSEL bandwidth and reliability on device current yield a very steep tradeoff.

In order to ease this tradeoff, equalizing output stages have been proposed to extend the data rate for a given average current. As an example, Fig. 2 shows a VCSEL transmitter with a four-tap finite impulse response (FIR) equalizer consisting of one pre-cursor, one main, and two post-cursor taps implemented by summing current sources at the output node. Five parallel data bits, $D[4:0]$, are routed to the taps, where they are shifted one bit time with respect to the clock phases to implement the necessary filter delays. At each tap, a multiplexer serializes the five parallel input bits and drives a differential output stage which steers current between the VCSEL and dummy diode-connected thick-oxide nMOS devices that are connected to a separate 2.8 V LVdd supply. This higher supply is necessary to support the 1.5 V VCSEL knee voltage. Eight-bit current mirror DACs bias the output stages to the desired current value. Because of the smaller current requirements of the pre/post-cursor taps, their muxes and output stages are set to one-fourth the size of the main tap to save power. A static DC current source, I_{DC} , is also used to bias the VCSEL above the threshold current to insure adequate bandwidth. This bias current and the leakage current from the tap driver transistors, I_{leak} , provide sufficient voltage drop across the VCSEL and dummy load to prevent excessive voltage stress on the output stage transistors. While the VCSEL's varying frequency response with current limits the performance of this linear equalizer for large extinction ratio modulation, the frequency response variations diminish with increasing average current due to the square-root relationship and a linear equalizer is effective in canceling intersymbol interference (ISI) for extinction ratios near 3 dB. In order to further improve performance, non-linear equalizers have been proposed with asymmetric current profiles for the different data transitions.

Electroabsorption Modulator Transmitters

An electroabsorption modulator is typically made by placing an absorbing quantum-well region in the intrinsic layer of a reverse-biased p-i-n diode. In order to produce a modulated optical output signal, light originating from a continuous-wave source laser is absorbed in an EA modulator depending on electric field strength through electro-optic effects such as the quantum-confined Stark effect or the Franz-Keldysh effect. These devices are often implemented as a waveguide structure where light is coupled in and

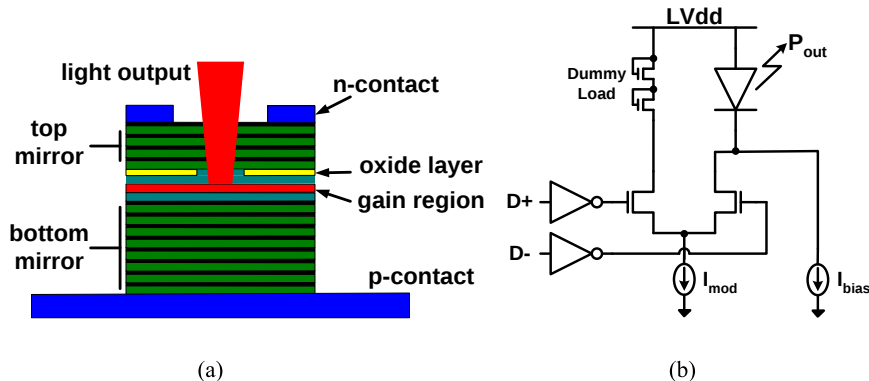


Fig. 1 Vertical-Cavity Surface-Emitting Laser: (a) device cross-section and (b) driver circuit.

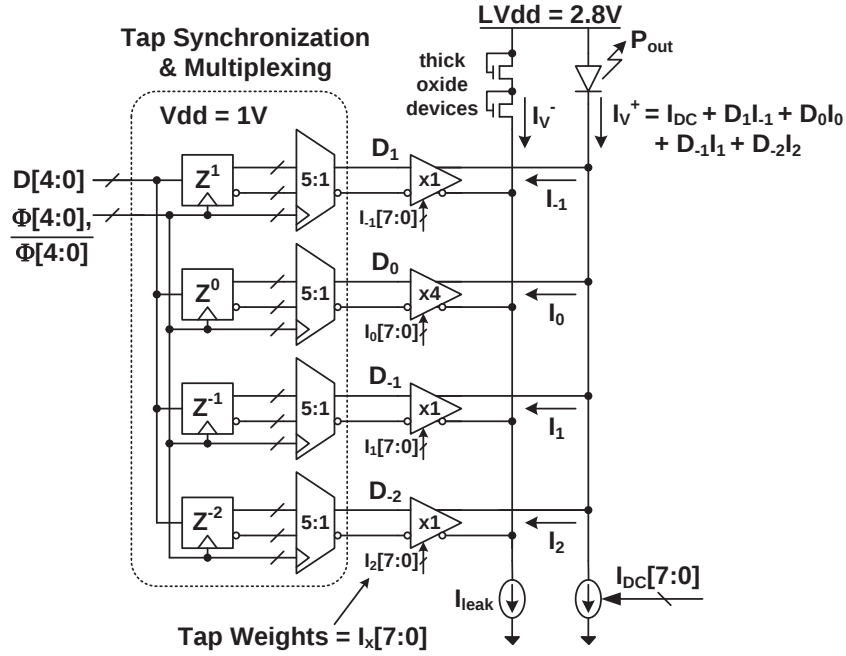


Fig. 2 VCSL transmitter with a four-tap FIR equalizer.

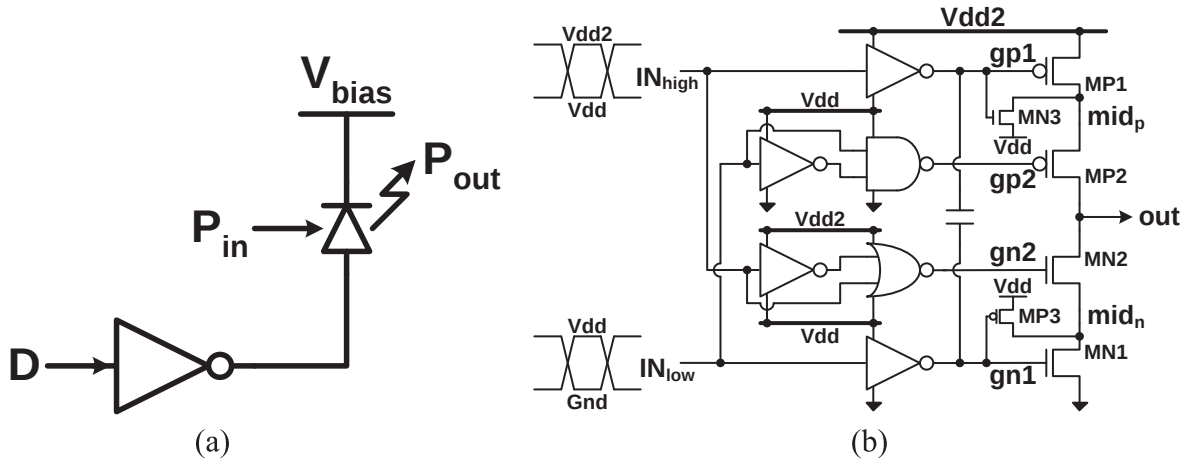


Fig. 3 Electroabsorption modulator drivers: (a) simple CMOS inverter and (b) high-swing pulsed-cascode output stage.

travels laterally through the absorbing multiple-quantum-well (MQW) region. Modulation is typically achieved by applying a static positive bias voltage to the n-terminal and driving the p-terminal with the high-speed signal.

Assuming tight integration with the driver circuitry, the EAM can be treated as a lumped-element device consisting of the diode capacitance and a parallel voltage-dependent photocurrent source. While conceptually the high-speed driver could be a simple CMOS inverter (Fig. 3(a)), the sub-1 V swings available in modern CMOS processes limit the achievable extinction ratio for relatively short lumped-element EAMs. Improved extinction ratios are possible with longer EAM device lengths. However, achieving high-speed operation with these longer devices necessitates traveling-wave driver topologies that often require low-impedance termination and a relatively large amount of switching current. Another option is to use a lumped element driver with a higher output swing than the nominal CMOS supply. The ability to drive the smaller devices as an effective lumped-element capacitor offers a huge power advantage when compared to longer controlled-impedance structures, as the CV^2f power is relatively low due to small device capacitance. One circuit which allows for this in a reliable manner is a pulsed-cascode driver, which offers a voltage swing of twice the nominal supply while using only core devices for maximum speed.

Fig. 3(b) shows the pulsed-cascode output stage which accepts both a “low” input IN_{low} that swings between the Gnd and the nominal chip V_{dd} and a “high” input IN_{high} with the same data value that has been level-shifted to swing between V_{dd} and V_{dd2} , where V_{dd2} is nominally twice the voltage of V_{dd} . Static-voltage overstress is eliminated in the output-stage cascode structure by

equally splitting the output voltage across the series transistors. Pulsing the gates of the cascode transistors (MN2 and MP2) during transitions with NAND- and NOR-pulse gates, respectively, allows this driver to eliminate the transient drain-source voltage (V_{DS}) overstress present in static-biased cascode drivers and prevents transistor degradation from hot-carrier injection.

Ring Resonator Modulator Transmitters

As shown in Fig. 4(a), a ring resonator consists of a top waveguide which couples a portion of the incoming continuous wavelength (CW) light into a ring waveguide. At the resonance wavelength, most of the incident light will be trapped due to the interference between the ring and top waveguide. This results in minimal power at the through port, as shown in Fig. 4(b). Adding a second bottom waveguide to the ring allows for the light coupled into the ring from the input port to be coupled out onto the bottom waveguide at the resonance wavelength, implementing a drop filter. By changing the ring waveguide's effective refractive index via the plasma dispersion effect, the resonance wavelength can be shifted to allow for modulation. Fig. 4(c) shows that if the resonance wavelength is aligned with the incident light wavelength, an adequate output extinction ratio is possible. However, if the high-Q device's resonance is off due to temperature sensitivity or manufacturing variations, the performance greatly diminishes (Fig. 4(d)).

The two most common silicon ring resonator modulators, carrier-injection and carrier-depletion devices, are implemented by forming a junction across the ring waveguide. Carrier-injection devices have an intrinsic ring waveguide surrounded by p+ and n+ regions that form the modulator's two electrical terminals. As they operate primarily in forward bias, this allows for significant changes in the intrinsic-region carriers with varying drive signals. While this provides large refractive index changes and high modulation depths, dynamic operation is limited by long minority carrier lifetimes. A pn junction is formed directly in the ring waveguide in carrier-depletion modulators. Higher speeds relative to a carrier-injection ring are generally achievable due to the

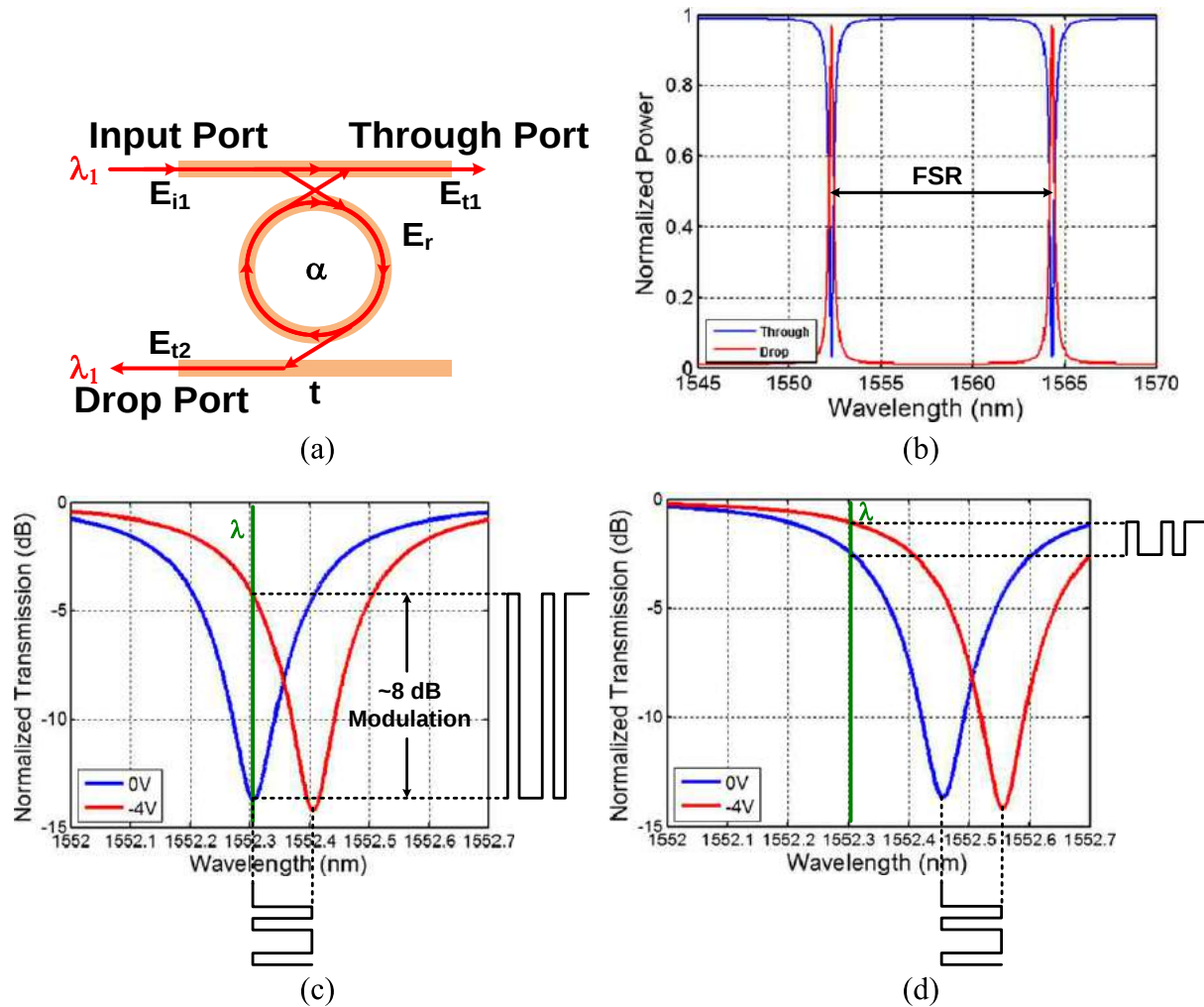


Fig. 4 Ring resonator device: (a) schematic, (b) through and drop port power versus wavelength, (c) modulator operation with 0 V bias resonance equal to the incident light wavelength, (d) modulator operation with a 160 pm resonance shift.

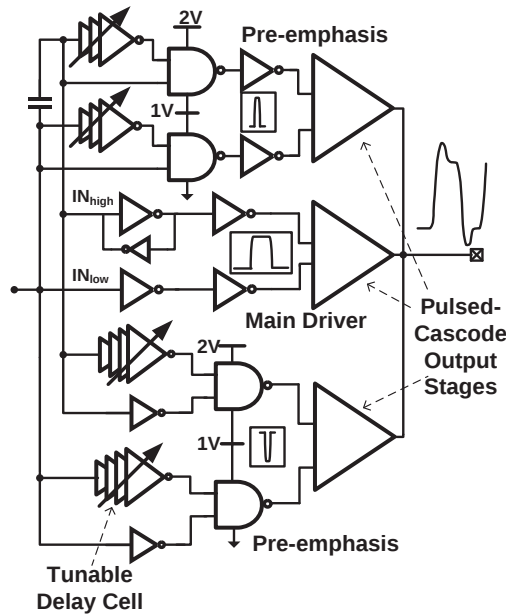


Fig. 5 Non-linear pre-emphasis ring modulator driver with $2V_{pp}$ output stage.

ability to rapidly change the depletion width. However, the modulation depth is limited due to the low doping concentration in the waveguide to avoid excessive loss.

The differing characteristics of carrier-injection and carrier-depletion ring modulators necessitate variations in the output stage design to enable high data rate operation. While carrier-injection modulators are capable of high extinction ratios, the operating speed is limited by relatively slow carrier dynamics in forward-bias and parasitic contact resistance in reverse-bias. This necessitates relatively large amounts of non-linear pre-emphasis to compensate for the differing time constants during rising and falling transitions. A single-ended driver optimized for carrier-injection modulators is shown in [Fig. 5](#). This design utilizes a pulsed-cascode output stage for robust high output swing and segments this output stage into main and independent rising/falling edge pre-emphasis drivers. Here the per-edge pre-emphasis strength is set with tunable delay elements that allow optimization of the transient response to be decoupled from the steady-state extinction ratio value. A 65 nm CMOS implementation of this design achieved an output swing of $2V_{pp}$ and a 9.2 dB extinction ratio at 9 Gb/s when driving a carrier-injection ring modulator.

Depletion-mode modulators have shorter carrier lifetimes and can easily achieve data rates >10 Gb/s. However, they typically exhibit low pn junction tunability that necessitates higher modulation voltages, require negative DC-biasing for high-speed operation, and need pre-emphasis optimized for non-linear optical dynamics important at data rates >20 Gb/s. One circuit which can achieve this is a differential version of the [Fig. 5](#) driver with on-die AC-coupling capacitors that allow for a constant negative-bias operation over large-swing dynamic operation. A 65 nm CMOS implementation of this design achieved $4.4V_{pp}$ differential swing and implemented asymmetric 2-tap equalization in the output stage to operate at 25 Gb/s with a 7 dB ER.

In order to ensure reliable operation over temperature and fabrication variations, resonance wavelength tuning can be achieved by adjusting the ring resonator's refractive index thermally, electrically, or by employing both techniques. Thermal tuning with integrated resistor heaters is most commonly employed, as it does not impact ring loss, but the tuning speed is limited by thermal time constants. Alternatively, electrical or bias-based tuning offers the potential for rapid tuning at the cost of some reduction in modulator extinction ratio. This approach has been shown to be effective with carrier-injection modulators, but is generally not considered effective for carrier-depletion modulators due to the reduced tunability.

[Fig. 6](#) shows an average-power-based dynamic thermal tuning loop where a drop-port with a waveguide PD has been added to a ring modulator to monitor the ring's power levels. An FSM automatically locks the ring to the optimal thermal bias point by comparing the average optical power with an on-chip reference voltage to produce error information for the control of a thermal DAC that drives a resistive heater placed close to the modulator. Successful demonstration of this dynamic tuning loop with through-port monitoring was achieved for 25 Gb/s operation of carrier-depletion modulators with integrated $1\text{ k}\Omega$ resistor heaters. This tuning loop was also modified to also tune the drop filters in a 25 Gb/s receiver front-end, with ring power monitoring achieved by adding peak detectors at the TIA output in order to not compete with the receiver's offset correction loop. A similar control loop was used for bias-based tuning of carrier-injection ring modulators, with the thermal DAC replaced by a bias DAC driving the modulator's anode terminal and the high-speed modulation applied to the cathode. This opens the possibility for dual-loop tuning with carrier-injection modulators, which involves a coarse-step thermal DAC and fine tuning with the bias DAC.

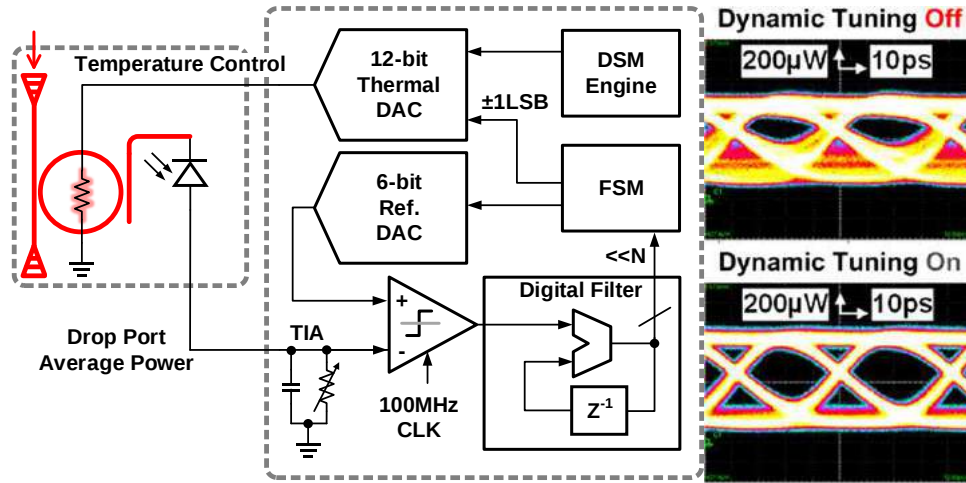


Fig. 6 Average-power-based dynamic thermal tuning loop to stabilize the ring resonance wavelength.

Mach-Zehnder Modulator Transmitters

Mach-Zehnder modulators (MZMs) are also refractive modulators which work by splitting the light to travel through two arms where a phase shift is developed that is a function of the applied electric field. The light in the two arms is then recombined either in phase or out of phase at the modulator output to realize the modulation. MZMs which use the free-carrier plasma-dispersion effect in pn diode devices to realize the optical phase shift have been integrated in CMOS processes and have demonstrated operation in excess of 10 Gb/s. The modulator transfer characteristic with half wave voltage V_π is

$$\frac{P_{out}}{P_{in}} = \frac{1}{2} \left(1 + \sin \frac{\pi V_{swing}}{V_\pi} \right) \quad (2)$$

Unlike smaller modulators which are treated as lumped capacitive loads, the relatively low modulation efficiency requires MZM phase shifters that commonly exceed 1 mm in silicon photonic implementations. Thus, the lumped-element drivers that are commonly utilized for ring resonator and electroabsorption modulators are not well suited for MZM operation at high speed. Instead, as shown in Fig. 7(a), traveling-wave drivers are often used with a differential electrical output signal that is distributed along the MZM phase shifter length using a pair of transmission lines terminated with a low impedance. The best high-speed output signal integrity is achieved when the transmission line electrodes are designed such that the propagation velocity of the electrical signal matches that of the optical signal in the MZM waveguides. While conceivably large contrast ratios are possible by simply increasing the MZM phase shifter length, this is limited by high-frequency loss of the long transmission lines and also increased optical insertion loss caused by the long doped waveguides. In order to achieve the required phase shift and reasonable contrast ratio, long devices and large differential swings are required; often necessitating a separate voltage supply MV_{dd} . Thick-oxide cascode transistors are used to avoid stressing driver transistors with the high supply.

A driver which provides more flexibility in electrode design and allows for high-swing lumped element drivers is the multistage topology shown in Fig. 7(b). This design segments the total length of the MZM phase shifters into shorter regions which are driven by high-swing stacked inverter-style drivers. The segment length is chosen to have a self-resonance frequency higher than the signal bandwidth. As the high-frequency loss of these short segments is relatively small, higher contrast ratios are possible with high-swing drivers. A combination of passive wire delay elements and tunable-delay circuitry are employed to match the multistage driver delay with the optical propagation velocity. In this design, the speed and power efficiency is limited by distributing CMOS-level signals across long distances and the power consumed by the tunable delay elements. This driver topology generally achieves better power efficiency at low to moderate data rates due to the dominant CV^2f switching power component. While at the highest data rates, the limited contrast ratio traveling wave design becomes more power efficient.

Receiver Circuitry

Optical receiver front-ends generally determine the overall optical link performance, as their sensitivity sets the maximum data rate and amount of tolerable channel loss. Typical optical receivers use a photodiode to sense the high-speed optical power and produce an input current. This photocurrent is then converted to a voltage and amplified sufficiently for data resolution. In order to achieve increasing data rates, sensitive high-bandwidth receiver front-end circuits are necessary.

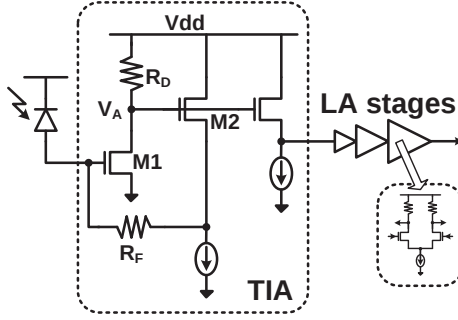


Fig. 9 Feedback TIA.

modifying a conventional common-gate input stage to a regulated cascode (RGC) architecture which employs active negative feedback gain to boost the input transconductance. This reduced input resistance pushes the input pole to a higher frequency, relaxing tradeoffs between TIA gain and bandwidth. However, conventional RGC topologies require additional voltage headroom due to the cascode topology. Moreover, extra power is required in the feedback stage in order to avoid excessive TIA frequency peaking and obtain sufficient noise performance. An interesting alternative approach to improve the input bandwidth involves employing a transformer-based RGC input stage which provides passive negative-feedback gain that enhances the effective transconductance of the input common-gate transistor. This allows for considerable bandwidth extension without significant noise degradation or power consumption.

Another wideband receiver frontend is the shunt-feedback TIA shown in **Fig. 9**, which decouples the transimpedance gain and input impedance by the amount of amplifier gain.

$$R_T = R_F \left(\frac{A}{1 + A} \right)$$

Feedback TIA Front-End :

$$\omega_{3dB} \approx \frac{1 + A}{R_F C_{in}}$$

where $A \approx g_{m1} (R_D \parallel r_{o1})$

(5)

The feedback TIA topology has the advantage that only the photodetector signal current flows through R_F , which provides the potential for large transimpedance gain without voltage headroom constraints. This structure allows for potentially both high transimpedance and bandwidth, provided that the amplifier in the TIA has a sufficient gain-bandwidth product, Af_A , to ensure system stability. However, as data rates scale, the TIA's amplifier gain-bandwidth must increase as a quadratic function in order to maintain the same effective TIA gain because of the inherent transimpedance limit.

$$R_T \leq \frac{Af_A}{2\pi C_{in} f_{3dB}^2}$$
(6)

TIA performance scaling is further limited by the lack of gain that is achieved in modern CMOS processes at nominal supply voltages due to both voltage headroom constraints and intrinsic transistor gain. This limits the maximum R_F value, which both reduces the transimpedance gain and increases the input-referred current noise. As data rates increase and less gain is realized in the input transimpedance stage, receiver sensitivity is improved with additional voltage or limiting amplifier (LA) stages that follow the TIA. High-performance optical receivers often use four or more differential amplifier stages in order to achieve adequate sensitivity, which can more than double the total optical receiver power consumption.

Bandwidth-Limited Frontends

The aforementioned optical receiver scaling issues have motivated researchers to investigate bandwidth-limited front-ends in order to improve sensitivity and power consumption. A receiver front-end architecture that eliminates linear high gain elements, and thus is less sensitive to the reduced gain in modern processes, is the integrating and double-sampling front-end shown in **Fig. 10**. The absence of high gain amplifiers allows for savings in both power and area and makes the integrating and double-sampling architecture advantageous for chip-to-chip optical interconnect systems where retiming is also performed at the receiver. The integrating and double-sampling receiver front-end demultiplexes the incoming data stream with parallel segments that include a pair of input samplers, a buffer, and a sense-amplifier. Two current sources at the receiver input node, the photodiode current and a current source that is feedback biased to the average photodiode current, supply and deplete charge from the receiver input capacitance respectively. For data encoded to ensure DC balance, the input voltage will integrate up or down due to the mismatch in these currents. A differential voltage, ΔV_b , that represents the polarity of the received bit is developed by sampling the input voltage at the beginning and end of a bit period defined by the rising edges of the synchronized sampling clocks $\Phi[n]$ and $\Phi[n+1]$ that are spaced a bit-period, T_b , apart. This differential voltage is buffered and applied to the inputs of an offset-corrected sense-amplifier which is used to regenerate the signal to CMOS levels.

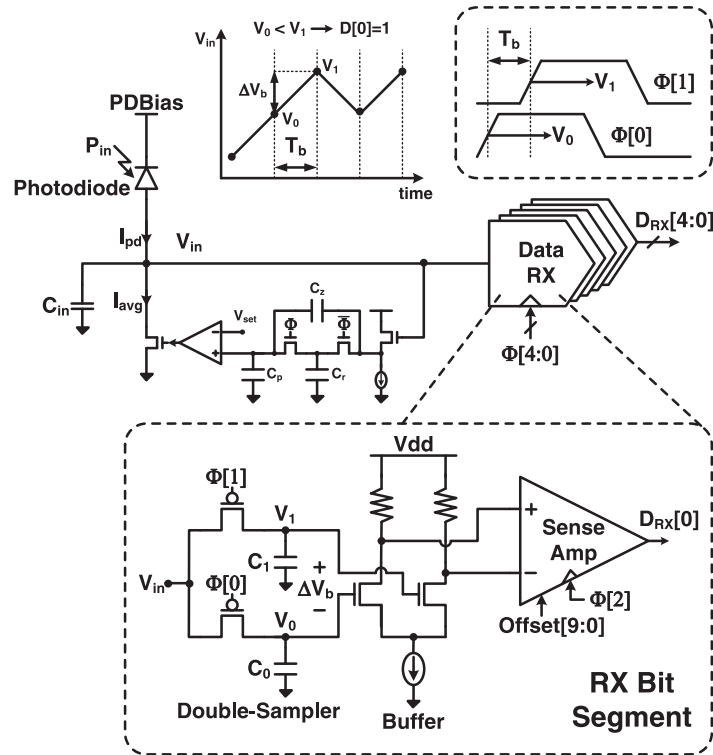


Fig. 10 Integrating and double-sampling receiver front-end.

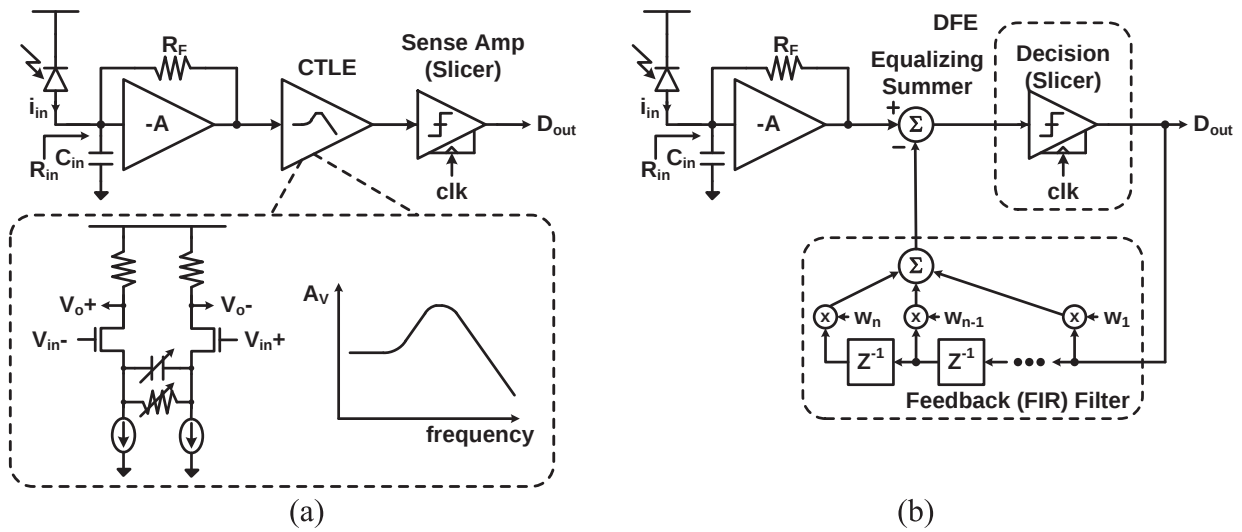


Fig. 11 Low-bandwidth TIA front-end followed by an equalizer: (a) continuous-time linear equalizer and (b) decision feedback equalizer.

One issue with this integrating and double-sampling front-end is that the input node saturates with long runlengths of consecutive bits. A solution to this is to implement a lossy integrator input node by utilizing a relatively large input resistor or low-bandwidth feedback TIA front-end to limit the low-frequency swing to a predictable value. A constant effective differential voltage is achieved by dynamically modulating the sense-amplifier's offset with a discrete-time FIR filter that follows the low-bandwidth input stage.

Fig. 11 shows another receiver front-end architecture which offers sensitivity improvement by utilizing a low-bandwidth feedback TIA followed by an equalizer circuit. At high data rates, the input-referred current noise of wideband feedback TIAs is often dominated by the feedback resistor R_F whose value is limited to achieve sufficient bandwidth. Intentionally increasing R_F results in lower noise which is integrated over a smaller bandwidth. However, significant intersymbol interference (ISI) exists at the

output of this low-bandwidth feedback TIA front-end. This ISI is then compensated by the subsequent equalizer block to achieve adequate voltage and timing margins at the data samplers.

A continuous-time linear amplifier (CTLE) is a simple and low-area equalizer block which implements a high-pass filter transfer function to restore the total receiver front-end bandwidth. While this equalizer doesn't require any additional sampling clocks, it does have to supply gain at frequencies close to the full signal data rate. Also, the CTLE amplifier introduces additional noise. Nonetheless, the ability to reduce the input R_F noise significantly allows for a net sensitivity improvement.

Another equalizer topology is the decision feedback equalizer (DFE). A DFE attempts to directly subtract ISI from the incoming signal by feeding back the resolved data to control the polarity of the equalization taps. Unlike linear equalization, a DFE doesn't directly amplify the input signal noise or cross-talk since it uses the quantized input values. However, there is the potential for error propagation in a DFE if the noise is large enough for a quantized output to be wrong. Also, due to the feedback equalization structure, the DFE cannot cancel pre-cursor ISI. The major challenge in DFE implementation is closing timing on the first tap feedback since this must be done in one bit period or unit interval (UI). Direct feedback implementations require this critical timing path to be highly optimized. Improvements in DFE speed are possible with loop-unrolling architectures which speculative summation with multiple data samplers and additional look-ahead logic to relax the critical timing path.

Further Reading

- Analui, B., Guckenberger, D., Kucharski, D., Narasimha, A., 2006. A fully integrated 20-Gb/s optoelectronic transceiver implemented in a standard 0.13- μm CMOS SOI technology. *IEEE Journal of Solid-State Circuits* 41, 2945–2955.
- Cevrero, A., Ozkaya, I., Francese, P.A., *et al.*, 2017. A 64 Gb/s 1.4pJ/b NRZ optical-receiver data-path in 14 nm CMOS FinFET. In: *IEEE International Solid-State Circuits Conference*, pp. 482–483.
- Emami-Neyestanak, A., Liu, D., Keeler, G., Helman, N., Horowitz, M., 2002. A 1.6 Gb/s, 3 mW CMOS receiver for optical communication. In: *IEEE Symposium on VLSI Circuits*, pp. 84–87.
- Li, C., Bai, R., Shafik, A., *et al.*, 2014. Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65-nm CMOS. *IEEE Journal of Solid-State Circuits* 49, 1419–1436.
- Li, C., Palermo, S., 2013. A low-power 26-GHz transformer-based regulated cascode SiGe BiCMOS transimpedance amplifier. *IEEE Journal of Solid-State Circuits* 48, 1264–1275.
- Li, D., Minoia, G., Repossi, M., *et al.*, 2014. A low-noise design technique for high-speed CMOS optical receivers. *IEEE Journal of Solid-State Circuits* 49, 1437–1447.
- Li, H., Xuan, Z., Titriku, A., *et al.*, 2015. A 25 Gb/s, 4.4 V-swing, AC-coupled ring modulator-based WDM transmitter with wavelength stabilization in 65 nm CMOS. *IEEE Journal of Solid-State Circuits* 50, 3145–3159.
- Palermo, S., Emami-Neyestanak, A., Horowitz, M., 2008. A 90 nm CMOS 16 Gb/s transceiver for optical interconnects. *IEEE Journal of Solid-State Circuits* 43, 1235–1246.
- Palermo, S., Horowitz, M., 2006. High-speed transmitters in 90 nm CMOS for high-density optical interconnects. In: *IEEE European Solid-State Circuits Conference*, pp. 508–511.
- Park, S., Yoo, H., 2004. 1.25-Gb/s regulated cascode CMOS transimpedance amplifier for gigabit ethernet applications. *IEEE Journal of Solid-State Circuits* 39, 112–121.
- Raj, M., Monge, M., Emami, A., 2016. Modelling and nonlinear equalization technique for a 20 Gb/s 0.77 pJ/b VCSEL transmitter in 32 nm SOI CMOS. *IEEE Journal of Solid-State Circuits* 51, 1734–1743.
- Soref, R., Bennett, B., 1987. Electrooptical effects in silicon. *IEEE Journal of Quantum Electronics* 23, 123–129.
- Temporiti, E., Ghilioni, A., Minoia, G., *et al.*, 2016. Insights into silicon photonics Mach–Zehnder-based optical transmitter architectures. *IEEE Journal of Solid-State Circuits* 51, 3178–3191.
- Young, I., Mohammed, E., Liao, J., *et al.*, 2010. Optical I/O technology for tera-scale computing. *IEEE Journal of Solid-State Circuits* 45, 235–248.
- Yu, K., Li, C., Li, H., *et al.*, 2016. A 25 Gb/s hybrid-integrated silicon photonic source-synchronous receiver with microring wavelength stabilization. *IEEE Journal of Solid-State Circuits* 51, 2129–2141.

Foundations of Coherent Transients in Semiconductors

Torsten Meier, University of Paderborn, Paderborn, Germany

Stephan W Koch, Philipps University, Marburg, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Electromagnetic irradiation of matter generates a coherent superposition of the excited quantum states. The attribute “coherent” relates to the fact that the material excitations originally have a well defined phase dependence which is imposed by the phase of the excitation source, often a laser in the optical regime. Macroscopically, the generated superposition state can be described as an optical polarization which is determined by the transition amplitudes between the participating quantum states.

The optical polarization is a typical non-equilibrium quantity that decays to zero when a system relaxes to its equilibrium state. Coherent effects are therefore only observable in a certain time window after pulsed photoexcitation, or in the presence of a continuous-wave (cw) beam. As *coherent transients* one usually refers to phenomena that can be observed during or shortly after pulsed laser excitation and critically depend on the presence of the induced optical polarization.

Many materials such as atoms, molecules, metals, insulators, semiconductors including bulk crystals, heterostructures, and surfaces, as well as organic and biological structures are studied using coherent optical spectroscopy. Depending on the particular system, the states participating in the optical transitions, the interactions among them, and the resulting time scale for the decay of the induced polarization may be very different. As a result, the time window during which coherent effects are observable can be as long as seconds for certain atomic transitions, or it may be as short as a few femtoseconds (10^{-15} s), e.g., for metals, surfaces, or highly excited semiconductors.

Coherent spectroscopy and the analysis of coherent transients has provided valuable information on the nature and dynamics of the optical excitations. Often it is possible to learn about the interaction processes among the photoexcitations and to follow the temporal evolution of higher-order transitions which are only accessible if the system is in a non-equilibrium state.

The conceptually simplest experiment which one may use to observe coherent transients is to time resolve the transmission or reflection induced by a single laser pulse. However, much richer information can be obtained if one excites the system with several pulses. A pulse sequence with well-controlled delay times makes it possible to study the dynamical evolution of the photoexcitations not only by time resolving the signal but also by varying the delay. Prominent examples of such experiments are pump–probe measurements, which usually are performed with two incident pulses, or four-wave mixing, for which one may use two or three incident pulses.

Besides the microscopic interaction processes, the outcome of an experiment is determined by the quantum mechanical selection rules for the transitions and by the symmetries of the system under investigation. For example, if one wants to investigate coherent optical properties of surface states one often relies on phenomena, such as second-harmonic or sum-frequency generation, which give no signal in perfect systems with inversion symmetry. Due to the broken translational invariance, such experiments are therefore sensitive only to the dynamics of surface and/or interface excitations.

In this article the basic principles of coherent transients are presented and several examples are presented. The basic theoretical description and its generalization for the case of semiconductors are introduced.

Basic Principles

In the absence of free charges and currents, Maxwell’s equations show that the electromagnetic field interacts with matter via the optical polarization. This polarization $\mathbf{P}$, or more precisely its second derivative with respect to time $\frac{\partial^2}{\partial t^2}\mathbf{P}$, appears as a source term in the wave equation for the electric field $\mathbf{E}$. Consequently, if the system is optically thin such that propagation effects within the sample can be ignored and if measurements are performed in the far field region, i.e., at distances that significantly exceed the characteristic optical wavelength λ , the emitted electric field resulting from the polarization is proportional to its second time derivative, $\mathbf{E} \propto \frac{\partial^2}{\partial t^2}\mathbf{P}$. Thus the measurement of the emitted field dynamics yields information about the temporal evolution of the optical material polarization.

Microscopically the polarization is determined by the transition amplitudes between the different states of the system. These may be the discrete states of atoms or molecules, or the microscopic valence and conduction band states in a dielectric medium, such as a semiconductor. In any case, the macroscopic polarization $\mathbf{P}$ is computed by summing over all microscopic transitions p_{cv} via $\mathbf{P} = \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv} + c.c.)$, where $\boldsymbol{\mu}_{cv}$ is the dipole matrix element which determines the strength of the transitions between the states v and c , and $c.c.$ denotes the complex conjugate. If ε_c and ε_v are the energies of these states, their dynamic quantum mechanical evolution is described by the phase factors $e^{-i\varepsilon_c t/\hbar}$ and $e^{-i\varepsilon_v t/\hbar}$, respectively. Therefore each p_{cv} is evolving in time according to $e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar}$. Assuming that we start at $t=0$ with $p_{cv}(t=0) = p_{cv,0}$, which may be induced by a short optical pulse, we have for the optical polarization $\mathbf{P}(t) = \Theta(t) \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv,0} e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar} + c.c.)$. Thus $\mathbf{P}(t)$ is given by a summation over microscopic transitions which all oscillate with frequencies proportional to the energy differences between the involved states. Hence, the optical

polarization is clearly a coherent quantity which is characterized by amplitude and phase. Furthermore, the microscopic contributions to $\mathbf{P}(t)$ add up coherently. Depending on the phase relationships, one may obtain either constructive superposition, interference phenomena like quantum beats, or destructive interference leading to a decay (dephasing) of the macroscopic polarization.

Optical Bloch Equations

The dynamics of photoexcited systems can be conveniently described by a set of equations, the optical Bloch equations, named after Felix Bloch (1905–1983) who first formulated such equations to analyze the spin dynamics in nuclear magnetic resonance. For the simple case of a two-level model the Bloch equations can be written as

$$i\hbar \frac{\partial}{\partial t} p = \Delta\varepsilon p + \mathbf{E} \cdot \boldsymbol{\mu} I \quad (1)$$

$$i\hbar \frac{\partial}{\partial t} I = 2\mathbf{E} \cdot \boldsymbol{\mu} (p - p^*) \quad (2)$$

Here, $\Delta\varepsilon$ is the energy difference and I the inversion, i.e., the occupation difference between upper and lower state. The field $\mathbf{E}$ couples the polarization to the product of the Rabi energy $\mathbf{E} \cdot \boldsymbol{\mu}$ and the inversion I . In the absence of the driving field, i.e., $\mathbf{E} \equiv 0$, Eq. (1) describes the free oscillation of $p \propto e^{-i\Delta\varepsilon t/\hbar}$ discussed above.

The inversion is determined by the combined action of the Rabi energy and the transition p . The total occupation N , i.e., the sum of the occupations of the lower and the upper states, remains unchanged during the optical excitation since light does not create or destroy electrons, it only transfers them between different states. If N is initially normalized to 1, then I can vary between -1 and 1 . $I = -1$ corresponds to the ground state of the system, where only the lower state is occupied. In the opposite extreme, i.e., for $I = 1$, the occupation of the upper state is 1 and the lower state is completely depleted.

One can show that Eqs. (1) and (2) contain another conservation law. Introducing $\mathcal{P} = p + p^* = 2\text{Re}[p]$ and $\mathcal{J} = i(p - p^*) = -2\text{Im}[p]$, which are real quantities and often named polarization and polarization current, respectively, one finds that the relation

$$\mathcal{P}^2 + \mathcal{J}^2 + I^2 = 1 \quad (3)$$

is fulfilled. Thus the coupled coherent dynamics of the transition and the inversion can be described on a unit sphere, the so-called Bloch sphere. The three terms $\mathcal{P}$, $\mathcal{J}$, and I define the components of the three-dimensional Bloch vector $\mathbf{S} = (\mathcal{P}, \mathcal{J}, I)$. With these definitions, one can reformulate Eqs. (1) and (2) as

$$\frac{\partial}{\partial t} \mathbf{S} = \boldsymbol{\Omega} \times \mathbf{S} \quad (4)$$

with $\boldsymbol{\Omega} = (-2\boldsymbol{\mu} \cdot \mathbf{E}, 0, \Delta\varepsilon)/\hbar$ and $\times$ denoting the vector product.

The structure of Eq. (4) is mathematically identical to the equations describing either the angular momentum dynamics in the presence of a torque or the spin dynamics in a magnetic field. Therefore, the vector $\mathbf{S}$ is often called pseudospin. Moreover, many effects which can be observed in magnetic resonance experiments, e.g., free decay, quantum beats, echoes, etc. have their counterparts in photoexcited optical systems.

The vector product on the right hand side of Eq. (4) shows that $\mathbf{S}$ is changed by $\boldsymbol{\Omega}$ in a direction perpendicular to $\mathbf{S}$ and $\boldsymbol{\Omega}$. For vanishing field the torque $\boldsymbol{\Omega}$ has only a z -component. In this case the z -component of $\mathbf{S}$, i.e., the inversion, remains constant, whereas the x - and y -components, i.e., the polarization and the polarization current, oscillate with the frequency $\Delta\varepsilon/\hbar$ on the Bloch sphere around $\boldsymbol{\Omega}$.

Semiconductor Bloch Equations

If one wants to analyze optical excitations in crystalline solids, i.e., in systems which are characterized by a periodic arrangement of atoms, one has to include the continuous dispersion (bandstructure) into the description. This can be done by including the crystal momentum $\hbar\mathbf{k}$ as an index to the quantities which describe the optical excitation of the material, e.g., p and I in Eqs. (1) and (2). Hence, one has to consider a separate two-level system for each crystal momentum $\hbar\mathbf{k}$. As long as all other interactions are disregarded the bandstructure simply introduces a summation over the uncoupled contributions from the different $\mathbf{k}$ states, leading to inhomogeneous broadening due to the presence of a range of transition frequencies $\Delta\varepsilon(\mathbf{k})/\hbar$.

For any realistic description of optical processes in solids, it is essential to go beyond the simple picture of non-interacting states and to treat the interactions among the elementary material excitations, e.g., the Coulomb interaction between the electrons and the coupling to additional degrees of freedom, such as lattice vibrations (electron-phonon interaction) or other bath-like subsystems. If crystals are not ideally periodic, imperfections, which can often be described as a disorder potential, need to be considered as well.

All these effects can be treated by proper extensions of the optical Bloch equations introduced above. For semiconductors the resulting generalized equations are known as semiconductor Bloch equations, where the microscopic interactions are included at a

certain level of approximation. For a two band model of a semiconductor, these equations can be written schematically as

$$i\hbar \frac{\partial}{\partial t} p_k = \Delta \varepsilon_k p_k + \Omega_k (n_k^c - n_k^v) + i\hbar \frac{\partial}{\partial t} p_k|_{corr} \quad (5)$$

$$i\hbar \frac{\partial}{\partial t} n_k^c = (\Omega_k^* p_k - \Omega_k p_k^*) + i\hbar \frac{\partial}{\partial t} n_k^c|_{corr} \quad (6)$$

$$i\hbar \frac{\partial}{\partial t} n_k^v = -(\Omega_k^* p_k - \Omega_k p_k^*) + i\hbar \frac{\partial}{\partial t} n_k^v|_{corr} \quad (7)$$

Here p_k is the microscopic polarization and n_k^c and n_k^v are the electron populations in the conduction and valence bands (c and v), respectively. Due to the Coulomb interaction and possibly further processes, the transition energy $\Delta \varepsilon_k$ and the Rabi energy Ω_k both depend on the excitation state of the system, i.e., they are functions of the time-dependent polarizations p_k and populations $n_k^{c/v}$. This leads, in particular, to a coupling among the excitations for all different values of the crystals momentum $\hbar \mathbf{k}$. Consequently, in the presence of interactions the optical excitations can no longer be described as independent two-level systems but have to be treated as a coupled many-body system. A prominent and important example in this context is the appearance of strong exciton resonances which, as a consequence of the Coulomb interaction, show up in the absorption spectra of semiconductors energetically below the fundamental band gap.

The interaction effects lead to significant mathematical complications since they induce couplings between all the different quantum states of a system and introduce an infinite hierarchy of equations for the microscopic correlation functions. The terms given explicitly in Eqs. (5)–(7) arise in a treatment of the Coulomb interaction on the Hartree-Fock level. Whereas this level is sufficient to describe excitonic resonances, there are many additional effects, e.g., excitation-induced dephasing due to Coulomb scattering and significant contributions from higher-order correlations like excitonic populations and biexcitonic resonances, which make it necessary to treat many-body correlation effects that are beyond the Hartree-Fock level. These contributions are formally included by the terms denoted as $|_{corr}$ in Eqs. (5)–(7). The systematic truncation of the many-body hierarchy and the analysis of controlled approximations is the basic problem in the microscopic theory of optical processes in condensed matter systems.

Examples

Radiative Decay

The emitted field originating from the non-equilibrium coherent polarization of a photoexcited system can be monitored by measuring the transmission and the reflection as function of time. If only a single isolated transition is excited, the dynamic evolution of the polarization and therefore of the transmission and reflection is governed by radiative decay. This decay is a consequence of the coupling of the optical transition to the light field described by the combination of Maxwell- and Bloch-equations. Radiative decay simply means that the optical polarization is converted into a light field on a characteristic time scale $2T_{rad}$. Here, T_{rad} is the population decay time, often also denoted as T_1 -time. This radiative decay is a fundamental process which limits the time on which coherent effects are observable for any photoexcited system. Due to other mechanisms, however, the non-equilibrium polarization often vanishes considerably faster.

The value of T_{rad} is determined by the dipole matrix element and the frequency of the transition, i.e., $T_{rad}^{-1} \propto |\mu|^2 \Delta \omega$, with $\Delta \omega = \Delta \varepsilon / \hbar$. The temporal evolution of the polarization and the emitted field are proportional to $e^{-i\Delta \omega t - t/(2T_{rad})}$. Usually one measures the intensity of a field, i.e., its squared modulus, which evolves as $e^{-t/T_{rad}}$ and thus simply shows an exponential decay, the so-called *free-induction decay*, see Fig. 1. T_{rad} can be very long for transitions with small matrix elements. For semiconductor quantum wells, however, it is on the order of only 10 ps, as the result of the strong light-matter interaction.

The time constant on which the optical polarization decays is often called T_2 . In the case that this decay is dominated by radiative processes we thus have $T_2 = 2T_{rad}$.

Superradiance

The phenomenon of superradiance can be discussed considering an ensemble of N two-level systems which are localized at certain positions $\mathbf{R}_i$. In this case Maxwell's equations introduce a coupling among all these resonances since the field emitted from any specific resonance interacts with all other resonances and interferes with their emitted fields. As a result, this system is characterized by N eigenmodes originating from the radiatively coupled optical resonances.

A very spectacular situation arises if one considers N identical two-level systems regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance, i.e., $\lambda = c/\Delta \omega$, where c is the speed of light in the considered material, see Fig. 2(a). In this case all emitted fields interfere constructively and the system behaves effectively like a single two-level system with a matrix element increased by $\sqrt{N}$. Consequently, the radiative decay rate is increased by N and the polarization of the coupled system decays N -times faster than that of an isolated system, see Fig. 2(b). This effect is called superradiance.

It is possible to observe superradiant coupling effects, e.g., in suitably designed semiconductor heterostructures. Fig. 2(c) compares the measured time-resolved reflection from a single quantum well (dashed line) with that of a Bragg structure, i.e., a multiple quantum-well structure which consists of 10 wells that are separated by $\lambda/2$ (solid line), where λ is the wavelength of the

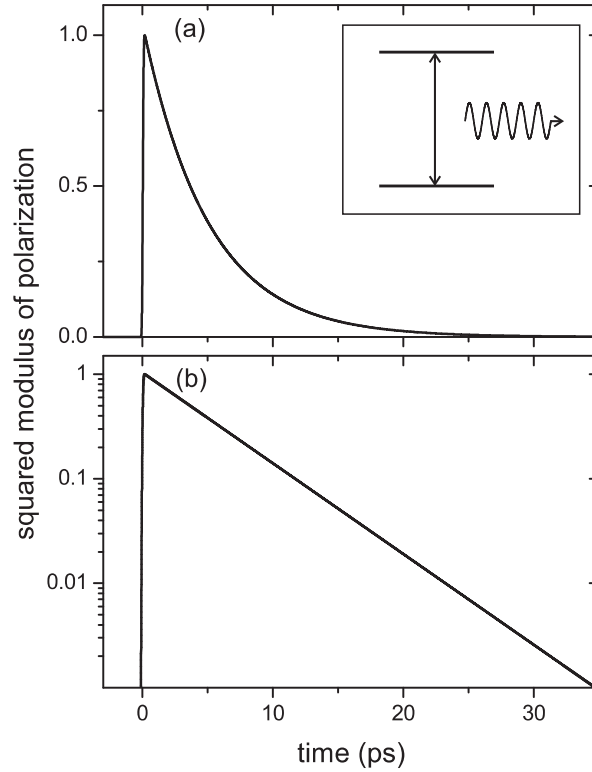


Fig. 1 A two-level system is excited by a short optical pulse at $t=0$. Due to radiative decay (inset) and possibly other dephasing mechanisms, the polarization decays exponentially as function of time with the time constant T_2 , the dephasing time. The intensity of the optical field which is emitted as a result of this decay is proportional to the squared modulus of the polarization, which falls off with the time constant $T_2/2$. The decay of the squared modulus of the polarization is shown in (a) on a linear scale and in (b) on a logarithmic scale. The dephasing time was chosen to be $T_2=10\text{ps}$ which models the radiative decay of the exciton transition of a semiconductor quantum well.

exciton resonance. For times greater than about 2ps the direct reflection of the exciting pulse has decayed sufficiently and one is left with the exponential decay of the remaining signal. **Fig. 2(c)** shows that this decay is much more rapid for the Bragg structure than for the single quantum well due to superradiance introduced by the radiative coupling among the quantum wells.

Destructive Interference

We now consider a distribution of two-level systems which have slightly different transition frequencies characterized by the distribution function $g(\Delta\omega - \bar{\omega})$ of the transition frequencies which is peaked at the mean value $\bar{\omega}$ and has a spectral width of $\delta\omega$, see **Fig. 3(a)**. Ignoring the radiative coupling among the resonances, the optical polarization of the total system evolves after excitation by a short laser pulse at $t=0$ proportional to $\int d\omega g(\Delta\omega - \bar{\omega}) e^{-i\Delta\omega t} \propto \tilde{g}(t) e^{-i\bar{\omega}t}$, where $\tilde{g}(t)$ denotes the Fourier transform of the frequency distribution function $g(\omega)$ and thus the optical polarization decays on a time scale which is inversely proportional to the spectral width of the distribution function $\delta\omega$. Thus the destructive interference of many different transitions frequencies results in a rapid decay of the polarization, see **Fig. 3(b)**.

In the spectral domain this rapid decay shows up as inhomogeneous broadening. Depending on the system under investigation, there are many different sources for such an inhomogeneous broadening. One example is the Doppler broadening in atomic gases or disorder effects such as well-width fluctuations in semiconductor quantum wells or lattice imperfections in crystals.

In the nonlinear optical regime it is under certain circumstances possible to reverse the destructive interference of inhomogeneously broadened coherent polarizations. For example, in four-wave mixing a second pulse may lead to a rephasing of the contributions with different frequencies, which results in the photon echo, see discussion below.

Quantum Beats

The occurrence of quantum beats can be understood most easily in a system where the total optical polarization can be attributed to a finite number of optical transitions. Let us assume for simplicity that all these transitions have the same matrix element. In this case, after excitation with a short laser pulse at $t=0$ the optical polarization of the total system evolves proportional to $\sum_i e^{-i\Delta\omega_i t}$. The finite number of frequencies results in a temporal modulation with time periods $2\pi/(\Delta\omega_i - \Delta\omega_j)$ of the squared modulus of the polarization which is proportional to the emitted field intensity. For the case of two frequencies the squared modulus of the

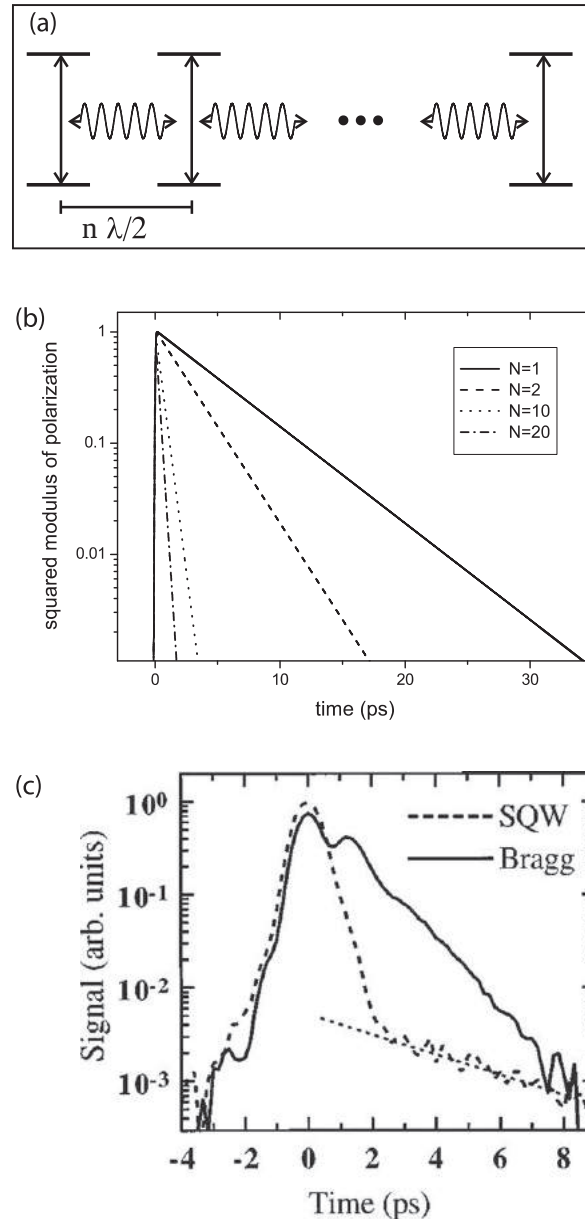


Fig. 2 (a) N identical two-level systems are depicted which are regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance. Due to their coupling via Maxwell's equations all fields emitted from the two-level systems interfere constructively and the coupled system behaves effectively like a single two-level system with a optical matrix element increased by $\sqrt{N}$. (b) The temporal decay of the squared modulus of the polarization is shown on a logarithmic scale. The radiative decay rate is proportional to N and thus the polarization of the coupled system decays N -times faster than that of an isolated system. This effect is called superradiance. (c) Measured time-resolved reflection of a semiconductor single quantum well (dashed line) and a $N=10$ Bragg structure, i.e., a multi quantum well where the individual wells are separated by $\lambda/2$ (solid line), on a logarithmic scale. Part (c) is taken from Fig. 5 of Haas, S., Stroucken, T., Hübner, M., *et al.* 1998. Phys. Rev. B 57, 14860.

polarization is proportional to $[1 + \cos((\Delta\omega_1 - \Delta\omega_2)t)]$, i.e., due to the interference of two contributions the polarization varies between a maximum and zero, see Fig. 4(b).

In the linear optical regime it is impossible to distinguish whether the optical transitions are uncoupled or coupled. As shown in Fig. 4, two uncoupled two-level systems give the same linear polarization as a three-level system where the two transitions share a common state. It is, however, possible to decide about the nature of the underlying transitions if one performs nonlinear optical spectroscopy. This is due to the fact that the so-called quantum beats, i.e., the temporal modulations of the polarization of an intrinsically coupled system, show a different temporal evolution as the so-called polarization interferences, i.e., the temporal modulations of the polarization of uncoupled systems. The former ones are also much more stable in the presence of inhomogeneous broadening than the latter ones.

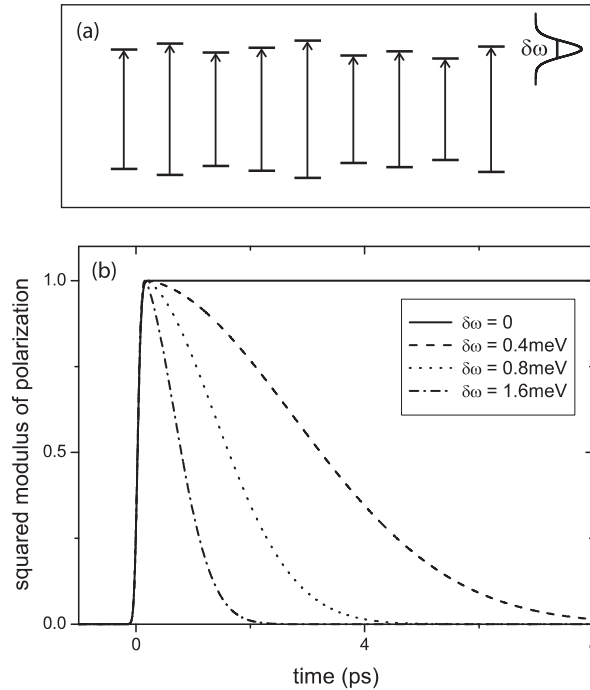


Fig. 3 (a) An ensemble of two-level systems is shown where the resonance frequencies are randomly distributed according to a Gaussian distribution function of width $\delta\omega$ around an average value. (b) The temporal dynamics of the squared modulus of the polarization of ensembles of two-level systems after excitation with a short optical pulse at $t=0$ is shown on a linear scale. Since the dephasing time is set to infinity, i.e., $T_2 \rightarrow \infty$, the polarization of an ensemble of identical two-level system ($\delta\omega=0$) does not decay. However, for a finite width of the distribution, $\delta\omega > 0$, the individual polarizations of the ensemble oscillate with different frequencies and, therefore, due to destructive interference the polarization of the ensemble decays as function of time. Since the Fourier transform of a frequency-domain Gaussian is a Gaussian in the time domain, the dashed, dotted, and dashed dotted line have a Gaussian shape with a temporal width which is inversely proportional to $\delta\omega$.

In semiconductor heterostructures, quantum-beat spectroscopy has been widely used to investigate the temporal dynamics of excitonic resonances. Also the coupling among different optical resonances has been explored in pump-probe and four-wave-mixing measurements.

Coherent Control

In the coherent regime the polarization induced by a first laser pulse can be modified by a second, delayed pulse. For example, the first short laser pulse incident at $t=0$ induces a polarization proportional to $e^{-i\Delta\omega t}$ and a second pulse arriving at $t=\tau > 0$ induces a polarization proportional to $e^{-i(\varphi)} e^{-i\Delta\omega(t-\tau)}$. Thus for $t \geq \tau$ the total polarization is given by $e^{-i\Delta\omega t}(1 + e^{-i\Delta(\varphi)})$ with $\Delta(\varphi) = (\varphi) - \Delta\omega\tau$. If $\Delta(\varphi)$ is an integer multiple of 2π the polarizations of both pulses interfere constructively and the amplitude of the resulting polarization is doubled as compared to a single pulse. If, on the other hand, $\Delta(\varphi)$ is an odd multiple of π , the polarizations of both pulses interfere destructively and the resulting polarization vanishes after the action of both pulses. Thus by varying the phase difference of the two pulses it is possible to coherently enhance or destroy the optical polarization, see Fig. 5(b).

One can easily understand the coherent control in the frequency domain by considering the overlap between the absorption of the system and the pulse spectrum. For excitation with two pulses that are temporally delayed by τ , the spectrum of the excitation shows interference fringes with a spectral oscillation period that is inversely proportional to τ . The positions of the maxima and the minima of the fringes depend on the phase difference between the two pulses. For constructive interference, the excitation spectrum is at a maximum at the resonance of the system, whereas for destructive interference the excitation spectrum vanishes at the resonance of the system, see Fig. 5(a).

Coherent control techniques have been applied to molecules and solids to control the dynamical evolution of electronic wavepackets and also the coupling to nuclear degrees of freedom. In this context, it is sometimes possible to steer certain chemical reactions into a preferred direction by using sequences of laser pulses which can be chirped, i.e., have a time-dependent frequency.

Coherent Photocurrents

Exciting an unbiased inversion symmetric semiconductor with laser light that resonantly above the band gap does not lead to electrical currents. This is due to the fact, that in such a situation one excites the same amount of electrons with positive as with negative (crystal) momentum $\hbar\mathbf{k}$, see Fig. 6(a). Therefore the electrons move with the same average velocity in any direction and

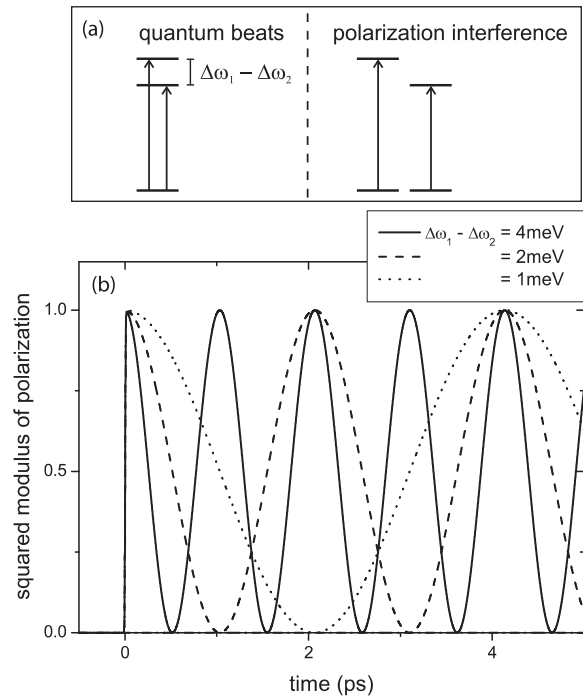


Fig. 4 (a) Two optical resonances with frequency difference $\Delta\omega_1 - \Delta\omega_2$ may be realized by either a three-level system or by two uncoupled two-level systems. After impulsive excitation with an optical pulse the linear polarization of both types of systems show a modulation of the squared modulus of the polarization with the time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$. For the intrinsically coupled three-level system these modulations are called quantum beats, whereas for an interference of uncoupled systems they are named polarization interferences. Using nonlinear optical techniques, e.g., four-wave mixing and pump probe, it is possible to distinguish quantum beats and polarization interferences since coupled and uncoupled systems have different optical nonlinearities. (b) The temporal dynamics of the squared modulus of the polarization is shown for systems with two resonances and frequency difference $\Delta\omega_1 - \Delta\omega_2$, neglecting dephasing, i.e., setting $T_2 \rightarrow \infty$. After excitation with a short optical pulse at $t=0$ the squared modulus of the polarization is periodically modulated with a time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$.

consequently no net photocurrent is generated since the currents in opposite directions cancel each other. It is, however, possible to coherently generate photocurrents by nonlinear optical excitation with laser fields that contain two frequencies ω and 2ω which fulfill $\hbar\omega < E_{\text{gap}}$ and $2\hbar\omega > E_{\text{gap}}$ where E_{gap} is the band gap. Such two-color laser fields couple the initial (valence band) and the final (conduction band) states by two quantum-mechanical excitation pathways, i.e., a single-photon transition of the 2ω field and a two-photon transition of the ω field. With this scheme it is possible to coherently control the amount of excitation by the phase difference between the ω and the 2ω fields and to realize asymmetric distributions of electrons and holes in momentum space, see Fig. 6(b), which correspond to the generation of photocurrents.

Such electrical currents generated by two-color laser excitation do not rely on any specific requirements or symmetry of the material and have been observed in several systems. Using the typical selection rules for III-V semiconductors it is possible to show that linearly polarized ω and the 2ω fields generate electrical currents if they are parallel polarized whereas orthogonal polarization directions lead to pure spin currents, i.e., a motion of spin-up and spin-down electrons in opposite directions.

Due to scattering processes, the optically-generated asymmetric momentum-space distributions relax rapidly and consequently the photocurrents decay on short, typically femtosecond, time scales. For the case of a quantum well, see Fig. 6(c), microscopic calculations including Coulomb and electron-phonon scattering predict decay times on the order of 100–200 fs for the charge and spin currents.

Transient Absorption Changes

In a typical pump-probe experiment one excites the system with a pump pulse (E_p) and probes its dynamics with a weak test pulse (E_t), see Fig. 7. With such experiments one often measures the differential absorption $\Delta\alpha(\omega)$ which is defined as the difference between the probe absorption in the presence of the pump $\alpha_{\text{pump on}}(\omega)$ and the probe absorption without the pump $\alpha_{\text{pump off}}(\omega)$.

For resonant pumping and for a situation where the pump precedes the test (positive time delays $\tau > 0$), the absorption change is usually negative in the vicinity of the resonance frequency $\Delta\omega$ indicating the effect of absorption bleaching, see Fig. 8(a) and (b). There may be positive contributions spectrally around the original absorption line due to resonance broadening and, at other spectral positions, due to excited state absorption, i.e., optical transitions to energetically higher states which are only possible if the system is in an excited state. The bleaching and the positive contributions are generally present in coherent and also in incoherent situations, where the polarization vanishes but occupations in excited states are present.

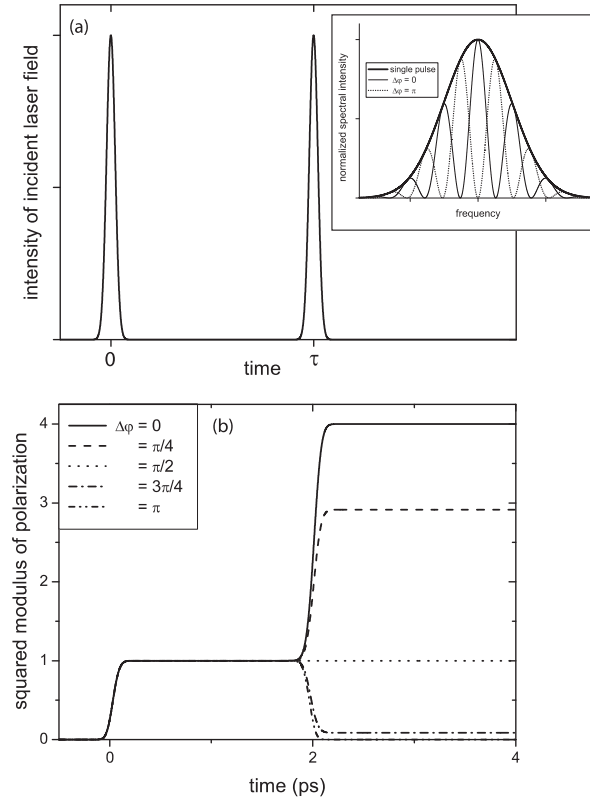


Fig. 5 (a) The interference of two Gaussian laser pulses separated by the time delay τ depends on their phase difference. As shown by the thick solid line in the inset, the spectral intensity of a single pulse is a Gaussian with a maximum at the central frequency of the pulse. The width of this Gaussian is inversely proportional to the duration of the pulse. The spectral intensity of the field consisting of both pulses shows interference fringes, i.e., is modulated with a spectral period that is inversely proportional to τ . As shown by the thin solid and the dotted lines in the inset, the phase of the interference fringes depends on the phase difference $\Delta(\phi)$ between the two pulses. Whereas for $\Delta(\phi)=0$ the spectral intensity has a maximum at the central frequency, it vanishes at this position for $\Delta(\phi)=\pi$. (b) The temporal dynamics of the squared modulus of the polarization is shown for a two-level system excited by a pair of short laser pulses neglecting dephasing, i.e., setting $T_2 \rightarrow \infty$. The first pulse excites at $t=0$ an optical polarization with a squared modulus normalized to 1. The second pulse, which has the same intensity as the first one, also excites an optical polarization at $t=2$ ps. For $t > 2$ ps the total polarization is given by the sum of the polarizations induced by the two pulses. Due to interference, the squared modulus of the total polarization depends sensitively on the phase difference $\Delta(\phi)$ between the two pulses. For constructive interference the squared modulus of the total polarization after the second pulse is four times bigger than that after the first pulse, whereas it vanishes for destructive interference. One may achieve all values in between these extremes by using phase differences $\Delta(\phi)$ which are no multiples of π . It is thus shown that the second pulse can be used to coherently control the polarization induced by the first pulse.

For detuned pumping the resonance may be shifted by the light field, as, e.g., in the optical Stark effect. Depending on the excitation configuration and the system, this transient shift may be to higher (blue shift) or lower energies (red shift), see Fig. 8(a) and (b). With increasing pump–probe time delay the system gradually returns to its unexcited state and the absorption changes disappear.

As an illustration we show in Fig. 8(c) experimentally measured differential absorption spectra of a semiconductor quantum well which is pumped spectrally below the exciton resonance. For a two-level system one would expect a blue shift of the absorption, i.e., a dispersive shape of the differential absorption with positive contributions above and negative contributions below the resonance frequency. This is indeed observed for most polarization directions of the pump and probe pulses. However, if both pulses are oppositely circularly polarized, the experiment shows a resonance shift in the reverse direction, i.e., a red shift. For the explanation of this behavior one has to consider the optical selection rules of the quantum well system. One finds that the signal should actually vanish for oppositely circularly polarized pulses, as long as many-body correlations are neglected. Thus in this case the entire signal is due to many-body correlations and it is their dynamics which gives rise to the appearance of the red shift.

Spectral Oscillations

For negative time delays $\tau < 0$, i.e., if the test pulse precedes the pump, the absorption change $\Delta\alpha(\omega)$ is characterized by spectral oscillations around $\Delta\omega$ which vanish as τ approaches zero, see Fig. 9(a). The spectral period of the oscillations decreases with increasing $|\tau|$. Fig. 9(b) shows measured differential transmission spectra of a multiple quantum-well structure for different negative time delays. As the differential absorption also the differential transmission spectra are dominated by spectral oscillations around the exciton resonance whose period and amplitude decrease with increasing $|\tau|$.

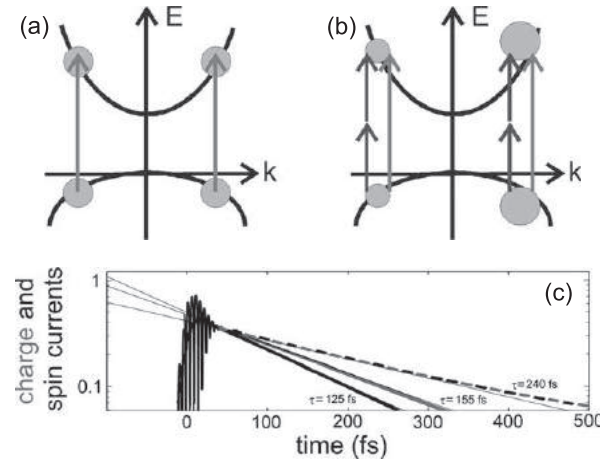


Fig. 6 (a) Schematic illustration of the electron and hole distributions in momentum space that are generated in the conduction and valence band, respectively, by an above band gap optical excitation. Due to the symmetry of the distributions the optical excitation does not lead to a photocurrent. (b) Schematic illustration of the electron and hole distributions in momentum space that are generated by the excitation with a two-color laser field containing the frequencies ω and 2ω . Since the valence and conduction band states are coupled by two quantum-mechanical pathways, i.e., a single and a two-photon transitions, it is possible to coherently control the excitations and to generate asymmetric electron and hole distributions which correspond to the generation of photocurrents. (c) Calculated dynamics of photocurrents that are excited in a GaAs quantum well 150 meV above the band gap by a 20 fs two-color laser field on a logarithmic scale. The simultaneous excitation by the ω and 2ω fields leads to rapid oscillations during the duration of the laser beams. After the excitation, the currents decay approximately exponentially on a 100 – 200 fs time scale. If only electron-LO-phonon scattering is considered, both the charge and spin currents decay with the same time constant, see dashed line. When also electron-electron scattering is included in the analysis, the spin current (dark grey line) decays more rapidly than the charge current (light grey line). Part (c) is taken from Fig. 2(a) of Duc, H.T., Meier, T., Koch, S.W., 2005. Phys. Rev. Lett. 95, 086606.

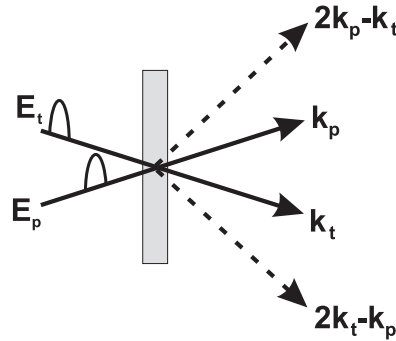


Fig. 7 Transient optical nonlinearities can be investigated by exciting the system with two time-delayed optical pulses. The pump pulse (E_p) puts the system in an excited state and the test (probe) pulse (E_t) is used to measure its dynamics. The dynamic nonlinear optical response may be measured in the transmitted or reflected directions of the pump and test pulses, i.e., $\pm k_p$ and $\pm k_t$, respectively. In a pump-probe experiment one often investigates the change of the test absorption induced by the pump pulse, by measuring the absorption in the direction k_t with and without E_p . One may also measure the dynamic nonlinear optical response in scattering directions like $2k_p - k_t$ and $2k_t - k_p$ as indicated by the dashed lines. This is what is done in four-wave-mixing and photon-echo experiments.

The physical origin of the coherent oscillations is the pump-induced perturbation of the free-induction decay of the polarization generated by the probe pulse. This perturbation is delayed by the time τ at which the pump arrives. The temporal shift of the induced polarization changes in the time domain leads to oscillations in the spectral domain, since the Fourier transformation translates a delay in one domain into a phase factor in the conjugate domain.

Photon Echo

In the nonlinear optical regime one may (partially) reverse the destructive interference of a coherent, inhomogeneously broadened polarization. For example, in four-wave mixing, which is often performed with two incident pulses, one measures the emitted field in a background-free scattering direction, see Fig. 5. The first short laser pulse excites all transitions at $t=0$. As a result of the inhomogeneous broadening the polarization decays due to destructive interference, see Fig. 3. The second pulse arriving at $t=\tau>0$ is able to conjugate the phases ($e^{i(\varphi)} \rightarrow e^{-i(\varphi)}$) of the individual polarizations of the inhomogeneously broadened system. The subsequent unperturbed dynamical evolution of the polarizations leads to a measurable macroscopic signal at $t=2\tau$. This photon

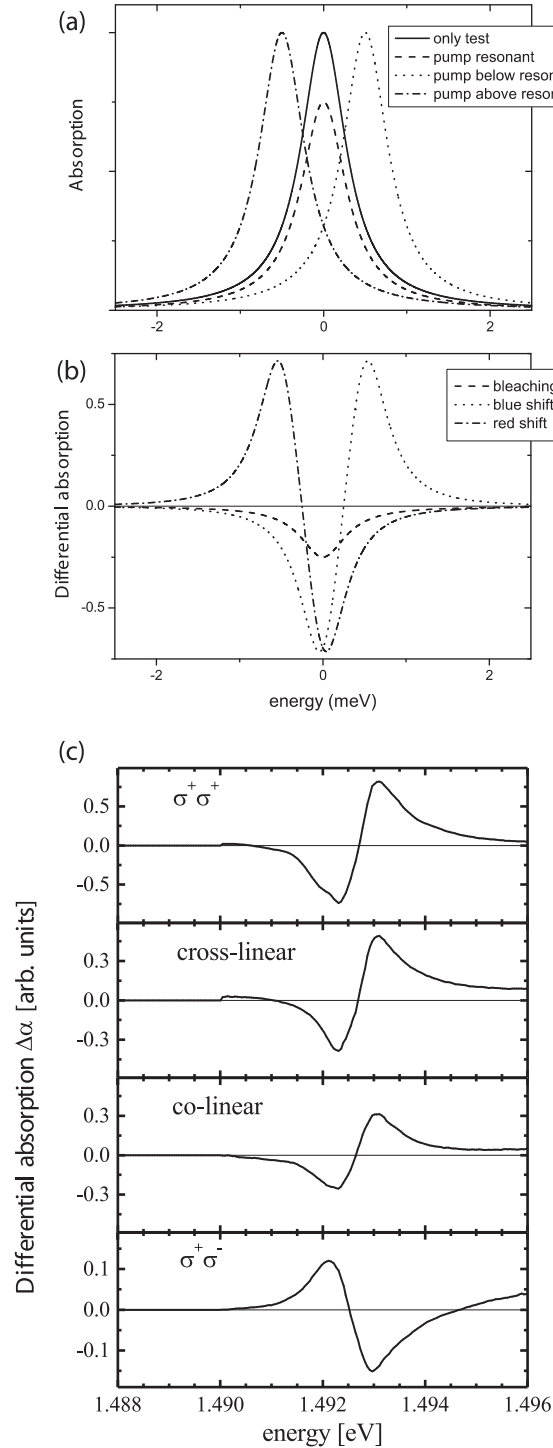


Fig. 8 (a) Absorption spectra of a two-level system. The solid line shows the linear optical absorption spectrum as measured by a weak test pulse. It corresponds to a Lorentzian line that is centered at the transition frequency $\Delta\omega$, which has been set to zero. The width of the line is inversely proportional to the dephasing time T_2 . In the presence of a pump pulse which resonantly excited the two-level system, the absorption monitored by the test pulse is reduced in amplitude, i.e., bleached, since the pump puts the system in an excited state (dashed line). In the optical Stark effect the frequency of the pump is non-resonant with the transition frequency. If the pump is tuned below (above) the transition frequency, the absorption is shifted to higher (lower) frequencies, i.e., shifted to the blue (red) part of the spectrum, see dotted (dashed-dotted) line. (b) The differential absorption obtained by taking the difference between the absorption in the presence of the pump and the absorption without pump. The dashed line shows the purely negative bleaching obtained using a resonant pump. The dotted and the dashed-dotted lines correspond to the dispersive shape of the blue and red shift obtained when pumping below and above the resonance, respectively. (c) Measured differential absorption spectra of a semiconductor quantum well which is pumped off-resonantly 4.5 meV below the 1s heavy-hole exciton resonance. The four lines correspond to different polarization directions of the pump and probe pulses as indicated. Part (c) is taken from Fig. 1 of Sieh, C., Meier, T., Jahnke, F., *et al.*, 1999. Phys. Rev. Lett. 82, 3112.

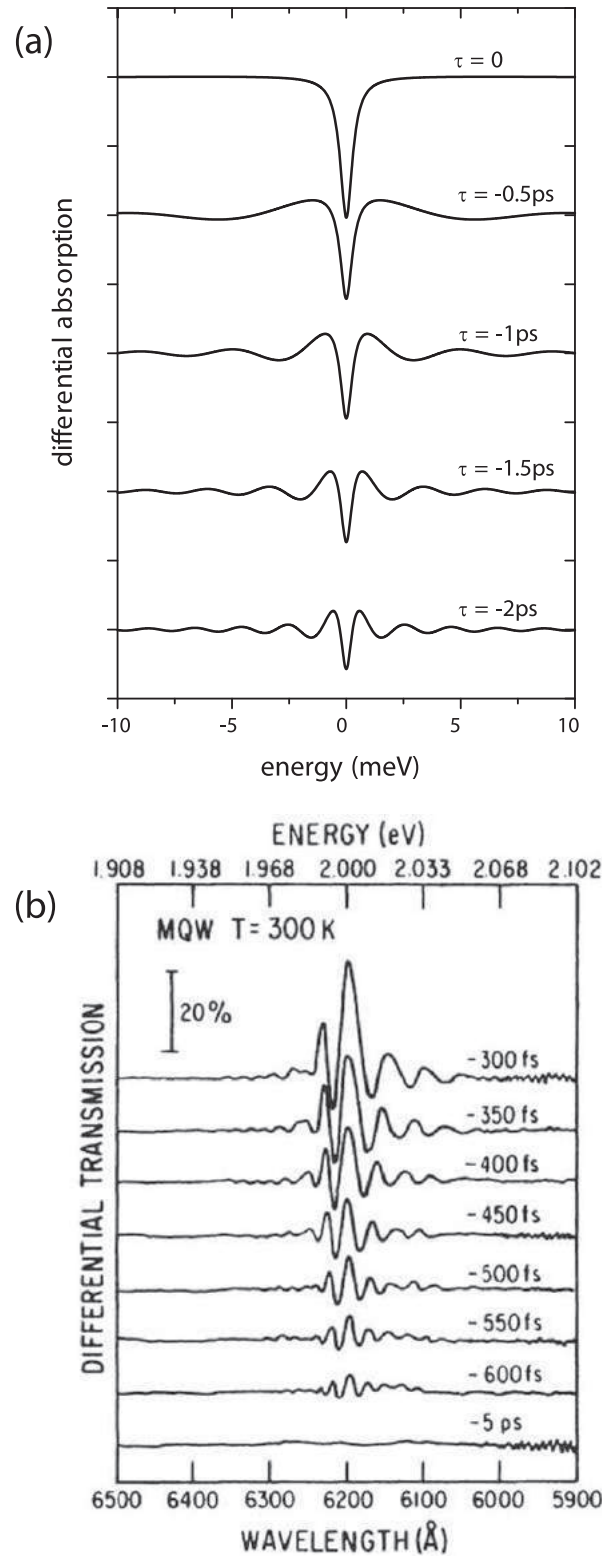


Fig. 9 (a) Differential absorption spectra are shown for resonant pumping of a two-level system for various time delays τ between the pump and the test pulse. When the pump precedes the test, i.e., $\tau < 0$, the differential absorption exhibits spectral oscillations with a period which is inversely proportional to the time delay. When τ approaches zero, these spectral oscillations vanish and the differential absorption develops into purely negative bleaching. (b) Measured differential transmission spectra of a multi quantum well for different negative time delays as indicated. Part (b) is taken from Fig. 3 of Sokoloff, J.K., Joffe, M., Fluegel, B., *et al.* 1988. Phys. Rev. B 38, 7615.

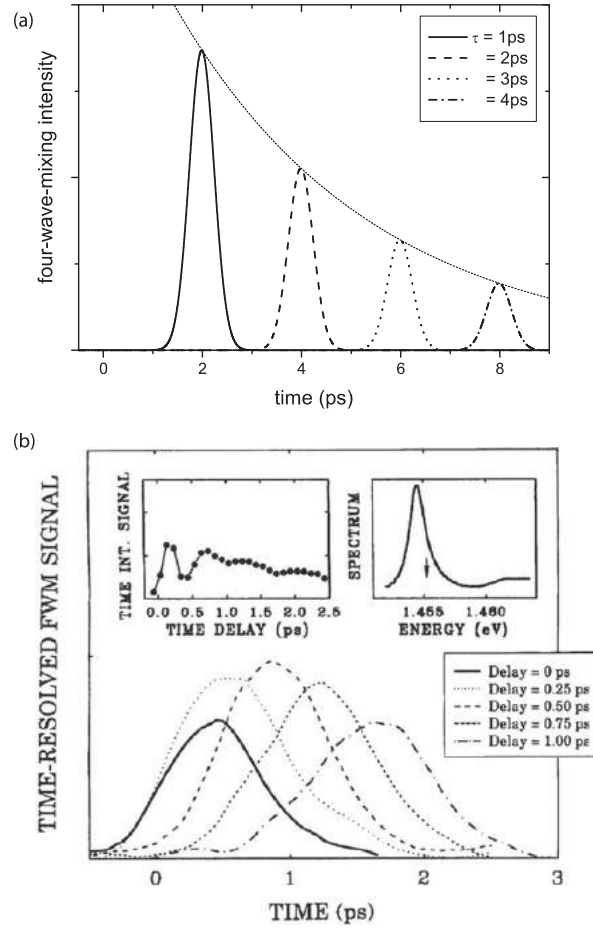


Fig. 10 (a) The intensity dynamics as measured in a four-wave-mixing experiment on an ensemble of inhomogeneously broadened two-level systems for various time delays τ of the two incident pulses. The first pulse excites the system at $t=0$. In the time period $0 < t < \tau$ the individual polarizations of the inhomogeneously broadened system oscillate with their respective frequencies and due to destructive interference the total polarization decays, cp. Fig. 3. The second pulse arriving at $t=\tau$ leads to a partial rephasing of the individual polarizations. Due to phase conjugation all individual polarizations are in phase at $t=2\tau$ which results in a macroscopic measurable signal, the photon echo. Due to dephasing processes the magnitude of the photon echo decays with increasing τ , see solid, dashed, dotted and dashed-dotted lines. Describing the dephasing via a T_2 time corresponds to an exponential decay as indicated by the thin dotted line. By measuring the decay of the echo with increasing τ , one thus gets experimental information on the dephasing of the optical polarization. (b) Time-resolved four-wave-mixing signal measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays as indicated. The origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t=\tau$. The insets show the corresponding time-integrated four-wave-mixing signal and the linear absorption. Part (b) is taken from Fig. 1 of Jahnke, F., Koch, M., Meier, T., *et al.* 1994. Phys. Rev. B 50, 8114.

echo occurs since at this point in time all individual polarizations are in phase and add up constructively, see Fig. 10(a). Since this rephasing process leading to the echo is only possible as long as the individual polarizations remain coherent, one can analyze the loss of coherence (dephasing) by measuring the decay of the photon echo with increasing time delay.

Time-resolved four-wave-mixing signals measured for different time delays on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure are presented in Fig. 10(b). Compared to Fig. 10(a), the origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t=\tau$. Due to the inhomogeneous broadening the maximum of the signal is shifting to longer times with increasing delay. Further details of the experimental results, in particular, the exact position of the maximum and the width of the signal, cannot be explained on the basis of non-interacting two-level systems, but require the analysis of many-body effects in the presence of inhomogeneous broadening.

Multidimensional Fourier Transform Spectroscopy

Exciting a semiconductor with N short laser pulses from different directions may lead to a nonlinear signal that depends on real time and $N-1$ time delays between the incident pulses. The idea of multidimensional Fourier transform spectroscopy is to Fourier transform the time-domain signal with respect to some or all of the time and time delay arguments to the frequency domain.

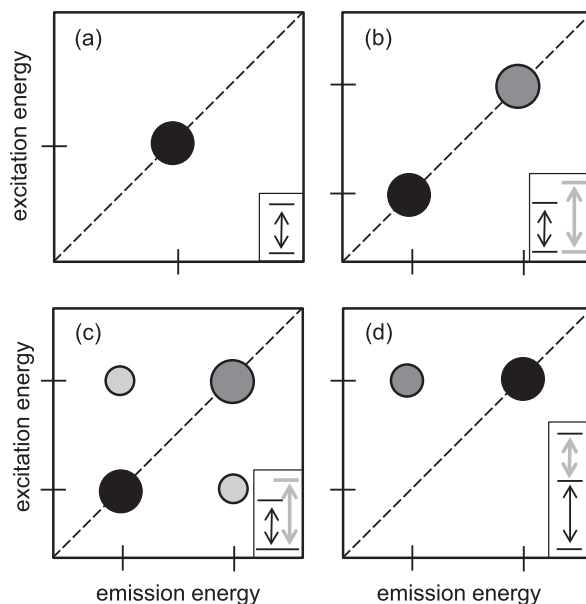


Fig. 11 Schematic illustration of two-dimensional Fourier-transform spectroscopy signals for a few-level systems. (a) For the case of a single two-level system the absorption and emission take place at the same frequency which leads to a single peak on the diagonal of the two-dimensional signal. (b) Two uncoupled two-level systems absorb and emit at two frequencies and therefore correspond to two peaks on the diagonal. (c) Similarly to case (b), a three-level system can also absorb and emit at two frequencies. However, for a three-level system the coupling among the transitions via the common ground state leads to the appearance of two additional off-diagonal peaks. (d) When the excitation to an energetically higher state is only possible if the system is already excited, this transition frequency appears as an additional off-diagonal peak in the emission. Such a situation can be realized in systems that contain, e.g., exciton and biexciton resonances.

The so-obtained multi-dimensional frequency-domain signal depends on up to N frequencies and contains correlations among the frequency components which arise due to the dynamics in the respective time delays. Such multi-dimensional signals have been shown to provide a transparent way to unravel couplings and interaction processes that determine the complex nonlinear dynamics of several investigated systems.

Fourier transform spectroscopy is frequently performed in four-wave-mixing set-ups where the system is excited by two or three short laser pulses. Fourier transforming with respect to time and time delays provides signals that depend on two or three frequency arguments corresponding to two- or three-dimensional spectroscopy, respectively. A detailed analysis of such signals and theoretical modelling is able to provide valuable information on the nature and the coupling among the optical excitations, their homogeneous and inhomogeneous broadening, many-body effects, etc.

In the most simple example of two-dimensional Fourier transform spectroscopy, one may interpret the two frequencies arguments as excitation and emission frequencies. Optical transitions that can be excited directly from the ground state and contribute to the nonlinear emission appear as peaks on the diagonal corresponding to identical excitation and emission frequencies, see Fig. 11(a) and (b) for the cases of a single and two two-level systems, respectively. Consequently, inhomogeneous broadening, cp. Section 3.3, leads to a broadening of the signal along the diagonal. A coupling of optical transitions, e.g., by sharing a common state as in a three-level system, leads to off-diagonal peaks appearing in Fourier transform spectroscopy, see Fig. 11(c). Also transitions that are only possible in an already excited system lead to off-diagonal peaks that describe emission at frequencies where the system has no linear absorption, see Fig. 11(d).

See also: Applications in Semiconductors. Excitons

Further Reading

- Allen, L., Eberly, J.H., 1975. *Optical Resonance and Two-Level Atoms*. New York: Wiley.
- Baumert, T., Helbing, J., Gerber, G., *et al.*, 1997. Coherent control with femtosecond laser pulses. *Adv. Chem. Phys.* 101, 47–82.
- Bergmann, K., Theuer, H., Shore, B.W., 1998. Coherent population transfer among quantum states of atoms and molecules. *Rev. Mod. Phys.* 70, 1003–1025.
- Bloembergen, N., 1965. *Nonlinear Optics*. New York: Benjamin.
- Brewer, R.G., 1977. Coherent optical spectroscopy. In: Balian, *et al.* (Eds.), *Course 4 in Les Houches Session XXVII, Frontiers in Laser Spectroscopy 1*. Amsterdam: North-Holland.
- Cohen-Tannoudji, C., Dupont-Roc, J., Grynberg, G., 1989. *Photons and Atoms*. New York: Wiley.
- Ernst, R.R., Bodenhausen, G., Wokaun, A., 1987. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. New York: Oxford Science.

- Haug, H., Koch, S.W., 2009. Quantum Theory of the Optical and Electronic Properties of Semiconductors, 5th ed Singapore: World Scientific.
- Koch, S.W., Peyghambarian, N., Lindberg, M., 1988. Transient and steady-state optical nonlinearities in semiconductors. *J. Phys. C* 21, 5229–5250.
- Macomber, J.D., 1976. The Dynamics of Spectroscopic Transitions. New York: Wiley.
- Meier, T., Duc, H.T., Vu, Q.T., *et al.*, 2008. Ultrafast dynamics of optically-induced charge and spin currents in semiconductors. *Adv. Solid State Phys.* 46, 199–210.
- Mukamel, S., 1995. Principles of Nonlinear Optical Spectroscopy. New York: Oxford.
- Peyghambarian, N., Koch, S.W., Mysyrowicz, A., 1993. Introduction to Semiconductor Optics. Englewood Cliffs, NJ: Prentice-Hall.
- Pötz, W., Schroeder, W.A. (Eds.), 1999. Coherent Control in Atoms, Molecules, and Semiconductors. Dordrecht: Kluwer.
- Schäfer, W., Wegener, M., 2002. Semiconductor Optics and Transport Phenomena. Berlin: Springer.
- Shah, J., 1999. Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures, 2nd ed Berlin: Springer.
- Stone, K.W., Gundogdu, K., Turner, D.B., *et al.*, 2009. Two-quantum 2D FT electronic spectroscopy of biexcitons in GaAs quantum wells. *Science* 324, 1169–1173.
- Yeazell, J.A., Uzer, T. (Eds.), 2000. The Physics and Chemistry of Wave Packets. New York: Wiley.
- Zhang, T., Kuznetsova, I., Meier, T., *et al.*, 2007. Polarization-dependent optical two-dimensional fourier transform spectroscopy of semiconductors. *Proc. Natl Acad. Sci. USA* 104, 14227–14232.

Nonlinear Optics in Disordered Media: Anderson Localization

Arash Mafi, University of New Mexico, Albuquerque, NM, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Is Anderson localization preserved, enhanced, weakened, or destroyed in the presence of nonlinearity? In this short article, we will learn that there may not be a straightforward answer to this question. A survey of the literature published on this subject reveals that the answer to this rather basic question can best be described as inconclusive at this point. Of course, it should not be a surprise that the potentially chaotic behavior as a result of the nonlinear dynamics is largely behind the present state of debate in the literature on the interplay between disorder and nonlinearity in the context of Anderson localization.

This article presents a survey of some highlights in the literature over the past few years with the hope that an eventual conclusive answer can be found to such a fundamental question in a not so distant future. The question can be cast in a limited form dealing only with the optical Kerr effect, which is of greater interest to the photonics community, or in a more general context to explore and differentiate the impact of various forms of nonlinearity on localization. The interplay between disorder and nonlinearity has been explored over the years in various contexts and different disciplines. Some inevitable overlap remains in the existing literature belonging to different fields of study, primarily due to similarities among the dynamical equations of different physical systems. For example, the Gross–Pitaevskii equation describing the Bose–Einstein condensate is similar to the nonlinear Schrödinger equation describing the propagation of light in a nonlinear optical medium.

Anderson localization in general, and even Anderson localization of light in particular are broad research topics. In the optical domain, the most conclusive results on localization have so far been obtained in the context of transverse Anderson localization of light, where light is transversely localized in a waveguide-like geometry with a disordered transverse profile and propagates freely in a longitudinally uniform medium. Anderson localization is more readily observed in this transverse form as explained in the next section. Moreover, the transversely localized and longitudinally propagating platform allows for a relatively long-distance interaction of light with the optical medium, resulting in an appreciable nonlinear phase shift. As such, we will primarily focus on results that are more relevant to the impact of nonlinearity on transverse Anderson localization. For more details, we refer the interested reader to an excellent review article on this subject by Fishman, Krivolapov, and Soffer and references therein.

Anderson Localization of Light

Anderson localization is the absence of diffusive wave transport in highly disordered scattering media. Its origin dates back to a theoretical study conducted by P. W. Anderson in 1958 who investigated the behavior of spin diffusion and electronic conduction in random lattices. It was soon realized that because the novel localization phenomenon is due to the wave nature of the quantum mechanical electrons scattering in a disordered lattice, it can also be observed in other coherent wave systems, including classical ones.

The fact that Anderson localization was deemed possible in non-electronic systems was encouraging, given that the observation of disorder-induced localization was shown to be inhibited by thermal fluctuations and nonlinear effects in electronic systems. Subsequently, localization was studied in various classical systems including in acoustics, elastics, electromagnetics, and optics. It was also recently investigated in various quantum optical systems, such as atomic lattices and propagating photons.

General studies of Anderson localization have revealed that waves in one-dimensional (1D) and two-dimensional (2D) unbounded disordered systems are always localized. However, in order for a three-dimensional (3D) random wave system to localize, the scattering strength needs to be larger than a threshold value. This statement is often cast in the form of $kl^* \sim 1$, where k is the effective wavevector in the medium, and l^* is the wave scattering transport length. This is referred to as the Ioffe–Regel condition and shows that in order to observe Anderson localization, the disorder must be strong enough such that the wave scattering transport length becomes on the order of the wavelength. The Ioffe–Regel condition is notoriously difficult to satisfy in 3D disordered optical media. For the optical field to localize in 3D, very large refractive index contrasts are required that are not generally available in low-loss optical materials. The fact that Anderson localization is hard to achieve in 3D optical systems may be a blessing in disguise; otherwise, no sunlight would reach the earth on highly cloudy days.

Transverse Anderson Localization of Light

Unlike 3D lightwave systems in which the observation of the localization is prohibitively difficult, observation of Anderson localization in a quasi-2D optical system is readily possible, as was first shown by Abdullaev *et al.* and De Raedt *et al.* In particular, De Raedt *et al.* showed that 2D Anderson localization can be observed in a dielectric with transversely random and longitudinally uniform refractive index profile. An optical field that is launched in the longitudinal direction tends to remain localized in the transverse plane as it propagates in the longitudinal direction in a transversely random dielectric medium. This behavior was dubbed transverse Anderson localization of light.

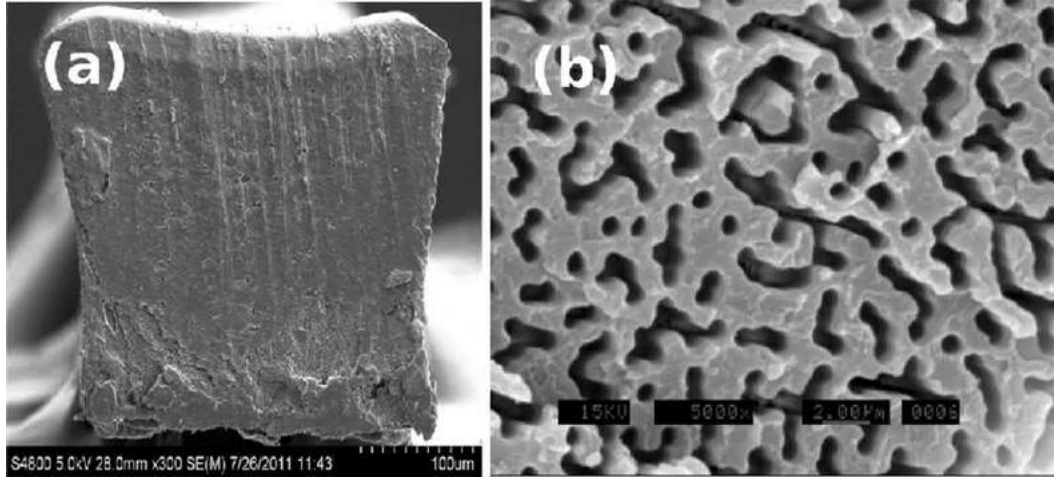


Fig. 1 (a) Cross section view of the polymer Anderson localization optical fiber developed by Karbasi *et al.* with a nearly square profile and an approximate side width of 250 μm ; (b) zoomed-in scanning electron microscope image of a 24 μm wide region on the tip of the fiber exposed to a solvent to differentiate between PMMA and PS polymer components, where feature sizes are around 0.9 μm and darker regions are PMMA. Adapted with permission from Karbasi, S., Hawkins, T., Ballato, J., Koch, K.W., Mafi, A., 2012a. Transverse Anderson localization in a disordered glass optical fiber. *Opt. Mater. Express* 2, 1496–1503. doi:10.1364/OME.2.001496; Karbasi, S., Mir, C.R., Yarandi, P.G., *et al.* 2012c. Observation of transverse Anderson localization in an optical fiber. *Opt. Lett.* 37, 2304–2306. doi:10.1364/OL.37.002304.

In their pioneering work, Schwartz *et al.* wrote the transversely disordered and longitudinally invariant refractive index profiles in a photorefractive crystal using a laser beam. They used another probe beam to investigate the transverse localization behavior. Their experiment was quite interesting as it allowed them to vary the disorder level by controlling the laser illumination of the photorefractive crystal in a controlled fashion to observe the onset of the transverse localization and the change in the localization radius as a function of the disorder level. The transverse localization of the beam and the free longitudinal propagation due to the longitudinal invariance of the dielectric medium strongly resembled the optical waveguides; therefore, applications of Anderson localization for light propagation in an optical fiber-like medium seemed an appealing extension of these ideas.

In 2012, Karbasi *et al.* reported the first observation of transverse Anderson localization in an optical fiber. In order to obtain large refractive index fluctuations required for a strong transverse localization, they randomly mixed 40,000 pieces of polymethyl methacrylate (PMMA) fiber with refractive index of 1.49, and 40,000 pieces of polystyrene (PS) fiber with refractive index of 1.59, and redrew the random stack to a 250- μm -wide optical fiber with random index fluctuations equal to 0.1 and an approximate transverse random feature size equal to 0.9 μm (see Fig. 1). The large index contrast of 0.1 ensured that both the beam localization radii and the sample-to-sample fluctuations over the ensemble of localized beam radii are sufficiently small to ensure that disordered fiber operates as a genuine optical fiber – they observed that the localized beam radii launched at different transverse positions over the facet of Anderson localizing fiber were all small and were nearly identical. Later, Karbasi *et al.* reported the first observation of Anderson localization in a silica optical fiber as well. The reported glass-air disordered fiber was made from “satin quartz”, which is a porous artisan glass. After drawing the preform, the airholes (bubbles) in the glass were stretched to form the hollow air-rods required for transverse Anderson localization. The large draw ratio sufficiently preserved the longitudinal invariance, without significant disturbance over typical lengths used in the experiments.

Transverse Anderson Localization and Kerr Nonlinearity in Theory and Experiment

Concrete experimental results on the interplay between disorder and nonlinearity are rather sparse. Even in the very few available published work in the literature, one does not get a clear sense on the extent to which Anderson localization is affected by nonlinearity. This is most likely due to the large parameter space and the possibility of many configurations under which the studies can be performed. In this article, we review the results of two pioneering works by Schwartz *et al.* and Lahini *et al.*, involving both experiment and theory, and then examine further theoretical considerations on this issue. The interaction between disorder and other forms of nonlinearity (non-Kerr) will be discussed in the next section.

Earlier, we mentioned the numerical and experimental work of Schwartz *et al.* that resulted in the observation of transverse Anderson localization of light. The authors also investigated the transverse Anderson localization of light in the presence of Kerr nonlinearity, both numerically and experimentally. The defining equation for the nonlinear propagation of light is the nonlinear Schrödinger equation (NLSE) which includes a Kerr nonlinearity term:

$$i \frac{\partial A}{\partial z} + \frac{1}{2n_0 k_0} [\nabla_T^2 A + k_0^2 (n^2 - n_0^2) A] + k_0 n_2 |A|^2 A = 0 \quad (1)$$

where A is slow-varying amplitude of the optical field, k_0 is the vacuum wavevector, z is the propagation distance, T represents the transverse coordinates (x and y here), $n(x,y)$ is the transversely disordered refractive index profile, n_0 is the mean value of the refractive index around which index fluctuation occur, and n_2 is the nonlinear index, which is positive for self-focusing and negative for self-defocusing nonlinearity.

As a case study, the authors considered a disordered lattice where the maximum contribution of the nonlinear term to the index change " $\max(|n_2| \times |A|^2)$ " was assumed to be a maximum of 15% of the index contrast of the underlying periodic waveguide. They also varied the disorder level from 0 to 30%, where the disorder level was defined as the magnitude of random index fluctuations relative to the index contrast of the underlying periodic waveguide. Using numerical simulations, they observed that over this range, the self-defocusing nonlinearity ($n_2 < 0$) results in a moderate (nearly negligible) widening of the average beam profile. They reported that after a short propagation distance, the beam broadens and the remaining propagation is essentially as if the medium were linear, primarily because the intensity is much lower and the nonlinear effects do not play a role anymore. In contrast, the self-focusing nonlinearity ($n_2 > 0$) resulted in a substantial reduction of the average localized beam diameter. The enhancement of localization due to the self-focusing nonlinearity was particularly noticeable when the disorder level was less than 15%. They observed that at high disorder levels, where localization takes place, the difference between linear and nonlinear propagation is reduced, and the behavior is dominated primarily by disorder.

The experiments were performed at different nonlinear strengths. Consider α as the ratio between the peak intensity of the probe beam and the maximum intensity of the lattice-forming beams in the photorefractive crystal. The experiments were carried out for values of $\alpha = 1, 2$, and 3. The statistical analysis of the localized beam radius clearly confirmed the expected reduction in the average beam radius due to the self-focusing nonlinearity. They observed that self-focusing enhances localization, altering the intensity profile from diffusive-like to exponentially decaying. They increased the nonlinearity (increased α) and observed that the intensity profile narrowed down accordingly. At $\alpha = 3$, the output beam profile resembles the input profile, suggesting the formation of a soliton.

Similar results were reported by Lahini *et al.* using disordered 1D waveguide lattices, as shown in Fig. 2. Their experiment consisted of a 1D lattice of $N=99$ coupled optical waveguides patterned on an AlGaAs substrate. The nonlinear Schrödinger equation governing the propagation of light in this coupled lattice system can be expressed as

$$i \frac{\partial U_n}{\partial z} = \beta_n U_n + C(U_{n+1} + U_{n-1}) + \gamma |U_n|^2 U_n \quad (2)$$

where n is the index labeling lattice sites (waveguides), U_n is the optical field amplitude at site n , β_n is the propagation constant associated with the n th waveguide, C is the tunneling rate between adjacent sites, and z is the longitudinal space coordinate. γ characterizes the nonlinear Kerr effect. Disorder was introduced to the lattice by randomly changing the width of each waveguide, resulting in the randomization of the propagation constants over a range of $\beta_0 \pm \Delta$, where β_0 is the propagation constant for a mean value of the waveguide width. Δ/C is defined as the disorder strength, with the implicit assumption that the coupling coefficient between adjacent waveguides is nearly identical.

Light was injected into one or a few waveguides at the input, and light intensity distribution was measured at the output. In the linear regime, Lahini *et al.* identified that highly localized eigenmodes exist near the top edge of the propagation constant band with a flat-phase profile and near the bottom edge of the band with a highly varying-phase profile, having phase flips between adjacent sites. The modes near the middle of the band were not as localized and also showed some phase variations. In the weak nonlinear regime, Lahini *et al.* observed that nonlinearity enhances localization in flat-phased modes and induces delocalization in the staggered modes. This behavior is explained as follows: the presence of the weak nonlinearity perturbatively shifts (increases) the value of the propagation constant of each localized mode. For the flat-phased modes, the nonlinearity shifts the modes outside the original linear spectrum. However, for the staggered, which belong to the bottom edge of the propagation constant band, a perturbative increase in the value of the propagation constant shifts it further inside the original linear spectrum. Therefore, the propagation constant of a staggered mode can cross and resonantly couple with other modes of the lattice, resulting in delocalization.

The effect of nonlinear perturbations on localized eigenmodes was studied by Lahini *et al.* experimentally via exciting a pure localized mode and increasing the input beam power. The intensities were kept far below the self-focusing threshold in the

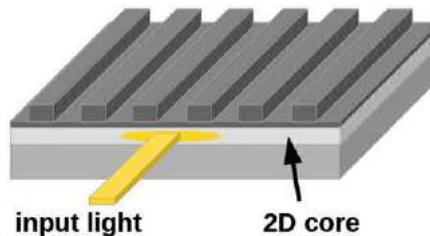


Fig. 2 Schematic view of the disordered lattice waveguide used in the experiment by Lahini *et al.* Reproduced from Lahini, Y., Avidan, A., Pozzi, F., *et al.* 2008. Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. *Phys. Rev. Lett.* 100, 013906. doi:10.1103/PhysRevLett.100.013906.

periodic lattice. They observed that some of the localized modes exhibited significant response to nonlinearity. The experiments showed that weak positive nonlinearity tends to further localize flat phased localized modes, but tends to delocalize staggered modes, consistent with the theoretical prediction above. The results obtained by Lahini *et al.* do not contradict the general observations of Schwartz *et al.*; rather, they argue that the enhanced localization due to nonlinearity by Schwartz *et al.* is only applicable to certain optical modes.

We would like to point out an early work by Pertsch *et al.* in 2004 who experimentally investigated light propagation in a disordered two-dimensional array of mutually coupled optical fibers. In the linear case they observed that light either spreads in a diffusive manner or localizes at a few sites. The absence of Anderson localization was most likely due to the fact that the disorder was not sufficiently high to localize the field whitening the transverse dimensions of the fiber. However, they reported that for high excitation power diffusive spreading is arrested by the focusing nonlinearity and a discrete soliton is formed.

Further Theoretical Considerations

In 2008, Pikovsky and Shepelyansky presented a somewhat different account of the interaction between disorder and nonlinearity. In their analysis, they considered a discrete nonlinear Schrödinger equation with a third-order diagonal Kerr nonlinear term, isomorphic to Eq. (2), to study the temporal evolution of quantum wavefunction on coupled lattices in the form of

$$i \frac{\partial \psi_n}{\partial t} = E_n \psi_n + \beta |\psi_n|^2 \psi_n + V(\psi_{n+1} + \psi_{n-1}) \quad (3)$$

where β characterizes nonlinearity, V is the hopping matrix element and is deterministic and identical for all terms. The disorder is introduced through the diagonal terms, where they are assumed to be randomly and uniformly distributed in the range $-W/2 < E_n < W/2$, and

$$\sum_n |\psi_n|^2 = 1 \quad (4)$$

is assumed. For optical waveguides, this equation describes the propagation of light in a coupled waveguide lattice, where the coupling strengths among the waveguides are identical but the propagation constants of individual waveguides are randomized by manipulating the geometry or the refractive index of each individual waveguide. Therefore, this conforms well with the problem studied by Lahini *et al.* described earlier. The normalization condition means that the total optical power remains unchanged as the light propagates through the coupled waveguide lattice. The authors verified that in the absence of nonlinearity ($\beta=0$) and in the presence of weak disorder, all eigenstates are exponentially localized as expected from Anderson localization theory with the localization length $l \approx 96(V/W)^2$ at the center of the energy band.

Pikovsky and Shepelyansky considered the spreading of a field that is initially localized at the central lattice point with $|\psi_n(0)|^2 = \delta_{n,0}$. Of course, this problem is fully studied in the limit of no disorder with $W=0$ and a general value of nonlinearity $\beta \neq 0$, e.g., in the context of nonlinear light propagation in discrete lattices, where diffraction and nonlinearity can give rise to interesting physics including diffraction-free propagation and self-localized states or discrete solitons. The opposite limit of no nonlinearity with $\beta=0$ in the presence of disorder $W \neq 0$ is also well studied in the context of Anderson localization. The intermediate regime is the focus of the work presented by Pikovsky and Shepelyansky.

Intuitively speaking, one may think that in a nonlinear disordered coupled waveguide system, the dynamics of the beam is initially influenced by nonlinearity; and as the beam spreads, the effect of nonlinearity becomes weaker and the disorder dynamics takes over. Therefore, one should always expect Anderson localization after a sufficiently long time. The analysis by Pikovsky and Shepelyansky is presented in view of earlier work by Sanchez-Palencia *et al.* who similarly argued for the localization of the field at large time scales, consistent with the intuitive view discussed above. However, Sanchez-Palencia *et al.* argued that the high-momentum cutoff of the Fourier transform of the correlation function for 1D speckle potentials can change localization from exponential to algebraic. Per argument presented in by Pikovsky and Shepelyansky, at first glance, one may think that Kerr nonlinear effects may be favored strongly by localized field configurations because if the field spreads over N lattice sites, the conservation of power implies that the Kerr nonlinearity should scale as $|\psi_n|^2 \sim 1/N$. However, Pikovsky and Shepelyansky argue that the nonlinear frequency shift $\beta |\psi_n|^2 \sim \beta/N$ should be compared with the characteristic distance between the frequencies of the exponentially localized modes, which scales as $1/N$; therefore, the effect of nonlinearity persists and does not quantitatively depend on the width of the field distribution and is omnipresent.

Using a theoretical argument, they suggested that in the presence of non-vanishing nonlinearity, the beam spreading characterized by

$$\sigma^2 = \sum_n (n - \langle n \rangle)^2 |\psi_n|^2 \quad (5)$$

should follow a subdiffusive behavior where $\sigma^2 \propto t^\alpha$ with $\alpha=2/5$. They verified their theoretical argument via numerical integration of Eq. (3) and monitoring the results, up to $t=10^8$. For the numerical simulation they used the boundary condition $\psi_n(t=0) = \delta_{n,0}$, where $n=0$ represents the middle waveguide, and the integration is performed by the operator splitting method. In a sample

set of simulations they chose the nonlinearity strength to be $\beta=V$ for two cases: case 1 with $W=2$; and case 2 with $W=2$. As expected, the initial expansion was ballistic for either case, but after some time t_0 , the expansion became subdiffusive in either case. They fit the subdiffusive expansion to $\sigma \propto t^\alpha$ over the range $t_0 < t < 10^8$ to obtain an estimate for the subdiffusive exponent. In case 1, for different instances of randomness, they obtained $0.32 \leq \alpha \leq 0.39$; and for case 2 they reported $0.28 \leq \alpha \leq 0.41$. Upon averaging over 8 independent realizations, they reported a fit of the form $57.5 \times t^{0.344}$ for case 1 and $8.7 \times t^{0.306}$ for case 2 over the subdiffusive ranges.

Pikovsky and Shepelyansky analyzed the probability distribution $w_n = |\psi_n|^2$ over lattice sites n ; in the absence of nonlinearity they observed an exponential decay with an exponent consistent with the Anderson localization. In the presence of nonlinearity, however, they observed a rather flat distribution in the vicinity of $n=0$ where the width of the flat distribution depended on the value of β . As expected, the (low-intensity) tails of the distribution always followed a decay exponent consistent with the linear theory of Anderson localization. Another important observation reported by Pikovsky and Shepelyansky was the presence of a certain critical strength β_c for nonlinearity, beyond which the reported delocalization behavior happens. The numerical solutions suggested a value of $\beta_c \approx 0.1$ for this threshold value. However, Pikovsky and Shepelyansky argued that the threshold-like behavior might not be perfect and a slow spreading may persist all the way down to very small values of nonlinearity albeit at an extremely slow rate that could not be detected in the numerical integration scheme.

In another study published by Kopidakis *et al.* on the spreading of an initially localized wave packet, the authors confirmed that while there are many initial conditions such that the second moment of the norm in Eq. (5) and energy density distributions diverge with time in agreement with Pikovsky and Shepelyansky, the participation number of a wave packet does not diverge simultaneously. The participation number is defined as

$$P = \frac{(\sum_n |\psi_n|^2)^2}{\sum_n |\psi_n|^4} \quad (6)$$

which is an alternative measure of the wave spreading to Eq. (5). Kopidakis *et al.* prove this result analytically for norm-conserving models that satisfy Eq. (4) with strong enough nonlinearity. They showed that initially localized wave packets with a large enough amplitude cannot spread to arbitrarily small amplitudes. The consequence is that a part of the initial energy must remain well focused at all times.

A later report by Fishman *et al.* provide a somewhat different account of the impact of nonlinearity on Anderson localization in disordered lattices. Fishman *et al.* present a thorough survey of the many subtleties involved regarding the interaction of nonlinearity and disorder. The conclusion is that the situation can best be described as inconclusive at this point. For example, when using the numerical simulations, they caution that Eq. (3) is chaotic with an exponential sensitivity to numerical errors. Because of the possibly chaotic motion due to the presence of nonlinearity, Fishman *et al.* argue that the numerical solutions of Eq. (3) are not actual solutions. In order to draw conclusions similar to those by Pikovsky and Shepelyansky, one must assume that the numerical solutions are statistically similar to the correct solutions with no real theoretical support for this assumption. For long-time (or long-distance) propagation in waveguides, it is not clear that reducing the t step size can control the cumulative numerical error, given that the limit of zero t step may be singular. Details of these arguments can be found in the review article by Fishman *et al.*

Other Forms of Nonlinearity

Other forms of nonlinearity besides Kerr can also interact with the disorder-induced localization. Recently Leonetti *et al.* explored the impact of thermally induced nonlinearity on Anderson localization for a beam of light propagating in the polymer Anderson localizing optical fiber of Fig. 1. They reported the observation of a self-focusing action occurring in the disordered fiber with the binary index distribution which was triggered by a defocusing thermal nonlinearity. The larger light absorption strength in PMMA than PS results in an inhomogeneous temperature distribution. The higher temperature in PMMA translates into a decrease of its refractive index. The result is an increased refractive index mismatch and stronger localization. In effect, they demonstrated that transversely localized modes shrink when the pump intensity is increased despite the fact that $n_2 < 0$ for the polymers, which seems quite counter-intuitive in the first glance.

In a subsequent publication, the authors provided further evidence of this behavior by analyzing the direct relation between the optical intensity and the localization length. In their experiments, they probed the behavior of light by using both a broadband (a femtosecond Ti: Sapphire laser at 800 nm wavelength with 80 nm bandwidth) and a monochromatic laser (solid-state continuous Nd: YAG laser at 1064 nm). They observed that the broadband light beam injected in the fiber is dispersed at the output into a series of peaks corresponding to localized modes. This output spectrum is modified when the input intensity in the system is increased, demonstrating that the nonlinearity also affects the modes structure. Importantly, they found that the spatial extension of the individual modes decreases with fluence due to the thermal nonlinearity. They repeated the experiment using a monochromatic continuous wave (CW) laser, activating only a single spatial mode, and observed similar self-focusing behavior with the optical power. By comparing their observations with standard homogeneous polymer fibers, they proved that the disorder has turned the defocusing medium into a focusing one.

Further Reading

- Abdullaev, S., Abdullaev, F.K., 1980. On propagation of light in fiber bundles with random parameters. *Radiofizika* 23, 766–767.
- Abrahams, E., 2010. 50 years of Anderson Localization. World Scientific.
- Anderson, P.W., 1958. Absence of diffusion in certain random lattices. *Phys. Rev.* 109, 1492–1505. doi:10.1103/PhysRev.109.1492.
- De Raedt, H., Lagendijk, A., de Vries, P., 1989. Transverse localization of light. *Phys. Rev. Lett.* 62, 47–50. doi:10.1103/PhysRevLett.62.47.
- Fishman, S., Krivolapov, Y., Soffer, A., 2012. The nonlinear schrödinger equation with a random potential: results and puzzles. *Nonlinearity* 25, R53.
- Ioffe, A.F., Regel, A.R., 1960. Non-crystalline, amorphous, and liquid electronic semiconductors. *Prog. Semicond.* 4, 237–291.
- John, S., 1991. Localization of light. *Phys. Today* 44, 32–40.
- Kopidakis, G., Komineas, S., Flach, S., Aubry, S., 2008. Absence of wave packet diffusion in disordered nonlinear systems. *Phys. Rev. Lett.* 100, 084103. doi:10.1103/PhysRevLett.100.084103.
- Lagendijk, A., Tiggelen, B.V., Wiersma, D.S., 2009. Fifty years of Anderson localization. *Phys. Today* 62, 24–29.
- Lahini, Y., Avidan, A., Pozzi, F., *et al.*, 2008. Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. *Phys. Rev. Lett.* 100, 013906. doi:10.1103/PhysRevLett.100.013906.
- Leonetti, M., Karbasi, S., Mafi, A., Conti, C., 2014a. Experimental observation of disorder induced self-focusing in optical fibers. *Appl. Phys. Lett.* 105, 171102. doi:10.1063/1.4900781.
- Leonetti, M., Karbasi, S., Mafi, A., Conti, C., 2014b. Observation of migrating transverse Anderson localizations of light in nonlocal media. *Phys. Rev. Lett.* 112, 193902. doi:10.1103/PhysRevLett.112.193902.
- Mafi, A., 2015. Transverse Anderson localization of light: A tutorial. *Adv. Opt. Photon* 7, 459–515. doi:10.1364/AOP.7.000459.
- Pertsch, T., Peschel, U., Kobelke, J., *et al.*, 2004. Nonlinearity and disorder in fiber arrays. *Phys. Rev. Lett.* 93, 053901. doi:10.1103/PhysRevLett.93.053901.
- Pikovsky, A.S., Shepelyansky, D.L., 2008. Destruction of anderson localization by a weak nonlinearity. *Phys. Rev. Lett.* 100, 094101. doi:10.1103/PhysRevLett.100.094101.
- Sanchez-Palencia, L., Clement, D., Lugan, P., *et al.*, 2007. Anderson localization of expanding Bose–Einstein condensates in random potentials. *Phys. Rev. Lett.* 98, 210401. doi:10.1103/PhysRevLett.98.210401.
- Schwartz, T., Bartal, G., Fishman, S., Segev, M., 2007. Transport and Anderson localization in disordered two-dimensional photonic lattices. *Nature* 446, 5255.
- Segev, M., Silberberg, Y., Christodoulides, D.N., 2013. Anderson localization of light. *Nat. Photon.* 7, 197–204.

Second-Harmonic Generation in Two-Dimensional Materials

Myrta Grüning, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

Main Definitions and Background

Two-dimensional (2D) materials are structures in which atoms are arranged periodically in two of the three spatial directions and which have (sub-)nanometric dimension in the third direction. Furthermore, when more than one layer is present, the chemical bonding between atoms in the periodic directions are much stronger (covalent or ionic) than those between layers (van der Waals). 2D materials are synthesized mostly by mechanical exfoliation of three-dimensional van der Waals layered materials. Some of the more common materials, such as monolayers of graphene, h-BN, Gallium chalcogenides (e.g. GaTe, GaSe) and transition metal dichalcogenides (e.g. MoS₂, WSe₂) have been obtained as well by chemical vapour deposition (CVD) or by van der Waals epitaxy (VDWE) (see e.g. [Butler et al. \(2013\)](#)). These processes produce polycrystalline structures with crystal flakes that have one to few layers thickness and with dimensions of 10 μm to 70–80 μm. The flakes which have typically a triangular shape – reflecting the crystal symmetry – are oriented in different directions and separated by grain boundaries. A variety of spectroscopic techniques are in place to characterize the crystal orientation, layers number and stacking, grain boundaries of these polycrystalline structures (see e.g. [Butler et al. \(2013\)](#)).

Second-harmonic generation (SHG) is a nonlinear optical process in which a crystal irradiated with laser light of a given frequency ω generates radiation at the second-harmonic frequency 2ω (see [Boyd \(2008\)](#)). In the context of 2D materials, SHG is relevant as it is the basis of spectroscopic techniques to structurally characterize the samples obtained e.g. by exfoliation or CVD. In addition, there is a general interest in optical properties of 2D materials which are expected to be characterized by strong excitonic effects because of the geometrical confinement and poor dielectric screening. Finally technologically there is an interest in the possibility of employing 2D materials with strong SHG as on-chip optical frequency converters.

We recall here some key relations for the SHG which are needed in the discussion that follows. The electric polarization field $\mathbf{P}$ can be expanded in powers of the total electric field $\mathbf{E}$ as (we assume the dipole approximation, if not otherwise stated, see [Shen \(2003\)](#)):

$$\mathbf{P}(\omega) = \epsilon_0 [\chi^{(1)}(\omega)\mathbf{E}(\omega) + \chi^{(2)}(\omega; \omega_1, \omega_2)\mathbf{E}(\omega_1)\mathbf{E}(\omega_2) + \dots + \chi^{(n)}(\omega; \omega_1, \dots, \omega_n)\mathbf{E}(\omega_1)\dots\mathbf{E}(\omega_n) + \dots] \quad (1)$$

where the coefficients $\chi^{(n)}$ are the (non)linear optical susceptibilities and ϵ_0 is the permittivity of vacuum. The SHG corresponds to the second order term when $\omega_1 = \omega_2$ and thus $\omega = \omega_1 + \omega_2 = 2\omega_1$. Since $\mathbf{P}$ and $\mathbf{E}$ are vector, $\chi^{(2)}$ is a third-rank tensor, $\chi_{ijk}^{(2)}$, with the first index i referring to the direction of the polarization field and the remaining indexes, jk to the direction of the electric field:

$$P_i^{(2)}(2\omega) = \epsilon_0 \chi_{ijk}^{(2)} E_j(\omega) E_k(\omega) \quad (2)$$

In [Eq. \(2\)](#), $P_i^{(2)}$ is the second order contribution to the polarization field. Symmetry properties of the crystal are reflected in the structure of the second order susceptibility tensor. For example in centrosymmetric crystals, i.e. with an inversion center, all elements of the tensor are identically zero (this implies that the SHG from the electric dipole is identically zero. The measured SHG in centrosymmetric crystals is very small, but not identically zero, because of higher order multipole contributions).

The most commonly studied 2D crystals belongs to the D_{3h} symmetry point-group: the crystal structure is invariant under rotation of 120° around an axis perpendicular to the plane of the 2D crystal (an example is given in [Fig. 1](#)). The D_{3h} symmetry point-group does not comprise the inversion operation, so crystals belonging to this point-group have a nonzero SHG. The second order susceptibility tensor in this case has only the $\gamma\gamma\gamma = -\gamma\chi\chi = -\chi\chi\gamma = -\chi\gamma\chi$ components different from zero where we labeled with x and y the armchair and the zig-zag directions (see [Fig. 1](#)). Note that graphene belongs to the D_{6h} symmetry point-group which includes the inversion operation and therefore does not have a dipole-allowed SHG.

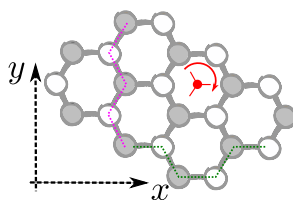


Fig. 1 Symmetry of 2D materials belonging to the D_{3h} symmetry point-group. As an example we consider a hexagonal structure with two atoms of different species. Only few units of the infinite periodic structure are drawn. The red dot represents the 3-fold rotation axis perpendicular to the plane of the 2D crystal. A rotation of 120° around this axis does not change the crystal structure. The three in-plane two-fold rotation axes are also represented as red lines with origin at the dot. The x axis is chosen parallel to one of the armchair directions (dotted green line) and the y parallel to one of the zig-zag directions (dotted magenta line).

SHG for Characterization of 2D Materials

SHG experiments on 2D materials are realized by shining laser light onto the sample on a substrate. The second-harmonic power (either reflected or transmitted) is then measured. Most of the SHG experiments on 2D materials are performed in the reflection geometry which is schematically shown in Fig. 2. The substrate is typically dielectric coated, e.g. SiO_2/Si . The substrate can be chosen to be transparent to both the fundamental and SH frequency so that the transmitted SH signal can be measured as well. Normally the sample can be both mechanically rotated around the out-of-plane direction and shifted in the two in-plane directions. Rotation allows to study dependence on the light polarization. Scans in the in-plane directions allows to take a SHG image of the sample. Because of the similarity of the problem, many of the concepts and techniques are reminiscent of concepts and techniques applied in SHG measurements of surfaces, films and interfaces.

Polarization Dependent Measurements

To illustrate the basic principle behind polarization dependent measurements we consider here a linearly polarized laser light with propagation direction perpendicular to the 2D crystal and the associated electric field forming an angle θ_0 with the armchair direction of the crystal which we assume to belong to the D_{3h} symmetry-point group. This is schematically depicted in Fig. 3. Because of the symmetry with respect the three-fold rotation axis, varying the field polarization of 120° leaves the SHG unchanged. That means that the parallel (along x') and perpendicular (along y') components of the SHG electric field E_{SH} vary as $\sin(3\theta + \theta_0)$ and $\cos(3\theta + \theta_0)$ when the sample is rotated by θ . The components of the SH power then, which is proportional to E_{SH}^2 , vary as $\sin^2(3\theta + \theta_0)$ and $\cos^2(3\theta + \theta_0)$. Plotting the reflected SH power as a function of the polarization angle on a polar grid results in a plot similar to that in Fig. 3 (see e.g. Kumar *et al.* (2013)). A fit of the measured data against the expected behaviour for the SH power gives the SH intensity and the initial direction of the electric field with respect to the crystal armchair direction θ_0 .

Then from polarization dependent measurements one can extract two important structural information on the 2D crystal: (1) whether it has a given/expected symmetry, (2) its orientation. Furthermore from the reflected power one can in principle deduce an absolute measure of the SHG, though this is not straightforward as discussed in Section "Absolute Measure of SHG."

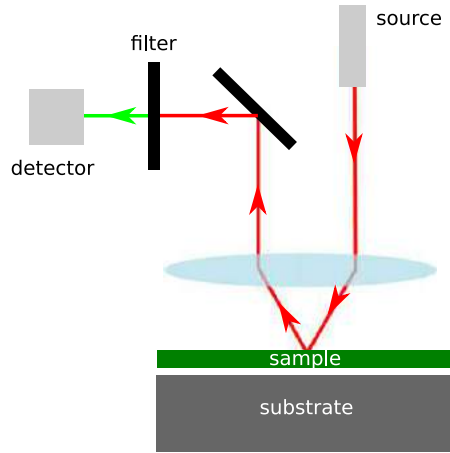


Fig. 2 Schematic set up of a SHG experiment in reflection geometry. The light pumped by the source is focus onto the sample by an objective which also collects the reflected light. Before being detected and analyzed from the detector, the fundamental and frequencies other than the SH frequency are blocked by a (set of) filter(s).

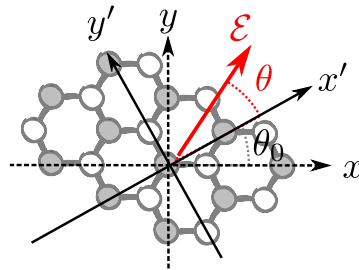


Fig. 3 Direction of electric field with respect to the 2D crystal sample. Initially, the electric field, polarized along the x axis, forms an angle θ_0 with the armchair direction of the crystal (x axis). Later the crystal is rotated, and forms an angle θ with the initial direction.

SHG Imaging

By scanning the SH power in the two in-plane directions of an sample illuminated with polarized (quasi-)monochromatic light, the SH can be rendered in a false-color image, where a scale of colors is associated to the (relative) SH intensity in either the x or y directions. If the substrate is chosen such that it has a negligible SHG, then high contrast images can be obtained. These images provide several information on the sample.

First, as the SH intensity does not change within a crystal, domains of the same color can be identified as crystal flakes. Second, because of the dependence on the electric field polarization discussed previously, crystals with a different orientation show a different SH intensity. Then those crystals can be distinguished from each other and their relative orientation can be estimated. However as it is evident from Fig. 4, crystals which form an angle of 60° with each other have the same color. Nevertheless it is still possible to distinguish them because the edges have a different electronic structure with respect to the rest of the crystal and, due to the sensitivity of the SHG to the electronic structure, have a different color (see e.g. Yin *et al.* (2014), Kumar *et al.* (2013)).

When compared with optical imaging, SHG imaging offers several advantages: stronger contrast with respect to the substrate, possibility of distinguish crystals with different orientation and of detecting edges. As discussed in the following section it also allows one to distinguish between different stacking geometries.

Dependence on Stacking and Number of Layers

The SHG is sensitive to the number of layers and to the stacking geometry. The most evident dependence on the number of layers occurs for 2D materials obtained from centrosymmetric 3D crystals (this is the case of several commonly studied crystals such as the transition metal dichalcogenides or the hexagonal-BN). For these 2D crystals a clear pattern is observed in the SHG of even and odd layers samples, with negligible SHG when the number of layers is even (Kim *et al.*, 2013; Li *et al.*, 2013). In fact, samples with an even number of layers have an inversion center and as discussed in Section “Main Definitions and Background” this is reflected in a zero dipole-allowed SHG. Multipole contributions to SHG are in fact several orders of magnitude smaller than the dipole one.

A simple interference model gives an alternative way to understand the even-odd dependence. Let's consider the SH electric field (that is the electric field driven by the SH polarization) of each crystal plane independently. Then for example the SH electric field parallel and perpendicular to the armchair direction in the i th layer are $E_{2\omega}^{\parallel} \propto \cos(3\theta_i)$ and $E_{2\omega}^{\perp} \propto \sin(3\theta_i)$ where θ_i is the angle between the applied electric field and the armchair direction. If we consider two layers, the total SH electric field is then given by the sum of the fields. The intensity – related to the square of the total electric field – then is given by:

$$I_{1+2} = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(3(\theta_1 - \theta_2)) \quad (3)$$

For identical layers $I_1 = I_2$, then when $\theta_1 - \theta_2 = 60^\circ$, I_{1+2} vanishes. The AB stacking, which gives a centrosymmetric crystal when the number of layers is even, corresponds indeed to an angle of 60° between the armchair directions of two successive layers (see Fig. 5). Instead, for crystals with an AA stacking $\theta_1 - \theta_2 = 0^\circ$ and I_{1+2} has its maximum value. The AA stacking occurs for example in GaSe and GaTe. Moreover it is possible to synthesize layered materials with a stacking different from that of the corresponding bulk layered material (Hsu *et al.*, 2014). Note that the interference model can be refined by considering the SH electric fields as a radiation fields which is equivalent to account for multipole contributions. Then also for an even number of layers in the AB stacking one obtains a small, but nonzero SHG.

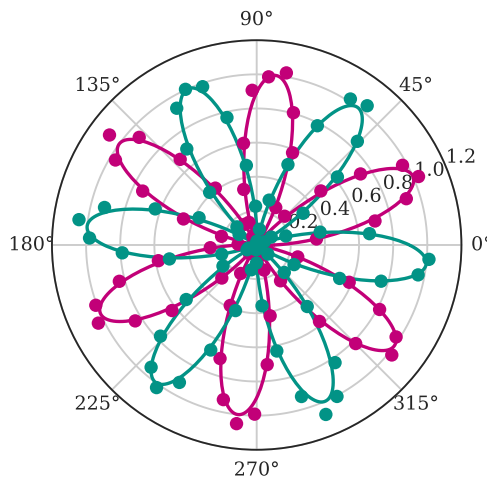


Fig. 4 Polar plot of the SH power versus the direction of the incident electric field, θ in Fig. 3. Reflecting the D_{3h} symmetry the parallel (magenta continuous line) and perpendicular (teal continuous line) components of the SH power are proportional to $\sin^2(3\theta + \theta_0)$ and $\cos^2(3\theta + \theta_0)$ respectively. By fitting the experimental data (dots) to the ideal behaviour the initial direction of the field θ_0 and absolute SH power are obtained. This plot is illustrative of typical plots obtained from experiments, but the data is from a real experiment. Dots have been obtained by adding a small random number to the sinusoidal curves.

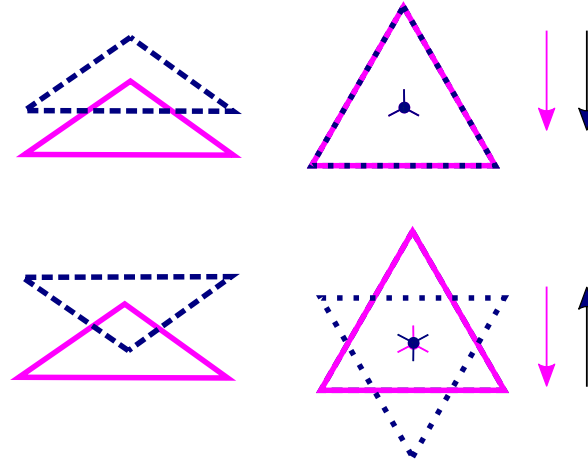


Fig. 5 Two possible stacking arrangements for two layers represented by equilateral triangles (magenta: bottom layer, blue: top layer). Equilateral triangles have been chosen because they are invariant under the same symmetry operations as 2D crystals belonging D_{3h} symmetry point-group. Top: angle between the two armchair directions of the layer is 0° , this is the so called *aa* stacking, side view on the left and top view on the right. When considering each layer independently, the radiated SH electric fields have the same direction, thus they sum up. Bottom: angle between the two armchair directions of the layer is 60° , this is the so called *AB* stacking (or Bernal stacking), side view on the left and top view on the right. When considering each layer independently, the radiated SH electric fields have opposite direction, thus they cancel. This is consistent with the system being centrosymmetric, i.e. invariant under the inversion operation (the inversion center is along the main rotation axis, at the midpoint between the two layers) and thus having no dipole allowed SHG.

Eq. (3) can be used to extract the stacking geometry of bilayers from polarization dependent measures. On the other hand by adding more layers (and knowing the stacking), one can predict the dependence of the SH intensity on the number of layers. As detailed above, from calculations based on the interference model (or modifications) one expects that in case of *AB* stacking samples with an even number of layers have negligible SHG, while samples with an odd number of layers have all similar SHG. Instead in the case of few layers non-centrosymmetric crystals, e.g. with *AA* stacking, the SH intensity is expected to grow quadratically. However the interference model— even when perfectioned by accounting for absorption and multiple reflections— captures the dependence on the number of layers only partially as it neglects the effects of interlayer interactions on the electronic structure. In particular for very few layers (typically < 10), confinement effects are stronger and the screening less effective. These effects can change both the absorption and SHG spectral shape and onset and ultimately can be captured only by electronic structure calculations. Indeed in experimental studies it is generally observed that a quadratic dependence hold for 10–100 layers, but for < 10 layers the behaviour changes (see e.g. Tang *et al.* (2016)).

Absolute Measure of SHG

A measure of the absolute SHG intensity is not straightforward. From Eq. (1) the SH susceptibility is extracted from the change in the polarization field P of the sample as a function of the total electric field E .

In practice what is measured is the SH radiation intensity (or power) and what is known the intensity (or power) of the applied field. The relation between the two quantities is derived from the wave equation for the electric field driven by the second order polarization field and using Eq. (2) to introduce the SH susceptibility and the total electric field. Then the SH radiation intensity is calculated from the corresponding field as (Boyd (2008)). Because of the quadratic dependence of the polarization on the electric field (Eq. (2)), the SH radiation intensity (power) depends quadratically on the intensity (power) of the applied field. The coefficient of proportionality contains the square of the SH susceptibility and other factors depending on the sample i.e. the thickness, the presence of a substrate, etc. The model assumed to describe the sample influences the experimental estimate for the SH susceptibility as discussed in Section “Models for the Sample.”

Furthermore, as it is the case for the SHG in bulk materials, the intensity of the applied field requires the detailed knowledge of the spatial and temporal dependence of the laser beam. Commonly the dependence on the intensity of the applied field is thus eliminated by measuring the SH power of the sample relatively to a reference material for which the SH susceptibility is accurately known (as e.g. quartz or beta barium borate).

Table 1 collects results from recent experimental measures of the SHG in 2D materials and reports parameters and characteristics that may influence the estimate, which are discussed below (Sections “Models for the Sample, Role of Substrate and Role of Excitonic Resonances”). The out-of-resonance values are usually comparable with those of nonlinear crystals, such as the beta barium borate and potassium titanyl phosphate, commonly used as frequency converters. At resonance, strong to very strong (one to three order of magnitude larger than conventional nonlinear crystal) is reported. In the subsection below aspects that may influence significantly the SHG experimental estimate are detailed.

Table 1 Measure of the SH for 2D materials [monolayer (ML), bilayer (BL), trilayer (TL) and few layers (FL)] from recent experiments. For each material beside the SH is reported how the material was synthesized, the substrate (thickness in parenthesis), excitation wavelength, whether excitonic resonances are excited, which model has been assumed to extract the SH from the measurements and the publication. Considering the number of parameters on which the estimates depend, those should be consider as order of magnitude estimates

	SH (pm/V)	Synthesis	Substrate	Wavelegth (nm)	Resonance	Model	
MoS ₂ ML	6	CVD	Si/SiO ₂ (300 nm)	static limit	No	sheet	PRB 92, 159901 (2015)
	40	CVD	Si/SiO ₂ (300 nm)	1100–2000	Yes (A,B excitons)	sheet	PRB 92, 159901 (2015)
	25–30	CVD	PET/fused silica	1360	Yes (A exciton)	sheet	APL 107, 13113 (2015)
	35–40	CVD	PET/fused silica	1240	Yes (B exciton)	sheet	APL 107, 13113 (2015)
	30–100	Exfoliated	amorphous quartz	680–1080	Yes (C exciton)	sheet	PRB 87, 201401 (2013)
	1×10^5	Exfoliated	Si/SiO ₂ (90 nm)	810	Yes (C exciton)	bulk	PRB 87, 161403 (2013)
	5×10^3	CVD	Si/SiO ₂ (90 nm)	810	Yes (C exciton)	bulk	PRB 87, 161403 (2013)
	150	CVD	sapphire	810	Yes (C exciton)	sheet	ACS Nano 8, 2951 (2014)
	160	Exfoliated	fused silica	810	Yes (C exciton)	sheet	Nano Lett 13, 3329 (2013)
	430	not rep.	not rep.	1600	No	bulk	JACS 137, 7994
MoS ₂ TL	10–50	Exfoliated	amorphous quartz	680–1080	Yes (C exciton)	sheet	PRB 87, 201401 (2013)
MoSe ₂ ML	50	CVD	Si/SiO ₂	1200–1800	Yes (A,B excitons)	sheet	Ann. Phys. 528, 551 (2016)
WS ₂ ML	9×10^3	CVD	suspended, Si/SiO ₂	832	Yes (C exciton)	sheet	SciRep 4, 5530 (2014)
WSe ₂ ML	10×10^3	Exfoliation	Si/SiO ₂ (300 nm)	816	Yes (C exciton)	sheet	2D Mat 4, 045015 (2015)
GaSe ML	2.4×10^3	VDWE	fused silica	1210	No	bulk	JACS 137, 7994
	1.7×10^3	VDWE	fused silica	1350	No	bulk	JACS 137, 7994
	700	VDWE	fused silica	1600	No	bulk	JACS 137, 7994
GaSe BL	21–34	Exfoliated	Si/SiO ₂ (90 nm)	1030–1460	No	sheet	PRB 94, 125302 (2016)
	60	Exfoliated	glass	900–1080	No	bulk	Ang. Chem. 127, 1201 (2015)
GaSe TL	5–9	Exfoliated	Si/SiO ₂ (90 nm)	1030–1460	No	sheet	PRB 94, 125302 (2016)
	93	Exfoliated	glass	900–1080	No	bulk	Ang. Chem. 19, 1201 (2015)
GaTe FL	1×15	Exfoliated	Si/SiO ₂	1560	No	?	APL 108, 073103 (2016)
h-BN ML	10	Exfoliated	fused silica	810	No	sheet	Nano Lett 13, 3329 (2013)

Models for the Sample

For 2D materials two models are commonly used (see e.g. [Clark et al. \(2014\)](#)):

- One model assumes that the 2D material sample behaves as bulk and uses the same formula for the SH radiation electric field with the only difference that the phase mismatch is neglected since the thickness of the material is much smaller than the coherence length. With this assumption the relation between the intensity I^{SH} of the SH radiation and the intensity I of the applied electric field at frequency ω to extract the magnitude of the SH susceptibility $|\chi^{(2)}|$ reads:

$$I^{\text{SH}}(2\omega) = \frac{\omega^2 |\chi^{(2)}|^2 \Delta h^2 I^2(\omega)}{2n_{2D}^2(\omega) n_{2D}(2\omega) \epsilon_0 c^3} \quad (4)$$

where Δh is the sample thickness, n_{2D} is the refractive index of the 2D material and c is the speed of light.

- Another model instead considers the 2D material on a dielectric substrate with refractive index n_{sub} as a dipole sheet screened by the substrate and thus considers the electric field driven by the SH sheet polarization. With this assumption the relation between the applied and measured intensities reads:

$$I^{\text{SH}}(2\omega) = \frac{512\omega^2 \pi^2 |\chi_{\text{surf}}^{(2)}|^2 I^2(\omega)}{(n_{\text{sub}}(\omega) + 1)^6 \epsilon_0 c^3} \quad (5)$$

Here $|\chi_{\text{surf}}^{(2)}|$ is the surface second order nonlinear susceptibility ([Shen, 2003](#)) which is relate to the usual (volume) second order nonlinear susceptibility by:

$$\chi_{\text{surf}}^{(2)} = \int dz \chi^{(2)}$$

where the integral runs across the 2D material along the normal to the surface. If we neglect the dependence on z of the nonlinear susceptibility, $\chi_{\text{surf}}^{(2)} = \chi^{(2)} \Delta h$.

One important difference between the two models is that at the denominator of the sheet model (Eq. (5)) the refractive index of the dielectric substrate appears instead of that of the material (Eq. (4)). As the latter is typically much smaller than the former it has been argued that a difference up to three order of magnitude in the estimate of the SH susceptibility may arise.

Role of Substrate

Characteristics of the substrate plays a role in the measured intensity. Typically a dielectric coated substrate is chosen for which the bulk SHG is negligible. However even if the bulk SHG of the substrate is negligible, surface SHG can still be noticeable. In addition atoms may be trapped at the interface between the sample and the substrate. Those defects can modify the SHG of the sample. In particular charge-transfer defects can induce strong local fields and locally enhance the SHG signal.

Furthermore it has been reported that the measured SHG changes when changing the thickness of dielectric coating. This effect has been explained as an enhancement of the measured SHG due to the interference of multiple reflections from the substrate (Miyachi *et al.*, 2016), similar to the enhancement observed in a Fabry-Perot resonator discussed in Section “Optical Devices.”

Finally, while for measurements of the SHG of the sample, substrate effects are a disturbance, in applications one may want to actually enhance the SHG from the sample by choosing an appropriate substrate. In this context few experiments have explored effects of metal substrates as it is expected that strong local fields from surface plasmons could significantly enhance the SH intensity (Zeng *et al.*, 2015).

Role of Excitonic Resonances

Excitations are commonly interpreted in terms of electron-hole pairs. Pictorially, considering the electronic structure of a finite-gap material, an electron-hole pair is formed when an electron is excited from the valence to the conduction energy band. Within this picture the excited system then has an extra electron in the conduction, and a missing electron (or hole) in the valence. The electron and the hole can be considered as particles with opposite charge interacting through the Coulomb interaction and forming a bound state, the exciton. The physics of these bound states resembles that of the hydrogen atom in which the radius and the energies are modified by the electric screening from the surrounding electrons. The largest the screening, the largest the radius of the exciton (i.e. the more delocalized) and the smallest the binding energy. In the case of 2D materials one needs to account for the geometrical confinement that modifies the equations for the radius and binding energy: the binding energy of the ground state exciton is 4 times larger than that of the bulk counterpart and the radius smaller by a factor 2 (Haug and Koch, 1990). Then typically excitons in 2D materials are expected to be more localized and more bound than in their bulk counterpart. Because of the resemblance with the hydrogen atoms excitons are commonly addressed as $1s$, $2p$, etc. even though it should be noted that the symmetry properties of these excitations are different from those of the hydrogenic atomic orbitals and depends on the crystal symmetry.

When the laser frequency ω is such that 2ω matches the energy of one of these excitations the SHG is expected to be enhanced. Indeed experimentally enhancements up to three order of magnitude have been observed for transition metal dichalcogenides (Wang *et al.*, 2015) resulting in very strong SHG at excitonic resonances as can be seen in Table 1.

Electronic Structure Calculations

The SHG can be computed from the electronic structure of the material as (Shen, 2003):

$$\chi_{ijk}^{(2)}(\omega; \omega_1, \omega_2) = \int dk \sum_{v,c,c'} f_v(k) \left\{ \frac{\langle v k | \xi_i | c k \rangle \langle c k | \xi_j | c' k \rangle \langle c' k | \xi_k | v k \rangle}{[\omega - \omega_{cv}(k)][\omega_1 - \omega_{c'v}(k)]} + \frac{\langle v k | \xi_i | c k \rangle \langle c k | \xi_k | c' k \rangle \langle c' k | \xi_j | v k \rangle}{[\omega - \omega_{cv}(k)][\omega_2 - \omega_{c'v}(k)]} + \frac{\langle v k | \xi_k | c k \rangle \langle c k | \xi_j | c' k \rangle \langle c' k | \xi_i | v k \rangle}{[\omega + \omega_{c'v}(k)][\omega_2 + \omega_{cv}(k)]} \right. \\ \left. + \frac{\langle v k | \xi_j | c k \rangle \langle c k | \xi_k | c' k \rangle \langle c' k | \xi_i | v k \rangle}{[\omega + \omega_{c'v}(k)][\omega_1 + \omega_{cv}(k)]} + \frac{\langle v k | \xi_j | c k \rangle \langle c k | \xi_i | c' k \rangle \langle c' k | \xi_k | v k \rangle}{[\omega_1 - \omega_{cv}(k)][\omega_2 + \omega_{c'v}(k)]} + \frac{\langle v k | \xi_k | c k \rangle \langle c k | \xi_i | c' k \rangle \langle c' k | \xi_j | v k \rangle}{[\omega_1 + \omega_{c'v}(k)][\omega_2 - \omega_{cv}(k)]} \right\} \quad (6)$$

The main quantities in Eq. (6) are the single particles wavefunctions $|ak\rangle$ where $a=v$ corresponds to valence and $a=c$ to conduction states, and k is the crystal momentum. ω_{cv} is the difference between the energies of the conduction and valence states. ξ_i is the dipole operator in the i direction, defined consistently with periodic boundary conditions.

Single particle wavefunctions and energies can be obtained from any electronic structure approach. In practice most of nowadays calculations are based on single-particle energies and wavefunctions from a density-functional theory calculations. The main advantages of this approach is computational ease, availability of computer codes while not relying on ad-hoc or experimental parameters. On the other hand, regarding the latter point it has to be noted that due to the underestimation of band gap inherent to this approach, often in calculations the knowledge of the material experimental band-gap is used (as well the band gap obtained from other approaches, such as the GW approximation can be used).

The role of electronic structure calculations of SHG is mainly (1) to connect features in the SHG spectrum to the electronic structure of the material and (2) to provide estimates for the absolute SHG. In the context of 2D materials connecting spectral features to the electronic structure can be important for instance to understand and interpret changes of the SHG with the number

of layers. On the other hand, an estimate of SHG is relevant considering the experimental difficulties in measuring the SHG. Comparison with the theory can provide a more firmly ground for the assumptions made in extracting the experimental SHG.

Eq. (6) is derived considering non-interacting electrons, in what is commonly addressed as independent particle model. More precisely, electron correlation is included in the ground-state calculations from which the electronic structure is extracted, but neglected in the response of the system to external perturbation.

Descriptions beyond the independent particle model include local fields and excitonic effects. The former effect is due to the microscopic electric fields from crystal inhomogeneity which locally counteract the applied electric field. The latter effect is due to the electron-hole long range interaction as discussed in Section “Role of Excitonic Resonances.” Modifications of Eq. (6) so to introduce these effects are nontrivial.

At present very few calculations of the SHG of 2D material exists including local fields and excitonic effects (see e.g. Grüning and Attacalite (2014), Trolle *et al.* (2014)). For transition metal dichalcogenides available calculations confirm the enhancement of the SHG at excitonic resonances and the strong SHG of these materials. As a side result of these works, calculations at the independent particle level are typically underestimating the SHG at excitonic resonances by a factor 2–3. Considering the experimental uncertainties in the determination of the SHG, calculations at the independent particle level can then be considered in general sufficient to provide a reasonable estimate of the SHG.

Optical Devices

As seen in Section “Absolute Measure of SHG” transition metal dichalcogenides have a significant SHG in the frequency range of Ti: Sapphire laser and Gallium chalcogenides have a significant SHG for wavelengths near 1560 nm relevant for telecommunications. This strong SHG can potentially be exploited for on-chip optical devices (e.g. frequency converter). One problem in the realization of such devices is that in spite of the intrinsically strong SHG the length scale available for light-matter interaction is subnanometric, so that the resulting signal is weak. One proposed solution is to embed the 2D material in a resonant optical micro Fabry-Perot cavity. Prototypes of devices built on this principle showed enhancement of the SHG from one to three order of magnitude (Day *et al.*, 2016; Yi *et al.*, 2016).

Furthermore, a prototype of optical data storage has been proposed (Zeng *et al.*, 2015) which exploits the dependence of the SHG intensity on the number of layers in MoS₂. In particular for a coated metallic substrate the SHG is found to increase with number of layers up to a certain thickness (17 nm for gold) and then to decrease again. Then, by choosing the number of layers exhibiting the strongest SHG, information can be stored by locally reducing the number of layers and can be read out by detecting the SHG intensity.

References

- Boyd, R.W., 2008. Nonlinear Optics. Academic Press.
- Butler, S.Z., Hollen, S.M., Cao, L., *et al.*, 2013. Progress, challenges, and opportunities in two-dimensional materials beyond graphene. *ACS Nano* 7 (4), 2898–2926.
- Clark, D.J., Senthilkumar, V., Le, C.T., *et al.*, 2014. Strong optical nonlinearity of cvd-grown MoS₂ monolayer as probed by wavelength-dependent second-harmonic generation. *Phys. Rev. B* 90, 121409.
- Day, J.K., Chung, M.-H., Lee, Y.-H., Menon, V.M., 2016. Microcavity enhanced second harmonic generation in 2d MoS₂. *Opt. Mater. Express* 6 (7), 2360–2365.
- Grüning, M., Attacalite, C., 2014. Second harmonic generation in *h*-bn and MoS₂ monolayers: role of electron-hole interaction. *Phys. Rev. B* 89, 081102.
- Haug, H., Koch, S.W., 1990. Quantum Theory of the Optical and Electronic Properties of Semiconductors. World Scientific.
- Hsu, W.-T., Zhao, Z.-A., Li, L.-J., *et al.*, 2014. Second harmonic generation from artificially stacked transition metal dichalcogenide twisted bilayers. *ACS Nano* 8 (3), 2951–2958.
- Kim, C.-J., Brown, L., Graham, M.W., *et al.*, 2013. Stacking order dependent second harmonic generation and topological defects in *h*-bn bilayers. *Nano Letters* 13 (11), 5660–5665.
- Kumar, N., Najmaei, S., Cui, Q., *et al.*, 2013. Second harmonic microscopy of monolayer MoS₂. *Phys. Rev. B* 87, 161403.
- Li, Y., Rao, Y., Mak, K.F., *et al.*, 2013. Probing symmetry properties of few-layer MoS₂ and *h*-bn by optical second-harmonic generation. *Nano Letters* 13 (7), 3329–3333.
- Miyauchi, Y., Morishita, R., Tanaka, M., *et al.*, 2016. Influence of the oxide thickness of a SiO₂/Si(001) substrate on the optical second harmonic intensity of few-layer MoSe₂. *Japan. J. Appl. Phys.* 55 (8), 085801.
- Shen, Y.R., 2003. The Principles of Nonlinear Optics. Wiley Inter-science.
- Tang, Y., Mandal, K.C., McGuire, J.A., Lai, C.W., 2016. Layer- and frequency-dependent second harmonic generation in reflection from gase atomic crystals. *Phys. Rev. B* 94, 125302.
- Trolle, M.L., Seifert, G., Pedersen, T.G., 2014. Theory of excitonic second-harmonic generation in monolayer MoS₂. *Phys. Rev. B* 89, 235410.
- Wang, G., Marie, X., Gerber, I., *et al.*, 2015. Giant enhancement of the optical second-harmonic emission of wse₂ monolayers by laser excitation at exciton resonances. *Phys. Rev. Lett.* 114, 097403.
- Yi, F., Ren, M., Reed, J.C., *et al.*, 2016. Optomechanical enhancement of doubly resonant 2d optical nonlinearity. *Nano Letters* 16 (3), 1631–1636.
- Yin, X., Ye, Z., Chenet, D.A., *et al.*, 2014. Edge nonlinear optics on a MoS atomic monolayer. *Science (New York, NY)* 344 (6183), 488–490.
- Zeng, J., Yuan, M., Yuan, W., *et al.*, 2015. Enhanced second harmonic generation of MoS₂ layers on a thin gold film. *Nanoscale* 7, 13547–13553.

Parity-Time Symmetry in Optics

Mercedeh Khajavikhan, University of Central Florida, Orlando, FL, United States

Mohammad-Ali Miri and Andrea Alu, University of Texas at Austin, Austin, TX, United States

Demetrios N Christodoulides, University of Central Florida, Orlando, FL, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Refractive index, gain and loss play a pivotal role in light propagation. Modulating the refractive index as a means to mold the flow of light dates back to antiquity when rock crystals like the “Loupe of Sargon” were used to produce fire by focusing sun rays. Since then, index crafting has reached such level of sophistication that artificial structures as diverse as optical fibers, photonic crystals, and meta-materials became possible. Optical amplification or gain was realized much later, followed by the discovery of the laser, and has enabled numerous applications in many areas of science and technology. However, this is not the case for attenuation. Loss, unlike the other two processes, is still perceived to be a foe, an undesirable attribute that should be avoided or compensated at all costs. It is perhaps for this particular reason that up until recently the simultaneous use of index, gain, and loss as a viable route to achieve new optical behavior has been generally overlooked.

At a first glance, deliberately intermixing gain with loss may appear counterintuitive and perhaps pointless. Recently, however, there has been a growing evidence within the field of non-Hermitian optics, indicating that the presence of *exceptional points* (phase transition points) arising in judiciously designed parity-time symmetric systems can be utilized to build structures with customized properties and functionalities- previously thought to be unattainable. In what follows, we first provide a mathematical description of parity-time symmetry both in the context of quantum mechanics and optics, and then continue by overviewing some of the current developments concerning the use of PT-symmetry for laser mode management and sensing, while also looking into their anomalous scattering properties for designing invisibility cloaks.

Spontaneous Symmetry Breaking and Exceptional Points

The field of parity-time (PT) symmetry began in 1998 following the discovery by Bender and Boettcher, who first indicated that a wide class of non-Hermitian Hamiltonians can exhibit entirely real spectra, provided that they commute with the parity-time ($\hat{P}\hat{T}$) operator (Bender and Boettcher, 1998). While the use of non-Hermitian operators has been suggested in the past as a means to incorporate loss in quantum mechanics, Bender and Boettcher’s work has been largely perceived as counterintuitive, since it went against the commonly held view that real eigenvalues are only associated with Hermitian observables. Starting from the aforementioned premise, one can directly show that a necessary (but not sufficient) condition for PT symmetry to hold is that the complex potential involved in such a Hamiltonian should satisfy $V(x) = V^*(-x)$. In other words, the real part of the complex potential must be an even function of position, while its imaginary component should be anti-symmetric. In such pseudo-Hermitian configurations, the eigenfunctions are no longer orthogonal, i.e., $\langle m|n \rangle \neq \delta_{mn}$, and hence the vector space is skewed. Even more intriguing is the possibility for a sharp symmetry-breaking transition – once a non-Hermiticity parameter exceeds a certain critical value. In this latter regime, the Hamiltonian and the $\hat{P}\hat{T}$ operator no longer display the same set of eigenfunctions (even though they commute) and as a result, the eigenvalues of the system become partially, or entirely, complex. This broken PT-symmetry phase is associated with a so-called exceptional point (EP) or a non-Hermitian degeneracy.

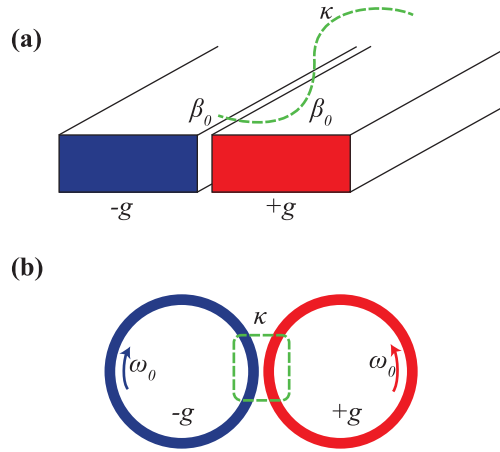
While the implications of these mathematical theories in quantum mechanics are still a matter of debate (since the realization of complex potentials in quantum mechanics is inherently out-of-reach), later it has been recognized that photonics can provide a fertile ground where PT symmetry concepts can be experimentally investigated (Makris *et al.*, 2008). The transition from quantum mechanics to optics can be formally justified by considering the isomorphism between the Schrödinger equation and the paraxial wave equation (see Table 1). In this regard, the complex refractive index profile plays the role of an optical potential: $k_0(n_R(x) + in_I(x))$. In optical systems, PT symmetry demands that the spatial distribution of the refractive index is an even function ($n_R(x) = n_R(-x)$), whereas the imaginary component (representing gain or loss) is an odd function of position ($n_I(x) = -n_I(-x)$). One can also show that in high contrast settings, where the electrodynamic problem must be treated fully vectorially, the condition for PT symmetry is expressed through the complex permittivity, e.g., $\epsilon(r) = \epsilon^*(-r)$. Nonetheless, since in most systems the imaginary part of the complex refractive index function is considerably smaller than its real counterpart, the PT symmetry condition in these two scenarios is almost equivalent.

To elucidate some the physics of exceptional points, here we provide a simple example of a two-level non-Hermitian PT symmetric system. In optics, this may be realized by two coupled resonators or waveguides, one exhibiting gain, while the other one experiencing an equal amount of loss. In the case of two cavities, energy exchange occurs in time, while for waveguides it takes place in space. Schematics of these configurations are depicted in Fig. 1.

Here we consider the case of two identical evanescently coupled cavities as depicted in Fig. 1(b). In order to simplify our analysis, we assume both resonators are single moded. For this system, the field evolution dynamics are described by the following

Table 1 Isomorphic relations between the Schrödinger equation in quantum mechanics and the paraxial wave equation in optics

	Quantum mechanics	Optics
Governing Equation	Schrödinger equation $i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x)\psi$	Paraxial wave equation $-i \frac{\partial E}{\partial z} = \frac{1}{2k} \frac{\partial^2 E}{\partial x^2} + k_0 n(x)E$
Wave function	$\psi(x, t)$	$E(x, z)$
Eigenvalues	Energy states	Propagation constants
Potential energy	Potential $V(x) = V_R(x) + iV_I(x)$	Refractive index $K_0 n(x) = K_0 n_R(x) + iK_0 n_I(x)$
PT-symmetry necessary condition	$V_R(x) = V_R(-x)$ $V_I(x) = -V_I(-x)$	$n_R(x) = n_R(-x)$ $n_I(x) = -n_I(-x)$

**Fig. 1** Two realizations of two-level parity-time symmetric optical systems, (a) a pair of coupled waveguides, one subjected to gain and the other to loss, trade energy along the direction of propagation at a rate of κ , (b) a pair of cavities, one experiencing gain and the other loss, exchange energy in time at a rate of κ .

set of equations:

$$\begin{cases} i \frac{da}{dt} - i \frac{g}{2} a + \kappa b = 0 \\ i \frac{db}{dt} + i \frac{g}{2} b + \kappa a = 0 \end{cases} \quad (1)$$

Here a and b represent the modal field amplitudes in the gain and loss cavities, respectively, g stands for the gain/loss coefficient, and κ is the coupling strength. This system supports two supermodes with distinct characteristics that depend entirely on the ratio between gain/loss (g) and coupling (κ). When $g/2\kappa < 1$, the two eigenvalues are real and are given by $\lambda_{1,2} = \pm \cos(\theta)$ where $\theta = \sin^{-1}(g/2\kappa)$ and the corresponding eigenvectors are $|1\rangle = [1 e^{i\theta}]^T$ and $|2\rangle = [1 - e^{-i\theta}]^T$. Note that these eigenvectors are not orthogonal in spite of the fact that the spectrum is real. In this case, none of the two modes experiences a net gain or loss, instead they oscillate, i.e., they remain neutral.

The spectral response of this PT-symmetric two-level system drastically changes as soon the gain-loss contrast exceeds the coupling strength ($g/2\kappa > 1$). In this regime, the eigenvalues are expressed by $\lambda_{1,2} = \pm \sinh(\theta)$ where $\theta = \cosh^{-1}(g/2\kappa)$, and their corresponding non-orthogonal eigenvectors turn out to be $|1\rangle = [1 i e^{\theta}]^T$ and $|2\rangle = [1 - i e^{-\theta}]^T$, indicating that the symmetry of each mode is broken in the sense that one of them resides mostly in the gain cavity while the other one in the lossy resonator. As a result, one of the eigenmodes enjoys amplification while the other experiences attenuation. Obviously, the supermode that mostly occupies the gain/loss region is amplified/attenuated. Fig. 2 displays the complex component of the eigenvalues as a function of $g/2\kappa$. The transition point $g/2\kappa = 1$ between these two different regimes is called the *exceptional point*. At this crossing, something quite remarkable happens. The dimensionality of the vector space is abruptly reduced from 2 to 1. This is evident from the fact that the two eigenvectors become entirely identical $[1 i]^T$ and neither oscillate nor exponentially varying. Instead the modal amplitudes depend algebraically on time ($\sim C_0 + C_1 t$).

Initial theoretical and experimental work carried out by several groups indicates that light propagating through non-Hermitian or PT-synthetic media can exhibit surprisingly unusual behavior. Among the numerous properties that such systems can display are: absorption enhanced transmittivity (Guo *et al.*, 2009; Rüter *et al.*, 2010), unidirectional invisibility (Lin *et al.*, 2011),

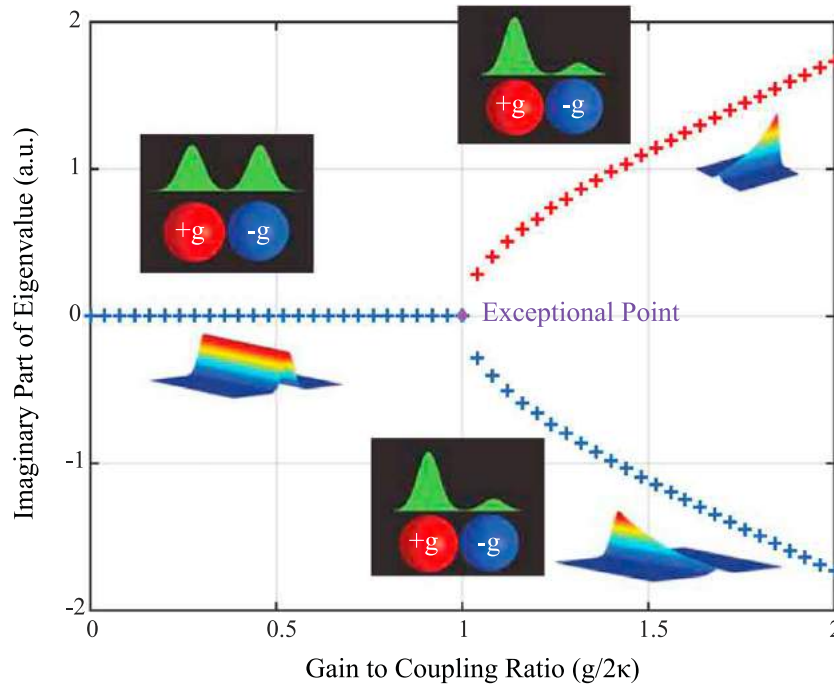


Fig. 2 The imaginary part of the eigenvalue vs. the change in the imaginary part of refractive index (gain-loss contrast) is depicted for a PT two-level system. At the exceptional point the symmetry of the modes breaks and two supermodes emerge, one amplifying and the other one attenuating.

unidirectional light transport (Peng *et al.*, 2014), topological chirality (Doppler *et al.*, 2016; Xu *et al.*, 2016), single mode lasing (Hodaie *et al.*, 2014; Feng *et al.*, 2014), and enhanced sensitivity at the exceptional point (Wiersig, 2014). In most cases, one may design the system close to the exceptional point such that one desired attribute can be extracted whereas other undesired properties remain below the threshold breaking. In what follows we provide an overview of the current developments in this field. In particular, we explain how the selective breaking of parity-time symmetry in multimode laser cavities can result in single mode operation. We will then look into the sensitivity of a system that is biased at an exceptional point and the anomalous scattering properties of PT-symmetric arrangements.

Mode Management in PT-Symmetric Lasers

Since the invention of diode lasers in 1962, there has been an ongoing effort to design optical cavities that funnel the maximum output power into a single spatiotemporal mode. Ultimately, a cavity becomes intrinsically single-moded only when its dimension is comparable to the wavelength of light. Most of the time, this criterion severely limits the amount of coherent emission that can be generated by a laser. On the other hand, when the active region is enlarged beyond this limit, most semiconductor lasers, because of their inhomogeneous lineshape, become susceptible to multimode operation something that eventually leads to output power fluctuations, instabilities, and lower beam quality. This explains why developing a compact, single mode laser has been the subject of numerous research activities over the past few decades. Such efforts have since resulted in several ingenious laser designs. For example, single transverse mode operation can be achieved using broad area slab coupled waveguide lasers (SCOWL). On the hand, distributed feedback (DFB), distributed Bragg reflector (DBR), and vertical cavity surface emitting lasers (VCSEL) have been designed to operate in a single longitudinal mode.

In recent years, the stringent requirements in photonic integrated circuits, have set new challenges for on-chip light sources. Not only such lasers must occupy a small size and exhibit a low threshold, they should also couple easily to the rest of the photonic network, preferably through a bus waveguide. In addition, cavity designs that rely on multi-step growth processes must be avoided. One family of lasers that can address these challenges are micro-ring lasers. Microring resonators have small footprint and as lasers they expected to show low thresholds because of their high quality factors. However, even a microring with a radius on the order of few tens of micrometer, tends to support several longitudinal modes across the gain bandwidth of most III-V active gain materials. Up until recently, the only systematic way to enforce single mode operation in such arrangements was through the incorporation of dispersive vertical gratings. Such a technique, however, seriously deteriorates the quality factor of the ring; hence increasing the threshold and reducing the slope efficiency. In what follows we will describe how selective PT symmetry breaking can be used to enforce single mode operation in microring lasers. This technique was first proposed by Miri *et al.* (2012) in the context of broad area lasers and amplifiers. In 2014, the use of PT symmetry for controlling the longitudinal mode contents of microring resonators

has been proposed by two independent groups. In one approach, the ring is patterned by a parity-time-symmetric grating structure (Feng *et al.*, 2014). In the other work, the active ring is paired with a lossy cavity to establish parity-time symmetry (Hodaie *et al.*, 2014). In what follows we will focus on the latter technique because of its simplicity in terms of avoiding additional gratings.

Single Longitudinal Mode Lasing

PT symmetry breaking can be elegantly exploited to establish single longitudinal mode operation in inherently multi-moded microring lasers. One way to do so, is by utilizing a coupled arrangement of two structurally identical ring resonators; one experiencing gain, and the other one an equal amount of loss. In general, when two fully identical resonators (both subjected to gain, loss, or neither) are placed next to one another, the degeneracy between their respective modes will be broken. The frequency splitting $\Delta\omega$ of the resulting supermode doublets $\omega_n^{(1,2)}$ is directly proportional to the coupling coefficient, i.e., $\Delta\omega = 2\kappa$ (Fig. 3(b)). On the other hand, in a PT-symmetric arrangement (like that depicted in Fig. 3(c)), the relationship between eigenfrequencies and coupling is by nature different, as indicated in the following expression:

$$\omega_n^{(1,2)} = \omega_n \pm \sqrt{\kappa_n^2 - g_n^2} \quad (2)$$

Consequently, any pair of modes, whose gain/loss contrast (g_n) remains below the coupling coefficient (κ_n), undergoes bounded neutral oscillations. However, as soon as g_n exceeds κ_n , a conjugate pair of lasing/decaying modes emerges. Consequently, the judicious placement of this PT threshold will allow a complete suppression of all non-broken mode pairs in favor of a single amplified mode associated with the aforementioned conjugate pair (Fig. 3(c)). As the imaginary parts of the eigenvalues diverge, degeneracy between their real parts is restored.

Even in the absence of PT-symmetry, any resonator with a spectrally non-uniform gain distribution $g(\omega)$ can in principle exhibit single-mode operation, provided that the loss exceeds the gain for all but one resonance. However, in this regime, the amplification cannot surpass the gain contrast $g_{\max} = g_0 - g_1$, where g_0 refers to the gain of the principal mode and g_1 to that of the next-strongest competing resonance. Obviously, this approach will impose severe constraints on the operating parameters, especially in the case of cavities having wide gain linewidths and/or closely spaced resonator modes, where $g_{\max}$ is very small. On the other hand, in a PT-symmetric setting, the coupling κ now plays the role of a virtual loss, and all undesirable modes must fall below its corresponding threshold. According to Eq. (2) we find that in this case the maximum achievable differential gain is:

$$g_{\max,PT} = \sqrt{g_0^2 - g_1^2} = g_{\max} \cdot \sqrt{\frac{g_0/g_1 + 1}{g_0/g_1 - 1}} \quad (3)$$

Given that $g_0 \geq g_1$, a selective breaking of PT symmetry can therefore systematically increase the available amplification for single-mode operation. As a matter of fact, the square-root behavior of this enhancement (a direct outcome of PT symmetry breaking), as characterized by the enhancement factor $G = g_{\max,PT}/g_{\max}$, is capable of providing substantially higher selectivity, especially when the initial contrast between adjacent modes is small ($g_1 \rightarrow g_0$). Consequently, this approach naturally exhibits a type of broadband self-adaptive single mode selection that is, in principle, applicable to any active resonator configuration. It should be stressed that this type of mode discrimination is resilient and remains valid even when PT symmetry does not exactly hold for all modes involved. On a fundamental level, the mechanism is related to the presence of an exceptional point in the system.

To experimentally verify these findings, active ring resonators based on InGaAsP quantum wells were defined using electron beam lithography and reactive ion etching (RIE). The gain and loss regions were determined by selective optical pumping at 1064 nm. The resonators were tested by placing them within a circularly symmetric Gaussian pump beam to implement three pumping configurations (single ring, evenly pumped double ring, and PT-symmetric ring arrangement). Accordingly, the effective pump powers were calculated from the geometric overlap between the active medium and the pump profile. Fig. 4(a,b) illustrate the behavior of a single ring (radius 10 μm , ring width 500 nm, height 210 nm) when exposed to an effective peak pump power of 2.5 mW (15 ns pulses with a repetition rate of 290 kHz). Under these conditions, at least four modes contribute significantly to lasing in the isolated ring. When two such rings are both placed at a distance of 200 nm from one another and exposed to the same pump power, one can clearly see the coupling-induced mode splitting (Fig. 4(c,d)), which occurs symmetrically around the resonance wavelengths of each ring in isolation. In this coupled regime, both structures are contributing equally. Once PT-symmetry is established, by withholding the pump from one of the resonators (Fig. 4(e,f)), lasing occurs exclusively in the active ring, where single-mode operation is now achieved. The presence of the lossy ring serves to suppress the unwanted longitudinal modes with a contrast exceeding 20 dB.

Comparing the light-light curves of these three lasing arrangements reveals that the slope efficiency in all scenarios remains virtually the same- indicating that the presence of lossy cavity in the PT-symmetric arrangement has no effect on the efficiency and output power. Moreover, the spectrally resolved light-light curves that compare the power in the fundamental mode of the single and the PT double ring resonators show that the PT system indeed offers superior performance, since the emission from the single ring includes contributions from several modes. This mechanism of mode selectivity using parity-time symmetry is robust in terms of fabrication inaccuracies and can accommodate active media with wide gain spectra. Moreover, as the occurrence of PT symmetry breaking is exclusively determined by the relation between net gain and coupling, the proposed arrangement is by nature self-adapting.

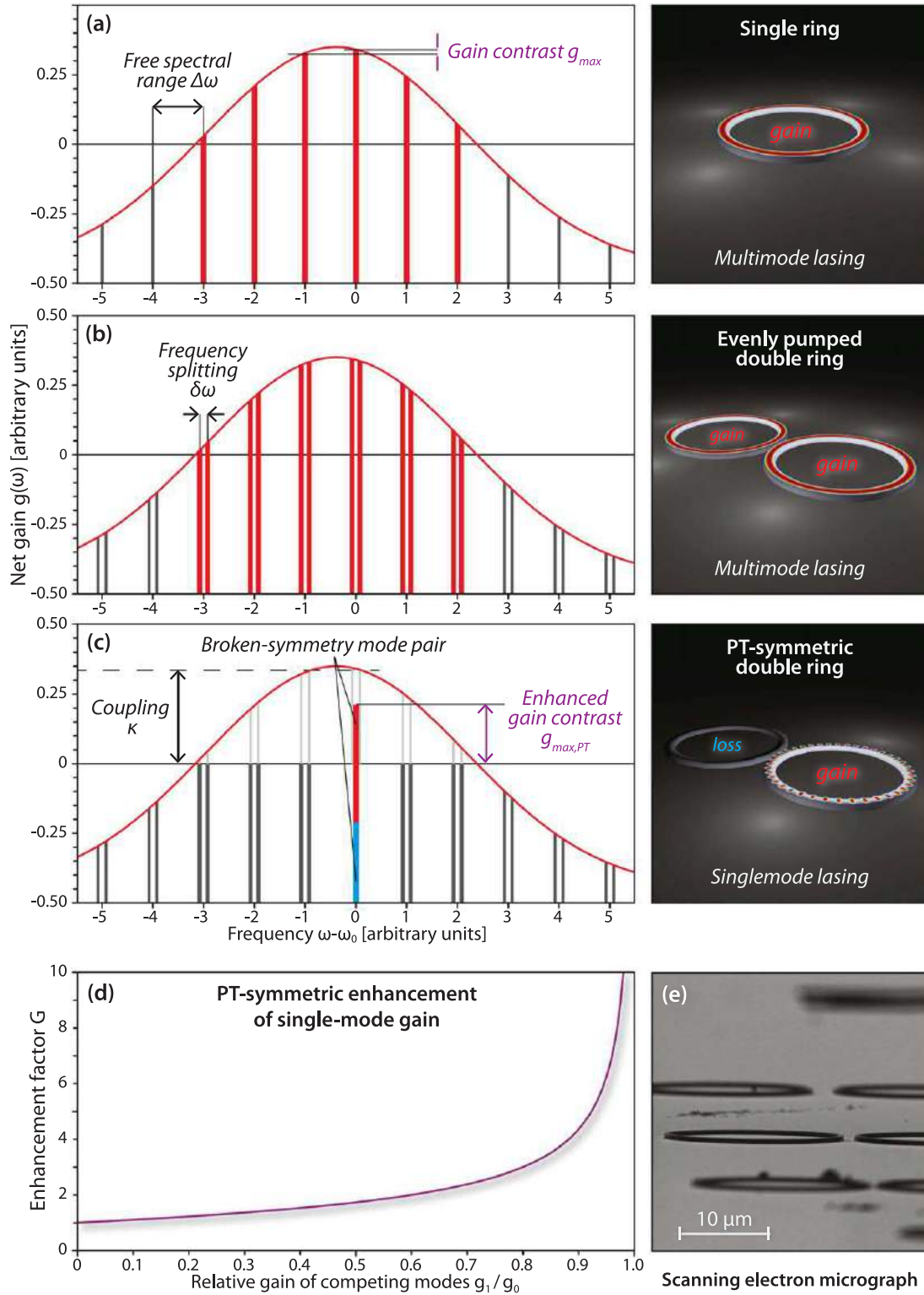


Fig. 3 Schematic principle of mode suppression in PT-symmetric microring lasers. (a), (b) In a coupled arrangement of two identical and evenly pumped rings, mode pairs emerge. (c), (d) PT-symmetry can be exploited to enforce stable single mode operation in otherwise multi-moded cavities. Reproduced from Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time-symmetric microring lasers. *Science* 346, 975–978.

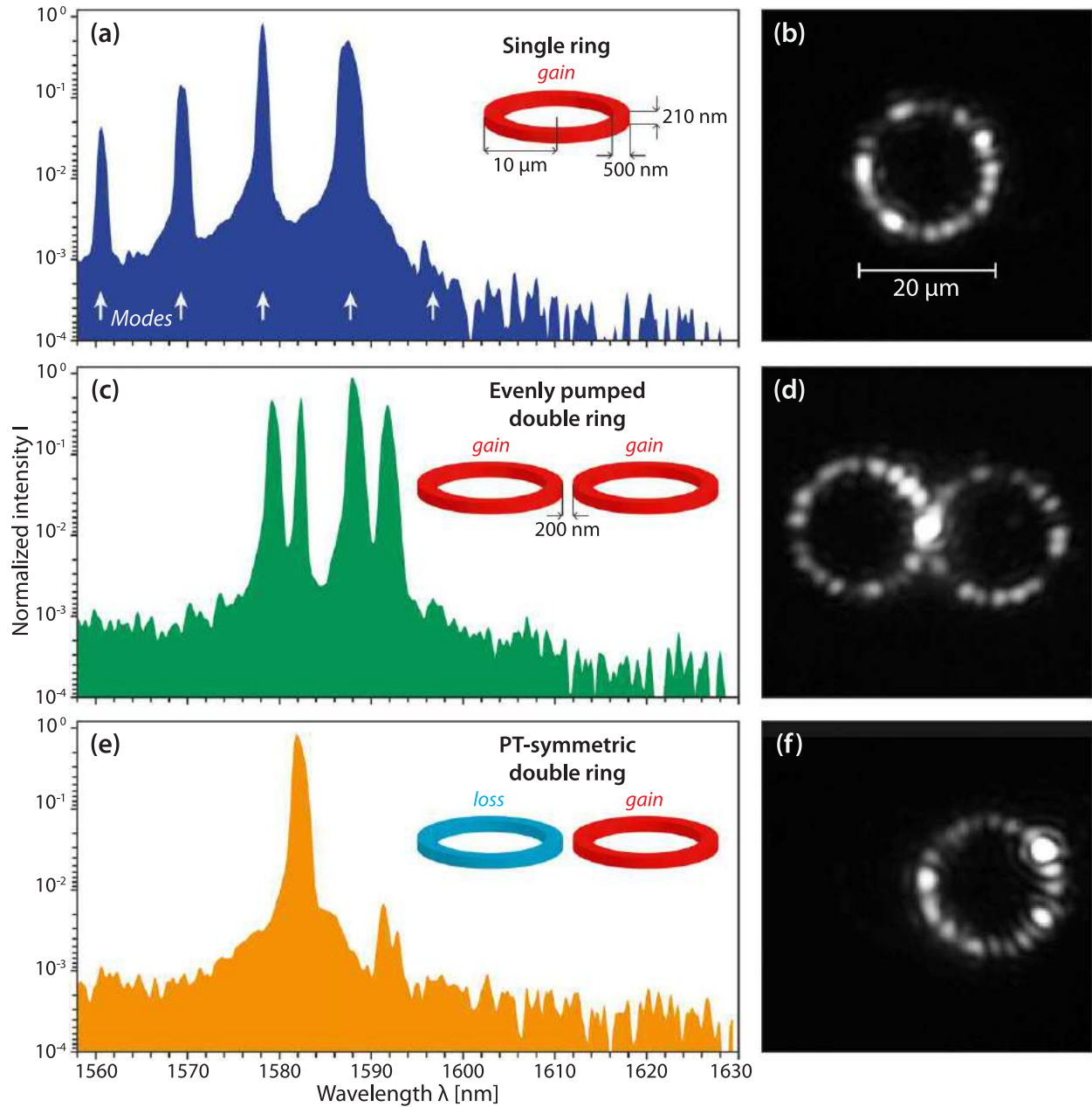


Fig. 4 Experimental observation of mode suppression by PT-symmetry breaking. (a) Emission spectrum of a single resonator. (b) Corresponding intensity pattern within the ring as observed from scattered light. (c) Spectrum obtained from an evenly pumped pair of such rings. (d) The intensity pattern shows that both resonators equally contribute. (e) Single-moded spectrum under PT-symmetric conditions. The mode suppression ratio exceeds 20 dB. (f) Lasing exclusively occurs in the active resonator. Reproduced from Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time-symmetric microring lasers. *Science* 346, 975–978.

Laser Transverse Mode Filtering

PT-symmetry can also be utilized in promoting the fundamental transverse mode in broad-area and multi-moded microring lasers (Hodaei *et al.*, 2016). In fact, as shown in Eq. (2), the virtual threshold at the exceptional point $g/\kappa=1$ introduces an additional degree of freedom: the coupling constant κ between the active and the lossy cavity mediated by the evanescent overlap of their respective modes. As it is well known, higher-order spatial modes systematically exhibit stronger coupling coefficients due to their lower degree of confinement. Consequently, in a PT symmetric arrangement, as the gain increases (when $g>\kappa$), the fundamental mode is the first in line to break its symmetry, thus experiencing a net amplification. On the other hand, for this same gain level, the rest of the modes retain an unbroken symmetry and therefore remain entirely neutral.

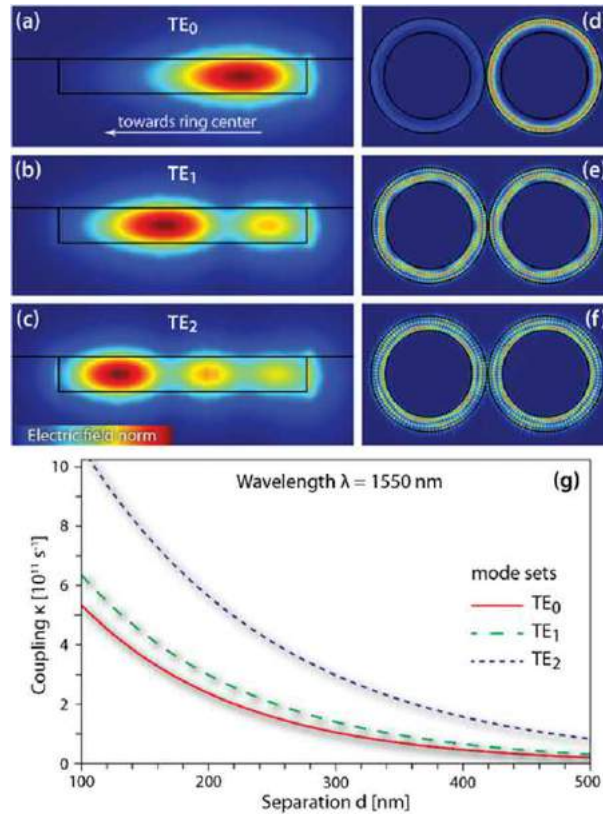


Fig. 5 (a)–(c) Intensity distributions in a microring resonator with a cross section $0.21 \mu\text{m} \times 1.5 \mu\text{m}$ and a radius of $R=6 \mu\text{m}$ as obtained by finite element simulations for the first three transverse modes, (d)–(f) intensity distribution of these same modes within the PT symmetric ring resonators. While the TE_0 mode operates in the broken PT symmetry regime and lases, all other modes remain in their exact PT phase and therefore they stay neutral, and (g) Exponential decay of the temporal coupling coefficients κ with cavity separation d . Higher order modes exhibit a much larger coupling coefficient than their lower-order counterparts. Reproduced from Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. *Laser & Photonics Reviews* 10 (3), 494–499.

Fig. 5(a)–(c) depicts the transverse intensity profiles of different spatial modes supported by a single microring resonator with a radius of $R=6 \mu\text{m}$ and waveguide dimensions of $0.21 \mu\text{m} \times 1.5 \mu\text{m}$. The curvature of the ring imposes a radial potential gradient, which deforms the mode fields into whispering-gallery-like distributions. Whereas the centroid of all modes shift towards the ring center, the exponential decay outside the ring still grows strongly with the mode order. As shown in **Fig. 5(d)–(f)**, for a certain coupling coefficient, set by the distance between the two rings, the transverse TE_0 field is the only mode to break its PT symmetry while all the higher order modes (TE_1 , TE_2) are still in the exact PT phase and hence occupy both rings equally. This behavior is also evident in **Fig. 5(g)** where the coupling strength between different transverse modes is depicted as a function of the separation between the two rings. For a fixed distance, the coupling coefficient increases with the order of the transverse mode, since the effective indices of higher order modes lie closer to that of the surrounding medium, allowing for stronger evanescent interactions across the cladding region. This trend persists for all wavelengths, and in conjunction with the difference in confinement enables PT symmetry breaking to be employed as a mode-selective virtual loss.

Fig. 6(a) and **(b)** illustrate the transition from multimode behavior in a broad area microring laser to single mode operation in a twin-ring configuration as enabled by preferential PT symmetry breaking. These experiments were conducted in high-contrast active ring resonators based on quaternary InGaAsP (Indium-Gallium-Arsenide-Phosphide) multiple quantum wells embedded in SiO_2 (silicon dioxide). The gain bandwidth of the active medium spans the spectral region between 1260 and 1590 nm. Whereas the quantum wells are present in all wave-guiding sections, gain and loss is provided by selectively pumping the respective rings (pump wavelength: 1064 nm). **Fig. 6(a)** shows the spectrum obtained from a single microring laser. As expected from the differences in confinement between mode sets, only the TE_0 and TE_1 and TE_2 modes undergo lasing. In an isolated ring, a decrease of the overall pump power does not yield single mode operation. On the other hand, when the active ring of the PT-symmetric double-ring arrangement is supplied with the same pump power as in **Fig. 6(a)**, the TE_1 and TE_2 modes are readily eliminated (see **Fig. 6(b)**) as they fall below the PT breaking threshold and therefore experience zero net gain. The corresponding resonances are suppressed with a fidelity of over 30 dB down to the noise floor of the measurement.

The PT transverse mode filtering mechanism described above is fundamentally different from the standard spatial mode filtering techniques that are typically used in lasers. For example, intra-cavity apertures or spatial filters have long been used to

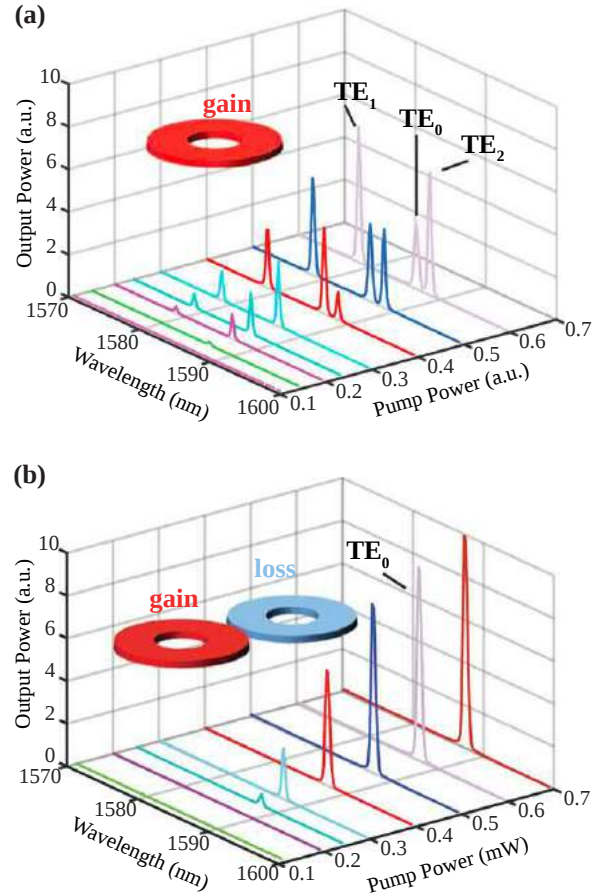


Fig. 6 (a) A single microring resonator lases in various TE₀, TE₁ and TE₂ resonances, (b) in contrast, a PT-symmetric double ring arrangement reliably suppresses all the TE₁ and TE₂ modes with a contrast exceeding 30 dB. Reproduced from Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. *Laser & Photonics Reviews* 10 (3), 494–499.

limit the presence of higher order transverse modes by introducing differential loss between modes. However, in such schemes, all modes including the fundamental one, experience loss, though to a varying degree. Therefore, the resulting mode suppression obtained by such techniques is often incremental and comes at the expense of undesirable losses in the fundamental mode. In contrast, the abrupt onset of PT symmetry breaking at the exceptional point relocates the selected mode to the active region, whereas all higher order modes remain equally distributed between loss and gain regions. Consequently, this method offers a high degree of mode discrimination, without any detrimental impact on the overall lasing efficiency or output power.

Another advantage of PT-symmetric laser cavities is their robustness with respect to nonlinearity-induced instabilities. Conventional broad-area weakly-guiding single-mode lasers often are subject to severe filamentation even at moderate intensities. In contrast, the influence of such effects on the above system remains negligible. The observed robustness against nonlinear perturbations can be intuitively understood by keeping in mind that in the microring systems examined here, the index contrast between the core and cladding regions is significant and hence for all practical purposes is not affected by nonlinearity-induced index changes. For example, for circulating powers in the range of tens of mW, the resulting change in the refractive index due to self-focusing properties of the semiconductor gain material is estimated to be at most on the order of 10^{-3} . This level of nonlinearity will be an issue only for much more densely packed sets of transverse modes, e.g., when the width of the waveguides involved is substantially increased.

In conclusion, single spatial mode operation can be effectively enforced in parity-time symmetric broad area coupled cavities. Unlike most other schemes for laser transverse mode control, this method establishes mode selectivity through coupling to a lossy cavity without compromising the optical power extracted from the fundamental mode. Following the initial experiments, more recently, this approach has been tested in the context of two coupled waveguide lasers. This technique is versatile, scalable and can be applied to a wide range of broad area laser systems.

Sensing at Exceptional Points

Degeneracy occurs ubiquitously in nature. Within the context of eigenvalue problems, this property is encountered in a broad range of physical disciplines and is behind some of the most intriguing phenomena observed in both classical and quantum

physics. In terms of applications, through lifting an existing degeneracy, a small perturbation can lead to a detectable splitting between the corresponding eigenvalues. This principle has been utilized as a measurement tool in many and diverse settings. Similarly, in optics, breaking the degeneracy of resonant frequencies can be used as a means to detect small changes in the refractive index and/or loss. Along these lines, ultrahigh Q microcavities have been developed that nowadays are regarded as one of the most promising photonic arrangements for a variety of extreme sensing applications, ranging from single molecule detection to recording minute rotation rates.

The action of a perturbation on the degeneracy of a conservative (Hermitian) system is well understood. A point in parameter space at which such a degeneracy occurs is called a diabolic point (DP). At such points, a disturbance of strength ε leads to energy shifts and splittings that are at most on the order of ε . In non-Hermitian settings, however, another type of degeneracy is possible: that associated with an exceptional point. At this point, the dimensionality of the arrangement is abruptly reduced and as a result, the reaction of a system to perturbations around an EP is more drastic as compared to that at a DP. In fact, a perturbation of strength ε acting on a second-order EP (when only two eigenvalues and eigenvectors merge) so happens to result to an energy splitting that is proportional to $\varepsilon^{1/2}$. This implies that the sensitivity of a sensing set-up can be enhanced by several orders of magnitude (since $\varepsilon^{1/2} \gg \varepsilon$ for small perturbations) by exploiting the physics of EPs. Furthermore, one can show that the enhanced sensing of non-Hermitian systems is not limited to that offered by $\varepsilon^{1/2}$. By using even higher-order exceptional points (EP-Ns), it is possible to further boost the sensitivity up to $\varepsilon^{1/N}$. For example, an EP system of order $N=5$ can “amplify” a disturbance on the order of 10^{-10} up to 10^{-2} , an eight order improvement over that expected from a Hermitian sensor.

Exceptional points appear in various non-Hermitian settings, where there is coupling between resonances of the system while they are subjected to a spatially non-uniform gain and/or loss distribution. As of today, however, perhaps one of the most elegant ways to systematically generate them is through PT symmetry. In applications pertaining to sensing, a key advantage of PT-symmetric systems is their high sensitivity to external perturbations. Such a response not only facilitates the experimental observation of the square-root behavior as suggested above, but it can also provide a practical route to increase the sensitivity of microcavity sensors. Fig. 7 depicts schematics of such non-Hermitian sensor arrangements comprising of coupled resonators. The cavities are identical in shape and size, but are subject to different levels of gain and/or loss. The contrast between the amplification levels experienced by the cavities can be introduced, for example, through preferential pumping. Although strictly speaking, such systems may not be invariant under the simultaneous action of parity (P) and time (T) operators, they can nevertheless be transformed into becoming PT symmetric once they are gauged through a constant gain/loss bias. Fig. 7(d) shows how the sensitivity in such arrangements can be enhanced depending on the order of the exceptional point involved.

The mathematical explanation behind the enhanced sensitivity expected from a PT symmetric coupled cavity configuration (when biased at an exceptional point) can be obtained, for example, by considering the modal behavior of the structure depicted in Fig. 1(b). In general, each ring, when uncoupled, can support a number of longitudinal modes both in the clockwise and counterclockwise directions. Without loss of generality, here we limit our analysis to a single longitudinal mode in one direction. We also assume that the cross-section of the rings are designed so as to support only the fundamental TE mode. In order to analyze the sensitivity response, we assume that a small perturbation is applied to the ring resonator subjected to gain. In this respect, the interplay between the electric modal fields in the two identical rings can be effectively described through a set of

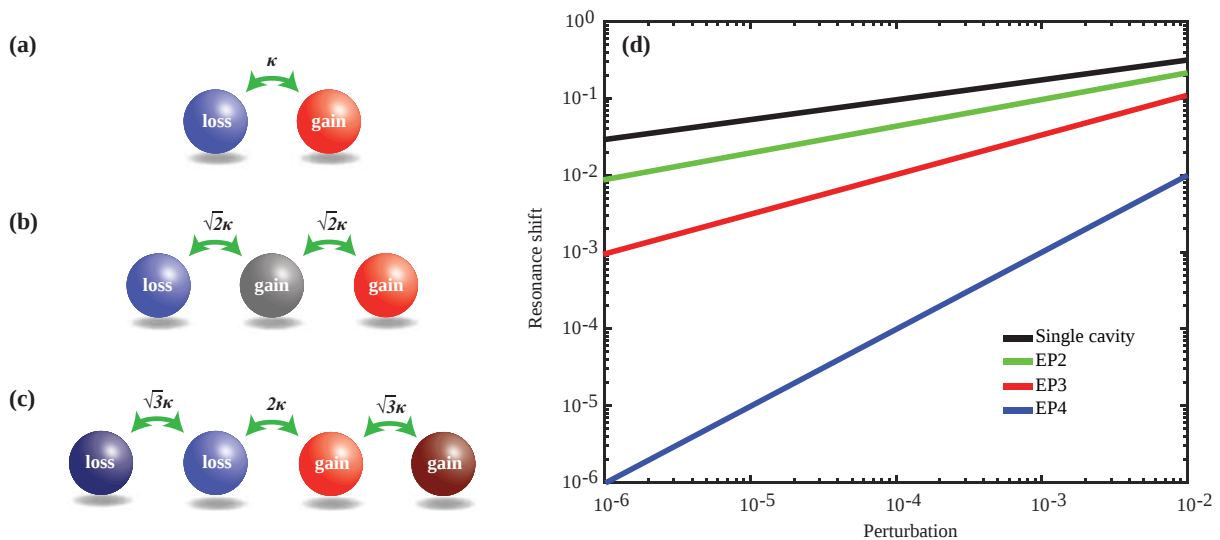


Fig. 7 A schematic of three photonic PT molecules exhibiting (a) 2nd, (b) 3rd, and (c) 4th order exceptional points. Red cavities exhibit gain, blue loss, and gray are neutral. (d) The dependence of sensitivity to perturbation strength.

time dependent coupled equations:

$$\begin{cases} \frac{da}{dt} = -i(\omega_0 + \epsilon)a + \kappa b + g_1 a \\ \frac{db}{dt} = -i\omega_0 b + \kappa a + g_2 b \end{cases} \quad (4)$$

Here, the cavities are treated as lumped resonators, a and b are modal amplitudes for a particular longitudinal resonance, κ and $g_{1,2}$ represent the coupling factor between the resonators and their associated gain/loss values, respectively, and ϵ is an intentionally induced perturbation. The two eigenvalues associated with this set of equations are given by the following expression:

$$\omega_{1,2} = \omega_0 + \epsilon/2 \pm \sqrt{\kappa^2 - (\Delta g/2 + i\epsilon/2)^2} \quad (5)$$

where $\Delta g = |g_1 - g_2|$ represents the gain-loss contrast between the resonators. If $\kappa = \Delta g/2$, the system is known to be at the exceptional point. Under this condition, Eq. (2) can be rewritten as:

$$\omega_{1,2} = \omega_0 + \epsilon/2 \pm i\sqrt{\epsilon\kappa/2} \quad (6)$$

In the absence of any perturbation ($\epsilon = 0$), the two frequencies are expected to be exactly equal to ω_0 . This is a large departure from a standard Hermitian system in which the two resonances are expected to split by 2κ . Furthermore, when operating at the exceptional point ($\kappa = \Delta g/2$), an induced perturbation ϵ leads to a $\Delta\omega_{PT} = \sqrt{2\kappa\epsilon}$ splitting in the real part of the eigen-frequencies. Therefore, by monitoring the splitting between the two resonances in the spectral domain, one can effectively determine the magnitude of the applied perturbation. Clearly, in this two-level system, the corresponding wavelengths diverge according to a square-root function of the perturbation ($\Delta\lambda \sim \epsilon^{1/2}$). Compared to a Hermitian or Hermitian-like single cavity configuration, the aforementioned square root behavior can substantially amplify the response of the system to a small perturbation (Chen *et al.*, 2017; Hodaei *et al.*, 2017). This approach has been recently suggested as a means to devise ultrasensitive microscale gyroscopes (Ren *et al.*, 2017).

Anomalous Scattering Properties of PT-Symmetric Arrangements

At a first sight, it may seem that balanced regions of gain and loss in a PT-symmetric system cannot affect the propagation of light in such settings, as the net amplification/attenuation in such systems is zero. However, as shown in several recent studies, the inter-mixing of gain and loss can significantly alter the dynamics of light in PT systems and can lead to a host of intriguing phenomena (Miri *et al.*, 2016). In order to shed light on anomalous scattering properties of PT-symmetric arrangements, here we focus on a PT-symmetric grating defined as follows (Lin *et al.*, 2011; Regensburger *et al.*, 2012):

$$n(z) = n_0 + n_R \cos(k_B z) + i n_I \sin(k_B z) \quad (7)$$

where, $k_B = 2\pi/\lambda_B$ and λ_B is the grating period. Clearly, this refractive index profile satisfies the necessary condition of PT symmetry, i.e., $n^*(-z) = n(z)$. Under weak refractive index and gain/loss modulations, i.e., $n_R, n_I \ll n_0$, this system is analytically solvable and its dynamics is governed by the coupling between the forward and backward propagating waves, as well as the gain/loss contrast. To show this, we consider a solution of the form $E(x, t) = E_f(x, t)e^{-i(\omega_0 t - k_0 n_0 z)} + E_b(x, t)e^{-i(\omega_0 t + k_0 n_0 z)}$, where E_f and E_b represent the slowly varying envelopes of the forward and backward propagating waves, ω_0 is a central frequency and $k_0 = \omega_0/c$ is the free space wavenumber. It is straightforward to show that this ansatz leads to the following simple coupled mode equations:

$$\left(\frac{\partial E_f}{\partial z} - \frac{1}{v} \frac{\partial E_f}{\partial t} \right) = +i(\kappa + g)e^{-i2\delta z} E_b \quad (8a)$$

$$\left(\frac{\partial E_b}{\partial z} + \frac{1}{v} \frac{\partial E_b}{\partial t} \right) = -i(\kappa - g)e^{+i2\delta z} E_f \quad (8b)$$

where, in these relations, $v = c/n_0$ represents the phase velocity in the background medium, $\delta = k_0 n_0 - k_B/2$ shows the detuning in the grating-assisted phase matching of the forward and backward propagating waves, and finally $\kappa = k_0 n_R/2$ and $g = k_0 n_I/2$ are two coupling coefficients associated with the real and imaginary parts of the complex grating. An interesting property of the coupled mode Eq. (8) is an asymmetry in the coupling between the forward and backward waves; according to these relations, the forward wave is coupled to the backward wave at rate $\kappa + g$, while the reverse process occurs at rate $\kappa - g$. Quite interestingly, for $\kappa = g$, the backward wave becomes completely independent from the forward. As a result, the complex grating will not have any influence on a wave propagating toward the left.

Under the ansatz of $(E_f, E_b) = (F_0 e^{-i\delta z}, B_0 e^{i\delta z}) e^{i\Omega t} e^{i(Kz - \Omega t)}$, the band structure of the PT grating is found to be:

$$\Omega^2/v^2 = K^2 + \kappa^2 - g^2 \quad (9)$$

This dispersion diagram is shown in Fig. 8 for different ratios of g/κ . According to this figure, for $\kappa > g$ there is a finite band gap $\Delta\omega = 2v\sqrt{\kappa^2 - g^2}$. By increasing the gain/loss, the band gap reduces such that for $\kappa = g$ the dispersion relation becomes the same as a the background medium, i.e., $\Omega/v = \pm K$. Finally, for $\kappa < g$, the dispersion diagram flips and regions with complex eigen-frequencies, associated with amplification and attenuation, emerge.

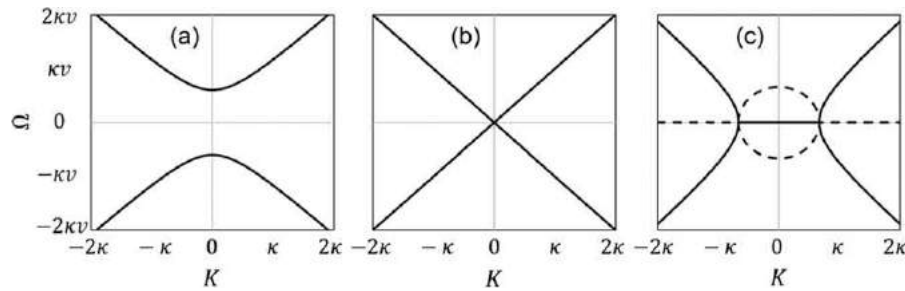


Fig. 8 The dispersion diagram of a PT-symmetric grating for (a) $g=0.8\kappa$, (b) $g=\kappa$, and (c) $g=1.2\kappa$. The dashed line in part (c) shows the imaginary part of Ω .

Such periodic PT-symmetric structures can exhibit surprising behavior such as unidirectional invisibility and unusual reflection characteristics. More specifically, light propagating in such a system can experience reduced or enhanced reflections (with a coefficient that can even exceed unity) depending on the direction of light propagation. This can be understood by considering that the left–right symmetry is now broken in this PT configuration and propagation is no longer the same when the sequence of the gain and loss regions is exchanged. Even more interesting is what happens at the PT threshold: light waves entering the structure from one side do not experience any reflection and can fully traverse the grating with complete transmission. Given that this occurs without acquiring any phase imprint from the PT grating, the periodic structure is essentially invisible. The situation is very different when light is incident from the other side.

See also: Nonlinear Optical Phase Conjugation

References

- Bender, C.M., Boettcher, S., 1998. Real spectra in non-Hermitian Hamiltonians having PT symmetry. *Physical Review Letters* 80 (24), 5243.
- Chen, W., Özdemir, S.K., Zhao, G., Wiersig, J., Yang, L., 2017. Exceptional points enhance sensing in an optical microcavity. *Nature* 548, 192–196.
- Doppler, J., Mailybaev, A.A., Böhm, J., *et al.*, 2016. Dynamically encircling an exceptional point for asymmetric mode switching. *Nature* 537 (7618), 76–79.
- Feng, L., Wong, Z.J., Ma, R.M., Wang, Y., Zhang, X., 2014. Single-mode laser by parity-time symmetry breaking. *Science* 346 (6212), 972–975.
- Guo, A., Salamo, G.J., Duchesne, D., *et al.*, 2009. Observation of PT-symmetry breaking in complex optical potentials. *Physical Review Letters* 103 (9), 093902.
- Hodaei, H., Hassan, A.U., Wittek, S., *et al.*, 2017. Enhanced sensitivity in photonic molecule systems using higher order non-Hermitian exceptional points. *Nature* 548, 187–191.
- Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. *Laser & Photonics Reviews* 10 (3), 494–499.
- Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time-symmetric microring lasers. *Science* 346, 975–978.
- Lin, Z., Ramezani, H., Eichelkraut, T., *et al.*, 2011. Unidirectional invisibility induced by PT-symmetric periodic structures. *Physical Review Letters* 106 (21), 213901.
- Makris, K.G., El-Ganainy, R., Christodoulides, D.N., Musslimani, Z.H., 2008. Beam dynamics in PT symmetric optical lattices. *Physical Review Letters* 100 (10), 103904.
- Miri, M.A., Eftekhari, M.A., Facao, M., *et al.*, 2016. Scattering properties of PT-symmetric objects. *Journal of Optics* 18 (7), 075104.
- Miri, M.A., LiKamWa, P., Christodoulides, D.N., 2012. Large area single-mode parity-time-symmetric laser amplifiers. *Optics Letters* 37 (5), 764–766.
- Peng, B., Özdemir, S.K., Lei, F., *et al.*, 2014. Parity-time-symmetric whispering-gallery microcavities. *Nature Physics* 10 (5), 394–398.
- Regensburger, A., Bersch, C., Miri, M.A., *et al.*, 2012. Parity-time synthetic photonic lattices. *Nature* 488 (7410), 167–171.
- Ren, J., Hodaei, H., Harari, G., *et al.*, 2017. Ultrasensitive micro-scale parity-time-symmetric ring laser gyroscope. *Optics Letters* 42 (8), 1556–1559.
- Rüter, C.E., Makris, K.G., El-Ganainy, R., *et al.*, 2010. Observation of parity-time symmetry in optics. *Nature Physics* 6 (3), 192–195.
- Wiersig, J., 2014. Enhancing the sensitivity of frequency and energy splitting detection by using exceptional points: Application to microcavity sensors for single-particle detection. *Physical Review Letters* 112 (20), 203901.
- Xu, H., Mason, D., Jiang, L., Harris, J.G.E., 2016. Topological energy transfer in an optomechanical system with exceptional points. *Nature* 537 (7618), 80–83.

Laser-Induced Damage in Optical Materials

Wolfgang Rudolph and Luke A Emmert, University of New Mexico, Albuquerque, NM, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Shortly after the invention of the laser, in the early 1960s, it became apparent that the ability of optical materials to withstand high laser intensity and fluences was of utmost importance for advances in laser development and nonlinear optics. Despite considerable progress in the past decades (Wood, 2003; Ristau, 2015) the physics and materials science of the interaction of high-intensity lasers with optical materials is still a very active and challenging research area. Laser-induced damage (LID) and laser ablation rest on similar physical foundations and describe the phenomena of irreversible material modification. Laser induced damage of surfaces and films is often associated with the removal of material.

The defining parameter is the laser-induced damage threshold (LIDT). Its value is often published as a fluence (J cm^{-2}) for pulsed laser systems and as an intensity (W cm^{-1}) for continuous wave (CW) lasers. The CW LIDT is valid for exposures long enough to reach steady-state temperatures. Typical time scales range from microseconds to seconds depending on the spot size. Solutions to the thermal diffusion equation reveal that the maximum temperature reached is proportional to the incident power divided by the beam radius (Ristau, 2015, see Chapter 2) for uniform bulk and surface absorption. Pulse LID is the subject of this article. LIDT values depend on material type (metals, semiconductor, insulators), topology (bulk, surface, thin film, multilayer), preparation (surface polishing, film deposition method, annealing, etc.), as well as laser parameters (wavelength, pulse duration, spot size, repetition rate, etc.). They also depend on how the damage test is performed. For example, if each test pulse illuminates a new sample site (1-on-1 test) or if several (S) pulses illuminate the same sample spot (S-on-1) (ISO 11254-2, 2011). Being transparent dielectrics have the highest LIDTs of all materials for lasers from the near infrared to the near ultraviolet spectral region.

Fig. 1 illustrates the material response of a dielectric sample as a function of the incident fluence. Often the LIDT behavior of a material is characterized by the probability that damage is observed when a pulse of certain fluence is incident, $P(F)$. There is a fluence range that separates catastrophic damage (for instance a surface crater) from reversible material modifications such as refractive index changes. The width of this transition region ΔF is narrow for fs pulses and becomes increasingly broader for ps, ns, and longer pulses. Different physical origins of damage are responsible – deterministic damage initiation for short pulses governed by intrinsic material properties and probabilistic, extrinsic-defect controlled damage for longer pulses. It should be noted that the physical mechanisms responsible for optical damage also enable laser machining and material processing.

The LIDT F_{th} is defined as the minimum fluence for damage to occur, i.e., $P(F \leq F_{th}) = 0$ (Porteus and Seitel, 1984). An optic will remain undamaged provided the incident fluence $F < F_{th}$. Because of the statistical nature of laser-induced damage there are practical difficulties in determining F_{th} so a damage fluence F_d , where $P(F = F_d) = 0.5$, is often used in the scientific literature. For short pulse LID, where the transition region ΔF is narrow, F_{th} and F_d are nearly the same. LIDT can be different for 1-on-1 and S-on-1 tests. Typically, the multiple pulse LIDT is smaller. There are also cases where the damage threshold increases when the sample is illuminated with a laser. This laser conditioning is used in high-power laser systems by gradually ramping up the laser power. Annealing of defects is a possible mechanism enabling laser conditioning (Bercegol, 1998).

Table 1 lists LIDTs for representative materials and pulse durations. Scaling laws (for dielectrics in particular) are known to extrapolate to other materials and pulse durations (Ristau, 2015, see chapter 5).

To discuss issues related to LIDT we refer to the schematic diagram of the morphology of an optical surface (or film) shown in Fig. 2.

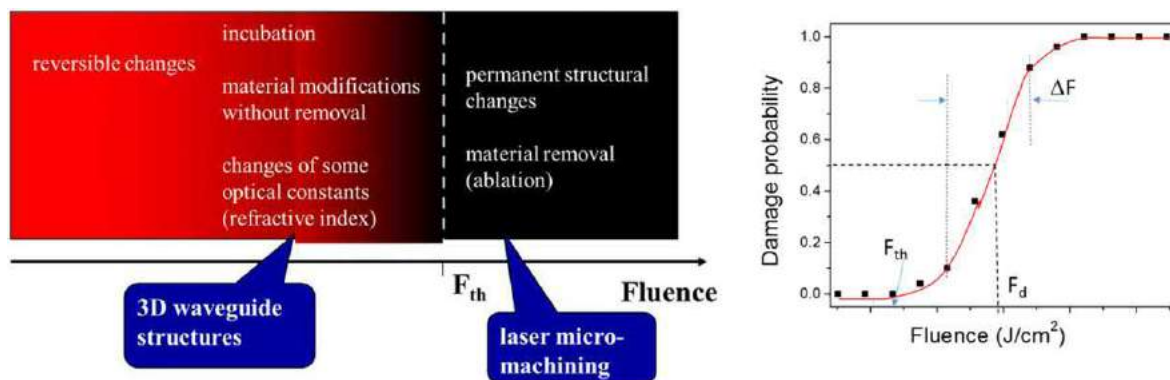


Fig. 1 Response of a dielectric material as a function of incident pulse fluence. The graph on the right shows a typical damage probability as a function of incident pulse fluence measured with femtosecond pulses that probe intrinsic material properties (F_{th} , F_d , and ΔF are defined in the text.).

Table 1 Representative 1-on-1 LIDTs for selected optical components, pulse durations and wavelengths

Optics	Pulse duration	Wavelength	LIDT (J cm^{-2})	References
Fused silica (bulk)	100 fs	800 nm	~ 4	Lenzner <i>et al.</i> (1998)
	10 ns	1064 nm	5000	Smith and Do (2008)
Dielectric HR mirror	100 fs	800 nm	0.2–1.0	Stolz <i>et al.</i> (2009)
	10 ns	1064 nm	5–140	Stolz <i>et al.</i> (2008)
Dielectric polarizer	10 ns	1064 nm	18	Stolz and Runkel (2012)
Metal (Al) film	100 fs	800 nm	0.3	Sun <i>et al.</i> (2015)

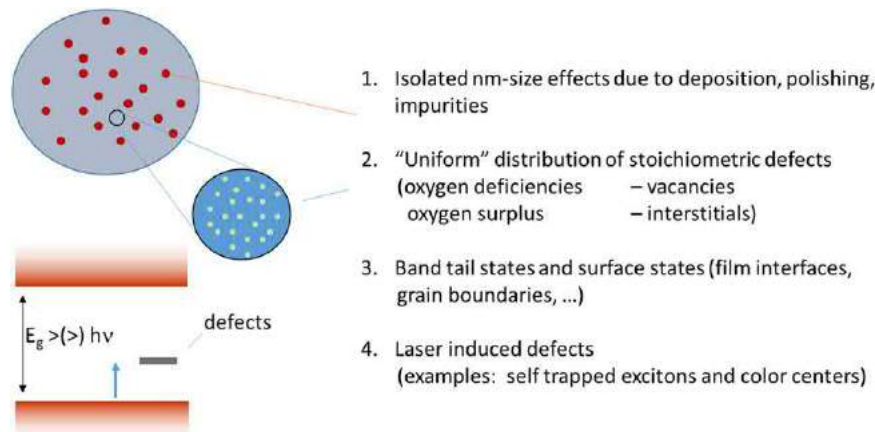


Fig. 2 Schematic diagram of the morphology of an optical wide-gap material. In high-quality materials, the density of defects of type 1 is low. Defects of type 2 and 3 are considered intrinsic. They determine macroscopic material properties probed by an incident laser. Additional defects (type 4) can be created during sub-threshold laser irradiation, for example, by a train of pulses. Defects can introduce states within the bandgap of the host material, which makes optical excitation more efficient, lowering the LIDT.

A laser excitation spot will typically contain several defects of type 1. Controlled by such statistically distributed defects, long-pulse LIDTs are stochastic and depend on spot size (DeShazer *et al.*, 1973). For sufficiently small spot sizes LIDTs representing intrinsic material properties are more likely to be observed. When measuring LIDTs of bulk materials care has to be taken to avoid or properly take into account propagation effects like self-focusing and dispersive effects in particular for short pulses. Characterization of damage controlling defects remains a challenge. Nondestructive techniques often lack sensitivity when the defect density is low.

Because of their importance for high power optical components dielectrics have received a lot of attention and are the focus of this encyclopedia entry. The physical processes leading to visible damage depend on the material and illumination conditions. In all cases a sufficiently large energy density has to be deposited first, which then triggers the material response that leads to the damage signature. This is illustrated in Fig. 3.

Fundamental Processes of Laser-Induced Damage

The intrinsic laser-induced damage mechanism of wide bandgap dielectrics is rather well understood (Emmert and Rudolph, 2015 and references therein). From the femtosecond to the nanosecond pulse regime, LIDT is reached when the excitation pulse has created an electron plasma of sufficient energy density in the conduction band. The processes leading to this electron plasma involve multiphoton ionization (interplay of multiphoton absorption and tunneling) of valence band electrons and impact ionization. The relative importance of these two processes has long been a question (Jones *et al.*, 1989) and remains an active research topic even today (Rajeev *et al.*, 2009; Karras *et al.*, 2011; Mouskeftaras *et al.*, 2013).

Damage will occur if the density of energy deposited into the conduction band plasma reaches some critical value. Its exact value is not *a priori* clear and depends, among other things, on the pulse duration. If damage were purely thermal, the critical energy would be related to the evaporation enthalpy. This, however, is rarely the case. A large enough density of conduction band electrons (electron plasma) can destabilize the lattice – similar to the excitation of a molecule to a non-binding state.

Dielectric breakdown used in modeling LID provides a useful proxy for the critical energy as a damage criterion (Bloembergen, 1974). The electron plasma can linearly absorb light through “free” carrier absorption also known as inverse Bremsstrahlung. With increasing absorption, the skin depth in which energy is deposited becomes smaller amplifying the rate of the energy deposition even more. Dielectric breakdown occurs at the plasma density (i.e., critical electron density) at which the plasma frequency

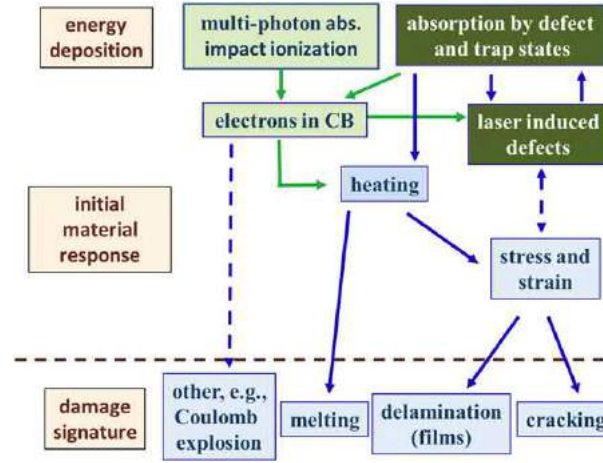


Fig. 3 Laser-induced damage in optical materials. Energy deposition is followed by an initial material response leading to visible, irreversible changes.

coincides with the laser frequency. The success of this criterion also stems from the fact that because of the exponential increase in plasma density during the incident pulse, the predicted threshold pulse fluence depends only logarithmically on the numerical value of the critical electron density.

Rate equations that model the electron density in the conduction have been quite successful in explaining the main features of intrinsic LIDT in particular in the short pulse regime. The principal structure of such a rate equation is of the form

$$\frac{d}{dt}n_e = K(I) + A(I, n_e) - L(n_e) \quad (1)$$

where n_e is the conduction band electron density and I is the laser intensity. The first term, $K(I)$, is the multiphoton ionization rate. It represents a combination of multiphoton absorption and tunneling described by Keldysh (1965). The second term, $A(I, n_e)$, describes impact (avalanche) ionization. This is the process where a highly excited electron in the conduction band (CB) transfers its energy to a valence band electron through collision, promoting it to the CB. This collision results in two electrons near the bottom of the CB. Repeating this process leads to exponential growth of n_e (avalanche ionization). This term is often approximated by an_e (from the flux-doubling model, which assumes that impact ionization takes place as soon as a CB electron has acquired sufficient energy (Stuart *et al.*, 1996)), where a is the avalanche ionization coefficient. This coefficient is usually taken as constant, but attempts have been made to measure an intensity dependence (Rajeev *et al.*, 2009; Karras *et al.*, 2011). The last term in Eq. (1) accounts for relaxation of electrons out of the conduction band. This term can represent either uni- or bi-molecular recombination with holes or existing trap states.

Using the concept of dielectric breakdown when a critical electron density n_{crit} is reached for LID, this rate equation predicts the observed subpicosecond pulse duration and bandgap scaling (Mero *et al.*, 2005) for subps pulses. Refinements to the rate equations are (1) multiple rate equations/multiple levels in the conduction band (Rethfeld, 2004); (2) effect of plasma on the local field in a film (Gallais *et al.*, 2010); and (3) estimation of energy deposition (Gallais *et al.*, 2015). Measurements and solutions to Eq. (1) for dielectric films and surfaces suggest that in the subps regime the LIDT scales approximately linearly with the material bandgap and as a power law with respect to pulse duration, τ^κ with $\kappa \approx 0.3 \dots 0.4$, for 1-on-1 damage (Mero *et al.*, 2005).

Multiple pulse, S-on-1, LIDTs need to take into account accumulating laser-induced material changes. This material *incubation* is apparent in the dependence of LIDT on the number of pulses in the test that illuminate each sample site. As the number of pulses increases, the damage threshold decreases. Possible underlying mechanisms are that (1) during the pulse train laser-induced defects and other material modifications increase the absorption and/or (2) the critical energy for damage decreases due, for example, to defect caused weakening of bonds. The latter can be expressed as a decrease in a critical enthalpy G , for example, evaporation enthalpy that leads to damage. In a simple model, the threshold fluence for pulse number S is determined by the absorption (coefficient) of the sample and the critical enthalpy after the previous pulse, $\alpha(S-1)$ and $G(S-1)$, respectively (Sun *et al.*, 2015):

$$F_{th}(S) \approx \frac{G(S-1)}{\alpha(S-1)} \quad (2)$$

Fig. 4 shows a typical incubation curve $F_{th}(S)$.

To explain incubation on a microscopic level we refer to the energy level diagram in Fig. 5. At fluences just below the LIDT a subcritical electron density is generated in the conduction band. Between pulses most of these electrons relax to the valence band, but some remain trapped in midgap native and laser-induced states. The electrons can be re-excited by subsequent laser pulses and enhance the generation of the conduction band plasma. The process repeats with each pulse until the midgap states become

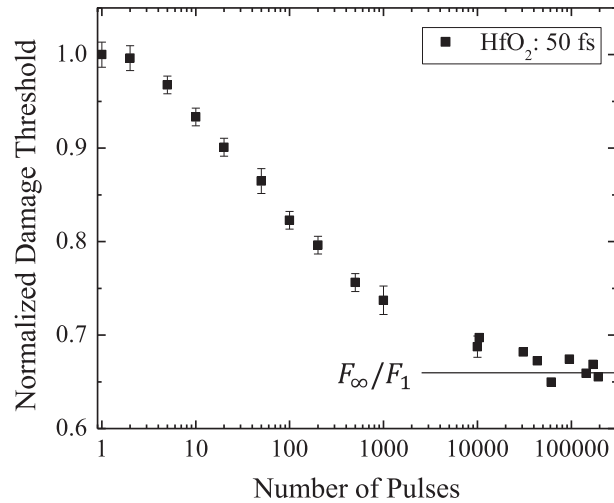


Fig. 4 Measured LIDT of a HfO_2 film as a function of the number of 50-fs pulses illuminating one and the same sample spot, normalized to the single-pulse LIDT F_1 . The first ($S-1$) pulses modify (incubate) the material leading to a reduction of F_{th} for the last pulse (number S) that triggers catastrophic damage. The optic can be used as long as the laser fluence $F < F_\infty$, i.e., the multiple pulse LIDT. Reproduced from Nguyen, D.N., Emmert, L., Mero, M., *et al.*, 2008. The effect of annealing on the subpicosecond breakdown behavior of hafnia films. In: Proc. of SPIE, vol. 7132, p. 71320N.

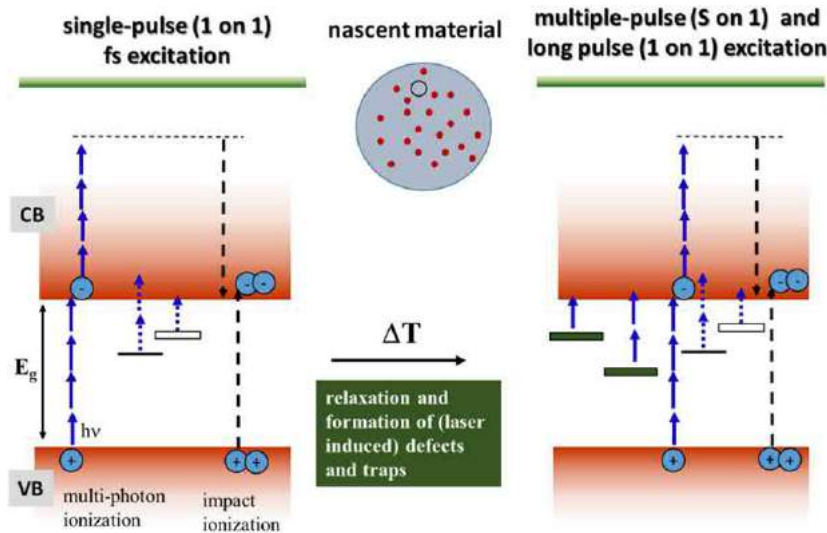


Fig. 5 Energy diagram of optical excitation and relaxation of dielectric materials. It is assumed here that the laser probes intrinsic material properties. The bandgap E_g exceeds the single photon energy. Optical excitation of electrons from valence band (VB) to conduction band (CB) is therefore possible only through a multi-photon process. The left group of processes is important for 1-on-1 events and short pulses; the right group plays a role for long pulses and pulse trains. Relaxation out of the CB can occur on a subps time scale. There is a certain probability that defects like self-trapped excitons and color centers are created (laser induced defects) during a long pulse or train of pulses.

saturated or damage occurs. This phenomenon can be modeled by adding additional rate equations (Emmert *et al.*, 2010) to include the various electronic states and transitions (see Fig. 5). In the case of fused silica, the microscopic states responsible for incubation are well known. Electrons in the conduction band combine with valance band holes to form self-trapped excitons (STEs). In most cases these relax through a non-radiative mechanism, but sometimes the STE will relax into an E' center and a non-bridging oxygen. The generation of these defects has been exploited to create a variety of optical structures in bulk glass (Beresna *et al.*, 2014). In many oxide films, the trap states have not yet been identified, but oxygen vacancies are predicted (Cavartin *et al.*, 2006).

Defect-Induced Damage

At pulse durations on the order of picoseconds and longer, a second class of extrinsic defects often controls the optical performance limits even in high-quality materials (Papernov, 2015). Defects in dielectric oxides can be, for example, clusters of partially

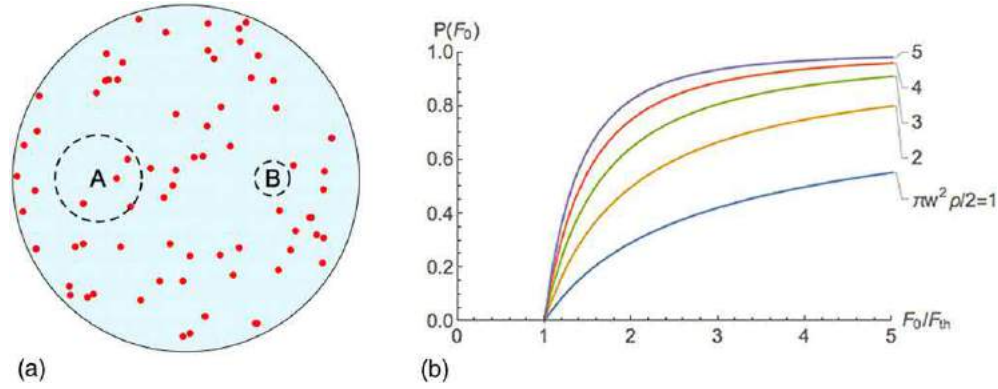


Fig. 6 (a) Schematic of an optical sample with randomly distributed extrinsic defects (dots) and incident laser beams of spot size A and B (dashed circles). The large spot (A) is likely to contain one or more defects and measure LIDT controlled by defects. A smaller beam (B) will often probe the intrinsic material that has a higher damage threshold. (b) $P(F_0)$ as a function of F_0/F_{th} for different values of the product $\pi w^2 \rho / 2$. The defect density and/or beam radius increases from bottom to top.

oxidized material with metal-like properties, impurities, or residues from surface polishing, cf. defects of type 1 in Fig. 2. Metal-like defects, for example, lack a bandgap and thus single photon absorption of long pulse and CW lasers can be efficient. One damage mechanism here can be the creation of a localized, hot, high-pressure plasma that eventually explodes, ejecting material. The fact that damage is controlled by defects is the main cause for the wide range of reported LIDT values.

As pointed out previously, the presence of such defects leads to an excitation spot size dependent LIDT. A large beam spot is more likely to illuminate defects that determine the damage threshold. A sufficiently small spot is likely to probe the intrinsic, defect-free material. This behavior is illustrated in Fig. 6(a). In general, a material is described by a defect distribution $\bar{\rho}(F)$. The product $\bar{\rho}(F)dF$ gives the density (areal or volume) of defects whose damage threshold is in the range from F to $F + dF$. This distribution also depends on the wavelength of the excitation laser (Jensen *et al.*, 2009).

A consequence of defect-initiated damage is that LIDT is probabilistic rather than deterministic. For fluences below the intrinsic LIDT, the probability of damage is equivalent to the probability that a defect is located in a region of the beam where the fluence is sufficiently large to initiate damage. Given a defect distribution and a beam profile, the probability of damage can be predicted. An example that is often used is an ensemble of identical defects described by a damage threshold F_{th} and density ρ illuminated by a Gaussian beam of radius w ($1/e^2$ intensity) and peak fluence F_0 . In this case, the damage probability is of the form

$$P(F_0) = 1 - (F_{th}/F_0)^{-\pi w^2 \rho / 2} \quad (3)$$

Note that the product in the exponent ($\pi w^2 \rho$) is the average number of defects within a circle of radius w . The shape of the function $P(F_0)$ for different values of $\pi w^2 \rho$ is illustrated in Fig. 6(b).

A variety of nano- and microstructures can be responsible for defect-initiated damage. Subsurface cracks can result from the cutting and polishing of optical surfaces. Cracks have sharp edges that can lead to local field enhancements or act as local absorbers due to surface states. An important defect in multilayers, which are typical for optical mirrors and beam splitters, are nodules (Tench *et al.*, 1994). These are particles embedded in the multilayer stack during deposition. Large nodules can disrupt the multilayer leading to mechanically weak interfaces and local intensification of the electric field. Laser conditioning has proved particularly successful in mediating the performance limiting effect of nodules (Bercegol, 1998).

Finally, defect-initiated damage even occurs in material that appears free of structural defects. This observation is attributed to nanoabsorbers that may result from local substoichiometric or metal-like regions. These nanometer-scale absorbers trigger an expansion of the absorption volume at intensities above the LIDT which leads to micron-scale craters (Lange *et al.*, 1987; Papernov and Schmid, 2002; Demos *et al.*, 2013).

Measuring LIDT

The LIDT of materials and optical components is measured by destructive damage tests. A variety of tests have been developed to address both engineering and scientific goals. On the one hand LIDT measurements are used to determine the operational limits of commercial products and/or for qualifying an optic for use in a system. In this case a witness sample is often used. But also, as a scientific tool, damage testing is used for characterizing materials. Examples of scientific interest include pulse duration scaling discussed earlier, bandgap scaling, incubation, and the distribution and characterization of defects. LIDT tests reveal material imperfections that are often not detected by other (non-destructive) diagnostic measurements.

Traditional damage tests are based on the evaluation of whether damage occurred following illumination of the sample with a train of pulses. This evaluation can be done either immediately by observation of damage with a scatter probe or an in-situ microscope or later using post mortem microscopy. Alternatively, measuring the ablation crater diameter as a function of fluence

has also been applied successfully for LIDT tests. Extrapolation to the fluence where the crater diameter is zero gives the LIDT for 1-on-1 (Liu, 1982) and S-on-1 experiments (Sun *et al.*, 2015). The following are few common examples of LIDT measurements methods.

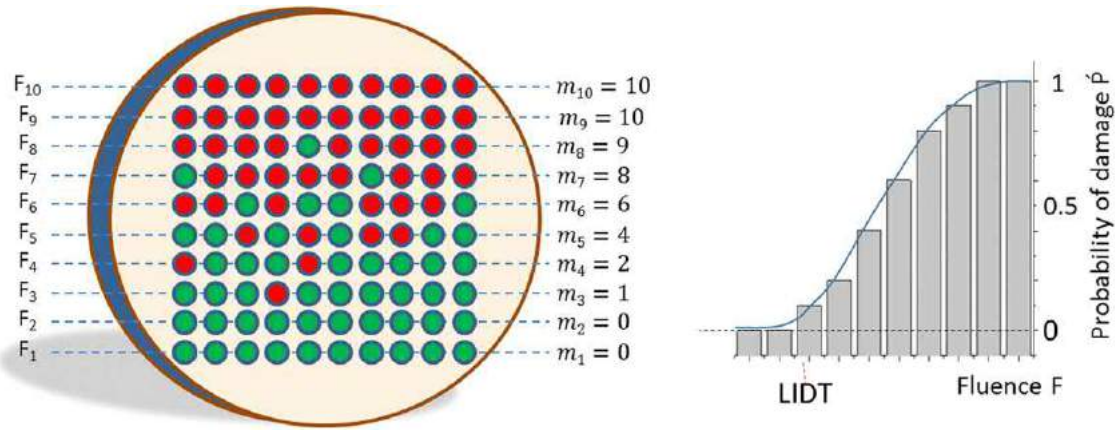


Fig. 7 Illustration of the ISO standard for LIDT tests. The damage probability versus fluence is measured. A tested optic (left) is divided into a number of tests sites and tested at various fluences F_i . The result of each test is either damage (red) or no damage (green). For each fluence F_i , the number of damage events m_i is counted. The probability is estimated by dividing the number of damage events by the number of tests at the given fluence. The probability vs. fluence (right) is plotted and LIDT is determined by extrapolating the curve to zero probability.

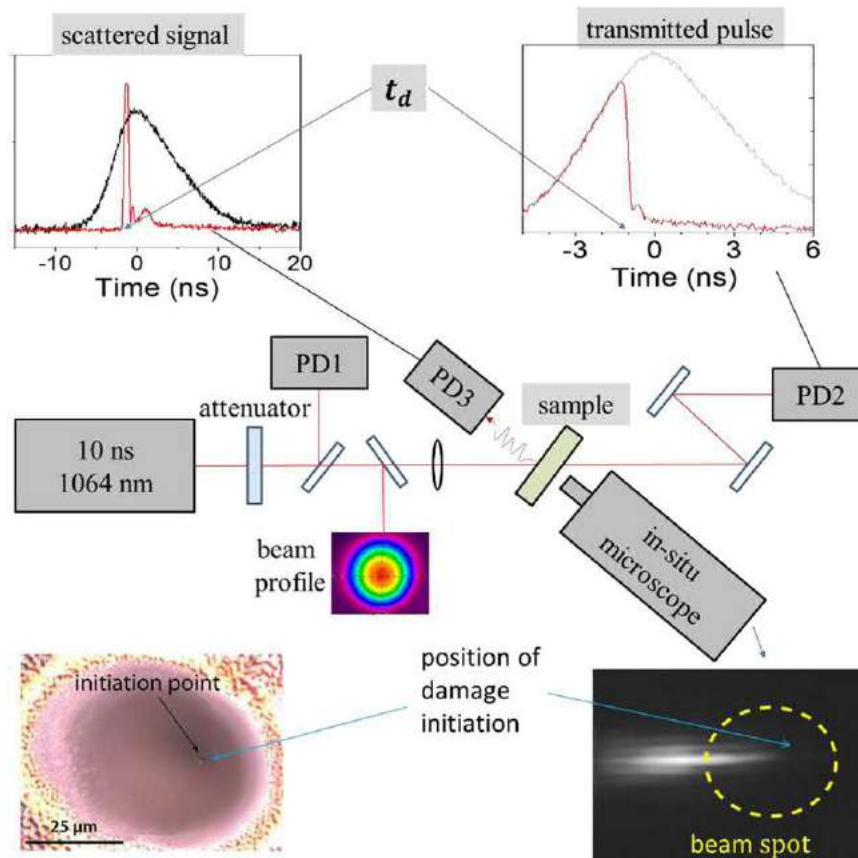


Fig. 8 Experimental setup for a nanosecond (or longer) pulse STEREO-LID test. The onset of damage and the location of damage initiation within the excitation spot are monitored for each event. Defect controlled damage leads to the ejection of a material jet as seen in the inset. Reproduced from Xu, Y., Emmert, L.A., Rudolph, W., 2015. Spatio-Temporally REsolved Optical Laser Induced Damage (STEREO LID) technique for material characterization. Opt. Express 23, 21607–21614.

An early and intuitive methodology is the R-on-1 test. In this test, each site is tested repeatedly with increasing fluence until damage occurs. While still used occasionally, the R-on-1 method does not account for laser conditioning and incubation discussed previously and therefore the test depends on the details of how the fluence is increased.

The ISO LIDT standard (ISO 11254-2, 2011) is based on a measurement of damage probability (Porteus and Seitel, 1984). The method is illustrated in Fig. 7. The damage probability is plotted as a function of the incident fluence and the damage threshold F_{th} is determined by extrapolating the curve to zero probability. Strictly speaking, this extrapolation requires knowledge of the defect density distribution function. Since this distribution is rarely known, it is done typically with a linear fit. It is important that a site is tested just once even if no damage occurred, because the material might have changed. The shape of the damage probability curve provides qualitative information about the optic. For small beam sizes, a narrow fluence range between the 0% and 100% damage probability is an indication of intrinsic damage.

The ISO protocol provides a balance between practical aspects of component testing and obtaining material data for scientific studies. It is poor at quantifying sparse fluence limiting defects. This issue, however, is especially important for large area optics. For such applications, raster scan methods using large spots have been developed (Borden *et al.*, 2005). At the National Ignition Facility a 1-mm test beam is scanned over a 1-cm² area. The tests are overlapping in order to illuminate the entire square. The scans are repeated with increasing fluence, like the R-on-1 method, but this is considered to mimic the laser conditioning that would occur for the optic under realistic operation conditions. Finally, to reduce the effect of outliers, the LIDT is defined as the fluence at which 10 or more damage events are observed.

The previous methods all share the common feature that they test for damage or no damage and therefore set an upper bound on the true damage threshold. A method that measures the damage fluence/intensity during a pulse is Spatio-Temporally REsolved Optical Laser-Induced Damage (STEREO-LID (Xu *et al.*, 2015a)) so named because it measures both the time during the pulse and the location within the beam spot where damage is initiated. Using the spatial/temporal intensity profile of the pulse, the critical fluence/intensity for each test event can be determined (Fig. 8).

Using STEREO-LID an optic is tested at multiple sites, like the ISO test, but all at an incident fluence sufficient to ensure a damage event. Each test event (one pulse at one sample site) determines where within the beam profile the damage initiating defect was located and when during the pulse “damage” occurred.

From these data one can derive the following: (1) the LIDT of performance-limiting defects: These are simply the lowest LIDT observed in the data set; (2) the defect distribution function $\bar{p}(F)$: this function can be retrieved without *a priori* assumptions by relating it to a histogram of the damage fluences of the data set (Xu *et al.*, 2015b); (3) the damage probability $P(F)$: This is done by a straightforward simulation of the ISO test using $\bar{p}(F)$. It should be noted that the same number of test sites provide not only new information about the sample compared to the ISO test, but also produce the ISO data with much greater accuracy.

See also: Ultrafast and Intense-Field Nonlinear Optics

References

- Bercegol, H., 1998. What is laser conditioning? A review focused on dielectric multilayers. In: Proc. of SPIE, vol. 3578, pp. 421–426.
- Beresna, M., Gecevicius, M., Kazansky, P.G., 2014. Ultrafast laser direct writing and nanostructuring in transparent materials. Adv. Opt. Photon 6. doi:10.1364/AOP.6.000293.
- Bloembergen, N., 1974. Laser-induced electric breakdown in solids. IEEE J. Quant. Electron. 10, 375–386.
- Borden, M.R., Foltz, J.A., Stolz, C.J., *et al.*, 2005. Improved method for laser damage testing coated optics. In: Proc. of SPIE, vol. 5991, p. 59912A.
- Demos, S.G., Negres, R.A., Raman, R.N., Rubenchik, A.M., Feit, M.D., 2013. Material response during nanosecond laser induced breakdown inside of the exit surface. Laser Photonics Rev. 7, 444–452.
- DeShazer, L.G., Newnam, B.E., Leung, K.M., 1973. Role of coating defects in laser-induced damage to dielectric thin films. Appl. Phys. Lett. 23, 607–609.
- Emmert, L.A., Mero, M., Rudolph, W., 2010. Modeling the effect of native and laser-induced states on the dielectric breakdown of wide band gap optical materials by multiple subpicosecond laser pulses. J. Appl. Phys. 108, 043523.
- Emmert, L.A., Rudolph, W., 2015. Femtosecond laser-induced damage in dielectric materials. In: Ristau, D. (Ed.), Laser-Induced Damage in Optical Materials. Boca Raton, FL: CRC Press/Taylor & Francis Group, pp. 127–151.
- Gallais, L., Douti, D.-B., Commandré, M., *et al.*, 2015. Wavelength dependence of femtosecond laser-induced damage threshold of optical materials. J. Appl. Phys. 117, 223103.
- Gallais, L., Mangote, B., Commandré, M., *et al.*, 2010. Transient interference implications on the subpicosecond laser damage of multidiellectrics. Appl. Phys. Lett. 97, 051112.
- Gavartin, J.L., Ramo, D.M., Shluger, A.L., Bersuker, G., Lee, B.H., 2006. Negative oxygen vacancies in HfO₂ as charge traps in high-k stacks. Appl. Phys. Lett. 89, 082908.
- ISO 11254-2:2011. Laser and laser related equipment – Test methods for laser-induced damage threshold – Part 2: Threshold determination. International Standard, 2011. International Organization for Standardization.
- Jensen, L., Schrammeyer, S., Jupe, M., Blaschke, H., Ristau, D., 2009. Spot-size dependence of the LIDT from the NIR to the UV. In: Proc. of SPIE, vol. 7504, p. 75041E.
- Jones, S., Braunlich, P., Casper, R., Shen, X., Kelly, P., 1989. Recent progress on laser-induced modifications and intrinsic bulk damage of wide-gap optical materials. Opt. Eng. 28, 1039–1068.
- Karras, C., Sun, Z., Nguyen, D.N., Emmert, L.A., Rudolph, W., 2011. The impact ionization coefficient in dielectric materials revisited. In: Proc. of SPIE, vol. 8190, p. 819028.
- Keldysh, L., 1965. Ionization in the field of a strong electromagnetic wave. Sov. Phys. JETP 20, 1307–1314.
- Lange, M.R., McIver, J.K., Guenther, A.H., 1987. Anomalous absorption in optical coatings. Nat. Bur. Stand. (U.S.) Spec. Publ. 746, 515–528.
- Lenzner, M., Krüger, J., Sartania, S., *et al.*, 1998. Femtosecond optical breakdown in dielectrics. Phys. Rev. Lett. 80, 4076–4079.
- Liu, J.M., 1982. Simple technique for measurements of pulsed Gaussian-beam spot sizes. Opt. Lett. 7, 196–198.
- Mero, M., Liu, J., Rudolph, W., Ristau, D., Starke, K., 2005. Scaling laws of femtosecond laser pulse induced breakdown in oxide films. Phys. Rev. B 71, 115109.

- Mouskeftaras, A., Guizard, S., Fedorov, N., Klimentov, S., 2013. Mechanisms of femtosecond laser ablation of dielectrics revealed by double pump–probe experiment. *Appl. Phys. A* 110, 709–715.
- Papernov, S., 2015. Defect-induced damage. In: Ristau, D. (Ed.), *Laser-Induced Damage in Optical Materials*. Boca Raton, FL: CRC Press/Taylor & Francis Group, pp. 25–73.
- Papernov, S., Schmid, A.W., 2002. Correlations between embedded single gold nanoparticles in SiO₂ thin film and nanoscale crater formation induced by pulsed-laser radiation. *J. Appl. Phys.* 92, 5720–5728.
- Porteus, J.O., Seitel, S.C., 1984. Absolute onset of optical surface damage using distributed defect ensembles. *Appl. Opt.* 23, 3796–3805.
- Rajeev, P.P., Gertsvolf, M., Corkum, P.B., Rayner, D.M., 2009. Field dependent avalanche ionization rates in dielectrics. *Phys. Rev. Lett.* 102, 083001.
- Rethfeld, B., 2004. Unified model for the free-electron avalanche in laser-irradiated dielectrics. *Phys. Rev. Lett.* 92, 187401.
- Ristau, D. (Ed.), 2015. *Laser-Induced Damage in Optical Materials*. Boca Raton, FL: CRC Press/Taylor & Francis Group.
- Smith, A., Do, B., 2008. Bulk and surface laser damage of silica by picosecond and nanosecond pulses at 1064 nm. *Appl. Opt.* 47, 4812–4832.
- Stolz, C.J., Ristau, D., Turowski, M., Blaschke, H., 2009. Thin film femtosecond laser damage competition. In: *Proc. of SPIE*, vol. 7504, p. 75040S.
- Stolz, C.J., Runkel, J., 2012. Brewster angle polarizing beam splitter laser damage competition: p polarization. In: *Proc. of SPIE*, vol. 8530, p. 85300M.
- Stolz, C.J., Thomas, M.D., Griffin, A.J., 2008. BDS thin film damage competition. In: *Proc. of SPIE*, vol. 7132, p. 71320C.
- Stuart, B., Feit, M.D., Herman, S., *et al.*, 1996. Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Phys. Rev. B* 53, 1749–1761.
- Sun, Z., Lenzner, M., Rudolph, W., 2015. Generic incubation law for laser damage and ablation thresholds. *J. Appl. Phys.* 117, 073102.
- Tench, R.J., Chow, R., Kozlowski, M.R., 1994. Characterization of defect geometries in multilayer optical coatings. *J. Vac. Sci. Technol. A* 12, 2808–2813.
- Wood, R.M., 2003. *Laser Damage in Optical Materials*. Bristol: IOP Publishing/Taylor & Francis Group.
- Xu, Y., Emmert, L.A., Rudolph, W., 2015a. Spatio-TEmporally REsolved Optical Laser Induced Damage (STEREO LID) technique for material characterization. *Opt. Express* 23, 21607–21614.
- Xu, Y., Emmert, L.A., Rudolph, W., 2015b. Determination of defect densities from spatiotemporally resolved optical-laser induced damage measurements. *Appl. Opt.* 54, 6813–6817.

Four-Wave Mixing

L Canioni and L Sarger, University of Bordeaux, Talence, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Light waves propagate along straight lines in a vacuum and in materials. However, their amplitude and phase can only be affected by the intrinsic properties of the sample, such as absorption and refraction index. Experimental work from the early 1960s, has deviated from this simple approach, mainly through propagating intense light beams in crystals and glasses. Among many puzzling new effects, frequency conversion and beam coupling have triggered intense studies and opened a whole new field, theoretically and experimentally. These various phenomena have been classified according to the order of their optical non-linearity. The second order mainly occurs in frequency generation and has already been discussed in previous articles of this encyclopedia. Therefore, this article will be devoted to the higher order, in which self focusing, coupling, and mixing of light beams will be explained. We will illustrate some simple ideas with a convenient model before describing a few applications.

The framework of electromagnetic wave interaction with matter is embedded in the Lorentz force. The oscillating electric field drives the components, ions, and electrons, but due to the large mass differences, only the electron movement is strongly perturbed. In dielectric media, these carriers are connected with static bonds such that the applied electric field modulates the spatial position of equilibrium of the charges. As a result, a dipolar moment is created and oscillates quasi-adiabatically at very high frequencies (between 10^{14} and 10^{17} Hz) for an applied optical wave. These collective oscillations of the induced dipoles compose the so-called matter polarization, proportional to the driving electric field amplitude. Magnetic field interaction is usually much weaker and so can be neglected in nonmagnetic media.

As an electromagnetic wave, lasers are a coherent source of very high intensity, i.e., large electric fields. For example, in standard pulsed lasers focused on a target, the light intensity can easily reach 10 Tw/cm^2 , corresponding to an electric field of the order of magnitude of the internal bonding field (10^{10} V/m). At this level of interaction, dipolar moments are no longer strictly proportional to the applied electric field, as depicted in Fig. 1. In the weak field regime, the simple perturbation approach displaces the carriers in a power series of the applied electric field. The polarization is thus developed in series, in respect to this electric field.

Formalism

Relation Between Field and Polarization: The Response Function

Although numerous works have been published with a frequency description of the material polarization, we will discuss here the study of nonlinearity in the time domain in order to follow the physical reality. In most processes, both spaces are equal and can be selected at will, as long as the entire requirements are fully understood; the electromagnetic field is classically described.

In the time domain, the polarization can be developed as

$$P(r, t) = P_{\text{Linear}}(r, t) + P_{\text{Nonlinear}}(r, t) \quad (1)$$

where the linear part of the polarization can be written as

$$P_{\text{Linear}}(r, t) = \epsilon_0 \int_{R^3} \int_{-\infty}^{\infty} R^{(1)}(r_1, t_1) E(r - r_1, t - t_1) dt_1 dr_1 \quad (2)$$

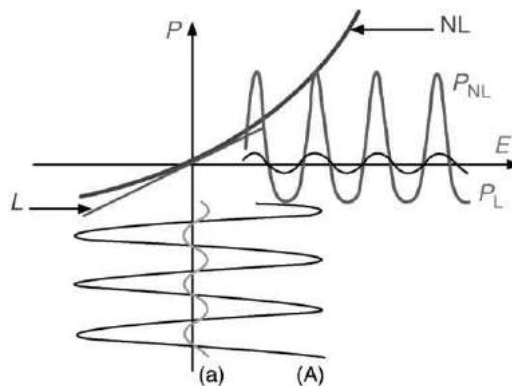


Fig. 1 Polarization for various field amplitude: (a) Weak field case, linear response. (A) Strong field, distortion and nonlinearity.

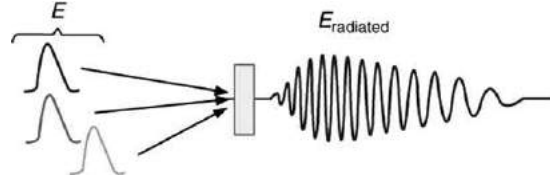


Fig. 2 Three incoming beams E_1 , E_2 , E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 .

This general expression is rather complex but shows that the polarization, due to collective movement of carriers in a macroscopic volume (thousands of dipoles) centered at position r , always depends on the applied electric field. This expression is also nonlocal as dipoles–dipoles interactions play an important role.

In general, for the nonlinear (NL) part of the polarization, an approximation of the local interaction is described for simplicity. For a second-order interaction, one can write:

$$P_{\text{NL}}^{(2)}(r, t) = \epsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(2)}(t_1, t_2) E(r, t - t_2) E(r, t - t_2 - t_1) dt_1 dt_2 \quad (3)$$

For the other series expansion – at third order, one obtains

$$P_{\text{NL}}^{(3)}(r, t) = \epsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t_1, t_2, t_3) E(r, t - t_3) E(r, t - t_3 - t_2) E(r, t - t_3 - t_2 - t_1) dt_1 dt_2 dt_3 \quad (4)$$

This third-order polarization already combines three electric fields. This nonlinear polarization also radiates a fourth wave and is therefore the adopted framework for describing four-wave mixing experiments (Fig. 2).

As the physical properties of materials depend only on time intervals between the different pulse interactions, Eq. (4) can be rearranged as

$$P_{\text{NL}}^{(3)}(r, t) = \epsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t - \tau_1, t - \tau_2, t - \tau_3) E(r, \tau_1) E(r, \tau_2) E(r, \tau_3) d\tau_1 d\tau_2 d\tau_3 \quad (5)$$

If the sample response time is much shorter than pulse duration, the so-called adiabatic limit (as for electronic cloud deformation excited distant from resonance) the polarization is written as

$$P_{\text{NL}}^{(3)}(r, t) = \epsilon_0 \sigma^{(3)} E(r, t) E(r, t) E(r, t) \quad (6)$$

We will see later how this expression is useful for third-harmonic generation and optical Kerr effects, etc. This expression is accurate for materials excited at a distance from resonances and care must be taken to reserve such a model for corresponding processes. In this case, the frequency approach is quite simple too and both spaces, time or Fourier, can be chosen at will.

For very long response time, such as orientational or structural molecular reorganization, one can use the Born–Oppenheimer approximation and present the nonlinear polarization as

$$P_{\text{NL}}^{(3)}(r, t) = \epsilon_0 E(r, t) R'(t) \quad (7)$$

Similar is linear polarization, where the response function takes into account the changes of the molecular structure under the light intensity stress:

$$R'(t) = \int_{-\infty}^{+\infty} d^{(3)}(t - t_1) E(r, t_1) E(r, t_1) dt_1 \quad (8)$$

The material response will then be mostly driven by the field amplitude rather than the frequency.

Properties of Nonlinear Susceptibilities

The third-order nonlinear susceptibility $\chi^{(3)}$ is the Fourier transform of the response function R and has proved to be a most practical tool in optical nonlinear theory. The symmetry properties of the material lead to simplification on the tensor of rank 4 (fourth rank tensor), involved in four-wave mixing:

$$P_i^{(3)}(\omega_m) = \epsilon_0 \sum_{n_1, \dots, n_p} \sum_{i_1, \dots, i_p} \chi_{i i_1, \dots, i_p}^{(3)}(\omega_m; \omega_{n_1}, \dots, \omega_{n_p}) E_{i_1}(\omega_{n_1}) \dots E_{i_p}(\omega_{n_p}) \quad (9)$$

The nonlinear susceptibility tensor must account for all the spatial and symmetrical order of the sample. This will eventually reduce the number of independent elements from a possible 81.

Wave Equation in Nonlinear Regime

From Maxwell's equations, with a sample without free carriers and current, one can deduce the wave equation in the time domain, as:

$$\Delta E - \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} - \frac{1}{c^2} \frac{\partial^2 P_L}{\partial t^2} = \frac{1}{c^2} \frac{\partial^2 P_{NL}}{\partial t^2} \quad (10)$$

In the frequency domain, this equation reduces to

$$\Delta \tilde{E}(r, \omega) + k(\omega)^2 \tilde{E}(r, \omega) = -\mu_0 \omega^2 \tilde{P}_{NL}(r, \omega) \quad (11)$$

In order to solve these (coupled) wave equations in relation to four-wave mixing, we must develop the electric field and the polarization in a slowly varying envelope approximation, using the following conventions:

$$E_i(r, t) = \frac{1}{2} \left(A_i(r, t) e^{i(\omega_i t - k(\omega_i) z)} + \bar{A}_i(r, t) e^{-i(\omega_i t - k(\omega_i) z)} \right) \quad (12)$$

$$P_{NL}(r, t) = \frac{1}{2} \left(p_{nl}(r, t) e^{i(\omega t - k(\omega) z)} + \bar{p}_{nl}(r, t) e^{-i(\omega t - k(\omega) z)} \right) \quad (13)$$

Within the framework of small perturbations, diffraction can be neglected and the wave equation can be easily reduced to one dimension:

$$\frac{\partial A}{\partial z} = \frac{i\omega}{2n(\omega)\epsilon_0 c} p^{NL}(z, \omega) \exp(-ik(\omega)z) \quad (14)$$

where ω represents the sum or difference permutations over all the frequencies ω_i , thus reflecting the law of energy conservation.

In the case of four-wave mixing, one has to solve four coupled equations for the four interacting waves. Using the polarization [Eq. \(13\)](#), one gets:

$$\frac{\partial A_j}{\partial z} = \frac{i\omega_j}{2n(\omega_j)c} \chi^3 A_1^{(*)} A_2^{(*)} A_3^{(*)} \exp\left(\sum_{i=1,2,3} \exp(\pm ik_i(\omega_i)z)\right) \exp(-ik(\omega_j)z) \quad (15)$$

This set of generic equations respectively represent indifferently, either the three incoming beams ($j = 1, 2, 3$), or the radiated field E_4 at ω . The algebra used here associates at each complex conjugate field (*) with the counter propagating wave ($-k$). The amplitude variation of the j field at frequency ω_j is due to the coupling of the three other fields within the sample. The accumulated field is nevertheless always strongly dependent of the phase term and will vanish except when this phase term is zero. This is known as the phase matching condition which has to be fulfilled to ensure an efficient energy conversion. A more general approach of this condition can be found in the corresponding article of this encyclopedia.

Example and Selected Applications

When inserting the full material polarization (linear and nonlinear) into the propagation equation, even restricted at third order, a whole set of processes must be considered. All of them are described by the generic equation already presented, but only a few very important processes – with increasing complexity – will be described here.

Optical Kerr Effect

We consider one single beam incident at frequency ω on a sample. The non-linear polarization is given by

$$P_{NL}^{(3)}(r, t) = \frac{\epsilon_0}{8} \chi^{(3)} (E(r, t)^3 + E^*(r, t)^3 + 3E(r, t)^2 E^*(r, t) + 3E(r, t) E^*(r, t)^2) \quad (16)$$

The polarization at frequency ω reduces to

$$P_{NL}^{(3)}(r, t) = \frac{3\epsilon_0}{4} \chi^{(3)} |E(r, t)|^2 E(r, t) \quad (17)$$

This term has the same phase as the incident beam and, therefore, the phase matching is automatic. Using [Eqs. \(12\)](#) and [\(13\)](#), one can rewrite the wave equation for the optical Kerr effect in the thin sample approximation:

$$\frac{\partial A}{\partial z} = \frac{3i\omega}{8nc} \chi^{(3)} |A|^2 A \quad (18)$$

The solution of this equation is

$$A(z) = A(0) \exp\left(\frac{3i}{8nc} \chi^{(3)} |A(0)|^2 z\right) \quad (19)$$

Along the beam, the amplitude of the input pulse stays constant for this process but its phase is affected by a nonlinear shift proportional to power density I .

In literature, one can find the concept of n_2 , the sample nonlinear index, in the following formula:

$$\psi_{NL}(z) = \frac{\omega}{c} n_2 I z \quad (20)$$

In this way of describing the total phase along the beam, the index of refraction in the nonlinear regime can be expressed more conveniently as

$$n = n_0 + n_2 I \quad (21)$$

One of the most significant demonstrations of this effect is the self-focusing phenomenon. The usual Gaussian transverse structure of a laser beam presents high intensity at the center of the beam, therefore creating an index variation in the isotropic sample due to this optical Kerr effect. An induced lens can then be applied to this refractive index gradient, following the field distribution. As a result, the beam focuses itself along the propagation and can result in a complete breakdown.

Third Harmonic Generation

In the third harmonic generation, i.e., the third-order nonlinear process where the frequency of the generated wave is three times the frequency of the input wave at ω , only two waves are in the sample (at ω and 3ω). For the nonlinear wave, the master equation is then:

$$\frac{\partial A_{3\omega}}{\partial z} = \frac{3i\omega}{8n(3\omega)c} \chi^{(3)} A_{\omega}^3 \exp((3ik(\omega) - ik(3\omega))z) \quad (22)$$

Here, propagation effects have to be carefully considered as these waves do travel at different velocities and will eventually be out of phase. Obviously this discrepancy of phase velocities precludes the co-propagation and strongly affects the global efficiency. The only way to ensure a coherent generation of the harmonic wave throughout the material is to fulfill the relation:

$$k(3\omega) = 3k(\omega) \quad (23)$$

This is the so-called phase matching condition. Although difficult because the usual dispersion of materials is very large for such a frequency difference, this can be satisfied using either geometrical arrangement or, if possible, using the birefringent properties of the material. The efficiency of this method is poor and a better approach is implemented when two $\chi^{(2)}$ processes are cascaded (doubling and sum frequency mixing).

Degenerate Four-Wave Mixing and Phase Conjugation

One of the most popular and intriguing interactions is depicted in Fig. 3. Where the sample is excited by three fields at the same frequency ω in this particular geometry: two-waves, named pump beams E_1 and E_2 , are exactly counter-propagating with opposite wave vectors of $+k_1$ and $k_2 = -k_1$, while the third wave (E_3 arrives at the sample with a given wave vector of k_3 .

Of possible couplings, one contribution E_4 corresponds to a re-emitted beam along $k_4 = -k_3$, i.e., opposite to the E_3 wave.

The third-order nonlinear polarization is given here by

$$P_{NL}^{(3)}(r, t) = \frac{\epsilon_0}{8} \chi^{(3)} (E_1(r, t) + E_1^*(r, t) + E_2(r, t) + E_2^*(r, t) + E_3(r, t) + E_3^*(r, t))^3 \quad (24)$$

This E_4 wave (along $-k_3$) originates in this product from source terms including:

$$E_1(r, t)E_2(r, t)E_3^*(r, t) \quad (25)$$

and thus present an evolution driven by:

$$\frac{\partial A_4}{\partial z} = \frac{3i\omega}{4n(\omega)c} \chi^{(3)} A_1 A_2 A_3^* \quad (26)$$

where the phase matching condition is automatically fulfilled. Interesting facts can be described using this expression. First, the radiated amplitude is proportional to the complex conjugate of the E_3 field. This nonlinear process is not only able to reverse the direction of propagation but also the whole phase of an arbitrary incoming beam of light. This 'phase conjugator' can be

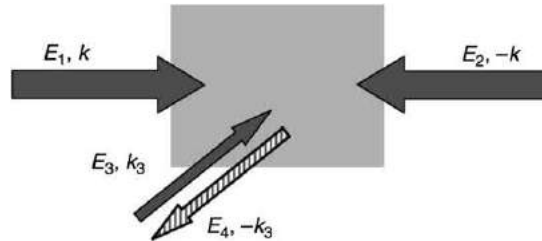


Fig. 3 Two counter-propagating fields (E_1 , E_2) and the third field E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 opposite to the E_3 wave.

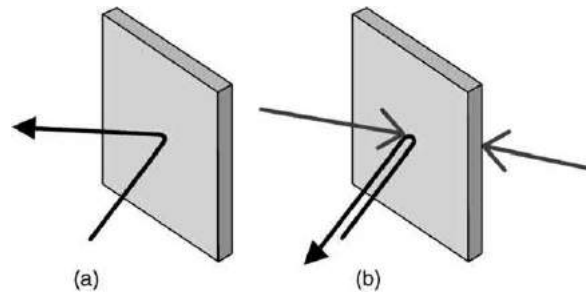


Fig. 4 (a) Snell refraction in a conventional mirror; (b) Phase conjugate mirror.

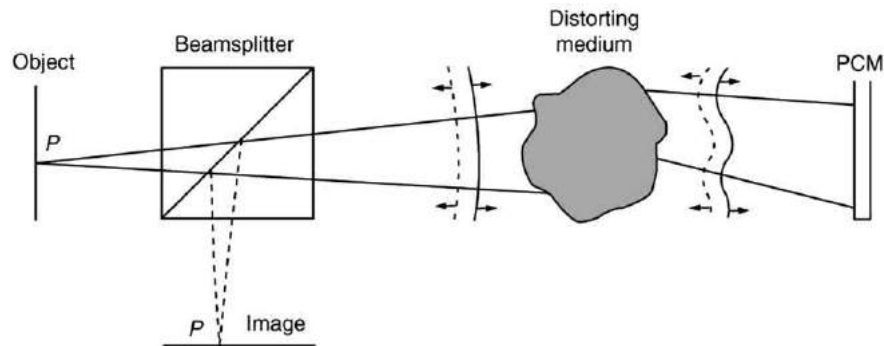


Fig. 5 A basic two-pass system for imaging and distortion compensation using optical phase conjugation.

considered as a kind of mirror with very unusual reflection properties. Unlike a conventional mirror, where a ray is redirected according to the ordinary law of reflection, a phase-conjugate mirror (PCM) retro-reflects all incoming rays back to their origin (Fig. 4).

Second, the reflectivity of this mirror depends on the susceptibility and the amplitude of the first two fields, usually called pump beams. This will allow reflectivity that is much higher than the unity, unlike with conventional mirrors.

The remarkable reflection properties of the PCM have found many important applications. The most useful undoubtedly is related to distortion correction. If the image information has been distorted by the transmitting medium on its way to the PCM, then these aberrations will be corrected when the reflected signal retraces its original path through the medium. Through this amazing 'healing' property of optical phase conjugation, a high-quality optical beam can be double passed through a distorting medium, such as an imperfect imaging device or a turbulent atmosphere, without any loss of beam quality. Fig. 5 shows the basic two-pass geometry for imaging and aberration correction using a PCM.

The spherical wave from the object point P is aberrated by the distorting medium. The wave-front is reversed by the PCM and a second passage through the same distorting path restores the initial field. The resulting spherical wave converges to the image point P' , separated from P by the beamsplitter.

Numerous other applications for PCM and the various underlying nonlinear optical interactions have been proposed. These include effects as diverse as parametric oscillation, optical tracking and pointing, spatial and temporal image processing, optical computing, optical filtering, etc. The PCM is also successfully implemented for aberration correction in high-power laser resonators.

Nondegenerate Four-Wave Mixing

If the frequencies of the impinging fields build a nonlinear polarization with an oscillating term close to absorption (or emission) frequencies of the sample, the nonlinear susceptibility increases tremendously. The radiated field amplitude changes by orders of magnitude and allows spectroscopic studies of the sample. The most popular process used to perform gas spectroscopy is CARS, an acronym for coherent anti-Stokes Raman scattering (or spectroscopy). In this case, the three incoming beams have different frequencies and the phase matching condition is eventually fulfilled in the 3D space.

Usually, two beams at the same frequency ω_P (pump beams), are mixed with a third beam at a tunable frequency ω_S , according to the energy scheme in Fig. 6. As the radiated frequency ω_{CARS} is higher, the Raman spectroscopic notation is used and the subscripts 'S' hold for 'Stokes' and 'P' for 'anti-Stokes'.

Whenever $\omega_P - \omega_S = \omega_{\text{molecule}}$, then an increase of the CARS intensity by many orders of magnitude is observed. Therefore, a Raman spectrum can be obtained by continuously changing, the Stokes laser and simultaneously recording the intensity of the

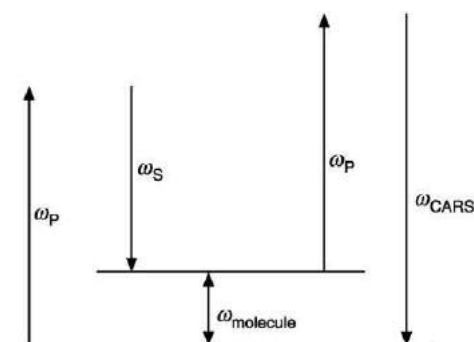


Fig. 6 Energy diagram for a CARS experiment. Four-wave mixing efficiency increases when the frequency difference $\omega_P - \omega_S$ is close to a resonant frequency of the sample, ω_{molecule} .

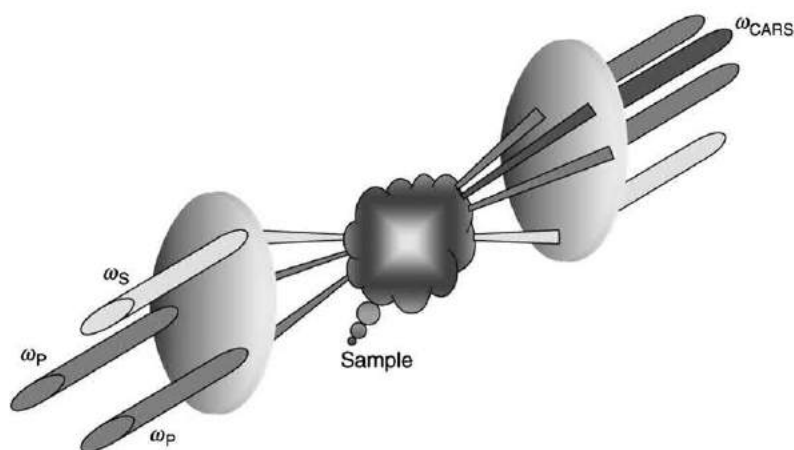


Fig. 7 Layout of the CARS beams near a sample (a flame here). This particular arrangement is called folded (or 3D) Boxcars.

CARS beam. This procedure is called scanning CARS, an interesting technique insofar as it allows taking high-resolution spectra, as well as measuring temperature of the sample while scanning over the Boltzmann distribution of the electronic ground state.

Fig. 7 displays a classical geometrical arrangement where the phase matching condition fully determines the directionality of the radiated field at ω_{CARS} . Although the alignment of the setup can become tedious, this technique has been widely used, even in hostile industrial environments. The coherent aspect of this four-wave mixing process, imbedded in the phase matching condition, has a huge benefit over classical spectroscopic techniques and thus isotropic techniques, such as laser-induced fluorescence or Raman spectroscopy.

The CARS process is an interference process comparable to diffraction of a grating. The two fields at ω_P and ω_S form a laser-induced moving grating and the third field, at ω_P , undergoes a Doppler shift at ω_{CARS} . This scattered field undergoes a coherent amplitude addition in the Bragg direction.

Conclusion

In this article, we have discussed a set of optical nonlinear processes occurring at the third order of the development in series of the material polarization. At this order, wave mixing obviously shows most significant behavior and has been extensively studied and applied in many scientific areas. The general framework presented here can be further investigated using some books listed in Further Reading below.

See also: Nonlinear Optical Phase Conjugation

Further Reading

Bloembergen, N., 1996. *Nonlinear Optics*. Singapore: World Scientific Pub Co.

- Bloom, D.M., Bjorklund, G.C., 1977. Conjugate wave-front generation and image reconstruction by four-wave mixing. *Applied Physics Letters* 31, 592–594.
- Born, M., Wolf, E., 1999. *Principles of Optics*. Cambridge, UK: Cambridge University Press.
- Butcher, P.N., Cotter, D., 1991. *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.
- Eaton, D.F., 1991. Nonlinear optical materials. *Science* 253, 281–287.
- Fisher, R.A., 1984. *Optical Phase Conjugation*. San Diego, CA: Academic.
- Flytzanis, C., 1975. Theory of nonlinear optical susceptibilities. In: Rabin, H., Tang, C.L. (Eds.), *Quantum Electronics*, vol. 1. New York: Academic Press, pp. 9–207.
- Hellwarth, R.W., 1977. Generation of time-reversed wave fronts by nonlinear refraction. *Journal of the Optical Society of America* 67, 1–3.
- Jackson, J.D., 1998. *Classical Electrodynamics*. New York: Wiley.
- Marburger, J.H., 1975. Self-focusing: Theory. *Progress in Quantum Electronics* 4, 35–110.
- Mitra, R., Habashy, T.M., 1984. Theory of wave-front-distortion correction by phase conjugation. *Journal of the Optical Society of America A* 1, 1103–1109.
- Rockwell, D.A., 1988. A review of phase-conjugate solid-state lasers. *IEEE Journal of Quantum Electronics* 24, 1124–1140.
- Shen, Y.R., 2002. *The Principles of Nonlinear Optics*. New York: Wiley.
- Yariv, A., 1977. Compensation for atmospheric degradation of optical beam transmission. *Optical Communication* 21, 49–50.
- Yariv, A., 1989. *Quantum Electronics*, 3rd edn. New York: Wiley.
- Yariv, A., Pepper, D.M., 1977. Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing. *Optical Letters* 1, 16–18.
- Zel'dovich, B.Ya., Pilipetsky, N.F., Shkunov, V.V., 1985. *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Kramers–Krönig Relations in Nonlinear Optics

M Sheik-Bahae, The University of New Mexico, Albuquerque, NM, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

α	linear absorption coefficient	n_2	nonlinear refractive index coefficient,
α_2	nonlinear absorption coefficient	$\Theta(t)$	coefficient of optical Kerr effect
β	two-photon absorption coefficient	$\wp$	step function
$\chi^{(n)}$	n th-order nonlinear optical susceptibility	2PA	principal value
n	linear refractive index	KK relations	two-photon absorption
			Kramers–Krönig relations

Since their introduction nearly 75 years ago, the Kramers–Krönig (KK) dispersion relations have been widely appreciated and applied in the analysis of linear optical systems. Because they are a consequence of strict causality, the KK relations apply not only to optical systems, but also to any linear, causal system such as electrical networks and particle scattering. In this article, we review the formulation and application of these relations in nonlinear optical systems. Simple logical arguments are used to derive dispersion relations that relate the nonlinear absorption coefficient to the nonlinear refraction coefficient. More general formalisms are then derived that apply to all nonlinear susceptibilities including the harmonic generating cases. Examples of recent successful application of these dispersion relations in analyzing various nonlinear materials will be presented.

The mathematical formalism of the KK dispersion relations in nonlinear optics was studied in the formative days of the field. The great usefulness of these relations was appreciated only recently, however, when they were used to derive the dispersion of the optical Kerr effect in solids from the corresponding nonlinear absorption coefficients, including two-photon absorption.

Before examining the details of KK relations in nonlinear optical systems, it is instructive to revisit the linear dispersion relations and their derivation based on the logic of causality. We will begin this task by introducing the definition of the linear as well as nonlinear susceptibilities $\chi^{(n)}$. In most nonlinear optics texts, the total material polarization (P) that drives the wave equation for the electric field (E) is expressed as

$$P_i(t) = \epsilon_0 \int_{-\infty}^{\infty} R_{ij}^{(1)}(t-t_1) E_j(t_1) dt_1 + \epsilon_0 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_{ijk}^{(2)}(t-t_1, t-t_2) E_j(t_1) E_k(t_2) dt_1 dt_2 + \dots$$

$$+ \epsilon_0 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_{ijkl}^{(3)}(t-t_1, t-t_2, t-t_3) E_j(t_1) E_k(t_2) E_l(t_3) dt_1 dt_2 dt_3 + \dots$$
(1)

where $R^{(n)}$ is defined as the n th-order, time-dependent response function or time-dependent susceptibility. The subscripts are polarization indices indicating, in general, the tensor nature of the interactions. The summation over the various indices $j, k, l, \dots$ is implied for the various tensor elements of $R^{(n)}$. Upon Fourier transformation, we obtain:

$$\vec{P}_i(\omega) = \epsilon_0 \int_{-\infty}^{\infty} d\omega_1 \chi_{ij}^{(1)}(\omega_1) \vec{E}_j(\omega_1) \delta(\omega - \omega_1) + \epsilon_0 \int_{-\infty}^{\infty} d\omega_1 \int_{-\infty}^{\infty} d\omega_2 \chi_{ijk}^{(2)}(\omega_1, \omega_2) \vec{E}_j(\omega_1) \vec{E}_k(\omega_2) \delta(\omega - \omega_1 - \omega_2)$$

$$+ \epsilon_0 \int_{-\infty}^{\infty} d\omega_1 \int_{-\infty}^{\infty} d\omega_2 \int_{-\infty}^{\infty} d\omega_3 \chi_{ijkl}^{(3)}(\omega_1, \omega_2, \omega_3) \vec{E}_j(\omega_1) \vec{E}_k(\omega_2) \vec{E}_l(\omega_3) \delta(\omega - \omega_1 - \omega_2 - \omega_3), \dots$$
(2)

where δ is the Dirac delta-function. Here the $E(\omega)$ are Fourier transforms of the corresponding electric field. The n th-order susceptibility is defined as the Fourier transform of the n th-order response function:

$$\chi_{ijk\dots n}^{(n)}(\omega_1, \omega_2, \dots, \omega_m) = \int_{-\infty}^{+\infty} d\tau_1 \int_{-\infty}^{+\infty} d\tau_2 \dots \int_{-\infty}^{+\infty} d\tau_n R_{ijk\dots m}^{(n)}(\tau_1, \tau_2, \dots, \tau_m) e^{i(\omega_1 \tau_1 + \omega_2 \tau_2 + \dots + \omega_m \tau_m)}$$
(3)

For simplicity, we drop the polarization indices $i, j, \dots$, and thus ignore the tensor properties of $\chi^{(n)}$ as well as the vector nature of the electric fields.

Let us for the moment concentrate on the linear polarization alone and derive the linear KK relations for the first-order susceptibility $\chi^{(1)}(\omega)$. For this, we rewrite Eq. (3) for $n=1$:

$$\chi^{(1)}(\omega) = \int_{-\infty}^{\infty} R^{(1)}(\tau) e^{-i\omega\tau} d\tau$$
(4)

(As defined above, $\chi^{(1)}(\omega)$ and $R^{(1)}(\tau)$ are not a strict Fourier transform pair because of a missing factor of 2π .) Causality means that the effect cannot precede the cause. This can be restated mathematically as:

$$R^{(1)}(t) = R^{(1)}(t) \Theta(t)$$
(5)

i.e., the response to an impulse at $t=0$ must be zero for $t<0$. Here $\Theta(t)$ is the Heaviside step function defined as $\Theta(t)=1$ for $t>0$ and $\Theta(t)=0$ for $t<0$. Upon Fourier transforming this equation, the product in the time domain becomes a convolution integral in

frequency space

$$\begin{aligned}
 \chi^{(1)}(\omega) &= \chi^{(1)}(\omega) \left[\frac{\delta(\omega)}{2} + \frac{i}{2\pi\omega} \right] \\
 &= \frac{\chi^{(1)}(\omega)}{2} + \frac{i}{2\pi} \wp \int_{-\infty}^{\infty} \frac{\chi^{(1)}(\omega')}{\omega - \omega'} d\omega' \\
 &= \frac{1}{i\pi} \wp \int_{-\infty}^{\infty} \frac{\chi^{(1)}(\omega')}{\omega' - \omega} d\omega'
 \end{aligned} \tag{6}$$

which is the KK relation for the linear optical susceptibility. The symbol $\wp$ stands for the Cauchy principal value of the integral. The KK relation is thus a restatement of the causality condition (5) in the frequency domain. Taking the real part we have,

$$\Re\{\chi^{(1)}(\omega)\} = \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\Im\{\chi^{(1)}(\omega')\}}{\omega' - \omega} d\omega' \tag{7}$$

Taking the imaginary part of Eq. (6) leads to a similar relation relating the imaginary part to an integral involving the real part. It is conventional to write the optical dispersion relations in terms of the more familiar quantities of refractive index, $n(\omega)$, and absorption coefficient, $\alpha(\omega)$. For $|\chi^{(1)}| \ll 1$ then $n - 1 = \Re\{\chi^{(1)}\}/2$ and $\alpha = \omega \Im\{\chi^{(1)}\}/c$, and Eq. (7) is transformed into

$$n(\omega) - 1 = \frac{c}{\pi} \wp \int_0^{\infty} \frac{\alpha(\omega')}{\omega' 2 - \omega^2} d\omega' \tag{8}$$

where we additionally used the reality conditions of $n(\omega) = n(-\omega)$, and $\alpha(\omega) = \alpha(-\omega)$ to change the lower integral limit to 0. More rigorous analysis shows that Eq. (8) is general and valid for any value of $|\chi^{(1)}|$. Although the KK dispersion relations and the extent of their applications in linear optics are well understood, some confusion sometimes exists about their applications to nonlinear optics. Causality clearly holds for both linear and nonlinear systems. The question is: what form do the resulting dispersion relations take in a nonlinear system? The linear Kramers–Krönig relations were derived from linear system theory, so it would appear to be impossible to apply the same logic to a nonlinear system. The key insight is that one can linearize the system. This is illustrated in Fig. 1 where a linear (and of course, causal) optical material is transformed into a ‘new’ linear system that now contains the material and an external perturbation denoted by ξ . Although we are interested in perturbations of an optical nature, this formalism is general under any type of perturbation. It is important to appreciate the fact that our new system is causal even in the presence of the perturbation. This allows us to write down a modified form of the Kramers–Krönig relation linking the index of refraction to the absorption:

$$[n(\omega) + \Delta n(\omega; \xi)] - 1 = \frac{c}{\pi} \wp \int_0^{\infty} \frac{\alpha(\omega') + \Delta \alpha(\omega'; \xi)}{\omega' 2 - \omega^2} d\omega' \tag{9}$$

which, after subtracting the linear relation between n and α leaves a relation between the changes in index and absorption:

$$\Delta n(\omega; \xi) = \frac{c}{\pi} \wp \int_0^{\infty} \frac{\Delta \alpha(\omega'; \xi)}{\omega' 2 - \omega^2} d\omega' \tag{10}$$

where ξ denotes a general perturbation. An equivalent relation also exists whereby the change in absorption coefficient can be calculated from the change in the refractive index. It is essential that the perturbation be independent of frequency of observation, ω' , in the integral (i.e., the excitation ξ must be held constant as ω' is varied).

Eq. (10) has been used to determine refractive changes due to ‘real’ excitations such as thermal and free-carrier nonlinearities in semiconductors. In those cases, ξ denotes either ΔT (change of temperature) or ΔN (change of free-carrier density), respectively. In the former case, one calculates the refractive index change resulting from a thermally excited electron–hole plasma and the temperature shift of the band edge. For cases where an electron–hole plasma is injected (e.g., optically), the change of absorption gives the plasma contribution to the refractive index. In this case, the ξ parameter in Eq. (10) is taken as the change in plasma density regardless of the mechanism of generation or the optical frequency.

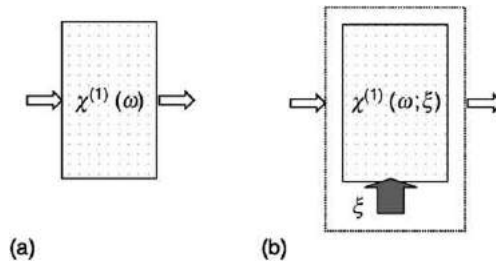


Fig. 1 (a) A causal linear system obeying KK relations. (b) The system in (a) when externally perturbed by ξ . The dotted box now represents our new linear causal system whose altered $\chi^{(1)}$ obeys the KK relations.

Let us now extend this formalism to the case where the perturbation is virtual, occurring at an excitation frequency Ω that is below any material resonance. To the lowest order in the excitation irradiance I_Ω , we write

$$\Delta\alpha(\omega; \zeta) = \Delta\alpha(\omega; \Omega) = 2\alpha_2(\omega; \Omega)I_\Omega \quad (11)$$

and

$$\Delta n(\omega; \zeta) = \Delta n(\omega; \Omega) = 2n_2(\omega; \Omega)I_\Omega \quad (12)$$

where n_2 and α_2 are the nonlinear refractive index and absorption coefficients of the material, respectively. By definition, these coefficients are related to the third-order nonlinear susceptibility $\chi^{(3)}(\omega_1, \omega_2, \omega_3)$ via

$$n_2(\omega; \Omega) = \frac{3}{4\varepsilon_0 n_0(\omega) n_0(\Omega) c} \Re \left\{ \chi^{(3)}(\omega, -\Omega, \Omega) \right\} \quad (13)$$

and

$$\alpha_2(\omega; \Omega) = \frac{3\omega_a}{2\varepsilon_0 n_0(\omega) n_0(\Omega) c^2} \Im \left\{ \chi^{(3)}(\omega, -\Omega, \Omega) \right\} \quad (14)$$

We can therefore write the dispersion relations between α_2 and n_2 :

$$n_2(\omega; \Omega) = \frac{c}{\pi} \mathcal{P} \int_0^\infty \frac{\alpha_2(\omega'; \Omega)}{\omega' 2 - \Omega^2} d\omega' \quad (15)$$

Note that even when the degenerate $n_2(\omega) = n_2(\omega; \omega)$ is desired (at a given ω), the dispersion relation requires that we should know the nondegenerate absorption spectrum $\alpha_2(\omega'; \omega)$ at all frequencies ω' .

Let us pause here and discuss some physical mechanisms that can be involved for a given system of interest. Consider a material characterized by an optical resonance occurring at, say ω_0 (i.e., a degenerate two-level system). For a solid, this resonance can be regarded as that of the fundamental energy gap; $\omega_0 = \omega_g = E_g/\hbar$ in a two-band system. Now, let us examine how the presence of an optical excitation at $\Omega < \omega_0$ can alter the absorption spectrum (at a variable probe ω'). In the quantum mechanical picture, this gives rise to a 'new' material whose perturbed wave functions are 'dressed' by the intensity and frequency of the applied optical field. The lowest-order correction to the absorption is given by $\alpha_2(\omega'; \Omega)$ which involves three major physical processes. Recalling that $\Omega < \omega_0$, these processes include (1) two-photon absorption (2PA) when $\omega' + \Omega \rightarrow \omega_0$ and (2) Raman-induced absorption when $\omega' - \Omega \rightarrow \omega_0$, both implying an absorption of a photon at the probe frequency ω' (i.e., $\alpha_2 > 0$). The third process can be identified as resulting from the blue-shift (for $\Omega < \omega_0$) of the resonance (known as the quadratic optical Stark effect) caused by the excitation field. For our two-level system, the latter results in a decrease followed by an increase in absorption in the vicinity of ω_0 . An example of the overall absorption changes due to such processes is shown in Fig. 2 where $\alpha_2(\omega'; \Omega)$ is qualitatively plotted for a degenerate two-level system. We should note that the relative magnitude of each contribution as well as the width and shape of the resonances are chosen arbitrarily for the purpose of illustration. Using the KK relation in Eq. (15), we can now arrive at the nonlinear index coefficient $n_2(\omega; \Omega)$. The result of this transformation is also given in Fig. 2. The above simple example elucidates the key concepts involving the relationship between nonlinear absorption and refraction in materials for third-order processes. These concepts, when applied more rigorously to semiconductors, have been successful in predicting the sign, magnitude, and

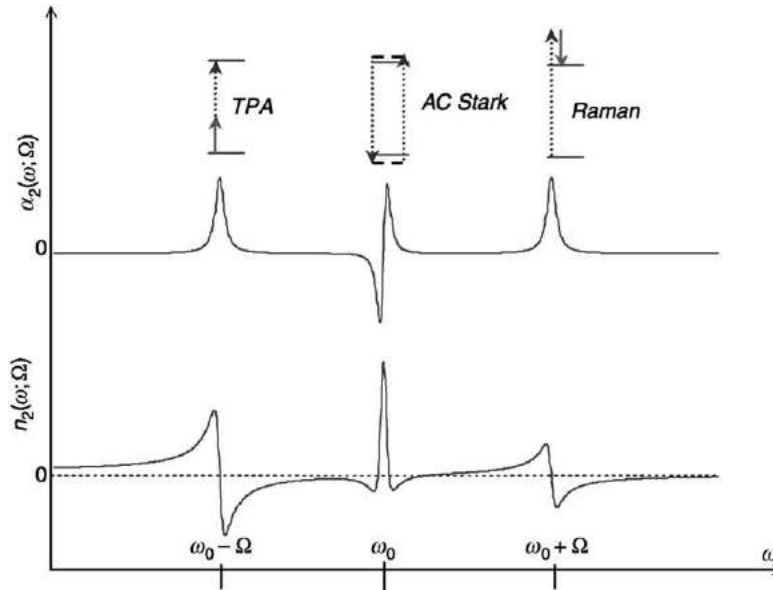


Fig. 2 Upper trace: the nonlinear absorption coefficient in a fictitious 'degenerate' two-level system. Lower trace: the resulting nonlinear refractive index obtained using the KK relations. The insets show the three possible physical mechanisms involved.

dispersion of n_2 due to the anharmonic motion of bound electrons. This will be briefly discussed later. Returning to the mathematical foundation of KK relations, we use Eqs. (13) and (14) to write Eq. (15) in terms of the nonlinear susceptibility $\chi^{(3)}$:

$$\Re\{\chi^{(3)}(\omega_1, \omega_2, -\omega_2)\} = \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\Im\{\chi^{(3)}(\omega', \omega_2, \omega_2)\}}{\omega' - \omega_1} d\omega' \quad (16)$$

The above dispersion relation for $\chi^{(3)}$ was obtained using the physical and intuitive arguments that followed the linearization scheme depicted in Fig. 1. General dispersion relations can be formulated following a mathematical procedure that is similar to the derivation of the linear KK relations. In this case we apply the causality condition directly to the n th-order nonlinear response $R^{(n)}$. For example, without loss of generality, we can write

$$R^{(n)}(\tau_1, \tau_2, \dots, \tau_n) = R^{(n)}(\tau_1, \tau_2, \dots, \tau_n) \Theta(\tau_j) \quad (17)$$

and then calculate the Fourier transform of this equation. Here j can apply to any one of the indices $1, 2, \dots, n$. Following the same procedure as for a linear response, we obtain

$$\chi^{(n)}(\omega_1, \omega_2, \dots, \omega_j, \dots, \omega_n) = \frac{-i}{\pi} \wp \int_{-\infty}^{\infty} \frac{\chi^{(n)}(\omega_1, \omega_2, \dots, \omega', \dots, \omega_n)}{\omega_j - \omega'} d\omega' \quad (18)$$

By separating the real and imaginary parts of this equation, we get the generalized Kramers–Krönig relation pairs for a non-degenerate, n th-order nonlinear susceptibility:

$$\Re\{\chi^{(n)}(\omega_1, \omega_2, \dots, \omega_j, \dots, \omega_n)\} = \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\Im\{\chi^{(n)}(\omega_1, \omega_2, \dots, \omega', \dots, \omega_n)\}}{\omega' - \omega_j} d\omega' \quad (19)$$

and

$$\Im\{\chi^{(n)}(\omega_1, \omega_2, \dots, \omega_j, \dots, \omega_n)\} = -\frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\Re\{\chi^{(n)}(\omega_1, \omega_2, \dots, \omega', \dots, \omega_n)\}}{\omega' - \omega_j} d\omega' \quad (20)$$

In particular, for $\chi^{(3)}$ processes having $\omega_1 = \omega_a$, $\omega_2 = \omega_b$, and $\omega_3 = \omega_c$, this becomes identical to Eq. (16).

Note that in describing the nonlinear susceptibilities, no special attention was given to the harmonic generating susceptibility $\chi^{(N)}(N\omega) \equiv \chi^{(N)}(\omega, \omega, \dots, \omega)$, i.e., the susceptibility generating the N th harmonic at $N\omega$. It turns out that in addition to the KK relations given by Eqs. (19) and (20), the real and imaginary parts of $\chi^{(N)}(N\omega)$ can also be related in different sets of dispersion integrals that involve only the degenerate forms of the susceptibilities. A more general yet simple analysis gives the most general form of KK relations for any type of $\chi^{(n)}$:

$$\chi^{(n)}(\omega_1 + p_1\omega, \omega_2 + p_2\omega, \dots, \omega_m + p_m\omega) = \frac{1}{i\pi} \wp \int_{-\infty}^{\infty} \frac{\chi^{(n)}(\omega_1 + p_1\Omega, \omega_2 + p_2\Omega, \dots, \omega_m + p_m\Omega)}{\Omega - \omega} d\Omega \quad (21)$$

for all $p_1, p_2, \dots, p_m \geq 0$. Setting $\omega_1 = \omega_2 = \dots = \omega_m \equiv 0$, and $p_1 = p_2 = \dots = p_m = 1$ in Eq. (21) yields an interesting form of the KK relations for the N th-harmonic susceptibilities:

$$\Re\{\chi^{(N)}(N\omega)\} = \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\Im\{\chi^{(N)}(N\omega')\}}{\omega' - \omega} d\omega' \quad (22)$$

These dispersion relations have allowed calculations of $\chi^{(2)}(2\omega)$ and $\chi^{(3)}(3\omega)$ in semiconductors using full band structures.

At the beginning of this article, it was noted that all the KK relations for nonlinear optics were known in the early days of the field. Their application in unifying nonlinear absorption (in particular two-photon absorption) and the optical Kerr effect (n_2) in solids came only much later. More recent work demonstrated that the KK relations are a powerful analytical tool in nonlinear optics. Following the picture of a degenerate two-level system shown in Fig. 2, a simple two-band model has been used to calculate the nonlinear absorption coefficient, $\alpha_2(\omega_1; \omega_2)$, resulting from three mechanisms: 2PA, the Raman absorption process, and the ac Stark effect. The optical Kerr coefficient $n_2(\omega_1; \omega_2)$ was then calculated using Eq. (15). Of particular practical interest is the degenerate case ($\omega_1 = \omega_2 = \omega$), from which the 2PA coefficient $\beta(\omega) = \alpha_2(\omega; \omega)$ can be extracted. Fig. 3 depicts the calculated dispersion of n_2 and β as a function of $\hbar\omega/E_g$ where E_g is the bandgap energy of the solid. The dispersion of n_2 and its sign reversal shown in Fig. 3 has been observed experimentally in many optical solids.

Finally, let us discuss a related implication of causality in nonlinear optics. The KK dispersion relations are traditionally derived in terms of internal material parameters such as susceptibility, absorption coefficient, and refractive index. Similar to the case of electrical circuits, one can obtain dispersion relations that apply to an external transfer function of the system that relates an input signal to an output signal. In this case, the dispersion of the transfer function includes system structure as well as the intrinsic dispersion of the material. As an optical (and linear) example, consider a Fabry–Perot etalon. The optical transmission of this system has well-known spectral features that are primarily caused by structural dispersion (i.e., interference) in addition to the intrinsic dispersion of the material. Causality still demands that the transmitted signal has a phase variation whose value and dispersion can be determined using a KK relation linking the real and imaginary parts of the transmission coefficient. In other words, the KK relations provide a spectral correlation between the real and imaginary components of the transfer function which in turn may translate to a spectral correlation between the phase and amplitude of the transmitted signal. However, the variations in phase do not necessarily imply the presence of a varying index of refraction, nor does an amplitude variation suggest the existence of material absorption (dissipation). Ultimately, this implies that any mechanism causing a variation in amplitude (including reflection, scattering, or absorption) must be accompanied by a phase variation. (One should note that the reverse of

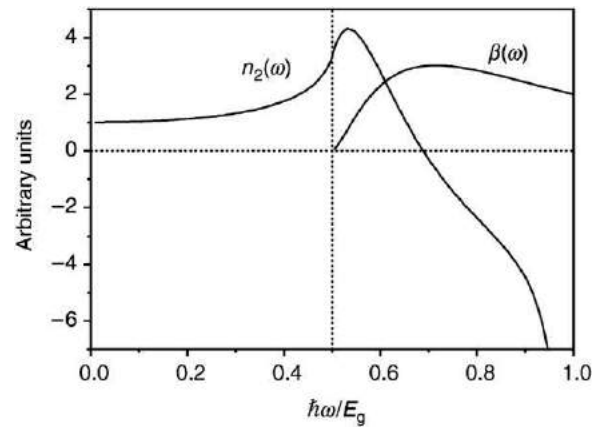


Fig. 3 The two-photon absorption coefficient in semiconductors (β) calculated for a two-band model. The resultant nonlinear refractive index (n_2) obtained using a KK transformation of the calculated nondegenerate nonlinear absorption coefficient includes all major mechanisms.

the previous statement is not necessarily true; i.e., a variation in phase does not have to be accompanied by an amplitude modulation.)

In nonlinear optics with the ‘black box’ approach of Fig. 1, the optical perturbation ξ (with frequency Ω) can render an amplitude variation in the probe (at ω) using various frequency mixing schemes in a noncentrosymmetric material (i.e., with nonzero $\chi^{(2)}$). For instance, the probe at ω can be depleted by nonlinear conversion to $\omega_{\text{sum}} = \omega + \Omega$ via sum-frequency generation involving $\chi^{(2)}(\omega, \Omega)$ and/or to $\omega_{\text{diff}} = \omega - \Omega$ via difference-frequency generation involving $\chi^{(2)}(\omega, -\Omega)$. Such a conversion (or depletion) should be accompanied by a phase variation according to the KK dispersion relations. This type of nonlinear phase modulation is known as a $\chi^{(2)}\text{:}\chi^{(2)}$ cascaded nonlinearity. Such cascaded processes are routinely (and more simply) analyzed with Maxwell’s equations governing the propagation of beams in a second-order nonlinear material. The KK relations, however, provide an interesting physical perspective of the process. We find that cascaded second-order nonlinearities are yet another manifestation of causality in nonlinear optics.

Further Reading

- Bassani, F., Scandolo, S., 1991. Dispersion-relations and sum-rules in nonlinear optics. *Physical Review B—Condensed Matter* 44, 8446–8453.
- Caspers, P.J., 1964. Dispersion relations for nonlinear optics. *Physical Review A* 133, 1249–1251.
- Hutchings, D.C., Sheik-Bahae, M., Hagan, D.J., Van Stryland, E.W., 1992. Kramers–Krönig relations in nonlinear optics. *Optical and Quantum Electronics* 24, 1–30.
- Kogan, S.M., 1963. On the electromagnetics of weakly nonlinear media. *Soviet Physics JETP* 16, 217–219.
- Nussenzveig, H.M., 1972. *Causality and Dispersion Relations*. New York: Academic Press.
- Price, P.J., 1964. Theory of quadratic response functions. *Physical Review* 130, 1792–1797.
- Ridener, F.L.J., Good, R.H.J., 1975. Dispersion relations for nonlinear systems of arbitrary degree. *Physical Review B* 11, 2768–2770.
- Sheik-Bahae, M., Van Stryland, E.W., 1999. Optical nonlinearities in the transparency region of bulk semiconductors. In: Garmire, E., Kost, A. (Eds.), *Nonlinear Optics in Semiconductors I* 58. Academic Press, pp. 257–318.
- Sheik-Bahae, M., Hutchings, D.C., Hagan, D.J., Van Stryland, E.W., 1991. Dispersion of bound electronic nonlinear refraction in solids. *IEEE Journal of Quantum Electronics* 27, 1296–1309.
- Smet, F., Vangroenendael, A., 1979. Dispersion-relations for N -order non-linear phenomena. *Physical Review A* 19, 334–337.
- Toll, J.S., 1956. Causality and the dispersion relation: logical foundations. *Physical Review* 104, 1760–1770.

Nonlinear Optical Phase Conjugation

BY Zeldovich, University of Central Florida, Orlando, FL, USA

© 2005 Elsevier Ltd. All rights reserved.

Phase conjugation (PC) beams and their applications are best illustrated by Fig. 1, which shows the passage of a coherent planewave incident upon a transparent optical element with inhomogeneities of refractive index (Fig. 1(a)). Reversibility of light propagation implies the existence of such an 'anti-distorted' beam, which reverts back to a planewave after reversing through the same inhomogeneities (Fig. 1(b)). A real-valued monochromatic optical wave, $E_{\text{real}}(\mathbf{R}, t)$, may be represented by the complex amplitude $E(\mathbf{R})$ via $E_{\text{real}}(\mathbf{R}, t) = 0.5[E(\mathbf{R})\exp(-i\omega t) + E^*(\mathbf{R})\exp(i\omega t)]$, where $E^*(\mathbf{R})$ represents the complex conjugate of $E(\mathbf{R})$:

$$\begin{aligned} E(\mathbf{R}) &= |E(\mathbf{R})|\exp[i\varphi(\mathbf{R})] \\ E^*(\mathbf{R}) &= |E(\mathbf{R})|\exp[-i\varphi(\mathbf{R})] \end{aligned} \quad (1)$$

Mathematical expression of time-reversal, $E_{\text{real}}(\mathbf{R}, t \rightarrow E_{\text{real}}(\mathbf{R}, -t)$, becomes the exchange $E(\mathbf{R}) \leftrightarrow E^*(\mathbf{R})$ for monochromatic waves. The reversibility of propagation means that the complex propagating field $E(\mathbf{R})$ (for example, planewave $\exp(i\mathbf{k} \cdot \mathbf{R})$), is a wave equation solution equivalent to $E^*(\mathbf{R})$ (in this example, $\exp(-i\mathbf{k} \cdot \mathbf{R})$). Conjugation of complex amplitude means reversal of the sign of the phase, and a mixed label PC has been coined and nowadays is firmly established. Terms such as 'wave front reversal' and 'generation of time-reversed replica of the beam' are also used to describe PC.

Fig. 1 also illustrates one of the most important applications of PC. If the element in question is a laser-type amplifier, then double-passage allows the extraction of energy from the optically inhomogeneous laser medium in the form of a perfectly collimated beam of diffraction-limited divergence. A PC device may also serve as a mirror of a laser resonator, resulting in the same beam-correction purpose.

One of the most practical and robust methods of PC is based on stimulated Brillouin back-scattering (SBS). The incident beam illuminates the SBS-active transparent medium, for example, liquids (CS_2 , CCl_4 , acetone, etc.), compressed gases (CH_4 , SF_6 , etc.), or solids (fused or crystalline quartz, glass, etc.). This 'pump' beam, $E(\mathbf{R}) \equiv E(\mathbf{r}, z)$, must possess well-developed transverse inhomogeneities of intensity $|E(\mathbf{r}, z)|^2$. Here, $\mathbf{r} = \{x, y\}$ is the part of coordinate vector transverse with respect to the central direction z of the beam in question, and $\mathbf{R} = \{\mathbf{r}, z\}$. These inhomogeneities may constitute a narrow focal waist in the case of weakly distorted focused beams, speckle-inhomogeneities in cases of strong distortions, or a combination thereof; see Fig. 2, where solid lines symbolize the 'rays' of the incident pump.

Spontaneous scattering of the pump results in a multitude of possible transverse profiles $S(\mathbf{r}, z = 0)$ of the signal, whose 'rays' are depicted via the dotted lines on Fig. 2. The signal is amplified exponentially due to SBS processes. This exponent in a simplified

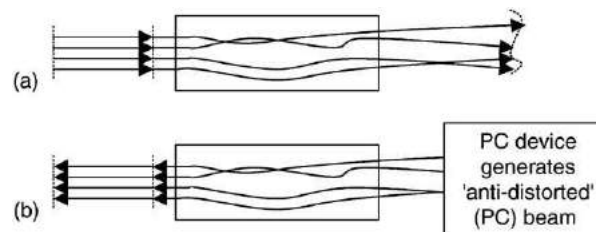


Fig. 1 (a) Collimated beam is distorted by propagation through inhomogeneous medium. (b) Conjugate, a.k.a. antidistorted beam becomes collimated after backward passage through the same medium.

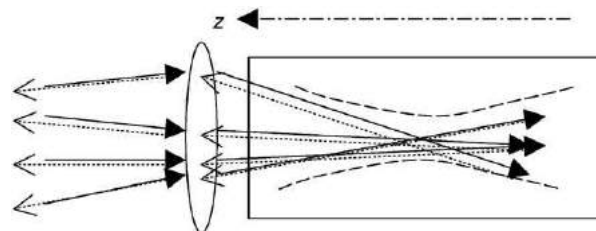


Fig. 2 Stimulated Brillouin Scattering (SBS) method of PC. Solid lines symbolize the rays of incident 'beam', while dotted lines depict the rays of the signal amplified by the SBS process.

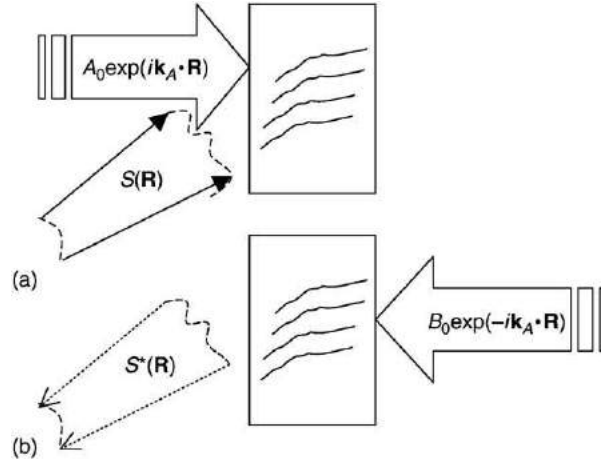


Fig. 3 (a) Recording of a hologram in a photosensitive medium by a reference wave A and signal wave S. (b) Read-out of that hologram by the wave B counter-propagating to A results in generation of the conjugated signal S^* .

form may be represented as $|S(\mathbf{r}, z)|^2 \propto \exp(\int g dz)$, with the gain $g(z)$ being the result of transverse overlapping of the intensity profiles of signal and pump:

$$g(z) \approx \text{const} \cdot \langle |E(\mathbf{r}, z)|^2 \cdot |S(\mathbf{r}, z)|^2 \rangle / \langle |S(\mathbf{r}, z)|^2 \rangle \quad (2)$$

Thus, the similarity of the intensity profile of the signal to that of the pump is rewarded by the exponential preference in the output signal level. However, this is not enough to guarantee that the output backward-scattered signal $S(\mathbf{r}, z)$ is phase conjugate with respect to the pump. It is here that the mutual reversibility of propagation of strongly inhomogeneous signal and pump becomes crucial. Namely, good transverse overlapping is sustained along the entire interaction length z , only if $S(\mathbf{r}, z) \propto E^*(\mathbf{r}, z)$. This constitutes the ‘discrimination mechanism’ of SBS-PC: the ‘conjugate mode’ of the signal has a z -sustained advantage in exponential gain and, under appropriate conditions, suppresses all nonconjugate signal configurations in competition for pump power. Nonlinear-optical wave theory of this discrimination process has been developed further.

The holographic mechanism of PC has a different nature. It may take the form of static holography or dynamic holography; the latter essentially may be considered as a part of nonlinear optics. Here is a simplified description of one of the variants. Suppose one wants to obtain a conjugate replica of the incident monochromatic signal, whose complex profile is $S(\mathbf{R})$. A plane reference wave $A(\mathbf{R}) = \exp(i\mathbf{k}_A \cdot \mathbf{R})$ is also directed to the registering medium. If $S(\mathbf{R})$ and $A(\mathbf{R})$ are mutually coherent, then an interference pattern of intensity is produced $\{S^*(\mathbf{R})A(\mathbf{R}) + \text{compl.conj.}\}$. The medium records this pattern in the form of the modulation of refractive index and/or of the absorption coefficient, $\delta\epsilon(\mathbf{R}) \propto (c_1 + ic_2)\{S^*(\mathbf{R})A(\mathbf{R}) + \text{c.c.}\}$ (Fig. 3(a)). At the ‘reconstruction’ stage, the hologram is illuminated by another plane reference wave, $B(\mathbf{R}) = B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$. In the specific case, when $\mathbf{k}_B = -\mathbf{k}_A$, the source δD_{conj} of the reconstructed wave becomes, as shown by Fig. 3(b):

$$\delta D_{\text{conj}} = (c_1 + ic_2)[S^*(\mathbf{R})A(\mathbf{R})] \cdot B(\mathbf{R}) \propto S^*(\mathbf{R})(c_1 + ic_2)A_0B_0 \quad (3)$$

Reversibility of propagation laws guarantees that this source will indeed excite a conjugate wave with high efficiency. Both processes, recording of the interference pattern and reconstruction of the conjugate wave, may occur simultaneously, if one deals with the dynamic holography. The same process may also be described as nonlinear-optical four-wave mixing (FWM) via cubic nonlinearity $\chi^{(3)}$:

$$\delta D_{\text{conj}}(\mathbf{R}) = \chi^{(3)} S^*(\mathbf{R}) A(\mathbf{R}) B(\mathbf{R}) = \chi^{(3)} S^*(\mathbf{R}) A_0 B_0 \quad (4)$$

with the four waves; A, B, S, S^* .

Important characteristics of a PC device are: fidelity of conjugation, efficiency of returning back-reflected power/energy, and reaction time or build-up time. The fidelity parameter must show how close the output of the device $E_{\text{out}}(\mathbf{r})$ is to the perfectly conjugate profile $S^*(\mathbf{R})$ of the incident signal (up to an arbitrary constant complex factor). Among other definitions, the square modulus of the normalized transverse overlapping integral of the fields is often discussed:

$$f = \frac{|\int \int E_{\text{out}}(x, y) S(x, y) dx dy|^2}{(\int \int |E_{\text{out}}(x, y)|^2 dx dy \cdot \int \int |S(x, y)|^2 dx dy)} \quad (5)$$

To achieve good (close to 1) fidelity of PC in the holographic or FWM scheme, one has to guarantee that the reference waves, $A_0 \exp(i\mathbf{k}_A \cdot \mathbf{R})$ and $B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$ are exactly conjugate to each other, $\mathbf{k}_A = -\mathbf{k}_B$. That, in turn, requires the absence of any distortions in the holographic or nonlinear medium. To the contrary, there are very modest requirements on the phase inhomogeneities of the medium in the SBS scheme of PC, and for that reason it is usually labeled as a scheme of self-phase conjugation.

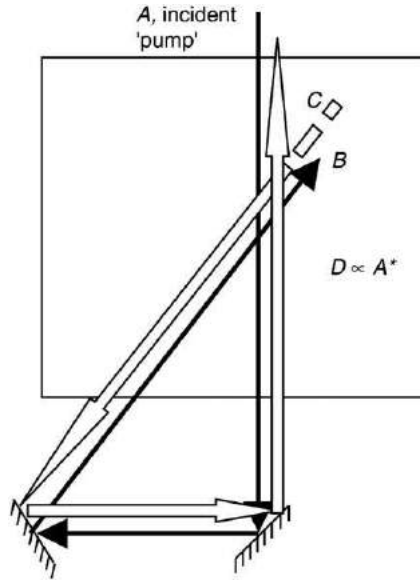


Fig. 4 Tail-biting scheme of phase conjugation.

A hybrid scheme, Brillouin-assisted FWM, has been suggested and implemented, where mutually conjugate reference waves are produced via SBS-PC, while FWM, in a separate medium, exploits SBS nonlinearity. This scheme has produced extremely high, up to about 10^{10} , reflectivity, accompanied by an extremely low-noise detection of incoming signals.

A number of other PC schemes were first realized with the use of an important class of materials for nonlinear optics: photo-refractive crystals (PRC). These are crystals where ionization of the dopants by the incident light creates interference pattern of charge separation. The linear electro-optic (Pockels) effect transforms the resulting patterns of the electrostatic field into patterns of refractive index, thus creating dynamic holograms. A remarkable property of these crystals is that most of them act as very good electrical insulators. Therefore, the only source of conductivity, which tends to erase the hologram, is photoconductivity, the value of the latter being proportional to the intensity of incident light itself. For this reason, the steady-state strength of the hologram turns out to be independent of the intensity. It is the build-up time that must be increased, if intensity is low. One can achieve various processes in PRC, including those similar to the $\chi^{(3)}$ -type optical Kerr effect, stimulated scattering, etc.

The simplest use of PRC for generation of PC waves is to devise a FWM scheme (see Fig. 3). However, a multitude of other, nontrivial schemes have been devised and implemented. The most impressive use of PRC for PC is based on interactions of the beams in the vicinity of a corner of a rectangular crystal, typically BaTiO₃, these interactions being of the stimulated scattering type. To honor J. Feinberg's cat, whose image was reconstructed in the first demonstration, this scheme is universally called 'cat conjugator'. It turned out to be a very reliable and robust scheme of self-phase-conjugation (SPC).

Another important group of the SPC scheme is called 'tail-biting'. One of the variants of the latter is shown in Fig. 4. Incident 'pump' wave $A(\mathbf{R})$ whose transverse profile one wants to conjugate, enters the medium, and then is redirected by mirrors back into the same medium, to 'bite itself'. For the purpose of discussion, this redirected pump is labeled as wave $B(\mathbf{R})$ on its second arrival into the medium. Spontaneous scattering results in a seed for the amplification of the wave $C(\mathbf{R})$ via the process of stimulated scattering (SS) $A(\mathbf{R}) \rightarrow C(\mathbf{R})$.

Wave $C(\mathbf{R})$ is also redirected into the medium by the same mirrors, and is labeled as wave $D(\mathbf{R})$ on its second arrival into the medium. Both $C(\mathbf{R})$ and $D(\mathbf{R})$ are depicted in Fig. 4 by white arrows. Wave $D(\mathbf{R})$ is also amplified due to the SS process $B(\mathbf{R}) \rightarrow D(\mathbf{R})$.

As it follows from the macroscopic description of SS process, volume gratings of refractive index are recorded in the medium, $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R}) + B^*(\mathbf{R})D(\mathbf{R})]$. Among all the transverse profiles for the seed $C(\mathbf{R})$ the one that happens to be proportional to $B^*(\mathbf{R})$ will be reflected by the mirrors into the medium in the form $D(\mathbf{R}) \propto A^*(\mathbf{R})$. The latter case is again a consequence of the reversibility of propagation laws, where both gratings have the same profile and add coherently in the form $A^*(\mathbf{R})B^*(\mathbf{R})$. What is even more important, the grating $B^*(\mathbf{R})D(\mathbf{R}) \propto B^*(\mathbf{R})A^*(\mathbf{R})$ serves to close the feedback loop for seeding the appropriate $C(\mathbf{R})$. This, and other variants of tail-biting schemes, have also proved to be reliable and robust. Quite often a laser amplifier is inserted in the path $A \rightarrow B$ and $C \rightarrow D$, thus enhancing the feedback.

Another important scheme is traditionally called 'double PC'; albeit it may be better described as 'mutual PC'. Fig. 5 should help in understanding that scheme. Consider the beam $A(\mathbf{R})$ incident to a medium active with respect to the stimulated-scattering-type process $A(\mathbf{R}) \rightarrow C(\mathbf{R})$. This process originates from a random seed and results in a fan of different waves $C(\mathbf{R})$. Such a process is characteristic of PRCs and is a consequence of recording dynamic gratings of refractive index $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R})]$. Suppose another beam $B(\mathbf{R})$, typically incoherent with respect to $A(\mathbf{R})$, illuminates the medium from the other side, and also is engaged in the fanning process, this time $B(\mathbf{R}) \rightarrow D(\mathbf{R})$, with the grating of refractive index $\delta n(\mathbf{R}) \propto -i[B^*(\mathbf{R})D(\mathbf{R})]$ involved. Among all the

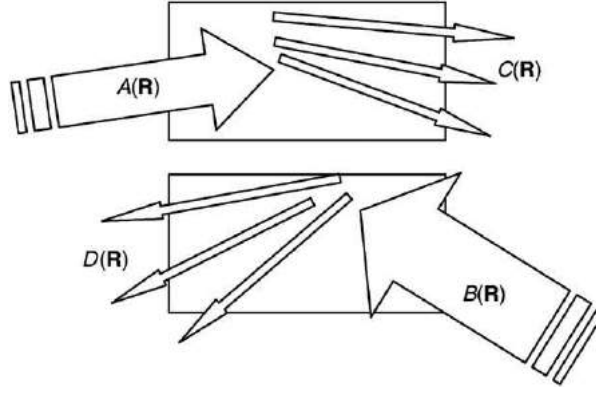


Fig. 5 Towards the mechanism of double PC (mutual PC).

transverse profiles for the seed grating $C(R)A^*(R)$, the one that happens to be proportional to $B^*(R)A^*(R)$ will be supported by similar profiles in the $B(R) \rightarrow D(R)$ scattering. In this way, two fanning processes enhance each other and close the feedback loop. An important consequence is that one obtains the following output waves:

$$\begin{aligned} C(R) &= \text{const}_1 \times B^*(R) \\ D(R) &= \text{const}_2 \times A^*(R) \end{aligned} \quad (6)$$

One can say that the output wave $C(R)$ presents a conjugate replica of $B(R)$ -beam, while the output wave $D(R)$ presents a conjugate replica of $A(R)$ -beam; hence the label 'mutual PC'. In some experiments the incident beams, $A(R)$ and $B(R)$, were of somewhat different colors (i.e., from different lasers), or their pulses did not overlap in time in the medium. It is the grating memory that 'informs' one beam about the presence of the other. The mechanism of cat conjugator is considered to be a combination of tail-biting and double PC schemes.

Pumped laser material with saturable gain often serves as an important nonlinear-optical medium for implementing PC. One of the evident advantages is that possible low efficiency of properly nonlinear process may be compensated by the gain in the same piece of material.

Considerable attention is paid nowadays to the chirp reversal of optical pulses that are used in fiber optical communications. Such chirped pulses, $S(t) = \exp(-i\omega_0 t + i\beta t^2/2 - t^2/\tau_1^2)$, with the value of the 'chirp' $d\omega/dt = -\beta$ and an increased pulse duration τ_1 , arise as a result of propagation in a fiber which possesses group velocity dispersion (GVD). Pulse stretching and chirp are the consequences of different propagation times for the different frequency constituents of the pulse. If some device changes the sign of that chirp, then subsequent propagation of the pulse, through a piece of fiber with the same GVD, restores the duration of the pulse to original shorter value τ_0 . A scheme has been suggested to use nonlinear optical FWM in a $\chi^{(3)}$ -medium or three-wave mixing (ThWM) in a $\chi^{(2)}$ -medium:

$$\begin{aligned} \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) A(t) B(t) \\ \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) C(t) \end{aligned} \quad (7)$$

Indeed, complex conjugation of the time dependence of the signal's field is equivalent to the change of the chirp sign. The second (ThWM) expression is written assuming that the reference wave $C(t) \propto \exp(-2i\omega_0 t)$, so that it has the frequency double that of the signal. The same process of ThWM has been shown to yield the conjugation of transverse phase profile of the beams, but for a number of reasons it was not applied.

As for the applications of PC, one of them was discussed above (Fig. 1: double-pass scheme of a laser amplifier). Lasers with one or two PC mirrors, or with more complicated PC scheme, promise generation of high-transverse-quality beams with the use of realistic laser media, the latter inevitably possessing thermal and other distorting inhomogeneities. Another important potential application is free-space optical communications: Earth–Earth or Earth–satellite through atmosphere, or satellite–satellite communications. An 'interrogating' beam may be sent in one direction through an imperfect optical system and/or through turbulent atmosphere. Conjugation of transverse profile of the 'interrogating' beam at the other end of the communication link leads to an almost perfect redirection of the reflected beam towards the 'beacon' source, while time modulation may carry the information.

Readout of information data from a holographic storage is almost always performed in the regime of reading the conjugate wave, since it corrects for the most part of aberrations of optical systems in question.

To conclude, one may use a citation from the two consecutive *Scientific American* popular papers on PC: 'at present the number of applications would seem to be limited only by imagination.'

Reviews of various aspects of phase conjugation may be found in monographs and *Scientific American* papers listed below. The author of the present article was introduced to PC by his colleague V. V. Ragulskii, and has greatly benefited from collaboration with V. V. Ragulskii and V. V. Shkunov.

See also: Adaptive Optics. Information Theory in Imaging

Further Reading

- Bespalov, V.I., Pasmanik, G.A., 1994. Nonlinear Optics and Adaptive Laser Systems. Commack, NY: Nova Science Publishers.
- Feinberg, J., 1986. Self-pumped, continuous wave phase conjugator using internal reflection. *Optics Letters* 7, 486.
- Fisher, R.A. (Ed.), 1983. Optical Phase Conjugation. New York: Academic Press.
- Gower, M., Proch, D. (Eds.), 1994. Optical Phase Conjugation. Berlin: Springer Verlag.
- Hellwarth, R.W., 1982. Optical beam conjugation by stimulated backscattering. *Optical Engineering* 21, 257.
- Nosatch, O.Y., Popovichev, V.I., Ragulskii, V.V., Faizullov, F.S., 1972. Compensation of phase distortions in an amplifying medium by a 'Brillouin mirror'. *JETP Letters* 16, 435.
- Pepper, D.M., 1986. Applications of optical phase conjugation. *Scientific American* 254, 74–83.
- Shkunov, V.V., Zeldovich, B.Y., 1985. Optical phase conjugation. *Scientific American* 253, 54–59.
- Yariv, A., 1978. Phase conjugate optics and real-time holography. *IEEE Journal of Quantum Electronics* 14, 650.
- Zeldovich, B.Y., Mamaev, A.V., Shkunov, V.V., 1995. *Speckle-Wave Interactions in Application to Holography and Nonlinear Optics*. Boca Raton, FL: CRC Press.
- Zeldovich, B.Y., Popovichev, V.I., Ragulskii, V.V., Faizullov, F.S., 1972. On relation between wavefronts of reflected and exciting radiation in stimulated Brillouin scattering. *JETP Letters* 16, 109.
- Zeldovich, B.Y., Shkunov, V.V., Pilipetsky, N.F., 1985. *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Photorefraction

M Cronin-Golomb, Tufts University, Medford, MA, USA

B Kippelen, University of Arizona, Tucson, AZ, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

γ	Amplitude gain [cm^{-1}]	N_D^i	Ionized donor density [m^{-3}]
E_0	Applied dc field [Vm^{-1}]	E_Q	Limiting space charge field [Vm^{-1}]
j	Current [Am^{-2}]	μ	Mobility [$\text{m}^2 \text{V}^{-1} \text{sec}^{-1}$]
ε	Dielectric permittivity [Farads m^{-1}]	E_μ	Mobility field [Vm^{-1}]
E_D	Diffusion field [Vm^{-1}]	I	Normalized intensity [$\text{V}^2 \text{m}^{-2}$]
N_A	Electron acceptor density [m^{-3}]	s	Photo-ionization coefficient [kg sec^{-2}]
e	Electron charge [C]	τ	Photorefractive time constant [sec]
N_D	Electron donor density [m^{-3}]	r	Pump ratio [1]
k_g	Grating wavenumber [m^{-1}]	ω	Radian frequency [sec^{-1}]
Γ	Intensity gain [cm^{-1}]	γ_R	Recombination coefficient [$\text{m}^3 \text{sec}^{-1}$]
ℓ	Interaction length [cm]	A	Slowly varying optical electric field [Vm^{-1}]

Introduction and Basics of the Photorefractive Effect

In 1966 Ashkin and coworkers were pursuing research on optical devices using the electro-optic material lithium niobate. They noticed that the refractive index of lithium niobate would change when it was exposed to laser light, and upset the expected operation of their devices. They called the effect optical damage. Shortly thereafter, Chen, LaMacchia, and Fraser reported on the use of the optical damage effect for holographic data storage. Thus began the field of photorefractive nonlinear optics, which has been used in various applications such as real-time holography, optical data storage, optical image amplification, nondestructive testing, distortion compensation by phase conjugation, pattern recognition, and radar signal processing. Many inorganic and organic materials have been investigated for their photorefractive effects, including ferroelectrics, semiconductors, and sensitized polymers. The most well-known inorganic materials are lithium niobate, bismuth silicon oxide, barium titanate, and strontium barium niobate. Most organic photorefractive materials are based on polymeric photoconductors such as those used in xerography that are doped with electro-active molecules, some plasticizers, and sensitizers.

While in its broadest interpretation, the photorefractive effect occurs whenever light incident on a material causes a refractive index change, one usually applies the term to electro-optic index changes resulting from optically generated space charge fields. Materials that are photorefractive in this sense share the following properties

- high transmission at the operating wavelengths;
- linear electro-optic coefficients or orientational Kerr effects;
- charge carriers that become mobile when optically excited;
- trapping centers for these charge carriers to enable spatially non-uniform redistribution of charge.

Consider two beams from the same laser crossing inside a photorefractive material such as barium titanate. The interference pattern will have bright and dark fringes. Charge carriers are excited where the light is bright, then drift and diffuse to regions of relative darkness where they preferentially recombine into trapping centers. In this way, a net excess charge develops in the dark regions, and a net deficit of charge develops in the bright regions. The spatially varying charge distribution has an electric field associated with it and this electric field causes a spatially varying refractive index profile. Because the space charge, its field, and resulting refractive index have the same spatial periodicity as the original interference pattern we have a holographic phase grating. The diffraction efficiency of the hologram can easily approach 100% for materials such as barium titanate and strontium barium niobate which have high electro-optic coefficients. Likewise, such high diffraction efficiencies are easily obtained in 100 micrometer-thick photorefractive polymer films.

The Standard Rate Equation Model

The development of photorefractive gratings can be modeled using standard semiconductor rate equations. Fig. 1 shows two beams incident symmetrically on a photorefractive crystal or polymer. They form an interference pattern whose intensity may be written:

$$I(x) = I_0(1 + m \cos(k_g x)) \quad (1)$$

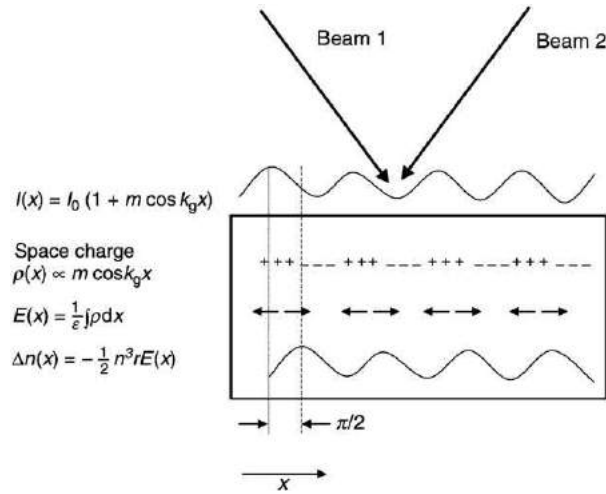


Fig. 1 Diagram of two-beam coupling process showing optical interference pattern, space charge, resultant electric field, and $\pi/2$ phase shifted electro-optically induced refractive index grating.

where x is the direction perpendicular to the interference fringes, I_0 is the average intensity, m is the modulation index, and k_g is the wavenumber of the interference pattern.

The crystal may be considered a wide bandgap semiconductor containing electron donors in the bandgap with density N_D and electron acceptors with density N_A . It is assumed that some electrons ionized from the donors permanently occupy these acceptors so that when charge in the crystal is distributed uniformly in the dark, the number density of ionized donors N_D^i is equal to N_A . Likewise, polymers contain donor and acceptor-like molecules that can be neutral or ionized. The spatially varying light distribution ionizes the donors at the following rate, assuming that $N_D^i \ll N_D$:

$$\frac{\partial N_D^i}{\partial t} = s I N_D - \gamma_R n_e N_D^i \quad (2)$$

where s is a photoionization coefficient, γ_R is a recombination parameter, and n_e is the density of excited charge carriers, which we assume here to be electrons. The model can easily be generalized to include holes. We also use the equation of charge conservation:

$$\frac{\partial N_D^i}{\partial t} = \frac{\partial n_e}{\partial t} - \frac{1}{e} \nabla \cdot \mathbf{j} \quad (3)$$

where e is the charge of an electron, and $\mathbf{j}$ is the current in the conduction band.

$$\mathbf{j} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e \quad (4)$$

where μ is the electron mobility, k_B is Boltzmann's constant, and T is the temperature. The electric field obeys Gauss's law:

$$\nabla \cdot \mathbf{E} = -e(n_e + N_A - N_D^i)/\epsilon \quad (5)$$

where N_A is the density of acceptors. These equations may be linearized by approximating the electron density, ionized donor density, and electric field with their first Fourier components:

$$\begin{aligned} E &= E_0 + \frac{1}{2} (E_1 \exp(ik_g x) + E_1^* \exp(-ik_g x)) \\ N_D^i &= N_{D0}^i + \frac{1}{2} (N_{D1}^i \exp(ik_g x) + N_{D1}^{i*} \exp(-ik_g x)) \\ n_e &= n_{e0} + \frac{1}{2} (n_{e1} \exp(ik_g x) + n_{e1}^* \exp(-ik_g x)) \end{aligned} \quad (6)$$

This assumption is strictly valid only when the modulation index m is much less than unity. Otherwise, a generalization to higher orders in the Fourier series is required. However, the linearized theory is sufficient to illustrate the most important features of the photorefractive effect. The solution for the space charge field E_1 for the case when the interference pattern is applied at time $t=0$ is

$$E_1 = m \frac{-iE_Q(E_0 + iE_D)}{E_0 + i(E_D + E_Q)} (1 - \exp(-t/\tau)) \quad (7)$$

where E_0 is an externally applied or photovoltaic field (if any), E_D is the diffusion field

$$E_D = \frac{k_B T k_g}{e} \quad (8)$$

and E_Q is the limiting space charge field

$$E_Q = \frac{eN_A}{\epsilon k_g} \quad (9)$$

Some photorefractive crystals, most notably LiNbO_3 , exhibit the photovoltaic effect in which optical illumination induces a dc field across the crystal.

The sinusoidally varying space charge field E_1 operates through the linear electro-optic effect with effective coefficient r to produce a sinusoidal variation in the refractive index n of the crystal:

$$n(x) = n_0 + \frac{1}{2} (n_1 \exp(ik_g x) + n_1^* \exp(-ik_g x)) \quad (10)$$

where

$$n_1 = -\frac{1}{2} r n_0^3 E_1 \quad (11)$$

The effective electro-optic coefficient r may be found from tensor analysis of the electro-optic tensor and the vector space charge and optical fields. Notice that there is a 90-degree phase shift between the interference pattern and the refractive index grating when E_0 is zero. The time constant τ is given by

$$\tau = \frac{N_A}{sN_D I_0} \frac{E_0 + i(E_D + E_\mu)}{E_0 + i(E_D + E_Q)} \quad (12)$$

where E_μ is a mobility field

$$E_\mu = \frac{\gamma_R N_A}{\mu k_g} \quad (13)$$

When E_0 is nonzero, there is an oscillatory component to the time constant.

In contrast to the case of ordinary optical nonlinearities such as the optical Kerr effect, in which the nonlinear coupling strength is proportional to the optical intensity, the steady state strength of the photorefractive effect is independent of optical intensity while the time constant is inversely proportional to intensity in the basic model described above. The time constant varies with the photoconducting performance of a given material. In the fastest polymeric and inorganic materials it is, at the time of writing, of the order of a few milliseconds at 1 W cm^{-2} of incident optical intensity.

At low intensities (below the equivalent dark intensity, 1 W cm^{-2} in barium titanate), the above equations need to be modified to include the effect of dark conductivity. Even in the dark, there will be a few mobile charge carriers in the conduction band that tend to erase the grating. This will result in the effect appearing more Kerr-like, except still with the 90-degree phase shift between the index grating and the interference pattern.

Coupled Wave Equations

The change in refractive index n_1 can be large enough to produce substantial interactions between the writing beams. The writing beams generate a phase grating that diffracts the beams into each other. The grating influences the writing beams, which in turn influence the grating. In the cases where the diffusion field dominates, for example when the externally applied or photovoltaic field E_0 is absent, one beam is amplified by in-phase diffraction of the other beam from the grating. As shown in Fig. 2, this amplification results from a 90-degree phase shift due to diffraction from a phase grating coupled with a -90 -degree phase shift from the spatial phase shift between the interference pattern and the index grating. The second beam is attenuated by destructive

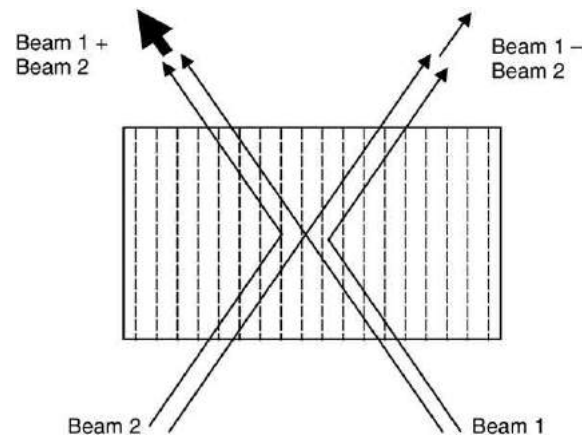


Fig. 2 Two-beam coupling amplification. Beam 1 is amplified by constructive interference. Beam 2 is de-amplified by destructive interference.

interference with the first beam diffracted by the grating. These interactions can be modeled by standard coupled wave theory. Let the electric field amplitude associated with the j th beam be

$$E_j(\mathbf{r}, t) = \mathbf{e}[A_j(\mathbf{r})\exp(i(\mathbf{k}_j \cdot \mathbf{r} - \omega t)) + A_j^*\exp(-i(\mathbf{k}_j \cdot \mathbf{r} - \omega t))] \quad (14)$$

where $\mathbf{e}$ is the polarization unit vector. Using the scalar wave equation

$$\nabla^2 E + k^2 E = 0 \quad (15)$$

and the slowly varying envelope approximation

$$\frac{d^2 A_j}{dz^2} \ll k \frac{dA_j}{dz} \quad (16)$$

we find

$$\begin{aligned} \frac{dA_1}{dz} &= +\gamma \frac{A_1 |A_2|^2}{I_0} \\ \frac{dA_2^*}{dz} &= -\gamma \frac{A_2^* |A_1|^2}{I_0} \end{aligned} \quad (17)$$

with the coupling constant γ given by

$$\gamma = -\frac{i\omega\Delta n_m}{c \cos\vartheta} \quad (18)$$

where ϑ is the half-angle between the writing beams and Δn_m is given by

$$\Delta n_m = -\frac{1}{2}n_0^3 r \frac{E_1}{m} \quad (19)$$

Eq. (17) shows that beam 1 is amplified and beam 2 is attenuated. That beam 1 is amplified instead of beam 2 is a result of the choice of crystal orientation and hence the sign of γ . These nonlinear equations can be solved for normalized intensity $I_j = |A_j|^2$:

$$\begin{aligned} I_1(z) &= \frac{I_0}{1 + (I_2(0)/I_1(0))\exp(+\Gamma z)} \\ I_2(z) &= \frac{I_0}{1 + (I_1(0)/I_2(0))\exp(-\Gamma z)} \end{aligned} \quad (20)$$

where $\Gamma = \gamma + \gamma^*$ is the intensity coupling constant. In the limit where $I_1 \ll I_2$, the weak beam, beam 1, experiences exponential amplification. This amplification can be used to build optical oscillators with many of the same properties as conventional laser oscillators. The solution can be generalized to the case where γ has an imaginary component due to an applied or photovoltaic field E_0 . Linear absorption can also be included.

Materials

Photorefractive materials may be separated into two broad classes: inorganic and organic. The first materials investigated were inorganic oxides and semiconductors. Their success led to efforts to endow more easily produced organic materials with the required photoconductivity, charge trapping centers, and electro-optic coefficients.

Inorganic Materials

The first requirement in a classic photorefractive material is a linear electro-optic coefficient, such as appears in sillenites such as bismuth silicon oxide $\text{Bi}_{12}\text{SiO}_{20}$ and bismuth germanium oxide $\text{Bi}_{12}\text{GeO}_{20}$. These were some of the first photorefractive materials used for image processing and phase conjugation applications. However, their electro-optic coefficients (a few pm/V) are not large enough to easily give rise to large diffraction efficiencies, or to photorefractive oscillators. Materials at a temperature near a phase transition generally have higher electro-optic coefficients because their crystal structures are on the verge of changing. They are extremely susceptible to the effect of any external influence such as the application of electric fields including photorefractive space charge fields. That is why ferroelectric materials are good candidates for photorefractive materials. These include barium titanate BaTiO_3 , potassium niobate KNbO_3 , and strontium barium niobate $\text{Sr}_x\text{Ba}_{1-x}\text{Nb}_2\text{O}_6$. The mean free path of charge transport is less than that in sillenites, so they require more photons to reach a steady state. Therefore, ferroelectric materials are typically less sensitive than sillenites.

The second requirement is for photoconductivity. This requires the existence of photoexcitable charge carriers, either from valence band to conduction band or from intraband trapping centers. The latter source of photocarriers is used most often because the absorption length of light whose energy is greater than the bandgap is usually only a few micrometers. This places a severe restriction on the beams' interaction length ℓ . Thus there has been considerable research on suitable dopants, most commonly Fe^{2+} and Fe^{3+} . These dopants also act as trapping centers.

Lithium niobate is an example of a material with a large photovoltaic effect. When illuminated, this crystal develops a large dc field, which acts to enhance the grating strength. In those cases where the photorefractive effect is not wanted, such as in the design of lithium niobate waveguide devices, the photovoltaic effect can be greatly diminished by the addition of MgO to the crystal melt during growth.

For photorefractivity in the near infrared, semiconductors such as gallium arsenide and indium phosphide have been used with success. These materials also have the advantage that they can be grown in layered structures to produce, for example, multiple quantum wells that can be used to tailor the characteristics of optical excitation and charge transport.

Organic Materials

First-generation photorefractive polymers were designed to mimic the properties of their inorganic counterparts. Owing to their rich structural flexibility, organic synthetic materials with suitable charge transport, trapping, linear electro-optic effects and low optical absorption, were developed. This approach did build on the know-how developed previously in making photoconducting polymers for xerography and electro-optic polymers for optical modulation. Current polymers are based on an orientational photorefractive effect that leads to higher refractive index changes compared with traditional photorefractive materials. In materials with orientational photorefractivity, the refractive index change is produced by the field induced reorientation of anisotropic conjugated molecules that possess a permanent dipole moment and that have a high polarizability anisotropy. The photorefractive space charge field together with the applied field will periodically reorient these molecules and produce a periodic refractive index modulation through an orientational Kerr effect. The build-up and dynamics of this space-charge field are similar to those of traditional photorefractive crystals and can be described by Eqs. (2)–(7). The time constant of the hologram is a convolution of the build-up time of the space-charge and the orientational diffusion time of the dipolar molecules in the total electric field. In contrast to crystals, polymers are nearly amorphous and the transport properties are described by charge hopping processes instead of band-type transport in crystals. Consequently, the photoconducting properties of polymers are strongly field dependent and photorefractivity is mainly observed under strong applied voltages. Numerous polymer composites have been developed using the hole transport polymer poly(N-vinylcarbazole). Several materials with a refractive index modulation amplitude of the order of a percent and two-beam coupling constants $\gamma > 100 \text{ cm}^{-1}$ have been reported. New polymer composites are continuously being tested. Photorefractivity is also studied in other organic materials including organic crystals, liquid crystals, nanocomposites such as polymer-dispersed liquid crystals, or hybrid materials such as sol-gels.

Applications

Holographic Data Storage

Holographic data storage takes advantage of the Bragg selectivity of thick gratings. This allows many holograms to be superimposed in the same small volume, typically of the order of one cubic centimeter. A page of data is displayed on a spatial light modulator and a laser beam passing through the modulator is holographically recorded in the crystal with a reference beam at a specific angle. Many pages of data can be superimposed by recording many holograms with angularly multiplexed reference beams. Other multiplexing schemes are implemented by changing the shape of the wavefront of the reference beam. In principle, the upper limit of recording density is determined by the wavelength of light: one bit per cubic wavelength. If the recording wavelength is $0.5 \text{ }\mu\text{m}$, then one cubic centimeter can contain 1000 gigabytes of data. In practical circumstances, when noise is taken into consideration, the capacity is more realistically of the order of one gigabyte if a 1000×1000 spatial light modulator is used.

Distortion Compensation by Phase Conjugation

Photorefractive materials can be used to make high-reflectivity phase conjugate mirrors. The phase conjugate of a laser beam is produced when a hologram of the beam is read by another beam traveling in the opposite direction to the original reference beam. The phase conjugate reconstruction is a time-reversed copy of the original beam. If the original beam has passed through distorting optics, then the phase conjugate beam will retrace the path of the original beam through the distortion and emerge in its undistorted original state. In the standard nomenclature of phase conjugation, the input beam is called the signal, or probe, and the two reference beams are called the pumps. The output beam at $z=0$ is called the phase conjugate beam and has zero amplitude at its input at $z=\ell$, where ℓ is the interaction length. Applications for phase conjugation exist, for example, in signal transmission through distortions and laser cavity design. If a phase conjugate mirror is used as a cavity mirror, then the effects of intracavity distortions are removed.

Since the photorefractive gratings can be very strong, the diffraction efficiency of the counterpropagating reference beam can be so high that the phase conjugate reflectivity can exceed unity. The simplest generalization of Eq. (17) to the four-wave mixing phase conjugation case is when only transmission gratings are important, as occurs in many circumstances depending on the mutual

coherence properties of the beams, and the material's properties. The coupled wave equations are

$$\begin{aligned}\frac{dA_1}{dz} &= \gamma \frac{(A_1A_4^* + A_2^*A_3)A_4}{I_0} \\ \frac{dA_2^*}{dz} &= \gamma \frac{(A_1A_4^* + A_2^*A_3)A_3^*}{I_0} \\ \frac{dA_3}{dz} &= -\gamma \frac{(A_1A_4^* + A_2^*A_3)A_2}{I_0} \\ \frac{dA_4^*}{dz} &= -\gamma \frac{(A_1A_4^* + A_2^*A_3)A_1^*}{I_0}\end{aligned}\quad (21)$$

In the undepleted pumps approximation, $(I_1, I_2 \gg I_3, I_4)$, the equations become linear and the solution for phase conjugate reflectivity $R = I_3(0)/I_4(0)$ is

$$R = \frac{\sinh[\gamma\ell/2]}{\cosh[(\gamma\ell + \ln r)/2]^2} \quad (22)$$

where $r = I_2/I_1$ is the ratio between the intensities of the pumping beams.

The fact that the reflectivity of the phase conjugate mirror can be greater than unity means that we can build an optical oscillator bounded by a regular mirror and a phase conjugate mirror only. Not only does it not require any additional optical gain, but it also compensates for intracavity distortions. The regular mirror can have any shape, provided that it is sufficiently reflective.

Self-pumped Phase Conjugate Mirrors

If a laser beam passes through a crystal placed inside an optical cavity formed by two facing mirrors, light scattered by imperfections in the crystal can be amplified through the photorefractive effect. The cavity provides feedback and optical oscillation can build up. The oscillation beams pump the crystal as a phase conjugate mirror for the incident beam, thus forming a self-pumped phase conjugate mirror. The feedback can even be provided by total internal reflection in the crystal in which case the crystal by itself can become a phase conjugate mirror.

Pattern Recognition and Image Filtering

Photorefractive wave mixing can be used to perform pattern recognition by matched filtering. One example would be to identify a tank in a battlefield scene, another would be to identify all of the occurrences of a particular word on a page of text. Suppose the input beams contain the corresponding pictorial information, such as might be obtained by passing the beams through an image-bearing transparency or spatial light modulator. Eq. (21) shows that the source for beam 3 contains a term proportional to the product of the amplitudes of the three input beams. If lenses are placed in the input beams so that the crystal receives the Fourier transforms of those beams, then the output beam at the crystal, beam 3, will be proportional to the product of the Fourier transforms of the input beams 1, 2, and 4. A lens may then be used to perform the inverse Fourier transform of the product of the Fourier transforms of the input beams, producing an output proportional to beam 1 convolved with beam 2 correlated with beam 4. If beam 1 is a point source before its Fourier transforming lens, then it will be a plane wave at the crystal. Beam 3 after inverse Fourier transformation by its lens will be the correlation of beams 4 and 2. For example, suppose we want to find all the occurrences of a particular word, say 'optics' in a given page of text. Then we would make a transparency of the word 'optics' and place that in input beam 4. We would then place an image of the page of text in beam 2. Beam 3 would then contain a field with bright spots at the places containing the word 'optics' in the original text.

The real-time holographic recording properties of photorefractive materials can also be exploited in medical imaging applications by performing time-gated holography. In this method, a hologram is formed by the temporal overlap in the photorefractive sample of a reference pulse and the first-arriving (ballistic photons) light from a stretched image-bearing pulse that has propagated through a scattering medium. The filtering of the useful photons from the scattered ones is achieved by reconstructing the hologram formed with the ballistic photons in a four-wave mixing geometry.

Optical Limiting, the Novelty Filter, and Laser Ultrasonic Inspection

The attenuation of beam 2 in Eq. (17) can be used in several applications including optical limiting and novelty filtering. If one wants to protect a sensor from high-intensity laser radiation, then one could split a small portion of the input beam directed at the sensor and use it as beam 1 in a photorefractive recording setup with the input beam acting as beam 2. If the laser intensity is above the equivalent dark intensity such that the optically excited charge density is greater than the thermally excited charge density, the photorefractive effect will be activated and the input beam will be de-amplified by destructive interference with beam 1, thus protecting the sensor. In materials with high gain-length products ($\gamma\ell > 1$), separate provision of beam 1 is unnecessary

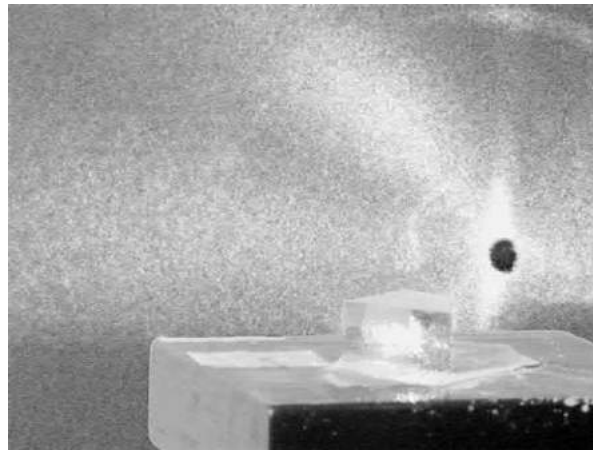


Fig. 3 Photorefractive barium titanate exhibiting amplified scattering. A helium neon laser beam is incident from the lower left, passes through the crystal to a screen where it is blacked out to prevent saturation of the camera. The screen shows brushes of amplified scattering, or fanning.

because light scattered from the input beam by imperfections in the material and other scattering centers will be greatly amplified, often to such an extent that the input beam is almost completely attenuated. This photorefractive amplification of scattered radiation is known as the fanning effect, because the amplified scattered light often appears as a fan, or brush of light, as can be seen in Fig. 3. The effect can also be used to make a novelty filter, which transmits only the moving portion of a scene. The crystal is only fast enough to respond to the slowly changing or static components of an image-bearing beam. Since the grating formed is Bragg-matched only to the slow component, the grating will only attenuate that slow component. Any rapidly changing parts of a scene pass through the crystal unattenuated. Such a filter is useful for picking out moving objects in a static cluttered background, for example a tank on a battlefield, or a micro-organism swimming against a stationary background.

A related application is in laser ultrasonic inspection. Defects in industrial material processing such as welding have characteristic ultrasonic signatures. The part under test is pinged by a pulsed laser and a probe laser is reflected from the part. As the part is shaken by the ultrasound, speckle in the reflected beam vibrates. A photorefractive recording is made of the speckle beam. Electrodes are placed on the photorefractive material, so that any optically induced currents can be detected. If the speckle pattern is not moving, or is moving more slowly than the speed of response of the photorefractive material, the photorefractive grating will be essentially at steady state: the drift currents balance the diffusion currents so there is no net current. There is no signal as the part moves through the process line. However, if the speckle beam is moving faster than the response time of the material, it will move photoexcited charge back and forth past the quasistatic grating and generate a net current for detection via the electrodes.

Adaptive Signal Processing

The relatively slow speed of photorefractive devices can be used to advantage in radio-frequency (RF) signal processing, such as signal extraction and coherent combination of signals from antenna arrays. The signal extraction application depends on grating competition in two-beam coupling. Suppose we wish to separate signals on two different RF carrier frequencies ω_1 and ω_2 , respectively. The combined signal is applied to an optical carrier beam using a high-speed modulator. The resulting optical field may be represented as

$$S_1(t)\exp(i\omega_1 t) + S_2(t)\exp(i\omega_2 t) \quad (23)$$

It is used to pump a unidirectional ring resonator so that grating competition allows oscillation only on the strongest component of the combined signal, say S_1 . The output of the oscillator is proportional to the extracted component S_1 . The effectiveness of the intersignal competition is enhanced by placing another photorefractive material in the cavity. A portion of the intracavity beam is picked off by a beamsplitter and used as a two-beam coupling pump in the second material. The crystal is oriented so that the photorefractive grating diffracts the picked-off beam back into the cavity. The return of the picked-off component is most effective for the stronger component S_1 thus decreasing its loss compared to that of the weaker component S_2 . This beamsplitter/crystal combination is known as a reflexive coupler.

Photorefractive Solitons

Under favorable conditions, a single beam incident on a photorefractive crystal will induce a positive refractive index change at the center of the beam. This provides a self-focussing tendency that counteracts the beam's divergence due to diffraction. When these two effects balance each other, the beam can propagate with a constant diameter. Such a beam is known as a spatial soliton in

analogy with temporal solitons in optical fibers and can occur when there is a component of the photorefractive response due to drift. The drift component of the photorefractive effect appears when a dc field E_0 is applied to the material. Potential applications are optically written waveguides and couplers.

See also: Nonlinear Optical Phase Conjugation

Further Reading

- Coufal, H.J., Psaltis, D., Sincerbos, G.T. (Eds.), 2000. Holographic Data Storage. Berlin: Springer.
- Gunter, P., Huignard, J.-P. (Eds.), 1988. Photorefractive Materials and their Applications I. Berlin: Springer-Verlag.
- Gunter, P., Huignard, J.-P. (Eds.), 1989. Photorefractive Materials and their Applications II. Berlin: Springer-Verlag.
- Nalwa, H.S., Miyata, S. (Eds.), 1997. Nonlinear Optics of Organic Molecules and Polymers. Boca Raton: CRC Press.
- Nolte, D.D. (Ed.), 1995. Photorefractive Effects and Materials. Kluwer.
- Pepper, D.M., Feinberg, J., Kukhtarev, N.V., 1990. The photorefractive effect. *Scientific American* 263, 62.
- Solymar, L., Webb, D.J., Grunnet-Jepsen, A., 1996. The Physics and Application of Photorefractive Materials. Oxford: Clarendon Press.
- Yeh, P., 1993. Introduction To Photorefractive Nonlinear Optics. New York: Wiley.
- Yu, F., Yin, S. (Eds.), 2000. Photorefractive Optics: Materials, Properties, and Applications. San Diego: Academic Press.

Ultrafast and Intense-Field Nonlinear Optics

AL Gaeta, Cornell University, Ithaca, NY, USA

RW Boyd, University of Rochester, Rochester, NY, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The tremendous development of high-powered femtosecond laser systems since the 1980s, has opened up new areas of research in nonlinear optics, plasma physics, atomic and molecular dynamics, and intense-field physics. For many of these applications, it is important to understand how ultrashort light pulses propagate through a medium under conditions in one or more of the processes in which nonlinear optics play an important role.

The starting point for the modeling of light propagation under these conditions is Maxwell's wave equation for the electric field $E(\mathbf{r}, t)$ which is given in Gaussian units as

$$\nabla^2 E - \frac{\partial^2 E}{\partial t^2} = 4\pi \frac{\partial^2 P}{\partial t^2} \quad (1)$$

where the $P(\mathbf{r}, t)$ is the polarization inside the medium. Typically, the polarization is separated into a term P_l that depends linearly on the field E and into a term P_{nl} that depends nonlinearly on the applied field. The Fourier transform of the linear polarization can be expressed as $\tilde{P}_l(\mathbf{r}, \omega) = \chi^{(1)}(\omega) \tilde{E}(\mathbf{r}, \omega)$, where $\tilde{E}(\mathbf{r}, \omega)$ is the Fourier transform of electric field $E(\mathbf{r}, t)$. For the laser-matter interactions that we consider here, we assume that the linear susceptibility $\chi^{(1)}(\omega)$ is real, in which case the wave equation can be expressed as

$$\nabla^2 \tilde{E} + \left[\frac{n_l(\omega)\omega}{c} \right]^2 \tilde{E} = -\frac{4\pi\omega^2}{c^2} \tilde{P}_{nl} \quad (2)$$

where $n_l(\omega) = \sqrt{1 + 4\pi\chi^{(1)}(\omega)}$ is the frequency-dependent linear refractive index of the medium.

For light pulses that are longer than an optical period, the envelope description is valid and the electric field can be described by an amplitude envelope $A(\mathbf{r}, t)$ and a carrier frequency ω_0 such that

$$E(\mathbf{r}, t) = A(\mathbf{r}, t) e^{i(k_0 z - \omega_0 t)} + c.c. \quad (3)$$

where $k_0 = k(\omega_0) = n_0 \omega_0 / c$ is the wavevector amplitude and $n_0 = n_l(\omega_0)$. This envelope description is advantageous for performing analytical and numerical studies of pulse propagation. However, for sufficiently short laser pulses, where the shape of the envelope function does not depend on the carrier phase of the carrier wave, such a description is no longer applicable. Nevertheless the envelope description can be made valid for pulses that are nearly as short as a single cycle or, alternatively, that have spectral bandwidths that are comparable to the central frequency ω_0 . To derive an equation for the spatio-temporal evolution of the pulse envelope, the relation for the electric field is substituted into the wave equation (Eq. (2)) and the following two approximations are made: i) $k_0 \partial A / \partial z \ll \partial^2 A / \partial z^2$, which signifies that the envelope varies slowly in space over a distance compared to the central wavelength; ii) $v_{ph} \sim v_{gr}$ where $v_{ph} = c/n_0$ is the phase velocity and $v_{gr} = (dk/d\omega)^{-1}$ is the group velocity. This latter approximation is invariably well satisfied when the frequencies contained in the laser field are highly nonresonant with any transition frequencies of the medium.

For an input pulse at $z=0$ with a peak amplitude A_0 , and characteristic widths in space and time given by w_0 and τ_p , respectively, the equation for the normalized amplitude $u(r_\perp, z, t) = A(r_\perp, z, t) / A_0$ can be expressed as

$$\frac{\partial u}{\partial \xi} = \frac{i}{4} \left(1 + \frac{i}{\omega_0 \tau_p} \frac{\partial}{\partial \tau} \right)^{-1} \nabla_\perp^2 u - i \sum_{n=2} \frac{L_{ds}}{n! L_{ds}^{(n)}} \frac{\partial^n u}{\partial \tau^n} + i \left(1 + \frac{i}{\omega_0 \tau_p} \frac{\partial}{\partial \tau} \right) p^{nl} \quad (4)$$

where $\tau = (t - z/v_g) / \tau_p$ is the normalized retarded time for the pulse traveling at the group velocity v_g , $L_{ds}^{(n)} = \tau_p^n / \beta_n$ is the n th-order dispersion length, $\beta_n (n \geq 2)$ is the n th-order dispersion constant, $L_{ds} = |L_{ds}^{(2)}|$ is the dispersion length, $\zeta = z / L_{df}$ is the normalized distance, $L_{df} = k w_0^2 / 2$ is the diffraction length, p_{nl} is the normalized nonlinear polarization, and $\nabla_\perp^2$ is the transverse Laplacian. The presence of the operator $(1 + i\partial/\omega_0 \partial\tau)$ in the diffraction term ($\nabla_\perp^2 u$) of Eq. (4) leads to space-time focusing, while its presence in the nonlinear term results in self-steepening and both these terms arise from not making the slowly varying envelope approximation in time (i.e., $k_0 \partial A / \partial t \ll \partial^2 A / \partial t^2$) in deriving Eq. (4). For a self-consistent analysis of pulse propagation in the nonlinear regime, both the effects of space-time focusing and self-steepening must be included.

Nonlinear Refractive Index

For an isotropic medium in the highly nonresonant limit, the third-order term is the lowest-order contribution to the nonlinear susceptibility. This term gives rise to the nonlinear refractive index, that is, the index of refraction depends on the local intensity of

the laser field. For many materials there are two contributions to the nonlinear refractive index: i) an instantaneous part that arises from the electronic response of the medium; and ii) a noninstantaneous contribution due to the nuclear motion of the molecules (i.e., the Raman contribution). For such a medium, the nonlinear polarization may be expressed as

$$p_{\text{nl}}(\zeta, \tau) = \frac{L_{\text{ds}}}{L_{\text{nl}}} \left[(1-f)|u(\zeta, \tau)|^2 + f\gamma_R \int_{-\infty}^{\tau} d\tau' R(\tau - \tau') |u(\zeta, \tau')|^2 \right] u(\zeta, \tau) \quad (5)$$

where $L_{\text{nl}} = (c/\omega_0 n_2 I_0)$ is the nonlinear length, $I_0 = n_0 c |A_0|^2 / 2\pi$ is the peak input intensity, f is the fraction of the Raman contribution to the nonlinear refractive index, and $R(\tau) = \{ [1 + (\Omega_R \tau_R)^2] / \Omega_R \tau_R \} \exp(-\tau/\tau_R) \sin(\Omega_R \tau)$ is the Raman response function, τ_R is the characteristic Raman response time, Ω_R is the characteristic Raman frequency, and $\gamma_R = \tau_p / \tau_R$. For example, for fused silica $f=0.15$, $\tau_R = 50$ fs, and $\Omega_R \tau_R = 4.2$. For a noble gas such as Xe, there is no Raman contribution and $f=0$.

Self Focusing, Supercontinuum Generation, and Filamentation

The presence of the nonlinear refractive index with $n_2 > 0$ can lead to self-focusing of a laser beam. For sufficiently long pulses such that dispersion, self-steepening, and space-time focusing effects can be neglected, it is found that laser beams with input powers greater than the critical power $P_{\text{cr}} = \alpha \lambda^2 / 4\pi n_0 n_2$ will undergo catastrophic self-focusing collapse. The dimensionless parameter $\alpha \geq 1.86$ depends on the spatial profile of the input beam and for a Gaussian input beam is given by $\alpha = 1.9$, in which case the ratio of the input power P to the critical power satisfies the relation $P/P_{\text{cr}} = 1.055 L_{\text{df}} / L_{\text{nl}}$. Extensive studies have been made on the dynamics of laser beams undergoing self-focusing. For example, it has been shown that the shape of the collapsing beam evolves to a radially symmetric profile as it approaches the collapse point and that the total power contained in the collapsing portion always corresponds to the minimum value (i.e., $\alpha \sim 1.86$) regardless of the initial power in the beam.

For light pulses shorter than a picosecond, the role of material dispersion plays an important role and completely alters the dynamics of the self-focusing process. These dispersive effects lead to a temporal splitting of the pulse into two pulses and the arrest of its collapse. At even higher powers, other phenomena can occur, such as 'optical shock' formation at the rear edge of the pulse, due to self-steepening and space-time focusing. Shock formation leads to the emission of a broad spectrum of radiation extending from the ultraviolet to the mid-infrared, known as supercontinuum generation (SCG). This phenomenon was first observed in 1970, and since then it has been observed in many different solids, liquids, and gases, under a wide variety of experimental conditions.

If the peak intensity of the pulse approaches intensities of $< 10^{13}$ W/cm², either through self-focusing or external focusing, multiphoton ionization occurs and a plasma is formed in the medium. (See section below on Plasma Nonlinearities and Relativistic Effects.) The generation of the plasma lowers the refractive index and effectively counteracts the self-focusing process, resulting in the spatial confinement of the pulse for distances far beyond what would be allowed by ordinary linear diffraction and has been observed in gases, liquids, and solids. A striking example of this apparent self-waveguiding is the observation of 'light strings' in air which can extend more than 10 km into the atmosphere. This phenomenon was first observed with 100 fs laser pulses in the near infrared ($\lambda = 800$ nm). Researchers found that pulses with energies greater than 10 mJ undergo self-focusing collapse in air and produce a highly intense ($> 10^{13}$ W/cm²) 100-micron-diameter light filament tens of meters long.

Multiphoton Absorption

Multiphoton absorption is a process in which an atom or molecule makes a transition from a ground state to an excited state by means of the simultaneous absorption of N photons. In the lowest order of perturbation theory, such a process can be described by means of a susceptibility of order $(2N-1)$, that is, by $\chi^{(2N-1)}$. Alternatively, this process can be described in terms of an N -photon cross-section $\sigma^{(N)}$ defined such that the transition rate per atom is given by

$$R^{(N)} = \sigma^{(N)} I^N \quad (6)$$

where I is the intensity of the laser field. Quantum mechanical expressions for the N -photon cross-sections are readily obtained. One finds, for instance, that

$$R_{\text{ng}}^{(2)} = \left| \sum_m \frac{\mu_{\text{nm}} \mu_{\text{mg}} E^2}{\hbar^2 (\omega_{\text{mg}} - \omega)} \right|^2 2\pi \rho_f (\omega_{\text{ng}} - 2\omega)$$

$$R_{\text{og}}^{(3)} = \left| \sum_{mn} \frac{\mu_{\text{on}} \mu_{\text{nm}} \mu_{\text{mg}} E^3}{\hbar^3 (\omega_{\text{ng}} - 2\omega) (\omega_{\text{mg}} - \omega)} \right|^2 2\pi \rho_f (\omega_{\text{og}} - 3\omega) \quad (7)$$

In each of these expressions, the quantity ρ_f represents the density of final states, or equivalently the atomic lineshape function, evaluated at the N -photon transition frequency.

Optical Damage

High-intensity laser fields can produce unwanted damage to optical materials. As a point of reference, the threshold for laser damage to fused silica at a wavelength of 1.05 micrometers for a pulse of 30 ps duration corresponds to an intensity of 230 GW/cm² or a fluence of 7 J/cm². Over a wide range of pulselengths (approximately 1 ps to 1 μs), the threshold intensity for laser damage decreases with pulse length T as $T^{-1/2}$ and correspondingly the threshold fluence for laser damage increases with pulse length $T^{1/2}$. In this range of pulse durations, the dominant mechanism of laser damage is avalanche breakdown. In this process, free electrons are accelerated by the laser field until they acquire sufficient energy to impact-ionize other atoms in the sample. These additional electrons are similarly accelerated and create still more free electrons. The combined action of the breaking of chemical bonds and the deposition of heat energy leads to the fracturing of the optical material. For pulses shorter than 1 ps, processes such as multiphoton absorption and multiphoton dissociation contribute to the mechanism of optical damage. For laser pulses longer than approximately 1 μs (including continuous wave laser beams), the dominant damage mechanism is direct heating of the optical material by linear absorption.

High-Harmonic Generation

Let us consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects are traditionally modeled using a power-series expansion, such as

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \dots \quad (8)$$

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{at} = e/a_0^2 = 2 \times 10^7$ statvolt/cm = 6×10^9 V/cm. This field strength corresponds to a laser intensity of $I_{at} = 4 \times 10^{16}$ W/cm², which constitutes the threshold intensity for exciting nonperturbative nonlinear optical response.

One of the consequences of excitation with intensities comparable to the atomic unit of intensity I_{at} , is the occurrence of high-harmonic generation. In a typical experimental arrangement, a gas jet is irradiated by high-intensity laser radiation, and all odd harmonics of the fundamental laser frequency, up to some maximum value N_{max} , are observed. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with N_{max} as large as 341.

The phenomenon of high-harmonic generation can be understood in terms of a simple physical model. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse, and each time the electron passes near the nucleus it undergoes strong acceleration and emits a short pulse of radiation. This radiation will occur in the form of a train of short pulses; the spectrum of the radiation will be the square of the Fourier transform of this pulse train, which will contain the odd harmonics of the oscillation frequency up to some maximum frequency, that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made quantitative to show that the maximum harmonic number is given by

$$N_{max}\hbar\omega = 3.17K + U_p \quad (9)$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ionization energy of the atom.

Plasma Nonlinearities and Relativistic Effects

The process of multiphoton ionization can liberate a sufficient number of free electrons to transform the optical medium into a plasma, that is, a fully or partially ionized gas. The process of plasma formation is described by the equation

$$\frac{dN_e}{dt} = \frac{dN_i}{dt} = (N_T - N_i)\sigma^{(N)}I^N - rN_eN_i \quad (10)$$

where N_e is the number density of electrons, N_i is the number density of ions, N_T is the total number of atoms (both ionized and un-ionized) in the material, and r is the electron-ion recombination coefficient. The optical properties of plasmas are very different from those of typical dielectric materials; the plasma contribution to the dielectric constant is given by

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \quad (11)$$

where $\omega_p = \sqrt{4\pi N e^2/m}$ is known as the plasma frequency.

Nonlinear effects can occur in the propagation of light through a plasma. One example is the nonlinear response resulting from the relativistic change in mass of the electron due to the large velocity that it can attain in the field of an intense

laser beam. Detailed consideration of this effect shows that the nonlinear change in refractive index can be described as $\Delta n = n_2 I$ where

$$n_2 = \frac{2\pi\omega_p^2 e^2}{n_0^2 m^2 c^3 \omega^4} = \frac{1}{2\pi n_0^2} \left(\frac{\omega_p}{\omega}\right)^2 1.1 \times 10^{-26} \frac{\text{cm}^2}{\text{W}} \quad (12)$$

Nonlinear Quantum Electrodynamics

One can imagine an electric field so intense that it could lead to the spontaneous creation of electron–positron pairs. Such a field would have a strength of the order of $E_{\text{QED}} = mc^2/e\lambda_c$ where $\lambda_c = \hbar/mc$ is the reduced Compton wavelength of the electron. The intensity of a light beam with a peak field amplitude of E_{QED} is $I_{\text{QED}} = 4 \times 10^{29} \text{ W/cm}^2$. Detailed consideration shows that even for fields weaker than I_{QED} there will be a field-induced change in the dielectric tensor given by

$$\varepsilon_{ik} = \delta_{ik} + \frac{e^4 \hbar}{45\pi m^4 c^7} [2(E^2 - B^2)\delta_{ik} + 7B_i B_k] \quad (13)$$

Because of the unusual tensor properties of this relation, it displays a different polarization dependence than typical optical nonlinearities. Nonetheless, to an order of magnitude one can describe the strength of this response as

$$n_2 = \frac{7}{15c} \frac{e^2}{\hbar c} \frac{1}{E_{\text{QED}}^2} = 5.6 \times 10^{-34} \text{ cm}^2/\text{W} \quad (14)$$

See also: Attosecond Spectroscopy

Further Reading

- Agostini, P., Fabre, F., Mainfray, G., Petite, G., Rahman, N.K., 1979. Free-free transitions following six photon ionization of xenon. *Physics Review Letters* 42, 1127.
- Alfano, R.R. (Ed.), 1989. *The Supercontinuum Laser Source*. New York: Springer-Verlag.
- Bloembergen, N., 1997. A brief history of light breakdown. *Journal of Nonlinear Optical Physics and Materials* 6, 377.
- Boyd, R.W., 2003. *Nonlinear Optics*, Second Edition. Section 12.5. Amsterdam: Academic Press.
- Braun, A., Korn, G., Liu, X., Du, D., Squier, J., Mourou, G., 1995. Self-channeling of high-peak-power femtosecond laser pulses in air. *Optical Letters* 20, 73.
- Brabec, T., Krausz, F., 2000. Intense few-cycle laser fields: Frontiers of nonlinear optics. *Reviews of Modern Physics* 72, 545.
- Chang, Z., Rundquist, A., Wang, H., Murnane, M.M., Kapteyn, H.C., 1997. Generation of coherent soft X-rays at 2.7nm using high harmonics. *Physics Review Letters* 79, 2967.
- Corkum, P.B., 1993. Plasma perspective on strong field multiphoton ionization. *Physics Review Letters* 71, 1994.
- Corkum, P.B., Rolland, C., Srinivasanrao, T., 1986. Supercontinuum generation in gases. *Physics Review Letters* 57, 2268.
- Gaeta, A.L., 2003. Collapsing light really shines. *Science* 301, 54.
- Kasparian, J., Rodriguez, M., Mejean, G., Yu, J., Salmon, E., Wille, H., Bourayou, R., Frey, S., André, Y.-B., Mysrowicz, A., Sauerbrey, R., Wolf, J.-P., Woste, L., 2003. White light filaments for atmospheric analysis. *Science* 301, 61.
- Lewenstein, M., Balcou, P., Ivanov, M.Y., L'Huillier, A., Corkum, P.B., 1994. Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49, 2117.
- Max, C.E., Arons, J., Langdon, A.B., 1974. Self-modulation and self-focusing of electromagnetic waves in plasmas. *Physics Review Letters* 33, 209.
- Moll, K.D., Fibich, G., Gaeta, A.L., 2003. Self-similar wave collapse: observation of the Townes profile. *Physics Review Letters* 90, 203902.
- Monot, P., Auguste, T., Gibbon, P., Jakober, F., Mainfray, G., Dulieu, A., Louis-Jacquet, M., Malka, G., Miquel, J.L., 1995. Experimental demonstration of relativistic self-channeling of a multiterawatt laser pulse in an underdense plasma. *Physics Review Letters* 74, 2953.
- Rairoux, P., Schillinger, H., Niedermeier, S., *et al.*, 2000. *Applied Physics B—Lasers Optics* 71, 573.
- Ranka, J.K., Schirmer, R.W., Gaeta, A.L., 1996. Observation of pulse splitting in nonlinear dispersive media. *Physics Review Letters* 77, 3783.
- Rothenberg, J.E., 1992. Pulse splitting during self-focusing in normally dispersive media. *Optical Letters* 17, 583.
- Sprangle, P., Tang, C.-M., Esarey, E., 1987. Relativistic self-focusing of short-pulse radiation beams in plasmas. *IEEE Transactions on Plasma Science* 15, 145.
- Stuart, B.C., Feit, D., Rubenchik, A.M., Shore, B.W., Perry, M.D., 1995. Laser-induced damage in dielectrics with nanosecond to subpicosecond pulses. *Physics Review Letters* 74, 2248.
- Wagner, R., Chen, S.-Y., Maksemchak, A., Umstadter, D., 1997. Electron acceleration by a laser wakefield in a relativistically self guided channel. *Physics Review Letters* 78, 3125.
- Wood, R.M., 1986. *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

Electromagnetically Induced Transparency

JP Marangos, Imperial College of Science, Technology and Medicine, London, UK

© 2005 Elsevier Ltd. All rights reserved.

Introduction

In this article we examine how the process of electromagnetically induced transparency (EIT) can be used to increase the efficiency of non-linear optical frequency mixing schemes. We illustrate the basic ideas by concentrating upon the enhancement of resonant four-wave mixing schemes in atomic gases. To start we introduce the quantum interference phenomenon that gives rise to EIT. The calculation of EIT effects in a three-level atomic medium using a density matrix approach is presented. Next we examine the modified susceptibilities that result and that can be used to describe nonlinear mixing, and show how large enhancements in nonlinear conversion efficiency can result. Specific examples are briefly discussed along with a further discussion of the novel aspects of refractive index and pulse propagation modifications that result from EIT. The potential benefits to nonlinear optics of electromagnetically induced transparency are discussed. In an article of this nature it is not possible to credit all of the important contributions to this field, but we include a bibliography for further reading indicating some of the seminal contributions.

Electromagnetically induced transparency is the term applied to the process by which an otherwise opaque medium can be rendered transparent through laser-induced quantum mechanical interference processes. The linear susceptibility is substantially modified by EIT. Absorption is cancelled due to destructive interference between excitation pathways. The dispersion of the medium is likewise modified such that where there is no absorption the refractive index is unity and there can be a large value of normal dispersion leading to low values of group velocity. In contrast to the linear susceptibility the nonlinear susceptibility involving these same resonant laser fields will undergo constructive rather than destructive interference. This leads to a strong enhancement in nonlinear frequency mixing. For a normally transparent medium the nonlinear optical couplings will be small unless large-amplitude fields are applied. The most important consequence of EIT, in contrast to the usual case when a medium is transparent, is that the dispersion is not vanishing and the nonlinear couplings can be very large.

To understand how this comes about we must consider the system illustrated in Fig. 1(a) where because of the resonant condition satisfied for the applied fields the atom can be simplified to just three-levels. The important parameters in this model system are that state $|2\rangle$ is metastable and has a dipole-allowed coupling to state $|3\rangle$. The coupling between states $|1\rangle$ and $|3\rangle$ is also dipole-allowed and as a consequence state $|3\rangle$ can decay radiatively to either $|1\rangle$ or $|2\rangle$. Let a coupling field be applied resonantly to the $|2\rangle - |3\rangle$ transition, which may be cw or pulsed but should in the latter case be transform-limited. We define the Rabi coupling as $\Omega_{23} = \frac{\mu_{23}E}{\hbar}$ where E is the laser electric field amplitude and μ_{23} is the dipole moment of the transition. The Rabi frequency needs to be larger than the inhomogeneous broadening of the sample. A second field, the probe, typically of much lower intensity, is then applied in resonance with the $|1\rangle - |3\rangle$ transition. If the condition $\Omega_{23} \gg \Omega_{13}$ is satisfied then it is convenient to replace the bare atomic states $|2\rangle$ and $|3\rangle$ with the dressed states (see Fig. 1(b)):

$$|a\rangle = \frac{1}{\sqrt{2}}[|2\rangle + |3\rangle] \quad (1a)$$

$$|b\rangle = \frac{1}{\sqrt{2}}[|2\rangle - |3\rangle] \quad (1b)$$

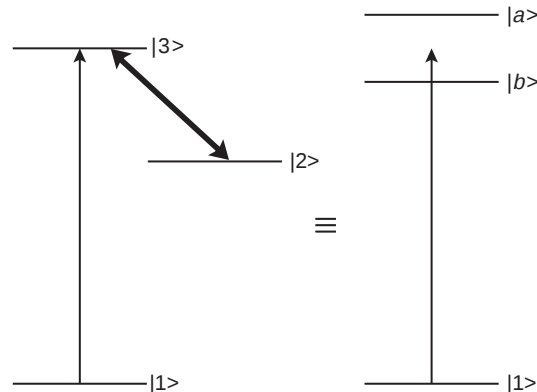


Fig. 1 A three-level atomic system coupled to laser fields in the lambda configuration is shown on the left-hand side. In the limit of a strong coupling field this is equivalent to the dressed states $|a\rangle$ and $|b\rangle$ coupled to the ground state by the probe field alone as shown on the right-hand side.

Transitions from state $|1\rangle$ induced by the probe field to this pair of near-resonant dressed states are subject to exact cancellation at resonance if $|2\rangle$ is perfectly metastable. This is because the only contribution to the excitation amplitude comes from the matrix elements $\langle 1|\mathbf{r}\cdot\mathbf{E}|3\rangle$ as the $|1\rangle - |2\rangle$ transition is dipole forbidden. The contributions from the equally detuned states $|a\rangle$ and $|b\rangle$ thus enter into the overall amplitude with opposite signs and equal magnitude as can be seen by inspection of Eqs. (1a) and (1b). This leads to cancellation of the absorption amplitude. This type of absorption cancellation is well known and closely related to the so-called *dark states*.

Theoretical Treatment of EIT in a Three-Level Medium

It was realized by several workers in the 1970s that laser-induced interference effects could lead to a cancellation of absorption at certain frequencies. To gain a more quantitative understanding of the effects of the coupling field upon the optical properties of a dense ensemble of three-level atoms we require a treatment that computes the optical susceptibility of the medium. A treatment originally carried out by Harris *et al.* for a Λ scheme similar to that illustrated in Fig. 1 was the first to derive the modified susceptibilities that will be discussed below. In that treatment the state amplitudes in the three-level atom were solved in the steady-state limit and from these the linear and nonlinear susceptibilities (see below) are then derived. In what follows we adopt a density matrix treatment as employed by various workers. This readily allows the inclusion of dephasing as well as population decay terms. The critical parameters in this system are the strengths of the fields (described in terms of the Rabi couplings), the detuning of the fields from resonance $\Delta\omega_{13} = \omega_{13} - \omega_p$ and $\Delta\omega_{23} = \omega_{23} - \omega_c$ (see Fig. 2), the radiative decays from $|3\rangle$ to $|1\rangle$ and $|2\rangle$, γ_c and γ_d and from $|2\rangle$ to $|1\rangle$ γ_a (although the latter is anticipated to be smaller). Extension to other configurations of the three states, such as a V or ladder scheme is implicit within this general treatment.

Fig. 2 illustrates the system considered. This treatment will also address the nonlinear response in a four-wave mixing scheme completed by the two-photon resonant coupling applied between state $|1\rangle$ and $|2\rangle$.

We write the interaction Hamiltonian as:

$$H = H_0 + V \quad (2)$$

where H_0 is the unperturbed Hamiltonian of the system and is written as

$$H_0 = \hbar\omega_1|1\rangle\langle 1| + \hbar\omega_2|2\rangle\langle 2| + \hbar\omega_3|3\rangle\langle 3| \quad (3)$$

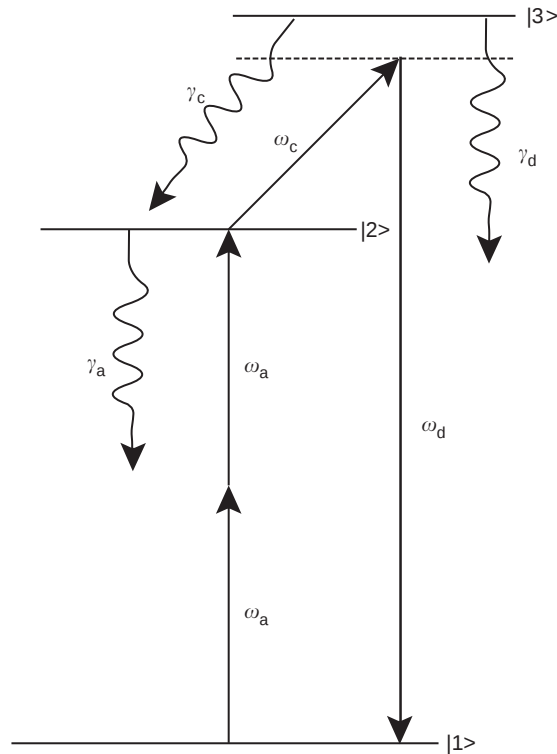


Fig. 2 A four-wave mixing scheme incorporating the three-level lambda system in which electromagnetically induced transparency has been created for the generated field ω_d by the coupling field ω_c . The decay rates γ_i from the three atomic levels are also shown. For a full explanation of this system see text.

and V is the interaction Hamiltonian and can be expressed

$$V = \hbar\Omega_a e^{-i\omega_a t} |2\rangle \langle 1| + \hbar\Omega_c e^{-i\omega_c t} |2\rangle \langle 3| + \hbar\Omega_d e^{-i\omega_d t} |3\rangle \langle 1| + c.c. \quad (4)$$

Note that the Rabi frequencies in the equation can be described as $\hbar\Omega_{ij} = \mu_{ij}|E(\omega_{ij})|$, where μ_{ij} is the dipole moment of the transition between states $|i\rangle$ and $|j\rangle$, the Rabi frequency Ω_a is a two-photon Rabi frequency that characterizes the coupling between the laser field a and the atom for this two-photon transition. We have assumed for simplicity that $\omega_a = \omega_b = \omega_{12}/2$.

Assuming all the fields lie close to the resonance, the rotating wave approximation can be applied to the interaction picture and the interaction Hamiltonian V^I is given as

$$V^I = \hbar\Omega_a e^{i\Delta_a t} |2\rangle \langle 1| + \hbar\Omega_c e^{i\Delta_c t} |2\rangle \langle 3| + \hbar\Omega_d e^{i\Delta_d t} |3\rangle \langle 1| + c.c. \quad (5)$$

where Δ_a , Δ_c and Δ_d refer the detunings of the fields and can be written as:

$$\begin{aligned} \Delta_a &= \omega_{12} - 2\omega_a \\ \Delta_c &= \omega_{32} - \omega_c \\ \Delta_d &= \omega_{13} - \omega_d \end{aligned} \quad (6)$$

For the evaluation of the density matrix with this interaction V^I , the Schrödinger equation can be restated in terms of the density matrix components. This form is called the Liouville equation and can be written as:

$$\hbar \frac{\partial}{\partial t} \rho_{ij}(t) = -i \sum_k H_{ik}(t) \rho_{kj}(t) + i \sum_k \rho_{ij}(t) H_{ik}(t) + \Gamma_{ij} \quad (7)$$

where Γ_{ij} represents phenomenologically added decay terms (i.e. spontaneous decays, collisional broadening, etc.). This formalism leads to a set of nine differential equations for nine different density matrix elements that describe the three-level system.

To remove the optical frequency oscillations, a coordinate transform is needed and to incorporate the relevant oscillatory detuning terms into the off-diagonal elements we make the substitution:

$$\begin{aligned} \tilde{\rho}_{12} &= \rho_{12} e^{-i\Delta_a t} \\ \tilde{\rho}_{23} &= \rho_{23} e^{-i\Delta_c t} \\ \tilde{\rho}_{31} &= \rho_{31} e^{-i\Delta_d t} \end{aligned} \quad (8)$$

This operation conveniently eliminates the time dependencies in the rate equations and the equations of motion for the density matrix are given by:

$$\begin{aligned} \frac{\partial}{\partial t} \rho_{11} &= i \frac{1}{2} \Omega_a \tilde{\rho}_{12} + i \frac{1}{2} \Omega_d \tilde{\rho}_{13} - i \frac{1}{2} \Omega_a^* \tilde{\rho}_{21} - i \frac{1}{2} \Omega_d^* \tilde{\rho}_{31} + \Gamma_{11} \\ \frac{\partial}{\partial t} \rho_{22} &= i \frac{1}{2} \Omega_c^* \tilde{\rho}_{23} + i \frac{1}{2} \Omega_a \tilde{\rho}_{21} - i \frac{1}{2} \Omega_c \tilde{\rho}_{32} - i \frac{1}{2} \Omega_a \tilde{\rho}_{12} + \Gamma_{22} \\ \frac{\partial}{\partial t} \rho_{33} &= i \frac{1}{2} \Omega_c^* \tilde{\rho}_{23} - i \frac{1}{2} \Omega_d^* \tilde{\rho}_{31} + i \frac{1}{2} \Omega_d \tilde{\rho}_{13} - i \frac{1}{2} \Omega_c \tilde{\rho}_{32} + \Gamma_{33} \\ \frac{\partial}{\partial t} \rho_{23} &= i \frac{1}{2} \Omega_c \tilde{\rho}_{22} - i \frac{1}{2} \Omega_c \tilde{\rho}_{33} - i \Delta_a \tilde{\rho}_{23} + i \frac{1}{2} \Omega_d \tilde{\rho}_{21} - i \frac{1}{2} \Omega_c \tilde{\rho}_{13} + \Gamma_{23} \\ \frac{\partial}{\partial t} \rho_{21} &= i \Omega_a \tilde{\rho}_{22} + \Omega_a \tilde{\rho}_{33} + i \Omega_d \tilde{\rho}_{23} - i \Delta_c \tilde{\rho}_{21} - i \frac{1}{2} \Omega_c \tilde{\rho}_{31} - i \frac{1}{2} \Omega_a \tilde{\rho}_{31} + \Gamma_{21} \\ \frac{\partial}{\partial t} \rho_{31} &= i \frac{1}{2} \Omega_d \tilde{\rho}_{22} - i \Omega_d \tilde{\rho}_{33} + i \Omega_d \tilde{\rho}_{23} - i \Delta_c \tilde{\rho}_{21} - i \frac{1}{2} \Omega_c \tilde{\rho}_{31} - i \frac{1}{2} \Omega_a \tilde{\rho}_{31} + \Gamma_{31} \end{aligned} \quad (9)$$

Using the set of equations above and the relation $\tilde{\rho}_{ij} = \tilde{\rho}_{ij}^*$ we obtain equations for $\tilde{\rho}_{12}$, $\tilde{\rho}_{32}$, and $\tilde{\rho}_{13}$. For the incoherent population relaxation the decays can be written:

$$\begin{aligned} \Gamma_{11} &= \gamma_a \rho_{22} + \gamma_d \rho_{33} \\ \Gamma_{22} &= -\gamma_a \rho_{22} + \gamma_c \rho_{33} \\ \Gamma_{33} &= -(\gamma_c + \gamma_d) \rho_{33} \end{aligned} \quad (10)$$

and for the coherence damping:

$$\begin{aligned} \Gamma_{21} &= -\left\{ \frac{1}{2} [\gamma_a + \gamma_c] + \gamma_{21}^{\text{col}} \right\} \rho_{21} \\ \Gamma_{23} &= -\left\{ \frac{1}{2} [\gamma_a + \gamma_c + \gamma_d] + \gamma_{23}^{\text{col}} \right\} \rho_{23} \\ \Gamma_{31} &= -\left\{ \frac{1}{2} [\gamma_d + \gamma_c] + \gamma_{31}^{\text{col}} \right\} \rho_{31} \end{aligned} \quad (11)$$

where γ_{ij}^{col} represents collisional dephasing terms which may be present.

This system of equations can be solved by various analytical or numerical methods to give the individual density matrix elements. Analytical solutions are possible if we can assume for instance that $\Omega_c \gg \Omega_a, \Omega_d$ and that there are continuous-wave fields so a steady-state treatment is valid. For pulsed fields or in the case where the generated field may be significant numerical solutions are in general required.

We are specifically interested in the optical response to a probe field at ω_d (close to resonance with the $|1\rangle - |3\rangle$ transition) that is governed by the magnitude of the coherence ρ_{13} . We find ρ_{13} making the assumption that only the coupling field is strong, i.e., that $\Omega_c \gg \Omega_a, \Omega_d$ holds. From this quantity the macroscopic polarization is obtained from which the susceptibility can be computed. The macroscopic polarization P at the transition frequency ω_{13} can be related to the microscopic coherence ρ_{13} via the expression:

$$P_{13} = N\mu_{13}\rho_{13} \quad (12)$$

where N is the number of equivalent atoms in the ground state within the medium, and μ_{13} is the dipole matrix element associated with the transition. In this way real and imaginary parts of the linear susceptibility χ at frequency ω can be directly related to ρ_{13} via the macroscopic polarization since the latter can be defined as:

$$P_{13}(\omega) = \epsilon_0 \chi(\omega) E \quad (13)$$

where E is the electric field amplitude at frequency ω_d . The linear susceptibility (real and imaginary parts) is given by the following expressions:

$$\text{Re}\chi_D^{(1)}(-\omega_d, \omega_d) = \frac{|\mu_{13}|^2 N}{\epsilon_0 \hbar} \left[\frac{-4\Delta_{21}(|\Omega_2|^2 - 4\Delta_{21}\Delta_{32}) + 4\Delta_{31}\gamma_a^2}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2} \right] \quad (14)$$

$$\text{Im}\chi_D^{(1)}(-\omega_d, \omega_d) = \frac{|\mu_{13}|^2 N}{\epsilon_0 \hbar} \left[\frac{8\Delta_{21}\gamma_d + 2\gamma_a(|\Omega_2|^2 + \gamma_d\gamma_a)}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2} \right] \quad (15)$$

We now consider the additional two-photon resonant field ω_a . This leads to a four-wave mixing process that generates a field at ω_d (i.e., the probe frequency). The nonlinear susceptibility that describes the coupling of the fields in a four-wave mixing process is given by the expression:

$$\chi_D^{(3)}(-\omega_d, \omega_a, \omega_a, \omega_c) = \frac{\mu_{23}\mu_{13}N}{6\epsilon_0 \hbar^3 ((\Delta_{21} - j\gamma_a/2)(\Delta_{31} - j\gamma_d/2) - (|\Omega_c|^2/4))^*} \sum_i \mu_{i1}\mu_{i3} \left(\frac{1}{\omega_i - \omega_a} + \frac{1}{\omega_i - \omega_b} \right)$$

Nonlinear Optical Processes

The dependence of the susceptibilities upon the detuning is plotted in [Fig. 3](#). In this plot the effects of inhomogeneous (Doppler) broadening are also incorporated by computing the susceptibilities over the inhomogeneous profile (see below for more details).

By inspection of [Eq. \(14\)](#) we see that the absorptive loss at the minimum of $\text{Im}[\chi^{(1)}]$ varies as γ_a/Ω_c^2 . This dependence is a consequence of the interference discussed above. In the absence of interference (i.e., just simply Autler-Townes splitting) the minimum loss would vary as γ_d/Ω_c^2 . Since $|3\rangle - |1\rangle$ is an allowed decay channel (in contrast to $|2\rangle - |1\rangle$) it follows that $\gamma_b \gg \gamma_a$ and so the absorption is much less as a consequence of EIT. $\text{Re}[\chi^{(1)}]$ in [Eq. \(15\)](#) and [Fig. 3](#) is also modified significantly. The resonant value of refractive index is equal to the vacuum value where the absorption is a minimum, the dispersion is normal in this region with a gradient determined by the strength of the coupling laser, a point we will return to shortly. For an unmodified system the refractive index is also unity at resonance but in that case there is high absorption and steep anomalous absorption. Reduced group velocities result from the steep normal dispersion that accompanies EIT. Inspection of the expression [\(16\)](#) and [Fig. 3](#) shows that $\chi^{(3)}$ is also modified by the coupling field. The nonlinear susceptibility depends upon $1/\Omega_c^2$ as is expected for a laser dressed system, however in this case there is not destructive but rather constructive interference, between the field-split components. This result is of great importance: it ensures that the absorption can be minimized at frequencies where the nonlinear absorption remains large.

As a consequence of constructive interference the nonlinear susceptibility remains resonantly enhanced whilst the linear susceptibility vanishes or becomes very small at resonance due to destructive interference. Large nonlinearity accompanying vanishing absorption (transparency) of course match conditions for efficient frequency mixing as a large atomic density can then be used. Moreover the dispersion (controlling the refractive index) also vanishes at resonance; this leads to perfect phase matching (i.e., zero wavevector mismatch between the fields) in the limit of a simple three-level system. As a result of these features a large enhancement of the conversion efficiency in this type of scheme can be achieved.

To compute the generated field strength Maxwell's equations must be solved using these expression for the susceptibility to describe the absorption, refraction, and nonlinear coupling in the medium. We will treat this within the slowly varying envelope approximation since the fields are cw or nanosecond pulses. To be specific we assume that the susceptibilities are time independent, i.e., that we are in the steady-state (cw) limit. We make the assumptions also that there is no pump field absorption and

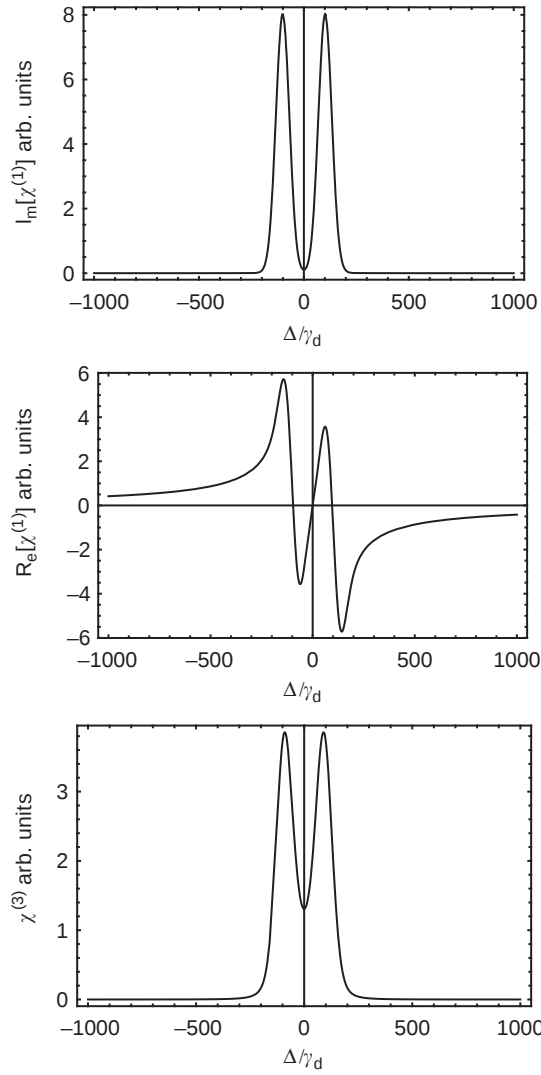


Fig. 3 Electromagnetically induced transparency is shown in the case of significant Doppler broadening. The Doppler averaged values of the imaginary $\text{Im}[\chi^{(1)}]$ and real $\text{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. The Doppler width is taken to be $20\gamma_d$, $\Omega_c = 100\gamma_d$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

that we have plane waves. Under these assumptions the generated field amplitude A_d is given in terms of the other field amplitudes A_i by:

$$\frac{\partial}{\partial z} A_d = i \frac{\omega_d}{4c} \chi^{(3)} A_a^2 A_c e^{-i\Delta k_d z} - \frac{\omega_d}{2c} \text{Im}[\chi^{(1)}] A_d + i \frac{\omega_d}{2c} \text{Re}[\chi^{(1)}] A_d \quad (16)$$

where the wavevector mismatch is given by:

$$\Delta k_d = k_d + k_c - 2k_a \quad (17)$$

The wavevector mismatch will be zero on resonance for the three-level atom considered in this treatment. In fact the contribution to the refraction from all the other atomic levels must be taken into account whilst computing Δk_d and it is implicit that these make a finite contribution to the wavevector mismatch.

We can solve this first-order differential equation with the boundary condition $A_d(z=0) = 0$ to find the generated intensity $I(\omega_d)$ after a length z :

$$I(\omega_d) = \frac{3n\omega_d^2}{8Z_0c^2} |\chi^{(3)}|^2 |A_a|^4 |A_c|^2 \frac{\left\{ 1 + e^{-\frac{\omega_d}{c} \text{Im}[\chi^{(1)}] z} - 2e^{-\frac{\omega_d}{c} \text{Im}[\chi^{(1)}] z} \cos\left(\Delta k + \frac{\omega_d}{2c} \text{Re}[\chi^{(1)}] z\right) \right\}}{\frac{\omega_d^2}{4c^2} \text{Im}[\chi^{(1)}]^2 + \left\{ \Delta k + \frac{\omega_d}{2c} \text{Re}[\chi^{(1)}] \right\}^2}$$

where Z_0 is the impedance of free space. This expression is quantitatively correct for the case of the assumptions made. More generally the qualitative predictions and general scaling remain valid in the limit where the pulse duration is significantly longer than the time required to traverse the medium. Note that both real and imaginary parts of $\chi^{(1)}$ and $\chi^{(3)}$ play an important role, as

we would expect for resonant frequency mixing. The influence of that part of the medium refraction which is due to other levels is contained in the terms with Δk . In the case of a completely transparent medium this becomes a major limit to the conversion efficiency.

Propagation and Wave-Mixing in a Doppler Broadened Medium

Doppler shifts arising from the Maxwellian velocity distribution of the atoms in the medium lead to a corresponding distribution in the detunings for the various members of the atomic ensemble. The response of the medium, as characterized by the susceptibilities, must therefore include the Doppler effect by performing a weighted sum over possible detunings. The weighting is determined by the Gaussian form of the Maxwellian velocity distribution. From this operation the effective values of the susceptibilities at a given frequency are obtained, and these quantities can be used to calculate the generated field. This step is of considerable practical importance as in most up-conversion schemes it is not possible to achieve Doppler-free geometries and the use of laser cooled atoms, although in principle possible, limits the atomic density that can be employed. The interference effects persist in the dressed profiles providing the coupling laser Rabi frequency is comparable to or larger than the inhomogeneous width. This is because the Doppler profile follows a Gaussian distribution which falls off much faster in the wings of the profile than the Lorentzian profile due to the natural broadening.

In the case considered with weak probe field, excited state populations and coherences remain small. The two-photon transition need not be strongly driven (i.e., a small two-photon Rabi frequency can be used) but a strong coupling laser is required. The coupling laser must be intense enough that its Rabi frequency is comparable to or exceeds the inhomogeneous widths in the system (i.e., Doppler width), and a laser intensity of above 1 MW cm^{-2} is required for a typical transition. This is trivially achieved even for unfocused pulsed lasers, but does present a serious limit to cw lasers unless a specific Doppler-free configuration is employed. The latter is not normally suitable for a frequency up-conversion scheme if a significant up-conversion factor is required, e.g., to the vacuum ultraviolet (VUV); however recent experiments report significant progress in cw frequency up-conversion using EIT and likewise a number of other possibilities, e.g., laser-cooled atoms and standing-wave fields, have been proposed.

A transform-limited single-mode laser pulse is essential for the coupling laser field since a multimode field will cause an additional dephasing effect on the coherence, resulting in a deterioration of the quality of the interference. In contrast, whilst it is advantageous for the field driving the two-photon transition to be single mode (in order to achieve optimal temporal and spectral overlap with the EIT hole induced by the dressing laser), this is not essential since this field does not need to drive the coherence responsible for interference.

When a pulsed laser field is used additional issues must be considered. The group velocity is modified for pulses propagating in the EIT large reductions, e.g., by factors down to $<c/100$, in the group velocity have been observed. Another consideration beyond that found in the simple steady-state case is that the medium can only become transparent if the pulse contains enough energy to dress all the atoms in the interaction volume. The minimum pulse energy to prepare a transparency is:

$$E_{\text{preparation}} = \frac{f_{13}}{f_{23}} NL \hbar \omega \quad (18)$$

where f_{ij} are the oscillator strengths of the transitions and NL the product of the density and the length. Essentially the number of photons in the pulse must exceed the number of atoms in the interaction volume to ensure all atoms are in the appropriate dressed state. This puts additional constraints on the laser pulse parameters.

Up-conversion to the UV and vacuum UV has been enhanced by EIT in a number of experiments. Only pulsed fields have so far been up-converted to the VUV with EIT enhancement. The requirements on a minimum value of $\Omega_c > \Delta_{\text{Doppler}}$ constrains the conversion efficiency that can be achieved because the $1/\Omega_c^2$ factor in Eq. (17) ultimately leads to diminished values of $\chi^{(3)}$. The use of gases of higher atomic weight at low temperatures is therefore highly desirable in any experiment utilizing EIT for enhancement of four-wave mixing to the VUV. Conversion efficiencies, defined in terms of the pulse energies by E_d/E_a or E_d/E_c of a few percent have been achieved using the EIT enhancement technique. It is typically most beneficial to maximize the conversion efficiency defined by the first ratio since ω_a is normally in the UV and is the lower energy of the two applied pulses.

Nonlinear Optics with a Pair of Strong Coupling Fields in Raman Resonance

An important extension of the EIT concept occurs when two strong fields are applied in Raman resonance between a pair of states in a three-level system. Considering the system illustrated in Fig. 1 we can imagine that both applied fields are now strong. Under appropriate adiabatic conditions the system evolves to produce the maximum possible value for the coherence $\rho_{12}=0.5$. Adiabatic evolution into the maximally coherent state is achieved by adjusting either the Raman detuning or the pulse sequence (counter to intuitive order). The pair of fields may also be in single-photon resonance with a third level, in which case the EIT-like elimination of absorption will be important. This situation is equivalent to the formation of a darkstate, since neither of the two strong fields is absorbed by the medium. For sufficiently strong fields the single-photon condition need not be satisfied and a maximum coherence will still be achieved.

An additional field applied to the medium can participate in sum- or difference-frequency mixing with the two Raman resonant fields. The importance of the large value of coherence is that it is the source polarization that drives the new fields generated in the frequency mixing process. Complete conversion can occur over a short distance that greatly alleviates the constraints usually set by phase-matching in nonlinear optics. Recently near unity conversion efficiencies to the far-UV were reported in an atomic lead system where maximum coherence had been created. In a molecular medium large coherence between vibrational or rotational levels has also been achieved using adiabatic pulse pairs. Efficient multi-order Raman sideband generation has been observed to occur. This latter observation may lead the way to synthesizing very short duration light pulses since the broadband Raman sideband spectrum has been proved to be phase-locked.

Pulse Propagation and Nonlinear Optics for Weak CW Fields

In a Doppler-free medium a new regime can be accessed. This is shown in Fig. 4 where the possibility now arises of extremely narrow transparency dips since very small values of Ω_c are now sufficient to induce EIT. The widths of these features are typically subnatural and are therefore accompanied by very steep normal dispersion, which corresponds to a much reduced group velocity. The ultraslow propagation of pulses is one of the most dramatic manifestations of EIT in this regime. Nonlinear susceptibilities are now very large as there is constructive interference controlling the value and the splitting of the two states is negligible compared to their natural width. Nonlinear optics at very low light levels, i.e., at the few-photon limit, is possible in this regime.

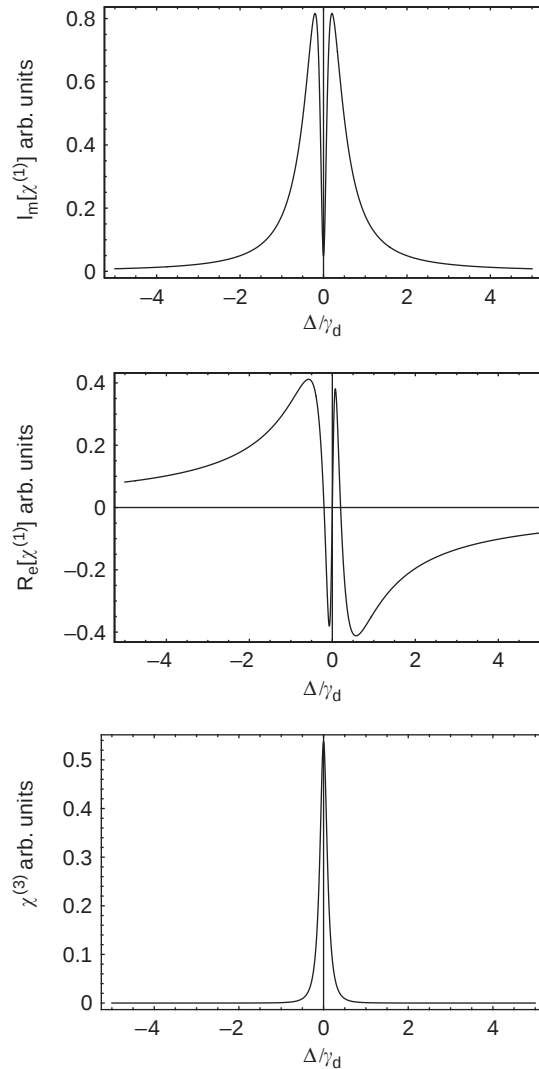


Fig. 4 Electromagnetically induced transparency is shown in the case where there is no significant Doppler broadening. The values of the imaginary $\text{Im}[\chi^{(1)}]$ and real $\text{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. We take $\Omega_c = \gamma_d/5$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

Propagation of pulses is significantly modified in the presence of EIT. Fig. 4 shows the changes to $\text{Re}[\chi^{(1)}]$ that arise. Within the transparency dip there exists a region of steep normal dispersion. In the vicinity of resonance this is almost linear and it becomes reasonable to consider the leading term only that describes the group velocity. An analysis of the refractive changes has been provided by Harris who expanded the susceptibilities (both real and imaginary parts) of the dressed atom around the resonance frequency to determine various terms in $\text{Re}[\chi^{(1)}]$. The first term of the series (zero order) $\text{Re}[\chi^{(1)}](\omega_{13})=0$ corresponds to the vanishing dispersion at resonance. The next term $\partial[\text{Re}\chi^{(1)}](\omega)/\partial\omega$ gives the slope of the dispersion curve; at ω_{13} this takes the value:

$$\frac{\partial \text{Re}\chi^{(1)}(\omega_{13})}{\partial\omega} = \frac{|\mu_{13}|^2 N 4(\Omega_c^2 - \gamma_a^2)}{2\pi\epsilon_0(\Omega_c^2 + \gamma_a\gamma_d)} \quad (19)$$

The slope of the dispersion profile leads to a reduced group velocity v_g :

$$\frac{1}{v_g} = \frac{1}{c} + \frac{\pi}{\lambda} \frac{\partial \chi^{(1)}}{\partial\omega} \quad (20)$$

From the expression for $\partial\chi/\partial\omega$ we see that this slope is steepest (and so v_g minimum) in the case where $\Omega_c \gg \Gamma_2$ and $\Omega_c^2 \gg \Gamma_2 \Gamma_3$ but is still small compared to Γ_3 (i.e., $\Omega_c < \Gamma_3$) and hence $\partial\chi/\partial\omega \propto 1/\Omega_c^2$. In the limit of small Ω_c the following expression for v_g therefore holds:

$$v_g = \frac{\hbar c \epsilon_0 [\Omega_c]^2}{2\omega_d |\mu_{13}|^2 N} \quad (21)$$

Extremely low group velocities, down to a few meters per second, are achieved in this limit using excitation of the hyperfine split ground states in either laser cooled atomic samples or Doppler-free configurations in finite temperature samples. Recently similar light slowing has been observed in solids. Storage of the optical pulse within the medium has also been achieved by adiabatically switching off the coupling field and thus trapping the optical excitation as an excitation within the hyperfine ground states for which the storage time can be very long (> 1 ms) due to very low dephasing rates. Since the storage scenario should be valid even for single photons this process has attracted considerable attention recently as a means to enable quantum information storage.

Extremely low values of Ω_c are sufficient to induce complete transparency (albeit in a very narrow dip) and at this location the nonlinear response is resonantly enhanced. Very high efficiency nonlinear frequency mixing and phase conjugation at low light levels have been reported under these conditions. It is expected that high-efficiency nonlinear optical processes will persist to ultralow intensities (the few photon level) in an EIT medium of this type.

One example of highly enhanced nonlinear interactions is the predicted large values of the Kerr type nonlinearity (nonlinear refractive index). The origin of this can be seen by considering the steep dispersion profile in the region of the transparency dip in Fig. 4. Imagine that we apply an additional field ω_f , perhaps a very weak one, at a frequency close to resonance between state $|2\rangle$ and a fourth level $|4\rangle$. The ac Stark shift caused by this new field to the other three level, although small, will have a dramatic effect upon the value of the refractive index because of the extreme steepness of the dispersion profile. Essentially even a very small shift of the resonant wavelength causes a large change in the refractive index for a field applied close to this frequency. It is predicted that strong cross-phase modulations will be created in this process between the fields ω_f and ω_d , even in the quantum limit for these fields. This is predicted to lead to improved schemes for quantum nondemolition measurements of photons through the measurement of the phaseshifts they induce (through cross-phase modulation) on another quantum field. This type of measurement has direct application in quantum information processing.

Further Reading

- Arimondo, E., 1996. Coherent population trapping in laser spectroscopy. *Progress in Optics* 35, 257–354.
Harris, S.E., 1997. Electromagnetically induced transparency. *Physics Today* 50, 36–42.
Harris, S.E., Field, J.E., Imamoglu, A., 1990. Non-linear optical processes using electromagnetically induced transparency. *Physical Review Letters* 64, 1107–1110.
Harris, S.E., Hau, L.V., 1999. Non-linear optics at low light levels. *Physical Review Letters* 82, 4611–4614.
Hau, L.V., Harris, S.E., Dutton, Z., Behroozi, C.H., 1999. Light speed reduction to 17 metres per second in an ultra-cold atomic gas. *Nature* 397, 594–598.
Marangos, J.P., 2001. Electromagnetically induced transparency. In: Bass, M. (Ed.), *Handbook of Optics*, vol. IV, ch. 23. New York: McGraw-Hill.
Merriam, A.J., Sharpe, S.J., Xia, H., *et al.*, 1999. Efficient gas-phase generation of vacuum ultra-violet radiation. *Optics Letters* 24, 625–627.
Schmidt, H., Imamoglu, A., 1996. Giant Kerr nonlinearities obtained by electromagnetically induced transparency. *Optics Letters* 21, 1936–1938.
Scully, M.O., 1991. Enhancement of the index of refraction via quantum coherence. *Physical Review Letters* 67, 1855–1858.
Scully, M.O., Suhail Zubairy, M., 1997. *Quantum Optics*. Cambridge, UK: Cambridge University Press.
Sokolov, A.V., Walker, D.R., Yavuz, D.D., *et al.*, 2001. Femtosecond light source for phase-controlled multi-photon ionisation. *Physical Review Letters* 87, 033402-1.
Zhang, G.Z., Hakuta, K., Stoicheff, B.P., 1993. Nonlinear optical generation using electromagnetically induced transparency in hydrogen. *Physical Review Letters* 71, 3099–3102.

Nonlinear Optics, Basics: Nomenclature and Units

MP Hasselbeck, The University of New Mexico, Albuquerque, NM, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

CARS	Coherent anti-Stokes Raman scattering	OPO	Optical parametric oscillation/oscillator
c.c.	Complex conjugate	2PA, TPA	Two-photon absorption
CSRS	Coherent Stokes Raman scattering	PCM	Phase conjugate mirror
DFG	Difference frequency generation	QPM	Quasi phase matching
DFM	Difference frequency mixing	RIKES	Raman-induced Kerr effect spectroscopy
DFWM, FWM	(Degenerate) Four wave mixing	RSA	Reverse saturable absorption
DRO	Doubly resonant OPO	SBS	Stimulated Brillouin scattering
EIT	Electromagnetically induced transparency	SFG	Sum frequency generation
ESA	Excited state absorption	SHG	Second harmonic generation
GVD	Group velocity dispersion	SIT	Self-induced transparency
GVM	Group velocity mismatch	SRO	Singly resonant OPO
NLA	Nonlinear absorption	SRS	Stimulated Raman scattering
NLR	Nonlinear refraction	SRWS	Stimulated Rayleigh wing scattering
OFID	Optical free-induction decay	SVEA, SVAA	Slowly varying envelope (amplitude) approximation
OPA	Optical parametric amplification/amplifier	THG	Third harmonic generation
OPG	Optical parametric generation/generator	TPF	Two-photon fluorescence
OPL	Optical power limiter		

Nomenclature Associated with the Excitation Light

Particular care must be used when characterizing the excitation beam in nonlinear optical experiments compared to linear measurements. By definition, nonlinear optical phenomena depend on the electric field to high order. The higher the order, the more sensitive the observed behavior depends on the input. Extracting a representative nonlinear coefficient from an experiment, for example, becomes progressively more difficult as the order of the optical nonlinearity increases. In other words: the errors associated with optical beam characterization get magnified by the order of the nonlinearity under study.

There is an extensive nomenclature for characterizing the light (almost always laser light) interacting with the nonlinear optical medium. Different descriptions may be used depending on whether the excitation light is pulsed, continuous, or a continuous train of pulses. When a laser beam is constant or continuous, it is often described by its 'cw power'. 'Cw' stands for 'continuous wave', an acronym taken from the nomenclature of electronics. The cw power of a laser is determined by placing a power meter in the path of a beam. Repetitively pulsed lasers can be characterized in the same way, provided the response time of the power meter is slower than the pulse separation period. This will almost certainly be the case with a continuously pumped mode-locked laser such as a dye laser, fiber laser, or Ti:sapphire laser, which produce pulses at repetition frequencies of tens of megahertz. Cw power may also be suitable for describing pulsed flashlamp lasers, gas discharge lasers, or any laser that is excited in a periodic fashion.

When the optical output can be distinguished as individual pulses, additional metrics are used. A pulse, by definition, exists during a window of time, i.e., there is a time when light is present and a time when light is absent. How one characterizes the time light is present – the pulse duration – is critically important and not always obvious. There are a myriad of ways the pulse temporal envelope can manifest itself; common examples include square pulses, triangular pulses, Gaussian pulses, and hyperbolic secant pulses. These names refer to the mathematical waveforms that map the pulse envelope as a function of time.

A pulse waveform that exhibits symmetry about the peak in its temporal envelope is commonly characterized by its 'full-width, half-maximum', abbreviated FWHM. One obtains the full-width, half-maximum by locating the two points on the pulse profile that are at half the peak value. The temporal separation of these two points is the FWHM. One uses this measure because in a strict mathematical sense, waveforms such as the Gaussian or hyperbolic-secant exist even at times $t = \pm \infty$. A less common term is the 'half-width, half-maximum' or HWHM. As the name implies, the HWHM is exactly half the FWHM value. Not all light-pulses are temporally symmetric, however, so greater detail may be needed when giving a mathematical description of asymmetric pulses.

Light pulses are also characterized by their energy and peak power. The temporal envelope recorded by a so-called square-law detector (all laboratory detectors are square-law detectors) shows the power of the pulse as a function of time, provided the detector response time is sufficiently fast. Integrating this waveform gives the pulse energy. It makes no sense to talk about the energy of a cw beam of light, unless that beam is composed of repetitive, distinguishable pulses. Each pulse in the train carries a distinct amount of optical energy.

'Ultrafast' or 'ultrashort' laser pulses generally refer to light pulses that are of such small duration they cannot be measured directly with detectors and oscilloscopes because of bandwidth limitations. To infer the pulse duration, so-called intensity autocorrelation measurements are often made. The temporal power profile is then deduced from the autocorrelation signal with the appropriate mathematical conversion factor. Equally important is the spectrum of a short pulse. The optical bandwidth determines the lower limit of temporal compression; such ideally compressed pulses are said to be 'transform limited'. If different components of the spectrum can be distinguished at specific positions on the pulse temporal profile, then the pulse is 'chirped'.

Nonlinear optical effects take place when light is concentrated on a target – there is a cross-section or 'footprint' of the beam on the medium. The interaction of the light and material takes place in a region called the beam area, beam cross-section, focal area, or focal volume. One must then define the appropriate interaction area or volume.

Specification of the beam area allows one to define two important quantities used frequently in nonlinear optics. The power/area is known as irradiance or power density. This quantity is commonly called the intensity, although the strict radiometric definition of intensity is power/(solid angle). The second important quantity is energy/area, which is called the energy density or fluence.

The convention in nonlinear optics is to define the optical electric field as:

$$\mathbf{E}(\mathbf{z}, t) = \frac{1}{2} E_0 \exp[i(\mathbf{k} \cdot \mathbf{z} - \omega t)] + \text{c.c.} \quad (1)$$

where E_0 is the peak field, $\mathbf{k}$ is the propagation vector, ω is the angular frequency of the light, and c.c. stands for complex conjugate. This represents a forward- and backward-propagating (in the z -direction) infinite, transverse plane wave. Other definitions are also used. In SI/mks units and using the convention of Eq. (1), the irradiance (I in units of W/m^2) inside a material of refractive index n is related to the electric field vector of the light ($\mathbf{E}$ in units of V/m) as follows:

$$I = \frac{1}{2} c n \epsilon_0 |\mathbf{E}|^2 \quad (2)$$

Here c is the speed of light (2.998×10^8 m/s) and ϵ_0 is the permittivity of free space (8.854×10^{-12} F/m). Nonlinear optics has an unfortunate tradition of mixing mks and cgs units, resulting in a lot of confusion. Irradiance is usually expressed in units of W/cm^2 . One can calculate the peak electric field (in units of V/cm) of a laser beam given its irradiance (in units of W/cm^2) by using this simple formula:

$$E_0 = 38.82 \sqrt{\frac{I}{n}} \left(\frac{\text{V}}{\text{cm}} \right) \quad (3)$$

The root-mean-square value of the electric field is obtained by replacing the factor 38.82 by 27.45. In Gaussian/cgs units, irradiance is expressed in units of $\text{ergs}/(\text{cm}^2 \text{ sec})$ and is related to the field as:

$$I = \frac{cn}{8\pi} |\mathbf{E}|^2 \quad (4)$$

where the field is in units of statvolt/cm and $c = 2.998 \times 10^{10}$ cm/sec. Conversion between mks and cgs for the electric field is $3 \times 10^4 \text{ V/m} = 1 \text{ statvolt/cm}$; irradiance is converted using $1 \text{ erg}/(\text{cm}^2 \text{ sec}) = 1 \times 10^{-3} \text{ W/m}^2$. The following point must be emphasized: depending on how the electric field is defined, there can be factors of 2 or even 4 discrepancies in the irradiance values quoted by different authors. The common definitions are used here, although they are certainly not universal.

A very common realization of the spatial irradiance profile of a laser beam is the radially symmetric Gaussian function:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \quad (5)$$

where I_0 is the peak irradiance at the center of the beam ($r=0$) and w is the 'spot size'. The spot size changes continuously as the beam propagates and the minimum spot size is known as the waist. One often hears the spot size referred to as the radius, but the radius of a Gaussian beam traditionally denotes the curvature of the phase front. By definition, the phase front radius is infinite at the waist. The factor of 2 appearing in the argument of the exponential in Eq. (5) stems from the fact that the spot size w is conventionally defined for the electric field of the Gaussian beam. When the field is squared to obtain the irradiance, the factor of 2 appears. Using this function for the irradiance spatial profile leads to the common parameter '1/e² diameter'. The irradiance falls to 1/e² (0.1353) of its peak value when $r=w$.

The Gaussian spatial profile or 'Gaussian beam' results when the laser has been designed to operate in the lowest-order transverse mode, usually denoted TEM₀₀. The TEM₀₀ Gaussian beam is particularly convenient because the same relative power profile is maintained in the near- and far-field, whether or not the beam is collimated or focused.

The cross-sectional area of a TEM₀₀ Gaussian beam normally incident on a surface is found as follows. The power in the beam is obtained by integrating the irradiance profile (I) over the surface:

$$P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I dA \quad (6)$$

Referring to Eq. (5), the integral becomes:

$$P = \int_0^{2\pi} d\theta \int_0^{\infty} r dr I_0 \exp\left(-\frac{2r^2}{w^2}\right) = I_0 \frac{\pi w^2}{2} \quad (7)$$

Hence the effective area of a TEM₀₀ Gaussian beam normally incident on a surface is:

$$A = \frac{\pi w^2}{2} \quad (8)$$

There are many situations where the Gaussian formulation of Eq. (5) is not suitable. Beams produced by unstable laser resonators, for example, are not Gaussian. A multitransverse mode beam profile obtained from a stable resonator does not usually lend itself to a simple mathematical characterization.

Nomenclature Associated with the Nonlinear Optical Medium

In nonlinear optics texts, the subject is often introduced by writing the macroscopic polarization in Maxwell's equations as a power series expansion in the electric field. This approach, proposed by Bloembergen and coworkers in the early 1960s, has been spectacularly successful for interpreting experiments, though it has also led to confusion in the definition of nonlinear optical coefficients. The confusion, which stems from arbitrary definitions and nomenclature used by different authors, cannot be resolved here. The reader should, however, be alert for these discrepancies. Comparison of results published by different laboratories requires that the fundamental equations for extracting the nonlinear coefficients from their data are known. As the discipline has matured, the nonlinear optics community largely recognized the ambiguities and confusion; precise definitions of terms and coefficients are now commonly provided in the research literature.

It should also be noted that the power-series framework may not always be the best formulation for modeling and characterizing nonlinear optical effects. A carefully designed optical limiter, for example, may exhibit an abrupt decrease of transmission at a specific 'threshold irradiance' or 'threshold fluence', perhaps at a given laser pulse duration. The physical phenomenon or phenomena driving this behavior may not readily succumb to characterization by an n th-order coefficient in a power series expansion. In this situation it may be appropriate to specify a threshold optical input parameter. Another example is a homogeneously broadened saturable absorber, in which the absorption of an optical medium decreases with increasing input irradiance. The irradiance-dependent absorption coefficient α (refer to Eq. (26)) is characterized by a 'saturation irradiance' I_{sat} :

$$\alpha(I) = \frac{\alpha_0}{1 + \frac{I}{I_{\text{sat}}}} \quad (9)$$

where α_0 is the linear absorption coefficient.

Nonlinear Susceptibility

The power-series expansion of the macroscopic polarization is the standard approach for modeling nonlinear optical behavior and categorizing the various phenomena. A common way to write the polarization, nonlinear in the electric field, is (mks units):

$$\mathbf{P} = \epsilon_0(\chi^{(1)} \cdot \mathbf{E} + \chi^{(2)} : \mathbf{E}\mathbf{E} + \chi^{(3)} : \mathbf{E}\mathbf{E}\mathbf{E} + \dots) \quad (10)$$

where the polarization $\mathbf{P}$ is a time- and space-dependent vector and the terms $\chi^{(n)}$ are the various orders of the nonlinear susceptibility (the cgs equation is obtained by dropping the free-space permittivity coefficient ϵ_0). A second-order effect is associated with $\chi^{(2)}$, a third-order with $\chi^{(3)}$, and so on. In general, there are distinguishable electric field vectors (in Eq. (10), the fields have not been individually designated for clarity). The susceptibility terms are generally tensors, which means the medium is sensitive to the orientation of the input fields. The dots in Eq. (10) indicate tensor products. The input field vectors can have different frequencies and the presence of complex conjugates in Eq. (1) indicates that the frequency components will have both + and - signs, i.e. the j th electric field term will have associated frequency factors $\exp(\pm i\omega_j t)$.

The nomenclature used here can be illustrated by example. In the case of the second-order nonlinear polarization (mks units):

$$\mathbf{P}_j = \epsilon_0 \chi_{jkl}^{(2)} : \mathbf{E}_k \mathbf{E}_l \quad (11)$$

Note that subscripts have been introduced. The indices correspond to Cartesian space vectors. This equation says that the polarization in the j -direction results from the tensor product of the appropriate $\chi^{(2)}$ with the input fields at $\mathbf{E}_k$ and $\mathbf{E}_l$. The j th component of polarization has an oscillation frequency that is determined by mixing of the input fields. In this example, the indices j , k and l can each take the value of x , y , and z . Let each input field be at one of two possible frequencies, call them ω_1 and ω_2 . There are an enormous number of permutations (81 to be exact) of Eq. (11). One term is:

$$P_x(\omega_1 + \omega_2) = \epsilon_0 \chi_{xyz}^{(2)}(\omega_1 + \omega_2) E_z(\omega_1) E_y(\omega_2) \exp[-i(\omega_1 + \omega_2)t] \quad (12)$$

Another possibility is:

$$P_z(\omega_2 - \omega_1) = \epsilon_0 \chi_{zyx}^{(2)}(\omega_2 - \omega_1) E_y(\omega_2) E_x(-\omega_1) \exp[-i(\omega_2 - \omega_1)t] \quad (13)$$

Eq. (12) corresponds to sum frequency generation ($\omega_1 + \omega_2$) and Eq. (13) shows difference frequency generation ($\omega_2 - \omega_1$).

Not explicitly written on the right-hand side of the above equations is the exponential containing the mixing of the propagation vectors. To maximize the nonlinear polarization, which is the source of nonlinear behavior in Maxwell's equations, the

vector sum $\sum \mathbf{k}_j$ for *all* the interacting fields should be close to zero. Arranging the propagation vectors to accomplish this is known as ‘phase matching’.

It is important to point out that there are eight more terms describing the nonlinear polarization at $P_x(\omega_1 + \omega_2)$ in addition to Eq. (12); likewise for $P_z(\omega_2 - \omega_1)$. Also note that the indices k and l may be identical, hence a nonlinear polarization driven by $\chi_{xy}^{(2)}$ for example, is allowed. Higher-order nonlinear polarizations get progressively more complicated. The general third-order nonlinear polarization is

$$\mathbf{P}_j = \epsilon_0 \chi_{jklm}^{(3)} : \mathbf{E}_k \mathbf{E}_l \mathbf{E}_m \quad (14)$$

Sometimes the second-order nonlinear susceptibility is written with the coefficient d instead of $\chi^{(2)}$. The relation between the d -coefficient and $\chi^{(2)}$ depends arbitrarily on how it is defined; a common definition is

$$d_{jkl} = \frac{1}{2} \chi_{jkl}^{(2)} \quad (15)$$

but the reader is cautioned that the factor $\frac{1}{2}$ is sometimes missing. In the preceding discussion it was pointed out how there can, in principle, be a large number of tensor components involved in the expansion of the nonlinear polarization. Simplification occurs when it is realized that there is no discernible physical difference between the frequency terms ω_j and $-\omega_j$ and that ordering of the fields in Eq. (14) – which suggests a time order in the arrival of the fields – is irrelevant. These symmetry arguments show that $d_{jkl} = d_{jlk}$, which reduces the number of independent d -coefficients from 81 to 18. A simpler subscript notation is then introduced that uses the integers 1–6 to represent pairs of Cartesian components k and l . This reduces the number of subscripts from three to two. An example of this notation is $d_{xxz} = d_{14}$.

In certain situations, one can take advantage of crystal symmetry to further reduce the complicated summations to a single scalar coefficient for the nonlinear susceptibility, which is referred to as ‘ d -effective’. For these specialized cases, the nonlinear polarization is

$$\mathbf{P} = \epsilon_0 d_{\text{eff}} \mathbf{E}_1 \mathbf{E}_2 \quad (16)$$

In SI/mks units, the polarization is in units of coul/m². The second-order susceptibility (i.e. $\chi^{(2)}$, d_{jkl}) must therefore have units of m/volt. The units of $\chi^{(3)}$ are m²/volt², $\chi^{(4)}$ is m³/volt³, and so on. In Gaussian/cgs units, $\chi^{(2)}$ has dimensions cm/statvolt, $\chi^{(3)}$ will be cm²/statvolt², etc. although sometimes in the Gaussian system *all* the coefficients $\chi^{(n)}$ are discussed with shorthand ‘electrostatic units’ or ‘esu’.

Complex Quantities

In linear optics, the susceptibility has real and imaginary parts

$$\chi^{(1)} = \chi_{\text{real}}^{(1)} + i\chi_{\text{imaginary}}^{(1)} \quad (17)$$

The complex linear index is, in turn, written as the sum of real and imaginary components, derived from the linear susceptibility as follows (mks units):

$$\sqrt{1 + \chi^{(1)}} = n_0 + i\kappa \quad (18)$$

where n_0 is the linear refractive index and κ is the imaginary term leading to absorption of light. In general, the nonlinear susceptibility $\chi^{(n)}$ is also a complex number:

$$\chi^{(n)} = \chi_{\text{real}}^{(n)} + i\chi_{\text{imaginary}}^{(n)} \quad (19)$$

As in linear optics, the complex notation is a bookkeeping method that conveniently accounts for what is known as ‘resonant enhancement’. Nonlinear optics research has revealed that the nonlinear susceptibility $\chi^{(n)}$ can be a strong function of frequency. Specifically, this quantity will be strongly enhanced when sums and/or differences of the photon energies in the interacting light beams coincide with quantum mechanical energy resonances in the material. When a resonance condition is achieved, $\chi^{(n)}$ will be dominated by its imaginary component. Far from the resonances, $\chi^{(n)}$ behaves more like a real quantity. These complex terms directly enter the wave equations describing how light propagates in a nonlinear medium and are particularly important for $\chi^{(3)}$; the complex quantity $\chi^{(3)}$ determines whether light will be refracted or absorbed and by how much. The real part of $\chi^{(3)}$ drives nonlinear refraction while the imaginary portion characterizes nonlinear absorption (e.g., two-photon absorption) and the inverse effect – gain.

Nonlinear Refraction

The most common manifestation of nonlinear refraction arises from the third-order nonlinear susceptibility, the so-called optical Kerr effect. In this case, the refractive index is linearly proportional to the irradiance (I) of a monochromatic light beam. The irradiance-dependent refractive index is

$$n(I) = n_0 + n_2 I \quad (20)$$

where n_0 is the linear, irradiance-independent refractive index and n_2 is the nonlinear refractive index coefficient. Because n is a dimensionless quantity, n_2 must be in units of area/power.

The optical Kerr effect is usually written in mks units with the expression shown in Eq. (20). In cgs units, the common (but by no means universal) convention is to write:

$$n(E) = n_0 + \frac{1}{2} \tilde{n}_2 |E|^2 \quad (21)$$

where the field is in units of statvolts/cm. The nonlinear coefficient $\tilde{n}_2$ (cgs) must therefore be in units of $\text{cm}^2/\text{statvolt}^2$, which is sometimes abbreviated to 'esu'. The conversion for n_2 between these two equations is:

$$\tilde{n}_2(\text{cgs}) = \frac{n_0 c}{40\pi} n_2(\text{mks}) \quad (22)$$

where $n_2(\text{mks})$ is in units of m^2/W and c is in m/sec. Also useful is the relation between n_2 and $\chi^{(3)}$, written in mks units as:

$$\Delta n = n_2 I = \frac{\text{Re}(\chi^{(3)})}{4\epsilon_0 n^2 c} I \quad (23)$$

where 'Re' denotes the real part of $\chi^{(3)}$. In cgs units we have:

$$\Delta n = \frac{1}{2} \tilde{n}_2 |E|^2 = \frac{4\pi^2 \text{Re}(\chi^{(3)})}{n^2 c} I \quad (24)$$

where I is in units of $\text{ergs}/(\text{cm}^2 \text{ sec})$ defined by Eq. (4). The reader is again advised that different authors will show discrepancies of 2, 4, or even 8 when writing these equations. The susceptibilities $\chi^{(3)}$ are related as follows:

$$\tilde{\chi}^{(3)}(\text{cgs}) = \frac{9 \cdot 10^8}{4\pi} \chi^{(3)}(\text{mks}) \quad (25)$$

If the coefficients n_0 and n_2 exist, does the nomenclature imply there are terms n_1 , n_3 , and others? These coefficients are certainly allowed, but not often seen in discussions of nonlinear refraction. When one refers to nonlinear refraction, the common understanding is that the index depends linearly on irradiance, which is conveniently modeled by Eq. (20). But there are other physically relevant situations to consider. The coefficient n_1 , for example, describes the linear electro-optic effect in which the change in index is linearly proportional to the electric field (although it is called the linear electro-optic effect, it is actually derived from the second-order nonlinear susceptibility). The electric field can be the oscillating field of a laser beam, for example, or a dc field applied to an electro-optic crystal (e.g., Pockel's cell). When other terms are added to Eqs. (20) or (21), the units of the coefficients must be chosen to keep the equation dimensionless.

In the preceding discussion, there is an implication that the nonlinear index is an instantaneous function of the irradiance or field. Although the material system can never respond instantaneously, this is a good approximation in many situations. Sometimes it is not. Consider a pulsed laser beam that heats the material. When the local temperature increases, the linear refractive index can change. A short, Q-switched laser can have a pulse duration far less than a microsecond, while the temperature change it induces can last many orders of magnitude longer. This means optical modification of the refractive index may persist long after the exciting pulse has vanished, i.e., when the irradiance is at zero.

Another example is when a laser beam promotes electrons from their ground state to higher energy states of the material system. These excited electrons may modify the refractive index of the material. In general, the excited electrons remain in high-energy states for some period of time before relaxing to the ground level – if this time is longer than the excitation, the change of the refractive index persists beyond the time the laser beam is present. In these situations, the model of nonlinear refraction suggested by the above equations is inaccurate. One cannot obtain the nonlinear refraction from a simple algebraic analysis. A system of time-dependent equations – describing the dynamics of excitation and relaxation – must be solved using numerical procedures. Such phenomena are sometimes loosely categorized as 'dynamic linear optical effects' or 'effective third-order nonlinearities'.

Nonlinear Absorption

Consider a single-frequency light beam passing through an optically absorbing region of length L . For simplicity, neglect reflections caused by surfaces that may define the region of interest, i.e., ignore reflections at surfaces that may be located on the optical axis z at points $z=0$, $z=L$, or any other point in the path. Linear absorption means that the optical power extracted from the light beam as it traverses the absorbing medium is a direct function of the power at a given point. This is described mathematically by an elementary, linear differential equation known as Beer's law:

$$\frac{d}{dz} I(z) = -\alpha I(z) \quad (26)$$

The constant of proportionality is α , which is the linear absorption coefficient. Eq. (23) is solved by direct integration:

$$\int_{I(0)}^{I(L)} \frac{dI}{I} = -\alpha \int_0^L dz \quad (27)$$

Table 1 Dimensions of optical absorption coefficients

Process	Coefficient	Units
Linear absorption	α, K_1	$(\text{length})^{-1}$
Two-photon absorption	β, K_2	$\text{length}/\text{power}$
Three-photon absorption	γ, K_3	$\text{length}^3/\text{power}^2$
Four-photon absorption	K_4	$\text{length}^5/\text{power}^3$

This has the solution:

$$\frac{I(L)}{I(0)} = \exp(-\alpha L) \quad (28)$$

which means the irradiance decreases exponentially as a function of propagation distance in a linearly absorbing medium.

In the nonlinear regime, we don't expect a classical Beer's law model to hold. By definition, the absorption will be a nonlinear function of irradiance at a given point. One makes the following power-series expansion to describe nonlinear absorption of monochromatic light:

$$\frac{d}{dz}I = -\alpha I - \beta I^2 - \gamma I^3 - \dots \quad (29)$$

The coefficient β corresponds to a two-photon absorption process and γ is the coefficient of three-photon absorption. What about four-photon and even higher-order processes? These are rarely encountered, but certainly possible. To make these distinctions, [Eq. \(29\)](#) is sometimes written with coefficients K_i or α_i instead of the sequential Greek alphabet:

$$\frac{d}{dz}I = -K_1 I - K_2 I^2 - K_3 I^3 - K_4 I^4 - \dots \quad (30)$$

The units of the various absorption coefficients must maintain the dimensional consistency of [Eqs. \(29\)](#) and [\(30\)](#), which is irradiance/length or power/(length)³. These are shown in [Table 1](#).

If there is linear absorption, the nonlinear processes at that wavelength are often (but not always) negligibly weak. This means we have the following inequalities:

$$K_1 I \gg K_i I^i, \quad \text{where } i \geq 2 \quad (31)$$

Nonlinear absorption is generally observed in the wavelength region where the medium is transparent to low-irradiance light, i.e., where linear absorption is negligible. It should be emphasized that there are situations where linear absorption is large and in fact a crucial component of the aggregate nonlinear effect. We will return to this point later in the discussion. For the moment, we neglect linear absorption. If the energy of the incident photons and energy levels of the system permit it, the lowest-order nonlinear process is two-photon absorption, described by the equation

$$\frac{d}{dz}I = -K_2 I^2 \quad (32)$$

which can be solved by elementary integration:

$$\frac{I(L)}{I(0)} = \frac{1}{1 + I(0)K_2 L} \quad (33)$$

Unfortunately, the situation is rarely this convenient. The above integration has ignored the fact that nonlinear absorption may enable significant linear absorption. This is possible because nonlinear absorption promotes electrons from low- to high-energy states in the medium. The change in excited electron density associated with absorption of light (both linearly and nonlinearly) is called 'photocarrier generation'. These photocarriers (e.g., photo-electrons) may modify the linear absorption as well as the refractive properties of the material. If two-photon absorption causes the linear absorption to increase, for example, the assumption that allowed us to ignore K_1 when writing [Eq. \(32\)](#) ceases to be valid. While [Eq. \(32\)](#) was appropriate at the very start of the light-matter interaction, the generation of photocarriers after the passage of time may change that condition. The behavior of the system is therefore time-dependent, i.e., it is dynamic. Simple analytic solutions are almost always not possible when dealing with nonlinear absorption.

The time rate of change of the photocarrier population (N) is modeled by the following equation:

$$\frac{\partial}{\partial t}N = \frac{K_1 I}{\hbar\omega} + \frac{K_2 I^2}{2\hbar\omega} + \frac{K_3 I^3}{3\hbar\omega} + \frac{K_4 I^4}{4\hbar\omega} + \dots - \text{recombination} - \text{diffusion} \quad (34)$$

where ω is the angular frequency ($\omega = 2\pi f$) of the incident light. Also included are place-holders for loss processes such as recombination and diffusion, which may themselves have complicated mathematical descriptions. The coefficients K_i have exactly the same units, dimensions, and interpretation as in [Eq. \(30\)](#) and [Table 1](#); in the spirit of the preceding discussion, these coefficients may be time dependent. Note that the dimensions of [Eq. \(30\)](#) describe absorption of light (irradiance/length), while [Eq. \(34\)](#) models the photocarrier population density (length⁻³ time⁻¹). The units of these two equations are very different.

Further Reading

- Bloembergen, N., 1992. *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.
- Boyd, R.W., 1991. *Nonlinear Optics*. Boston, MA: Academic Press.
- Levenson, M.D., Kano, S., 1988. *Introduction to Nonlinear Laser Spectroscopy*. Boston, MA: Academic Press.
- Miller, A., Miller, D.A.B., Smith, S.D., 1981. Dynamic nonlinear optical processes in semiconductors. *Advances in Physics* 30, 697–800.
- Mills, D.L., 1998. *Nonlinear Optics: Basic Concepts*. New York: Springer.
- Newell, C., Moloney, J.V., 1992. *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.
- Sauter, E.G., 1996. *Nonlinear Optics*. New York: Wiley.
- Shen, Y.R., 1984. *The Principles of Nonlinear Optics*. New York: Wiley.
- Sutherland, R.L., 1996. *Handbook of Nonlinear Optics*. New York: Marcel Dekker.
- Yariv, A., 1991. *Optical Electronics*, 3rd edn. New York: Holt, Rinehart, and Winston.
- Yariv, A., 1989. *Quantum Electronics*. New York: Wiley.
- Zernike, F., Midwinter, J.E., 1973. *Applied Nonlinear Optics*. New York: Wiley.

Raman Spectroscopy

R Withnall, University of Greenwich, Chatham, UK

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Inelastic light scattering, the optical analogue of Compton scattering, had been predicted to occur by Smekal in 1923, but it was Chandrasekar Venkataraman Raman who provided the first experimental demonstration of the phenomenon in February 1928. Only a few months later in 1928 the Russian scientists, Landsberg and Mandelstam observed inelastic light scattering from a quartz crystal. In spite of initial claims to the contrary, this was the same effect that had been observed by Raman, although the Russian scientists did not acknowledge this, preferring to call the effect 'combinatorial scattering'. It was not until the 1970s that the effect became known as Raman scattering everywhere, including Russia.

As typically practised, Raman spectroscopy involves laser excitation of a sample and measurement of the wavelength and intensity distribution of the scattered Raman light. When a microscope is used for delivering laser excitation and/or collecting the inelastically scattered light, the technique is referred to as Raman microscopy. However, it is important to recognize that, apart from the differences in sampling configuration, there are no fundamental differences between Raman microscopy and Raman spectroscopy; the terms merely identify the different sampling techniques.

Since its invention in 1960, major advances in the technology of the laser have occurred, resulting in the wide variety of laser light sources that are currently available for Raman spectroscopy. Characteristic properties of laser systems for Raman microscopy will be described in this chapter, following an introduction to the technique and its applications in the next section below. Then, the commonly used laser sources for Raman microscopy are categorized according to the wavelength regions of their emissions, along with their merits and drawbacks for each application. The chapter will consider mainly continuous wave laser sources, because pulsed laser beams that are tightly focused with microscope optics give very high irradiance that would destroy most samples.

Raman Microscopy

The first experimental Raman microscopes were developed in the mid-1970s by two independent groups, and it was not long before the first generation of Raman microscopes were subsequently commercialized. Many of these early instruments simply consisted of commercial optical microscopes coupled to Raman spectrometers, as shown in Fig. 1. It was recognized that the use of a microscope for sampling in this way can facilitate the Raman examination of tiny particles of micrometer dimensions.

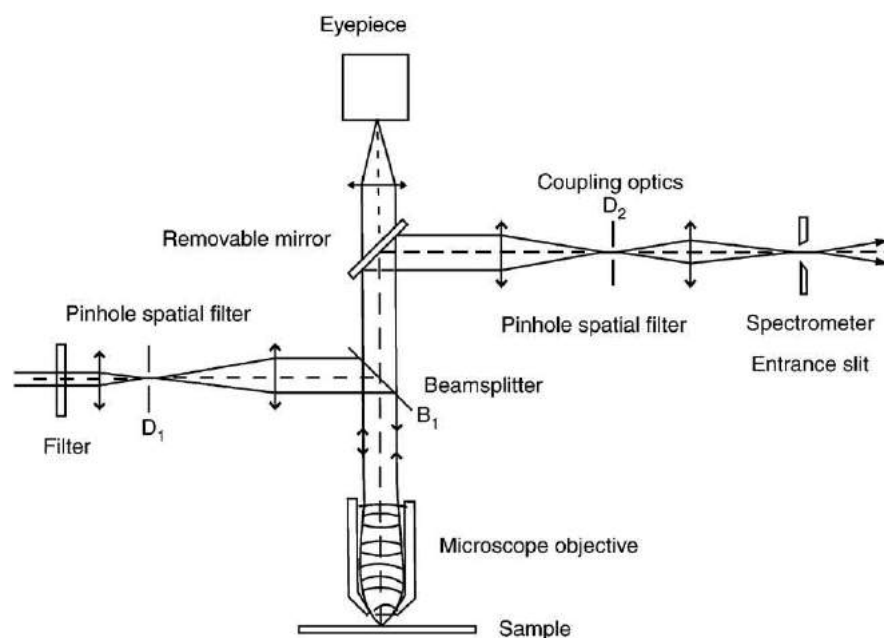


Fig. 1 Diagram of the optical layout of a visible Raman microscope. Reproduced from Turrell G, Delhay M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy. Development and Applications*, 2. London: Academic Press.

The optical lay-out in Fig. 1 shows an infinity corrected microscope. The advantage of using this, rather than a standard tube length microscope, is that its length can be extended in order to incorporate additional optical elements that are required for coupling to the Raman excitation and collection optics. In the optical layout, the incident laser beam is spatially filtered by a pinhole, D_1 , in order to remove the diffraction rings and give a point source. The laser light is then partially reflected by beamsplitter, B_1 , and focused on to the sample by the microscope objective. The back-scattered Raman radiation is collected by the same microscope objective, partially transmitted by the beamsplitter, and directed into the entrance slit of the Raman spectrometer by means of coupling optics. The accurate location of an aperture, D_2 , at the focal point improves the spatial resolution and allows depth profiling of transparent samples. The two pinholes, D_1 and D_2 , act as spatial filters and are referred to as confocal diaphragms. The Raman microscope is described as being confocal, because out-of-focus light, collected from outside the focal volume, is rejected.

The use of the beamsplitter, in the optical path of early Raman microscope designs, gives rise to significant losses of the precious Raman scattered light. If a 50:50 beamsplitter is used then half of the incoming laser light is lost by transmission at B_1 and, more significantly, half of the Raman scattered light is lost by reflection at B_1 . As long as a higher laser power can be employed, it is better to use a 90:10 beamsplitter that transmits 90% of the light and reflects 10%; then, only 10% of the incoming laser light is directed by B_1 on to the sample, but 90% of the Raman scattered light is transmitted by B_1 towards the detector and only 10% of the Raman light is lost by reflection at B_1 . The efficiency of a beamsplitter can be defined as the product of the reflectivity at the laser wavelength and the transmission at the Raman-shifted wavelength. Thus, conventional 50:50 and 90:10 beamsplitters would have efficiencies of 25% and 9%, respectively. Recently, however, holographic notch filters have been developed which can be used as higher efficiency beamsplitters. These typically have a laser rejection contrast ratio of better than 10^4 , reflecting ca. 90% of the laser excitation and transmitting ca. 90% of the Raman scattered light, thus giving an efficiency of ca. 81%. Due to these considerations, higher throughput, commercial Raman microscope designs usually incorporate a holographic notch filter.

The importance of Raman microscopy stems from the fact that it is the only microanalytic method available today, by use of which it is possible to identify, or characterize, small particles of micrometer dimensions *in situ*. It is also advantageous that no sample preparation is required when performing Raman microscopy.

The fundamental limit of the lateral spatial resolution (i.e., for a diffraction limited focus) is the separation at which the maximum of one Airy disc function just touches the first minimum of an adjacent Airy disc, given by:

$$\text{Lateral spatial resolution} = \frac{1.22\lambda}{2NA} \quad (1)$$

where λ is the wavelength of the light and NA is the numerical aperture of the microscope objective.

The axial Airy disc function determines the resolution in the longitudinal direction. A good estimate of the axial resolution limit for low numerical aperture is given by:

$$\text{Axial spatial resolution} = \frac{2\lambda}{(NA)^2} \quad (2)$$

When the sample is heterogeneous and exhibits fluorescence that is not evenly distributed within the sample, a region of the sample can be selected with the microscope that shows the minimum amount of fluorescence. If the fluorescence is intrinsic to the sample itself, then it is possible to use the shift subtract technique and/or temporal discrimination between the fluorescence and the Raman scattering. The latter can be achieved by Kerr gate fluorescence rejection, in which a pulse of light is used to close a Kerr gate shutter. The problem with this approach, however, is that it uses a laser pulse, and even pulses of modest energy will have extremely high irradiance when focused to a tiny spot by a microscope objective. Such laser pulses would inevitably destroy the sample under a microscope.

Another approach for reducing fluorescence is to use near infrared excitation, which is low enough in energy so that absorption cannot occur to promote electronic transitions. The technique of Fourier Transform Raman (FTR) spectroscopy often offers the best chance of obtaining Raman spectra from fluorescent samples for this reason. Low energy excitation of wavelength equal to 1064 nm, provided by a Nd:YAG laser, is normally used for FTR. Disadvantages are that the Raman scattering efficiency is low relative to that obtained with visible excitation, due to the ν^4 dependency (ν^4 refers to Rayleigh scattering of the Raman light). This is compensated to some extent by employing a high throughput (Jacquinot advantage) interferometer. However, this works better for macro rather than micro samples. This is because the coupling of a microscope to an FT Raman spectrometer has a fundamental drawback; the microscope is throughput limited, so the high throughput advantage of the FTR spectrometer cannot be realized in FTR microscopy. The image of the Jacquinot stop (the large circular aperture which is the entrance to the interferometer) at the sample plane typically has a diameter of a few hundred μm , which is much larger than particles with diameters of ca. 1 μm having similar dimensions to the waist diameter of the 1064 nm light excitation at its diffraction-limited focus. Consequently the throughput advantage of the interferometer is not fully exploited and there is a trade-off of spatial resolution with signal-to-noise at the detector. For this reason, the typical spatial resolution that is achieved is in the range of 15–100 μm rather than 4 μm , as claimed in the literature. Bruker has commercialized an FTR microscope by coupling an optical microscope to an FTR spectrometer with near infrared optical fibers.

An approach that offers more promise for reducing fluorescence and achieving spatial resolution close to the diffraction limit is to use near infrared excitation in conjunction with dispersive Raman microscopy. For example, semiconductor lasers operating

from 785 to 852 nm can be used in conjunction with sensitive multichannel silicon-based CCD arrays. If longer excitation wavelengths are necessary, in order to overcome the problems with fluorescence, then diode lasers can be used that emit at longer wavelengths in conjunction with dispersive Raman spectrometers equipped with multichannel near infrared (NIR) detectors, e.g., Ge- or InGaAs-array detectors.

The applications of Raman microscopy cover many areas, including material science, the earth sciences, environmental science, biology and medicine, forensic science, and even the analysis of artworks. These areas are too wide ranging to be described in detail here, but the interested reader is referred to the Further Reading section at the end of this article.

Characteristics of Laser Sources

The laser is an excellent light source for Raman spectroscopy to such an extent that the terms 'Raman spectroscopy' and 'laser Raman spectroscopy' are synonymous for all but the most esoteric experiments, for example, with synchrotron sources. Indeed, the advent of the laser was the stimulation for the renaissance of Raman spectroscopy in the 1960s, given its special properties, such as monochromaticity, high intensity, beam collimation, and coherence.

The characteristics of laser light and the advantages it offers to Raman microscopy are now considered.

Beam Quality

The light inside a laser tube is formed from a number of standing waves having distinct vibrational modes. There are a limited number of these modes transverse to the beam and these are characterized by the TEM_{pq} number (where p and q can be 0, 1, 2, ...) where TEM is an acronym for 'transverse electromagnetic'. When a laser is operating in its fundamental TEM_{00} mode, light rays are reflected on axes between the end mirrors of the laser cavity. The '00' indicates that there are no nodes in the beam profile (Fig. 2(a)), and the laser beam has a Gaussian intensity profile in the radial direction:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \quad (3)$$

where $I(r)$ is the irradiance at a radial distance r from the axis of the beam, I_0 is the axial irradiance, and w is the beam radius.

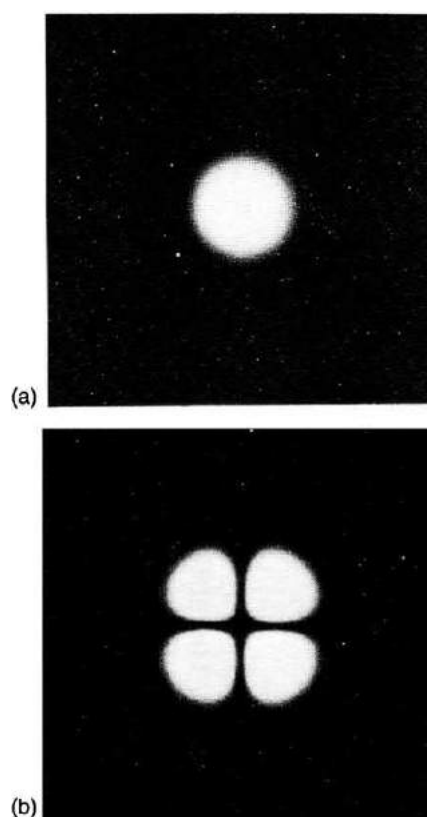


Fig. 2 Transverse electromagnetic modes formed with confocal concave mirrors, (a) TEM_{00} , and (b) TEM_{11} . Reproduced with permission from Young M (1977) *Optics and Lasers: an Engineering Approach*. Berlin, Heidelberg: Springer-Verlag.

For higher-order modes, a number of nodes are observed in the beam profile, which arise from off-axis light rays being reflected between the end mirrors, for example, the TEM₁₁ mode has two nodes which are mutually perpendicular (Fig. 2(b)).

The propagation characteristics of a Gaussian beam can be fully defined, either by the diameter of the beam waist or by the far-field divergence. Consequently, it is only necessary to know the diameter of the beam waist ($2w_0$), or the diameter of the beam ($2w_z$) at a longitudinal distance z from the waist, in order to determine the propagation characteristics of the beam:

$$w(z) = w_0 \sqrt{1 + \left(\frac{z}{z_R}\right)^2} \quad (4)$$

$$R(z) = z \left[1 + \left(\frac{z_R}{z}\right)^2 \right] \quad (5)$$

where the quantity $z_R = \pi w_0^2 / \lambda$ is known as the Rayleigh range of the beam, λ is the wavelength of the laser radiation, and $R(z)$ is the radius of curvature of the wavefront at a distance z from the beam waist.

The wavefront is planar at the minimum beam waist and the Rayleigh range is the distance from the beam waist to the location at which the wavefront is most curved (Fig. 3), the region from the waist to the Rayleigh range being the near field. Beyond approximately ten times the Rayleigh range, in the far field, the beam diverges as a cone with approximately straight sides. It can be seen by substituting $z = z_R$ into Eq. (4) that the beam diameter at the Rayleigh range is $\sqrt{2}$ times the waist diameter.

Unfortunately laser beams do not conform to pure Gaussian functions and therefore they do not propagate according to the above equations. As a result, a dimensionless beam propagation parameter was devised in the early 1970s. This parameter, known as the M^2 factor or the 'times diffraction limit', is based on the brightness theorem, which states that for any laser beam the product of the beam diameter and the far-field divergence is a constant. Thus, M^2 is defined as the ratio of the laser beam's multimode diameter-divergence product to the ideal diffraction-limited (TEM₀₀) beam diameter-divergence product:

$$M^2 = \left(\frac{2w_m \Theta_m}{2w_0 \theta_0} \right) \quad (6)$$

where Θ_m is the laser beam's multimode divergence, θ_0 is the theoretical diffraction-limited divergence, $2w_m$ is the laser beam's multimode waist diameter, and $2w_0$ is the ideal diffraction-limited beam waist diameter.

Alternatively, M^2 can be defined as the square of the ratio of the multimode beam diameter ($2w_m$) to the diffraction-limited beam diameter ($2w_0$):

$$M^2 = \left(\frac{2w_m}{2w_0} \right)^2 \quad (7)$$

The intensity of the beam has a Gaussian distribution in the radial direction, but the accepted definition of beam diameter is the distance at which the intensity has fallen to $1/e^2$ (i.e., ca. 13.5%) of its axial intensity, as can be seen by setting $r = w$ in Eq. (3) above. This definition only applies properly to Gaussian beams; for other beam profiles the diameter can be calculated using a second moment measurement of the entire beam.

The real beam, then, can be treated as Gaussian by substituting an artificial wavelength $M^2 \lambda$ into the equations that apply to an ideal diffraction-limited TEM₀₀ beam. The ideal TEM₀₀ Gaussian can be superimposed as an 'embedded Gaussian' on the optical

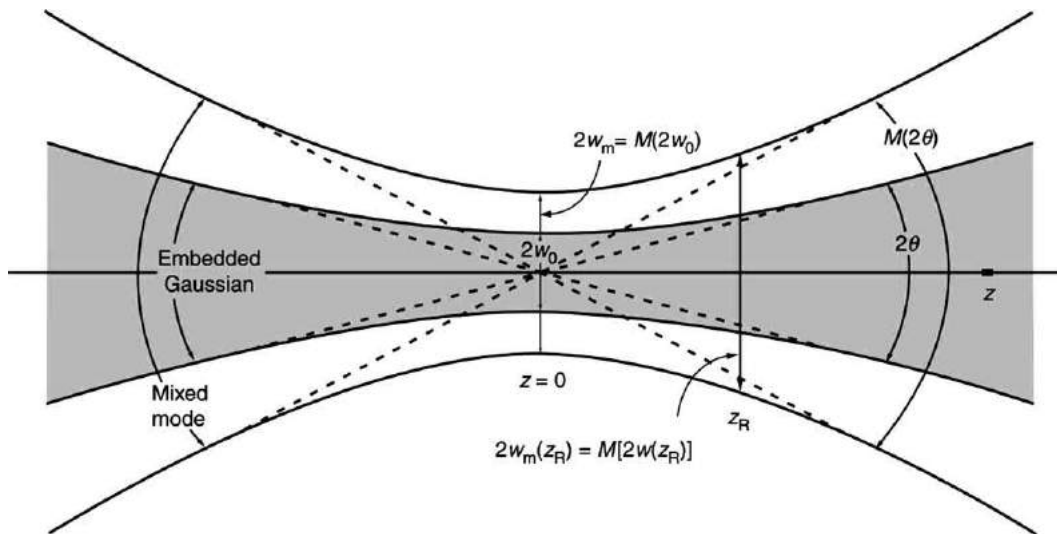


Fig. 3 The Rayleigh range and embedded Gaussian. Adapted from http://beammeasurement.mellesgriot.com/tut_m2.asp.

diagram of the real multimode beam (Fig. 3). It can be seen that the multimode beam has a waist diameter and a divergence that are both M times larger than those of the embedded Gaussian.

The M^2 factor has a value of unity for an ideal diffraction-limited TEM₀₀ beam, and values of greater than unity for all other beams; it can be as high as several hundred for a distorted beam of poor quality. This is an important parameter where a tightly focused laser spot (e.g., confocal Raman microscopy) or low divergence (e.g., remote sensing) are required.

Polarization

Although polarization is not an inherent property of laser light, lasers provide plane polarized light when the end windows of the laser tube are mounted at Brewster's angle (θ_B):

$$\theta_B = \tan^{-1} n \quad (8)$$

where n is the refractive index of the window material for the appropriate wavelength of laser light. For this angle, reflected light is completely plane polarized and consequently the laser light that is generated is also plane polarized. Brewster windows are often used inside the laser cavity in order to reduce reflection losses. This is important in low-gain lasers such as HeNe lasers where laser action could be prevented by reflection losses. Air-cooled ionic gas lasers typically deliver highly linearly polarised output with ratios exceeding 1000:1.

The polarization property of the exciting light is used a great deal in Raman spectroscopy. In combination with the polarization properties of the Raman scattered light, it can be used to determine the depolarization ratios, ρ , of Raman bands in the solution state. In solid state studies it can be used to determine the degree of orientation, for example, in polymeric fibers or films, and in single crystal studies it can be used to determine the symmetry classes of the Raman active vibrations. In such studies a linear polarizer is used to improve the degree of plane polarization of the laser light, and a half-wave plate is used to rotate the plane of polarization by 90° .

Plane polarized laser light can be converted to circularly polarized laser light which is required for Raman optical activity (ROA) studies. For example, plane polarized 514.5 nm laser light provided by an argon ion laser can be converted to circularly polarized light by an electro-optic modulator. As CCD detectors are usually employed in present-day ROA experiments and they have relatively long sampling times, slow polarization modulation is required which can be achieved by periodically rotating a quarter-wave plate or by applying a square quarter-wave voltage to a Pockels cell.

Whereas optical isomers or enantiomers give identical Raman spectra, they are distinguished in the ROA experiment by the sign of the ROA signal. Indeed, the sign can be used to determine the absolute configuration of a chiral molecule. Also the enantiomeric excess of mixtures of enantiomers can be determined in such experiments, with obvious applications in the pharmaceutical industry, and the conformations of biological molecules, for example, proteins, nucleic acids, sugars, and viruses, can be determined.

Up until now, the ROA technique has been practised by no more than a handful of research groups world-wide, but it is expected that this situation will change in the near future, since the first commercial ROA spectrometer (the ChiralRAMAN spectrometer) has recently been launched by Biotools Inc. This spectrometer makes use of a solid state laser as the source of the polarized laser light.

Beam Divergence

Referring to Fig. 3, the full divergence angle, Θ , for the fundamental TEM₀₀ Gaussian beam is given by:

$$\Theta = \lim_{z \rightarrow \infty} \frac{2w(z)}{z} = \frac{2\lambda}{\pi w_0} \quad (9)$$

From Eq. (4) it can be seen that the spot size increases linearly with the longitudinal distance z and, from Eq. (9), that it diverges at a constant cone angle, Θ . It can also be seen from Eq. (9) that the smaller the spot size, w_0 , the greater the beam divergence angle, Θ .

Indeed, a highly divergent beam is a problem with edge-emitting laser diodes due to the small dimensions of the light source. Furthermore, it is much smaller in the vertical direction than in the horizontal, giving rise to an elliptical output beam that diverges much more rapidly in the vertical direction than in the horizontal. The highly divergent, elliptical beam can be corrected, to an extent, with a cylindrical lens, but the inherent problem of a small, elliptical source can never be completely corrected.

In contrast, vertical-cavity surface-emitting semiconductor lasers (VCSELs) do not have the same limitations on beam divergence because the cavity is in the vertical direction and the light emission occurs from the surface, which has a much larger area than the light source of an edge emitter. Indeed, frequency-doubled green and blue semiconductor lasers are now available, having beams with M^2 values of less than 1.1 and beam divergences of a few milliradians. In terms of these beam characteristics, it appears that these lasers are beginning to challenge the argon ion laser.

Emission Linewidth

The laser linewidth limit, $\Delta\nu_L$, is given by the Schawlow–Townes expression:

$$\Delta\nu_L = \frac{2\pi h\nu(\Delta\nu_c)^2}{P_{\text{out}}} \quad (10)$$

where $\Delta\nu_c$ is the linewidth of the passive resonator, $h\nu$ is the photon energy, and P_{out} is the laser power output.

The power output, P_{out} , is equal to the number of photons, N_p , in the resonator times the energy per photon, $h\nu$, divided by the photon lifetime, τ_c :

$$P_{\text{out}} = \frac{N_p h\nu}{\tau_c} \quad (11)$$

Furthermore, the width $\Delta\nu$ of the resonance curve, at which the intensity has fallen off to half the maximum value, is given by:

$$\Delta\nu_c = (2\pi\tau_c)^{-1} \quad (12)$$

Substituting for P_{out} and $\Delta\nu_c$ from Eqs. (8) and (9) into Eq. (10) gives the limit for the linewidth, $\Delta\nu_L$:

$$\Delta\nu_L = \frac{\Delta\nu_c}{N_p} \quad (13)$$

The linewidth of a laser is characterized by its Q factor:

$$Q = \frac{\nu}{\Delta\nu} \quad (14)$$

where ν is the frequency of the laser line and $\Delta\nu$ is the linewidth.

Other useful relationships involving linewidth parameters are:

$$Q = \frac{\lambda}{\Delta\lambda} \quad (15)$$

and

$$\Delta\tilde{\nu} = -\frac{\Delta\lambda}{\lambda^2} \quad (16)$$

where λ and $\Delta\lambda$ are the wavelength and linewidth (in units of metres), respectively, and $\Delta\tilde{\nu}$ is the linewidth expressed as a wavenumber (in units of cm^{-1}).

The Q factor can be as high as 10^8 which is of great use in high resolution spectroscopy.

Visible Lasers

Helium-Neon Laser

The helium-neon laser is continuously pumped electrically using a high dc voltage of up to 1 kV. The gain medium consists of a gas mixture of about 10 parts helium to each part of neon at a pressure of about 3 Torr. The gain of this medium is extremely low, hence the requirement for Brewster-angle windows to eliminate reflection loss of the light polarized in the plane that includes the axis of the laser and the normal to the Brewster window. Due to this requirement, the light output is necessarily plane polarized.

The pumping excites helium atoms by electron impact, and resonant energy transfer to neon atoms then occurs via collisions of the gaseous atoms (Fig. 4). This creates a population inversion in the neon atoms, enabling the laser transition to occur at 632.8 nm. Following emission, the neon atoms decay nonradiatively down to their metastable $2p^53s^1$ level from which they decay back down to the ground state via collisional de-excitation with the walls of the tube. This mechanism restricts the maximum achievable power output to ca. 50 mW, because of the need to depopulate the metastable neon atoms by wall collisions.

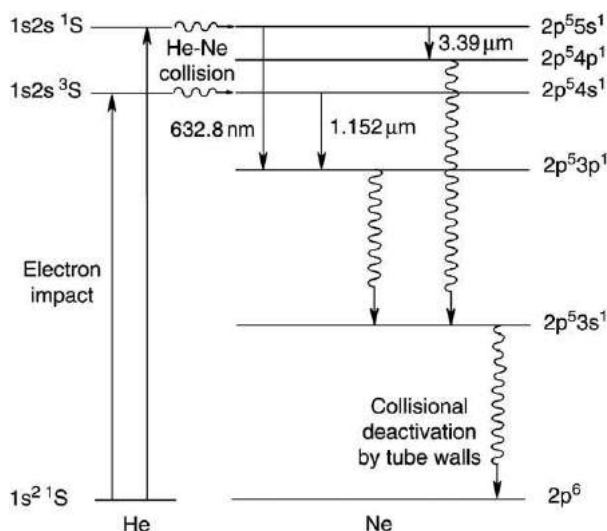


Fig. 4 Energy levels of the He-Ne laser.

An advantage of using helium-neon lasers for Raman spectroscopy is that they generally operate in the fundamental TEM₀₀ mode, which is critically dependent on the ratio of the length of the tube to its diameter. For this reason they are designed to have small tube diameters of around a few millimetres and lengths of 0.15–0.50 m, the small diameter also aiding collisional deactivation with the walls due to the relatively large surface to volume ratio of the tube. Another advantage of this laser for Raman spectroscopy is that the linewidth of the 632.8 nm emission line is ca. 1.5 GHz.

A disadvantage of the helium-neon laser is that it emits a large number of spontaneous emission lines originating from the excited neon atoms. These plasma emissions are observed in the Raman spectrum as sharp lines unless they are filtered out, for example, by a pre-monochromator or an interference filter. It should be mentioned, in passing, that these plasma lines are not always unwanted, because they can be very useful for wavenumber calibration of the Raman spectrometer.

The low output power of < 50 mW is not generally a disadvantage for Raman microscopy since higher laser powers can often cause photolytic or thermal degradation of the sample. This is due to the high irradiance at the sample when the laser light is focused to a tight spot by the microscope objective. The excitation wavelength of 632.8 nm is suitable for combined use of the laser with a silicon-based CCD detector in Raman spectroscopy, because large Stokes shifts in the 3000–4000 cm⁻¹ region lie well within the quantum efficiency curve of this detector. It is also possible, though less common, to operate the helium-neon laser on weaker transitions that include the green 543.5 nm line.

Argon Ion Laser

The argon ion laser was one of the first lasers to be discovered following the invention of the laser and up until now it has been used extensively for Raman spectroscopy, among many other applications. Excitation is provided by a continuous electrical discharge, and because of the high energy required to ionize the argon atoms and then promote the ions to an excited state, the efficiency of the laser is very small (ca. 0.1%). In spite of this, once population inversion has been achieved, the gain of the laser is very high and output powers of up to 25 W can be obtained for the strong lines at 488.0 and 514.5 nm and up to 50 W for multiline operation.

An advantage of the argon ion laser is that it can provide emission at more than 35 discrete wavelengths, the strongest of which are listed in Table 1. These lie in the green, blue, and near ultraviolet regions of the spectrum, a number of the ultraviolet lines only being obtained on the larger frame argon ion lasers. Discrete laser emission lines are selected by tuning a prism or grating inside the cavity. The argon ion laser can also be operated in multiline mode, for example, for pumping dye or Ti:sapphire lasers.

The linewidths of the argon ion emission lines at 488.0 and 514.5 nm are around 4.0 GHz. The high temperature laser tube has a diameter of approximately 12 mm and a length in the range of 0.1 to 1.8 m. A 240 V three phase power supply and ca. 35 A are required for pumping a medium power 5 W argon ion laser. The tube also requires water cooling at a flow rate of ca. 10 L/min and a pressure of 25 psi because of the large amount of heat dissipation.

Table 1 Continuous wave argon ion laser wavelengths and output powers for the Coherent Inc. Innova 300 argon ion laser system

Wavelength (nm)	Power (mW)
228.9	30
238.3	100
244.0	400
248.3	180
257.3	750
275.4	5
300–305.5	20
333.4	40
333.8	30
335.8	20
351.1	200
351.4	60
363.8	240
454.5	140
457.9	420
465.8	180
472.7	240
476.5	720
488.0	1800
496.5	720
501.7	480
514.5	2400
528.7	420

The five lines below 260 nm are frequency doubled.

For Raman microscopy smaller, air-cooled argon ion lasers, which only require a 240 V single phase power supply, can be used to provide output of a few hundred milliwatts on the 488.0 and 514.5 nm lines. For this application, the excitation is provided by argon ion lasers, which are designed to operate in the TEM₀₀ mode.

Until recently, the argon ion laser has been a workhorse as the most common excitation source for Raman spectroscopy, but it is now losing ground to solid state lasers. The principle reasons for this are that the latter are much more efficient and consequently do not in general have three phase power and water-cooling requirements. Nevertheless, argon ion lasers still have a niche when high excitation powers on the order of watts are required in conjunction with a good beam quality ($M^2 < 1.1$), for example, for Raman spectroscopy of gaseous samples. Another advantage of argon ion lasers over solid state lasers is their multiline capability.

Krypton Ion Laser

The krypton ion laser has useful discrete emission lines in the near ultraviolet, blue, yellow, red, and near infrared regions of the electromagnetic spectrum; the strongest lines of a Coherent Inc. Innova 400 krypton ion laser are listed in Table 2.

The argon ion and krypton ion lasers are close relatives, thus the large frame krypton ion laser has similar power and water-cooling requirements to those mentioned above for the argon ion laser. Also, like the argon ion laser, air-cooled models with lower power output are available. The characteristics of the two lasers are also similar; for example, the linewidths of the krypton ion transitions at 530.9, 568.2, and 647.1 nm are about 4.0 GHz, the laser tube has similar dimensions (0.1–2.0 m in length) and the laser can be operated in either TEM₀₀ or multimode.

The krypton ion laser is even less efficient than the argon ion laser, consequently lower-power outputs of up to about 20 W can be achieved when operated in multiline mode.

Mixed argon and krypton ion laser tubes are also commonly used that provide laser lines originating from both argon ion and krypton ion transitions. As with the individual argon ion and krypton ion lasers, sharp plasma lines, due to spontaneous emission from the rare gas ions, are observed in the Raman spectrum unless they are filtered out. If interference filters are used for this, each laser line requires its own filter that is tailored to the specific emission wavelength.

HeCd Laser

Laser lines at 325 and 442 nm, both capable of providing milliwatts of output power, can be obtained from the helium-cadmium laser. These emission lines result from electronic transitions in free cadmium atoms.

It is obviously advantageous to use blue, rather than longer wavelength, excitation (e.g., provided by the 442 nm line of the HeCd laser) for off-resonance Raman spectroscopy, due to the ν^4 dependence of the Raman light scattering efficiency, provided that the efficiencies of the illumination/collection optics and Raman spectrometer throughput are optimized in the blue region of the spectrum. Unfortunately, far more samples fluoresce when excited with blue light than with red or near infrared radiation, which explains why red or near infrared lasers (e.g., HeNe, semiconductor lasers) are more commonly used for general Raman applications.

Table 2 Continuous wave krypton ion laser wavelengths and output powers for the Coherent Inc. Innova 400 krypton ion laser system

<i>Wavelength (nm)</i>	<i>Power (mW)</i>
206.5	4
234	8
337.5–356.4	2000
406.7	900
413.1	1800
415.4	280
468.0	500
476.2	400
482.5	400
520.8	700
530.9	1500
568.2	1100
647.1	3500
676.4	900
720.8	45
752.5	1200
793.1–799.3	300

The 206.5 and 234 nm lines are frequency doubled.

Near Infrared Lasers

Although the traditional laser systems for Raman spectroscopy have been the argon ion, krypton ion, and helium-neon lasers for discrete excitation wavelengths, the diode laser has gained popularity over recent years.

The principal advantage of near infrared semiconductor lasers for Raman microscopy is that they generally excite less fluorescence in the Raman spectrum than visible lasers due to their longer wavelength. Commonly available wavelengths are 670, 785, 830, and 852 nm and, of these, the first two can be used in conjunction with silicon-based CCD detectors to give Stokes Raman shifts over the whole range of fundamental vibrations, up to 4000 cm^{-1} . However, for excitation wavelengths of 830 and 852 nm, only Stokes Raman bands in the fingerprint region can be detected with a silicon-based CCD detector because higher wavenumber Raman shifts approach the bandgap of the silicon semiconductor ($\lambda > \text{ca. } 1050\text{ nm}$).

The disadvantages of diode lasers are that they cannot supply high power of narrow linewidth in single-mode and they are susceptible to mode hopping which gives rise to a shift in the excitation wavelength. This latter drawback is particularly disadvantageous for Raman spectroscopy because it results in a shift in the wavenumber positions of the Raman bands. When good beam quality is not required, broad stripe, high power laser diodes can be used; these can have output powers that are greater than 1 W but their emission linewidths are equal to ca. 2 nm, which is far too wide for Raman spectroscopic applications. These laser diodes find applications as pumps of solid state lasers, for example, Nd:YAG, Nd:YVO₄ lasers, however. Additionally, amplified stimulated emission (ASE) gives an unwanted background signal that can be very broad (ca. 20–30 nm) and 0.1–1% of the intensity of the laser line. This necessitates the use of bandpass filters in order to reduce this unwanted background.

Although laser diodes having an output of less than 200 mW can operate in TEM₀₀ and in single longitudinal modes, mode hopping can occur due to optical feedback or fluctuations in environmental factors and this causes severe broadening of the linewidth. The sensitivity to optical feedback can be greatly reduced, hence the laser linewidth can be narrowed appreciably, by confining the frequency of the photons either in an internal cavity containing a small diffraction grating or in an external cavity. The former types of laser are sometimes called 'distributed feedback' (DFB) lasers because the feedback is distributed over the length of the grating, rather than occurring all at once at a mirror. The wavelength that is fed back is determined by the period of the grating. Usually, a DFB laser has a grating fabricated into the entire length of the laser. A variation referred to as a distributed Bragg reflector (DBR) laser has a distinct grating fabricated into the substrate on each side of the active area. The external cavity semiconductor laser (ECSL) is fabricated by placing the laser diode in a separate resonator like a conventional gas or solid-state laser. The DBR and ECSL lasers are now discussed.

Distributed Bragg Reflector Lasers

The DBR laser consists of a grating on each side of the active region (Fig. 5); these gratings act as mirrors having a reflectivity that is optimized at one particular wavelength, in addition to narrowing the laser linewidth. At present, DBR lasers are only commercially available having an excitation wavelength of 852 nm.

The emission linewidth is less than 4 MHz, which is suitable for the vast majority of Raman spectroscopic applications and the laser operates in single TEM₀₀ mode. The disadvantages of DBR lasers are that the maximum output power is restricted to around 150 mW, so an optical isolator is usually necessary in order to prevent external facet damage caused by external optical feedback, and only an excitation wavelength of 852 nm is currently available.

External-Cavity Semiconductor Lasers (ECSL)

ECSLs provide higher output power (up to ca. 1 W) and a wider range of excitation wavelengths (630 nm to around 850 nm), as well as being less expensive than DBR lasers. They use the semiconductor chip only as the gain medium and employ an external grating, both as the frequency selector and the reflective mirror. Specific excitation wavelengths are becoming widely adopted in Raman microscopy employing single grating spectrograph in conjunction with CCD detection, such as 670, 785, 830, and 852 nm. The ECSLs have emission linewidths as low as a few MHz, but in general they span the range 2 MHz–30 GHz, and can operate in a nearly diffraction-limited transverse mode.

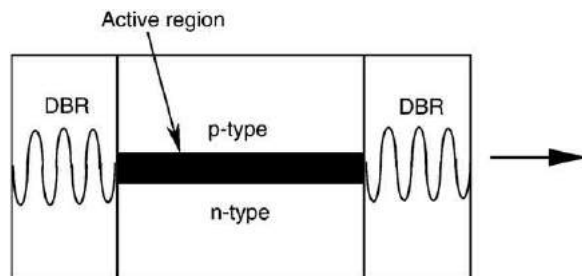


Fig. 5 Diagram of the cavity of the DBR laser. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

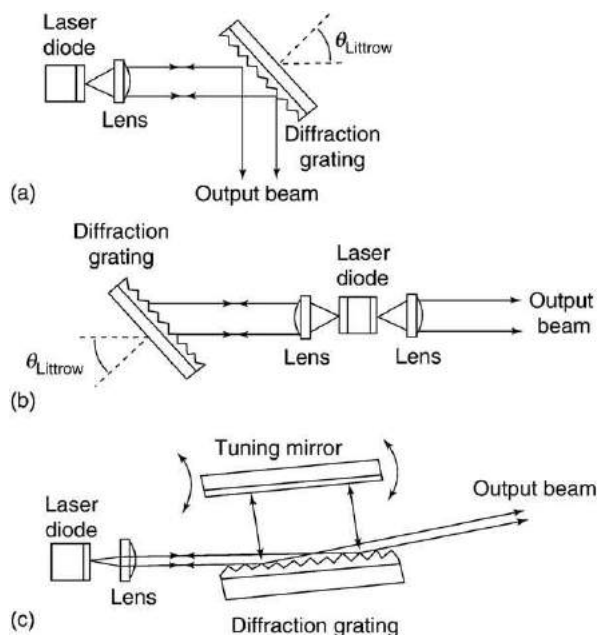


Fig. 6 Diagrams of the cavities of (a) Littrow type-I, (b) Littrow type-II, and (c) Littman external cavity diode lasers. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

Three different cavity designs employing a diffraction grating have been discussed in the literature: Littrow type-I, Littrow type-II, and Littman configurations. In the Littrow type-I design, the laser diode light is incident on the grating at an angle of incidence equal to θ_{Littrow} (Fig. 6(a)). The diffracted light in first order is re-directed back to the laser diode to provide feedback and the output is the zeroth order (i.e., specularly reflected) light. A disadvantage of this design is that the bandwidth of the grating is relatively large because a single pass geometry is used. In the Littrow type-II design, light from the rear facet of the laser diode is collimated by a lens and directed on to a grating at an angle of incidence equal to θ_{Littrow} (Fig. 6(b)). The diffracted light in first order is re-directed back to the laser diode, in a similar fashion to the Littrow type-I design, in order to provide optical feedback, and the output is the laser light emitted from the front facet. Disadvantages of the Littrow type-II design are the requirement for custom fabrication of antireflection coatings on both facets and the sensitivity to external optical feedback. In the Littman design, the laser diode light is collimated by a lens and directed on to a grating at a grazing angle of incidence. The diffracted light is reflected by a mirror and diffracted back to the laser diode by the grating in order to provide optical feedback (Fig. 6(c)). An advantage of this design is that the grating bandwidth is less than half that of the Littrow designs due to the double pass geometry, but a disadvantage is that the external cavity length is longer, resulting in a narrower spacing of the cavity modes. Another advantage of this design is that the tuning is accomplished by rotating the mirror instead of the grating, thus the alignment of the output beam is not altered when tuning. For Littrow type I and II and Littman ECSL designs, filtering of the ASE is required, as it can have an intensity of ca. 0.1 to 1% of the intensity of the laser line. Thus, these designs necessitate the incorporation of a bandpass filter (having a rejection of better than 10^{-4} at the ASE).

Nd:YAG Laser

Under normal conditions, the Nd:YAG laser oscillates on a transition in the near infrared at 1064 nm, and this is the excitation line that is used in most commercial Fourier Transform Raman spectrometers. The gain medium is a crystal of yttrium aluminium garnet ($\text{Y}_3\text{Al}_5\text{O}_{12}$, YAG) doped with ca. 1.0 mol% Nd^{3+} cations that substitute Y^{3+} ions in the cubic YAG lattice.

The Nd:YAG laser is a four-level system (Fig. 7), which has high gain and low threshold due to the narrow fluorescent linewidth of the laser transition. Absorption bands of the Nd^{3+} ions around 808 nm conveniently match the energy of commercially available high-power multimode diode lasers that serve as the pump. The Nd^{3+} ions decay nonradiatively to the upper laser level, the $^4\text{F}_{3/2}$ state, thereby creating a population inversion. This is because the lower laser level, the $^4\text{I}_{11/2}$ state, has no appreciable thermal population at room temperature, since it is $>2000\text{ cm}^{-1}$ above the $^4\text{I}_{9/2}$ ground state, and the $^4\text{F}_{3/2}$ excited state has a relatively long lifetime of 230 μs .

The high thermal conductivity of Nd:YAG enables the laser to operate at high power in either continuous wave (CW) or Q-switched modes, and the diode pumped solid state (DPSS) variety of the cw Nd:YAG laser can currently achieve an output power in excess of 20 W on the 1064 nm line. The laser can be designed to operate in the fundamental TEM_{00} mode and the full width at half maximum (FWHM) linewidth of the spontaneous emission of the 1064 nm laser transition is 120 GHz (ca. 0.45 nm). Advantages of the diode pumped Nd:YAG laser are that it is air-cooled and can be operated at 240 V single phase.

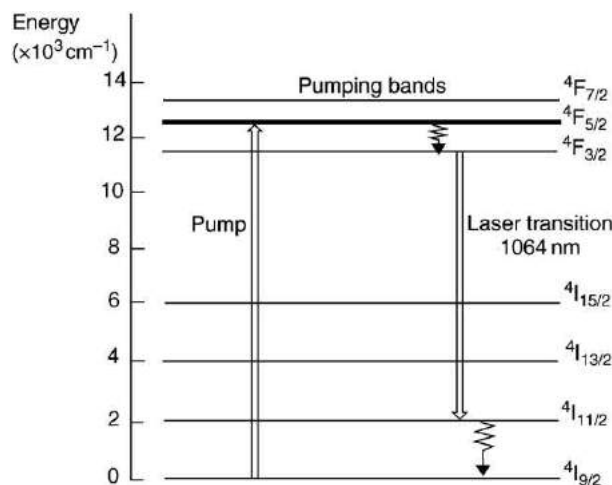


Fig. 7 Energy levels of the Nd:YAG laser.

Also, being all solid state and having a small footprint, it is robust and portable. The lifetime of the laser is dependent on the laser diodes used for pumping, but some Nd:YAG laser designs enable these to be replaced in the field.

The frequency-doubled Nd:YAG laser emitting green light, having a wavelength of 532 nm, is also commonly used nowadays for dispersive Raman spectroscopy, and an output power of 10 W on the 532 nm laser line is provided by commercially available frequency-doubled diode-pumped Nd:YAG lasers.

It is worth mentioning that Nd:YAG lasers have also been fabricated that emit either at 946 or 1330 nm, by suppressing the strong emission at 1064 nm and optimising the optics for the desired wavelength.

Nd:YVO₄ and Nd:YLF Lasers

An intracavity, frequency-doubled DPSS Nd:YVO₄ laser has recently become commercially available, and it has some advantages over the Nd:YAG laser including a larger stimulated emission cross-section and a higher absorption coefficient (along the extraordinary direction of the birefringent crystal). It has the same emission wavelength as the frequency-doubled Nd:YAG laser, i.e., 532 nm, and Nd:YVO₄ is the material of choice for cw end-pumped lasers having around 5 W output power. This is because the diode laser pump beam is tightly focused in the end-pumped system, but a small waist diameter cannot be retained over a distance of more than a few millimeters, consequently a high absorption coefficient and gain are very beneficial.

The gain medium of the Nd:YLF laser consists of a crystal of yttrium lithium fluoride (YLF) that is doped with Nd³⁺ ions on the Y³⁺ cation sites. Unlike the Nd:YAG and Nd:YVO₄ lasers, the emission does not occur at 1064 nm; instead the ⁴F_{3/2}–⁴I_{11/2} emission occurs at wavelengths of either 1047 or 1053 nm, depending on the polarization that is selected. The former is due to extraordinary polarized light, whereas the latter is due to ordinary polarization, and either of these emission wavelengths can be selected using an intracavity polarizer. The Nd:YLF laser can offer benefits in Q-switched operation when the longer fluorescence lifetime (480 μs) of Nd³⁺ ions in the ⁴F_{3/2} state enables a higher energy to be stored for the same number of pump laser diodes.

Ti:Sapphire Laser

The Ti:sapphire laser is tunable over the approximate range of 670–1070 nm with the peak of the gain curve at ca. 800 nm. In the gain medium, ca. 0.1% by weight Ti³⁺ is doped into a crystal of sapphire grown by the Czochralski method. The Ti³⁺ ions substitute for Al³⁺ ions in the Al₂O₃ of the sapphire and the laser emission is due to the ²E–²T₂ transition of the Ti³⁺ cation, which has a 3d¹ valence electronic configuration (Fig. 8). The laser has a large stimulated emission cross-section but the fluorescence lifetime of the upper laser level (²E state) is quite short (3.2 μs), thus the laser is usually laser (e.g., by argon ion or frequency-doubled Nd:YAG or Nd:YVO₄ lasers) rather than flashlamp pumped because a very high pump flux is required.

The absorption and emission bands are broad and widely separated, due to the vibronic coupling between the Ti³⁺ host and the Al₂O₃ lattice (Fig. 9). The lower laser level is any one of the vibronic levels of the ²T₂ state. Following the laser transition, the Ti³⁺ ions decay from the upper vibronic levels of the ²T₂ state down to the lower vibronic levels. Hence, this is a four-level laser.

The broad spontaneous emission linewidth of the ²E–²T₂ laser transition is around 100 THz, the output power can be as high as 50 W, and the laser can be operated in either TEM₀₀ or multimodes.

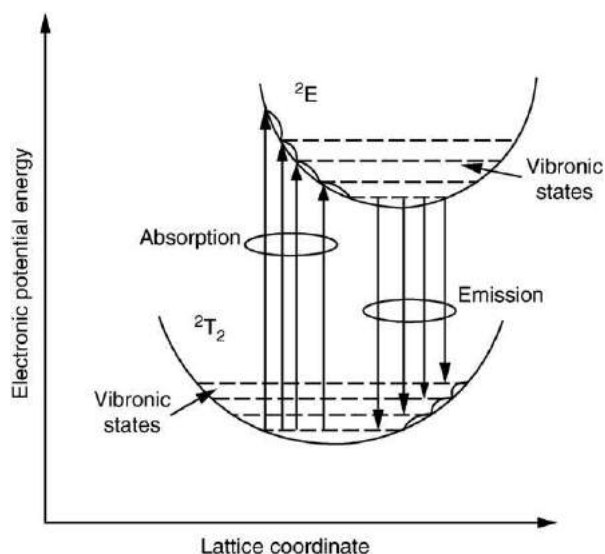


Fig. 8 Energy levels of the Ti:sapphire laser. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

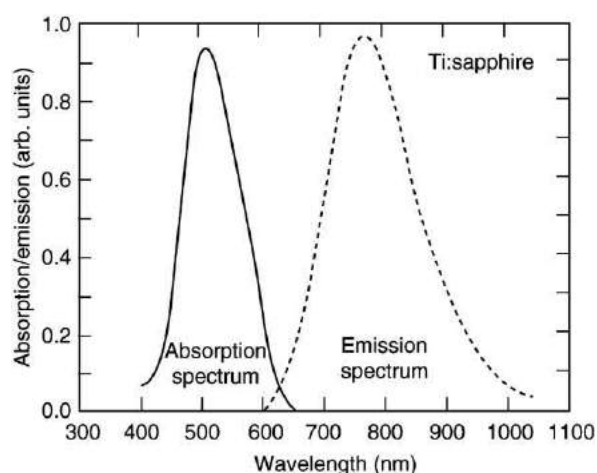


Fig. 9 Absorption and emission spectra of the Ti^{3+} ion in sapphire. Al_2O_3 . Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

UV Lasers

UV laser sources offer numerous advantages over visible laser sources for Raman spectroscopy. A major consideration is that many analytes have absorptions in the near UV, making them amenable to resonance Raman spectroscopic studies. The signal enhancement (up to 10^6), that can be achieved in resonance Raman spectra, results in a large increase in detection sensitivity. Furthermore, some vibrational modes, which normally give rise to weak bands in the off resonance Raman spectrum, can show strong enhancement in the UV excited resonance Raman spectrum. A good example of this is the amide II band of peptides and proteins, which is due to a combination of N–H in plane bending and C–N stretching modes of the peptide linkage. This band is normally weak in the Raman spectrum but strong in the infrared spectrum; however, the amide II band can show strong enhancement in the UV excited resonance Raman spectrum and this can be useful for determining secondary structure in proteins. A further benefit of UV excited Raman spectroscopy is that fluorescence can often be avoided as it tends to occur at a lower energy, outside the Stokes Raman spectral window.

It can be advantageous to use cw rather than pulsed excitation for UV resonance Raman spectroscopy, if the analyte has a long excited state lifetime relative to the pulsewidth of the pulsed laser. This is because the concentration of analyte molecules in the ground state is significantly depleted during pulsed excitation, due to Raman saturation giving a lower signal to noise ratio in the Raman spectrum than for the case of cw excitation. It has been found that pulse energy flux densities should be less than

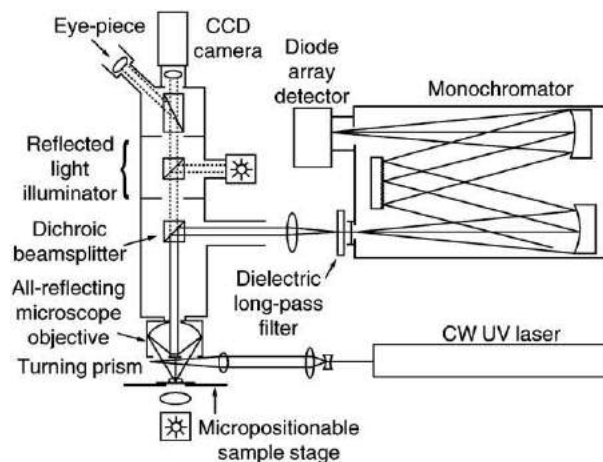


Fig. 10 Diagram of the optical layout of a UV Raman microscope. Adapted from Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86.

1 mJ/cm² and flowing or spinning samples should be used, in order to ensure that nonlinear phenomena and saturation effects do not occur.

For UV Raman microscopy, where the laser beam is focused into a small spot of micrometer dimensions on the sample, it is essential to avoid using pulsed lasers delivering high peak powers. This is because such pulsed lasers can cause dielectric breakdown, and even at lower peak powers they can cause nonlinear effects and Raman saturation phenomena. Consequently, with UV Raman microscopy, one is restricted to the use of cw or quasicontinuous laser excitation.

It is only recently that high throughput UV Raman microspectrometers have been developed. It has been demonstrated for visible Raman spectroscopy that a significantly higher throughput can be achieved by employing a single stage spectrograph with a notch filter instead of a double or triple monochromator. In the UV region, blocking the Rayleigh scattering is a problem because notch filters are not currently available. However, it has been discovered that this can be overcome by introducing two modifications to the design of the optical layout of a visible Raman microscope. First, two novel dielectric longpass filters were used instead of a notch filter in order to reject the elastically scattered light. Second, an all-reflecting Cassegrain microscope objective was used, instead of a lens, in order to block the specular reflection, and thereby further reduce the Rayleigh scattering background. These design modifications have enabled a single stage spectrograph with a holographic grating to be used to disperse the Stokes Raman radiation in a UV Raman microspectrometer (Fig. 10).

In the optical layout of Fig. 10 it can be seen that the laser excitation is not focused by the Cassegrain microscope objective because an epi-illumination configuration is not employed. Instead the laser is focused by a separate lens and directed on to the sample with a turning prism located directly underneath the Cassegrain objective. This minimizes loss of laser beam throughput or scattered light throughput to the spectrometer at the expense of spatial resolution, since the laser light is focused by a longer focal length lens than the microscope objective. An additional difference between the optical layouts of the visible and UV Raman microscopes shown in Figs. 1 and 10, respectively, is that pinhole spatial filters are not shown in the latter. However, better axial spatial resolution could be achieved if this UV Raman microscope was made confocal by introducing a pinhole and lens.

Frequency-Doubled Argon Ion Laser

The frequency doubled argon ion laser is a popular choice among UV laser sources. The cw UV laser contains a nonlinear optical beta-barium borate (BBO) crystal, which is located within the laser cavity; this crystal frequency doubles the strong Ar⁺ lines to give five UV lines below 260 nm which are listed in Table 1 for the Coherent Inc. Innova 300 Ar⁺ ion laser system. Of these UV lines, the 228.9 nm line is almost ideal for resonance Raman enhancement of tyrosine and tryptophan residues in proteins and the 244.0 nm line is well suited for studying tyrosinate groups. The 244.0 nm line has also been used to excite selectively UV resonance Raman spectra from spatially resolved areas of biological samples within the nucleus of a single cell, from DNA in particular, using low laser powers and short acquisition times to keep sample damage to a minimum under the microscope.

Frequency-Doubled Krypton Ion Laser

Like the cw frequency-doubled argon ion laser mentioned above, the cw frequency-doubled krypton ion laser contains a nonlinear optical BBO crystal, which is located within the laser cavity; this crystal frequency doubles the strong Kr⁺ lines to give 206.5 and 234 nm UV lines which are listed in Table 2 for the Coherent Inc. Innova 400 Kr⁺ ion laser system. The 206.5 nm laser line is useful for exciting resonance Raman spectra within the π - π^* amide transition of the peptide backbone of proteins, and the enhanced protein amide vibrational bands can be used to determine the protein secondary structure.

HeCd Laser

The 325 nm HeCd excitation line, in the near ultraviolet, can be used for combined micro-Raman/photoluminescence studies of, for example, semiconductors. These lasers are suitable for low-power applications, typically having output powers in the 1–100 mW range, and they can be designed to operate in TEM₀₀ single-mode or multimode.

Quasi-CW Mode-Locked Ti:Sapphire Laser

The second harmonic of the mode-locked Ti:sapphire laser (ca. 350–500 nm), third harmonic (ca. 233–333 nm), and fourth harmonic (ca. 200–250 nm), are all available with conventional nonlinear crystals.

Typically, the Ti:sapphire crystal is pumped by a cw argon ion laser, for example, as the excitation source for a UV Raman microscope. This quasicontinuous laser system produces 2–3 ps pulses at a 76 MHz repetition rate and is continuously tunable over the 200–300 nm range. The typical power output is ca. 50 mW at 250 nm, ca. 10 mW at 240 nm, and 1 mW at 230 nm. In future, it is anticipated that a frequency-doubled Nd:YAG laser will be increasingly favored as the pump, creating an all-solid-state laser source.

Conclusions

Laser systems for Raman microscopy have been described in this article. Although the differences in the terms 'Raman spectroscopy' and 'Raman microscopy' only refer to whether a macro or micro sampling configuration is used, this does have a bearing on the types of laser systems that are employed. This is because the latter imposes a restriction on using low duty pulsed lasers with high pulse peak power, as both spatial and temporal confinement of the laser excitation can lead to dielectric breakdown of the sample, nonlinear phenomena, or Raman saturation. Thus, for this reason, one is limited to cw lasers or quasi-cw lasers for Raman microscopy whereas there is no such restriction for Raman spectroscopy.

The quality of the laser beam, measured by the M^2 value or 'times diffraction limit' influences the ultimate spatial resolution that can be achieved in Raman microscopy.

In the visible region, argon ion, krypton ion, and frequency-doubled Nd:YAG lasers have good beam quality and high power, with helium-neon and helium-cadmium lasers also having good beam quality but lower power; this makes these lasers good light sources for visible Raman microscopy.

For FT Raman microscopy using Nd:YAG excitation, there is a lack of fluorescence interference from a wide variety of samples, but sensitivity and spatial resolution are compromised in comparison to visible dispersive Raman microscopy.

Red and near infrared semiconductor lasers operating in TEM₀₀ single mode can have good beam quality, approaching that of gas lasers, but they are power limited. Obviously, when good beam quality is not necessary, for example, for optically pumping other laser sources, high-power diode lasers can be used in multimode operation. Solid state red and near infrared laser sources are becoming very popular for routine Raman microscopic analysis due to their compactness, robustness, and economical power consumption.

Another reason for the popularity of the red and near infrared diode laser sources for Raman microscopy is that fluorescence is less of a problem due to the lower energy of the light, compared to conventional green 514.5 nm excitation. Although there is a penalty, due to the dependency of the scattering efficiency on the ν^4 term, present day silicon-based CCD arrays can exhibit excellent detection quantum efficiency of Stokes shifted Raman radiation, across the whole vibrational spectrum, when using standard 670 and 785 nm diode lasers and, at least across the fingerprint region, when using 830 and 852 nm diode lasers.

For ultimate spatial resolution in the Raman microscopic experiment, short wavelength excitation is advantageous because it is fundamentally possible to focus to a smaller diffraction-limited laser spot. In this regard, it would be preferable to use UV Raman microscopy where fluorescence is also not as problematic as it is in the visible region. However, in spite of recent improvements in UV laser sources such as the intracavity frequency-doubled argon and krypton ion cw lasers, there are current instrumental limitations on the throughput efficiency of the UV Raman microscope imposed by the lack of availability of suitable notch filters and low stray light and high throughput optics.

It is anticipated that in the future all-solid-state, quasi-cw, diode pumped, mode-locked Ti:sapphire lasers operating on their third or fourth harmonics will become increasingly popular for providing tunable UV excitation affording good beam quality. It is now also possible to fabricate hollow cathode metal ion deep UV lasers, 10–15 cm in length, 2–4 cm in diameter, weighing 50–100 g and requiring only 2–3 W of electrical power. These lasers have low beam quality (M^2 equal to ca. 18), but they make possible the development of portable UV fluorescence imaging and Raman microprobes for geobiological exploration of terrestrial and extraterrestrial environments.

Further Reading

Andrews, D.L., 1990. *Lasers in Chemistry*, 2nd edition Berlin, Heidelberg: Springer-Verlag.

Asher, S.A., Bormett, R.W., Chen, X.G., *et al.*, 1993. UV resonance Raman spectroscopy using a new cw laser source: Convenience and experimental simplicity. *Applied Spectroscopy* 47, 628–633.

- Bell, S.E.J., Bourguignon, E.S.O., O'Grady, A., Villaumie, J., Dennis, A.C., 2002. Extracting Raman spectra from highly fluorescent samples with "Scissors" (SSRS, Shifted-Subtracted Raman Spectroscopy). *Spectroscopy Europe* 14, 17.
- Best, S.P., Clark, R.J.H., Withnall, R., 1992. Non-destructive pigment analysis of artefacts by Raman microscopy. *Endeavour* 16, 66.
- Boustlary, N.N., Manoharan, R., Dasari, R.R., Feld, M.S., 2000. Ultraviolet resonance Raman spectroscopy of bulk and microscopic human colon tissue. *Applied Spectroscopy* 54, 24–30.
- Delhaye, M., Barbillat, J., Aubard, J., Bridoux, M., Da Silva, E., 1996. Instrumentation. In: Turrell, G., Corset, J. (Eds.), *Raman Microscopy: Developments and Applications*. London: Academic Press. chap. 3.
- Delhaye, M., Dhamelincourt, P., 1975. Raman microprobe and microscope with laser excitation. *Journal of Raman Spectroscopy* 3, 33.
- Derbyshire, A., Withnall, R., 1999. Pigment analysis of portrait miniatures using Raman microscopy. *Journal of Raman Spectroscopy* 30, 185.
- Gordeyev, S.A., Nikolaeva, G.Y., Prokhorov, K.A., Withnall, R., Dunkin, I.R., Shilton, S.J., 2001. Super-selective polysulfone hollow fiber membranes for gas separation: assessment of molecular orientation by Raman spectroscopy. *Laser Physics* 11, 82–85.
- Hayward, C.L., Best, S.P., Clark, R.J.H., Ross, N.L., Withnall, R., 1994. Polarised single crystal Raman spectroscopy of sinhalite, MgAlBO_4 . *Spectrochimica Acta* 50A, 1287–1294.
- Hendra, P., Jones, C., Warnes, G., 1991. *Fourier Transform Raman Spectroscopy*. London: Ellis Horwood.
- Hitz, B., 2003. Solid state lasers are gunning for argon ion's place. *Photonics Spectra*. 54–58.
- Holtz, J.S.W., Bormett, R.W., Chi, Z., *et al.*, 1996. Applications of a new 206.5 nm continuous wave laser source: UV Raman determination of protein secondary structure and CVD diamond material properties. *Applied Spectroscopy* 50, 1459–1468.
- <http://www.btools.com>.
- Koechner, W., Bass, M., 2003. *Solid-State Lasers*. New York: Springer-Verlag.
- Landsberg, G., Mandelstam, L., 1928. Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen. *Naturwiss* 16, 557.
- Matousek, P., Towrie, M., Stanley, A., Parker, A.W., 1999. Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. *Applied Spectroscopy* 53, 1485.
- Nakashima, S., Okumura, H., Yamamoto, T., Shimidzu, R., 2004. Deep-ultraviolet Raman microspectroscopy: Characterization of wide-gap semiconductors. *Applied Spectroscopy* 58, 224–229.
- Pajcini, V., Munro, C.H., Bormett, R.W., Witkowski, R.E., Asher, S.A., 1997. UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51, 81–86.
- Pallister, D.M., Morris, M.D., 1993. In: Morris, M.D. (Ed.), *Microscopic and Spectroscopic Imaging of the Chemical State*. New York: Marcel Dekker. chap. 1.
- Pan, M.-W., Benner, R.E., Johnson, C.W., Smith, L.M., 2000. Near-IR lasers rejuvenate Raman spectroscopy. *Optoelectronics World*. S5–S10.
- Pan, M.-W., Benner, R.E., Smith, L.M., 2002. Continuous lasers for Raman spectrometry. In: Chalmers, J.M., Griffiths, P.R. (Eds.), *Handbook of Vibrational Spectroscopy*. Chichester, UK: Wiley.
- Raman, C.V., Krishnan, K.S., 1928. A new type of secondary radiation. *Nature* 121, 501.
- Rosasco, G.J., 1980. Raman microprobe spectroscopy. In: Clark, R.J.H., Hester, R.E. (Eds.), *Advances in Infrared and Raman Spectroscopy*, vol. 7. Chichester: Wiley.
- Rosasco, J., Etz, E.S., Cassatt, W.A., 1975. The analysis of discrete fine particles by Raman spectroscopy. *Applied Spectroscopy* 29, 396.
- Schawlow, A.L., Townes, C.H., 1958. Infrared and optical masers. *Physics Review* 112, 1940.
- Smekal, A., 1923. Sur Quantentheorie der Dispersion. *Naturwiss* 11, 873.
- Storrie-Lombardi, M.C., Hug, W.F., McDonald, G.D., Tsapin, A.I., Neelson, K.H., 2001. Hollow cathode ion lasers for deep ultraviolet Raman spectroscopy and fluorescence imaging. *Review of Scientific Instruments* 72, 4452–4459.
- Turrell, G., Delhaye, M., Dhamelincourt, P., 1996. Characteristics of Raman microscopy. In: Turrell, G., Corset, J. (Eds.), *Raman Microscopy: Development and Applications* 2. London: Academic Press.
- Young, M., 1977. *Optics and Laser: An Engineering Approach*. Berlin, Heidelberg: Springer-Verlag.
- Yu, Y.-M., Nam, S., Byungsung, O., *et al.*, 2002. Resonant Raman scattering in ZnS epilayers. *Materials Chemistry and Physics* 78, 149–153.

Spatial Heterodyne

Mark F Spencer, Albuquerque, NM, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Many optics and photonics applications benefit from the use of spatial heterodyne, which the open literature often refers to as spatial-heterodyne interferometry, spatial-heterodyne detection, digital-holographic detection, digital holography, or coherent detection. This is said because spatial heterodyne provides us with an estimate of the complex field (i.e., both the amplitude and wrapped phase). As such, this encyclopedia article reviews the fundamentals of spatial heterodyne, so that future research efforts can benefit from the material presented within.

Spatial heterodyne, in practice, results from the interference between a signal and a reference. As described in Fig. 1, we can create the signal using the active illumination of an object. Note that the light for this active illumination comes from a master-oscillator (MO) laser. Also note that the surface of the object is often considered to be optically rough compared with the wavelength of the incident light; thus, the receiving aperture of an imaging system collects the resultant speckle (i.e., the constructive and destructive interference which results from the laser-object interaction and propagation to the receiving aperture). To make it all work, we need to then interfere the (signal) light from the speckle with the (reference) light from a local oscillator (LO). This LO, in general, is split off from the MO laser. We can then record the resultant interference pattern or hologram with a focal-plane array (FPA) and digitize it for digital-signal processing. It is this ability to perform digital-signal processing which makes spatial heterodyne practical and such a versatile tool within the optics and photonics research communities.

To ensure that the signal interferes with the reference, it is advantageous to use a MO laser with a coherence length that is greater than twice the distance to the object. For tactical applications, however, this specification becomes difficult to achieve. We can then attempt to path-length match the reference with the signal, for example, using fiber-optic relays and continuous-wave illumination with a coherence length that is greater than twice the depth of the object (McManamon, 2015). With that said, if path-length matching is impractical (and it often is), we can then use pulsed illumination and incur the efficiency losses associated with decreased fringe visibility in our digital holograms. In this regime, it is desirable (although not required) to use a MO laser with a coherence length that is greater than the pulse length (i.e., use transform-limited pulses).

Doppler shifts, fluorescence, and depolarization from the laser-object interaction, in addition to mismatches (in the reference-signal spatial overlap, pulse timing, etc.), extinction, and noise are other sources for efficiency losses that, in practice, degrade fringe visibility. With this in mind, this encyclopedia article uses a scalar formulation (and the assumptions therein) to study three separate recording geometries often used with spatial-heterodyne applications. As described in Fig. 1, these recording geometries include the off-axis image plane recording geometry (IPRG), the off-axis pupil plane recording geometry (PPRG), and the on-axis phase shifting recording geometry (PSRG). Each of these recording geometries has its benefits and drawbacks depending on the spatial-heterodyne application of interest. Thus, the goal for the following analysis is to study these recording geometries in detail, so that we can ultimately develop closed-form expressions for their signal-to-noise ratios (SNRs).

In what follows, we will explore the background information needed to investigate the three recording geometries described in Fig. 1. In particular, we will review the optics relationships needed to propagate from the object plane to the FPA and the photonics principles needed to record digital holograms with the FPA. Following these two background sections, there are detailed sections on the three recording geometries described in Fig. 1. A final section then contains an illustrative comparison of the different recording geometries and an appendix provides example MATLAB[®] code. The goal throughout is to show that we can process spatial-heterodyne measurements to obtain an estimate of the complex field. Provided the complex field, we can then pursue the spatial-heterodyne application of interest (e.g., wave-front sensing, coherent imaging, etc).

Background Material: Part I

With Fig. 1 in mind, we need to further define the experimental parameter space. To help orient the reader, Fig. 2 pictorially describes the various planes of interest within the analysis. The reader should note that the entrance pupil of our proposed imaging system effectively collimates the monochromatic light from the object plane, whereas the exit pupil effectively focuses the monochromatic light to form the image plane at focus.

In the background material that follows, we will develop a relationship that says that the pupil and image planes of our proposed imaging system form a Fourier-conjugate pair. As such, we can use spatial heterodyne to exploit this relationship and obtain an estimate of the desired complex field. Depending on the spatial-heterodyne application of interest (e.g., wavefront sensing, coherent imaging, etc.), this estimate is what makes spatial heterodyne such a versatile tool within the optics and photonics research communities.

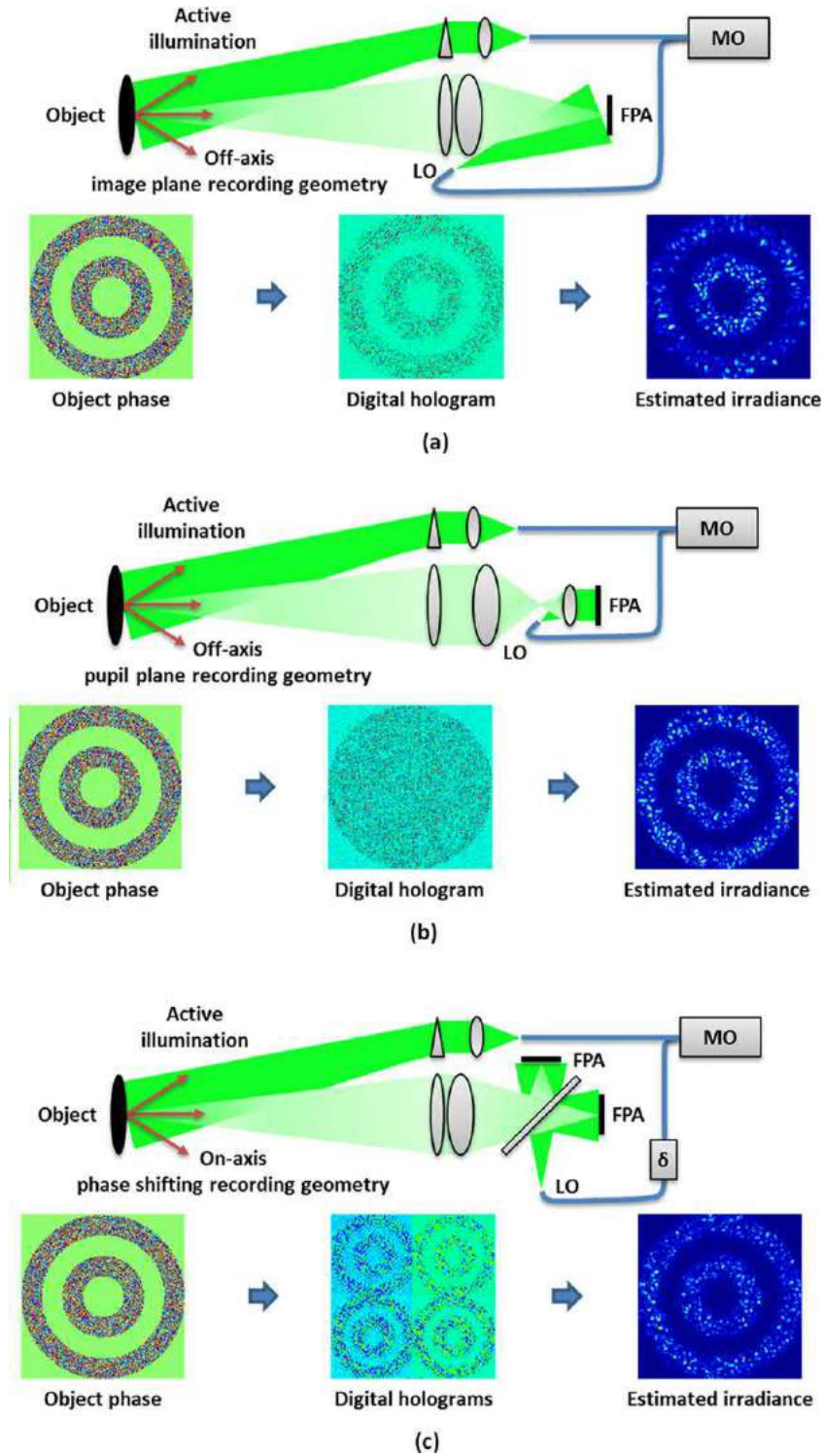


Fig. 1 A description of three spatial-heterodyne recording geometries.

Object Plane

Depending on the spatial-heterodyne application of interest, the object plane of the proposed imaging system takes on different forms. For example, given a point-source object (which is desirable for wavefront sensing),

$$U_O^+(x_0, y_0) = a_S e^{-j k z_0} \delta(x_0) \delta(y_0) \quad (1)$$

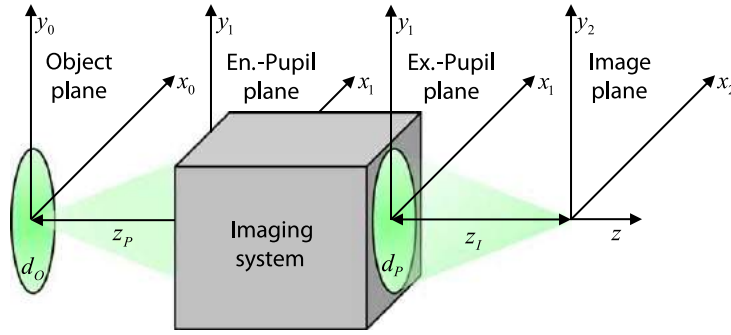


Fig. 2 A description of an imaging system.

where $U_O^+(x_0, y_0)$ is the complex field leaving the object plane, a_s is a real constant, $k=2\pi/\lambda$ is the angular wavenumber, λ is the wavelength, z_p is the distance to the object plane from the pupil plane, and $\delta(x)$ is an impulse function (cf. Eq. (80) in Appendix A). This point-source object gives rise to an ideal-spherical wave which propagates through the various planes of interest in our proposed imaging system (cf. Fig. 2). In the image plane, the irradiance associated with our imaged point source becomes a very special quantity known as the point-spread function (PSF) and is a quantity that we can measure with our FPA. In the presence of no aberrations, the PSF is an Airy disk (given a circular entrance/exit pupil), whereas in the presence of isoplanatic phase aberrations, the rings of the Airy disk start to wash out. For spatial-heterodyne applications, the sampling of the PSF with the FPA pixels becomes an important variable within the analysis and is a point that we will explore in more detail in the section on the off-axis IPRG.

For an extended object (which is desirable for coherent imaging),

$$U_O^+(x_0, y_0) = R_O(x_0, y_0) U_O^-(x_0, y_0) \quad (2)$$

where $R_O(x_0, y_0) = U_O^+(x_0, y_0)/U_O^-(x_0, y_0)$ is the object-plane complex reflectance function and $U_O^-(x_0, y_0)$ is the complex field entering the object plane. Let's assume that we actively illuminate the object plane with a uniform-amplitude, on-axis plane wave, so that

$$U_O^-(x_0, y_0) = A_s e^{-jkz_p} \quad (3)$$

where A_s is a complex constant. In addition, let's assume that the surface of the object that we are actively illuminating is optically rough compared to the wavelength of the monochromatic light (Spencer, 2014). If we then assume that the optically rough surface is delta correlated (Goodman, 2007),

$$R_O(x_0, y_0) = O(x_0, y_0) e^{j\phi_1(x_0, y_0)} \quad (4)$$

where $O(x_0, y_0)$ is the generalized complex object function and $\phi_k(x, y)$ is the k th realization of uniformly distributed, real-valued random numbers in the interval $[-\pi, \pi]$. In practice, Eqs. (2)–(4) provide us with a phase-screen model for rough-surface scattering that gives the correct speckle-correlation statistics upon propagation from the object plane to the pupil plane. For spatial-heterodyne applications, the sampling of the speckle-correlation statistics with the FPA pixels becomes an important variable within the analysis and is a point that we will explore in more detail in the section on the off-axis PPRG.

Pupil Plane

Using the convolution form of the Fresnel diffraction integral, we can represent the complex field $U_P^-(x_1, y_1)$ entering the pupil plane as

$$U_P^-(x_1, y_1) = \frac{e^{jkz_p}}{j\lambda z_p} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_O^+(x_0, y_0) \exp\left(j\frac{k}{2z_p} [(x_1 - x_0)^2 + (y_1 - y_0)^2]\right) dx_0 dy_0 \quad (5)$$

where again, $k=2\pi/\lambda$ is the angular wavenumber, λ is the wavelength, z_p is the distance to the pupil plane from the object plane, and $U_O^+(x_0, y_0)$ is the complex field leaving the object plane (cf. Eqs. (1) and (2) above).

As mentioned previously, the entrance pupil of the proposed imaging system effectively collimates the monochromatic light from the object plane, whereas the exit pupil effectively focuses the monochromatic light to form the image plane at focus (cf. Fig. 2). With this in mind, we can represent the complex field $U_P^+(x_1, y_1)$ leaving the pupil plane as

$$U_P^+(x_1, y_1) = T_P(x_1, y_1) U_P^-(x_1, y_1) \quad (6)$$

where

$$T_P(x_1, y_1) = \frac{U_P^+(x_1, y_1)}{U_P^-(x_1, y_1)} = \exp\left[-j\frac{k}{2z_p} (x_1^2 + y_1^2)\right] \exp\left[-j\frac{k}{2z_i} (x_1^2 + y_1^2)\right] P(x_1, y_1) \quad (7)$$

is the pupil-plane complex transmittance function, z_I is the distance to the image plane from the pupil plane, and

$$P(x_1, y_1) = \text{cyl}\left(\frac{\sqrt{x_1^2 + y_1^2}}{d_p}\right) e^{j\phi(x_1, y_1)} \quad (8)$$

is the generalized complex pupil function. In Eq. (8), $\text{cyl}\left(\frac{\sqrt{x_1^2 + y_1^2}}{d_p}\right)$ is a cylinder function (cf. Eq. (81) in Appendix A), d_p is the exit-pupil diameter, and $\phi(x_1, y_1)$ is the isoplanatic phase function, which represents all of the linear, shift-invariant phase aberrations present within the imaging system (Gaskill, 1978).

Substituting Eq. (7) into Eq. (6), we arrive at the following relationship:

$$U_P^+(x_1, y_1) = \exp\left[-j\frac{k}{2z_I}(x_1^2 + y_1^2)\right] U_P(x_1, y_1) \quad (9)$$

Here,

$$U_P(x_1, y_1) = \exp\left[-j\frac{k}{2z_P}(x_1^2 + y_1^2)\right] P(x_1, y_1) U_P^-(x_1, y_1) \quad (10)$$

is the pupil-plane complex field (i.e., the complex field that exists in the pupil plane of the proposed imaging system). It is important to note that $U_P(x_1, y_1)$ is being multiplied by a positive-thin-lens complex transmittance function in Eq. (9). This multiplication will allow us to form an image at focus upon propagation from the pupil plane to the image plane – our going-in disposition.

Image Plane

Provided Eq. (9), we can now account for the image-plane complex field $U_I(x_2, y_2)$ (i.e., the complex field that exists in the image plane of the proposed imaging system). In particular,

$$U_I(x_2, y_2) = \frac{e^{jkz_I}}{j\lambda z_I} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_P^+(x_1, y_1) \exp\left(j\frac{k}{2z_I}[(x_2 - x_1)^2 + (y_2 - y_1)^2]\right) dx_1 dy_1 \quad (11)$$

which again makes use of the convolution form of the Fresnel diffraction integral. Substituting Eq. (9) into Eq. (11), we arrive at the following relationship:

$$U_I(x_2, y_2) = \frac{e^{jkz_I}}{j\lambda z_I} \exp\left[j\frac{k}{2z_I}(x_2^2 + y_2^2)\right] \mathcal{F}\{U_P(x_1, y_1)\}_{x_x = \frac{x_2}{\lambda z_I}, y_y = \frac{y_2}{\lambda z_I}} = \frac{e^{jkz_I}}{j\lambda z_I} \exp\left[j\frac{k}{2z_I}(x_2^2 + y_2^2)\right] \tilde{U}_P\left(\frac{x_2}{\lambda z_I}, \frac{y_2}{\lambda z_I}\right) \quad (12)$$

This relationship says that the pupil plane and image plane form a Fourier-conjugate pair. Moving forward we can use spatial heterodyne to exploit this relationship and obtain an estimate, $\hat{U}_P(x_1, y_1)$ or $\hat{U}_I(x_2, y_2)$, of the desired complex field, $U_P(x_1, y_1)$ or $U_I(x_2, y_2)$, that exist in the pupil or image plane of our proposed imaging system.

Background Material: Part II

In the background material that follows, we will derive a relationship for the hologram photoelectron density that, in practice, a FPA records via spatial-heterodyne measurements. Provided Fig. 3, it is important to note that the spatial-heterodyne recording geometry is what ultimately determines where we place the FPA within the analysis. For example, in the off-axis IPRG we must place the FPA in the image plane of the proposed imaging system, whereas in the off-axis PPRG we must place the FPA in the pupil plane (or a plane that is conjugate to the pupil plane). The on-axis PSRG is unique in the sense that we do not have to go to the Fourier plane in order to obtain an estimate. Therefore, we can place the FPA in either the image or pupil plane of our proposed imaging system. In subsequent sections, we will explore the differences and similarities between the various spatial-heterodyne recording geometries. With that said, it is important to acknowledge their subtle differences up front, so that the following background material becomes more tractable.

Hologram Irradiance

In units of Watts per square meter ($\text{W}\cdot\text{m}^{-2}$), we can determine the real-valued hologram irradiance $i_H(x, y)$ that is incident on the FPA as

$$i_H(x, y) = |U_S(x, y) + U_R(x, y)|^2 = |U_S(x, y)|^2 + |U_R(x, y)|^2 + U_S(x, y)U_R^*(x, y) + U_R(x, y)U_S^*(x, y) \quad (13)$$

Here, $U_S(x, y)$ and $U_R(x, y)$ are, respectively, the signal and reference complex fields that are incident on the FPA and the superscript * denotes complex conjugate. It is important to note that in Eq. (13) we have made the assumption that the spatial-heterodyne

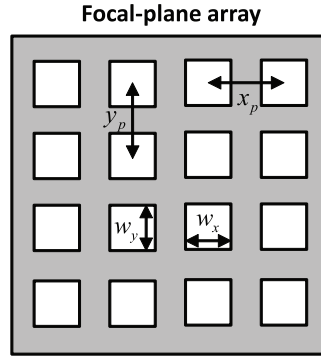


Fig. 3 Nominal layout for a focal-plane array or FPA.

measurements are fast enough so that $U_S(x, y)$ and $U_R(x, y)$ do not change significantly over the integration time of the FPA. Consequently, there is no need to include a time-dependent variable within the analysis.

For all intents and purposes, the FPA will convert the hologram irradiance $i_H(x, y)$, which is in an analog form, into a form that is suitable for digital-signal processing. Following the approach taken by Gaskill (1978), let's assume that "digitization" is to take place at sampling intervals of x_p and y_p , which are, respectively, the x - and y -axis pixel pitches of the FPA (cf. Fig. 3). At any particular FPA pixel, we can then obtain an estimate, $\hat{i}_H(x, y)$, of the hologram irradiance, $i_H(x, y)$, by computing its average value over the active area of the FPA pixel, which is centered at $x = nx_p$ and $y = my_p$, where $n = 1, 2, \dots, N$ and $m = 1, 2, \dots, M$. Specifically,

$$\hat{i}_H(nx_p, my_p) = \frac{1}{w_x w_y} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_H(x', y') \text{rect}\left(\frac{x' - nx_p}{w_x}\right) \text{rect}\left(\frac{y' - my_p}{w_y}\right) dx' dy' \quad (14)$$

where w_x and w_y are, respectively, the x - and y -axis pixel widths of the FPA, and $\text{rect}(x)$ is a rectangle function (cf. Eq. (78) in Appendix A).

Hologram Mean Number of Photoelectrons

At any particular FPA pixel, the real-valued hologram mean number of photoelectrons $\bar{m}_H(nx_p, my_p)$ results from random photoevents (i.e., when the FPA pixel creates an electron-hole pair or liberates a photoelectron (Saleh and Teich, 2007)). Therefore, over the integration time of the FPA,

$$\bar{m}_H(nx_p, my_p) = \frac{\eta \tau}{h\nu} w_x w_y \hat{i}_H(nx_p, my_p) \quad (15)$$

where η is the quantum efficiency, τ is the integration time, $h\nu$ is the quantized photon energy, and $w_x w_y$ is the active area of the FPA pixel. It is important to note that in Eq. (15) we have made the assumption that the measurements are without noise.

Following the approach taken by Tippie (2012), noise is additive. As such,

$$\bar{m}_H^+(nx_p, my_p) = \mathcal{P}\{\bar{m}_H(nx_p, my_p)\} + \sigma_r n_1(nx_p, my_p) \quad (16)$$

where $\mathcal{P}\{\circ\}$ denotes a Poisson-noise operator, σ_r is the read-noise standard deviation, and $n_k(x, y)$ is the k th realization of real-valued, zero-mean, unit-variance Gaussian random numbers. For most spatial-heterodyne applications, $\bar{m}_H(nx_p, my_p) \gg 1$ because of the use of an LO in creating the reference complex field $U_R(x, y)$. Thus, to a very good approximation, we can rewrite Eq. (16) as

$$\bar{m}_H^+(nx_p, my_p) \approx \bar{m}_H(nx_p, my_p) + \sigma_s(nx_p, my_p) n_2(nx_p, my_p) + \sigma_r n_1(nx_p, my_p) \quad (17)$$

where $\sigma_s(nx_p, my_p)$ is the shot-noise standard deviation. This approximation says that when $\bar{m}_H(nx_p, my_p) \gg 1$, our Poisson-distributed random process becomes a Gaussian-distributed random process with mean $\bar{m}_H(nx_p, my_p)$ and variance $\sigma_s^2(nx_p, my_p)$. Note that for Poisson-distributed random processes, the variance is equal to the mean (Dereniak and Boreman, 1996). In turn,

$$\sigma_s^2(nx_p, my_p) = \bar{m}_S(nx_p, my_p) + \bar{m}_R(nx_p, my_p) + \bar{m}_B(nx_p, my_p) \quad (18)$$

where $\bar{m}_S(nx_p, my_p)$, $\bar{m}_R(nx_p, my_p)$, and $\bar{m}_B(nx_p, my_p)$ are, respectively, the mean-number of photoelectrons associated with the signal, reference, and background illumination that is incident on the FPA. Also note that in Eq. (17), we have made the assumption that pixel to pixel both the shot and read noise result from statistically independent (i.e., delta correlated) random processes (Frieden, 2001). This assumption is a sound one since the sources for the shot noise (e.g., the object, LO, and sun) and read noise (i.e., the read out integrated circuit (ROIC) of the FPA) are physically separated from one another.

Hologram Photoelectron Density

At any particular FPA pixel, a real-valued hologram digital number $d_H(nx_p, my_p)$ results via an analog-to-digital (A/D) conversion (Janesick, 2007). From Eqs. (15)–(18), it then follows that

$$d_H^+(nx_p, my_p) = g_{A/D} \bar{m}_H^+(nx_p, my_p) + \sigma_q n_3(nx_p, my_p) \quad (19)$$

where $g_{A/D}$ is the A/D gain factor in units of digital number per photoelectron (DN-pe⁻¹) and σ_q is the quantization-noise standard deviation. In practice,

$$\sigma_q^2 = \frac{1}{12} \quad (20)$$

and the factor of 12 in the denominator comes from rounding to the nearest integer. Before moving on in the analysis, it is important to note that the effects of dark-current, 1/f, leakage, fano, and fixed-pattern noise are not included in Eq. (19). Under some circumstances, these noise effects might become important (Dereniak and Boreman, 1996; Saleh and Teich, 2007; Janesick, 2007); however, they are beyond the scope of the present analysis.

As previously stated, for most spatial-heterodyne applications, $\bar{m}_H(nx_s, my_s) \gg 1$ because of the use of an LO in creating the reference complex field $U_R(x, y)$ (cf. Eqs. (13)–(15)). For example, given an ideal reference with uniform irradiance, where $|U_R(x, y)|^2 = |A_R|^2$, we can determine the reference mean number of photoelectrons $\bar{m}_R(nx_p, my_p) = \bar{m}_R$ from the following relationship:

$$\bar{m}_R = \frac{\eta \tau}{h\nu} w_x w_y |A_R|^2 \quad (21)$$

Provided Eq. (21), we can then dictate that $\bar{m}_R \gg \bar{m}_S(nx_p, my_p)$, $\bar{m}_R \gg \bar{m}_B(nx_p, my_p)$, and $\bar{m}_R \gg g_{A/D}^{-1} \sigma_q$ within the constraints of the FPA pixel well depth ℓ (i.e., $\ell \geq \bar{m}_H(nx_p, my_p)$, so that the FPA pixels do not saturate). As such, $\sigma_s^2(nx_p, my_p) \approx \bar{m}_R$ (cf. Eq. (18)), and Eq. (19) simplifies, such that

$$d_H^+(nx_p, my_p) \approx g_{A/D} \left[\bar{m}_H(nx_p, my_p) + \sqrt{\bar{m}_R} n_2(nx_p, my_p) + \sigma_r n_1(nx_p, my_p) \right] \quad (22)$$

Dividing both sides of Eq. (22) by the A/D gain factor $g_{A/D}$, we obtain the following relationship:

$$\bar{m}_H^+(nx_p, my_p) \approx \bar{m}_H(nx_p, my_p) + \sigma_n n_4(nx_p, my_p) \quad (23)$$

where

$$\sigma_n^2 \approx \bar{m}_R + \sigma_r^2 \quad (24)$$

is the noise variance. This relationship says that we only have to account for photoelectrons within the analysis.

Neglecting the effects of pixel-edge diffusion in the FPA (Poon and Liu, 2014), the hologram photoelectron density $\rho_H(x_2, y_2)$, in units of pe-m⁻², is simply a sampled version of the analog form of Eq. (23). This declaration leads us to the following relationship:

$$\rho_H^+(x, y) = [\bar{m}_H(x, y) + \sigma_n n_4(x, y)] \frac{1}{x_p} \text{comb}\left(\frac{x}{x_p}\right) \frac{1}{y_p} \text{comb}\left(\frac{y}{y_p}\right) \text{rect}\left(\frac{x}{Nx_p}\right) \text{rect}\left(\frac{y}{My_p}\right) \quad (25)$$

where

$$\bar{m}_H(x, y) = \frac{\eta \tau}{h\nu} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_H(x', y') \text{rect}\left(\frac{x' - x}{w_x}\right) \text{rect}\left(\frac{y' - y}{w_y}\right) dx' dy' = \frac{\eta \tau}{h\nu} i_H(x, y) ** \text{rect}\left(\frac{x}{w_x}\right) \text{rect}\left(\frac{y}{w_y}\right) \quad (26)$$

is the analog form of Eq. (15) and $\text{comb}(x)$ is a comb function (cf. Eq. (79) in Appendix A). The reader should note that in Eq. (26), $**$ denotes a 2-D convolution, as defined in Eq. (83) in Appendix B. Moving forward we will use Eq. (25) to explore the differences and similarities between the various spatial-heterodyne recording geometries.

Off-Axis Image Plane Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the off-axis IPRG. With that said, we can represent the signal complex field $U_S(x, y)$ that is incident on the FPA as

$$U_S(x, y) = U_I(x_2, y_2) \quad (27)$$

where again, $U_I(x_2, y_2)$ is the image-plane complex field (cf. Eq. (12)). In addition, if by choice we inject an off-axis LO, then by choice we can represent the reference complex field $U_R(x, y)$ that is incident FPA as

$$U_R(x, y) = U_R(x_2, y_2) = A_R e^{jkz_l} \exp\left[j \frac{k}{2z_l} (x_2^2 + y_2^2)\right] \exp\left(j2\pi x_R \frac{x_2}{\lambda z_l}\right) \exp\left(j2\pi y_R \frac{y_2}{\lambda z_l}\right) \quad (28)$$

where here, A_R is a complex constant and (x_R, y_R) are the coordinates of the off-axis LO. It is important to note that the relationship given in Eq. (28) is the Fresnel approximation to a tilted spherical wave.

Fourier Plane

From Eqs. (1)–(28), we can gain access to an estimate, $\hat{U}_P(x_1, y_1)$, of the pupil-plane complex field (cf. Eq. (10)), $U_P(x_1, y_1)$, by going to the Fourier plane. To go to the Fourier plane, we first let $x=x_2=\lambda z_l v_x$ and $y=y_2=\lambda z_l v_y$. Next, we apply a 2-D inverse Fourier transformation, as defined in Eq. (85) in Appendix B, to the relationship contained in Eq. (25), so that

$$\mathcal{F}^{-1}\left\{\rho_H^+\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1} = \frac{1}{\lambda^2 z_l^2} \tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_l}, \frac{-y_1}{\lambda z_l}\right) = \left[\frac{\eta\tau}{h\nu} w_x \text{sinc}\left(\frac{w_x x_1}{\lambda z_l}\right) w_y \text{sinc}\left(\frac{w_y y_1}{\lambda z_l}\right) \mathcal{F}^{-1}\left\{i_H\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1} \right. \\ \left. + \sigma_n \mathcal{F}^{-1}\left\{n_4\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1}\right] ** \frac{1}{\lambda^2 z_l^2} \text{comb}\left(\frac{x_p x_1}{\lambda z_l}\right) \text{comb}\left(\frac{y_p y_1}{\lambda z_l}\right) ** \frac{N x_p}{\lambda z_l} \text{sinc}\left(\frac{N x_p x_1}{\lambda z_l}\right) \frac{M y_p}{\lambda z_l} \text{sinc}\left(\frac{M y_p y_1}{\lambda z_l}\right) \quad (29)$$

where $\text{sinc}(x)$ is a sinc function (cf. Eq. (82) in Appendix A). Taking a look at the remaining 2-D inverse Fourier transformations in Eq. (29), we obtain the following relationship with respect to the hologram irradiance $i_H(x, y)$ (cf. Eqs. (12), (13), and (28)):

$$\mathcal{F}^{-1}\left\{i_H\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1} = \frac{1}{\lambda^2 z_l^2} U_P(x_1, y_1) ** U_P^*(-x_1, -y_1) \\ + |A_R|^2 \delta(x_1) \delta(y_1) + \frac{A_R^*}{j \lambda z_l} U_P(x_1 - x_R, y_1 - y_R) - \frac{A_R}{j \lambda z_l} U_P^*(x_1 + x_R, y_1 + y_R) \quad (30)$$

This relationship contains the desired pupil-plane complex field $U_P(x_1, y_1)$.

To see that this last statement is true, let's analyze the various terms contained in the right-hand side of Eq. (30). The first term is an amplitude-scaled, 2-D autocorrelation. This term is centered on axis and is physically twice the circumference of the exit-pupil diameter d_p . The second term is also centered on axis and contains separable impulse functions (cf. Eq. (80) in Appendix A). These impulse functions are at the strength of the ideal reference with uniform irradiance (i.e., $|A_R|^2$). The last two terms form a complex-conjugate pair and contain the pupil-plane complex field $U_P(x_1, y_1)$, both scaled in amplitude and shifted off axis.

Substituting Eq. (30) into Eq. (29), we obtain the following relationship:

$$\tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_l}, \frac{-y_1}{\lambda z_l}\right) = \left\{ \frac{\eta\tau}{h\nu} w_x \text{sinc}\left(\frac{w_x x_1}{\lambda z_l}\right) w_y \text{sinc}\left(\frac{w_y y_1}{\lambda z_l}\right) \left[\frac{1}{\lambda^2 z_l^2} U_P(x_1, y_1) ** U_P^*(-x_1, -y_1) + |A_R|^2 \delta(x_1) \delta(y_1) \right. \right. \\ \left. \left. + \frac{A_R^*}{j \lambda z_l} U_P(x_1 - x_R, y_1 - y_R) - \frac{A_R}{j \lambda z_l} U_P^*(x_1 + x_R, y_1 + y_R) \right] + \sigma_n \mathcal{F}^{-1}\left\{n_4\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1} \right\} ** \text{comb}\left(\frac{x_p x_1}{\lambda z_l}\right) \text{comb}\left(\frac{y_p y_1}{\lambda z_l}\right) \\ ** \frac{N x_p}{\lambda z_l} \text{sinc}\left(\frac{N x_p x_1}{\lambda z_l}\right) \frac{M y_p}{\lambda z_l} \text{sinc}\left(\frac{M y_p y_1}{\lambda z_l}\right) \quad (31)$$

which is in units of pe. This relationship is remarkably physical, as the sampling theorem dictates that a sampled function becomes periodic upon finding its spectrum (Gaskill, 1978). In turn, the 2-D convolution with the separable comb functions causes the terms contained within the squiggly brackets in Eq. (31) to repeat at intervals of $\lambda z_l/x_p$ and $\lambda z_l/y_p$ along the x and y axes, respectively. This periodicity occurs because of the convolution-sifting property of the impulse function (Gaskill, 1978). Lastly, the final 2-D convolution with the separable narrow sinc functions serves to smooth out these repeated terms.

Sampling Quotient

To help simplify the analysis to a case that we can easily simulate using $N \times N$ computational grids, let's assume that the FPA has adjacent square pixels, so that $x_p=y_p=w_x=w_y=p$. In so doing, we can rewrite Eq. (31) in terms of the image-plane sampling quotient q_l , where

$$q_l = \frac{\lambda z_l}{p d_p} \quad (32)$$

For all intents and purposes, q_l is a measure for the number of FPA pixels across the half width of the PSF. Recall that for linear, shift-invariant imaging systems, the PSF is the irradiance associated with an imaged point source (Gaskill, 1978). In turn, the relationship given in Eq. (32) allows us to vary the sampling with the FPA pixels.

Using Eq. (32), we can rewrite Eq. (31) in terms of the image-plane sampling quotient q_l , such that

$$\tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_l}, \frac{-y_1}{\lambda z_l}\right) = \left\{ \frac{\eta\tau}{h\nu} p^2 \text{sinc}\left(\frac{x_1}{q_l d_p}\right) \text{sinc}\left(\frac{y_1}{q_l d_p}\right) \left[\frac{1}{\lambda^2 z_l^2} U_P(x_1, y_1) ** U_P^*(-x_1, -y_1) + |A_R|^2 \delta(x_1) \delta(y_1) \right. \right. \\ \left. \left. + \frac{A_R^*}{j \lambda z_l} U_P(x_1 - x_R, y_1 - y_R) - \frac{A_R}{j \lambda z_l} U_P^*(x_1 + x_R, y_1 + y_R) \right] + \sigma_n \mathcal{F}^{-1}\left\{n_4\left(\lambda z_l v_x, \lambda z_l v_y\right)\right\}_{x_1, y_1} \right\} ** \text{comb}\left(\frac{x_1}{q_l d_p}\right) \text{comb}\left(\frac{y_1}{q_l d_p}\right) \\ ** \frac{N^2}{q_l^2 d_p^2} \text{sinc}\left(\frac{N x_1}{q_l d_p}\right) \text{sinc}\left(\frac{N y_1}{q_l d_p}\right) \quad (33)$$

Here, $q_l d_p = \lambda z_l/p$ is the side length of the $N \times N$ computational grid in the Fourier plane. Note that as $N \rightarrow \infty$ (cf. Eq. (80) in Appendix A), we can make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolution in Eq. (33). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (33) becomes minimized; however, for small N the smoothing becomes more pronounced. Let's assume that $x_R=y_R=q_l d_p/4$, so that the last two terms within the square brackets in

Eq. (33) shift diagonally. When $q_I \geq 4$, the last two terms no longer overlap with the first two terms which are centered on axis. Correspondingly, when $2 \leq q_I < 4$, the last two terms are still resolvable within the side length of the $N \times N$ computational grid but overlap with the first term. This latter case allows for us to obtain more samples across the exit-pupil diameter d_p which in turn minimizes the smoothing caused by the final 2-D convolution in Eq. (33) but increases the noise sampling. Also note that this functional overlap becomes negligible when the amplitude of the reference $|A_R|$ is dominate (in comparison to the other amplitude terms).

Estimate

Provided Eqs. (32) and (33), we must shift the Fourier-plane data and apply a window function to obtain an estimate, $\hat{U}_P(x_1, y_1)$, of the pupil-plane complex field, $U_P(x_1, y_1)$. Specifically,

$$\hat{U}_P(x_1, y_1) = w(x_1, y_1) \tilde{\rho}_H^+ \left(\frac{-x_1 - x_R}{\lambda z_I}, \frac{-y_1 - y_R}{\lambda z_I} \right) \quad (34)$$

where $w(x_1, y_1)$ is the window function. In using Eq. (34), we must satisfy Nyquist sampling with the FPA pixels (Gaskill, 1978), so that the repeated terms within Eq. (33) do not overlap and cause significant aliasing. As such, the Nyquist rate is $q_I d_p = \lambda z_I / p$ and the Nyquist interval is $1/(q_I d_p) = p/(\lambda z_I)$ when $x_R = y_R = q_I d_p / 4$.

Moving forward let's assume that we satisfy Nyquist sampling. In addition, let's assume that $q_I \geq 2$, $|A_R|$ is dominant, and

$$w(x_1, y_1) = \text{cyl} \left(\frac{\sqrt{x_1^2 + y_1^2}}{d_p} \right) \quad (35)$$

In so doing, Eq. (34) simplifies, such that

$$\begin{aligned} \hat{U}_P(x_1, y_1) \approx q_I^2 d_p^2 \left[\frac{\eta \tau}{h \nu} p^2 \text{sinc} \left(\frac{x_1 + x_R}{q_I d_p} \right) \text{sinc} \left(\frac{y_1 + y_R}{q_I d_p} \right) \frac{A_R^*}{j \lambda z_I} U_P(x_1, y_1) + \sigma_n w(x_1, y_1) \mathcal{F}^{-1} \{ n_4(\lambda z_I v_x, \lambda z_I v_y) \}_{x_1 + x_R, y_1 + y_R} \right] \\ ** \delta(x_1) \delta(y_1) ** \frac{N^2}{q_I^2 d_p^2} \text{sinc} \left(\frac{N x_1}{q_I d_p} \right) \text{sinc} \left(\frac{N y_1}{q_I d_p} \right) \end{aligned} \quad (36)$$

The reader should note that as $N \rightarrow \infty$ (cf. Eq. (80) in Appendix A), we can again make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolutions in Eq. (36). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (36) becomes minimized (and the noise remains delta correlated); however, for small N the smoothing becomes more pronounced (and the noise is no longer delta correlated).

Assuming that N is large, the estimate $\hat{U}_P(x_1, y_1)$ simplifies (cf. Eq. (36)), such that

$$\hat{U}_P(x_1, y_1) \approx \frac{\hat{\eta}_I(x_1, y_1) \tau}{h \nu} p^2 A_R^* U_P(x_1, y_1) + \frac{\sigma_I}{\sqrt{2}} N_1(x_1, y_1) \quad (37)$$

where $\hat{\eta}_I(x_1, y_1)$ is the estimation efficiency in the off-axis IPRG, σ_I is the compressed-noise standard deviation in the off-axis IPRG, and $N_k(x, y)$ is the k th realization of complex-circular Gaussian random numbers with zero mean and unit variance for both the real and imaginary parts (hence the factor of $\sqrt{2}$ in the denominator). It is important to note that by Parseval's theorem (Gaskill, 1978), the total noise power in the windowed Fourier plane is equal to α_I times the original noise power in the image plane, where α_I is the ratio of the pupil area to the Fourier-plane area. Therefore,

$$\sigma_I^2 = \alpha_I \sigma_n^2 = \frac{\pi (d_p/2)^2}{(q_I d_p)^2} \sigma_n^2 = \frac{\pi}{4 q_I^2} \sigma_n^2 \quad (38)$$

which says that the noise variance σ_n^2 (cf. Eq. (24)) is compressed by a factor of α_I when performing spatial heterodyne in the off-axis IPRG.

Signal-to-noise ratio

We can determine the SNR S/N_{IPRG} for the off-axis IPRG as

$$S/N_{\text{IPRG}} = \frac{\mathcal{E}\{|\hat{U}_P(x_1, y_1)|^2\}}{\mathcal{V}\{\hat{U}_P(x_1, y_1)\}} \quad (39)$$

where $\mathcal{E}\{\circ\}$ denotes an expected-value operator and $\mathcal{V}\{\circ\}$ denotes a variance operator. Provided a strong reference, $|A_R|^2 \gg |A_S|^2$ and $|U_S(x, y)|^2 \approx |A_S|^2$. As such, we can determine the signal mean number of photoelectrons $\bar{m}_S(n_x, m_y) \approx \bar{m}_S$ from the following relationship:

$$\bar{m}_S = \frac{\eta \tau}{h \nu} p^2 |A_S|^2 \quad (40)$$

From Eqs. (37) and (38), it then follows that

$$\mathcal{E}\{|\hat{U}_P(x_1, y_1)|^2\} \approx \bar{m}_R \bar{m}_S \quad (41)$$

and

$$\mathcal{V}\{\hat{U}_P(x_1, y_1)\} \approx \sigma_I^2 \quad (42)$$

where $\bar{m}_R$ is the reference mean number of photoelectrons (cf. Eq. (21)). Substituting Eqs. (41) and (42) into Eq. (39), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\text{IPRG}} \approx \frac{4q_I^2}{\pi} \frac{\bar{m}_R \bar{m}_S}{\bar{m}_R + \sigma_r^2} \quad (43)$$

In writing this closed-form expression, we must assume that the estimation efficiency in the off-axis IPRG is equal to the quantum efficiency (i.e., $\hat{\eta}_I(x_1, y_1) = \eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\bar{m}_R \gg \sigma_r^2$, then the SNR S/N_{IPRG} for the off-axis IPRG simplifies (cf. Eq. (43)), such that

$$S/N_{\text{IPRG}} \approx \frac{4q_I^2}{\pi} \bar{m}_S \quad (44)$$

The open literature often refers to this approximation as the shot-noise limit or as obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ . In general, $\ell \geq \bar{m}_H(n x_s, m y_s)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

Off-Axis Pupil Plane Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the off-axis PPRG. In turn, we can represent the signal complex field $U_S(x, y)$ that is incident on the FPA as

$$U_S(x, y) = U_P(x_1, y_1) \quad (45)$$

where again, $U_P(x_1, y_1)$ is the pupil-plane complex field (cf. Eq. (10)). Furthermore, if by choice we inject an off-axis LO, then by choice we can represent the reference complex field $U_R(x, y)$ that is incident FPA as

$$U_R(x, y) = U_R(x_1, y_1) = A_R \exp\left(-j2\pi \frac{x_R}{\lambda z_p} x_1\right) \exp\left(-j2\pi \frac{y_R}{\lambda z_p} y_1\right) \quad (46)$$

where here, A_R is a complex constant and (x_R, y_R) are the coordinates of the off-axis LO. It is important to note that the relationship given in Eq. (46) is a tilted plane wave.

Fourier Plane

From Eqs. (1)–(26) and Eqs. (45) and (46), we can gain access to an estimate, $\hat{U}_I(x_1, y_1)$, of the image-plane complex field, $U_I(x_1, y_1)$ (cf. Eq. (10)), by going to the Fourier plane. To go to the Fourier plane, we first let $x=x_1$ and $y=y_1$. Then, we apply a 2-D Fourier transformation, as defined in Eq. (84) in Appendix B, to the relationship contained in Eq. (25), so that

$$\begin{aligned} \mathcal{F}\{\rho_H^+(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}} &= \tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) = \left[\frac{\eta\tau}{h\nu} w_x \text{sinc}\left(\frac{w_x x_2}{\lambda z_p}\right) w_y \text{sinc}\left(\frac{w_y y_2}{\lambda z_p}\right) \mathcal{F}\{i_H(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}}\right. \\ &\quad \left. + \sigma_n \mathcal{F}\{n_4(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}}\right] * \frac{1}{\lambda^2 z_p^2} \text{comb}\left(\frac{x_p x_2}{\lambda z_p}\right) \text{comb}\left(\frac{y_p y_2}{\lambda z_p}\right) * \frac{N x_p}{\lambda z_p} \text{sinc}\left(\frac{N x_p x_2}{\lambda z_p}\right) \frac{M y_p}{\lambda z_p} \text{sinc}\left(\frac{M y_p y_2}{\lambda z_p}\right) \end{aligned} \quad (47)$$

Taking a look at the remaining 2-D Fourier transformations in Eq. (47), we obtain the following relationship with respect to the hologram irradiance $i_H(x, y)$ (cf. Eqs. (10), (13), and (46)):

$$\begin{aligned} \mathcal{F}\{i_H(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}} &= \frac{1}{\lambda^2 z_p^2} \tilde{U}_P\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) * \tilde{U}_P^*\left(\frac{-x_2}{\lambda z_p}, \frac{-y_2}{\lambda z_p}\right) + \lambda^2 z_p^2 |A_R|^2 \delta(x_2) \delta(y_2) + A_R \tilde{U}_P\left(\frac{x_2 - x_R}{\lambda z_p}, \frac{y_2 - y_R}{\lambda z_p}\right) \\ &\quad + A_R^* \tilde{U}_P^*\left(\frac{x_2 + x_R}{\lambda z_p}, \frac{y_2 + y_R}{\lambda z_p}\right) \end{aligned} \quad (48)$$

This relationship contains a 2-D Fourier transformation of the pupil-plane complex field $U_P(x_1, y_1)$ which is proportional to the image-plane complex field $U_I(x_2, y_2)$ (cf. Eq. (12)).

Moving forward let's analyze the various terms contained in the right-hand side of Eq. (48). The first term is a 2-D auto-correlation. This term is centered on axis and is physically twice the circumference of the object diameter d_O . The second term is also centered on axis and contains separable impulse functions (cf. Eq. (80) in Appendix A). These impulse functions are at the strength of the ideal reference with uniform irradiance (i.e., $|A_R|^2$). The last two terms form a complex-conjugate pair and contain a 2-D Fourier transformation of the pupil-plane complex field $U_P(x_1, y_1)$, both scaled in amplitude and shifted off axis.

Substituting Eq. (48) into Eq. (47), we obtain the following relationship:

$$\begin{aligned} \tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) = & \left\{ \frac{\eta\tau}{h\nu} w_x \text{sinc}\left(\frac{w_x x_2}{\lambda z_p}\right) w_y \text{sinc}\left(\frac{w_y y_2}{\lambda z_p}\right) \left[\frac{1}{\lambda^2 z_p^2} \tilde{U}_p\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) * \tilde{U}_p^*\left(\frac{-x_2}{\lambda z_p}, \frac{-y_2}{\lambda z_p}\right) + \lambda^2 z_p^2 |A_R|^2 \delta(x_2) \delta(y_2) \right. \right. \\ & \left. \left. + A_R \tilde{U}_p\left(\frac{x_2 - x_R}{\lambda z_p}, \frac{y_2 - y_R}{\lambda z_p}\right) + A_R \tilde{U}_p^*\left(\frac{x_2 + x_R}{\lambda z_p}, \frac{y_2 + y_R}{\lambda z_p}\right) \right] + \sigma_n \mathcal{F}\{n_4(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}} \right\} \\ & * \frac{1}{\lambda^2 z_p^2} \text{comb}\left(\frac{x_p x_2}{\lambda z_p}\right) \text{comb}\left(\frac{y_p y_2}{\lambda z_p}\right) * \frac{N x_p}{\lambda z_p} \text{sinc}\left(\frac{N x_p x_2}{\lambda z_p}\right) \frac{M y_p}{\lambda z_p} \text{sinc}\left(\frac{M y_p y_2}{\lambda z_p}\right) \end{aligned} \quad (49)$$

In units of p , this relationship is physically relevant. Note that the sampling theorem dictates that a sampled function becomes periodic upon finding its spectrum (Gaskill, 1978). In turn, the 2-D convolution with the separable comb functions and the convolution-sifting property of the impulse function causes the terms contained within the squiggly brackets in Eq. (49) to repeat at intervals of $\lambda z_p/x_p$ and $\lambda z_p/y_p$ along the x and y axes, respectively. Also note that the final 2-D convolution with the separable narrow sinc functions serves to smooth out these repeated terms.

Sampling Quotient

To help simplify the analysis to a case that we can easily simulate using $N \times N$ computational grids, let's again assume that the FPA has adjacent square pixels, so that $x_p = y_p = w_x = w_y = p$. In so doing, we can rewrite Eq. (49) in terms of the pupil-plane sampling quotient q_p , where

$$q_p = \frac{\lambda z_p}{p d_O} \quad (50)$$

For all intents and purposes, q_p is a measure for the number of FPA pixels across the half width of the speckle-correlation radius (Goodman, 2007; Spencer, 2014). Recall that for monochromatic light, the speckle-correlation radius is proportional to the size of the speckles. In turn, the relationship given in Eq. (50) allows us to vary the sampling with the FPA pixels.

Using Eq. (50), we can rewrite Eq. (49) in terms of the pupil-plane sampling quotient q_p , so that

$$\begin{aligned} \tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) = & \frac{1}{p^2} \left\{ \frac{\eta\tau}{h\nu} p^2 \text{sinc}\left(\frac{x_2}{q_p d_O}\right) \text{sinc}\left(\frac{y_2}{q_p d_O}\right) \left[\frac{1}{\lambda^2 z_p^2} \tilde{U}_p\left(\frac{x_2}{\lambda z_p}, \frac{y_2}{\lambda z_p}\right) * \tilde{U}_p^*\left(\frac{-x_2}{\lambda z_p}, \frac{-y_2}{\lambda z_p}\right) + \lambda^2 z_p^2 |A_R|^2 \delta(x_2) \delta(y_2) \right. \right. \\ & \left. \left. + A_R \tilde{U}_p\left(\frac{x_2 - x_R}{\lambda z_p}, \frac{y_2 - y_R}{\lambda z_p}\right) + A_R \tilde{U}_p^*\left(\frac{x_2 + x_R}{\lambda z_p}, \frac{y_2 + y_R}{\lambda z_p}\right) \right] + \sigma_n \mathcal{F}\{n_4(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_p}, v_y = \frac{y_2}{\lambda z_p}} \right\} \\ & * \frac{1}{q_p^2 d_O^2} \text{comb}\left(\frac{x_2}{q_p d_O}\right) \text{comb}\left(\frac{y_2}{q_p d_O}\right) * \frac{N^2}{q_p^2 d_O^2} \text{sinc}\left(\frac{N x_2}{q_p d_O}\right) \text{sinc}\left(\frac{N y_2}{q_p d_O}\right) \end{aligned} \quad (51)$$

Here, $q_p d_O = \lambda z_p/p$ is the side length of the $N \times N$ computational grid in the Fourier plane. Note that as $N \rightarrow \infty$ (cf. Eq. (80) in the Appendix), we can make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolution in Eq. (51). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (51) becomes minimized; however, for small N the smoothing becomes more pronounced. Let's assume that $x_R = y_R = q_p d_O/4$, so that the last two terms within the square brackets in Eq. (51) shift diagonally. When $q_p \geq 4$, the last two terms no longer overlap with the first two terms which are centered on axis. Correspondingly, when $2 \leq q_p < 4$, the last two terms are still resolvable within the side length of the $N \times N$ computational grid but overlap with the first term. This latter case allows for us to obtain more samples across the object diameter d_O which in turn minimizes the smoothing caused by the final 2-D convolution in Eq. (51) but increases the noise sampling. Also note that this functional overlap becomes negligible when the amplitude of the reference $|A_R|$ is dominant (in comparison to the other amplitude terms).

Estimate

Provided Eqs. (50) and (51), we must shift the Fourier-plane data and apply a window function to obtain an estimate, $\hat{U}_l(x_2, y_2)$, of the image-plane complex field, $U_l(x_2, y_2)$. In particular,

$$\hat{U}_l(x_2, y_2) = w(x_2, y_2) \tilde{\rho}_H^+\left(\frac{x_2 + x_R}{\lambda z_p}, \frac{y_2 + y_R}{\lambda z_p}\right) \quad (52)$$

where $w(x_2, y_2)$ is the window function. In using Eq. (52), we must satisfy Nyquist sampling with the FPA pixels (Gaskill, 1978), so that the repeated terms within Eq. (51) do not overlap and cause significant aliasing. Thus, the Nyquist rate is $q_p d_O = \lambda z_p/p$ and the Nyquist interval is $1/(q_p d_O) = p/(\lambda z_p)$ when $x_R = y_R = q_p d_O/4$.

Assuming that we satisfy Nyquist sampling, let's assume that $q_p \geq 2$, $|A_R|$ is dominant, and

$$w(x_2, y_2) = \text{cyl}\left(\frac{\sqrt{x_2^2 + y_2^2}}{d_O}\right) \quad (53)$$

In turn, Eq. (52) simplifies, such that

$$\begin{aligned} \hat{U}_I(x_2, y_2) \approx & \frac{1}{p^2} \left[\frac{\eta\tau}{h\nu} p^2 \text{sinc}\left(\frac{x_2 + x_R}{q_P d_O}\right) \text{sinc}\left(\frac{y_2 + y_R}{q_P d_O}\right) A_R^* \tilde{U}_P\left(\frac{x_2}{\lambda z_P}, \frac{y_2}{\lambda z_P}\right) + \sigma_n w(x_1, y_1) \mathcal{F}\{n_4(x_1, y_1)\}_{v_x = \frac{x_2}{\lambda z_P}, v_y = \frac{y_2}{\lambda z_P}} \right] \\ & * * \frac{N^2}{q_P^2 d_O^2} \text{sinc}\left(\frac{Nx_2}{q_P d_O}\right) \text{sinc}\left(\frac{Ny_2}{q_P d_O}\right) \end{aligned} \quad (54)$$

The reader should note that as $N \rightarrow \infty$ (cf. Eq. (80) in Appendix A), we can again make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolutions in Eq. (54). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (54) becomes minimized (and the noise remains delta correlated); however, for small N the smoothing becomes more pronounced (and the noise is no longer delta correlated).

Assuming that N is large, the estimate $\hat{U}_I(x_1, y_1)$ simplifies (cf. Eq. (54)), such that

$$\hat{U}_I(x_2, y_2) \approx \frac{\hat{\eta}_P(x_2, y_2)\tau}{h\nu} A_R^* \tilde{U}_P\left(\frac{x_2}{\lambda z_P}, \frac{y_2}{\lambda z_P}\right) + \frac{\sigma_P}{\sqrt{2}} N_1(x_2, y_2) \quad (55)$$

where here, $\hat{\eta}_P(x_1, y_1)$ is the estimation efficiency in the off-axis PPRG, σ_P is the compressed-noise standard deviation in the off-axis PPRG, and $N_k(x, y)$ is again the k th realization of complex-circular Gaussian random numbers with zero mean and unit variance for both the real and imaginary parts (hence the factor of $\sqrt{2}$ in the denominator). It is important to note that by Parseval's theorem (Gaskill, 1978), the total noise power in the windowed Fourier plane is equal to α_P times the original noise power in the pupil plane, where α_P is the ratio of the image area to the Fourier-plane area. Therefore,

$$\sigma_P^2 = \alpha_P \sigma_n^2 = \frac{\pi(d_O/2)^2}{(q_P d_O)^2} \sigma_n^2 = \frac{\pi}{4q_P^2} \sigma_n^2 \quad (56)$$

which says that the noise variance σ_n^2 (cf. Eq. (24)) is compressed by a factor of α_P when performing spatial heterodyne in the off-axis PPRG.

Signal-to-noise ratio

We can determine the SNR S/N_{PPRG} for the off-axis PPRG as

$$S/N_{\text{PPRG}} = \frac{\mathcal{E}\{|\hat{U}_I(x_2, y_2)|^2\}}{\mathcal{V}\{\hat{U}_I(x_2, y_2)\}} \quad (57)$$

where again, $\mathcal{E}\{\circ\}$ denotes an expected-value operator and $\mathcal{V}\{\circ\}$ denotes a variance operator. From Eqs. (55) and (56),

$$\mathcal{E}\{|\hat{U}_I(x_2, y_2)|^2\} \approx \bar{m}_R \bar{m}_S \quad (58)$$

and

$$\mathcal{V}\{\hat{U}_I(x_2, y_2)\} \approx \sigma_P^2 \quad (59)$$

Here, $\bar{m}_R$ and $\bar{m}_S$ are the reference and signal mean number of photoelectrons, respectively, assuming a strong reference (cf. Eqs. (21) and (40)). Substituting Eqs. (58) and (59) into Eq. (57), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\text{PPRG}} \approx \frac{4q_P^2}{\pi} \frac{\bar{m}_R \bar{m}_S}{\bar{m}_R + \sigma_r^2} \quad (60)$$

In writing this closed-form expression, we must assume that the estimation efficiency in the off-axis PPRG is equal to the quantum efficiency (i.e., $\hat{\eta}_P(x_1, y_1) = \eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\bar{m}_R \gg \sigma_r^2$, then the SNR S/N_{PPRG} for the off-axis PPRG simplifies (cf. Eq. (60)), such that

$$S/N_{\text{PPRG}} \approx \frac{4q_P^2}{\pi} \bar{m}_S \quad (61)$$

Again, the open literature often refers to this approximation as the shot-noise limit or as obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ . In general, $\ell \geq \bar{m}_H(n_x, m_y)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

On-Axis Phase Shifting Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the on-axis PSRG. For this purpose, we can represent the signal complex field $U_S(x, y)$ that is incident on the FPA as

$$U_S(x, y) = U_P(x_1, y_1) \quad (62)$$

or

$$U_S(x, y) = U_I(x_2, y_2) \quad (63)$$

where again, $U_P(x_1, y_1)$ is the pupil-plane complex field (cf. Eq. (10)) and $U_I(x_2, y_2)$ is the image-plane complex field (cf. Eq. (12)). Moreover, if by choice we inject an on-axis LO, then by choice we can represent the reference complex field $U_R(x, y)$ that is incident FPA as

$$U_R^{(\delta)}(x, y) = U_R^{(\delta)}(x_1, y_1) = U_R^{(\delta)}(x_2, y_2) = A_R e^{-j\delta} \quad (64)$$

where here, A_R is a complex constant and δ is a real constant. It is important to note that the relationship given in Eq. (64) is an ideal reference with a uniform irradiance and a piston-phase shift (i.e., $|A_R|^2$ and δ , respectively).

Estimate

Provided Eqs. (62)–(64), Eq. (13) becomes

$$I_H^{(\delta)}(x, y) = |U_S(x, y)|^2 + |A_R|^2 + U_S(x, y)A_R^* e^{j\delta} + A_R e^{-j\delta} U_S^*(x, y) \quad (65)$$

This relationship says that via (simultaneous or sequential) piston-phase shifts, we can acquire four hologram-irradiance measurements with the FPA (Poon and Liu, 2014) to obtain an estimate, $\hat{U}_P(x_1, y_1)$ or $\hat{U}_I(x_2, y_2)$, of the desired complex field, $U_P(x_1, y_1)$ or $U_I(x_2, y_2)$. For example, if $\delta=0, \pi/2, \pi$, and $3\pi/2$, then

$$\begin{aligned} I_H^{(0)}(x, y) &= |U_S(x, y)|^2 + |A_R|^2 + U_S(x, y)A_R^* + A_R U_S^*(x, y) \\ I_H^{(\pi/2)}(x, y) &= |U_S(x, y)|^2 + |A_R|^2 + jU_S(x, y)A_R^* - jA_R U_S^*(x, y) \\ I_H^{(\pi)}(x, y) &= |U_S(x, y)|^2 + |A_R|^2 - U_S(x, y)A_R^* - A_R U_S^*(x, y) \\ I_H^{(3\pi/2)}(x, y) &= |U_S(x, y)|^2 + |A_R|^2 - jU_S(x, y)A_R^* + jA_R U_S^*(x, y) \end{aligned} \quad (66)$$

To remove the common terms (i.e., $|U_S(x, y)|^2 + |A_R|^2$), we can perform the following subtractions:

$$\begin{aligned} I_H^{(0)}(x, y) - I_H^{(\pi)}(x, y) &= 2U_S(x, y)A_R^* + 2A_R U_S^*(x, y) \\ I_H^{(\pi/2)}(x, y) - I_H^{(3\pi/2)}(x, y) &= j2U_S(x, y)A_R^* - j2A_R U_S^*(x, y) \end{aligned} \quad (67)$$

so that

$$\left[I_H^{(0)}(x, y) - I_H^{(\pi)}(x, y) \right] + j \left[I_H^{(3\pi/2)}(x, y) - I_H^{(\pi/2)}(x, y) \right] = 4A_R^* U_S(x, y) \quad (68)$$

From Eqs. (1)–(26) and Eq. (68), our estimate, in units of pe, becomes

$$\begin{aligned} \hat{U}_S(x, y) &\approx \frac{4\eta\tau}{S\hbar\nu} A_R^* U_S(x, y) * \text{rect}\left(\frac{x}{p}\right) \text{rect}\left(\frac{y}{p}\right) + 2\sigma_n N_2(x, y) \\ &\approx \frac{4\hat{\eta}(x, y)\tau}{S\hbar\nu} p^2 A_R^* U_S(x, y) + 2\sigma_n N_2(x, y) \end{aligned} \quad (69)$$

with the assumption that the FPA has adjacent square pixels, so that $x_p = y_p = w_x = w_y = p$. In Eq. (69), S is the total number of amplitude splits required to make the four hologram-irradiance measurements with the FPA (Rhoadarmer and Barchers, 2002). It is important to note that since we have four hologram-irradiance measurements with the FPA, the noise variance σ_n^2 (cf. Eq. (24)) is multiplied by a factor of 4 when performing spatial heterodyne in the on-axis PSRG.

Signal-to-noise ratio

We can determine the SNR S/N_{PSRG} for the on-axis PSRG as

$$S/N_{\text{PSRG}} = \frac{\mathcal{E}\{|\hat{U}_S(x, y)|^2\}}{\mathcal{V}\{\hat{U}_S(x, y)\}} \quad (70)$$

where again, $\mathcal{E}\{\circ\}$ denotes an expected-value operator and $\mathcal{V}\{\circ\}$ denotes a variance operator. From Eq. (69),

$$\mathcal{E}\{|\hat{U}_I(x_1, y_1)|^2\} \approx \frac{16}{S^2} \bar{m}_R \bar{m}_S \quad (71)$$

and

$$\mathcal{V}\{\hat{U}_I(x_1, y_1)\} \approx 4\sigma_n^2 \quad (72)$$

where here, $\bar{m}_R$ and $\bar{m}_S$ are the reference and signal mean number of photoelectrons, respectively, assuming a strong reference (cf. Eqs. (21) and (40)). Substituting Eqs. (71) and (72) into Eq. (70), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\text{PSRG}} \approx \frac{4}{S^2} \frac{\bar{m}_R \bar{m}_S}{\bar{m}_R + \sigma_r^2} \quad (73)$$

In writing this closed-form expression, we must assume that the estimation efficiency in the on-axis PSRG is equal to the quantum efficiency (i.e., $\hat{\eta}(x, y) = \eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\bar{m}_R \gg \sigma_r^2$, then the SNR S/N_{PSRG} for the on-axis PSRG simplifies (cf. Eq. (73)), such that

$$S/N_{\text{PSRG}} \approx \frac{4}{S^2} \bar{m}_S \quad (74)$$

Again, the open literature often refers to this approximation as the shot-noise limit or obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ . In general, $\ell \geq \bar{m}_H(n x_s, m y_s)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

Comparison of the Different Recording Geometries

In this section, we will explore the differences between the various recording geometries (studied in the previous sections) from a visual standpoint. For this purpose, let's assume that we have a point-source object (cf. Eq. (1)), so that the laser-object interaction gives rise to an ideal spherical wave upon propagation. Let's also assume that we have isoplanatic phase aberrations present (i.e., those that exist (or approximately exist) within the pupil plane of our imaging system (cf. Fig. 2)). As such (cf. Eqs. (8) and (10)),

$$\frac{U_P(x_1, y_1)}{A_s} = P(x_1, y_1) \quad (75)$$

and we can neglect the effects of speckle due to an optically rough object. Before moving on in the analysis, it is important to note that example MATLAB[®] code for the off-axis IPRG, the off-axis PPRG, and the on-axis PSRG is included in Appendix C. The readers interested in extended objects (in addition to point-source objects) will be able to visualize the associated physics (cf. Fig. 1) by executing these provided scripts.

To quantify performance, we will use the field-estimated Strehl ratio S_F , such that

$$S_F = \frac{|\mathcal{E}\{U_S(x, y) \hat{U}_S^*(x, y)\}|^2}{\mathcal{E}\{|U_S(x, y)|^2\} \mathcal{E}\{|\hat{U}_S(x, y)|^2\}} \quad (76)$$

where $U_S(x, y)$ and $\hat{U}_S(x, y)$ are the "truth" and "estimated" signal complex fields, respectively, and again, $\mathcal{E}\{\cdot\}$ is the expected-value operator. This performance metric bears some resemblance to a Strehl ratio, which in practice provides a normalized measure for performance (Spencer *et al.*, 2016). In Eq. (76), if $U_S(x, y) = \hat{U}_S(x, y)$, then $S_F = 1$. Else if $U_S(x, y) \neq \hat{U}_S(x, y)$, then $S_F < 1$. Thus, Eq. (76) is in line with the general understanding of a Strehl ratio, and provides a normalized measure for field-estimated performance. The reader should note that Eq. (76) ultimately stems from the following definition of the Strehl ratio:

$$S = \frac{|\mathcal{E}\{U_P(x_1, y_1)\}|^2}{\mathcal{E}\{|U_P(x_1, y_1)|^2\}} \quad (77)$$

Here, we have made use of the fact that the expected value of a pupil-plane field quantity is equivalent to the on-axis DC term of the 2D Fourier transformation of that pupil-plane field quantity.

Off-Axis Image Plane Recording Geometry

Fig. 4 shows the results for the off-axis IPRG. In the left-hand column, $q_I = 2$, whereas in the right-hand column, $q_I = 4$ (cf. Eq. (32)). Note that the computational grids were setup so that the image-plane grid length was set equal to the pupil-plane grid length (which was also set equal to the exit-pupil diameter $d = 0.3$ m) using 256×256 grid points. This setup required that $z_I = q_I d^2 / (N \lambda)$ when solving the Fresnel integral numerically (Spencer *et al.*, 2016). Also note that

- Fig. 4(a) and (b) show the 1 RMS isoplanatic phase aberrations that existed in the simulated pupil plane;
- Fig. 4(c) and (d) show the zero-mean digital holograms recorded with the off-axis IPRG;
- Fig. 4(e) and (f) show the amplitudes associated with the simulated Fourier plane; and
- Fig. 4(g) and (h) show the wrapped phases associated with the estimated complex field.

Here, $S_F = 0.943$ and $S_F = 0.984$ for the left-hand and right-hand columns of Fig. 4, respectively. To calculate the field-estimated Strehl ratio S_F (cf. Eq. (76)), the simulated Fourier plane was zero padded to the appropriate size, so that the windowed Fourier plane (i.e., the estimated complex field) contained 256×256 grid points. This zero padding ensured that the estimated complex field had the same number of grid points as the simulated pupil plane. Given that $\bar{m}_S = 4$ pe, $\bar{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe, the following SNRs were also calculated for the left-hand and right-hand columns of Fig. 4 (cf. Eqs. (43) and (44)): $S/N = 9$ and $S/N = 18$.

Off-Axis Pupil Plane Recording Geometry

Fig. 5 shows the results for the off-axis PPRG. Here, $q_P = 2$ in the left-hand column and $q_P = 4$ in the right-hand column (cf. Eq. (50)). The computational grids were setup with 256×256 grid points and the pupil-plane grid length was set equal to the exit-pupil diameter $d = 0.3$ m. It is important to note that

- Fig. 5(a) and (b) show the 1 RMS isoplanatic phase aberrations that existed in the simulated pupil plane;
- Fig. 5(c) and (d) show the zero-mean digital holograms recorded with the off-axis PPRG;

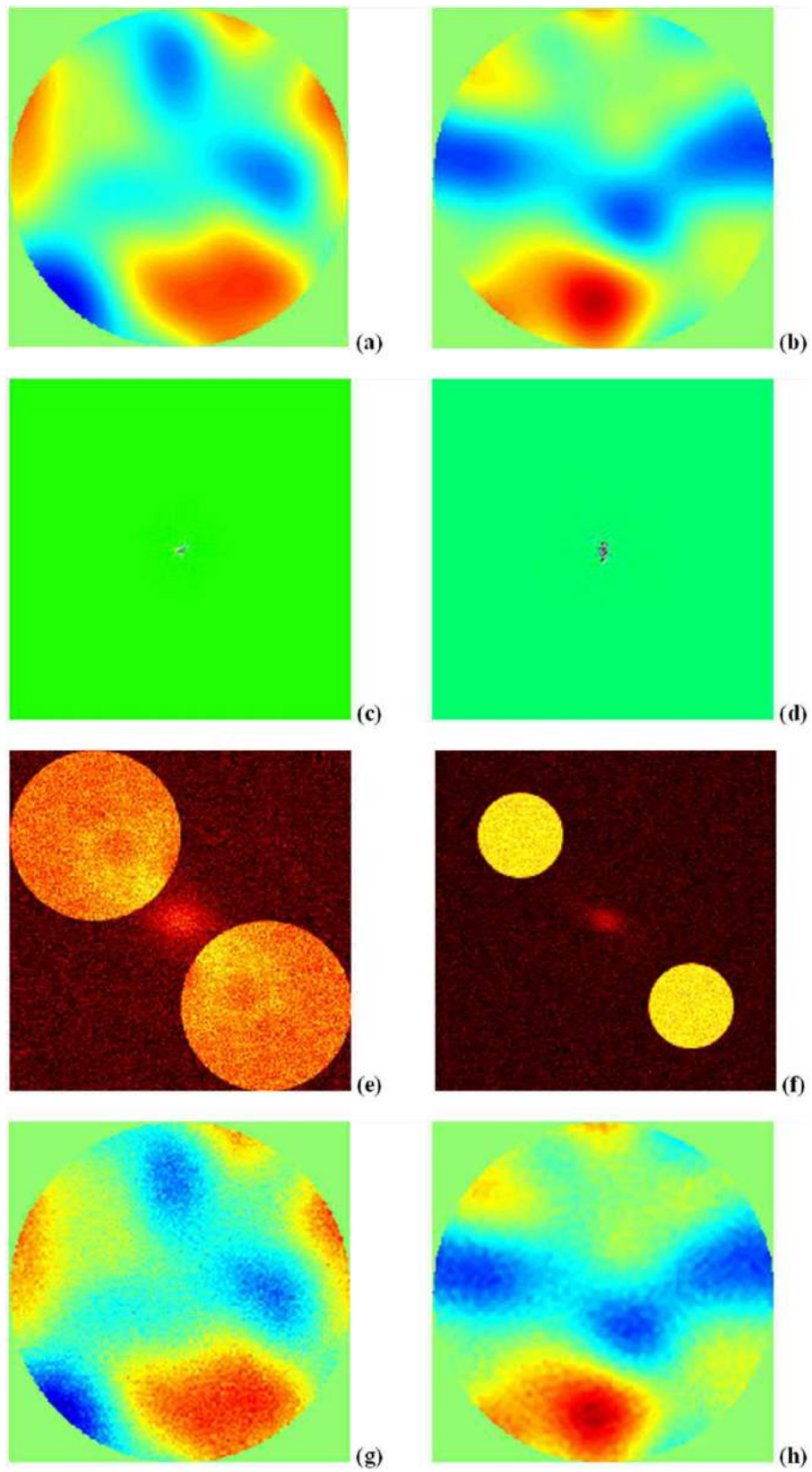


Fig. 4 Results for the off-axis image plane recording geometry or IPRG.

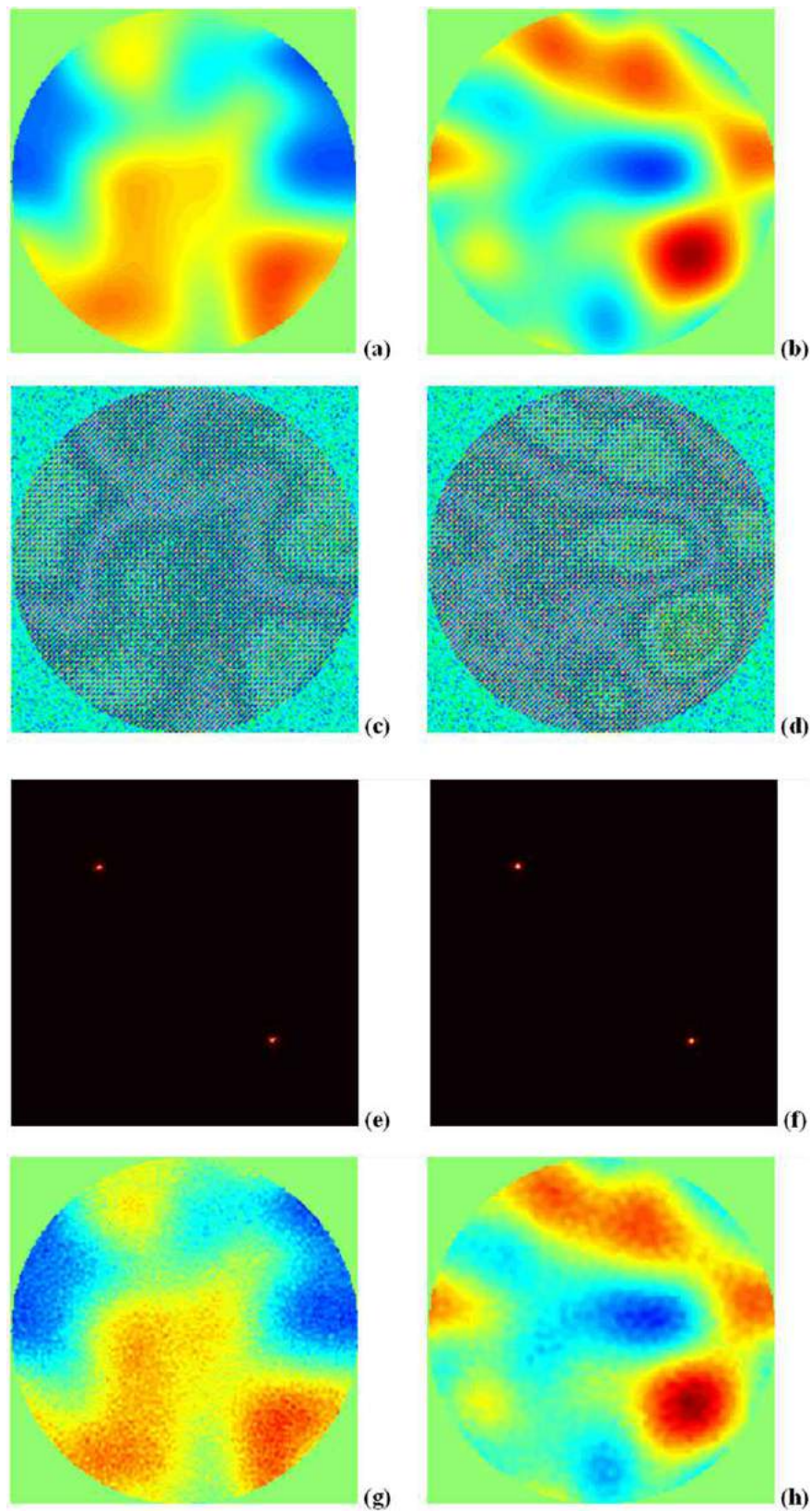


Fig. 5 Results for the off-axis pupil plane recording geometry or PPRG.

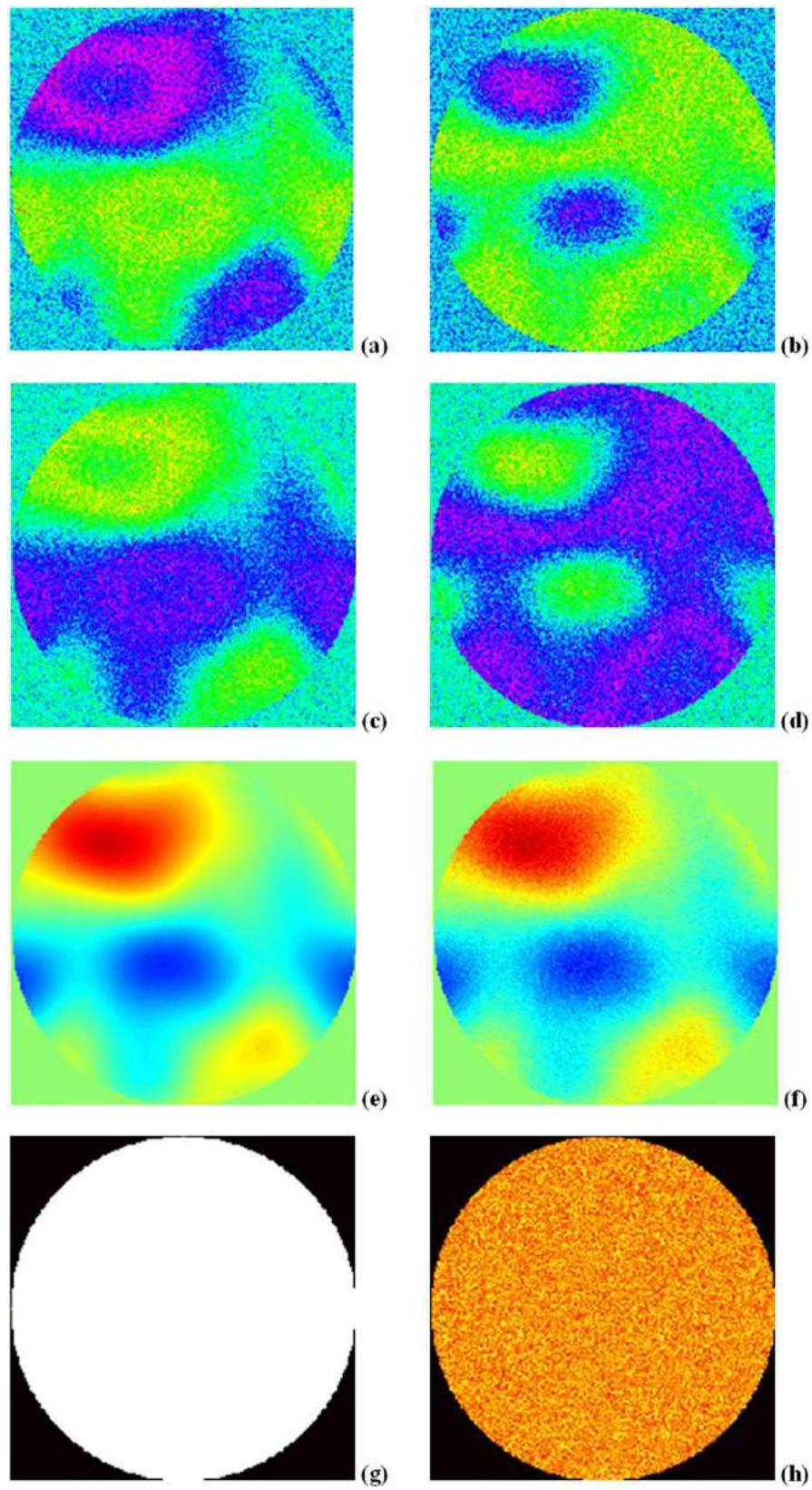


Fig. 6 Results for the on-axis phase shifting recording geometry or PSRG.

- **Fig. 5(e)** and **(f)** show the amplitudes associated with the simulated Fourier plane; and
- **Fig. 5(g)** and **(h)** show the wrapped phases associated with the estimated complex field.

To calculate the estimated complex field, the windowed Fourier plane was zero padded to 256×256 grid points and was numerically inverse Fourier transformed. This zero padding ensured that the estimated complex field had the same number of grid points as the simulated pupil plane for the purposes of computing field-estimated Strehl ratio S_F (cf. Eq. (76)), where here, $S_F=0.955$ and $S_F=0.986$ for the left-hand and right-hand columns of **Fig. 5**, respectively. Given that $\bar{m}_S = 4$ pe, $\bar{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe, the following SNRs were also calculated for the left-hand and right-hand columns of **Fig. 5** (cf. Eqs. (60) and (61)): $S/N=9$ and $S/N=18$.

On-Axis Phase Shifting Recording Geometry

Fig. 6 shows the results for the on-axis PSRG. Once again, the computational grids were setup with 256×256 grid points and the pupil-plane grid length was set equal to the exit-pupil diameter $d=0.3$ m. **Fig. 6(a)–(d)** show the four phase-shifted digital holograms, where $\delta=0, \pi/2, \pi$, and $3\pi/2$, respectively. Here, $S=1$, so that $\bar{m}_S = 4/S$ pe, $\bar{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe. Given these parameters, the following SNR was calculated for the results shown in **Fig. 6** (cf. Eqs. (60) and (61)): $S/N=14$. It is important to note that **Fig. 6(e)** and **(g)** show the wrapped phase and amplitude for the truth complex field, whereas **Fig. 6(f)** and **(h)** show the wrapped phase and amplitude for the estimated complex field. As such, the field-estimated Strehl ratio S_F (cf. Eq. (76)), was calculated as $S_F=0.947$.

Conclusion

Spatial heterodyne, in general, offers a distinct way forward to combat low SNRs that often occur when performing optics and photonics applications. In using spatial heterodyne, we can set the strength of the reference so that it boosts the signal above the read-noise floor of the FPA. As such, we can approach a shot-noise limited detection regime. This last statement is of course dependent on the parameters of the FPA, such as the pixel well depth. With that said, this encyclopedia article provides the toolset needed for future research efforts to use spatial heterodyne in their optics and photonics applications.

Appendix A

This appendix defines the rectangle, comb, impulse, cylinder, and sinc functions as

$$\text{rect}(x) = \begin{cases} 0, & |x| > 0.5 \\ 0.5, & x = 0.5 \\ 1, & |x| < 0.5 \end{cases} \quad (78)$$

$$\frac{1}{|w|} \text{comb}\left(\frac{x}{w}\right) = \sum_{n=-\infty}^{\infty} \delta(x - nw) \quad (79)$$

$$\delta(x) = \lim_{w \rightarrow 0} \frac{1}{|w|} p\left(\frac{x}{w}\right) \quad (80)$$

(where $p(x)$ is a pulse-like function (e.g., the rectangle function)),

$$\text{cyl}(\rho) = \begin{cases} 10 \leq \rho < 0.5 \\ 0.5, \rho = 0.5 \\ 0, \rho > 0.5 \end{cases} \quad (81)$$

(where $\rho = \sqrt{x^2 + y^2}$), and

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \quad (82)$$

respectively.

Appendix B

This appendix defines 2-D convolution as

$$V(x, y) * W(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} V(x', y') W(x - x', y - y') dx' dy' \quad (83)$$

(where x' and y' are dummy variables of integration), a 2-D Fourier transform as

$$\mathcal{F}\{V(x, y)\}_{v_x, v_y} = \tilde{V}(v_x, v_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} V(x, y) e^{-j2\pi(xv_x + yv_y)} dx dy \quad (84)$$

and a 2-D inverse Fourier transformation as

$$\mathcal{F}^{-1}\{\tilde{V}(v_x, v_y)\}_{x, y} = V(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{V}(v_x, v_y) e^{j2\pi(xv_x + yv_y)} dv_x dv_y \quad (85)$$

Appendix C

This appendix provides the MATLAB[®] code needed to explore the off-axis IPRG, the off-axis PPRG, and the on-axis PSRG.

Listing 1. MATLAB[®] code for the cylinder function.

```
function z = cyl(x, y)
% function z = cyl(x, y)
r = sqrt(x.^2 + y.^2);
z = double(r < 0.5);
z(r == 0.5) = 0.5;
return
```

Listing 2. MATLAB[®] code for the off-axis image plane recording geometry or IPRG.

```
% script_IPRG_final
clear all
close all
clc
%
N = 256; % number of grid points
d = 0.3; % exit-pupil diameter [m]
delta = d/N; % grid spacing [m]
wvl = 1e-6; % wavelength [m]
k = 2*pi/wvl; % angular wavenumber [rad/m]
q_I = 2.0; % IPRG quotient
ifov = wvl/d/q_I; % pixel field of view
z_I = d/(N*ifov); % propagation length [m] (FPA side = d)
sigma_r = 100; % read-noise standard deviation [pe]
l = 100e3; % FPA well depth [pe]
m_R = 0.75*l; % reference mean number of pe's [pe]
m_S = 4; % signal mean number of pe's [pe]
%
% Setup the pupil plane
%
[x1, y1] = meshgrid((-N/2 : N/2-1)*delta);
r1 = hypot(x1, y1);
%
phi = randn(N); % phase aberration [rad]
a = N/10*delta; % 1/e width [m]
g = exp(-r1.^2/(2*a^2)); % Gaussian function
phi = ... % smoothed phase aberration [rad]
real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
```



```

phi = ... % rms = 1 [rad]
(phi-mean(phi(:)))/std(phi(:));
P = ... % gen. complex pupil function
cyl(x1/d,y1/d).*exp(1i*phi);
%
% Let's assume a point-source object
%
U_P = P; % pup.-plane complex field [sqrt(W)/m]
%
figure(1); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_P)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_P),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_Pp = ... % leaving complex field [sqrt(W)/m]
exp(-1j*k/(2*z_I)*(x1.^2 + y1.^2)).*U_P;
%
figure(2); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pp)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pp),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Propagate from the pupil plane to the image plane
%
[x2,y2] = meshgrid((-N/2:N/2-1)*delta);
%
M = 2*N+rem(N,2); % double the number of grid points
U_Ppg = zeros(M); % pad with a guard band of 2
nn = round(N/2)+(1:N); % original coordinates
U_Ppg(nn,nn) = U_Pp; % embed in padded grid
% transform the embedded field
U_Ppgt = delta^2*fftshift(fft2(fftshift(U_Ppg)));
% create the propagation kernel
const = k/(2*z_I)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2)).^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ppgt = (kernelt*kernelt.').*U_Ppgt;
U_Ip = 1/(1j*wvl*z_I)*exp(1j*k*z_I) ...
*(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ppgt)));
% use original coordinates
U_I = U_Ip(nn,nn); % img.-plane complex field [sqrt(W)/m]
U_S = ... % signal complex field [sqrt(pe)]
sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(3); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
x_R = d*q_I/4; % x-axis reference origin
y_R = d*q_I/4; % y-axis reference origin
U_R = ... % reference complex field [sqrt(W)/m]
1/1j*exp(1j*k*z_I)*exp(1j*k*(x2.^2+y2.^2)/(2*z_I)) ...

```

```

        .*exp(1i*2*pi*y_R*(y2/(wvl*z_I))).*exp(1i*2*pi*x_R*(x2/(wvl*z_I)));
U_R = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(4); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the image plane on a FPA
%
m_H = abs(U_S + U_R).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (point-source object)']);
m_Hn = ... % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the following if you do not have the Statistics Toolbox
% m_Hn = ... % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_Hn(:))) ' [pe] (point-source object with noise)']);
%
m_H = m_H - mean(m_H(:)); % remove DC
m_Hn = m_Hn - mean(m_Hn(:)); % remove DC
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,jet);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,jet);
%
M = round(q_I*N); % desired number of grid points
m_Hng = zeros(M); % pad with a gaurdband
nn = round(M/2-N/2)+(1:N); % original coordinates
m_Hng(nn,nn) = m_Hn; % embed in padded grid
%
% Go to the Fourier plane to obtain an estimate
%
m_Hngt = ... % Fourier plane
    (1/delta)^2*ifftshift(ifft2(ifftshift(m_Hng)));
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hngt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hngt),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
nn = round(3*M/4-N/2)+(1:N); % estimate coordinates
U_Pe = ... % window Fourier plane
    cyl(x1/d,y1/d).*m_Hngt(nn,nn);
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
S_F = ... % calculate field-estimated Strehl

```

```

        abs(mean(U_P(:).*conj(U_Pe(:))))).^2 ...
        /(mean(abs(U_P(:)).^2).*mean(abs(U_Pe(:)).^2));
display(['Field-estimated Strehl Ratio      : ' num2str(S_F)]);
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
d0 = d;                                % diameter of the object [m]
O = ...                                % gen. complex object function
    cyl(x0/d0,y0/d0) ...
    - cyl(x0/(3*d0/4),y0/(3*d0/4)) ...
    - cyl(x0/(d0/2),y0/(d0/2)) ...
    + cyl(x0/(d0/4),y0/(d0/4));
phz = -pi + 2*pi*rand(N);               % unif. dist. random phase (-pi,pi]
R_0 = 0.*exp(1j*phz);                   % obj.-plane complex reflectance.
%
U_I = ...                               % img.-plane complex field [sqrt(w)/m]
    ifftshift(ifft2(ifftshift( fftshift(fft2(fftshift(R_0))) ...
    .* fftshift(fft2(fftshift((U_I)))))));
U_S = ...                               % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(8); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
x_R = d*q_I/4;                          % x-axis reference origin
y_R = d*q_I/4;                          % y-axis reference origin
U_R = ...                               % reference complex field [sqrt(w)/m]
    1/1j*exp(1j*k*z_I)*exp(1j*k*(x2.^2+y2.^2)/(2*z_I)) ...
    .*exp(1i*2*pi*y_R*(y2/(wvl*z_I))).*exp(1i*2*pi*x_R*(x2/(wvl*z_I)));
U_R = ...                               % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the image plane on a FPA
%
m_H = abs(U_S + U_R).^2;                 % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (extended object)']);
m_Hn = ...                              % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the following if you do not have the Statistics Toolbox
% m_Hn = ...                             % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoelectrons: ' ...

```

```

num2str(mean(m_Hn(:))) ' [pe] (extended object with noise)']);
%
m_H = m_H - mean(m_H(:));           % remove DC
m_Hn = m_Hn - mean(m_Hn(:));        % remove DC
%
figure(10); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,hsv);
%
% Go to the Fourier plane to obtain an estimate
%
m_Hnt = ...                          % Fourier plane
(1/delta)^2*ifftshift(ifft2(ifftshift(m_Hn)));
%
figure(11); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hnt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hnt),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = ...                             % estimate coordinates
round(N/4-round(N/q_I)/2)+(1:round(N/q_I));
M = length(nn);                      % estimate number of grid points
[xf,yf] = meshgrid((-M/2 : M/2-1)*d*q_I/N);
U_Pe = ...                           % window Fourier plane
cyl(xf/d,yf/d).*m_Hnt(nn,nn);
%
figure(12); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = round(N/2-M/2)+(1:M);           % desired coordinates
U_Peg = zeros(N);                   % pad with a gaurdband
U_Peg(nn,nn) = U_Pe;                % embed in padded grid
U_Ie = ...
(d*q_I/N)^2 * fftshift(fft2(fftshift(U_Peg)));
%
figure(13),clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(abs(R_0)); axis image; axis off; colormap(ax2,hot);
%
SNR = ...                            % calculate SNR
4*q_I*m_R*m_S/(pi*(m_R + sigma_r^2));
display(['PPRG signal-to-noise ratio      : ' num2str(SNR)]);

```

Listing 3. MATLAB® code for the off-axis pupil plane recording geometry or PPRG.

```

% script_PPRG_final
clear all
close all
clc
%
```

```

N = 256; % number of grid points
d = 0.3; % exit-pupil diameter [m]
delta = d/N; % grid spacing [m]
wvl = 1e-6; % wavelength [m]
k = 2*pi/wvl; % angular wavenumber [rad/m]
q_P = 2.0; % PPRG quotient
ifov = wvl/d/q_P; % pixel field of view
z_P = d/(N*ifov); % propagation length [m] (FPA side = d)
sigma_r = 100; % read-noise standard deviation [pe]
l = 100e3; % FPA well depth [pe]
m_R = 0.75*l; % reference mean number of pe's [pe]
m_S = 4; % signal mean number of pe's [pe]
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
d0 = d; % diameter of the object [m]
O = ... % gen. complex object function
    cyl(x0/d0,y0/d0) ...
    - cyl(x0/(3*d0/4),y0/(3*d0/4)) ...
    - cyl(x0/(d0/2),y0/(d0/2)) ...
    + cyl(x0/(d0/4),y0/(d0/4));
phz = -pi + 2*pi*rand(N); % unif. dist. random phase (-pi,pi]
R_0 = 0.*exp(1j*phz); % obj.-plane complex reflectance.
U_0 = ... % obj.-plane complex field [sqrt(W)/m]
    R_0*exp(-1j*k*z_P);
%
figure(1); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_0)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_0),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
% Propagate from the object plane to the pupil plane
%
[x1,y1] = meshgrid((-N/2:N/2-1)*delta);
r1 = hypot(x1,y1);
%
M = 2*N+rem(N,2); % double the number of grid points
U_Og = zeros(M); % pad with a guard band of 2
nn = round(N/2)+(1:N); % original coordinates
U_Og(nn,nn) = U_0; % embed in padded grid
% transform the embedded field
U_Ogt = delta^2*fftshift(fft2(fftshift(U_Og)));
% create the propagation kernel
const = k/(2*z_P)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2)).^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ogt = (kernelt*kernelt.').*U_Ogt;
U_Pg = k/(1j*2*pi*z_P)*exp(1j*k*z_P) ...
    *(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ogt)));
% use original coordinates
U_Pm = U_Pg(nn,nn); % entering complex field [sqrt(W)/m]
%
figure(2); clf;
ax1=subplot(1,2,1);

```

```

imagesc(abs(U_Pm)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pm),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
% Setup the pupil plane
%
phi = randn(N); % phase aberration [rad]
a = N/10*delta; % 1/e width [m]
g = exp(-r1.^2/(2*a^2)); % Gaussian function
phi = ... % smoothed phase aberrtaion [rad]
    real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
phi = ... % rms = 1 [rad]
    (phi-mean(phi(:)))/std(phi(:));
P = ... % gen. complex pupil function
    cyl(x1/d,y1/d).*exp(1i*phi);
U_P = ... % pup.-plane complex field [sqrt(w)/m]
    exp(-1j*k/(2*z_P)*(x1.^2 + y1.^2)).*P.*U_Pm;
U_S = ... % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(3); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
x_R = d*q_P/4; % x-axis reference origin
y_R = d*q_P/4; % y-axis reference origin
U_R = ... % reference complex field [sqrt(w)/m]
    exp(-1i*2*pi*y_R*(y1/(w1*z_P))).*exp(-1i*2*pi*x_R*(x1/(w1*z_P)));
U_R = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(4); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the pupil plane on a FPA
%
m_H = abs(U_S + U_R).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (extended object)']);
m_Hn = ... % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the follwoing if you do not have the Statistics Toolbox
% m_Hn = ... % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_Hn(:))) ' [pe] (extended object with noise)']);
%
m_H = m_H - mean(m_H(:)); % remove DC
m_Hn = m_Hn - mean(m_Hn(:)); % remove DC
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);

```



```

ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2, hsv);
%
M = round(q_P*N); % desired number of grid points
m_Hng = zeros(M); % pad with a gaurdband
nn = round(M/2-N/2)+(1:N); % original coordinates
m_Hng(nn,nn) = m_Hn; % embed in padded grid
%
% Go to the Fourier plane to obtain an estimate
%
m_Hngt = ... % Fourier plane
    delta^2*fftshift(fft2(fftshift(m_Hng)));
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hngt)); axis image; axis off; colormap(ax1, hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hngt), [-pi, pi]); axis image; axis off; colormap(ax2, jet);
%
nn = round(3*M/4-N/2)+(1:N); % estimate coordinates
U_Ie = ... % window Fourier plane
    cyl(x1/d, y1/d) .* m_Hngt(nn, nn);
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1, hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Ie), [-pi, pi]); axis image; axis off; colormap(ax2, jet);
%
SNR = ... % calculate SNR
    4*q_P*m_R*m_S/(pi*(m_R + sigma_r^2));
display(['PPRG signal-to-noise ratio : ' num2str(SNR)]);
%
% Let's assume a point-source object
%
U_P = P; % pup.-plane complex field [sqrt(W)/m]
U_S = ... % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(8); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1, hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S), [-pi, pi]); axis image; axis off; colormap(ax2, jet);
%
x_R = d*q_P/4; % x-axis reference origin
y_R = d*q_P/4; % y-axis reference origin
U_R = ... % reference complex field [sqrt(W)/m]
    exp(-1i*2*pi*y_R*(y1/(wvl*z_P))) .* exp(-1i*2*pi*x_R*(x1/(wvl*z_P)));
U_R = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1, hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R), [-pi, pi]); axis image; axis off; colormap(ax2, jet);
%

```

```

% Measure the pupil plane on a FPA
%
m_H = abs(U_S + U_R).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H(:))) ' [pe] (point-source object)']);
m_Hn = ... % add noise
        poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the following if you do not have the Statistics Toolbox
% m_Hn = ... % add noise
% m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_Hn(:))) ' [pe] (point-source object with noise)']);
%
m_H = m_H - mean(m_H(:)); % remove DC
m_Hn = m_Hn - mean(m_Hn(:)); % remove DC
%
figure(10); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,hsv);
%
% Go to the Fourier plane to obtain an estimate
%
m_Hnt = ... % Fourier plane
        delta^2*fftshift(fft2(fftshift(m_Hn)));
%
figure(11); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hnt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hnt),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = ... % estimate coordinates
        round(3*N/4-round(N/q_P)/2)+(1:round(N/q_P));
M = length(nn); % estimate number of grid points
[xf,yf] = meshgrid((-M/2 : M/2-1)*d*q_P/N);
U_Ie = ... % window Fourier plane
        cyl(xf/d,yf/d).*m_Hnt(nn,nn);
%
figure(12); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Ie),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = round(N/2-M/2)+(1:M); % desired coordinates
U_Ieg = zeros(N); % pad with a gaurdband
U_Ieg(nn,nn) = U_Ie; % embed in padded grid
U_Pe = ...
        (d*q_P)^2*fftshift(ifft2(fftshift(U_Ieg))).*cyl(x1/d,y1/d);
%
figure(13),clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
S_F = ... % calculate field-estimated Strehl
        abs(mean(U_P(:).*conj(U_Pe(:)))).^2 ...
        /(mean(abs(U_P(:)).^2).*mean(abs(U_Pe(:)).^2));
display(['Field-estimated Strehl ratio : ' num2str(S_F)]);

```

Listing 4. MATLAB® code for the on-axis phase shifting recording geometry or PSRG.

```

% script_PSRG_final
%
clear all
close all
clc
%
N = 256; % number of grid points
d = 0.3; % exit-pupil diameter [m]
delta = d/N; % grid spacing [m]
wvl = 1e-6; % wavelength [m]
k = 2*pi/wvl; % angular wavenumber [rad/m]
q_I = 4.0; % IPRG quotient
ifov = wvl/d/q_I; % pixel field of view
z_I = d/(N*ifov); % propagation length [m] (FPA side = d)
sigma_r = 100; % read-noise standard deviation [pe]
l = 100e3; % FPA well depth [pe]
m_R = 0.75*l; % reference mean number of pe's [pe]
S = 1; % number of amplitude splits
m_S = 4/S; % signal mean number of pe's [pe]
%
% Setup the pupil plane
%
[x1,y1] = meshgrid((-N/2 : N/2-1)*delta);
r1 = hypot(x1,y1);
%
phi = randn(N); % phase aberration [rad]
a = N/10*delta; % 1/e width [m]
g = exp(-r1.^2/(2*a^2)); % Gaussian function
phi = ... % smoothed phase aberration [rad]
    real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
phi = ... % rms = 1 [rad]
    (phi-mean(phi(:)))/std(phi(:));
P = ... % gen. complex pupil function
    cyl(x1/d,y1/d).*exp(1i*phi);
%
% Let's assume a point-source object
%
U_P = P; % pup.-plane complex field [sqrt(W)/m]
U_S = ... % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(1); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_R = 1; % reference complex field [sqrt(W)/m]
U_R1 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R1(:)).^2)) ' [pe] (1)']);
U_R2 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R2(:)).^2)) ' [pe] (2)']);
U_R3 = ... % reference complex field [sqrt(pe)]

```

```

    sqrt(m_R)*exp(-1i*pi)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R3(:)).^2)) ' [pe] (3)']);
U_R4 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*3*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R4(:)).^2)) ' [pe] (4)']);
%
figure(2); clf;
ax1 = subplot(2,2,1);
imagesc(angle(U_R1),[-pi,pi]); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(angle(U_R2),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(angle(U_R3),[-pi,pi]); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(angle(U_R4),[-pi,pi]); axis image; axis off; colormap(ax4,hsv);
%
% Measure the pupil plane on a FPA
%
m_H1 = abs(U_S + U_R1).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H1(:))) ' [pe] (point-source object 1)']);
m_H2 = abs(U_S + U_R2).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H2(:))) ' [pe] (point-source object 2)']);
m_H3 = abs(U_S + U_R3).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H3(:))) ' [pe] (point-source object 3)']);
m_H4 = abs(U_S + U_R4).^2; % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H4(:))) ' [pe] (point-source object 4)']);
%
figure(3); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4); axis image; axis off; colormap(ax4,hsv);
%
m_H1n = ... % add noise
    poissrnd(m_H1)+sigma_r*randn(size(m_H1));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H1n(:))) ' [pe] (point-source object with noise 1)']);
m_H2n = ... % add noise
    poissrnd(m_H2)+sigma_r*randn(size(m_H2));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H2n(:))) ' [pe] (point-source object with noise 2)']);
m_H3n = ... % add noise
    poissrnd(m_H3)+sigma_r*randn(size(m_H3));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H3n(:))) ' [pe] (point-source object with noise 3)']);
m_H4n = ... % add noise
    poissrnd(m_H4)+sigma_r*randn(size(m_H4));
display(['Hol. mean number of photoelectrons: ' ...
    num2str(mean(m_H4n(:))) ' [pe] (point-source object with noise 4)']);
%

```

```

figure(4); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1n); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2n); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3n); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4n); axis image; axis off; colormap(ax4,hsv);
%
U_Se = ((m_H1n-m_H3n) + 1j*(m_H4n-m_H2n)).*cyl(x1/d,y1/d);
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Se)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_Se),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
S_F = ... % calculate field-estimated Strehl
abs(mean(U_S(:).*conj(U_Se(:))))).^2 ...
/(mean(abs(U_S(:)).^2).*mean(abs(U_Se(:)).^2));
display(['Field-estimated Strehl Ratio : ' num2str(S_F)]);
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
d0 = d; % diameter of the object [m]
O = ... % gen. complex object function
cyl(x0/d0,y0/d0) ...
- cyl(x0/(3*d0/4),y0/(3*d0/4)) ...
- cyl(x0/(d0/2),y0/(d0/2)) ...
+ cyl(x0/(d0/4),y0/(d0/4));
phz = -pi + 2*pi*rand(N); % unif. dist. random phase (-pi,pi]
R_0 = 0.*exp(1j*phz); % obj.-plane complex reflectance.
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(R_0)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(R_0),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_Pp = ... % leaving complex field [sqrt(W)/m]
exp(-1j*k/(2*z_I)*(x1.^2 + y1.^2)).*U_P;
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pp)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pp),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Propagate from the pupil plane to the image plane
%
[x2,y2] = meshgrid((-N/2:N/2-1)*delta);
%
M = 2*N+rem(N,2); % double the number of grid points
U_Ppg = zeros(M); % pad with a guard band of 2
nn = round(N/2)+(1:N); % original coordinates
U_Ppg(nn,nn) = U_Pp; % embed in padded grid

```

```

% transform the embedded field
U_Ppgt = delta^2*fftshift(fft2(fftshift(U_Ppgt)));
% create the propagation kernel
const = k/(2*z_I)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2)).^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ppgt = (kernelt*kernelt.').*U_Ppgt;
U_Ip = 1/(1j*wvl*z_I)*exp(1j*k*z_I) ...
    *(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ppgt)));
% use original coordinates
U_I = U_Ip(nn,nn); % img.-plane complex field [sqrt(W)/m]
%
figure(8); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_I)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_I),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_I = ... % img.-plane complex field [sqrt(W)/m]
    ifftshift(ifft2(ifftshift(fftshift(fft2(fftshift(R_0)))) ...
    .* fftshift(fft2(fftshift((U_I)))) ));
U_S = ... % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_R = 1; % reference complex field [sqrt(W)/m]
U_R1 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R1(:)).^2)) ' [pe] (1)']);
U_R2 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R2(:)).^2)) ' [pe] (2)']);
U_R3 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R3(:)).^2)) ' [pe] (3)']);
U_R4 = ... % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*3*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoelectrons: ' ...
    num2str(mean(abs(U_R4(:)).^2)) ' [pe] (4)']);
%
figure(10); clf;
ax1 = subplot(2,2,1);
imagesc(angle(U_R1),[-pi,pi]); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(angle(U_R2),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(angle(U_R3),[-pi,pi]); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(angle(U_R4),[-pi,pi]); axis image; axis off; colormap(ax4,hsv);
%

```



```

% Measure the image plane on a FPA
%
m_H1 = abs(U_S + U_R1).^2;           % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H1(:))) ' [pe] (extended object 1)']);
m_H2 = abs(U_S + U_R2).^2;           % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H2(:))) ' [pe] (extended object 2)']);
m_H3 = abs(U_S + U_R3).^2;           % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H3(:))) ' [pe] (extended object 3)']);
m_H4 = abs(U_S + U_R4).^2;           % holgram mean number of pe's [pe]
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H4(:))) ' [pe] (extended object 4)']);
%
figure(11); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4); axis image; axis off; colormap(ax4,hsv);
%
m_H1n = ...                          % add noise
        poissrnd(m_H1)+sigma_r*randn(size(m_H1));
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H1n(:))) ' [pe] (extended object with noise 1)']);
m_H2n = ...                          % add noise
        poissrnd(m_H2)+sigma_r*randn(size(m_H2));
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H2n(:))) ' [pe] (extended object with noise 2)']);
m_H3n = ...                          % add noise
        poissrnd(m_H3)+sigma_r*randn(size(m_H3));
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H3n(:))) ' [pe] (extended object with noise 3)']);
m_H4n = ...                          % add noise
        poissrnd(m_H4)+sigma_r*randn(size(m_H4));
display(['Hol. mean number of photoelectrons: ' ...
        num2str(mean(m_H4n(:))) ' [pe] (extended object with noise 4)']);
%
figure(12); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1n); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2n); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3n); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4n); axis image; axis off; colormap(ax4,hsv);
%
U_Se = (m_H1n-m_H3n) + 1j*(m_H4n-m_H2n);
%
figure(13); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Se)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_Se),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
SNR = ...                            % calculate SNR
        4*m_R*m_S/(S^2*(m_R + sigma_r^2));
display(['PSRG signal-to-noise ratio      : ' num2str(SNR)]);

```

See also: Heterodyning

References

- Dereniak, E.L., Boreman, G.D., 1996. *Infrared Detectors and Systems*. New York: John Wiley and Sons.
- Frieden, B.R., 2001. *Probability, Statistical Optics, and Data Testing*, third ed. New York: Springer-Verlag.
- Gaskill, J.D., 1978. *Linear Systems, Fourier Transforms, and Optics*. New York: John Wiley and Sons.
- Goodman, J.W., 2007. *Speckle Phenomena in Optics Theory and Application*. Englewood: Roberts and Company.
- Janesick, J.R., 2007. *Photon Transfer*. Bellingham: SPIE Press.
- McManamon, P.F., 2015. *Field Guide to Lidar*. Bellingham: SPIE Press.
- Poon, T.-C., Liu, J.-P., 2014. *Introduction to Modern Digital Holography*. New York: Cambridge University Press.
- Rhoadarmer, T.A., Barchers, J.D., 2002. Noise Analysis for complex field estimation using a self-referencing interferometer wave front sensor. *s.l.* SPIE.
- Saleh, B.E.A., Teich, M.C., 2007. *Fundamentals of Photonics*, second ed. New York: John Wiley and Sons.
- Spencer, M.F., 2014. *The Scattering of Partially Coherent Electromagnetic Beam Illumination from Statistically Rough Surfaces*. PhD Dissertation, Air Force Institute of Technology, Volume ADA603227.
- Spencer, M.F., Raynor, R.A., Banet, M.T., Marker, D.K., 2016. Deep-turbulence wavefront sensing using digital-holographic detection in the off-axis image plane recording geometry. *Optical Engineering* 56 (3), 031213.
- Tippie, A.E., 2012. *Aberration Correction in Digital Holography*. PhD Dissertation, University of Rochester, Volume AS38.6635.

Further Reading

- Steinbock, M.J., Hyde IV, M.W., Schmidt, J.D., 2014. LSPV + 7, a branch-point-tolerant reconstructor for strong turbulence adaptive optics. *Applied Optics* 53 (18), 3821–3831.

3D Metrics for Airborne Topographic Lidar

Shea T Hagstrom and Myron Z Brown, The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The development and widespread use of airborne topographic lidar systems is one of the most significant changes within the remote sensing field in recent decades. Lidar is now commonly used for production of high-resolution foundation data such as Digital Elevation Models (DEMs) and bare-earth Digital Terrain Models (DTMs) as well as for mapping power lines and discovering ancient ruins under jungle canopy. Community standards for 3D lidar metrics have emerged to set minimum expectations for lidar data quality and to enable objective and quantitative evaluation and comparison of systems and capabilities.

Lidar data quality depends on sensor technology (e.g., well established linear-mode lidar and more recently emerging photon counting lidar), system design, calibration, operating environment, and post-processing such as registration of overlapping swaths and noise filtering. Metrics for assessment of 3D lidar point clouds must be general enough to account for all of these factors. Where possible, metrics should also be general enough to apply to 3D data derived from other sources such as multiple view imagery (Bosch *et al.*, 2016) and synthetic aperture radar (Zhu and Bamler, 2012), though we do not explicitly address these in this article. As lidar system designs continue to evolve, and as lidar technology continues to be applied in new ways, metrics for ensuring acceptable data quality are advancing as well to keep up. This article reviews lidar metrics currently in common use, highlights emerging challenges, and discusses recent work to begin to address those challenges.

Current Lidar Metric Standards

Minimum expectations for topographic mapping lidar metric performance are captured in recent community standards for lidar collection and government specifications for delivered lidar products. The United States Geological Survey (USGS) National Geospatial Program (NGP) released the Lidar Base Specification in 2014 to define minimum metric requirements for the vertical accuracy, data density, and completeness of lidar products for the National interagency 3D Elevation Program (3DEP). Similarly, the American Society for Photogrammetry and Remote Sensing (ASPRS) released the ASPRS Positional Accuracy Standards for Digital Geospatial Data in 2015 to specify vertical and horizontal accuracy requirements for data providers. While other organizations and agencies also independently define metric standards, these documents describe the current state of the industry in metric performance characterization of lidar products. The following discussion draws heavily from these documents to summarize commonly used lidar metrics to assess data density, completeness, and accuracy.

Density and Completeness

Data density is perhaps the most intuitive metric associated with a lidar point cloud, as it indicates what sizes of objects in a scene can be expected to be well represented. The terms nominal pulse spacing (NPS) and nominal pulse density (NPD) are commonly used to indicate lidar point spacing and density respectively on a measured ground surface within the area covered by data of a single flight line or scan. The lidar base specification also defines the terms aggregate nominal pulse spacing (ANPS) and aggregate nominal pulse density (ANPD) to indicate point spacing and density more generally for point clouds that include points from multiple flight lines. Furthermore, to properly characterize point spacing and density for lidar systems incorporating focal plane arrays, even more general terms would be required. The intent of any of these terms is to capture the equivalent of ground sample distance (GSD) in remote sensing digital photography, though lidar point spacing is generally irregular. These metrics are two-dimensional and intended primarily for conventional topographic lidar data collection. In the section on emerging capabilities, a more general three-dimensional metric for data density is discussed that is suitable for lidar collections consisting of multiple off-nadir viewpoints to increase coverage of objects in the scene.

In addition to nominal point density, the completeness and spatial distribution of points are also important in determining the utility of a point cloud. Gaps in lidar coverage are commonly termed data voids. The lidar base specification defines a data void as any area of missing points that is at least as large as $4 \times \text{ANPS}^2$ and indicates that these data voids are unacceptable in delivered lidar products unless caused by water bodies or low near infrared surface reflectivity. Regularity of the spatial distribution of points is also required. This is measured by producing a two-dimensional occupancy grid with cell size $2 \times \text{ANPS}$ and determining if at least 90% of cells are occupied. These metrics are also intended primarily for conventional topographic lidar. A more general discussion of three-dimensional completeness is provided in the section on emerging capabilities.

Absolute and Relative Vertical Accuracy

Absolute vertical accuracy of lidar elevation values is typically estimated by comparison with check point elevation values that have been independently measured in the lidar coverage area. ASPRS recommends that check points be at least three times more accurate than the required accuracy of the product and located in areas with low and uniform slope to avoid interpolation errors.

Highly accurate check points are typically obtained by Global Positioning System (GPS) survey. Control points used for calibration of a lidar system are also typically obtained by GPS survey and should not be included as check points. Absolute vertical accuracy is reported separately in non-vegetated and vegetated terrain and is commonly reported as root mean squared error (RMSE) for non-vegetated terrain, assuming a normal error distribution, or percentile confidence level for vegetated terrain.

Relative vertical accuracy refers to precision or repeatability and internal consistency of elevation measurements in a lidar product. Vertical precision is estimated as the variation in elevation values measured on a flat hard surface such as a building roof. Internal consistency is estimated as the variation in elevation values measured on non-vegetated terrain in overlapping swaths. ASPRS and USGS define standards for repeatability based on the maximum measured difference and for consistency based on both root mean squared difference (RMSD) and maximum difference. Sources of relative vertical error include range precision of the lidar instrument and registration error among overlapping lidar swaths, among others.

Absolute and Relative Horizontal Accuracy

Absolute horizontal accuracy of a lidar product is estimated by comparison with horizontal check points that have been independently measured in the lidar coverage area. Unlike vertical check points described above, horizontal check points must be located at points in the lidar coverage area for which the horizontal position may be very accurately measured independently and which are clearly identifiable in the lidar product itself. These check points are typically obtained by GPS survey and the locations clearly documented with photography. ASPRS defines horizontal accuracy standards for RMSE of X and Y coordinates independently, RMSE of radial accuracy, and percentile confidence level of radial accuracy.

Relative horizontal accuracy refers to point to point horizontal distance measurement accuracy within a geospatial product and may be estimated similarly to absolute horizontal accuracy using check points. ASPRS does not define an explicit standard for relative horizontal accuracy in lidar datasets, and absolute horizontal accuracy standards for lidar collection are sufficiently stringent that independently measuring relative horizontal accuracy would have limited utility. However, in situations for which ground control and GPS accuracy may be limited, such as remote lidar collection for humanitarian purposes, explicitly measuring relative horizontal error may be desirable. Methods for beginning to address these issues are discussed in the following section on emerging capabilities.

Emerging Capabilities and Challenges

Current lidar metrics standards have been developed for use with mature commercial linear-mode lidar technology, fixed nadir viewing angles, and accurate ground survey. New lidar technology, system designs, and use cases may require additional or revised metrics to accurately characterize error. For instance, more emphasis on measuring relative horizontal accuracy may be desirable for recent photon counting lidar systems which require post-processing for photon aggregation and noise filtering that can improve or degrade resolution. Similarly, comprehensive modeling of all error sources may be more desirable for lidar systems that collect from higher altitudes and with significantly off-nadir look angles such that some errors often considered negligible become important. Off-nadir data collection also enables enhanced foliage penetration and true three-dimensional mapping capabilities. Current metrics and conventional wisdom for topographic lidar data collection are insufficient to fully characterize these new collection capabilities. The following sections discuss preliminary efforts within the lidar research community to begin to address some of these recent challenges.

Defining a Lidar Error Model

Inaccuracies in lidar point cloud geospatial coordinates are primarily influenced by uncertainties in sensor position, sensor orientation, and range measurements as described by [Habib et al. \(2010\)](#). These uncertainties are commonly characterized for lidar system error budgets and sensor calibration processing; however, they generally are not readily available to end users of lidar data. While current standards may ensure that products meet minimum performance specifications, additional knowledge of specific uncertainties is desirable for an end user who wishes to merge or compare multiple datasets.

[Rodarmel et al. \(2015\)](#) have proposed a Universal Lidar Error Model (ULEM) that captures sensor-space parameter uncertainties and explains how to mathematically propagate errors into ground plane covariance predictions. The capability of ULEM to predict absolute horizontal and vertical accuracy of air-to-ground lidar data was demonstrated using check points surveyed with GPS combined with standard metric analysis methods. Horizontal or Circular Error at the 90th percentile (CE90) and vertical or linear error at the 90th percentile (LE90) were computed, consistent with National Geospatial-Intelligence Agency (NGA) practices, and approximately 90% of the estimated check point accuracies were determined to be within predicted bounds, indicating the potential utility of this model for practical use. [Fig. 1](#) shows the concept of the ULEM and how the error in horizontal and vertical accuracy leads to position uncertainty following an ellipsoidal shape.

An ongoing challenge in the adoption of a lidar error model is incorporating metadata required to support such models into standard lidar file formats. ULEM metadata variable length records (VLRs) have been developed that are compliant with the ASPRS-developed LAS specification. Similarly, ULEM metadata is defined in an optional header in the BPF3 specification adopted by the National Center for Geospatial Intelligence Standards in 2015. At present, ULEM is used within specialized government

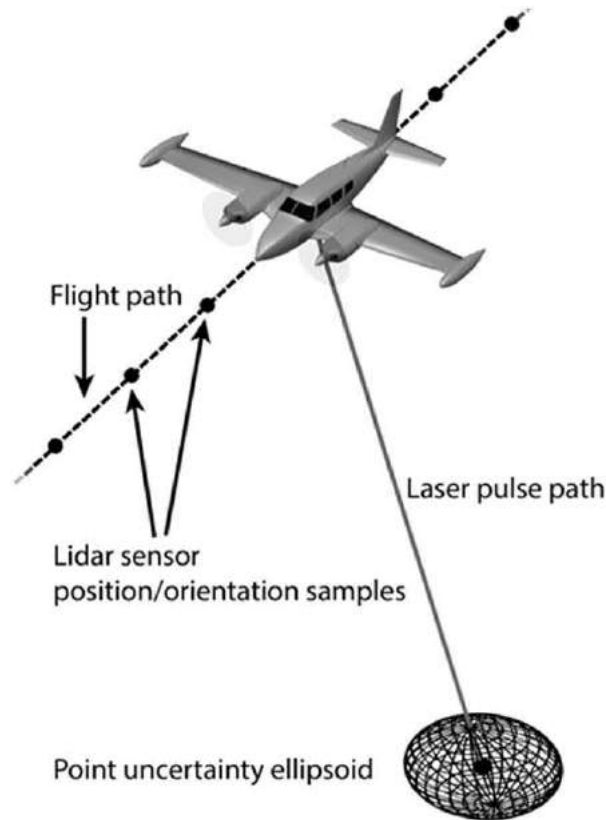


Fig. 1 Universal Lidar Error Model (ULEM) sensor-space error modeling.

communities. With the widespread proliferation of lidar and other 3D data from a variety of sources, development of common standards for metadata used to support error modeling and error propagation is becoming increasingly important.

Measuring Point Cloud Resolution

Horizontal resolution of a lidar point cloud refers to the minimum distance between two adjacent objects that are still resolvable as unique objects. Resolution is often assumed to be equivalent to NPS or GSD, but this is often not the case. For a conventional lidar system, horizontal resolution is largely determined by flight altitude and laser beam divergence. For a system with a focal plane array detector, resolution is also determined by pixel pitch. For photon-counting lidar, the resolution of the point cloud may be degraded by the coincidence processing algorithms used to filter noise. For any point cloud generated by aggregating multiple spatially coincident datasets, resolution may be degraded by sensor calibration error or by registration error. Measuring horizontal resolution explicitly can be an important factor in determining a sensor's capability to produce an interpretable 3D image.

Stevens *et al.* (2011) report methods for estimating resolution achievable by a lidar sensor. Among other methods, horizontal resolution was determined by calculating the contrast transfer function (CTF) based on lidar point densities both on and between pairs of closely spaced reflective panels from an array of panels with varying width and spacing. Visual inspection of panel resolvability in the point clouds was shown to be consistent with the Rayleigh criterion, indicating the utility of this approach. This method for determining resolution requires significant oversampling for reliable calculation and unique targets in the lidar coverage area. While this method provides a useful tool for lidar system calibration and characterization, it is not practical for general use to assess product quality. Methods for calculating the edge spread function (ESF) and line spread function (LSF) were also proposed for estimating horizontal precision using linear features such as building edges which would not require unique calibration targets in the lidar coverage area. However, for reliable calculation these methods also require significant oversampling, and the analytical assumptions associated with these methods do not hold for the broad range of modern lidar system designs. While resolution remains difficult to explicitly measure except in controlled environments, new methods for measuring relative horizontal accuracy are discussed in the next section that may be sufficient to characterize the horizontal precision of lidar point clouds.

Accounting for Relative Horizontal Accuracy

The ASPRS lidar calibration/validation Working Group led by USGS recognizes that current standards are not sufficient to capture the relative horizontal accuracy of lidar data. To address this shortcoming, Sampath *et al.* (2014) define data quality measures

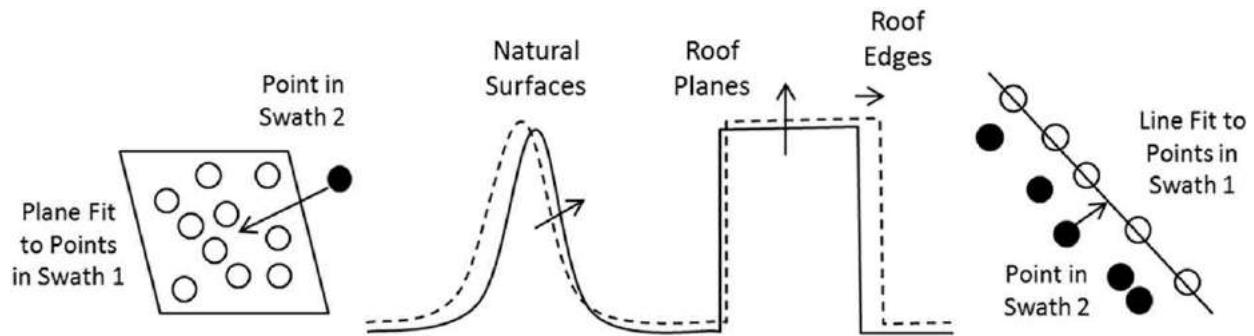


Fig. 2 Relative horizontal and vertical errors with overlapping lidar swaths capture point-to-plane differences for sloped surfaces and flat roof planes and point-to-line differences for roof edges.

(DQM) for determining inter-swath goodness of fit (**Fig. 2**) based on quality control methods proposed by [Habib et al. \(2010\)](#). These measures have been proposed to verify the quality of lidar system calibration and other sources of relative horizontal error in the data. For natural sloped terrain surfaces, a DQM measures the point to plane distance for conjugate points in overlapping swaths. Since a lidar system is unlikely to measure the same exact point in two swaths, a tangential plane is fit to each point in one swath and the point to plane distance is computed rather than point to point distance. For roof planes, conjugate planes are extracted from two swaths, and the distance from the centroid point from one swath to the plane from the other swath is measured. For roof edges, lines are fit to the conjugate edges and the distance from the centroid point in one line is measured to the conjugate line. These metrics do not rely on additional ground control or unique targets in the scene and so can be practically used for routine assessment of relative accuracy; however, they require multiple overlapping swaths which may not always be available. The working group is currently evaluating prototype tools that implement the DQMs before making formal recommendations for their use. Metrics like these will become increasingly important as more complex lidar system designs requiring more careful calibration and post-processing become commercially available.

Characterizing Foliage Penetration Performance

As reviewed by [Wulder et al. \(2012\)](#), multiple-return lidar provides a useful capability for measuring tree height, volume, and biomass for forestry inventory. Federal Emergency Management Agency (FEMA) [Lidar Specifications for Flood Hazard Mapping \(2009\)](#) suggest that high point densities may enable satisfactory data collection under foliage canopy. The USGS lidar base specification requires a minimum ANPD of two pulses per square meter (referred to by USGS as quality level 2) for lidar to be suitable for the 3DEP. For forestry applications requiring canopy penetration, 4–8 pulses per square meter (quality level 1 in the base specification) is commonly recommended in the literature. However, ANPD is insufficient to fully characterize foliage penetration performance.

For high obscuration environments (e.g., multiple layers of foliage canopy), [Burton \(2010\)](#) examines the value of line of sight angle diversity to improve the likelihood of canopy penetration. Both linear mode ([Roth et al., 2007](#)) and photon counting ([Vaidyanathan et al., 2007](#)) lidar systems have been developed that incorporate a gimbal to enable multiple off-nadir line of sight angles that better exploit the nonuniform gap structure in foliage canopy and improve foliage penetration performance. Current rules of thumb based on average point density also do not account for saturation in modern single photon detection lidar systems as described by [Henriksson \(2005\)](#) or the influence of filtering in coincidence processing as discussed in [Stevens et al. \(2011\)](#). To better account for the increasing variety of lidar system designs, explicit measurement of point density on the ground under foliage canopy is desirable to accurately characterize performance and determine data utility.

Defining Density and Completeness in 3D

Current two-dimensional point density metrics are useful for evaluating coverage and finding voids in relatively flat regions, but unfortunately do not work in areas with significant 3D structure because the sample density and completeness of vertical surfaces is ignored. No standard definition of 3D point density exists yet, making evaluation of coverage difficult. In addition, the presence of obscurants means that complete coverage requires sensing from multiple viewpoints, a consequence that is not captured by 2D metrics.

Prior works such as those of [Popescu and Zhao \(2008\)](#); [Pyysalo and Sarjakoski \(2008\)](#) have evaluated 3D point density by using a grid of volumetric cells, known as voxels, and counting the number of points within each cell to derive the density. However, this makes the assumption that the entire volume is evenly sampled by the lidar which is typically not true due to many effects including lidar scan patterns and obscurations. A lack of points within a volume may be due to either an absence of objects or a lidar system being unable to sense certain objects. Ambiguity between those two scenarios makes evaluating 3D completeness and voids difficult, if not impossible, using existing point-based methods.

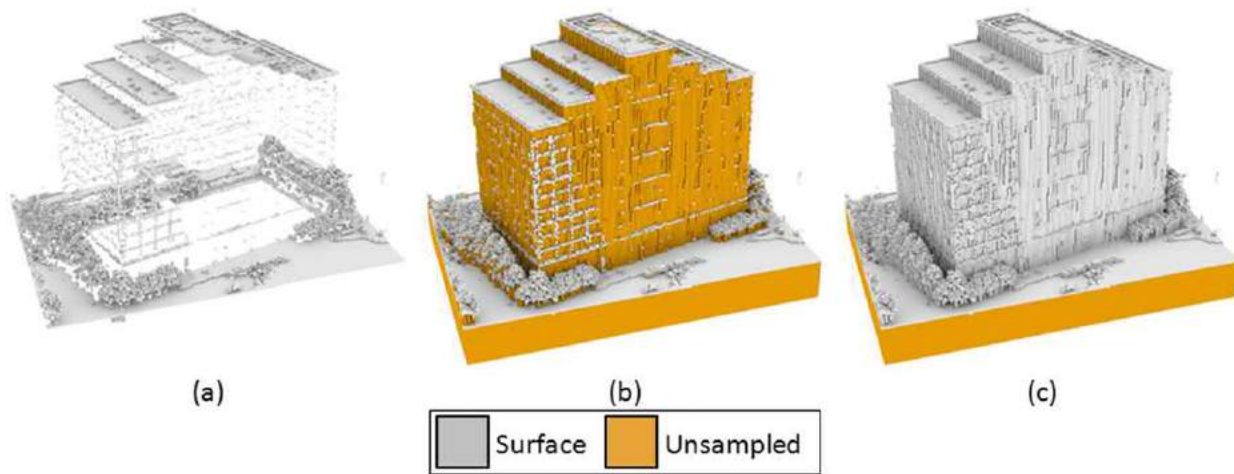


Fig. 3 Estimating 3D completeness of a dataset. The true surface of the model cannot be well-estimated from the sparse lidar points in (a). The orange unsampled areas in (b), identified using lidar path tracing, estimate the positions of missing surfaces. Given that a complete model should appear as (c), the completeness can be estimated by taking the ratio of the measured surface area to the estimated total surface area.

Several research groups have approached the problem of evaluating density by tracing the path of laser pulses through the dataset volume. The line segment path between the system and each point in the cloud point is computed using the coincident lidar instrument position records. Hagstrom (2014) defines the sampling density for a given voxel as the number of laser pulses intersecting that voxel divided by the voxel's geometric volume, regardless of whether the voxel contains any points itself. This is similar to the ASPRS two-dimensional definitions, where the requirement of using only first returns is effectively counting pulses rather than points.

A byproduct of this volumetric path tracing is that every voxel can be classified into one of three states: full due to containing points, known-empty due to containing no points but having pulses pass through, or unsampled due to being unreachable by the lidar (Fig. 3). This type of classification and knowledge of the unsampled regions has been used to improve methods such as object detection by Yapo *et al.* (2008) and change detection by Haas (2006).

Hagstrom (2014) build on this volumetric analysis idea to estimate completeness of a 3D dataset. Making the assumption that opaque objects always consist of an unsampled interior and known-empty exterior, a completely sampled object would always have at least one layer of full voxels between these two sections. Interfaces between adjacent full and known-empty voxels indicate a known surface, while interfaces between unsampled and known-empty voxels can only appear where surface points are missing. Taking the ratio of the missing surface area to the total surface area gives an estimate of the unsampled surface fraction (USF). While not yet adopted as an official metric, their work shows that 3D completeness can be estimated and that volumetric lidar representations can assist with analysis.

Evaluating 3D density and completeness remains an open problem but one which is increasingly important as lidar becomes more common and datasets increase in density. Though not yet standardized, volume-based metrics are one solution to evaluating 3D dataset quality, and show promise in this regard.

Conclusions

Common standards for metric assessment of lidar data are required to ensure that data quality is sufficient to meet end user needs and to enable objective and quantitative evaluation and comparison of lidar systems and capabilities as well as emerging non-lidar sources of 3D data. As lidar system designs continue to evolve and as lidar data continues to be used in new ways, community standards are also evolving to ensure that metrics continue to accurately characterize data quality. This article has reviewed current industry standards for metric evaluation of lidar data and discussed a few important emerging challenges and preliminary efforts toward addressing them.

See also: Multi-Dimensional Laser Radars

References

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery. In: Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop.

- Burton, R., 2010. Foliage penetration obscuration probability density function analysis from overhead canopy photos for gimbaled linear-mode and geiger-mode airborne lidar. In: Proceedings of the SPIE Laser Radar Technology and Applications XV.
- Haas, G., 2006. Three-dimensional change detection with the use of an evidence grid. ARL-TR-3916, Army Research Laboratory.
- Habib, A., Kersting, A.P., Bang, K.I., Lee, D.-C., 2010. Alternative methodologies for the internal quality control of parallel LiDAR strips. *IEEE Transactions on Geoscience and Remote Sensing* 48 (1), 221–236.
- Hagstrom, S., 2014. Voxel-based LIDAR analysis and applications. PhD Thesis, Rochester Institute of Technology.
- Henriksson, M., 2005. Detection probabilities for photon-counting avalanche photodiodes applied to a laser radar system. *Applied Optics* 44 (24), 5140–5147.
- Lidar Specifications for Flood Hazard Mapping. 2009. Appendix 4B: Airborne Light Detection and Ranging Systems, Federal Emergency Management Agency.
- Popescu, S.C., Zhao, K., 2008. A voxel-based lidar method for estimating crown base height for deciduous and pine trees. *Remote Sensing of Environment* 112 (3), 767–781.
- Pyysalo, U., Sarjakoski, T., 2008. Voxel approach to landscape modelling. *International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*. 37. Part B4.
- Rodarmel, C., Lee, M., Gilbert, J., *et al.*, 2015. The universal lidar error model. *Photogrammetric Engineering & Remote Sensing* 81 (7), 543–556.
- Roth, M., Hunnell, J., Murphy, K., Scheck, A., 2007. High-resolution foliage penetration with gimbaled lidar. In: Proceedings of the SPIE Laser Radar Technology and Applications XII.
- Sampath, A., Heidemann, H.K., Stensaas, G.L., Christopherson, J.B., 2014. ASPRS research on quantifying the geometric quality of lidar data. *Photogrammetric Engineering and Remote Sensing* 80 (3), 201–205.
- Stevens, J.R., Lopez, N.A., Burton, R.R., 2011. Quantitative data quality metrics for 3D laser radar systems. In: Proceedings of the SPIE Laser Radar Technology and Applications XVI.
- Vaidyanathan, M., Blask, S., Higgins, T., *et al.*, 2007. Jigsaw phase III: A miniaturized airborne 3-D imaging laser radar with photon-counting sensitivity for foliage penetration. In: Proceedings of the SPIE 6550, Laser Radar Technology and Applications XII.
- Wulder, M.A., White, J.C., Nelson, R.F., *et al.*, 2012. Lidar sampling for large-area forest characterization: A review. *Remote Sensing of Environment* 121.
- Yapo, T.C., Stewart, C.V., Radke, R.J., 2008. A probabilistic representation of LiDAR range data for efficient 3D object detection. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops.
- Zhu, X.X., Bamler, R., 2012. Demonstration of super-resolution for tomographic SAR imaging in urban environment. *IEEE Transactions on Geoscience and Remote Sensing* 50 (8), 3150–3157.

Further Reading

- ASPRS Positional Accuracy Standards for Digital Geospatial Data, 2015. The American Society for Photogrammetry & Remote Sensing, vol. 81, no. 3, pp. A1–A26.
- Baltsavias, E.P., 1999. Airborne laser scanning: Basic relations and formulas. *ISPRS Journal of Photogrammetry & Remote Sensing* 54, 199–214.
- Lidar Base Specification, 2014. Version 1.2, National Geospatial Program, USGS.
- Salvaggio, K., Salvaggio, C., Hagstrom, S., 2014. A voxel based approach for imaging voids in three-dimensional point clouds. In: Proceedings of the SPIE Geospatial InfoFusion and Video Analytics IV; and Motion Imagery for ISR and Situational Awareness II.

Multiple Input, Multiple Output, MIMO, Active Electro-Optical Sensing

Paul F McManamon, Exciting Technology LLC, Dayton, OH, United States

Jeffrey R Kraczek, University of Dayton: Electro-Optics, Dayton, OH, United States

© 2017 Elsevier Inc. All rights reserved.

Introduction

Weight and space constraints are frequently placed on active EO systems, including lidar systems, but high resolution imaging in cross range requires large optics due to diffraction. Large optics are frequently impractical to use because an imaging system that uses them has to be deeper than a small optical imager to maintain the same $f\#$. This contributes to large monolithic optical systems being heavy and expensive. In order to achieve high resolution at lighter weight and smaller size, methods have been developed to synthesize large apertures either by motion of a single smaller aperture or by an array of smaller apertures (Krause *et al.*, 2011; McManamon and Thompson, 2003).

Using multiple transmitter apertures and multiple receive apertures allows design flexibility not present in monolithic apertures. Synthetic aperture radar is comprised of a single moving aperture, so the transmit and receive apertures both move (Skolnik, 1980, 1990; Soumekh, 1999; Richards, 2005). A synthetic aperture lidar, SAL, uses motion of the single aperture to synthesize a larger effective aperture. For both SAR and SAL the synthetic aperture is almost twice as large as the actual flown distance, due to the angle of incidence being equal to the angle of reflection. This has been experimentally demonstrated over decades in the microwave region, and recently in the optical regime, when SAL, was demonstrated (Krause *et al.*, 2011; Beck *et al.*, 2005). Duncan provides a cross range resolution equation for spotlight-mode SAL:

$$\delta = \frac{R\lambda}{2L + D} \quad (1)$$

where D is the size of the aperture, R is the distance to the object, λ is the wavelength, and L is the distance moved (Duncan and Dierking, 2009). As a result, the effective size of a synthetic aperture based on motion from a diffraction point of view is:

$$D_{\text{eff}} = 2 * L + D_{\text{real}} \quad (2)$$

For RF systems the size of the real aperture is neglected, making the effective aperture twice as large as a monolithic aperture of width equal to the distance flown. The real aperture in RF systems is often on the order of a meter compared to the distance moved, which is multiple kilometers. In an EO system, the real aperture can be a significant fraction of the flight distance, so the D_{real} term is retained. For example, a SAL might have a real beam size of 15 cm and a flight distance of 1 m.

In order to synthesize a large aperture the returned field, including both amplitude and phase, must be measured or estimated. Optical carrier frequencies are much higher than can be measured directly by detectors. For example, a 1.5 μm wavelength has a carrier frequency of 200 THz, orders of magnitude higher than even the highest bandwidth detectors can measure. In order to measure phase as well as amplitude we will need to use a local oscillator, LO, to beat against the return signal. We can choose to have a local oscillator that aligns with the return signal, but uses a different frequency than the transmitter. The two return signals beat against each other to create an intermediate frequency, IF, return. If the LO and the transmitted signal are stable, then the IF return will have the same phase difference as the carrier return. This is called temporal heterodyne, and allows us to measure phase. Alternately we can use spatial heterodyne, sometimes called by the broader term of digital holography, to measure the return phase of the signal. For spatial heterodyne we use an LO that is the same frequency as the transmitter signal, but is offset in angle from the return. This creates a spatial beat, generating fringes across the receive plane, which allows us to measure spatial variations in phase. Both temporal and spatial heterodyne can use either pupil plane or image plane imaging to measure the return fields. When you measure the field, it is possible to convert from image to pupil plane and back, as well as to digitally adjust focus or correct for aberrations.

Conceptual Discussion of Multiple Input, Multiple Output Lidar

Instead of using motion to synthesize a larger effective aperture, such as done in Synthetic Aperture Lidar (SAL), Multiple Input, Multiple Output (MIMO) active EO sensing uses multiple physical sub-apertures to synthesize a larger effective aperture. An array of receive-only sub-apertures can synthesize an effective aperture as large as the receive array if the field across the array can be measured, or estimated. With an array of transmit and receive sub-apertures, even more flexibility is obtained, so long as it is possible on receive to identify which transmitter each photon initially came from. MIMO can be used in both dimensions, or in only one dimension. It can be combined with SAL, with SAL synthesizing a larger pupil plane aperture in one dimension and MIMO in the other dimension.

One effect from multiple transmit and receive sub-apertures is increased angular resolution. This is similar to the angular resolution increase from motion based synthetic aperture sensors. Instead of motion, an array of n sub-apertures that both transmit and receive can be used. Alternately we can have separate arrays for transmitting and for receiving. For 9 sub-apertures in a row, with both transmit and receive from each aperture, the array will have a diffraction limit consistent with a monolithic

aperture that is 1.89 times as large in diameter as the array. This is because 8 sub-apertures are equivalent to the distance moved, L , in Eq. (1) while 1 sub-aperture is equivalent to the real aperture, D_{real} . If transmission occurs from one sub-aperture in the middle of an array, the receive aperture array is effectively in its normal location. If the transmit beam is up one sub-aperture, it is as though the receive aperture were moved down one sub-aperture. If all of the sub-apertures transmit the result is something like what is shown in Fig. 1, where the lighter color linear arrays indicate the perceived location of the linear receive array, depending on which transmit sub-aperture photons are emitted from. The dark color column shows the actual location of the arrays. The lighter columns show the coordinate transformed location of the receive array, based on which aperture is emitting photons. We need to do a coordinate transformation as though all photons are emitted from a single sub-aperture before we can add the fields. The full extent of the linear array is 1.89 times larger, but it is sampled more near the middle of the effective aperture, represented by 9 samples in the middle down to one on either end. This is similar to an apodized aperture.

We could use fewer transmit apertures rather than using them all, and still have the same maximum resolution, but it would reduce the intensity at certain spatial frequencies. The apodizing effect would not be as significant in that case.

An alternate view of the increased resolution is provided in Fig. 2.

The main thrust of Tyler (2012) is to argue that we have multiple ways to obtain the same optical distances because of the multiple receive and transmit apertures. This allows us to solve for the difference in path lengths between the sub-apertures, and to adjust out that optical path difference. This can be done closed form, without iteration.

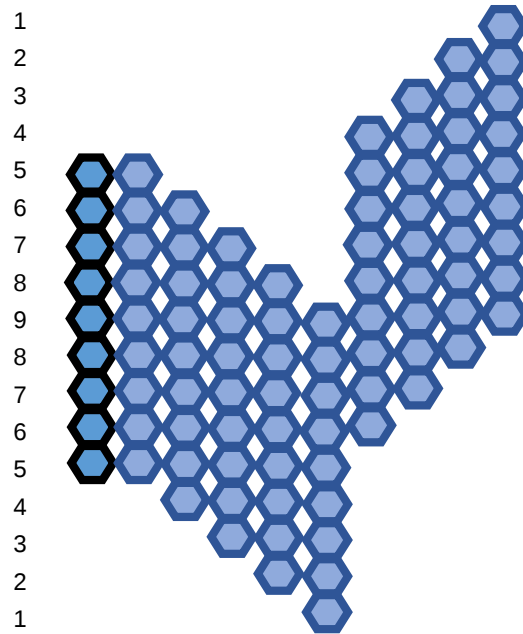


Fig. 1 Effective receiver aperture placement based on transmit sub-aperture utilized. Dark color shows actual location of the arrays. Light color shows perceived location. The number on the left shows how many received sub-apertures are perceived to be at that location.

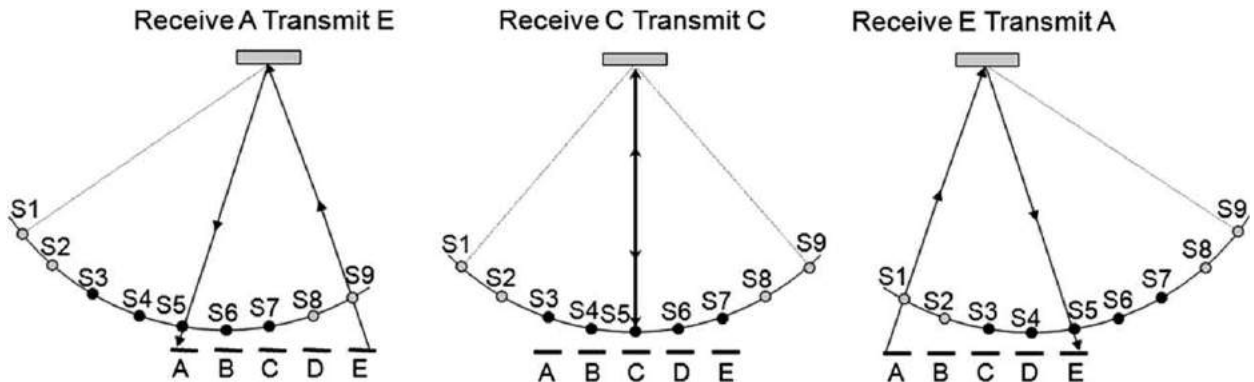


Fig. 2 Speckle measurements create a synthetic aperture that is almost twice as large as the array. The diagrams in this figure illustrate the nature of the speckle return from the object. Taken from Tyler, G.A., 2012. Accommodation of speckle in object-based phasing. Journal of the Optical Society of America A 29 (5), 722–733

Closed Form MIMO

Rabb *et al.* (2011) only use one receive aperture, but multiple transmit apertures. If transmitters are spaced at distances less than the size of a receive sub-aperture diameter, it provides sampling that can allow a closed form solution to differences in atmospheric path across the sub-aperture array (Fienup, 2000). An example of the resulting effective receiver pupil using transmitter diversity is shown in Fig. 3. Because of the overlap, one can solve in a closed-form manner for the phase between transmit sub-apertures, allowing its use to compensate for atmospheric phase disturbances in the pupil plane.

Rabb showed that he could calculate phase differences between transmit apertures by using multiple closely spaced transmit apertures with a single receive aperture, by using a linear systems of equations in conjunction with three transmit apertures. In this method, they used bivariate expansion to model intra-aperture phase aberrations. Similarly, Gunturk showed closed form phasing to remove intra-aperture aberrations for a single receive aperture using three transmit apertures but he used Zernike polynomials instead of a bivariate expansion (Gunturk *et al.*, 2012). The wave front difference was used to calculate the coefficients of up to seventh order Zernike polynomials to correct for the phase errors. Both of these methods deal with intra-aperture aberrations, but do not address the need for image phasing in a multi-receive aperture imaging system.

Krazcek *et al.* (2016) showed intra and inter-aperture phase aberrations can be solved in a closed form manner similar to Rabb and Gunturk. The intra-aperture aberrations are removed from the field first using an individual receive aperture and an array of three transmit apertures. Krazcek also considered a fourth transmitter widely separated from the three transmitter cluster. This will greatly extend the size of the synthetic aperture and will increase the frequency response cutoff of the system by approximately double, because the fourth transmitter aperture is placed about the width of the aperture array away from the set of 3 transmit apertures. Krazcek analytically calculated the MTF of this MIMO array, and then “experimentally” verified it in simulation using the slant edge MTF approach (Reichenbach *et al.*, 1991).

To perform inter-aperture closed form phasing, intra-aperture phase errors must be removed. The method used by Krazcek closely follows Gunturk’s work, by using three transmit apertures to remove the inter-aperture phase errors.

The left side of Fig. 4 shows an example of a multi-transmitter, single receive aperture system. The receive aperture is light green and the three transmitter cluster is magenta. The cluster is set to the vertices of an equilateral triangle with side lengths equal to the radius of the receive apertures. The right side of 4 shows the measured received fields, referenced to a common coordinate system. This is similar to Fig. 3, taken from Rabb. The overlap regions of the received fields in the pupil plane due to the different transmitter positions are shown in yellow and red. The outer edge of this image is the shape of the synthetic pupil. The synthetic pupil results from all the different fields being referenced to a common coordinate system. This larger synthetic pupil gives rise to a high resolution image. As described in Rabb *et al.* (2010), a shift in the transmitter location has an equal result on the movement of the placement of the received field in the synthetic pupil (Rabb *et al.*, 2010). This necessitates a coordinate change to compensate for different transmitter locations before the pupil plane fields can be coherently added. The synthetic pupil is built from the individual pupil fields being registered into the correct location in post processing. Pupil fields are assembled by Krazcek using known transmitter position since this work was done in simulation.

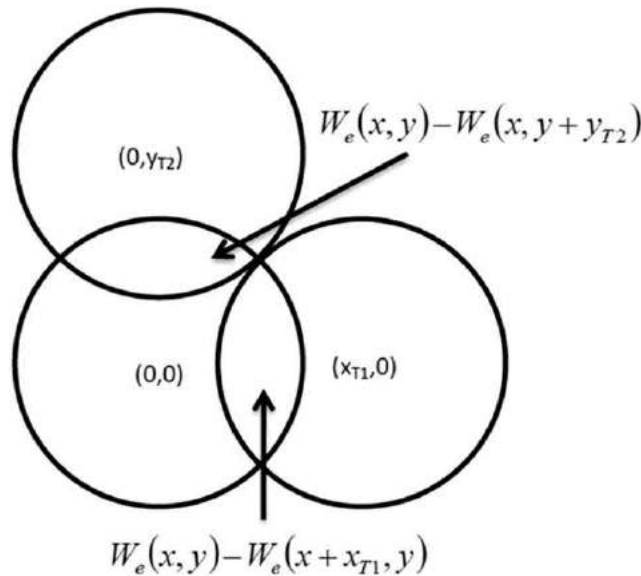


Fig. 3 Multiple sub-aperture overlap using 3 transmitters in a pattern about half a receive sub-aperture diameter apart. When images are adjusted to account for transmitter location, the receiver sub-aperture pupils overlap as shown. Reproduced from Rabb, D.J., Stafford, J.W., Jameson, D.F., 2011. Non-iterative aberration correction of a multiple transmitter system. *Optics Express* 19 (25), 25048–25056

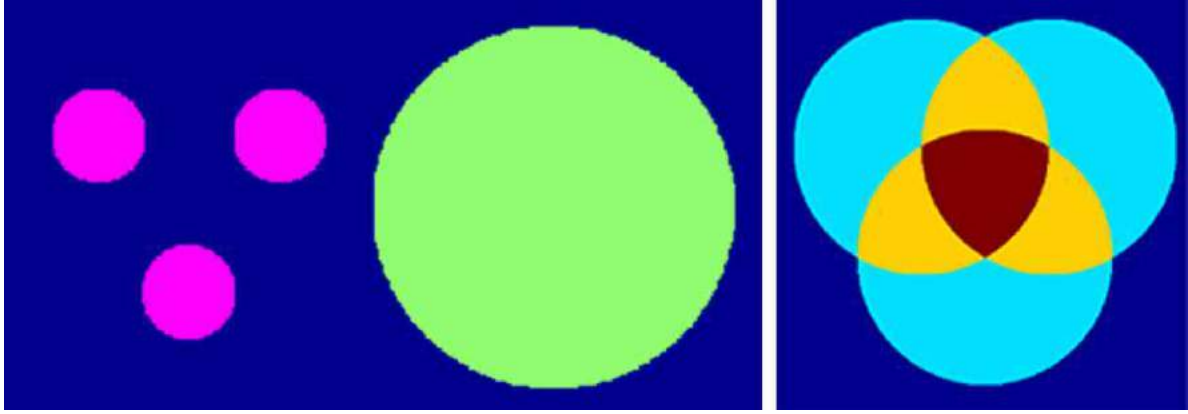


Fig. 4 Left: Single receive aperture, light green, with three transmitter apertures, magenta. Right: Field overlap due to a single receive aperture and three transmitter cluster.

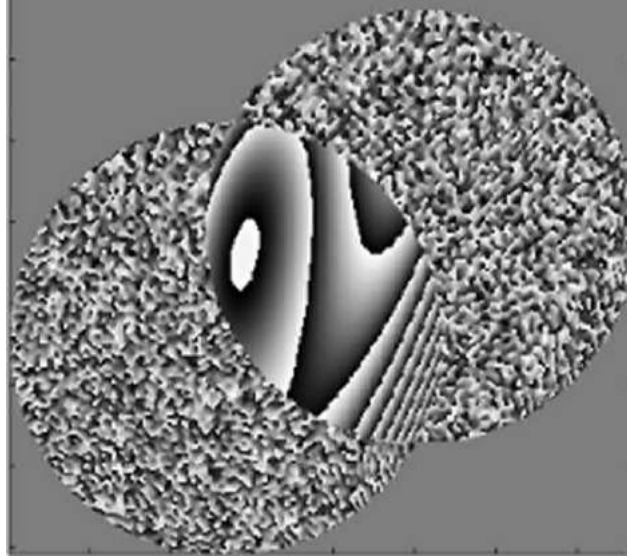


Fig. 5 Difference between the wave fronts of two overlapped pupil fields. The fringed section in the middle is where the two fields overlap.

The fields incident on the receive aperture can be described by

$$U_i(x, y) = P(x, y) \exp(j2\pi W_e(x, y)) U_b(x - x_i, y - y_i), \quad i = 1, 2 \quad (3)$$

where $P(x, y)$ is the pupil function of the receive aperture, $W_e(x, y)$ is the wave front error over the aperture, $U_b(x - x_i, y - y_i)$ is the received field with the subscript i referring to the i th transmitter. Defining $W_b(x, y)$ to be the wave front of $U_b(x, y)$, the received wave front is

$$W_i(x, y) = W_e(x, y) + W_b(x - x_i, y - y_i), \quad i = 1, 2 \quad (4)$$

The object wave front is not well defined and it is desirable to calculate the wave front aberrations independent of the intended target. As it is necessary to register the different fields based on the field shifts caused by transmitter positions into a synthetic pupil plane, the numerical dependence on the object wave front can be removed. Taking the difference of the overlapping areas in the received wave fronts gives

$$\begin{aligned} \Delta W(x, y) &= W_1(x - x_1, y - y_1) - W_2(x - x_2, y - y_2) = (W_e(x + x_1, y + y_1) + W_b(x, y)) \\ &\quad - (W_e(x + x_2, y + y_2) + W_b(x, y)) = W_e(x + x_1, y + y_1) - W_e(x + x_2, y + y_2) \end{aligned} \quad (5)$$

Due to the coordinate shift, the wave front difference is a difference between the wave front errors, which can be calculated and removed. [Fig. 5](#) shows the difference of the wave front of two overlapped pupil plane fields. The fringed area is where the two field overlap. Notice that this also removes speckle within the region of overlap.

Using the wave front difference, Gunturk is able to remove the intra-aperture phase errors using a set of over-determined linear equations, solving for the Zernike polynomials describing the phase errors across a circular pupil. His method does not work for inter-aperture phase errors because the Zernike coefficients are not necessarily the same between two apertures. The zeroth and first order, piston, tip, and tilt, cannot be solved for in the method describe above because they do not show up in the wave front difference using the same receive aperture. This is because piston is a flat phase and there is no difference in that term across the aperture. Tip and tilt terms have constant slopes across the aperture and therefore also cancel out when taking the difference of the two wave fronts in the same aperture.

A different method must be developed to correct for the inter-aperture tip-tilt-piston aberrations. The top of 6 shows a three transmitter and five aperture array. This is an example of a MIMO system that can be used for inter aperture closed form phasing. The bottom of 6 shows the pupil plane overlap pattern for the system shown in the top 6. Each individual receive aperture provides coverage in the same pattern as shown in 4. There is significantly less overlap between each aperture, but it still exists. The receive apertures should be located as close together as is physically permitted to allow for greater inter-aperture overlap. This greater overlap is beneficial in the registration process. We will define the wave front difference as the wave front error across the overlapped apertures. Assuming that the intra aperture phase errors have been removed, the wave front difference is now defined as

$$\Delta W(x, y) = W_{e,i1,j1}(x, y) - W_{e,i2,j2}(x, y) \quad i1 \neq i2, j1 \neq j2 \quad (6)$$

where $i1$ and $i2$ are different transmit apertures and $j1$ and $j2$ are different receive apertures (Fig. 6).

The left side of 7 shows an example of the wave front difference between two receive apertures. The wave front difference shows only the relative difference between the phase errors on the two apertures. On the left is the wave front difference after intra-aperture corrections have been made and on the right is the same wave front difference after tip-tilt corrections have also been made, resulting in just piston remaining. The gradient can be used on the area between phase wraps to gain an estimate of the tip and tilt between the two apertures. This estimate can then be used to correct tip and tilt. At this point, all that is left is piston. The right side of Fig. 7 shows an example of the wave front difference after all other phase errors are removed, resulting in just piston phase error being left. The value of the dark section is the piston phase. In a noise free environment both the piston and the tip and tilt corrections can be calculated with very few pixels. In a noisy system more pixels will need to be used. When correcting for these inter-aperture phase errors, one aperture is set as a reference to which all others are corrected. The corrections are daisy chained from the reference field to the field farthest away. As seen in 7, only two transmit apertures are required for inter-aperture phasing. Three transmit apertures are still used for intra-aperture phasing and to facilitate inter-aperture phasing with 2D receive aperture arrays.

Adding a fourth transmitter can increase the size of the synthetic pupil significantly, depending on the location of the fourth transmitter. The top of 8 shows a four transmitter, five receive aperture system. The bottom of 8 shows the shape of the synthetic pupil. The set of five circles on the left hand side comes from the fourth, widely separated, transmitter. Those fields do not need to be overlapped since they come from the same apertures that already had the intra and inter-aperture phase aberrations removed. The exception is that there needs to be overlap between at least one pupil field from the fourth transmitter with one of the pupil

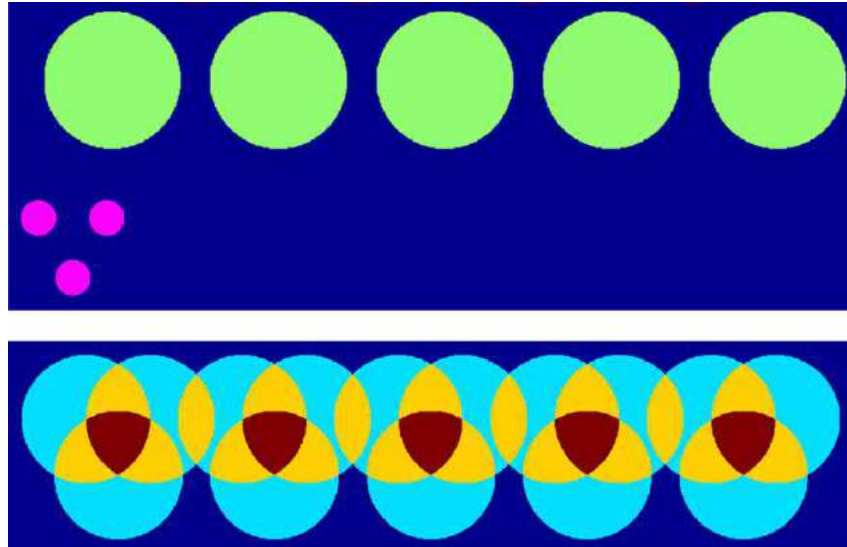


Fig. 6 Top: Example of a three transmitter and five aperture array. The red circles are unfilled receive apertures, the light green circles are filled receive apertures and the magenta circles are transmit apertures. Bottom: Example of the field overlaps from the three transmitter and five aperture array shown on top. The light blue areas are where only one field is present, the yellow are areas of two fields overlapping, and the red is where three fields overlap.

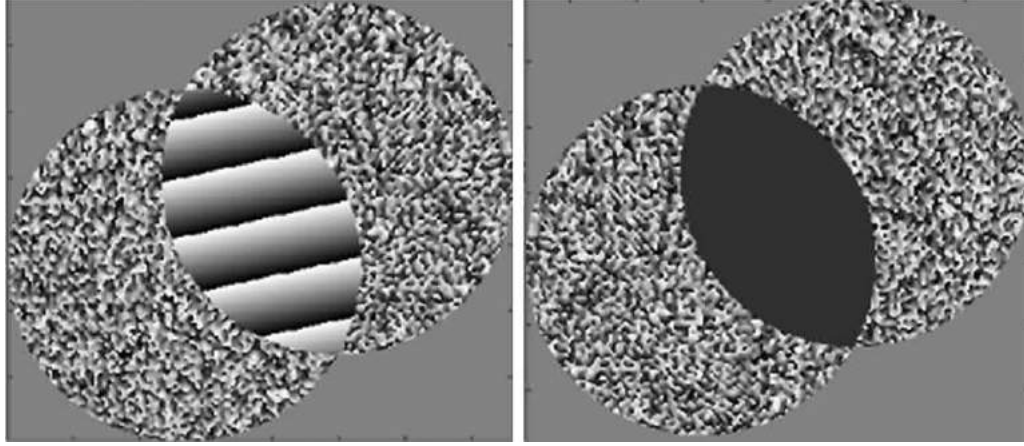


Fig. 7 Wave front difference between two pupils coming from two different apertures. Left: A Piston, tip, and tilt error remains. Right: Piston error remains.

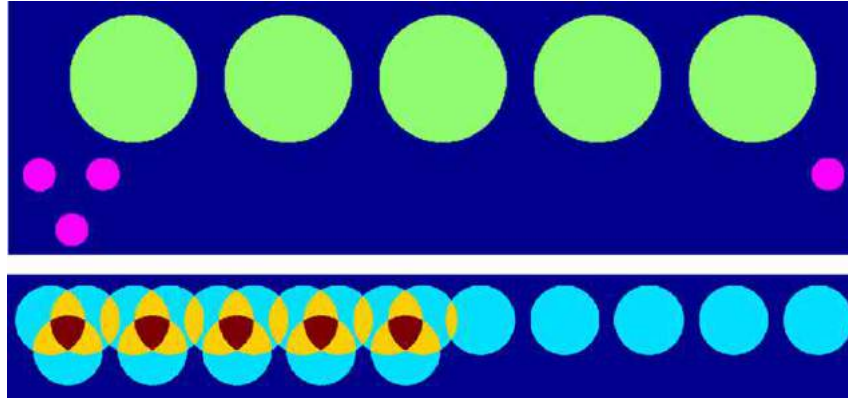


Fig. 8 Top: Example of a four transmitter and five aperture array. The red circles are unfilled receive apertures, the light green circles are filled receive apertures and the magenta circles are transmit apertures. Bottom: Example of the field overlaps from the four transmitter and five aperture array shown on top.

fields from the initial three transmitter cluster. In this example, the farthest left field from the fourth transmitter and the farthest right field from the transmitter cluster are overlapped. This overlap is used to obtain an additional piston correction for all pupil fields from the fourth transmitter and allow for field registration. Sub-pixel registration on the speckle in the pupil plane is used. It is assumed exact positions of all transmit apertures are known but overlap is maintained. An additional piston correction will be necessary to account for the different position of the fourth transmitter fields from where the initial inter-aperture measurements were made; movement of the pupil fields cause an additional piston term in the wave front difference, due to tip-tilt, that needs to be corrected (Fig. 8).

Fig. 9 shows the theoretical modulation transfer function (MTF) in the horizontal dimension for the synthetic aperture in 6 on the left and the synthetic aperture in 8 on the right. Notice that the cutoff frequency for the three transmitter and five aperture array is about half the cutoff of the array with four transmit apertures and five apertures. This is because the synthetic pupil size is about half as large in that dimension.

Trade Space of Current MIMO

MIMO techniques as described here can be implemented either using temporal or spatial heterodyne (digital holography) techniques. It will however be much easier to tag the emitted transmitter signals, allowing simultaneous transmission, if high bandwidth temporal heterodyne is used, since high bandwidth tagging schemes can then be used. RF MIMO techniques that use multiple simultaneous phase centers have been developed (Coutts *et al.*, 2006).

The angular resolution of an array of sub-apertures can be almost twice the resolution of the diffraction limit for a monolithic aperture. In addition, atmospheric turbulence between the object imaged and the sensor can be very quickly and accurately compensated because we do not need to use an iterative approach to finding the required optical path correction to compensate for the turbulence. This compensation is relatively straightforward for turbulence in the pupil plane, but will be more difficult for

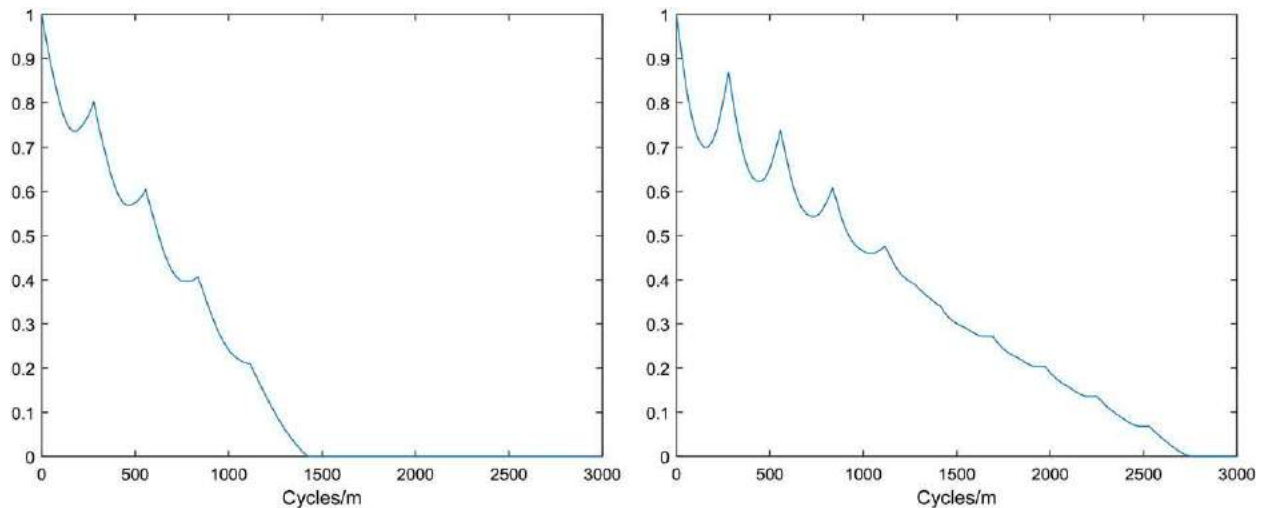


Fig. 9 MTF plots for the arrays shown in 5 and 7. The X-axis is in cycles per mm. Left: Horizontal MTF for array shown in 6. Right: Horizontal MTF for array shown in 7.

volume turbulence. A significant limitation is that the received signal is captured in a smaller receive aperture area. The required laser power will increase by the ratio of the area of the monolithic aperture to the received aperture array area. Also, if the angular resolution is greater than a monolithic aperture of the same size, the cross section of each pixel is lower because of the increased resolution. This results in either more required power, or lower signal to noise. Since each new transmit aperture adds power this can balance out.

Arrays of high temporal bandwidth detectors will be very helpful in implementing MIMO techniques for EO imaging. The papers cited sequenced through the multiple transmitters because they implemented MIMO using a digital holography/spatial heterodyne approach to imaging, using low bandwidth framing detector arrays. If high bandwidth detector arrays and a temporal heterodyne approach to imaging are used then it should be possible to simultaneously emit multiple tagged transmitter beams, and to have each receiver be able to distinguish the transmit aperture any photon came from. High-bandwidth detectors can allow high bandwidth modulations to be imposed on each transmitted beam, and sorted on receive. Temporal heterodyne arrays will need to be AC-coupled, or have high dynamic range, or have high sensitivity such that temporal heterodyne can be implemented using a relatively weak LO.

A second technical hurdle to overcome is volume turbulence. Techniques for calculating, and compensating for, volume turbulence still need to be developed.

MIMO technology will allow imaging with high angular resolution using a much lighter and more compact aperture arrays than a monolithic aperture. An array of small sub-apertures can be much thinner and lighter than a monolithic aperture. Also, a MIMO approach can achieve almost twice the diffraction limited angular resolution of a monolithic aperture.

To be really useful in freezing the atmosphere, MIMO should be implemented using high bandwidth detector arrays, which still need to be further developed. There will be a digital implementation requirement to conduct the required calculations. There will be a need to have many different optical trains, complicating the optical system. Also, narrow line lasers will need to be used to do either spatial or temporal heterodyne. Higher power lasers will be required to image a given area using a MIMO array compared with imaging with a monolithic aperture because the spatial frequency cutoff of the measured signal is being increased more rapidly than the area of the pupil plane receive aperture.

Conclusion

This technology is well suited for long range imaging applications from air or space. MIMO could be used in the cross range dimension along with motion-based synthetic aperture imaging.

In summary, MIMO approaches for active EO sensing can, at a minimum, increase the effective diameter of an aperture array by almost a factor of two, and can allow multiple sub-apertures on receive to be phased using a closed-loop calculation of the phase difference between sub-apertures. This can compensate for atmospheric turbulence at least at some locations between the sensor and the imaged object.

References

Beck, S.M., Buck, J.R., Buell, W.F., *et al.*, 2005. Synthetic-aperture imaging laser radar: Laboratory demonstration and signal processing. *Applied Optics* 44 (35), 7621–7629.

- Coutts, S., Cuomo, K., Mcharg, J., Robey, F., Weikle, D., 2006. Distributed coherent aperture measurements for next generation BMD radar. In: Fourth IEEE Workshop on Sensor Array and Multichannel Processing, pp. 390–393.
- Duncan, B.D., Dierking, M.P., 2009. Holographic aperture ladar. *Applied Optics* 48 (6), 1168–1177.
- Fienup, J.R., 2000. Phase error correction for synthetic-aperture phased-array imaging systems. In: *Proceedings of SPIE 4123, Image Reconstruction From Incomplete Data* 4123, pp. 47–55.
- Gunturk, B.K., Rabb, D.J., Jameson, D.F., 2012. Multi-transmitter aperture synthesis with Zernike based aberration correction abstract. *Optical Express* 20 (24), 5179–5186.
- Kraczek, J.R., Mcmanamon, P.F., Watson, E.A., 2016. High resolution non-iterative aperture synthesis. *Optical Express* 24 (6), 6229–6239.
- Krause, B.W., Buck, J., Ryan, C., *et al.*, 2011. Synthetic aperture ladar flight demonstration. In: *CLEO: Applications and Technology 2011: Laser Applications to Photonic Applications*.
- McManamon, P.F., Thompson, W., 2003. Phased array of phased arrays (PAPA) laser systems architecture. *Fiber and Integrated Optics Journal* 22 (2), 79–88.
- Rabb, D.J., Jameson, D.F., Stafford, J.W., Stokes, A.J., 2010. Multi-transmitter aperture synthesis. *Optical Express* 18 (24), 24937–24945.
- Rabb, D.J., Stafford, J.W., Jameson, D.F., 2011. Non-iterative aberration correction of a multiple transmitter system. *Optical Express* 19 (25), 25048–25056.
- Reichenbach, S.E., Park, S.K., Narayanswamy, R., 1991. Characterizing digital image acquisition devices. *Optical Engineering* 30 (2), 170–177.
- Richards, M.A., 2005. *Fundamentals of Radar Signal Processing: Chapter **. New York: McGraw-Hill.
- Skolnik, M.I., 1980. *Introduction to Radar Systems: Chapter 14*, second ed. New York, NY: McGraw-Hill.
- Skolnik, M.I., 1990. *Radar Handbook, Chapter 17*, by Roger Sullivan, second ed. New York, NY: McGraw-Hill.
- Soumekh, M., 1999. *Synthetic Aperture Radar Signal Processing With Matlab Algorithms*. New York, NY: Wiley.
- Tyler, G.A., 2012. Accommodation of speckle in object-based phasing. *Journal of the Optical Society of America A* 29 (5), 722–733.

InGaAs Linear-Mode Avalanche Photodiodes

Andrew S Huntington, Voxtel, Inc., Beaverton, OR, United States

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

a	Random variable for the number of primary photoelectrons injected into a multiplier	NEP	Photoreceiver noise expressed as an equivalent input optical power level
a_{bg}	Random variable for the number of primary background photoelectrons	NEPh	Photoreceiver noise expressed as an equivalent number of input photons
a_{dark}	Random variable for the number of primary dark current electrons	n_h	Count of holes within the depletion region of a junction
a_{signal}	Random variable for the number of primary signal photoelectrons	N_Q	Photoreceiver noise expressed as an equivalent number of electrons at the APD output
BW	A modulation frequency bandwidth in Hz	n_{th}	Detection threshold expressed as an equivalent number of electrons at the APD output
F	Excess noise factor by which APD output variance exceeds the deterministic gain case	P_{RX}	Photoreceiver output distribution scaled to electron count at APD output
I_{bg}	Background photocurrent measured at the APD's terminals	q	Elementary charge
I_{dark}	Dark current measured at the APD's terminals	QE	Quantum efficiency
$I_{primary}$	Primary current injected into a multiplier	Q_{signal}	Output signal charge in units of electrons
I_{signal}	Signal photocurrent measured at the APD's terminals	R	Responsivity
k	Impact ionization rate ratio for electrons and holes (slower rate over faster rate)	$S_{I\ TIA}$	Input-referred noise current spectral intensity of an amplifier in units of $A^2\ Hz^{-1}$
M	Mean avalanche gain factor	$S_{I\ total}$	Total noise current spectral intensity of a photoreceiver
m	Random variable for the per-electron gain factor	S_I	Noise current spectral intensity
n	Random variable for the total number of electrons output by a multiplier	t	Time
N_{CTIA}	Input-referred charge noise of an amplifier in units of electrons	Γ	Euler gamma function
n_e	Count of electrons within the depletion region of a junction	v_{se}	Electron saturation drift velocity
		v_{sh}	Hole saturation drift velocity
		w	Junction depletion width
		λ	Optical signal wavelength

Introduction

Linear-mode (LM) avalanche photodiodes (APDs) are used to preamplify signal photocurrent in photoreceivers where amplifier circuit noise is the dominant noise component. Since amplifier noise is a fixed quantity that is independent of the APD's gain, the photoreceiver's signal initially increases faster as a function of APD gain than its total noise, improving measures of sensitivity such as the signal-to-noise ratio (SNR). However, LM APDs amplify the shot noise associated with charge carrier generation in the diode junction and also generate excess shot noise associated with the stochastic gain process itself. Since an APD's shot noise increases faster with gain than the amplified signal photocurrent, optimal photoreceiver sensitivity occurs at some finite gain that is determined by factors such as the intensity of the amplifier noise, the magnitude of the APD's dark current, the intensity of the optical signal, and the statistics of the APD's gain process. The interplay of photoreceiver noise components leading to optimal sensitivity at finite gain is illustrated in Fig. 1.

InGaAs LM APD Structure and Manufacturing

Several binary and ternary III–V compound semiconductor alloys are plotted in Fig. 2 according to their crystal lattice parameters and their bandgaps at room temperature. In the diagram, the binary alloys are plotted as points, connected by arcs representing the range of ternary alloy compositions which result from blending pairs of binary alloys. $In_{0.53}Ga_{0.47}As$ is the alloy composition aligned vertically with the triangle representing InP, which means they share the same lattice parameter. $In_{0.53}Ga_{0.47}As$ is favored for use in short-wavelength infrared (SWIR) detectors because its direct and relatively narrow (~ 0.74 eV) bandgap results in efficient light absorption out to a cutoff wavelength near $1.7\ \mu m$ and because comparatively thick ($\sim \mu m$) single-crystal light absorption layers can be grown with very low defect density, lattice-matched to InP substrates. Although non-lattice-matched InGaAs alloy compositions can be grown on InP substrates to push spectral response further into the SWIR, the use of non-lattice-

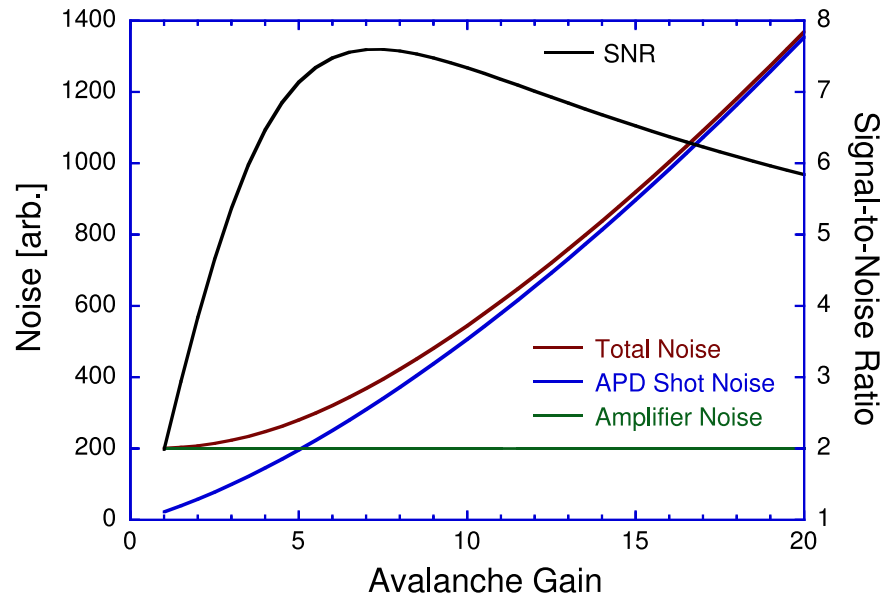


Fig. 1 Total noise, noise components, and signal-to-noise ratio of an avalanche photodiode (APD) photoreceiver vs. mean avalanche gain.

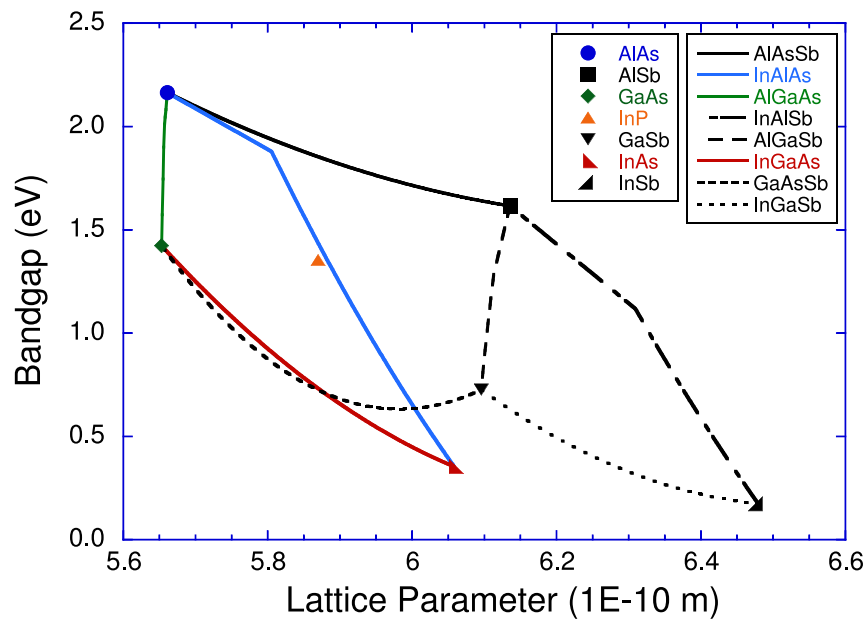


Fig. 2 Room temperature band gap vs. crystal lattice parameter for technologically important III-V compound semiconductor alloys.

matched alloys in APDs is not common, owing to the sensitive dependence of APD dark current on defect density. Dark current considerations also constrain selection of the alloys used for avalanche multiplication layers in InGaAs APDs to either InP itself or the lattice-matched ternary $\text{In}_{0.52}\text{Al}_{0.48}\text{As}$. Wider-bandgap alloys like InP or InAlAs are preferred for use in APD multiplication layers because their use minimizes dark current generation by band-to-band tunneling.

The single-crystal thin films from which InGaAs APDs are fabricated are grown on (100)-oriented or (100)-vicinal InP substrates by molecular beam epitaxy (MBE) or metal-organic chemical vapor deposition (MOCVD). Most InGaAs APDs share a common epitaxial structure in which an InGaAs light absorption layer and a charge carrier multiplication layer are separated by a space charge layer. Also called a field control layer, the charge layer assures that during operation, the electric field in the narrower-bandgap absorber will remain weak enough to limit dark current generation by tunneling when the electric field in the wider-bandgap multiplier is strong enough to generate a useful amount of avalanche gain. The charge layer of an InGaAs APD is the most critical to its function, and calibration of the charge layer's thickness and active doping concentration is the most difficult aspect of InGaAs APD manufacturing. Too many active dopant ions between absorber and multiplier result in an APD that reaches

avalanche breakdown at a reverse bias lower than that required to render it sensitive to optical signals, whereas too few result in excessive dark current generation by tunneling at reverse bias lower than that required to achieve useful avalanche gain. Common methods of doping calibration used in crystal growth labs, such as Hall effect measurements or electrochemical capacitance–voltage (ECV) profiling, typically do not have sufficient precision to calibrate InGaAs APD charge layers. Instead, prior to growth of production wafers, charge layer calibration is commonly accomplished by growth of complete APD layer stacks followed by quick-turn fabrication and current–voltage (I – V) characterization of test structures.

The separate absorption, charge, and multiplication (SACM) layer stack is sandwiched between anode and cathode layers. It is common to grow InGaAs APDs with the cathode side down, and often n-type InP substrates are selected so that electrical contact to the APD's cathode can be made through the wafer substrate. The ordering of the SACM layer stack between anode and cathode is determined by which alloy is selected for the multiplication layer. Because the absorption layer is physically separate from the multiplication layer, and because electrons drift toward the cathode and holes drift toward the anode of a reverse-biased diode, only one carrier polarity can be injected into the multiplier. The impact ionization rate of holes is higher in InP than that of electrons, but lower than that of electrons in InAlAs. Consequently, InP multipliers are grown on the anode side of the absorber, whereas InAlAs multipliers are grown on the cathode side of the absorber. The two common SACM stack orderings are illustrated in Fig. 3.

Detector elements are formed from the epitaxial thin film either by patterned diffusion to define the active area of each element, or by etching away the film surrounding each element. In the most common implementation of the diffused junction design, the anode layer of the APD thin film is not doped during crystal growth. Instead, the p-type dopant species – typically zinc – is introduced during wafer fabrication by diffusion through openings patterned in a dielectric mask layer that is deposited over the thin film. The maximum electric field strength inside the junction of a diode formed by patterned diffusion tends to occur at the edge of the diffusion because charge neutrality requires that the donor ions on the cathode side of the junction be balanced by an equal number of acceptor ions on the anode side. The upward curvature of the depletion region at the edge of the anode diffusion results in ionized acceptors facing ionized donors along a narrower frontage, concentrating the electric field (Fig. 4, left). Shallower anode diffusions result in tighter radii of curvature with higher maximum electric field strength. Multilevel diffusions or diffused guard rings surrounding the central anode can be used to minimize crowding of electric field lines at the perimeter of a

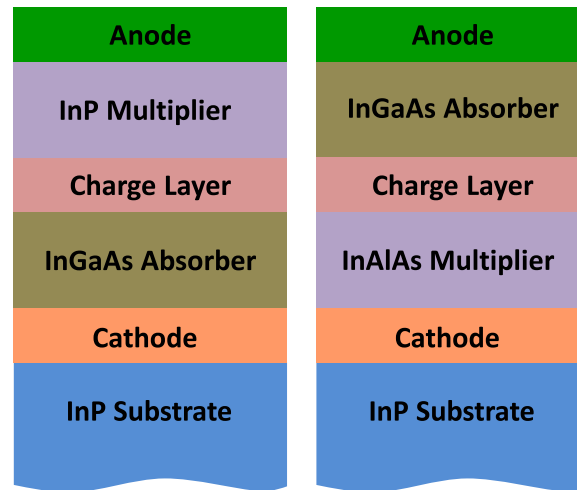


Fig. 3 Typical layer orderings for InP- and InAlAs-multiplier InGaAs avalanche photodiodes (APDs).

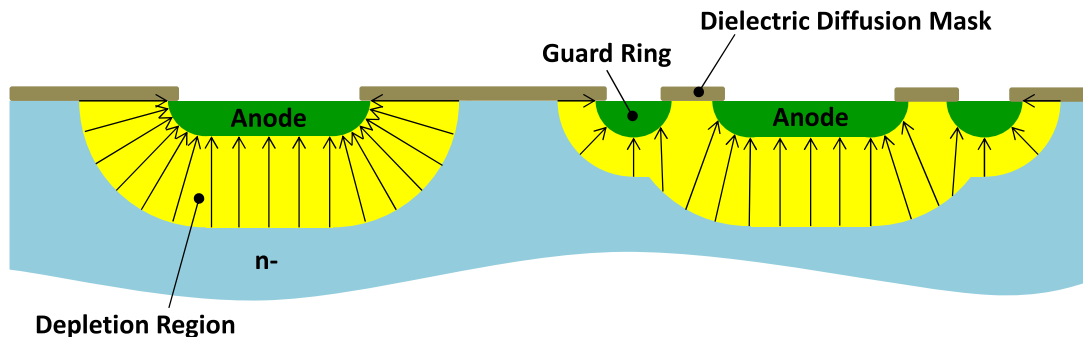


Fig. 4 Sketch of electric field lines in diffused junction photodiodes with (right) and without (left) guard rings.

diffused anode by reshaping the depletion region along the edge of the detector element, preventing edge breakdown (Fig. 4, right).

APDs fabricated by etching are made from thin films in which the top contact layer (typically the anode) is doped during crystal growth. Instead of defining the junction area by patterning the anode doping, the thin film outside the detector element area is etched away down to the substrate, leaving behind a mesa. Whereas the principal concern with diffused junction APDs is managing the electric field intensity at the perimeter of the patterned diffusion, it is sufficient that the sidewalls of an etched mesa APD be smooth, and slope gradually inward from the base of the mesa. Physical irregularities on the mesa sidewall – and especially notches where the slope of the mesa sidewall reverses, creating an overhang – can locally enhance the electric field and create a weak point at the junction perimeter. However, in general, the chemical cleanliness of the mesa sidewall is the principal manufacturing challenge for mesa APDs.

Most of an APD's dark current is generated in its InGaAs absorption layer because, having the narrowest bandgap in the structure, it presents the lowest energetic barrier to generation of charge carrier pairs, and because it is depleted during operation of the APD. In a diffused junction APD the InGaAs layer is safely buried under a wider-bandgap material like InP, and the depletion region intersects the wafer surface along a narrow band of wider-bandgap material at the anode perimeter. In contrast, a swath of depleted narrow-bandgap InGaAs is exposed along the sidewall of an etched mesa APD. Interruption of a crystal lattice by a surface creates a high density of mid-bandgap trap states. When the surface is of a narrow-bandgap semiconductor alloy and lies within the depletion region of a photodiode, a high dark current generation rate results. Chemical formation of materials with higher conductivity than the depleted diode junction is another problem at the mesa sidewall, as this can form a conductive path that shunts the diode. Both issues are addressed during mesa APD manufacturing by preparing mesa sidewalls that are as chemically clean – and devoid of sub-surface damage to the lattice – as possible, and then encapsulating the clean sidewall surface under an impermeable dielectric coating to protect it from environmental degradation. In some cases, chemical passivation treatments – often involving sulfide compounds – are applied to the mesa sidewall prior to encapsulation. Chemical passivation works by forming bonds at the semiconductor surface, the molecular orbitals of which lie outside the semiconductor bandgap, such that the resulting energy states do not behave as traps. The passivation treatment replaces trap states with inactive states and prevents traps from forming by tying up potentially reactive surface sites.

Fabrication of APDs also includes deposition of anode and cathode metal contact pads, and usually deposition of an anti-reflection coating. Different APD shapes and configurations are possible. The most common APD configuration is designed for illumination normal to the wafer surface, with light absorption occurring in a single pass through a comparatively thick ($\sim \mu\text{m}$) absorption layer. APDs of this class are commonly fabricated with active area diameters ranging from about 25 to 500 μm . Larger InGaAs APDs are seldom fabricated because detector capacitance, dark current, and manufacturing yield loss all scale linearly with junction area. However, hemispherical immersion lenses are sometimes affixed to APDs to increase their effective light-gathering area. APDs with thick single-pass absorbers may be top- or bottom-illuminated, depending upon whether the optical signal is delivered from the side of the wafer on which the epitaxial thin film was grown, or through the InP substrate. Top-illuminated InGaAs APDs tend to have superior responsivity at wavelengths shorter than the InP absorption band edge at about 950 nm, although the short-wavelength response of bottom-illuminated InGaAs APDs can be enhanced by physically removing the InP substrate. Bottom-illuminated APDs can be bump-bonded directly to application-specific integrated circuits (ASICs) to achieve lower interconnect parasitics than is possible with front-illuminated APDs of equal active area, and the bottom-illuminated configuration is always used for dense two-dimensional imaging arrays of APD pixels. A cross section through an example 25- μm diameter, bottom-illuminated, etched mesa APD pixel is sketched in Fig. 5 to illustrate the major structural parts of such a device. In Fig. 5 a ring-shaped cathode contact around the base of the APD mesa is connected to a pad on top of an adjacent inactive mesa for easy of bump-bonding of both anode and cathode connections. It is common for all the APD pixels of a multielement array to share a small number of common cathode connections.

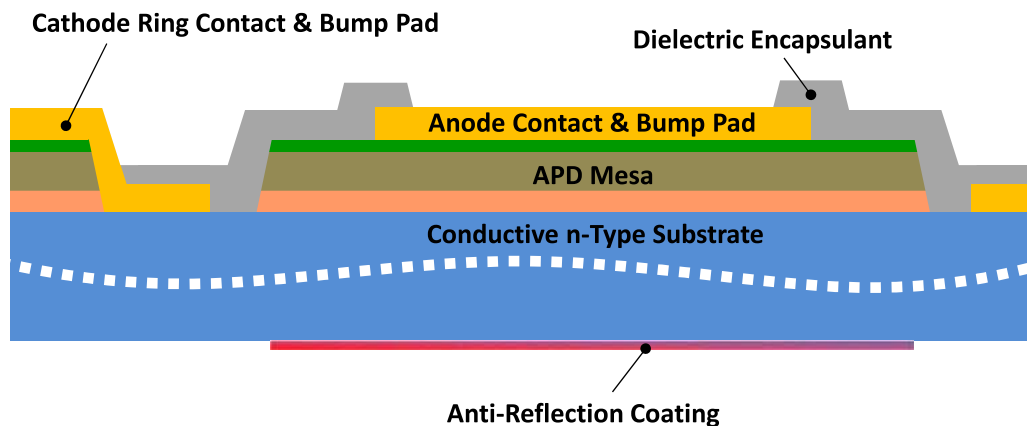


Fig. 5 Cross-sectional sketch of a representative etched mesa InGaAs avalanche photodiode (APD) pixel.

APDs designed for high-bandwidth operation (greater than a few GHz) must minimize their junction thickness in order to cut down the junction transit time. Minimizing the junction thickness minimizes the thickness of the InGaAs absorption layer, so alternative APD configurations like resonant cavity and waveguide designs are necessary to maintain high quantum efficiency (QE). Resonant cavity APDs are illuminated normal to the wafer surface but include mirrors above and below the standard SACM layer stack in order to create a vertical optical cavity that increases the effective number of passes an optical signal takes through the thin absorption layer. In waveguide APDs, the optical signal propagates in the plane of the thin film, so the absorption path length can be increased without increasing the junction transit time. Resonant cavity APDs have the advantage of comparatively easy optical coupling at the resonant wavelength, since the signal may be focused onto the surface of the wafer. Resonant cavity APDs can also be built with very small junction area, which limits capacitance. In comparison, waveguide APDs present the challenge of coupling the optical signal into an edge-facing facet of the in-plane waveguide, and potentially have higher capacitance owing to the larger footprint of the waveguide structure. However, the lattice-matched alloys from which epitaxial distributed Bragg reflectors (DBRs) can be formed for resonant cavity APDs do not provide very much refractive index contrast, necessitating gratings with a large number of periods which are difficult to manufacture. Also, resonant cavity APDs only work well at a single wavelength, since the DBR through which the signal enters is reflective at nonresonant wavelengths. Although these high-bandwidth APD configurations are potentially of interest for light and compact optical free space communications transceivers, APDs are not generally used for very high-bandwidth terrestrial telecommunications because erbium-doped fiber amplifiers (EDFAs) paired with gain-less photodiodes provide superior speed and sensitivity.

Some manufacturing issues have a bigger impact on dense APD pixel arrays than on single-element detectors. Localized morphological defects originate from thin film growth over dust or patches of native oxide remaining on imperfectly cleaned substrate surfaces, or from growth over regions on the substrate surface inadvertently damaged during the thermal desorption process that removes the native oxide. Fig. 6 is an amplitude mode atomic force microscope (AFM) image of an unusually large dust-related defect. Other morphological defects may be introduced during thin film growth as a metal droplet spat from an effusion cell, or as a patch of non-stoichiometric material created by a local temperature or pressure excursion outside the growth window for the compound semiconductor. When one of these defects occurs within the footprint of an APD's active area it can provide a current path that shunts the diode junction, causing the detector element to fail short. The area density of these defects on an APD wafer is usually on the order of 100 cm^{-2} , and APD active areas are small, so shorted single-element detectors can be screened out and rejected with very modest manufacturing yield loss. However, even when the probability is very low that any given detector element is defective, the likelihood of yielding large format APD arrays with 100% operable pixels drops rapidly with array format. For example, if the likelihood any given $30\text{-}\mu\text{m}$ -diameter pixel is defect-free is about 99.93%, the likelihood that every one of the 1024 pixels in a particular 32×32 -format array is defect-free is only about 48.5%; for 64×64 -format arrays the estimated yield of completely defect-free array die is only about 5.5%. Insofar as InGaAs APDs operate at voltages that greatly exceed the damage threshold of most CMOS processes and a shorted APD pixel drops most of that high voltage across the readout integrated circuit (ROIC) to which it is connected, it is usually necessary to engineer into the interconnect between APD and ROIC pixels a means of isolating shorted APD pixels.

Limited optical fill factor is another issue particular to APD arrays. Arrays of etched mesa pixels have the advantage of extremely good inter-pixel isolation but cannot make use of the portion of the optical signal which falls between pixel active areas. The dimensional tolerance of the semiconductor process used to pattern and etch pixel mesas determines some minimum separation between adjacent mesas. As the pixel pitch decreases, the dead area consumed by this minimum inter-mesa gap becomes a larger fraction of the total pixel footprint. The ratio of active area to pixel footprint defines an optical fill factor by which the signal is attenuated. Optical fill factor can be recovered by aligning a microlens array to an APD array, such that light incident on the footprint of each microlens element is reimaged onto the active area of the corresponding APD pixel. However, practical implementation can be challenging for finer pixel pitches when the array is to be compatible with small f -number external camera

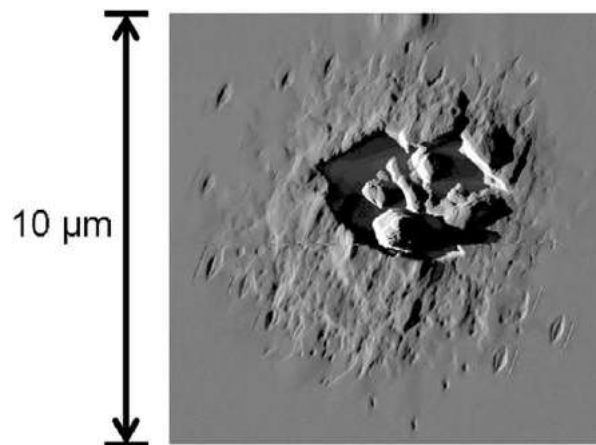


Fig. 6 Amplitude mode atomic force microscope (AFM) image of a large dust-related defect on an avalanche photodiode (APD) wafer surface.

optics of the sort commonly used in compact sensor systems. APD and microlens wafers thinner than about $300\text{ }\mu\text{m}$ – and, depending on size, die which are thinner than about half that – cannot be safely handled as independent pieces without high risk of breakage. Depending on pixel pitch, pixel size, and the f -number of external camera optics, a microlens focal length less than $100\text{ }\mu\text{m}$ may be required, necessitating use of special techniques such as substrate removal and attachment of temporary mechanical handling wafers during fabrication.

InGaAs LM APD Device Characteristics

Example current–voltage (I – V) and capacitance–voltage (C – V) characteristics for a $500\text{-}\mu\text{m}$ -diameter InGaAs LM APD are overlaid in Fig. 7, with reverse current plotted logarithmically against the left-hand axis and capacitance plotted linearly against the right-hand axis. The InGaAs APD in question is one of the larger to be produced commercially. APDs of smaller junction area have proportionally lower dark current and capacitance, although these parameters also depend somewhat on details of structure and manufacture. However, the characteristics in Fig. 7 exhibit features that are common to all InGaAs LM APDs.

Both the current and capacitance characteristics undergo an abrupt step at approximately 22 V reverse bias. This step occurs because the junction’s growing depletion region has finished penetrating the doped space charge layer which separates the APD’s multiplication and absorption layers. Prior to “punching-through” the charge layer, the junction’s depletion region does not extend into the narrow-bandgap InGaAs absorption layer, and a potential energy barrier prevents charge carriers generated in the absorber from reaching the junction and contributing to the diode’s reverse current. Reverse current jumps at punch-through because with the potential barrier pulled down, the junction abruptly becomes able to collect the charge carriers generated in the absorption layer. Capacitance drops abruptly at punch-through because the junction’s depletion region is able to widen into the substantially undoped absorption layer.

The vertical asymptote in the reverse current characteristics of Fig. 7 occurs at the APD’s avalanche breakdown voltage, the critical reverse bias at which the impact ionization process inside the APD junction becomes self-sustaining. Breakdown occurs because secondary charge carriers generated by impact ionization can themselves initiate further impact ionization in a branching chain reaction which need never terminate if the propensity to impact-ionize is high enough. A notional chain of impact ionization events is diagrammed in Fig. 8, where time progresses downward on the page and – because of their opposite electric charge – electrons are depicted drifting to the right and holes drift to the left. The starburst symbols from which both the initiating carrier and a new electron–hole pair emerge represent impact ionization events. The fact that impact ionization is a stochastic process is represented in Fig. 8 by the starburst outlines which do not result in additional secondary carriers – these are impact ionization events which might have happened but didn’t in the eventuality sketched. Because the multiplication layer is of finite width, carriers can drift out of the multiplier without triggering impact ionization. Below the APD’s breakdown voltage, every impact ionization chain eventually terminates in this way, so the avalanche gain is finite. Increasing the reverse bias applied to the APD’s terminals exponentially increases the impact ionization rate until, at the breakdown voltage, chains which never terminate become possible. When an APD is biased above its breakdown voltage it behaves like a low value resistor, and the current through the diode is limited by what its bias circuit can supply rather than by the APD junction characteristics. It is advisable to use

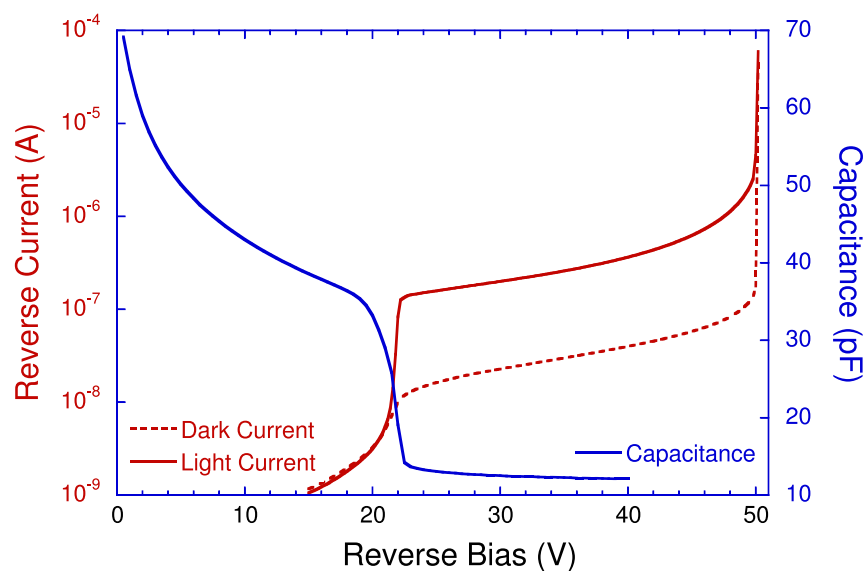


Fig. 7 Current–voltage (I – V) and capacitance–voltage (C – V) characteristics of a representative $500\text{-}\mu\text{m}$ InGaAs Linear-mode (LM) avalanche photodiode (APD).

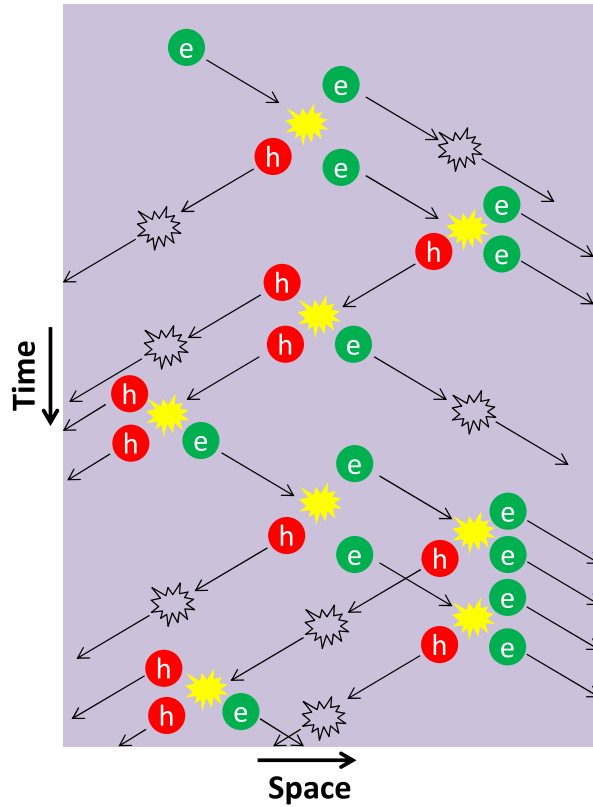


Fig. 8 Diagrammatic sketch of notional impact ionization chains inside an avalanche photodiode's (APD's) multiplier.

current-limited supplies when operating or testing APDs near breakdown, because most InGaAs LM APDs burn out under direct currents in the 1–10 mA range.

QE, avalanche gain (M), and spectral responsivity (R) are usually calculated from direct current (DC) I - V characteristics of the sort plotted in Fig. 7. Responsivity, the APD's ratio of output photocurrent to input optical power, expresses the linear relationship alluded to by the terminology LM APD, and contains factors of gain and QE:

$$R = 8.065544 \times 10^5 \times M \times QE \times \lambda \text{ (AW}^{-1}\text{)} \quad (1)$$

where λ is the optical wavelength in m.

Of the three parameters QE, gain, and responsivity, only responsivity can be measured directly and unambiguously for every InGaAs LM APD. DC responsivity measurements are made by using an optical power meter with a reference photodiode to calibrate the optical power delivered from a monochromatic, continuous-wave (CW) source such as a stabilized diode laser or a monochromator. The DC photocurrent characteristic is found by subtracting the dark I - V characteristic from the light I - V characteristic, so that at any given reverse bias the responsivity is the photocurrent divided by the optical power delivered to the APD's active area.

When operated at a fixed reverse bias, gain saturation causes an InGaAs LM APD's responsivity to drop if the optical signal intensity is increased beyond a certain level. Gain saturation occurs when the photocurrent density is high enough that the mobile charges carrying the current partially compensate the ionized dopant atoms of the APD junction. The impact ionization rate in the APD multiplier is an exponential function of the local electric field strength. Partial compensation of the ionized dopants weakens the electric field, reducing the ionization rate, and therefore the avalanche gain. Different APD designs exhibit gain saturation to different degrees, but in general, the onset of gain saturation occurs at lower optical signal levels when a given APD is operated at higher avalanche gain. This is because the space charge compensation mechanism depends on the charge density carried by the multiplied photocurrent, whereas the optical signal power only determines the primary photocurrent prior to multiplication. Fig. 9 illustrates this point with responsivity data from a 75- μm -diameter InGaAs LM APD. The APD was operated at different reverse biases under 1550 nm illumination at different optical power levels. The lowest operating point depicted, 67.5 V reverse bias, corresponds to an avalanche gain of about $M=10$. Very little gain saturation is observed for the 67.5 V responsivity curve from 1 nW through 10 μW of optical signal power, whereas the responsivity curve at 70.35 V reverse bias drops from nearly 80 A W^{-1} ($M \approx 80$) for a 1 nW signal to close to 10 A W^{-1} ($M \approx 10$) for a 100 μW signal.

InGaAs LM APDs are usually used to sense weak optical signals, but gain saturation has practical bearing on optical overload behavior and special situations such as optical heterodyne systems in which a weak signal is coherently mixed with a strong optical local oscillator. Gain saturation can also affect the accuracy of APD experiments meant to isolate some property of the

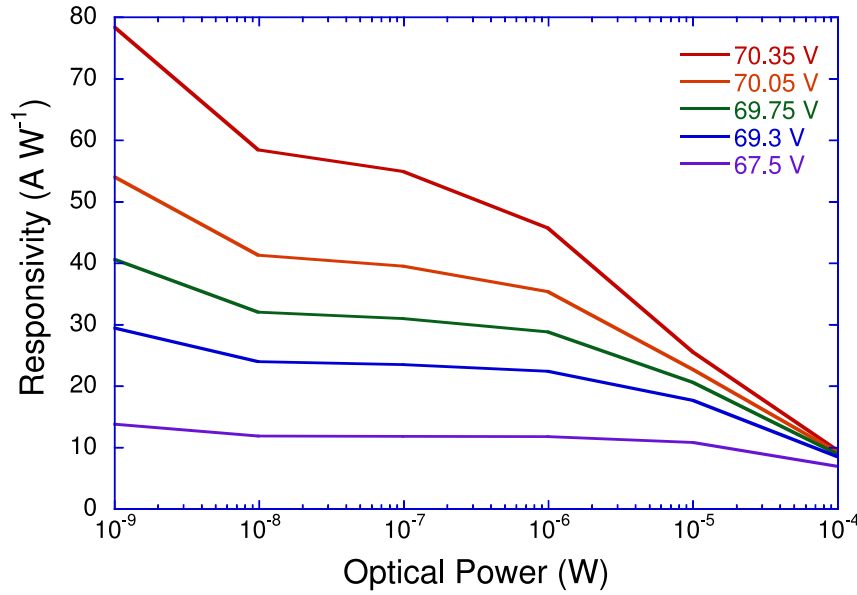


Fig. 9 Avalanche gain saturation as a function of optical power for a representative 75- μm InGaAs linear mode (LM) avalanche photodiode (APD).

photocurrent, in which a measurement taken in the dark is subtracted from the same measurement under strong optical illumination. Spurious measurements can result if the test signal is bright enough to saturate the APD's gain because then the APD's gain operating point is higher for the dark measurement than for the illuminated measurement, and not all of the light-versus-dark difference results from the photocurrent.

Although the DC responsivity of an APD is a commonly reported parameter, APDs are almost always used to detect rapidly modulated or pulsed optical signals. APD responsivity to modulated optical signals decreases with increasing modulation frequency because of the APD's fundamental impulse response, and because of low-pass filtering related to the APD's capacitance in combination with the characteristics of the amplifier which operates the APD. According to the Shockley–Ramo theorem, the instantaneous terminal current of an APD depends upon the instantaneous drift velocities of the electrons and holes in its junction, but can be approximated based on the instantaneous count of electrons and holes present in the junction (n_e and n_h), average electron and hole saturation drift velocities (v_{se} and v_{sh}), and the junction width (w):

$$i(t) \approx \frac{q}{w} [v_{se} n_e(t) + v_{sh} n_h(t)] \quad (\text{A}) \quad (2)$$

where q is the elementary charge.

The fundamental impulse response of an APD has a finite rise time because it takes a finite amount of time for the instantaneous population of secondary carriers in the junction to reach its maximum value. It takes time for primary photocarriers generated in the absorber to reach the multiplier, and once there, carriers which lack sufficient kinetic energy to impact ionize must move through a minimum displacement in the local electric field – the carrier's dead space – in order to pick up the ionization threshold energy. Once energetically active, carriers do not instantly impact ionize. Instead, active carriers travel a variable distance through the multiplier before ionizing – if they ever ionize – based on alloy composition and field-dependent impact ionization rates. Following impact ionization, both the original primary carrier and the newly generated secondary carriers must again pick up the ionization threshold energy before they can trigger new ionizations.

The fall time of an APD's fundamental impulse response is determined by how long impact ionization continues in its multiplier after the peak instantaneous carrier population is reached, and how far secondary carriers must travel from their point of generation in the multiplier to exit the junction. Electrons exit the junction at the cathode and holes exit at the anode, so depending upon the APD's epitaxial stack order (Fig. 3), one carrier type has a short distance to travel to exit the junction and the other carrier type must cross the absorption layer. The carrier type with further to travel carries most of the current. When an APD is operated at low gain, generation of secondary carriers may end before any of the secondary carriers of the type with further to travel have exited the junction, in which case the peak of the impulse response coincides with the end of impact ionization and the fall time is determined by the absorber transit time. However, when an APD is operated at higher gain, impact ionization may continue after the peak of the impulse response, extending the fall time of the APD.

Monte Carlo simulations of the impulse response of an APD operating at $M=12.3$ and $M=44.0$ are compared in Fig. 10. The APD that was simulated has a 1.5- μm -thick InGaAs absorber and a 0.3- μm -thick InAlAs multiplier; including various spacer layers and the charge layer, the total junction width of the simulated APD was 2.25 μm . Saturation drift velocities typical of electrons and holes in these alloys – respectively 10^5 m s^{-1} for electrons and $5 \times 10^4 \text{ m s}^{-1}$ for holes – were assumed. The spatial distribution of

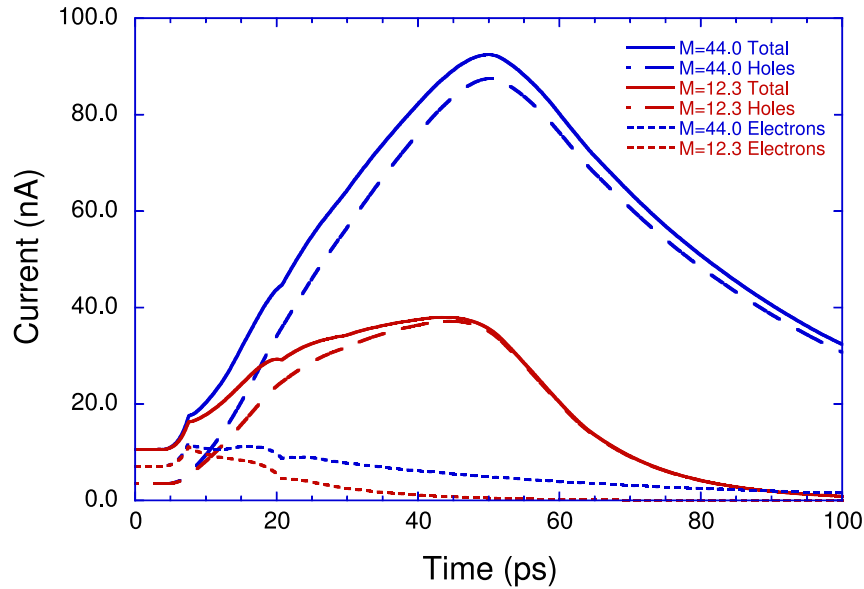


Fig. 10 Monte Carlo simulation of an InGaAs avalanche photodiode's (APD's) photocurrent impulse response.

primary photocarrier generation within the absorber, the dead space effect, and the stochastic nature of impact ionization are treated in this model, but with the approximation of a hard dead space and local field-dependent rather than carrier energy-dependent impact ionization rates, meaning that a carrier's propensity to impact-ionize is treated as abruptly assuming a value that depends on the local field strength once it picks up the ionization threshold energy, as opposed to gradually increasing as it accumulates kinetic energy in excess of the threshold. The approximation that all carriers travel at the average saturation drift velocity at all times is also made, rather than tracking the velocity of individual carriers as they are accelerated by the electric field and lose energy through scattering. These approximations – especially the values chosen for the saturation drift velocities – affect the scaling of the simulated impulse response, but **Fig. 10** illustrates the qualitative features that the carrier type which takes longest to leave the junction dominates the total current, and the fall time is extended at higher avalanche gain. The increase of impulse response duration with gain places a fundamental limit on APD gain-bandwidth product.

The impulse response of an InGaAs LM APD can be measured by exciting it with a laser pulse that is very short compared to the APD's junction transit time, and observing the resulting photocurrent pulse with an oscilloscope, but the input impedance of the scope in combination with the APD's junction capacitance low-pass filters the voltage waveform measured by the oscilloscope. InGaAs LM APD capacitance varies with junction area from less than 0.1 pF for small APD pixels to greater than 10 pF for the largest detector elements. Whereas smaller InGaAs LM APDs typically have sub-nanosecond impulse response, **Fig. 11** illustrates how the 12.3 pF capacitance of a 500- μm -diameter APD leads to much longer impulse response durations.

Similar low-pass filtering effects occur when an APD is connected to an amplifier circuit, and any given amplifier design provides gain over a limited frequency band, so the frequency response of an APD photoreceiver depends as much on the characteristics of the amplifier as on those of the APD. The highest gain-bandwidth products reported for experimental waveguide (Kinsey *et al.*, 2001) and resonant cavity (Lenox *et al.*, 1999) InGaAs LM APDs as individual components is about 300 GHz, but most commercially available InGaAs LM APD photoreceivers have 3 dB bandwidths in the range from tens of megahertz to a few gigahertz. Example plots of photoreceiver responsivity in unit of V W^{-1} (i.e., including a factor of the amplifier's transimpedance gain) versus modulation frequency are shown in **Fig. 12** for a photoreceiver assembled from a 75- μm -diameter InGaAs LM APD and a transimpedance amplifier (TIA) chip designed for 2.125 Gbps optical communications.

The photocurrent gain of an APD is the ratio between the photocurrent measured at its terminals to the primary photocurrent generated in its absorber. Avalanche gain is a critical input to APD sensitivity calculations because it affects both signal level and the intensity of excess multiplication noise. APD dark current is also commonly benchmarked to a particular reference gain, and empirical measurement of the APD's excess noise factor requires knowing the gain at which noise power measurements are collected. It is therefore unfortunate that the gain of an InGaAs LM APD often cannot be measured unambiguously.

Although terminal photocurrent is directly measured by subtracting dark from light I - V characteristics, measurement of the primary photocurrent can be ambiguous because photocarriers generated in the APD's absorber cannot contribute to the terminal current until the APD is biased above its punch-through voltage. Since the electric field in the APD's multiplier strengthens with increasing reverse bias regardless of whether or not punch-through has yet occurred, it is often the case that at the APD's punch-through voltage, when it first becomes possible to observe photocurrent, impact ionization is already generating avalanche gain in the multiplier.

Bias-dependent collection efficiency of primary photocarriers by the junction is a second factor which can make gain measurements ambiguous. Although the APD abruptly becomes sensitive to light at its punch-through voltage, the collection efficiency of primary photocarriers increases with increasing reverse bias until the InGaAs absorption layer is fully depleted. Since avalanche

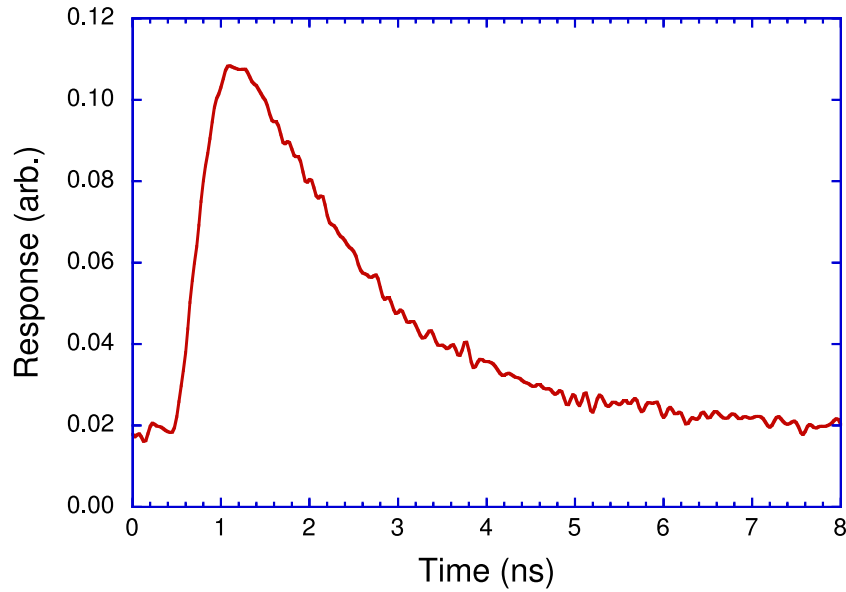


Fig. 11 Measured impulse response of a 500- μm InGaAs linear mode (LM) avalanche photodiode (APD).

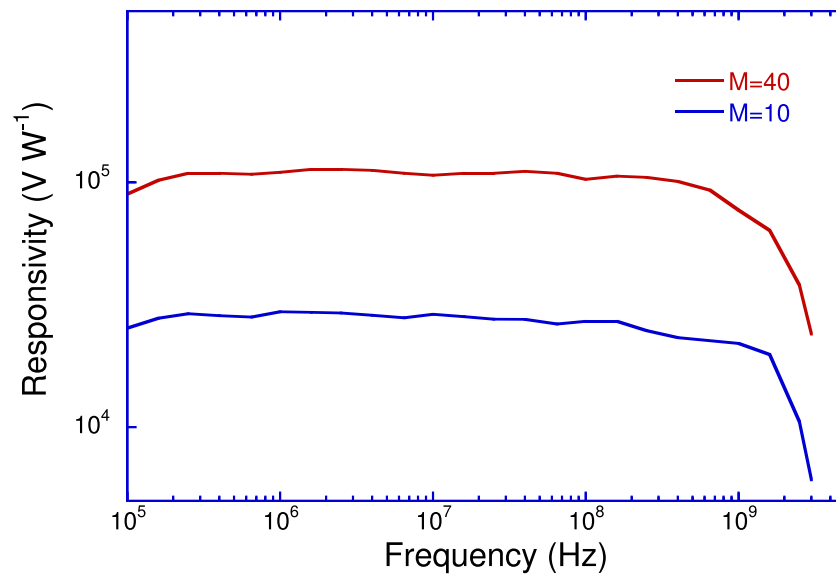


Fig. 12 Measured frequency response of an InGaAs linear mode (LM) avalanche photodiode (APD) photoreceiver assembled from a 75- μm APD and a transimpedance amplifier (TIA) designed for 2.125 Gbps telecommunications.

gain also increases with increasing reverse bias, both the supply of primary photocarriers reaching the multiplier and the gain factor by which they are multiplied increase as the reverse bias is raised past the APD's punch-through voltage. In order to avoid confusing increasing collection efficiency for avalanche gain, gain measurements based on the terminal photocurrent must be made relative to a reference point for which collection efficiency is already at its maximum value. Maximum collection efficiency is usually reached within a few volts of punch-through and is recognizable as a local minimum of the slope of the photocurrent characteristic above the punch-through voltage. The point of minimum photocurrent slope above punch-through can be attributed to the collection efficiency nearing its maximum because any contribution to the slope from avalanche gain increases monotonically with reverse bias. The photocurrent characteristic of an APD for which the primary photocurrent is directly observable has a slope above punch-through that approaches zero where collection efficiency is maximized and increases gradually thereafter as avalanche gain turns on. The point of minimum photocurrent slope above punch-through is also used as a reference point for relative gain measurements in APDs where avalanche gain is already active in the multiplier at punch-through, but in this case, the minimum slope is nonzero and an absolute measurement of the primary photocurrent cannot be made. [Fig. 13](#) replots on linear

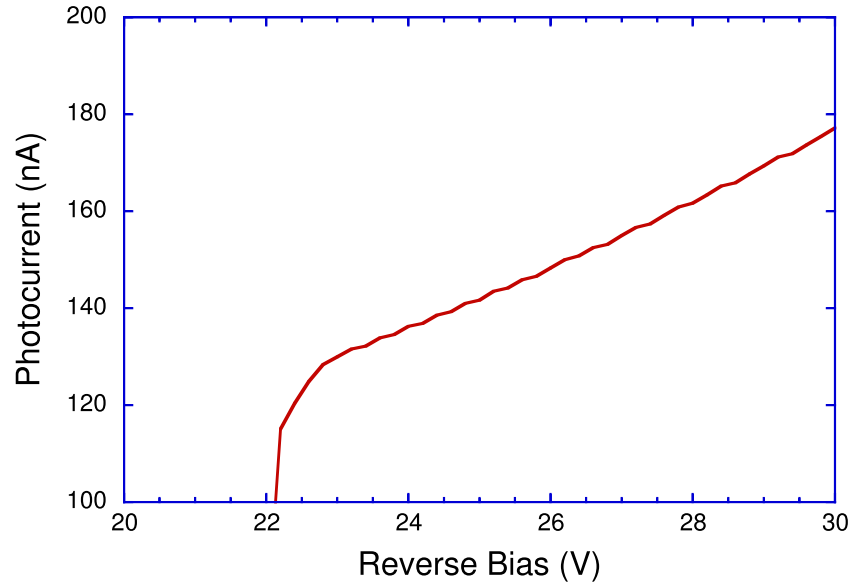


Fig. 13 Photocurrent characteristic of a representative 500-μm InGaAs linear mode (LM) avalanche photodiode (APD).

axes the photocurrent characteristic derived from Fig. 7. The minimum slope reference point occurs near 23 V reverse bias but is nonzero, from which one can infer that the multiplier is already generating some avalanche gain at punch-through.

If an InGaAs LM APD's primary photocurrent cannot be directly measured, that prevents exact measurement of its avalanche gain. In that circumstance, measurement of QE is similarly ambiguous, because the gain must be known in order to find the QE from a responsivity measurement. In this common circumstance, physical arguments or measurements on physically analogous photodiodes which lack a charge layer can be used to bound the plausible QE and gain, based on responsivity measurements. For instance, the bottom-illuminated 500-μm-diameter APD whose I - V characteristics were plotted in Figs. 7 and 13 has a 1.5-μm-thick InGaAs absorption layer. Ellipsometric measurements on pure InGaAs films give an absorption coefficient of 6700 cm^{-1} at 1500 nm. The minimum plausible QE at 1500 nm, when this APD is operated at 23 V reverse bias, can be estimated by assuming an antireflection coating that transmits 99% of the incident optical power, transmission losses of about 15% through the doped substrate, and absorption of about 63.4% of the optical signal in a single pass through the InGaAs layer, calculated from the known absorption coefficient under the assumption of near perfect collection efficiency. This results in a lower bound estimate of $\text{QE} = 53.3\%$. On the other hand, in this design the optical signal is intended to back-reflect from the underside of the APD's anode contact, passing through the absorber a second time. In the best case scenario the substrate absorption is only 5% and the back-reflection is lossless, giving a high estimate for the plausible QE of about $\text{QE} = 81.4\%$. Eq. (1) gives unity-gain responsivities at 1500 nm of $R = 0.64 \text{ A W}^{-1}$ in the worst case QE scenario and $R = 0.98 \text{ A W}^{-1}$ in the best case. In comparison, an empirical spectral responsivity measurement at 1500 nm with the APD biased at 23 V finds a responsivity of $R = 0.99 \text{ A W}^{-1}$, implying that at 23 V the APD is actually operating with a gain between $M = 1.0$ and $M = 1.5$. The continuous increase of photocurrent with reverse bias observed in Fig. 13 shows that by 23 V some avalanche gain is already occurring, which is consistent with the range estimated for the gain. Example responsivity and QE spectra for this APD that reflect the gain calibration ambiguity are plotted in Fig. 14.

Multiplied shot noise is the dominant noise source in an InGaAs LM APD. Because InGaAs LM APDs are exclusively used to sense fast optical signals, any significant $1/f$ noise generated by an InGaAs LM APD is usually at frequencies below the low-frequency cutoff of its amplifier. Some references include a Johnson–Nyquist noise term in their analysis of LM APD noise but this is unnecessary because diodes only have Johnson–Nyquist noise near zero bias, and because the series contact resistance of an APD is typically too low to contribute measureable noise. Sometimes noise is attributed to a hypothetical load resistor connected in parallel to the APD, but this circuit configuration is practically never employed. Instead, photoreceiver sensitivity analysis should consider the multiplied shot noise of the APD and an input-referred amplifier circuit noise term that is particular to the amplifier design.

An APD's multiplication noise results from the random variation of its avalanche gain. Analysis of APD noise is based on the counting distribution derived by Robert J. McIntyre (McIntyre, 1972). The probability that an APD multiplier operating at some average gain (M) will output a certain number of electrons (n) given an input of a certain number of primary electrons (a) is

$$P_{\text{McIntyre}}(n) = \frac{a \Gamma\left[\frac{n}{1-k} + 1\right]}{n(n-a)! \times \Gamma\left[\frac{nk}{1-k} + 1 + a\right]} \times \left[\frac{1+k(M-1)}{M}\right]^{a+\frac{n-k}{1-k}} \times \left[\frac{(1-k)(M-1)}{M}\right]^{n-a} \quad (3)$$

where k is the ratio of impact ionization rates for the two carrier types (the slower rate over the faster rate) and Γ is the Euler gamma function.

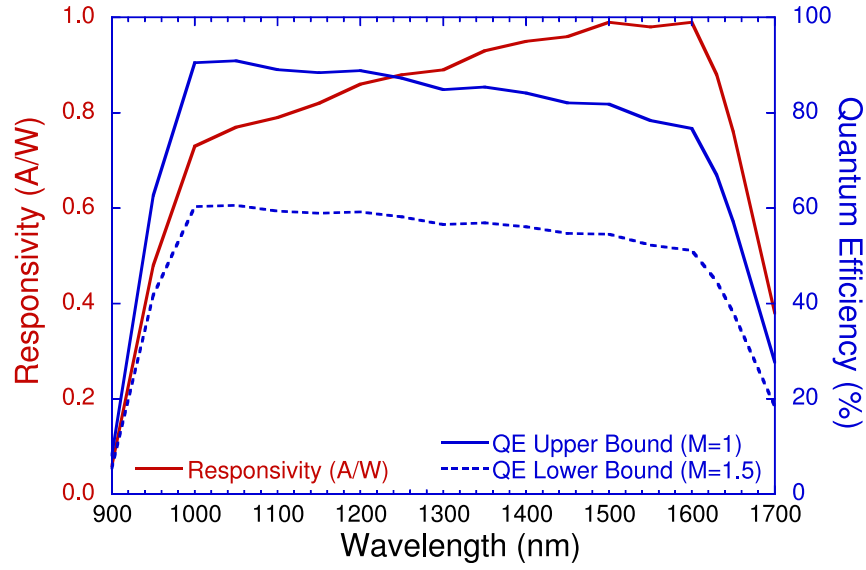


Fig. 14 Measured responsivity and upper and lower bounds on quantum efficiency (QE) for a representative 500- μm InGaAs linear mode (LM) avalanche photodiode (APD).

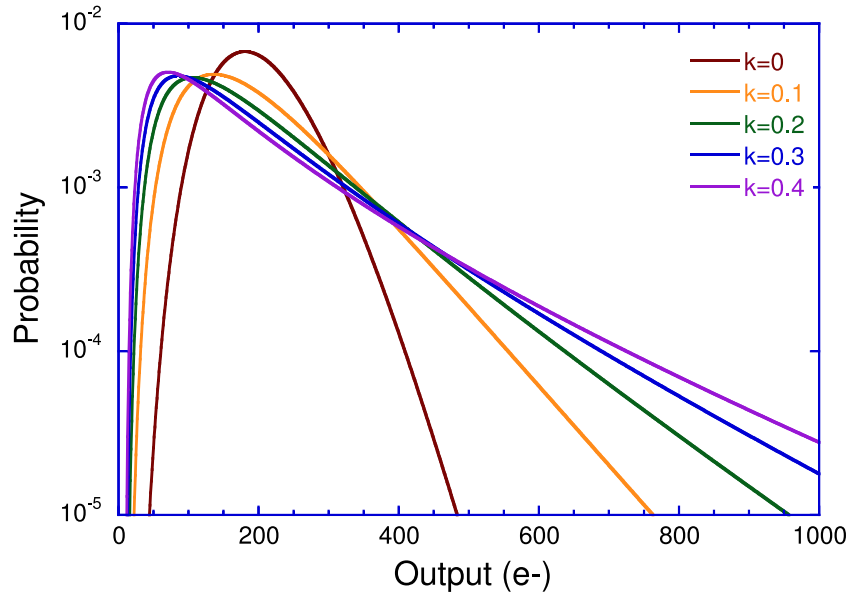


Fig. 15 Output count distributions for linear mode (LM) avalanche photodiodes (APDs) operating at $M=20$ and responding to 10 primary photoelectrons, but differing in terms of k .

McIntyre distributions for APDs operating at the same average gain ($M=20$) and illuminated by the same signal strength ($a=10$ primary photoelectrons) but differing in k are plotted in Fig. 15. These distributions illustrate the practical meaning of different values of k . For the same input signal strength and the same average gain, an APD with lower k will have a higher probability of detecting a signal and a lower probability of generating a false alarm.

These statements assume that the APD is employed in a photoreceiver equipped with a binary decision circuit that rejects inputs below a certain detection threshold, and that the mean signal photocurrent is larger than the mean dark current. In this common scenario, a single detection threshold is simultaneously in the high-output tail of the dark current distribution but comfortably lower than the bulk of the photocurrent distribution's probability density (Fig. 16), such that the longer tail of the high- k distribution increases the probability of false alarm but reduces signal detection probability by decreasing the distribution's median output value. The detection threshold is employed to reject false alarms arising from total photoreceiver noise, of which the APD's dark current is one component. At the same time, the detection threshold must not be set so high that it also rejects outputs arising from valid photocurrent signals. An output distribution with a higher median for a given input is desirable because

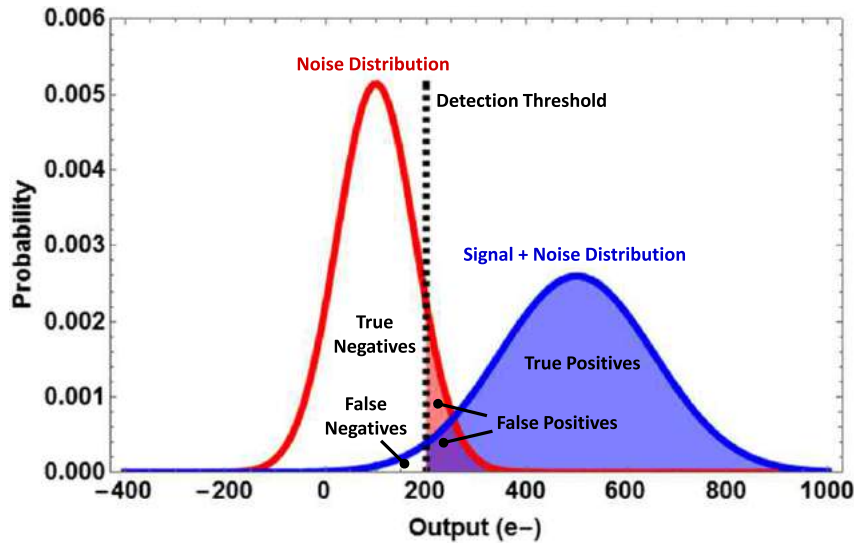


Fig. 16 Illustration of signal and noise distributions for an avalanche photodiode (APD) photoreceiver equipped with a thresholded decision circuit.

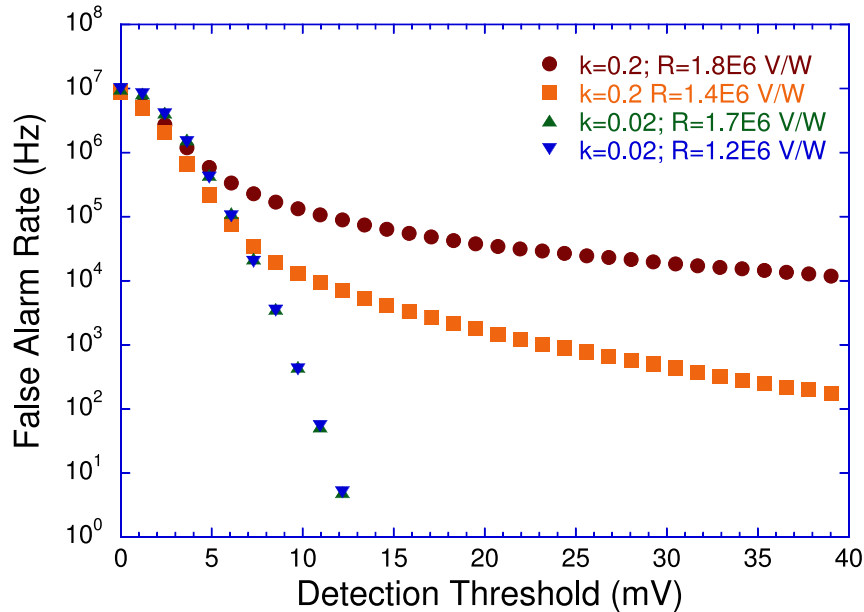


Fig. 17 Comparison of false alarm rate vs. detection threshold profiles for avalanche photodiode (APD) photoreceivers with different values of k .

the high median will allow one to set the detection threshold higher without sacrificing signal detection efficiency. On the other hand, a reduced likelihood of very high-output events will help minimize the false alarm rate (FAR) arising from “lucky” dark current electrons that happen to individually experience very high avalanche gain.

Photoreceiver performance actually depends upon the convolution of the APD’s output distribution with the amplifier’s Gaussian circuit noise and not just the APD’s output distribution, but the general observations about how k relates to signal detection and false alarm performance illustrated for the APD output distributions of Fig. 15 still hold in a more rigorous analysis. Fig. 17 is an empirical demonstration of the advantage a low- k APD conveys from the standpoint of FAR performance. FAR measurements are plotted in Fig. 17 for two otherwise identical APD photoreceivers, both operated with similar low-frequency responsivities to 1550 nm light (marked in V W^{-1} because of the TIA’s transimpedance), which employ 75- μm -diameter InGaAs LM APDs respectively characterized by $k=0.2$ (red square; orange rectangle) and $k=0.02$ (blue and green triangles). The FAR versus detection threshold behavior of the two receivers is similar at low threshold but diverges sharply at higher threshold, giving the photoreceiver with a $k=0.02$ APD a significant sensitivity advantage in applications that are intolerant of high FARs.

The exact McIntyre distribution often must be used to calculate quantities like FAR or bit error rate (BER) because those photoreceiver characteristics depend on the shape of the tail of the APD's output distribution, and the McIntyre distribution has positive skew. In most cases, the standard FAR and BER formulas which apply to photodiode-based receivers do not provide accurate results when applied to APD-based receivers – even when the standard deviation is adjusted to account for the APD's excess multiplication noise – because those formulas assume the probability density of false alarms falls off as a function of output amplitude with a Gaussian profile. In general, the only circumstances in which the Gaussian approximation provides accurate results are when APDs operate at low gain or with very low k . The FAR of InGaAs LM APD photoreceivers is better modeled by

$$\text{FAR} = \sqrt{\frac{2\pi}{3}} N_Q \text{BW} P_{\text{RX}}(n_{\text{th}}) \text{ (Hz)} \quad (4)$$

where N_Q is the noise of the photoreceiver (APD and amplifier) referenced to the output of the APD in units of electrons, BW is the bandwidth in Hz of the signal chain into the threshold comparator, and $P_{\text{RX}}(n_{\text{th}})$ is the McIntyre distribution of the APD's multiplied dark current convolved with the amplifier's input noise, evaluated at an electron count corresponding to the detection threshold.

The most common measures of photoreceiver sensitivity such as SNR and NEP define noise based on the standard deviation of an APD's output distribution, and so are not affected by the divergence of the tail of the McIntyre distribution from the Gaussian approximation, which normally becomes an issue a few standard deviations beyond the mean. For this type of analysis, it is sufficient to compute the variance of the APD's output. In units of electrons the APD output variance is found using the Burgess variance theorem:

$$\begin{aligned} \text{var}(n) &= M^2 \text{var}(a) + \langle a \rangle \text{var}(m) \\ &= M^2 F \langle a \rangle \text{ (e}^{-2}\text{)} \end{aligned} \quad (5)$$

where the excess noise factor (F) is defined as:

$$F \equiv \frac{\langle m^2 \rangle}{M^2} \quad (6)$$

In Eq. (5) the symbol m is a per-electron gain random variable, and the symbols a , M , and n have the same meaning as in Eq. (3). The noise factor of Eq. (6) is deemed “excess” because if APD gain was deterministic, Eq. (5) would simply read $\text{var}(n) = M^2 \langle a \rangle$. Eq. (5) assumes that the primary photocarrier count represented by a is generated by a Poisson process, so the substitution $\text{var}(a) = \langle a \rangle$ can be made.

For most linear-mode APDs, including InGaAs APDs, the excess noise factor has the gain-dependence derived by McIntyre for thick, uniform junctions (McIntyre, 1966):

$$F = M \left[1 - (1 - k) \left(\frac{M - 1}{M} \right)^2 \right] \quad (7)$$

In Eq. (7), the parameter k is the same ratio of impact ionization rates appearing in Eq. (3). When $k > 0$, k is the slope of the excess noise curve as a function of gain, in the limit of high gain. For single-carrier multiplication, $k = 0$, and $F \rightarrow 2$ in the limit of high gain. Another feature of single-carrier $k = 0$ multiplication is that avalanche breakdown cannot occur because the only carriers capable of impact ionizing all drift in the same direction across the junction. The gain curve of a $k = 0$ APD does not exhibit the vertical asymptote shown in Fig. 7, enabling stable operation at higher gain than a $k > 0$ APD. Different InGaAs LM APDs vary with respect to the value of k that characterizes their multiplication noise. Holes ionize more readily than electrons in InP, and InGaAs APDs with comparatively thick ($\sim \mu\text{m}$) InP multipliers typically have $k = 0.4$. The impact ionization rate of electrons is higher than that of holes in InAlAs; APDs with thick ($\sim \mu\text{m}$) InAlAs multipliers are characterized by $k = 0.3$ whereas those with InAlAs multipliers thinner than about $0.3 \mu\text{m}$ have $k < 0.2$ (Lenox *et al.*, 1998). InGaAs LM APDs with multipliers engineered to suppress hole multiplication have been reported in which $k = 0.04$ (Williams *et al.*, 2013).

In many applications it is common to work in units of current rather than electron count. The noise spectral intensity theorem for APDs that is equivalent to Eq. (5) is

$$S_I = 2 q M^2 F I_{\text{primary}} \text{ (A}^2\text{Hz}^{-1}\text{)} \quad (8)$$

where I_{primary} is the total unmultiplied current (the sum of signal and background photocurrent and dark current) prior to injection into the APD's multiplier. In principle, there can be dark current leakage paths in an APD which bypass the multiplying junction which would not be multiplied and therefore which would not be part of I_{primary} in Eq. (8). Also, in some exotic InGaAs LM APDs a substantial portion of the total dark current is generated inside the multiplier by tunneling rather than originating in the absorber, and so – because of a different path length through the multiplier – experiences a different amount of avalanche gain and contributes a different amount of multiplied shot noise than the photocurrent. However, in practice, dark current which originates in the absorber and which experiences the APD's full multiplication dominates in most commercially available InGaAs LM APDs, and the noise on that dark current is well modeled by Eq. (8).

As explained in connection to gain and QE measurements, unambiguous measurements of primary photocurrent cannot be made on an InGaAs LM APD if avalanche gain in its multiplier turns on at a lower reverse bias than its punch-through voltage. For APDs where I_{primary} is not observable, the best practice when making APD photoreceiver sensitivity calculations with Eq. (8) is to

use the terminal current in place of the quantity $M \times I_{\text{primary}}$, and to use the best estimate of the gain for the second factor of M . Uncertainty about I_{primary} also affects the accuracy of experimental measurements of the excess noise factor, because the most common method of measuring F involves measuring noise power spectral intensity, S_p , with a noise figure meter, converting S_p to S_i using an impedance that is characteristic of the test system but which must be measured, and then applying Eq. (8) to find F from S_i (which requires an estimate of M). For this reason, the most accurate measurements of F are made on special test structures which lack charge layers, and which do not include InGaAs absorbers. Instead, photocarriers are generated by illuminating the test structure with a short-wavelength laser, and both the primary current and the avalanche gain can be unambiguously measured because there is no charge layer to prevent collection of photocarriers at low reverse bias.

Eqs. (5) and (8), respectively, calculate multiplied shot noise variance expressed in units of electrons or in units of current; the square root gives the standard deviation, which is the most common measure of noise. The total noise of an APD photoreceiver assembled from a charge-integrating capacitive-feedback transimpedance amplifier (CTIA) is

$$\begin{aligned} N_Q &= \sqrt{N_{\text{CTIA}}^2 + (\langle a_{\text{dark}} \rangle + \langle a_{\text{bg}} \rangle + \langle a_{\text{signal}} \rangle) M^2 F} \\ &= \sqrt{N_{\text{CTIA}}^2 + \left[\frac{\tau}{q} (I_{\text{dark}} + I_{\text{bg}}) + Q_{\text{signal}} \right] M F} \quad (\text{e}^-) \end{aligned} \quad (9)$$

where N_{CTIA} is the input-referred charge noise of the CTIA in units of electrons, τ is the effective DC current integration time of the CTIA, I_{dark} and I_{bg} are, respectively, the dark current and background photocurrent as measured at the APD's terminals, and Q_{signal} is the product of the APD's QE and gain factor (M) with the optical signal level in units of photons.

The total noise current spectral density of an APD photoreceiver assembled from a resistive-feedback transimpedance amplifier (RTIA) is

$$\sqrt{S_{I_{\text{total}}}} = \sqrt{S_{I_{\text{TIA}}} + 2 q M F (I_{\text{dark}} + I_{\text{bg}} + I_{\text{signal}})} \quad (\text{AHz}^{-1/2}) \quad (10)$$

where $S_{I_{\text{TIA}}}$ is the input-referred noise current spectral intensity of the RTIA and I_{signal} is the product of the APD's spectral responsivity and the optical signal power. If the optical signal intensity is rapidly modulated rather than being quasi-CW, the RMS optical signal power is used.

The total noise calculated in Eq. (9) can be referred to the photoreceiver input, expressed in units of noise-equivalent photons, by dividing N_Q by the product of QE and M :

$$\text{NEPh} = \frac{Q_N}{\text{QE} \times M} \quad (\text{photons}) \quad (11)$$

Similarly, the total noise calculated in Eq. (10) can be referred to the photoreceiver input, expressed in units of noise-equivalent power, by dividing $S_{I_{\text{total}}}^{1/2}$ by the responsivity, R :

$$\text{NEP} = \frac{\sqrt{S_{I_{\text{total}}}}}{R} \quad (\text{W Hz}^{-1/2}) \quad (12)$$

The spectral density forms of current noise and NEP can be converted to in-band figures by multiplying either by $\text{BW}^{1/2}$.

Optical SNR, such as plotted in Fig. 1, is calculated as the ratio of the optical signal in photons to NEPh or in units of power to the in-band NEP.

References

- Kinsey, G.S., Campbell, J.C., Dentai, A.G., 2001. Waveguide avalanche photodiode operating at 1.55 μm with a gain-bandwidth product of 320 GHz. *IEEE Photonics Technology Letters* 13 (8), 842–844.
- Lenox, C., Nie, H., Yuan, P., *et al.*, 1999. Resonant-cavity InGaAs/InAlAs avalanche photodiodes with gain-bandwidth product of 290 GHz. *IEEE Photonics Technology Letters* 11 (9), 1162–1164.
- Lenox, C., Yuan, P., Nie, H., *et al.*, 1998. Thin multiplication region InAlAs homojunction avalanche photodiodes. *Applied Physics Letters* 73 (6), 783–784.
- McIntyre, R.J., 1966. Multiplication noise in uniform avalanche photodiodes. *IEEE Transactions on Electron Devices* ED-13, 164–168.
- McIntyre, R.J., 1972. The distribution of gains in uniformly multiplying avalanche photodiodes: Theory. *IEEE Transactions on Electron Devices* ED-19 (6), 703–713.
- Williams, G.M., Compton, M., Ramirez, D.A., Hayat, M.M., Huntington, A.S., 2013. Multi-gain-stage InGaAs avalanche photodiode with enhanced gain and reduced excess noise. *IEEE Journal of the Electron Devices Society* 1 (2), 54–65.

Very High Range Resolution Lidars

Zeb W Barber, Montana State University – Spectrum Lab, Bozeman, MT, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Light detection and ranging (lidar) or laser radar (ladar) technologies have already made high impacts in wide range of industrial, construction, scientific, and defense fields, and future applications of lidar will multiply as performance and cost improve. Consumer lidars already exist in the form of laser rangefinders for golfers and hunters, but the push toward autonomous vehicles and robots will increase the proliferation of lidar technology to unprecedented levels (Ackerman, 2016; Khoshelham and Zlatanova, 2016; see Section Relevant Websites). An important performance metric for all lidars is the resolution of the range measurement, which determines the suitability for different applications. Range resolution at the level of a meter to a few centimeters are the most common and are good for general range finding, mapping, and outdoor navigation applications. With very high range resolution lidar technology (as small as a few tens of micrometers) based on coherent frequency modulated continuous wave (FMCW) (De Groot and McGarvey, 1992) or concepts such as dual-comb coherent optical sampling (Coddington *et al.*, 2009); very different classes of applications are enabled. These include: noncontact metrology, high resolution 2D and 3D imaging, and coherent synthetic aperture imaging. This article focuses on very high resolution lidar technologies that enable range measurements with resolution of less than 1 cm. This is a convenient threshold as generally the lidar and ladar technologies required to achieve this level of range resolution are significantly different than more common brief-pulse, direct-detect lidars due to the difficulty of achieving sub-70 ps (~ 1 cm range difference) timing resolution.

Range Resolution Versus Range Precision

Before describing very high range resolution lidar technology, it is important to clarify the definition of range resolution. The definition of range resolution used in this article is consistent with that used in the radar community (see Section Relevant Websites). Namely, resolution is defined as the minimum range separation that two point targets of equal power can have and still be resolved (see Fig. 1). As all lidars and ladars are based upon measuring the round-trip time-of-flight of light between the ladar system and the target, range resolution is then fundamentally limited by the temporal bandwidth of the optoelectronics as $\Delta R \geq c/2B$, where B is the bandwidth and c is the speed of light. This definition is directly analogous to the Rayleigh resolution criterion used in optical imaging, where the image resolution is fundamentally limited by the spatial bandwidth of the imaging system. For example, in brief-pulse, direct-detection lidar, the fundamental timing (range) resolution is just set by the temporal duration, $\Delta\tau$, of the optical pulse (assuming the detector and electronics are sufficiently fast). For a Fourier limited Gaussian pulse the temporal width is just the inverse of the bandwidth as, $\Delta\tau = 1/B$. Given this temporal pulse width, if two targets are separated by less than $\Delta\tau$ in delay then echo pulses from each target merge into a single slightly wider return pulse and the targets are unresolved.

Unfortunately, this definition of range resolution is not used consistently throughout the literature or in manufacturer specifications. Often metrics better described as range precision (σ_R in Fig. 1) or the range grid/bin size (dR in Fig. 1) are specified as resolution. Range precision measures the statistical spread of multiple ideally identical range measurements of a single target. Using signal processing and estimation on a lidar signal with sufficient signal-to-noise ratio (SNR) the range precision achieved can be significantly better than the range resolution supported by the bandwidth. Part of the confusion of range resolution and range precision is that many lidars for 3D imaging are designed to only return a single range for each measurement, which blurs the

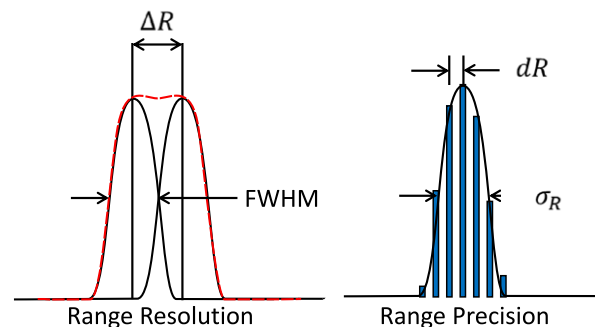


Fig. 1 Graphical definitions of range resolution – minimum separation to resolve two targets, and range precision – deviation of range estimates from a constant target.

distinction as the lidar can measure range displacements or differences in range to targets with precision much smaller than the resolution with sufficient SNR. However, the distinction is important as range resolution is an intrinsic property of the lidar system independent of the return power, while range precision is an extrinsic property dependent on the return power and resulting SNR. Therefore, when comparing different lidar technologies, range resolution provides a more fundamental comparison.

An advantage of resolving power in lidar is the ability to reduce the effect of range pulling from weaker nearby scatterers. An extreme example of range pulling can be observed with a common type of time-of-flight based ranging based on phase measurements of periodic intensity modulated optical signals (Fujima *et al.*, 1998). In these systems, the range to a target can be inferred by the period of the modulation and the amount of phase lag on the received light induced by the time-of-flight to the target and back. This type of time-of-flight distance measurement method has been used in surveying and metrology equipment for a long time and is now being used for 3D imaging (Iddan and Yahav, 2001) using special photonic mixing device, complementary metal-oxide-semiconductor (CMOS) cameras also known as “time-of-flight” cameras that are now provided by many vendors (see Section Relevant Websites). With a modulation frequency of 50 MHz (in the range common for these type of systems) and 12 bit resolution of the phase, the range could potentially be measured with a very good precision of $\sigma_R = c / (2f_{\text{mod}} \cdot 2^N) = 0.73$ mm. However, the drawback with this method is that there is essentially no range resolution whatsoever. The return phase of the received light is just the radio frequency (RF) coherent sum of all the scatterers in the scene. This means that if there are multiple returns in the range dimension (e.g., imaging through a window, screen, or foliage; or in a scene with a strong specular reflections which generates multiple path returns) the range will be a weighted average and very sensitive to stray scattering. Additionally, the range is only unambiguous out to one half of a wavelength of the modulation frequency (~ 3 m for 50 MHz), which limits the performance of the systems. In surveying and metrology applications these issues are resolved by using retro-reflecting targets to ensure the target of interest dominates the return signal reducing range pulling, and changing the modulation frequency slightly to implement a Vernier effect to resolve ambiguities.

Finally, range resolution is defined analogous to the Rayleigh criterion because many applications of interest require imaging, i.e., resolving multiple returns in the range dimension. For these applications, the resolving power of the lidar system is most important. A good example is the use of lidar in foliage and camouflage penetrating 3D imaging. In order to range gate out the cover, the lidar system must be able to resolve the return from the cover from the object beneath, which means the minimum distance between the cover and the object is proportional to the range resolution. Another example is the use of very high resolution lidar for optical metrology, where reflections from multiple surfaces need to be resolved to measure thicknesses and positions. However, the best example is the use of coherent lidar in synthetic aperture imaging lidar (SAIL or SAL) (Beck *et al.*, 2005). Here the Rayleigh analogy is complete as the image resolution in one dimension is directly set by the range resolution.

Very High Range Resolution Lidar Techniques

Brief-Pulse, Direct-Detect Lidar

For a lidar system to meet a threshold of 1 cm range resolution (timing resolution less than 66.7 ps), the corresponding bandwidth must be in excess of 15 GHz. To achieve this level of resolution with the straightforward brief-pulse, direct-detect lidar approach is very challenging, in particular if one is to simultaneously achieve high sensitivity (i.e., equivalent to longer ranges). Several type of pulsed laser sources with sub-100 ps pulse width exist to support the range resolution including microchip passively Q-switched lasers (Spühler *et al.*, 1999) and modulated master oscillator/fiber power amplifier (MOPA) systems. However, high energy lasers needed for long range lidar applications are less available due to the high peak powers this entails. On the detection side, bandwidths exceeding 50 GHz is achievable with modern optoelectronics developed for the telecommunication industry, however, these components meet severe trade-offs between sensitivity and bandwidth. Typically, these very fast linear PIN detectors are fiber coupled and use a 50 Ω load, which limits the NEP to approximately 10 pW/rHz when dark current and a 3 dB noise figure (NF) low noise amplifier are included. For a 15 GHz bandwidth pulse this translates to a unity SNR sensitivity of about 1 μ W, which is the equivalent of over 500 return photons per pulse. Linear avalanche photodiode (APD) based receivers can achieve a factor of 10 better, but rarely achieve bandwidth greater a few GHz. Operating APDs in Geiger mode can achieve single photon sensitivity and timing resolution of sub-50 ps (Butera *et al.*, 2016) and are the preferred approach for achieving the highest sensitivity and resolution for brief-pulse, direct-detect lidars. Yet, in general, approximately 70 ps/1 cm range resolution is the limit for brief-pulse, direct-detect lidar and prospects for further improvement are not promising due to electronic bandwidth limitations.

Streak Cameras

The use of streak cameras with brief-pulse laser illumination has been used with success for high resolution lidar (Gelbart *et al.*, 2002; Knight *et al.*, 1989; Yang *et al.*, 2012). Streak cameras are optical detectors that use time-to-space mapping mechanism to allow high time resolution. The most common type of streak cameras use a photocathode to generate free-electrons which are accelerated and deflected in space using time varying electric fields. Streak cameras can achieve very high timing resolution (as low as 200 fs; Hamamatsu, 2008) and sensitivity as low as a single photoelectron. However, there are trade-offs with high temporal resolution. This includes timing errors dominated by trigger jitter against the electric field sweep waveform. The timing jitter can be

significantly worse (> 50 ps) than the timing resolution of the streak camera reducing absolute ranging capability and making it difficult to average low-level signals to increase sensitivity. Another problem that arises with high timing resolution and sensitivity is reduced dynamic range caused by the need to run at high photomultiplier gain. Finally, as the range window/range resolution ratio is set by the spatial resolution of the camera, the range window to range resolution ratio (i.e., time bandwidth product) can be much smaller than with other approaches. However, a significant benefit of many streak cameras is the ability to use the second spatial dimension of the camera for line imaging or other modalities such as spectroscopy.

Coherent Mode-Locked Lasers

Coherent mode-locked lasers have emerged as a powerful tool for a large range of applications including various forms of lidars. For very high resolution lidar applications several different techniques (Coddington *et al.*, 2009; van den Berg *et al.*, 2015; Delfyett *et al.*, 2012) have been developed that utilize the coherence of the mode-locked laser rather than just relying on the mode-locked laser's ability to make short pulses of light. The key difference is that these methods utilize interferometric or coherent detection methods to resolve time differences much smaller than the timing resolution of the optical detector.

Mode-locked laser based multi-wavelength interferometry

This class of mode-locked laser based very high resolution lidar combines the spectral comb structure of the mode-locked laser with interferometry and spectrally resolving elements (van den Berg *et al.*, 2012, 2015; Wu *et al.*, 2016). The basic operation utilizes a Michelson interferometer to coherently interfere the mode-locked laser with itself. The output port of the interferometer is then spectrally dispersed in space to resolve each comb line (see Fig. 2). Each resolved comb line in this interferogram then becomes an individual CW interferometer obeying Michelson interference rule of $I(\lambda) = I_0 \{ \frac{1}{2} + \frac{1}{2} \cos(4\pi R/\lambda) \}$. If the repetition rate of the comb is not large (i.e., < 1 GHz) individually resolving each comb line with an optical grating spectrometer can be difficult, so some researchers have utilized an etalon based dispersive element known as a virtually imaged phased array (VIPA) (Shirasaki, 1996) that provides a high, but periodic, dispersion in another dimension to fully resolve the comb. Fully resolving the comb is important to eliminate the pulse structure of the mode-locked laser on the optical detector, without which the pulses would also need to be temporally overlapped to see the spectral interference.

The reason why this technique can be classified under the category of lidar ranging is that with a large number of comb lines a range profile can be constructed by a fast Fourier transform (FFT) of the interferogram. The range resolution is set by the bandwidth $\Delta R = c/(2N f_{\text{rep}})$, where N is the number of resolved comb lines. However, due to the equally spaced comb structure, the interferogram is discretely sampled at the repetition rate which forms a frequency domain Nyquist limit. The range profile is then ambiguous with a period of $\tau_{\text{max}} = 1/f_{\text{rep}} \Rightarrow R_{\text{max}} = c\tau_{\text{max}}/2 = c/2f_{\text{rep}}$. For example, a 1 GHz repetition rate leads to a 15 cm ambiguity period. These ambiguities can be resolved by changing the repetition rate of the laser to sample with a different repetition rate.

It is important to note that this technique relies on the coherence of the mode-locked laser to maintain contrast of the interferogram. This means the coherence length of each individual comb lines must be larger than the range to the target. However, this technique does not require coherence between the comb lines. A laser made-up of many independent lasers with tight tolerance of the line frequencies would work as well. This also means that the technique is robust to the temporal shape of the transmitted mode-locked laser pulse, however relative dispersion and/or line-to-line relative phase errors between the Tx/Rx path and the reference path are important. In addition to the laser coherence, this technique utilizes coherent detection which means the target must maintain holographic stability with the Tx/Rx system during the integration time of the camera.

Dual-comb interferometry based distance measurement

A second method of mode-locked laser based very high resolution lidar is based on an innovative detection method that uses a second mode-locked comb as a local oscillator rather than a copy of the signal laser (Coddington *et al.*, 2009; Wu *et al.*, 2016; Durán *et al.*, 2016; Zhang *et al.*, 2014). The second comb has a slightly different repetition rate so that the relative pulse delay on

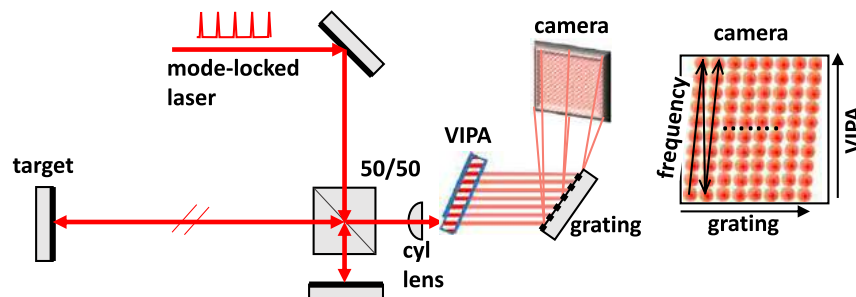


Fig. 2 A schematic of one possible realization of a mode-locked laser based multi-wavelength interferometric lidar system. Adapted with permission from van den Berg, S.A., Persijn, S.T., Kok, G.J.P., Zeitouny, M.G., Bhattacharya, N., 2012. Many-wavelength interferometry with thousands of lasers for absolute distance measurement. *Physics Review Letters* 108, 183901. doi:10.1103/PhysRevLett.108.183901.

the detector scans with a period equal to the inverse of the difference in the repetition rates ($T_{\text{meas}} = 1/\Delta f_{\text{rep}}$). If balanced homodyne detection is used the output is zero unless the pulses overlap in time. When the pulses do overlap in time the coherent interference determines the strength of the homodyne signal (see Fig. 3(a)). If the detector output is digitized with a sample rate synced to the local oscillator (LO) comb, then the output is a sampled version of the optical field of the signal laser with very high time resolution. Lidar ranging is simply accomplished by using one of the combs as a transmitter (see Fig. 4). The range profile then is a series of echo pulses in the dual-comb heterodyne output (or nonlinear sampling (Zhang *et al.*, 2014)), which not only provides the envelope of the pulse but also the pulse phase. The sampling resolution of the range profile is related to the repetition rates as $\Delta t = \Delta f_{\text{rep}} / (f_{\text{rep, sig}} \cdot f_{\text{rep, LO}})$. For example, with ~ 100 MHz repetition rates with an $\Delta f_{\text{rep}} = 1$ kHz the sampling resolution is 100 fs or effectively a 10 THz sampling rate. This time sampling condition sets the range resolution, but also must satisfy Nyquist criteria on the bandwidth of the comb to prevent aliasing. The Nyquist condition is $\Delta t \leq 1/2B$, where B is the bandwidth of the combs. This then sets the fundamental trade-offs between range resolution, unambiguous range window, and measurement rate (Wu *et al.*, 2014). These trade-offs often necessitate the use of long measurement periods and/or use of optical bandpass filters to limit the optical bandwidth.

In addition to the time domain description of this dual-comb approach, as shown in Fig. 3(b), the frequency domain description reveals it also as a massively multi-heterodyne detection technique (Coddington *et al.*, 2010). The difference in the repetition rate of the two combs causes the heterodyne beat frequency of adjacent comb modes to increase linearly with mode

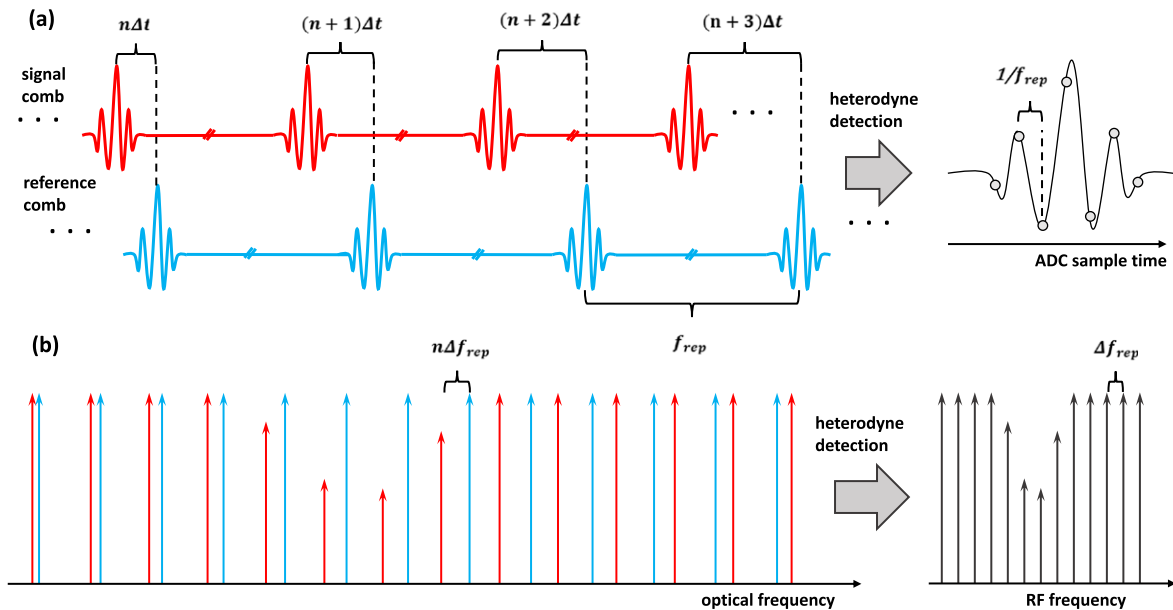


Fig. 3 Graphical description of the dual-comb approach in (a) the time domain as a cross-correlation and (b) the frequency domain as a multi-heterodyne signal showing how the optical comb translates to a radio frequency comb including individual line amplitudes. Adapted with permission from Coddington, I., Swann, W.C., Newbury, N.R., 2010. Coherent dual-comb spectroscopy at high signal-to-noise ratio. *Physical Review A* 82, 043817. doi:10.1103/PhysRevA.82.043817.

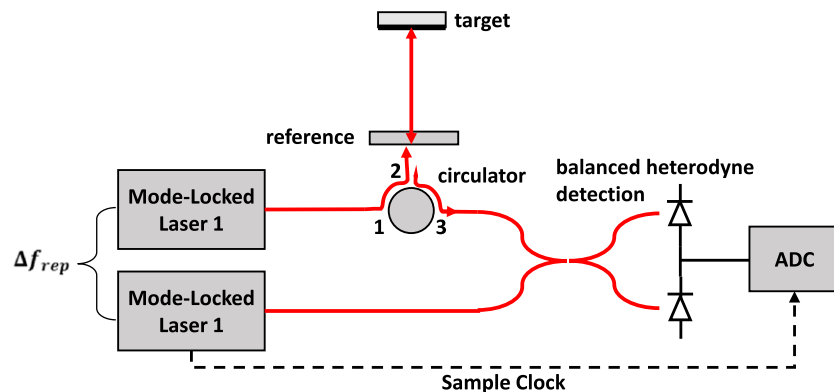


Fig. 4 Simple schematic setup used to perform dual-comb very high resolution ranging approach.

number. This effectively translates the phase and amplitude of the optical frequency comb onto a fine resolution RF comb with spacing equal to Δf_{rep} . Therefore if the amplitude of an optical comb line is reduced, for example, by a molecular absorption in the path, this will be observed on the corresponding RF comb line. While this RF comb can be obtained by use of an RF spectrum analyzer, an FFT of the time domain dual-comb signal provides a phase coherent spectrum. From this perspective the dual-comb approach is similar to Fourier transform infrared red (FTIR) spectroscopy, where the scanned delay between the signal and reference paths is provided by Δf_{rep} rather than the physical motion of an interferometer mirror.

Dual-comb interferometry has relatively strict requirements on the mutual coherence of the mode-locked lasers. The technique was originally demonstrated using highly stabilized mode-locked lasers (also known as optical frequency combs), where both the repetition rate and carrier offset degrees of freedom are locked to stable RF sources. This ensures that the relative phases of all the mode-locked lines are well defined during the measurement time and allows one to simply time average the interferogram on an oscilloscope to achieve improved SNR. An additional benefit of using highly stabilized comb lasers is that the optical phase of the return pulses in the interferogram allows interferometric distance measurement along-side the time-of-flight range measurement provided by the envelope. As demonstrated in Liu *et al.* (2011) high resolution lidar ranging can be achieved using free-running mode-locked lasers as long as the linewidth of the individual comb lines are less than the difference in the repetition rate, Δf_{rep} . However with this technique, the absolute phase of the pulse drifts with time necessitating more advanced signal processing techniques to perform averaging of the signal envelope which provides the time-of-flight lidar measurement. Other difficulties of this technique include: (1) The large optical pulse amplitudes, which can lead to nonlinearities in the photodetector and limited dynamic range of the measurement. This has been mitigated by using large fixed fiber dispersion to spread the pulse in time then using digital pulse compression in the signal processing stage. (2) The range ambiguity caused by the periodic pulse structure. This can be mitigated by making Vernier type measurements by changing the repetition rates of the lasers (Nakajima and Minoshima, 2015).

FMCW Lidar

The final very high range resolution lidar method that will be described is coherent FMCW lidar. FMCW ranging techniques have long been utilized in the RF domain for radar applications (Stove, 1992). The benefits of FMCW techniques include: continuous wave (as opposed to pulsed) output that provides higher energy per measurement at low peak powers which is also a better match to continuous waveform (CW) amplifiers, shot-noise-limited single-photon sensitivity, and high dynamic range. While many types of frequency modulated waveforms have been investigated, the dominant waveform type in use for both the radar and lidar domains is the linear frequency chirp (LFC). The LFC waveform is a linear ramp of the frequency versus time that provides a unique linear correspondence of time and frequency (see Fig. 5). Combined with coherent detection against a local copy of the LFC waveform (i.e., a LO), this structure produces a coherent beat signal, the frequency of which indicates the delay between the two LFC waveforms as $f(\tau) = \kappa\tau$, where κ is the chirp rate. Extending this to multiple range returns as in a radar or lidar system, the full range profile can be obtained from the spectrum of the coherently detected signal (generally in the form of an FFT of the digitized time domain signal). This procedure, known as stretched processing, makes a simple and bandwidth efficient signal processing system as compared to other coded waveform ranging techniques.

The main benefit of this stretched processing approach is that when the maximum relative delay of the signal versus the LO is small compared to the chirp time, T_c , the bandwidth of the coherent beat note is much less than the overall chirp bandwidth, B . This allows the use of relatively low bandwidth photodetectors and analog-to-digital converters (ADCs) while maintaining the full range resolution provided by the chirp bandwidth. Low bandwidth detectors have higher dynamic range and make it easier to achieve shot-noise-limited detection, while high speed ADCs are expensive and are improving only slowly (Walden, 1999) and the best speed versus dynamic range performance of ADCs are currently in the couple hundred megasample per second (MS/s) regime (see Section Relevant Websites). The trade-off with this bandwidth compression is to limit the maximum delay and/or range

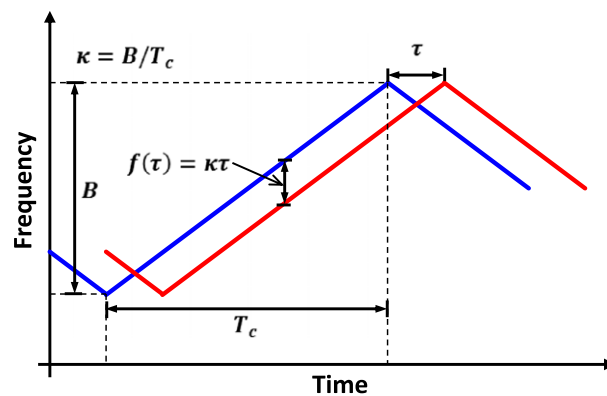


Fig. 5 Diagram of the time-to-frequency mapping performed by the coherent mixing of linear frequency chirps as in frequency modulated continuous wave (FMCW) lidar.

window. For very long range lidar applications this can be a significant issue. As an example, assume an LFC with $B = 100$ GHz of bandwidth and a chirp duration of $T_c = 1$ ms, which translates to a chirp rate of $k = 100$ MHz/ μ s. At this chirp rate, a 2 km range ($\tau = 13.3$ μ s round-trip delay) produces a frequency of ~ 1.33 GHz. This is much smaller than the 100 GHz chirp bandwidth, but still relatively large compared to the < 100 MHz regime in which one would like to operate. Additionally, to sample this high frequency signal at more than Nyquist requires significant memory and signal processing requirements. Assuming 3 GS/s sampling rate for the full T_c would produce 3 million samples that need to be processed with an FFT. For these reasons long range FMCW lidar systems may utilize hardware based range shifting mechanisms such as frequency shifters or multiple chirp sources to limit the bandwidth requirements of the photodetector, ADC and digital-signal-processing system.

A second issue with FMCW lidar that has not been brought up with the other coherent techniques is the relative motion of the target and Tx/Rx system which create Doppler shifts ($f_D = v_l/\lambda$) where v_l is the line-of-sight velocity. In direct-detect lidar systems only the optical power is detected and the optical frequency shifts caused by the Doppler shift are too small to greatly affect the measurement. For coherently detected systems, however, the Doppler shift translates in to RF frequency shifts (~ 645 KHz/(m/s) for 1550 nm light) of the detected signal. This has effects in all coherent lidar systems, but they are most consequential in FMCW lidar with LFC's because the frequency shift of the coherently detected signal directly translates to a shift in apparent range. This range-Doppler ambiguity can be resolved by making measurements at different chirp rates. One common method is to make measurements on both the up frequency chirp and the down frequency chirp of a triangle type chirp waveform as in Fig. 5. By changing the sign of the chirp rate, the relative sign of the Doppler frequency shift and range delay are reversed and the range-Doppler ambiguity can be resolved from the two measurements.

Very High Resolution FMCW Lidar

The rest of this article focuses on FMCW very high resolution lidar with discussion of the technical aspects of the technique, laser sources, and its application to imaging and metrology.

Chirp Nonlinearity

To take advantage of the bandwidth compressing ability of FMCW lidar in the very high resolution regime requires chirped laser sources with large frequency tuning range. While microwave photonic methods of generating optical chirps using microwave chirps and external modulators has been used successfully for resolutions of about 1 cm, the most successful demonstrations of very high resolution lidar to date have relied on the use of chirps created by directly tuning the output frequency of a CW laser. The main difficulty with this approach is that widely tunable lasers are generally not very stable and often do not tune linearly. An important aspect LFC based FMCW lidar is that the linear chirp enables the use of a simple FFT to compress the pulse and obtain a high resolution range profile. If the chirp is nonlinear then the linear mapping of frequency-to-time is not perfect and, as illustrated in Fig. 6, the time domain FMCW signal is no longer a single frequency sinusoid that can be compressed with a simple FFT. If the nonlinear form of the chirp is well-behaved and known it can be possible to compensate for the nonlinearities in the FMCW signal with additional signal processing. However, often the nonlinearities are unknown or vary from chirp-to-chirp, which means that some method is required to measure and correct the nonlinearities on each chirp.

The most common method for measuring the nonlinearities of a chirp from a tunable laser is to simply measure the FMCW signal from a single fixed delay, often from a fiber based reference interferometer. If the delay is fixed and not wavelength dependent (Barber *et al.*, 2010) then the FMCW signal should be a single sinusoid and the phase of the time domain signal should

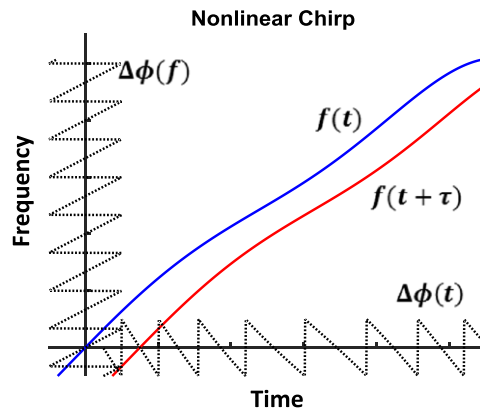


Fig. 6 Resulting FMCW signal from a nonlinear chirp. The blue and red lines represent the frequency of the imperfect chirp waveform vs. time. The dotted lines represent the relative phase of the interference signal in the time (along horizontal axis) and frequency domain (along the vertical axis).

advance perfectly linearly in time (ignoring phase wraps). Deviations from a single frequency sinusoid indicate local variations in the chirp rate with advancing phase indicating locally higher chirp rate and retarding phase indicating lower chirp rate. By tracking the phase of the FMCW signal the nonlinearities of the chirp can be determined. As illustrated in Fig. 6, with a fixed delay the relative phase of the two chirps referenced in the frequency domain is consistent because the relative phase (up to a constant overall offset phase) is just $\Delta\phi(f) = 2\pi f' t$, where f' is the instantaneous frequency. However, due to the chirp nonlinearities the time domain signal is not consistent with the frequency domain signal. An important aspect of this reference delay approach that should not be overlooked is that the interferometer signal is ambiguous to the sign of the chirp rate. This means that if the chirp is not monotonic there can be errors in tracking the phase when the chirp rate goes through zero. One way to break this ambiguity, is to add an acousto-optic modulator (AOM) to one arm of the interferometer. This shifts the beat note away from DC so that positive chirp rates show up as a positive frequency shift from the AOM frequency and negative chirp rates show up as negative shifts.

Active Chirp Laser Stabilization for FMCW Lidar

While this delay based chirp referencing method is common, how it is used varies. In very short range applications such as swept source optical coherence tomography (SS-OCT) (Huber *et al.*, 2005) the reference interferometer output is often just used to trigger samples of the ADC in the signal path at every zero phase crossing. This forces the time domain samples to be regular in frequency. Other works use the reference output to determine the chirp nonlinearities and use digital signal processing to correct the FMCW signal in post-processing (Beck *et al.*, 2005; Ahn *et al.*, 2005; Cabral and Rebordão 2007). The last approach – and what will be described in further detail here – is to use the reference interferometer output as an error signal in a phase-locked loop (PLL) to actively stabilize the frequency of the chirp (Roos *et al.*, 2009; Satyan *et al.*, 2009; Greiner *et al.*, 1998). In general, the sweep rate of most tunable lasers can be actively stabilized to achieve a linear frequency chirp suitable for lidar range measurements. However, the frequency tuning, free-running linewidth, and feedback characteristics of the particular laser will determine the details of the stabilization approach and the resulting chirp properties.

Broadband (5 THz) chirp stabilization

A first example of active chirp stabilization is that described in Roos *et al.* (2009). The laser used was developed for telecommunications test and measurement and is able to tune mod-hop free over an extremely large range (up to >80 nm from 1525 to 1605 nm). The relatively long external cavity and tunes mechanically by changing the angle of the feedback grating and so has a slow frequency sweep period, >500 ms, with a maximum chirp rate of ~10 THz/sec. Around this relatively slow sweep there are faster frequency tuning mechanisms including: a piezoelectric actuator that provides a few kHz bandwidth and 1 GHz of frequency tuning range; and direct feedback to the current of the laser diode which provides about 1 MHz feedback bandwidth. To stabilize this laser to a 70 m reference fiber interferometer required several layers of passive feedforward linearization and active stabilization using both the piezoelectric actuator for larger, slower deviations from linearity and direct current feedback for fast, small fluctuations from the finite linewidth of the laser. With both active and passive linearization a 5 THz chirp from this laser was linearized to less than 100 kHz rms (Barber *et al.*, 2011). This bandwidth provides less than 50 μm resolution at large ranges (several hundred meters limited by the rms deviations). In addition, one of the important results of this work was the realization that over this large of bandwidth (~40 nm) the wavelength dispersion of the fiber interferometer had a great effect on the linearity of the chirp. Therefore, dispersion compensating fiber was used in the reference interferometer to minimize the wavelength dispersion. Without dispersion management, the ability to perform pulse compression with a simple FFT for longer range (greater than a couple meters) FMCW lidar was compromised (Barber *et al.*, 2010).

For use in metrology applications where high accuracy distance measurements are needed, in addition to linearizing the chirp, the chirp rate also must be accurately calibrated. To provide calibration independent of a distant standard, the chirp rate of the laser could be calibrated by measuring the time versus optical frequency against a fixed set of widely separated frequency references. Two examples that were performed on this laser were: (1) The use of the NIST calibrated hydrogen cyanide (HCN) absorption lines as absolute frequency references as in Fig. 7, and (2) using the comb lines of a stabilized mode-locked laser as a frequency ruler (Barber *et al.*, 2011; Giorgetta *et al.*, 2010). The former method can achieve sub part per million calibration, and the latter could achieve part per billion calibration. The optical frequency comb based method was adapted in Baumann *et al.* (2014a,b, 2013) to provide dynamic high accuracy chirp nonlinearity measurements that could be used for post-processing based correction in high range resolution FMCW lidar based 3D metrology.

80 GHz distributed feedback based chirp stabilization

A second example of an actively linearized chirp laser is one based on a distributed feedback (DFB) laser (Satyan *et al.*, 2009). DFB lasers are common in the telecommunications industry due to their robust single-mode, narrow linewidth, and stable wavelength operation. While stable relative to other diode laser types, DFB lasers can be frequency tuned relatively rapidly and reliably using the laser current with a tuning rate on the order of 1 GHz/mA. A schematic for the active linearization and stabilization of a DFB chirp laser is shown in Fig. (8a). This setup is similar to that used in Ref. Roos *et al.* (2009) for the 5 THz tunable chirp laser except that the fiber delay line is shorter to better match the relatively larger linewidth of the DFB laser (~500 kHz) versus that of the external cavity diode laser (<50 kHz). However, this setup is more advanced in that the RF reference of the phase and frequency

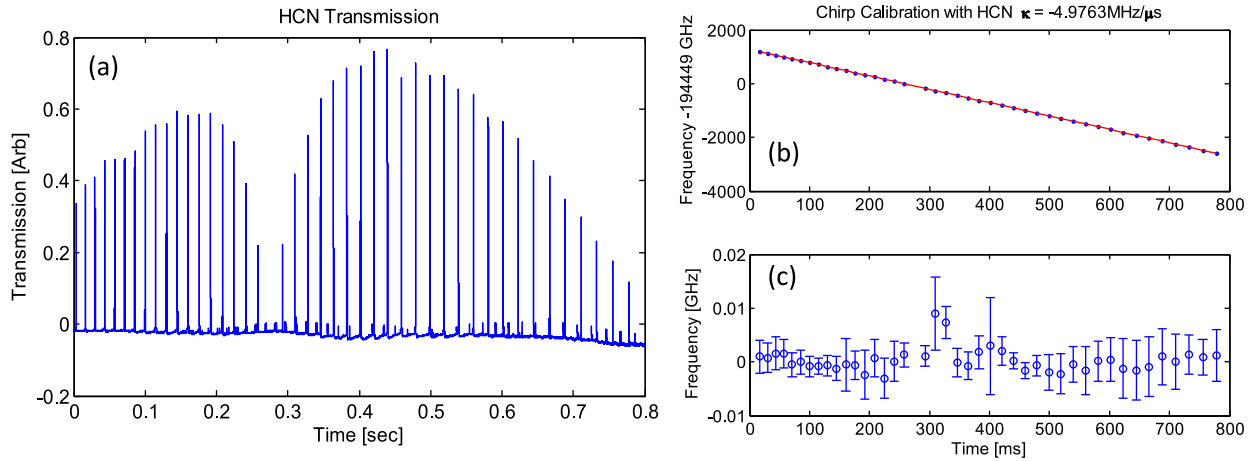


Fig. 7 (a) The transmission of a wideband chirp laser from 1530 to 1570 nm region through 10 Torr hydrogen cyanide (HCN) fiber coupled cell; (b) shows the measured HCN absorption peaks vs. chirp time use to calibrate the chirp rate and residual quadratic chirp of the 5 THz chirp laser; and (c) the residuals of the fit in (b). The error bars are the absolute frequency uncertainty provided by NIST (Gilbert *et al.*, 2005).

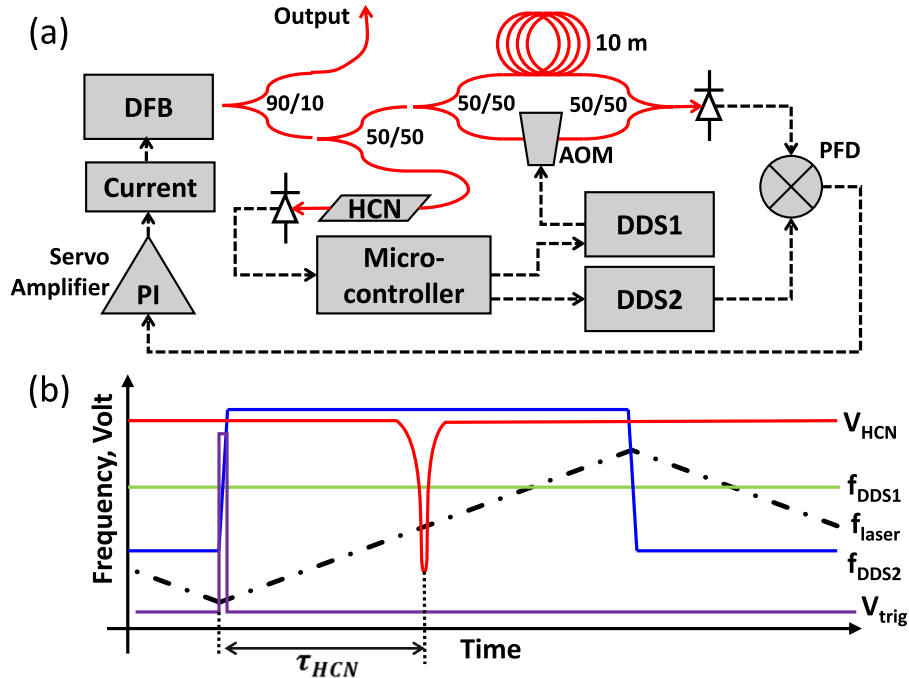


Fig. 8 (a) A schematic of a distributed feedback (DFB) chirp linearization system and center frequency stabilization based on an optical absorption line in hydrogen cyanide (HCN); and (b) a timing diagram representing how the different frequencies and voltages are used to stabilize the up-chirp, down-chirp, and the center frequency of the laser. A full description is provided in Section 80 GHz distributed feedback based chirp stabilization. DDS, direct digital synthesizer; PFD, phase and frequency detector; PI, proportional integral amplifier.

detector (PFD) used to generate the error signal for the PLL is derived from a highly controllable direct digital synthesizer (DDS) signal source. As in Roos *et al.* (2009) both the up-chirp and the return down-chirp of the laser are actively stabilized by switching the frequency of the reference to be shifted above or below the AOM frequency, respectively. However, instead of rapidly switching from one frequency to the other, here the reference frequency makes a controlled sweep between the two settings (see the f_{DDS2} line in Fig. 8(b)). This controlled sweep allows the PLL loop to remain locked at the turn-around point. This has the effect of making the center frequency of the chirp also stabilized to the fiber interferometer. This is in great contrast to the 5 THz laser chirp where the laser came completely unlocked at the turn-around points, leading to >100 MHz level jumps of the center frequency from chirp-to-chirp. Further investigations of the DFB laser, showed that while this locking around the corner provided good short term stability, drifts in the fiber length and possible timing instabilities in the triggering of the up and down sweeps could lead to slow drift of the center frequency of the laser. This long term drift could be controlled by slight (1 Hz level) adjustments of the

AOM frequency, but to provide truly long term absolute stability this setup implemented a slow digital feedback loop using an absorption line in a HCN spectroscopic cell as a frequency reference. The feedback loop was implemented using a microcontroller that measured the time difference, τ_{HCN} , between the start of the chirp and incidence of the absorption peak using a simple threshold comparator circuit. The AOM frequency is then adjusted through f_{DDS1} to stabilize the time difference. This stabilized center frequency is beneficial for SAL imaging or other interferometric applications that require the frequency of the chirps to be stable.

Applications of Very High Range Resolution Lidar

Aspherical optical metrology

The ability to perform lidar with resolutions at smaller than $50\ \mu\text{m}$ over large distances (greater than $10\ \text{m}$) lends itself very naturally to think about length metrology as an application. What is more, because FMCW lidar has the ability to resolve multiple range returns, new metrology concepts can be explored. In particular, if one applies the FMCW lidar system to an optical system the distance to every surface (even if buried inside) can be measured. This is similar to optical time domain reflectometry (OTDR) systems that are used in the fiber telecommunications industry to diagnose and find faults, but on a much finer scale of microns instead of meters. Additionally, while the range resolution of the FMCW lidar system is $50\ \mu\text{m}$, the range precision and accuracy on the measurement of distance can be much smaller due to the ability to find the center of a range peak much better than its width. Fig. 9(a) shows the range profile of a microscope slide placed in a slightly focused beam at about $2.8\ \text{m}$ of total range delay. The two peaks are well resolved by the $\approx 50\ \mu\text{m}$ range resolution at about $1.5\ \text{mm}$ apart, which is the optical thickness of the microscope slide. Both peaks have high SNR of greater than $35\ \text{dB}$. This high signal to noise allows a fit of the peaks to determine their centers to high precision. This can be confirmed by taking a series of measurements of the range profile. While the individual standard deviation of the fit to the individual range peaks is approximately $700\ \text{nm}$ (not shown), the thickness as measured by subtracting the two individual range measurements show less than $50\ \text{nm}$ standard deviation, which is approximately 1000 times less than the range resolution of an individual peak.

Extending this to three-dimensional metrology shows a potential application of very high resolution lidar to optical metrology.

Fig. 10 shows work performed to characterize highly aspherical optics. This work used an X-Y stage to make a grid of range profile

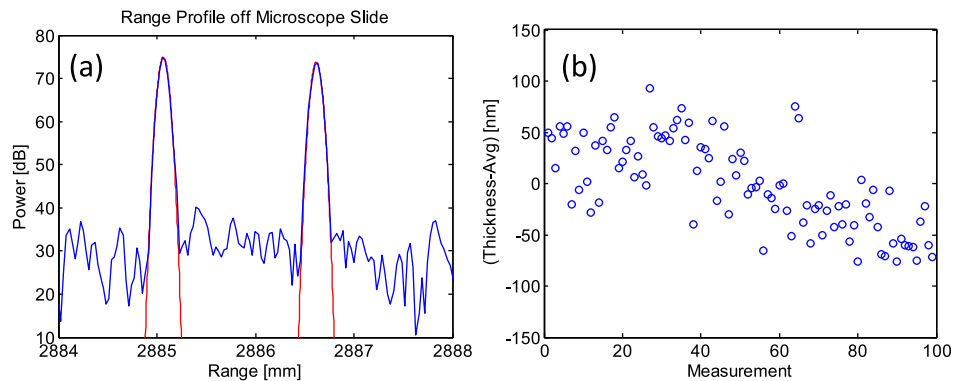


Fig. 9 (a) Range profile of the front and back surface of a microscope slide placed at about $2.8\ \text{m}$ range delay using FMCW lidar with an $\sim 5\ \text{THz}$ chirp and (b) series of optical thickness measurements taken by subtracting the range of the back and front surface. This series of measurements shows less than $50\ \text{nm}$ standard deviation.

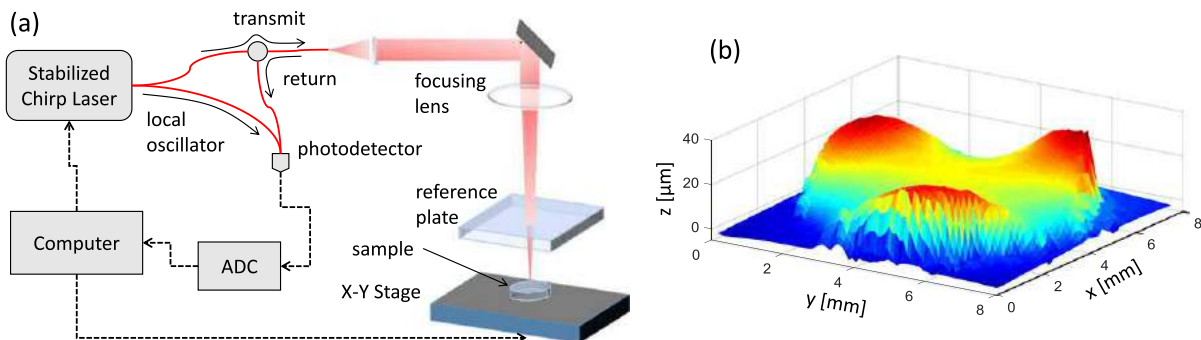


Fig. 10 (a) X-Y scanning setup to perform multi-surface profilometry and (b) 3D surface profile (vertical dimension exaggerated to show range precision) of complex aspheric optic provided by Wavesource Inc.

measurements on an optical flat whose surface was machined with a strong Zernike aberration function profile. This complex surface was machined by Wavesource Inc. as part of a research project to test the potential of FMCW lidar for optical metrology and as confirmation and testing of Wavesource Inc.'s diamond turning capability. The profile had steep slopes (nearly 50% grade at points), which make standard interferometric methods of characterization difficult. In addition to the reflection on the top surface, the high range resolution allows both the top and bottom surface to be measured simultaneously. To improve the precision and accuracy of the measurement a reference reflection near to the sample is used to remove any path length fluctuations that occur in the measurement chain. As it can be seen from Fig. 10(b) the $\sim 50\ \mu\text{m}$ surface elevations are highly resolved. Comparison with the predicted surface profile from the diamond turning process showed good quantitative agreement.

Noncontact 3D metrology

In addition to shorter range optical metrology, very high resolution lidar is also suited to 3D metrology applications over larger distances and volumes. In the 3D metrology space the highest accuracy volume metrology systems utilize interferometric based optical distance measurements that require some type of retro-reflective target to achieve high precision and accuracy ($\sim 25\ \mu\text{m}$) desired by the most demanding applications. The use of a retro-reflective targets forces the operator to manually contact the target to the part under measurement reducing productivity. The use of very high resolution FMCW lidar that can achieve high SNR from diffuse surfaces could allow noncontact and unsupervised metrological scanning. Indeed, 3D architectural and terrestrial lidar scanners (Petrie and Toth, 2008) have become one of the largest markets for lidars. Utilizing high resolution lidars in these systems to make more precise and accurate range measurements is an obvious application space.

In the area of 3D metrology with FMCW lidar, investigations have been made into the fundamental limits of precision when measuring the range to a diffuse surface (Baumann *et al.*, 2014a,b). One difficulty that was discovered was the influence of surface roughness and laser speckle on the FMCW ranging process that can lead to unpredictable range pulling with size on the order of the range resolution. A similar effect also can lead to pseudo-Doppler range pulling when the measurement beam moves rapidly across a surface. Despite these limits it was shown that with a 1 THz bandwidth laser these effects could be held to less than $10\ \mu\text{m}$ (Baumann *et al.*, 2014a,b).

Improving the range measurement in a 3D lidar scanner also necessitates that the transverse precision and accuracy of the laser scanner is commensurate with the range measurement. While angular scanning mechanisms can be made quite accurate, difficulties still arise when making precise noncontact measurements due to the size of the incident laser beam on the target. With the high precision and accuracy available with very high resolution FMCW lidar, researchers have investigated 3D metrology methods that utilize only distance measurements to position points in 3D space (Warden *et al.*, 2015; Warden, 2014; Mateo and Barber, 2015). These approaches use similar techniques to GPS where the distance between the point of interest and a few (or several) known points in 3D space is used to determine the position of the unknown point. One difficulty with these techniques is being able to provide measurement coverage in a large volume while maintaining sufficient SNR to make precise and accurate range measurements.

Advanced coherent imaging with FMCW lidar

One of the most exciting applications for coherent FMCW lidar is in the area of coherent imaging including Synthetic Aperture (Imaging) Lidar or Ladar (SAL or SAIL; Beck *et al.*, 2005; Bashkansky, 2002; Krause *et al.*, 2011; Crouch and Barber, 2012), holographic aperture lidar (HAL; Duncan and Dierking, 2009), sparse aperture imaging, digital holography, or combinations of these (Crouch *et al.*, 2015; Krause *et al.*, 2012). These coherent imaging methods seek to exploit the phase information obtained with coherent detection methods to improve the ability to image a scene and/or reduce resource requirements in an imaging system. As a direct optical analog to common radar methods, FMCW lidar is ideal for translating synthetic aperture radar (SAR) imaging methods to the optical domain. In fact, SAL was one of the motivations for the development of the 5 THz chirp laser described in Section Broadband (5 THz) chirp stabilization. An earlier investigation into SAL (Beck *et al.*, 2005), described the lack of suitable chirp laser sources as an impediment to the development of the technique. The benefit of very high resolution lidar FMCW lidar is that relatively high resolution images, i.e., large number of pixels, can be made with table-top sized experimental setups, making the investigation of SAL technique easier (Crouch and Barber, 2012).

SAL is a 2D range/cross-range imaging technique that provides imaging resolution within the diffracted limited beam of the transmitter. In one dimension the image resolution is provided by the range resolution of the lidar system, which does not depend on diffraction and means that high range resolution provides better image resolution. It also means that the target/scene must present range diversity to the lidar receiver, usually accomplished by illuminating the target at an angle. In the other dimension, the image resolution is provided by exploiting relative motion between the target/scene and the lidar transmitter/receiver. Relative motion perpendicular to the line-of-sight provides different points of view of the target and traces out a larger effective synthetic aperture for imaging. To successfully use this synthetic aperture, the system must coherently combine the information from the many range profile measurements. This necessitates the use of coherent lidar.

A basic design and simplified signal processing architecture for strip-map SAL is shown in Fig. 11. In strip-map SAL the motion of the Tx/Rx aperture, which is perpendicular to the ranging direction, also scans the Tx/Rx beam across the target. This limits the maximum size of the synthetic aperture in strip-map to the size of the beam on target. The data collection and signal processing system is relatively simple. At each aperture position the coherent FMCW lidar signal, before processing with an FFT to give a range profile, is stored in a 2D matrix called the phase history data. Because no lenses or range dimension FFT is used, this phase history data is a Fourier domain projection of the scene in the range dimension and the cross-range dimension (along the direction of motion or track). In the far-field limit, the SAL image can then be formed by a 2D FFT of the phase history data. In more physical

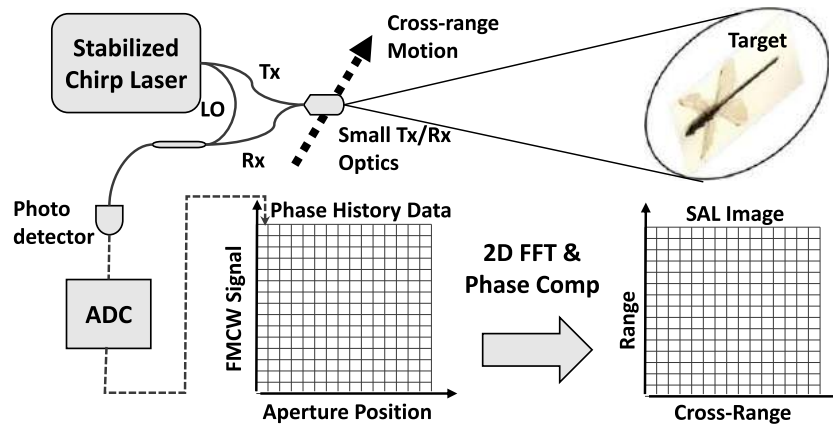


Fig. 11 Simple schematic of SAL setup and signal processing architecture.

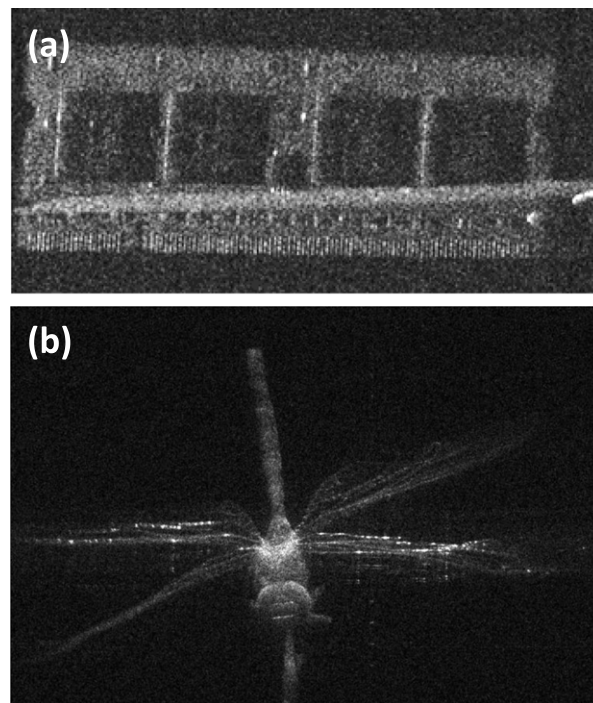


Fig. 12 Two SAL images made at $\sim 2\text{m}$ range using a 5 THz chirp laser (a) is a RAM memory chip and (b) is a dried dragonfly specimen. Republished with permission from Crouch, S.C., 2012. Synthetic aperture LADAR techniques. Master's Thesis, Montana State University – Bozeman, College of Letters & Science. Available at: <http://scholarworks.montana.edu/xmlui/handle/1/1125> (accessed 30.11.16).

situations, some phase compensation of the phase history data is needed first to account for not being infinitely far away. **Fig. 12** shows two high resolution images made on a table-top setup with the laser described in Section Broadband (5 THz) chirp stabilization.

An important aspect of SAL imaging is that it is a coherent process, which means that the phase of the phase history data is crucial. Yet, the phase of the FMCW signal is sensitive to the relative motion of the Tx/Rx aperture and the scene on the wavelength scale, i.e., holographically sensitive. While this would seem to make the SAL data collection task near impossible over any kind of distance or time, there are methods to compensate for the non-ideal motions. First, if there are known bright points in a scene these can be used as a phase reference. Second, algorithms developed for SAR such as the phase gradient autofocus (PGA) algorithm show remarkable ability to estimate and correct piston motion phase errors using the correlation of phase errors in the range dimension (Wahl *et al.*, 1994). The PGA algorithm has also shown to work quite well in the low SNR regime where the image signal less than an order of magnitude larger than the shot-noise background (Barber and Dahl, 2014). The proof of the feasibility for real-world SAL imaging is that air-to-ground SAL imaging has actually been demonstrated from an aircraft (Krause *et al.*, 2011), however work is still required for practical applicability.

Conclusions

The objective of this article was to introduce the reader to technologies, techniques, and applications in the area of very high range resolution (< 1 cm) lidar. In general, these techniques are not brief-pulse, direct-detection lidars. They do, however, often utilize coherent lidar techniques with lasers that have unique spectral domain coherence. Clever exploitation of this spectral coherence reduces the requirements for high speed optoelectronics allowing range resolution of tens of microns – the equivalent of THz bandwidths which is far beyond current electronics. The article then focused on coherent FMCW lidar to introduce interesting applications for very high resolution lidar in metrology and coherent imaging. While precision optical distance measurement techniques for metrology have been around for a long time, the difference between these older methods and the newer lidar techniques is the ability to fully resolve a range profile, not just single ranges. With the ability to resolve range profiles, multiple returns in the path of the lidar can be measured simultaneously, which provides unique capabilities in the area of optical metrology. In addition, the high sensitivity of these coherent lidar techniques provide the ability to measure distances to diffuse surfaces precisely and accurately. Lastly, very high resolution coherent FMCW lidar has been used to demonstrate coherent imaging modalities such as synthetic aperture lidar (SAL).

Acknowledgments

This article represents my understanding of the state of the field of very high resolution lidars. I have been working in this field for only 8 years and a lot of great work had been performed before I entered the field and continues to this day. About once a month I receive peer-review requests for articles in the area of coherent lidar and the field only seems to be growing. As one person, it is difficult to keep up to date on all the advances in even this relatively narrow field, and I am sure that I overlooked much excellent work in high resolution lidar. While I decided to heap all the errors, omissions, and opinions contained in this article onto myself as sole author, none of my knowledge of the field could have been gained without the help and assistance of many people including current and former students, colleagues and coworkers, industry and government collaborators, and acquaintances.

In particular I need to acknowledge: My former students Stephen Crouch and Ana Baselga Mateo who performed the work for their thesis research that led to my understanding of the issues and applications of very high resolution FMCW lidar. My colleagues at Bridger Photonics and Blackmore Sensors and Analytics, Dr. Peter Roos and Dr. Randy Reibel who introduced me to FMCW lidar and helped me earn my first research grants in the area and much sub-award funding over the years. Wavesource Inc. for collaborating on initial optical metrology measurements by providing samples and sub-award funding. My collaborators at NIST in Boulder, CO Ian Coddington, Esther Baumann, Fabrizio Giorgetta, and Nathan Newbury for showing initial interest in this area and then by greatly pushing the field of precision metrology with very high resolution FMCW lidar. Finally, I would like to thank collaborators in the AFRL/Rymm – LADAR Technology Branch.

Additional acknowledgment must go to Nate Newbury for providing permission to adapt Figure 3 from their excellent paper on dual-comb measurement methods. Thanks is required for Dr. Steven van den Berg of the Dutch Metrology Institute for giving permission to adapt Figure 2 from an article on his excellent work in mode-locked laser based multi-heterodyne interferometry.

Parts of this work were supported by grants through the Montana Board of Research and Commercialization Technology (MBRCT), also by the National Science Foundation through GOALI grant #1031211, and by the United States Air Force through a Young Investigator Program Award.

See also: Micro-Lidars for Short Range Detection and Measurement

References

- Ackerman, E., 2016. Cheap lidar: The key to making self-driving cars affordable. IEEE Spectrum: Technology, Engineering, and Science News. Available at: <http://spectrum.ieee.org/transportation/advanced-cars/cheap-lidar-the-key-to-making-selfdriving-cars-affordable> (accessed 30.11.16).
- Ahn, T.-J., Lee, J.Y., Kim, D.Y., 2005. Suppression of nonlinear frequency sweep in an optical frequency-domain reflectometer by use of Hilbert transformation. Applied Optics 44, 7630–7634. Available at: <https://www.osapublishing.org/abstract.cfm?uri=ao-44-35-7630> (accessed 30.11.16).
- Barber, Z.W., Babbitt, W.R., Kaylor, B., Reibel, R.R., Roos, P.A., 2010. Accuracy of active chirp linearization for broadband frequency modulated continuous wave lidar. Applied Optics 49, 213–219. Available at: <http://www.opticsinfobase.org/abstract.cfm?uri=ao-49-2-213> (accessed 26.01.15).
- Barber, Z.W., Dahl, J.R., 2014. Synthetic aperture lidar imaging demonstrations and information at very low return levels. Applied Optics 53, 5531–5537. doi:10.1364/AO.53.005531.
- Barber, Z.W., Giorgetta, F.R., Roos, P.A., et al., 2011. Characterization of an actively linearized ultrabroadband chirped laser with a fiber-laser optical frequency comb. Optics Letters 36, 1152–1154. Available at: <http://www.opticsinfobase.org/abstract.cfm?uri=ol-36-7-1152> (accessed 26.01.15).
- Bashkansky, M., 2002. Synthetic aperture imaging at 1.5 μ m: Laboratory demonstration and potential application to planet surface studies. SPIE 4849, 48–56. doi:10.1117/12.460767.
- Baumann, E., Deschênes, J.-D., Giorgetta, F.R., et al., 2014a. Speckle phase noise in coherent laser ranging: Fundamental precision limitations. Optics Letters 39, 4776–4779. doi:10.1364/OL.39.004776.

- Baumann, E., Giorgetta, F.R., Coddington, I., *et al.*, 2013. Comb-calibrated frequency-modulated continuous-wave lidar for absolute distance measurements. *Optics Letters* 38, 2026–2028. doi:10.1364/OL.38.002026.
- Baumann, E., Giorgetta, F.R., Deschênes, J.-D., *et al.*, 2014b. Comb-calibrated laser ranging for three-dimensional surface profiling with micrometer-level precision at a distance. *Optics Express* 22, 24914–24928. doi:10.1364/OE.22.024914.
- Beck, S.M., Buck, J.R., Buell, W.F., *et al.*, 2005. Synthetic-aperture imaging laser radar: Laboratory demonstration and signal processing. *Applied Optics* 44, 7621–7629. doi:10.1364/AO.44.007621.
- Butera, S., Vines, P., Tan, C.H., Sandall, I., Buller, G.S., 2016. Picosecond laser ranging at wavelengths up to 2.4 μm using an InAs avalanche photodiode. *Electronics Letters* 52, 385–386. doi:10.1049/el.2015.3995.
- Cabral, A., Rebordão, J., 2007. Accuracy of frequency-sweeping interferometry for absolute distance metrology. *Optical Engineering* 46. doi:10.1117/1.2754308.
- Coddington, I., Swann, W.C., Nenadovic, L., Newbury, N.R., 2009. Rapid and precise absolute distance measurements at long range. *Nature Photonics* 3, 351–356. doi:10.1038/nphoton.2009.94.
- Coddington, I., Swann, W.C., Newbury, N.R., 2010. Coherent dual-comb spectroscopy at high signal-to-noise ratio. *Physical Review A* 82, 043817. doi:10.1103/PhysRevA.82.043817.
- Crouch, S., Barber, Z.W., 2012. Laboratory demonstrations of interferometric and spotlight synthetic aperture lidar techniques. *Optics Express* 20, 24237. doi:10.1364/OE.20.024237.
- Crouch, S., Kaylor, B.M., Barber, Z.W., Reibel, R.R., 2015. Three dimensional digital holographic aperture synthesis. *Optics Express* 23, 23811. doi:10.1364/OE.23.023811.
- De Groot, P., McGarvey, J., 1992. Chirped synthetic-wavelength interferometry. *Optics Letters* 17, 1626–1628. Available at: <https://www.osapublishing.org/abstract.cfm?uri=ol-17-22-1626> (accessed 30.11.16).
- Delfyett, P.J., Mandridis, D., Piracha, M.U., *et al.*, 2012. Chirped pulse laser sources and applications. *Progress in Quantum Electronics* 36, 475–540. doi:10.1016/j.pquantelec.2012.10.001.
- Duncan, B.D., Dierking, M.P., 2009. Holographic aperture lidar. *Applied Optics* 48, 1168–1177. Available at: <http://www.opticsinfobase.org/ao/fulltext.cfm?uri=ao-48-6-1168&id=176694> (accessed 23.04.15).
- Durán, V., Andrekson, P.A., Torres-Company, V., 2016. Electro-optic dual-comb interferometry over 40 nm bandwidth. *Optics Letters* 41, 4190. doi:10.1364/OL.41.004190.
- Fujima, I., Iwasaki, S., Seta, K., 1998. High-resolution distance meter using optical intensity modulation at 28 GHz. *Measurement Science and Technology* 9, 1049. Available at: <http://iopscience.iop.org/article/10.1088/0957-0233/9/7/007/meta> (accessed 30.11.16).
- Gelbart, A., Redman, B.C., Light, R.S., Schwartzlow, C.A., Griffis, A.J., 2002. Flash lidar based on multiple-slit streak tube imaging lidar. *SPIE* 4723, 9–18. doi:10.1117/12.476407.
- Gilbert, S.L., Swann, W.C., Wang, C.-M., 2005. Hydrogen cyanide H13C14N absorption reference for 1530 nm to 1565 nm wavelength calibration—SRM 2519a. NIST Special Publication 260, 137. Available at: <http://132.163.4.18/srm/upload/SP260-137.pdf> (accessed 9.02.15).
- Giorgetta, F.R., Coddington, I., Baumann, E., Swann, W.C., Newbury, N.R., 2010. Fast high-resolution spectroscopy of dynamic continuous-wave laser sources. *Nature Photonics* 4, 853–857. doi:10.1038/nphoton.2010.228.
- Greiner, C., Boggs, B., Wang, T., Mossberg, T.W., 1998. Laser frequency stabilization by means of optical self-heterodyne beat-frequency control. *Optics Letters* 23, 1280–1282. doi:10.1364/OL.23.001280.
- Hamamatsu 2008. Guide to Streak Cameras, Hamamatsu Photonics K.K. Available at: http://www.hamamatsu.com/resources/pdf/sys/SHSS0006E_STREAK.pdf (accessed 31.10.16).
- Huber, R., Wojtkowski, M., Fujimoto, J.G., Jiang, J.Y., Cable, A.E., 2005. Three-dimensional and C-mode OCT imaging with a compact, frequency swept laser source at 1300 nm. *Optics Express* 13, 10523–10538. doi:10.1364/OPEX.13.010523.
- Iddan, G.J., Yahav, G., 2001. Three-dimensional imaging in the studio and elsewhere. *SPIE* 4298, 48–55. doi:10.1117/12.424913.
- Khoshelham, K., Zlatanova, S., 2016. Sensors for indoor mapping and navigation. *Sensors* 16. doi:10.3390/s16050655.
- Knight, F.K., Klick, D., Ryan-Howard, D.P., *et al.*, 1989. Laser radar reflective tomography utilizing a streak camera for precise range resolution. *Applied Optics* 28, 2196. doi:10.1364/AO.28.002196.
- Krause, B.W., Buck, J., Ryan, C., *et al.*, 2011. Synthetic aperture lidar flight demonstration. In: *Conference on Lasers and Electro-Optics, CLEO, IEEE*, pp. 1–2.
- Krause, B.W., Tiemann, B.G., Gatt, P., 2012. Motion compensated frequency modulated continuous wave 3D coherent imaging lidar with scannerless architecture. *Applied Optics* 51, 8745–8761. Available at: <http://www.osapublishing.org/abstract.cfm?uri=ao-51-36-8745> (accessed 2.08.15).
- Liu, T.-A., Newbury, N.R., Coddington, I., 2011. Sub-micron absolute distance measurements in sub-millisecond times with dual free-running femtosecond Er fiber-lasers. *Optics Express* 19, 18501. doi:10.1364/OE.19.018501.
- Mateo, A.B., Barber, Z.W., 2015. Multi-dimensional, non-contact metrology using trilateration and high resolution FMCW lidar. *Applied Optics* 54, 5911. doi:10.1364/AO.54.005911.
- Nakajima, Y., Minoshima, K., 2015. Highly stabilized optical frequency comb interferometer with a long fiber-based reference path towards arbitrary distance measurement. *Optics Express* 23, 25979. doi:10.1364/OE.23.025979.
- Petrie, G., Toth, C., 2008. Terrestrial laser scanners. In: Shan, J., Toth, C.K. (Eds.), *Topographic Laser Ranging and Scanning*. Boca Raton, FL: CRC Press, pp. 87–128. Available at: <http://www.crcnetbase.com/doi/abs/10.1201/9781420051438.ch3> (accessed 30.11.16).
- Roos, P.A., Reibel, R.R., Berg, T., *et al.*, 2009. Ultrabroadband optical chirp linearization for precision metrology applications. *Optics Letters* 34, 3692–3694. Available at: <http://www.opticsinfobase.org/abstract.cfm?uri=OL-34-23-3692> (accessed 26.01.15).
- Satyan, N., Vasilyev, A., Rakuljic, G., Leyva, V., Yariv, A., 2009. Precise control of broadband frequency chirps using optoelectronic feedback. *Optics Express* 17, 15991–15999.
- Shirasaki, M., 1996. Large angular dispersion by a virtually imaged phased array and its application to a wavelength demultiplexer. *Optics Letters* 21, 366–368. Available at: <https://www.osapublishing.org/abstract.cfm?uri=ol-21-5-366> (accessed 1.12.16).
- Spühler, G.J., Paschotta, R., Fluck, R., *et al.*, 1999. Experimentally confirmed design guidelines for passively Q-switched microchip lasers using semiconductor saturable absorbers. *Journal of the Optical Society of America B* 16, 376. doi:10.1364/JOSAB.16.000376.
- Stove, A.G., 1992. Linear FMCW radar techniques. *IEE Proceedings F – Radar Signal Process* 139, 343–350. doi:10.1049/ip-f-2.1992.0048.
- van den Berg, S.A., Persijn, S.T., Kok, G.J.P., Zeitoony, M.G., Bhattacharya, N., 2012. Many-wavelength interferometry with thousands of lasers for absolute distance measurement. *Physical Review Letters* 108, 183901. doi:10.1103/PhysRevLett.108.183901.
- van den Berg, S.A., van Eldik, S., Bhattacharya, N., 2015. Mode-resolved frequency comb interferometry for high-accuracy long distance measurement. *Science Reports* 5, 14661. doi:10.1038/srep14661.
- Wahl, D.E., Eichel, P.H., Ghiglia, D.C., Jakowatz, C.V., 1994. Phase gradient autofocus – A robust tool for high resolution SAR phase correction. *IEEE Transactions on Aerospace and Electronic Systems* 30, 827–835. doi:10.1109/7.303752.
- Walden, R.H., 1999. Analog-to-digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications* 17, 539–550. doi:10.1109/49.761034.
- Warden, M.S., 2014. Precision of frequency scanning interferometry distance measurements in the presence of noise. *Applied Optics* 53, 5800. doi:10.1364/AO.53.005800.
- Warden, M.S., Campbell, M., Hughes, B., Lewis, A., 2015. *GPS Style Position Measurement With Optical Wavelengths*. Washington, DC: OSA.
- Wu, H., Zhang, F., Liu, T., Balling, P., Ou, X., 2016. Absolute distance measurement by multi-heterodyne interferometry using a frequency comb and a cavity-stabilized tunable laser. *Applied Optics* 55, 4210. doi:10.1364/AO.55.004210.
- Wu, G., Zhou, Q., Shen, L., *et al.*, 2014. Experimental optimization of the repetition rate difference in dual-comb ranging system. *Applied Physics Express* 7, 106602. doi:10.7567/APEX.7.106602.

- Yang, H., Wu, L., Wang, X., *et al.*, 2012. Signal-to-noise performance analysis of streak tube imaging lidar systems I Cascaded model. *Applied Optics* 51, 8825. doi:10.1364/AO.51.008825.
- Zhang, H., Wei, H., Wu, X., Yang, H., Li, Y., 2014. Absolute distance measurement by dual-comb nonlinear asynchronous optical sampling. *Optics Express* 22, 6597. doi:10.1364/OE.22.006597.

Relevant Websites

- <https://www.wired.com/2015/09/laser-breakthrough-speed-rise-self-driving-cars/>
A.C.M.C.M. Business, Laser Breakthrough Could Speed the Rise of Self-Driving Cars, WIRED.
- <http://web.stanford.edu/~murrmann/adcsurvey.html>
Boris Murrmann: ADC Survey.
- <http://www.radartutorial.eu/01.basics/Range%20Resolution.en.html>
Radar Basics: Range Resolution.
- https://en.wikipedia.org/w/index.php?title=Time-of-flight_camera&oldid=751907785
Time-of-Flight Camera.

Multi-Dimensional Laser Radars

Vasyl V Molebny, Academy of Technological Sciences of Ukraine, Kiev, Ukraine

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Despite the word “ranging” in the term “laser detection and ranging” implies acquiring the information on range, the other parameters measured by laser radar can have its own dimensionality, for example, object orientation in space and its velocity, light scattering in the media, refraction of the media, temperature, humidity, gases concentration, etc. Some of them are single-dimensional, like temperature. Some can have several dimensions, like temperature distribution in space or temperature dependence on time. Other types of examples are combinations of several parameters distributed in space and time, like 3D velocity of an object having its 3D position and 3D orientation in space, both changing in time, thus climbing up to a hypothetical ten-dimensional (10D) laser radar system. Extending the popular term “3D (three-dimensional)”, Molebny and Steinvall (2014) introduced the term “multi-dimensional” meaning that the laser radar outputs the information defined by three or more independent descriptors, which can be regarded as dimensions. This publication will serve as a canvas for our article. Examples of information delivered by three-dimensional laser radars are illustrated in Table 1.

Examples of multi-dimensional laser radar systems are illustrated in Table 2.

Measurement of any of the dimensions can be provided with a single transmitter and a single receiver or with several of each of them. Moreover, there could be a combination of different laser radar systems or even a combination of laser radars with passive infrared systems or with microwave radars. Last decade brought hundreds of studies and corresponding publications on a variety of designs, applications, theories, experiments and field tests. We report here some representatives of them.

Stepping up the Dimensionality

The number of dimensions acquired by laser radar can be increased when adding new functions. For example, scanning the optical axis of a simple rangefinder and storing the measured range at different positions of the optical axis one can reconstruct the profile of the target thus getting its three-dimensional image.

The range information for reconstruction of the 3D image can be got not only by changing the direction of the optical axis sequentially in time (scanning), but by illuminating (flashing) the space in a wide field of view and receiving the information from all directions simultaneously.

In the first case, the range dimension is supplemented with two new dimensions (x, y) or (α, ϵ). In the second case, the two-dimensional image is supplemented with the range information. Here are the examples of these two approaches.

Table 1 Three-dimensional description of the parameters acquired by a laser radar

D_1	D_2	D_3	Comments
x	x	R	Range (R) in orthogonal (x, y, R) system of co-ordinates
x, α	y, ϵ	R, d	Range (R) or depth (d) in orthogonal (x, y, R), (x, y, d), or spherical (α, ϵ, R), (α, ϵ, d), system of co-ordinates
α	ϵ	V_r	Radial velocity in spherical (α, ϵ, V_r) system of co-ordinates
x, α	y, ϵ	A, f	Distribution of amplitude (A) or frequency (f) of vibrations on the target surface
f	t	A	Frequency (f) and amplitude (A) of vibrations of a point on the target surface in time (t)
x, α	y, ϵ	β, c, h, n, T, p	Distribution in space of scatter (β), gas concentration (c), humidity (h), refraction (n), temperature (T), polarization (p)

Table 2 Examples of multi-dimensional description of laser radar information (4D and more)

D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	Comments
α	ϵ	V_r	t							Radial velocity varying in time
x	y	T	V_x	V_y						Temperature and velocity distribution in a plane
x	y	z	V_x	V_y	V_z					Velocity distribution in a volume
x	y	z	V_x	V_y	V_z	t				Velocity distribution in the volume varying in time
α	ϵ	R	V_α	V_ϵ	V_R	α_x	ϵ_η	$(\varphi)_\zeta$	t	Range, velocity and object orientation varying in time

Ranging Supplemented by Two-Dimensional Scanning

One of the very first 3D laser radars developed under NASA contract used two-axis piezoelectric driven mirror scanners providing 350×350 resolvable scan elements (Flom, 1972). These elements, for both the transmitter and the receiver, were aligned and their positions were electronically controlled, thereby creating a synchronously scanned transmitter-receiver in a total search field of view 30×30 degrees. The wavelength of the GaAs laser was $0.9 \mu\text{m}$. Pulse repetition rate was controlled depending on the range to the target.

In 1960s, the antimissile defense (AMD) was a strategic problem at both sides of Atlantic. To improve the precision of the AMD radars, an experimental laser radar was built and tested (Molebny *et al.*, 2010) with impressive dimensions and parameters. It consisted of 196 lasers and the same number of range-gated photomultipliers (4 groups of 7×7 arrays, totaling a 14×14 array). Initial target designation from microwave radar was got with an error ± 3.5 arc. min. Therefore, the total field of view (7×7) arc. min. was split into (100×100) arc. sec. sub-zones, to be illuminated in sequence. Each ruby laser pulse had energy of 1 J pulse repetition frequency 10 Hz, pulse duration 30 ns.

In 1980s, an imaging CW heterodyne ladar was tested (Wang *et al.*, 1984). The CO_2 laser produces a 5 W beam at $10.6 \mu\text{m}$. The beam is split into a signal beam and a local oscillator beam. The signal beam is frequency shifted, intensity modulated at 15 MHz, and scanned across the field of view by a galvanometer scanner controlled by a microcomputer. The scanner covers 512 lines in a (12×12) -deg. field of view. From the received signal, 15 MHz is filtered, amplified, and detected to produce amplitude and phase data. These data are then recorded on a magnetic tape. The recorded data are read by a computer in the laboratory and presented on a video monitor. Reflectance has a speckled appearance, horizontal stripes in the range images represent 10 m ambiguity steps produced by the amplitude modulated ranging system. Range resolution was about 0.3 m at a range of 50 m.

Two-Dimensional Imaging Supplemented by Scanning in z Direction

The technology of scanning in z direction can be implemented by range gating – a technique widely used in microwave radars. In the case of laser radar, pulses of picosecond duration allow to section the space in very thin “slices”. High-speed cameras with a highly sensitive CCD and short gate times provide gated viewing with automatic gain control versus range, whereby foreground backscatter can be suppressed. Compensation can be automated for decreasing reflection with range as z (Flom, 1972), or an exponential damping of the light in absorbing media. A 3D laser radar prototype implementing this technology was reported (Busck and Heiselberg, 2004). The system uses a picosecond Q-switched Nd:YAG laser at 532 nm with a 32 kHz pulse repetition frequency, which triggers a (752×582) -pixel CCD camera.

Developing the idea of cutting the range in slices, a set of image slices at different distances can be reconstructed. Some of these slices can be removed (e.g., obscuring objects, camouflage, clutter, etc.). In one of the approaches of such processing, the first step consists in selecting the distances of interest in the histogram, for example, distances around the ranges of 2.5, 7.5, and 16.0 m (Fig. 1), where maximal numbers of pixels are concentrated. Additional processing can be made involving a 3D mesh to analyze the cloud points both in (x, y) and z directions. Processing for laser gated viewing was developed by TNO Defence, Security and Safety (Netherlands) (Bovenkamp and Schutte, 2010) based on Intevac laser gated viewer. 1.5- μm eye safe laser was used and a CMOS camera for gated viewing (minimum gate shift was 1.0 m, minimum gate width – 20 m). The results of image processing are demonstrated in Fig. 2.

Intevac’s Electron Bombarded Active Pixel Sensor (EBAPS) technology is based on a III-V semiconductor photocathode in proximity-focus with a high resolution, backside-thinned, CMOS chip anode. The electrons emitted by the photocathode are

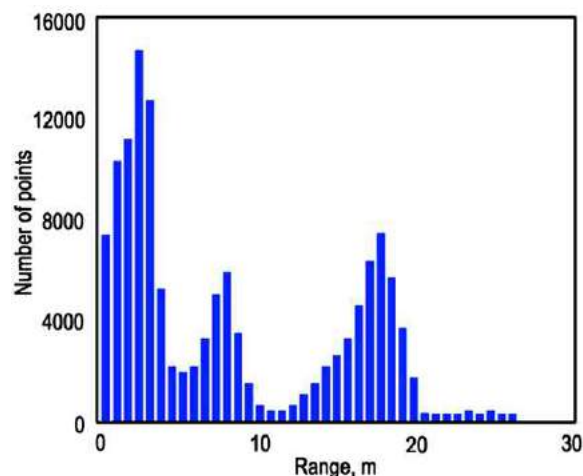


Fig. 1 Histogram of range image. Author's courtesy Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. Proc. SPIE7684, 768435.

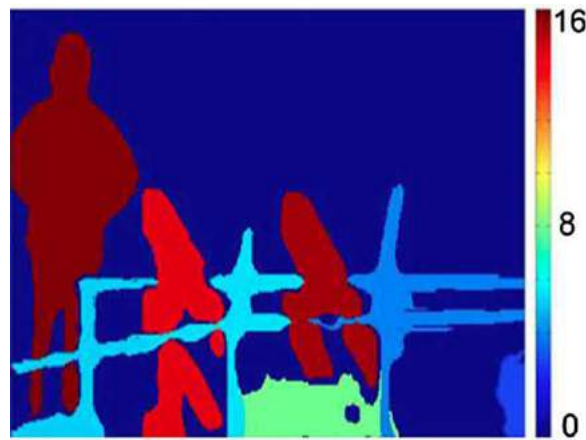


Fig. 2 2D projections of objects at different distances. Reproduced from Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. Proc. SPIE7684, 768435.

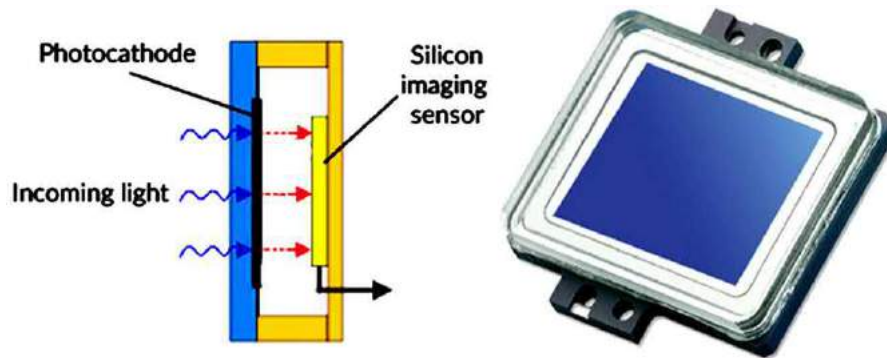


Fig. 3 Intevac's electron bombarded active pixel sensor (EBAPS). Reproduced from Electron Bombarded Active Pixel Sensor (EBAPS). Available at: <http://www.intevac.com>.

directly injected in the electron bombarded mode into the CMOS anode, where the electrons are collected, amplified and read-out to produce digital video directly out of the sensor (Electron Bombarded Active Pixel Sensor (EBAPS)) (Fig. 3).

It is the further development of the hybrid photomultiplier tube sensitive from 900 to 1300 nm referred to as an intensified photodiode (Bradbury *et al.*, 1997). It uses an active electron photocathode and a GaAs Schottky avalanche photodiode as an anode. Its first stage amplification consists of a very low noise electron-hole multiplication. This is then followed by traditional APD multiplication. These two processes are sufficient to overcome preamplifier noise, and the photodetector is therefore capable of photon counting.

Another option of 3D imaging with laser radar is the illumination of the whole field of view with one or small number of short pulses. This technology called "flash ladar" is based on range-gated imaging. Fig. 4 illustrates the time "slices" 1, 2, 3, ..., 20 operated by a (128×128) FPA (focal plane array) with a ROIC (read out and integration circuit). This flash ladar was developed by Northrop Grumman Aerospace Systems (Wong *et al.*, 2010). It can store an image every 0.5 ns. The ROIC captures 20 time lapsed images for each pulse which are then used to calculate the range estimates of the target area. Laser flash lamp pumps Nd: YAG OPO, 1.57 μm , 50 mJ, 6.7 ns, 30 Hz. Time per slice is 2.2 ns.

Advanced Scientific Concepts (ASC) developed the TigerEye flash 3D ladar video camera and DragonEye space camera (Stettner, 2010) that can assist almost any manned or unmanned vehicle with collision avoidance, navigation, etc. Each of 128×128 pixels is "triggered" independently, allowing capture of 16,384 range data points to generate a 3D point cloud image up to 30 frames per second.

Selex developed advanced infrared detectors for multimode active and passive imaging applications (Baker *et al.*, 2008) that can be electronically switched from a thermal imaging function to a laser gated imaging function so that passive thermal imaging, solar flux imaging and laser-gated imaging can be designed into one electro-optic system. The wavelength sensitivity of these detectors (APD arrays in HgCdTe) ranges from 1 μm to 4.5 μm which is promising for active multispectral imaging in the SWIR and MWIR regions.

Another promising technology of 3D imaging developed at Lincoln Lab (Albota *et al.*, 2002a,b), is photon counting. The technique is based on a 2D detector array. Rather than measuring intensity, these detectors measure the photon time-of-arrival.

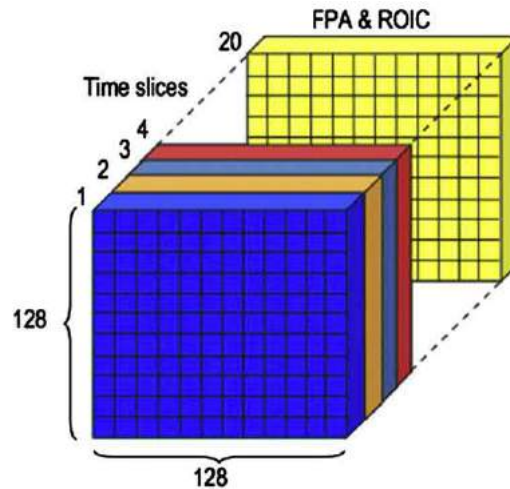


Fig. 4 Image building in a flash ladar. Reproduced from Wong, C.M., Logan, J.E., Bracikowski, C., Baldauf, B.K., 2010. Automated in-track and cross-track airborne flash LADAR image registration for wide-area mapping. Proc. SPIE7684, 768427.

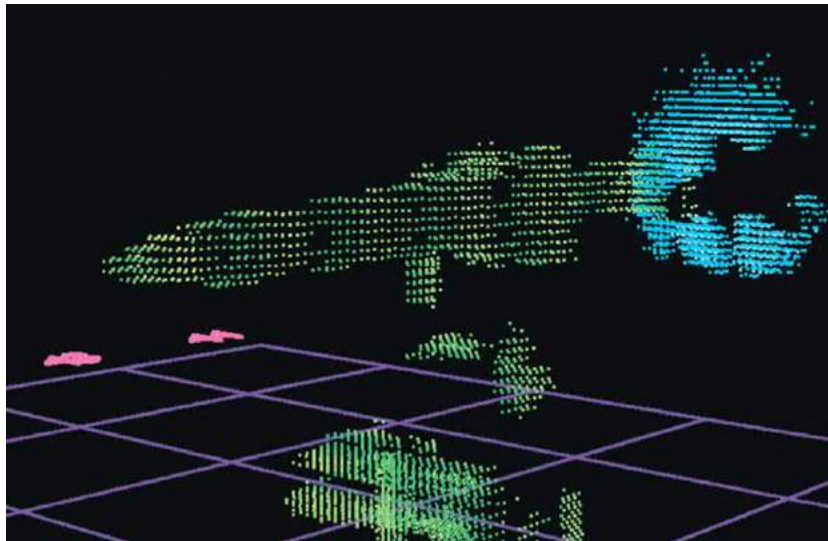


Fig. 5 Ladar reconstructed image of a missile model with a 32×32 APD array, 16 kHz repetition rate. Reproduced from Cho, P., Anderson, H., Hatch, R., Ramaswami, P., 2006. Real-time 3D ladar imaging. Lincoln Lab. J.16 (1), 147–164.

With each pixel coded in range, the ladar produces a 3D image (angle-angle-range). The illuminator is a passively Q-switched laser. The electro-optical receiver is a 4×4 array of avalanche photodiodes (APDs) that is passively quenched and gated to operate in photon-counting (Geiger) mode. Each of the 16 APD pixels in the detector array is linked to timing circuitry. The field of view of the 4×4 array is scanned to generate larger images with up to 128×128 pixels.

Doubled Nd:YAG laser microchip generates 250 ps pulses at 532 nm. It is pumped with 1.2 W at 808 nm. Repetition rate is 1 kHz, output energy per pulse is $3 \mu\text{J}$. This energy is enough to image the targets nearer than 1 km. 3D images at ranges beyond 1 km with a $30 \mu\text{J}$ per pulse source were also reported. In response to a single photoelectron, Geiger-mode APDs yield a fast, high-amplitude electrical pulse that triggers the timing circuitry. The next step at Lincoln Lab was designing a 32×32 APD array with 16 kHz repetition rate (Cho *et al.*, 2006). Single photon counting and linear mode photon counting made great progress in the last decade. An example of 3D imaging is demonstrated in Fig. 5.

For operation in the short-wavelength infrared (1.06 μm), arrays of InP-based avalanche photodiodes were developed in the 128×32 format (Verghese *et al.*, 2009).

One of the perspective applications of laser radar is identification capability based on measuring the range profiles reconstructed in the form of histograms from a combination with a detection sensor. The detection sensor gives the location of the threat, while the laser range profiler (LRP) can identify the target based on high range-resolution data. Heuvel van den *et al.* (2008) tested the LRP with the Infrared Search and Track (IRST) sensor on board of a naval vessel.

It was shown that range profiles of ships can be obtained up to a range of 10 km. The range profiles have a good correspondence with the geometry of the ship and do not vary rapidly with aspect angle like microwave radar range profiles. Fig. 6 shows the LRP prototype used for the MCG/8 NATO trial in Stavanger, Norway. Fig. 7 shows the 3D model of a frigate illuminated from the front side. The model was used to calculate the range profile. An example of the reconstructed range profile at a distance 3.6 km is shown in Fig. 8.

The profiler gives the range and can discriminate between false alarms and potential threats. False alarms that can limit IRST usefulness like birds, cloud edges, and sun glints can be easily discriminated from real targets by a laser range profiler.



Fig. 6 Laser Range Profiler for the MCG/8 trial. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

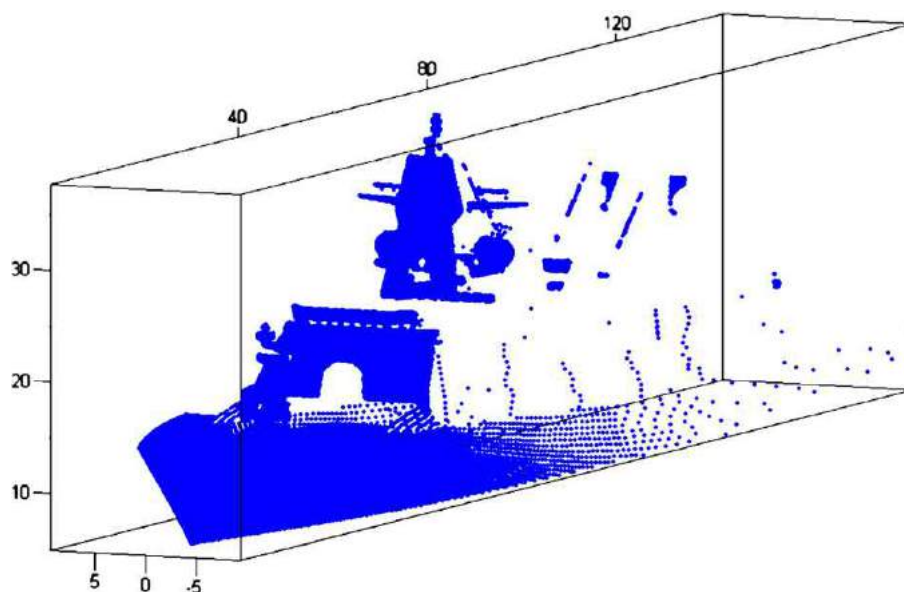


Fig. 7 Plot of data points used to generate a range profile. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

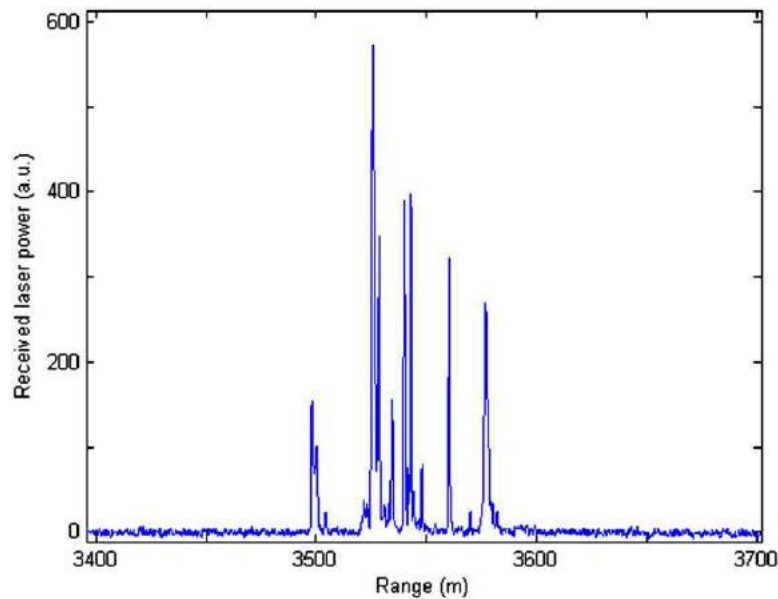


Fig. 8 Laser range profile of a ship. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. *Proc. SPIE* 6950, 69500V.



Fig. 9 Marine target sensing testbed with 1550-nm laser and two receivers. Below is optics for two cameras and Celestron telescope. Reproduced from Steinvall, O., Tulldahl, M., 2017. Laser range profiling for small target recognition. *Opt. Eng.* 56 (3), 031206.

The LRP has the favorable properties of a good range resolution in combination with low sea-surface clutter. These properties have been validated by experiments: a range resolution of 0.6 m has been achieved. Sea-surface reflection was negligible.

Steinvall and Tulldahl (2017) described similar technique for identification of small targets. They used an LRP based on a laser with a pulse width of 6 ns. They showed both experimental and simulated results for laser range profiling of small boats out to a 6–7-km range and a unmanned aerial vehicle (UAV) mockup at close range (1.3 km). The naval experiments took place in the Baltic Sea using many other active and passive electro-optical sensors in addition to the profiling system (Fig. 9). The UAV experiments showed the need for a high-range resolution, thus a photon counting system in addition to the more conventional profiler was used in the naval experiments. The typical resolution of 0.7 m obtainable with a conventional range finder type of sensor can be used for target classification with a depth structure over 5–10 m or more, but for smaller targets such as an UAV a high resolution (in the described case 7.5 mm) is needed to reveal depth structures and surface shapes. Fig. 10 shows an example of range profile from the boat at 0-deg course. The large peak is attributed to the fore and the next to the front of the cabin at a 2.5-m distance from the fore. The extent of the pulse shows a boat length exceeding 7 m.

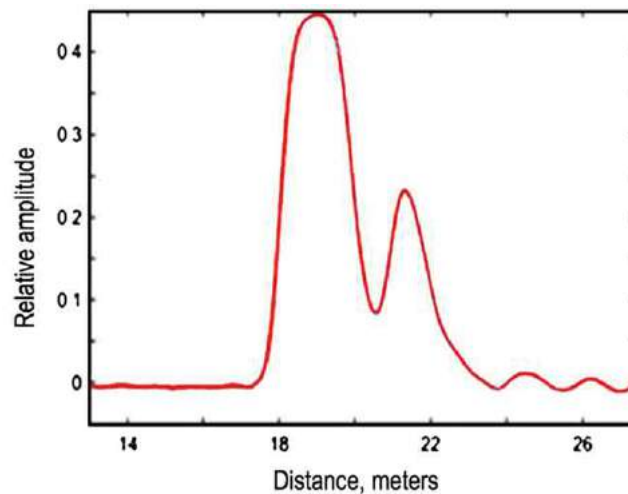


Fig. 10 The waveform referred as the pixel number when sweeping along the line of sight. Reproduced from Steinvall, O., Tulldahl, M., 2017. Laser range profiling for small target recognition. *Opt. Eng.* 56 (3), 031206.



Fig. 11 Subsonic advance cruise missile AGM 129A. Reproduced from Directory of U.S. Military Rockets and Missiles. Raytheon (General Dynamics) AGM-129 ACM. Available at: <http://www.designation-systems.net/dusrm/m-129.html>.

Adding the Velocity as a Dimension

Precision navigation to the designated target or to any other landing site on Earth or extraterrestrial objects, rendezvous and docking with orbiting spacecraft require more information than range and/or altitude. The solutions for navigation can be flying on a preprogrammed route, via image recognition of a series of radar photographs of the terrain or surface altitude profile. Early variants of the Tomahawk (Brigety, 2007) are equipped with an inertial navigation system (INS), Terrain Contour Mapping (TERCOM), and the Digital Scene Matching Area Correlator (DSMAC). TERCOM uses a highly sensitive radar altimeter embedded in the missile to identify surface features by their height. In the terminal phase, the weapon requires even more accurate guidance than TERCOM can provide. For this reason, Tomahawk relies on DSMAC which uses an optical sensor to scan the ground over which the missile is flying. The on-board computer compares the acquired image to the stored one and adjusts the missile's course so it is following the preplanned route. The missile uses also inertial navigation to update the navigation points.

A modified approach was implemented in the subsonic advance cruise missile (ACM) AGM 129A (Fig. 11). Its external shape was optimized for low observable characteristics. To provide a low detectability, the microwave radar was excluded from the set of sensors. For guidance, the AGM 129A uses an inertial navigation system together with a TERCOM enhanced by highly accurate speed updates from laser Doppler velocimeter. The accuracy is quoted between 30 and 90 m.

For the extraterrestrial applications, a Doppler lidar was developed by NASA under the Autonomous Landing and Hazard Avoidance Technology (ALHAT) project (Pierrottet *et al.*, 2009; Amzajerdian *et al.*, 2012). The lidar precision vector velocity data enable the navigation system to continuously update the vehicle trajectory toward the landing site. The ALHAT project included

tests on the Morpheus lander (Fig. 12) developed not to fly in space but to demonstrate technologies that could support future human or robotic exploration (Morpheus lander).

The prototype of Doppler lidar in ALHAT project measures three components of the velocity of which the velocity vector is calculated. A fiber laser is used to generate a very narrow line width, which is frequency modulated and amplified. The output is split into three components in order to distribute the power to three optical channels corresponding to the velocity vector components (Fig. 13).

Using the laser waveform modulated linearly with time the lidar obtains high-resolution range and velocity information. Fig. 14 shows the waveform's frequency content versus time, and the resulting intermediate frequency that contains the range and velocity information. The lidar design uses an optical homodyne receiver configuration, in which a portion of the transmitted beam serves as the reference local oscillator for the optical receiver. The prototype was preliminary tested on the Eurocopter AS350D helicopter. Its optical head is mounted inside a gimbal spherical shroud, pointing in the nadir position while collecting the data.

Interference of oppositely chirped pulses was exploited for simultaneous, range and Doppler velocity measurements (Henderson *et al.*, 1993). In this instrument, echo signals are Doppler shifted in frequency and are also delayed in time relative to



Fig. 12 Morpheus lander in the last moments of landing with ALHAT laser radar instruments. Reproduced from Morpheus lander. Available at: <https://morpheuslander.jsc.nasa.gov>.

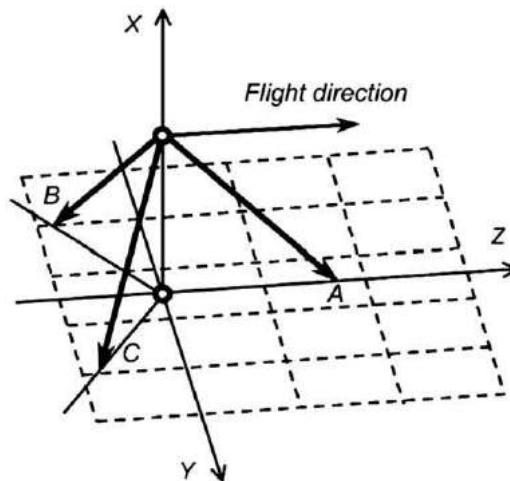


Fig. 13 Unit vectors describing the sensor geometry. Reproduced from Pierrottet, D., Amzajerdian, F., Petway, L., *et al.*, 2009. Flight test performance of a high precision navigation Doppler lidar. Proc. SPIE 7323, 732311.

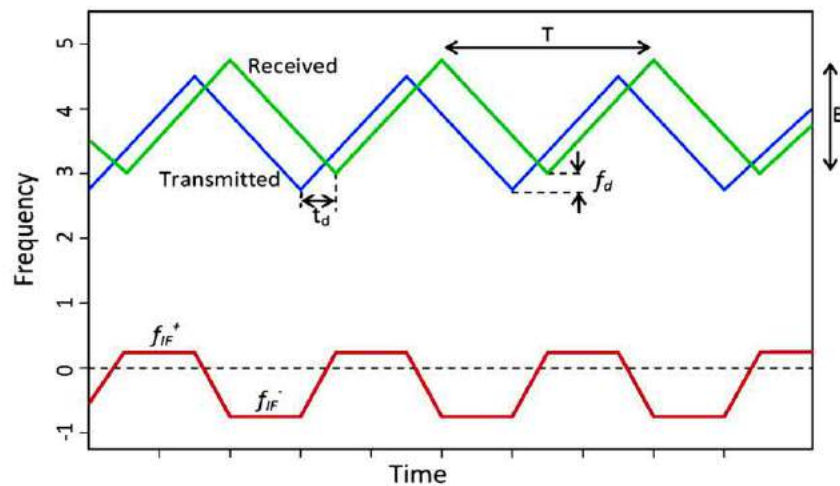


Fig. 14 Linearly modulated laser frequency (top). The difference of transmitted and received signals (lower trace) contains both range and velocity information. Reproduced from Pierrottet, D., Amzajerdian, F., Petway, L., *et al.*, 2009. Flight test performance of a high precision navigation Doppler lidar. *Proc. SPIE* 7323, 732311.



Fig. 15 SAVIS laser Doppler vibrometer. Reproduced from Lutzmann, P., Göhler, B., Putten, F., Hill, C.A., 2011. Laser vibration sensing: Overview and applications. *Proc. SPIE* 8186, 818602.

the reference pulse trains. This results in the generation of up and down beat tones. Sub millimeter resolution ranging was performed and simultaneous, range and Doppler velocity measurements were experimentally demonstrated using a target moving faster than 330 km/h.

Vibrations

The detection of vibrations at long range is of interest in a number of applications: different kinds of operating mechanisms (civil and military), civil infrastructure (bridges, dams, etc.), security (detection of acoustic signals induced in objects nearby), medicine, and so on.

Laser vibrometers typically use the Doppler effect and the coherent detection at any optical wavelength (1.5 μm , 2 μm , 10.6 μm). To get spatially resolved vibrational information, scanning and even multi-beam lasers are used. The review papers of Lutzmann *et al.* (2011), (2017) are focused on works at Fraunhofer IOSB target classification and identification, including camouflaged or partly concealed targets, detection of buried land mines, remote sensing of human breathing, some other applications. Authors used an in-house-developed coherent laser radar (Fig. 15). The sensor is all-fiber-based and operates at a wavelength of 1.5 μm . With an output power of almost 1 W, a spectral line width smaller than 1 kHz and a field of view of 75 μrad , this system is optimized for performing long-range measurements at distances of up to several kilometers.

One of the applications of the developed vibrometer was monitoring the wind turbines. Turbines tend to vibrate strongly, and laser Doppler vibrometry can usefully supplement the existing on-board sensors. The measurements were done on a medium-size wind turbine with a hub height of about 70 m with a rotor diameter of 25 m. The laser Doppler vibrometer was placed about 270 m away from the base point of the wind turbine. The aim was to measure the vibration characteristics of the mast and the nacelle. Fig. 16 shows the wind turbine with indication of points on the mast where the measurements were done and an example of a spectrogram of the signal recorded from a point in the upper point of the mast. Following the time axis one can see the first natural frequency which is about 0.33 Hz. To study the turbine in multiple points simultaneously, a multi-beam vibrometry system would shorten the measuring time.

It is of great interest to the security and military forces, to use remote vibration sensors to detect and assess human state of health or activity. Typical targets are battlefield or earth-quake casualties, and also humans perceived as potential threats. The LDV can give important hints about human activity and health.

The data from 1.5- μm vibrometer were collected from stationary subjects the laser beam being mostly directed on their bare necks at a distance of about 90 m. Fig. 17(a) displays strong pulse beat spikes recorded immediately after a short run with a pulse rate of 110 beats per minute, where the breathing is weakly overlaid by a rate of ~ 43 breaths per minute. The lower part (b) displays the signal sequence a little later, where a deep breathing is dominant (breathing rate: 25 breaths per minute) and the body has partly recovered.

An example (Lutzmann *et al.*, 2000) of the visualized distribution of vibrations of a functioning engine is shown in Fig. 18. Fig. 19 illustrates (Polytec 2006) the vibrations of a car's door (left) and of a turbine's blades (right).

Understanding the characteristics of the vibration of musical instruments is important in studying the acoustic characteristic and improving their manufacture technique. Fig. 20 demonstrates the acoustic vibrations of the violin acquired with the (Polytec 2006) vibrometer using the Doppler principle.

The drumhead has been studied using various methods. Zhang and Su (2005) used the method of Fourier Transform Profilometry (FTP) introduced by Takeda and Motoh (1983). The method is particularly requiring only one image of the deformed fringe pattern to reconstruct the 3D shape of the measured dynamic object. A sinusoidal structured pattern, which is projected onto the surface of a measured drum, is dynamically deformed with the vibration of the membrane. The sequence of the deformed fringe images is grabbed by a high-speed CCD camera, and then a series of wrapped phase maps are produced after the processing with Fourier transform, filtering and inverse Fourier transform. Fig. 21 shows two different phases of drum vibration.

Compared with the vibration pattern measurement by laser Doppler vibrometry, this method has lower accuracy, but has an obvious advantage that all the data of the whole membrane at one sampling instant can be obtained simultaneously, furthermore its information recording time can last along with the whole vibration process to demonstrate the whole vibration from uncreated to fade away.

To get a spatial distribution of vibrations on the surface of the investigated object, a scanning vibrometry is applied. It allows analysis of the structure with a very fine spatial resolution, not modifying its dynamic behavior, decreasing the testing time if a

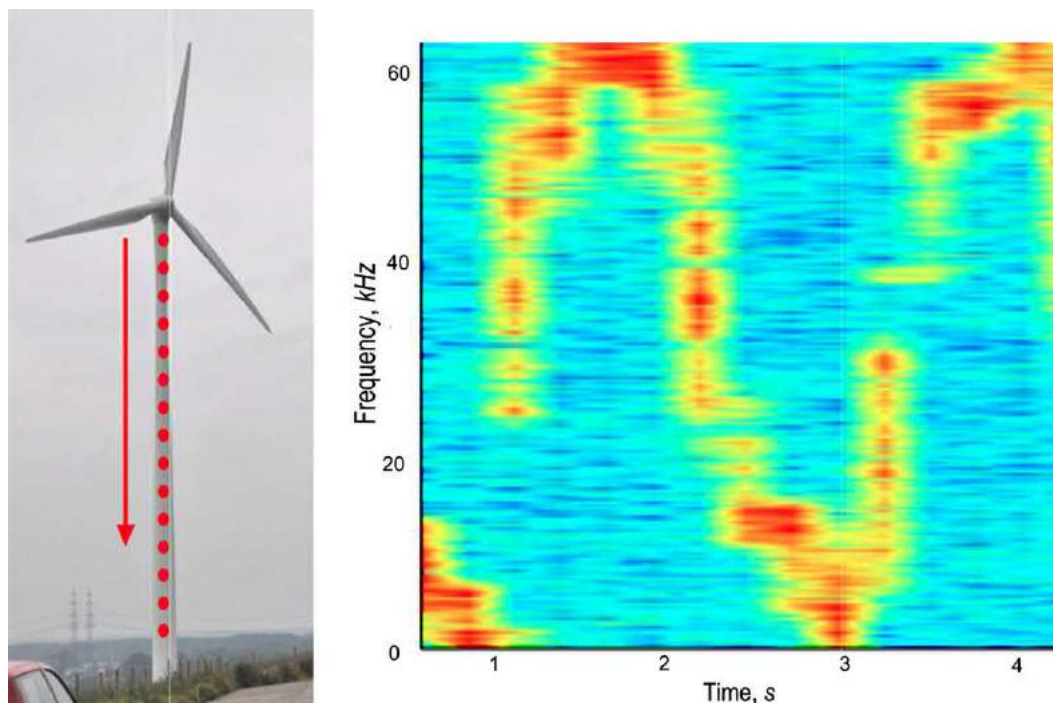


Fig. 16 Measurement of mast modes. Spectrograms of the signal recorded from three selected positions. Reproduced from Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. *Opt. Eng.* 56 (3), 031215.

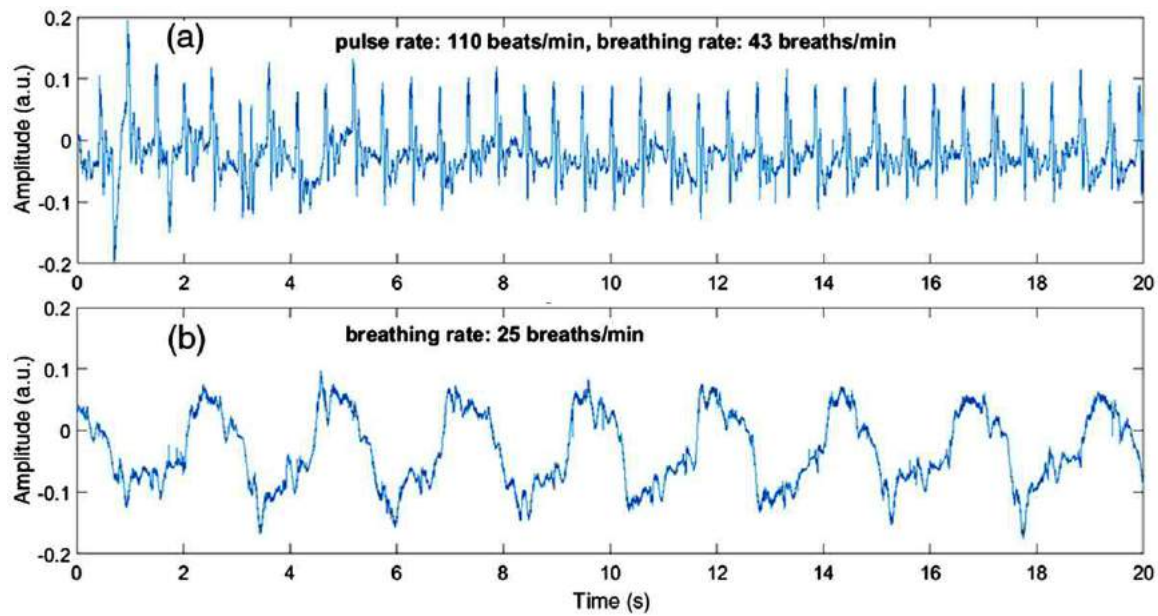


Fig. 17 LDV measurements of pulse beat and breathing. (a) Signal displaying the strong pulse beat spikes immediately after a short run; (b) signal soon after the run, with a deep and slower breathing. Reproduced from Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. *Opt. Eng.* 56 (3), 031215.

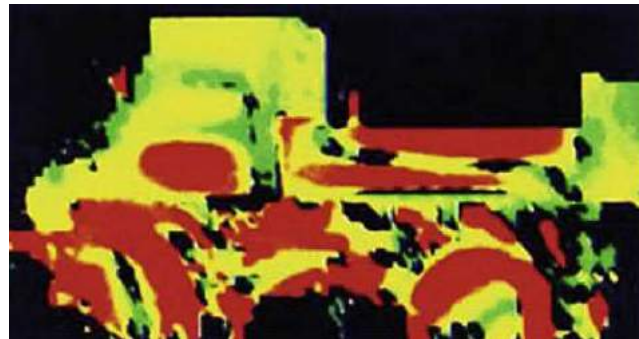


Fig. 18 Vibration intensity map of a functioning engine. Reproduced from Lutzmann, P., Frank, R., Ebert, R., 2000. Laser radar based vibration imaging of remote objects. *Proc. SPIE* 4035, 436–443.

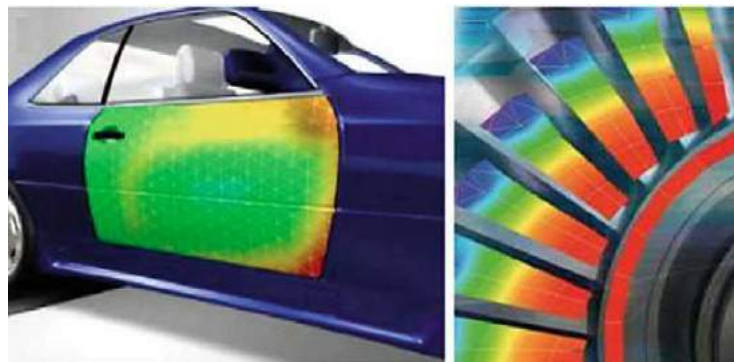


Fig. 19 Vibration intensity maps of a car's door (left) and of a turbine's blades (right). Reproduced from Polytec. Available at: <http://www.polytec.com>.

large number of measurement points is requested, and allowing measurement of vibrations even on hot surfaces. In some applications, laser vibrometry has also been coupled with laser excitation to develop a complete noncontact and nonintrusive modal analysis procedure.

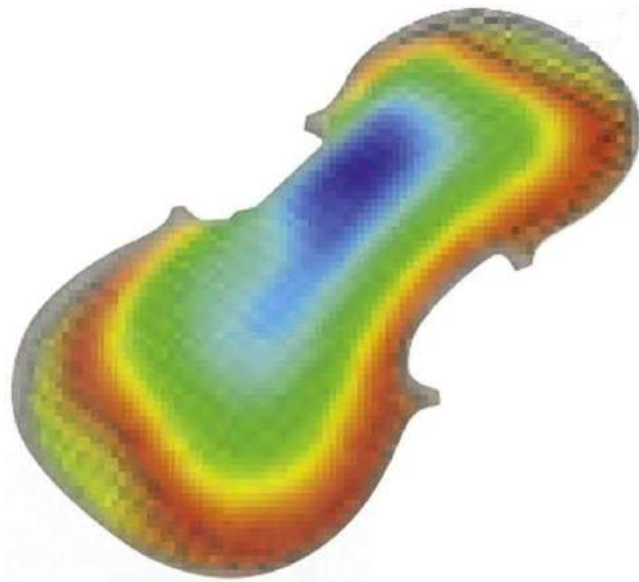


Fig. 20 Acoustic vibrations of a violin acquired with the Polytec vibrometer. Reproduced from Polytec. Available at: <http://www.polytec.com>.

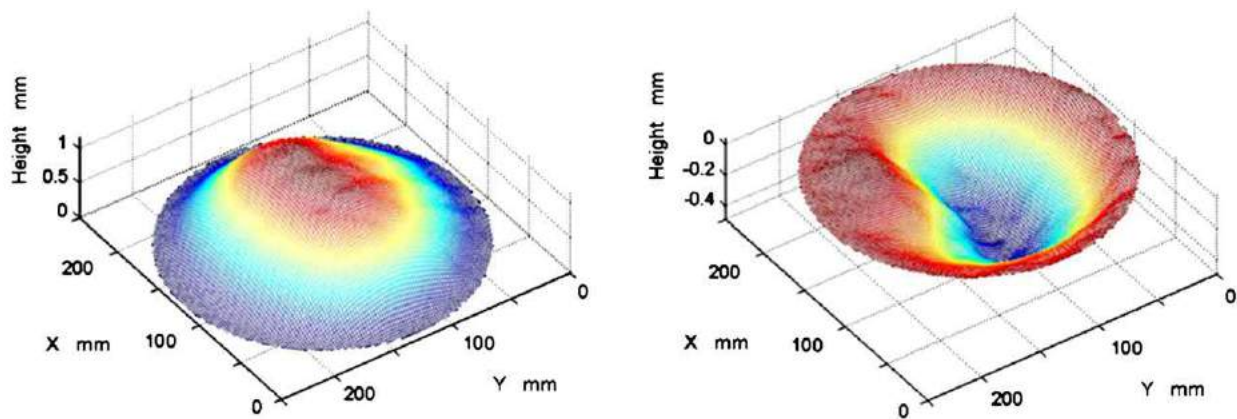


Fig. 21 Different phases of drum vibration. Reproduced from Zhang, Q., Su, X., 2005. High-speed optical measurement for the drumhead vibration. *Opt. Express* 13, 3110–3116.

Revel *et al.* (2011) investigated the problem of noise in the cabin of Agusta A109MKII. They described the use of a scanning laser Doppler vibrometer to measure vibrations inside the helicopter using its mock-up, and demonstrated the applicability of the technique for in-flight tests. The whole area was 430×315 mm. In the scanning tests, a 30×20 grid (corresponding to a spatial resolution of about 14×15 mm) was used. An additive mass was used in order to lower the center of mass of the vibrometry system and to gain stability. The comparison was also made with the vibrograms measured when the vibrometer was placed outside the mock-up. Fig. 22 is a summary of the tests at different resonances. They all refer to instantaneous amplitude of the velocity component at a certain frequency orthogonal to the surface with the bandwidth ± 10 Hz for the frequencies up to 1000 Hz and ± 30 Hz for the frequencies up to 5000 Hz.

Perea and Libbey (2016) demonstrated a laboratory system for measurement of three degrees of vibrational freedom simultaneously (Perea and Libbey, 2016). Heterodyne speckle imaging was exploited in which the optical speckle pattern was mixed with a coherent reference. From these data, three dimensions of surface motion are extracted. Axial velocity is measured by demodulating the received time-varying intensity of high amplitude pixels. Tilt, a gradient of surface displacement, is calculated by measuring speckle translation following extraction of the speckle pattern from the mixed signal.

Media Parameters as Dimensions

The diversity of laser radar applications for media investigation is impressive. Many of them use detection and processing of signals scattered in the medium of propagation, that can be not only the atmosphere of planets, but also the water and any other

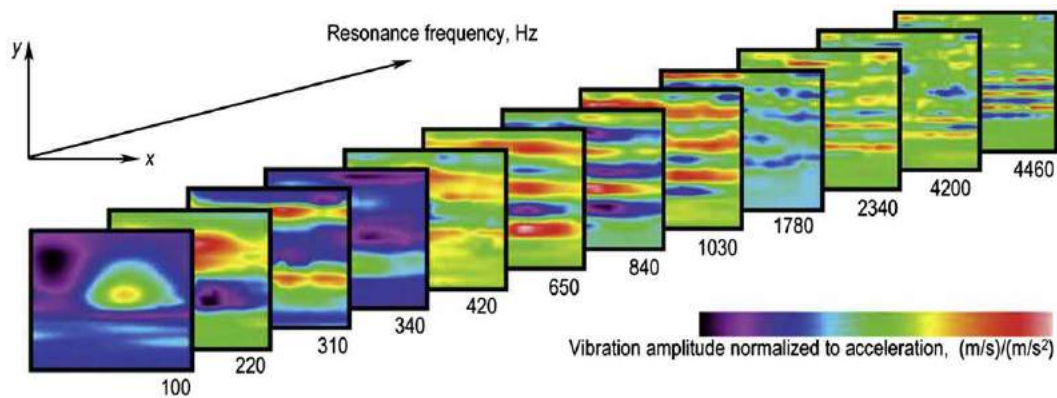


Fig. 22 Resonance vibration frequencies registered in the cabin of the helicopter mock-up. Summary of images reproduced from Revel, G.M., Castellini, P., Chiariotti, P., *et al.*, 2011. Laser vibrometry vibration measurements on vehicle cabins in running conditions: Helicopter mock-up application. Opt. Eng. 50, 101502.

medium of organic or non-organic origin. Examples could be like airborne atmospheric remote sensing (Palm *et al.*, 1994), cloud top remote sensing by airborne lidar (Spinhirne *et al.*, 1982), airborne high spectral resolution lidar that provides measurements of aerosol backscatter and extinction coefficients and aerosol depolarization at two wavelengths (Hair *et al.*, 2008), nanostructure and defect analysis using a simple 3D light-scatter sensor (Herffurth *et al.*, 2013), refraction distribution in the human eye (Molebny, 2013). Even a military rangefinder can be used for atmosphere sensing (Steinvall *et al.*, 2014). There are also many other applications.

Retro-scattering based on interaction of laser light with medium components brings another amount of information. Examples are as follows: Raman lidar for water vapor, temperature, aerosol, and cloud measurement (Reichardt *et al.*, 2012), dual field-of-view Raman lidar measurements for the retrieval of cloud microphysical properties (Schmidt *et al.*, 2013), lidar measurements of the Eyjafjallajökull ash cloud over the Netherlands (Donovan and Apituley, 2013), spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere (Liu and Yi, 2013), using Raman and elastic backscatter for measurement of temperature, density, pressure, moisture, and particle backscatter coefficient (Fraczek *et al.*, 2012).

Philbrick *et al.* (2010) experimentally showed how multi-wavelength backscatter measurements from Rayleigh and Raman lidar techniques can provide signals to be used to profile the properties of the atmospheric column. Rayleigh lidar signals provide backscatter coefficients, and Raman lidar signals backscattered from the major molecular components provide extinction profiles. The ratio of these simultaneous extinction and backscatter measurements are used to classify the aerosol type. In addition, a laser beam can be used to make bistatic and multistatic measurements of the polarization ratio of the scattering phase function. Analysis of multistatic measurements can be used to determine profiles of the aerosol number density, size, size distribution, and type. These parameters can be determined for spherical particles in the size range between about 20 nm and 20 μm . Information on aerosol type and shape can be supposed from determining the approximate refractive index of the scattering aerosols and by measuring the depolarization of the scattered radiation.

In a later publication, Philbrick and Hallen, (2017) discussed the Raman lidar measurements that provide profiles of several different tracers of spatial and temporal variations, which are excellent signatures for studies of dynamical processes in the atmosphere. An examination of Raman lidar data collected during the last four decades clearly showed signatures of atmospheric planetary waves, gravity waves, low-level jets, weather fronts, turbulence from wind shear at surfaces and at the interface of the boundary layer with the free troposphere. Water vapor profiles are important as a tracer of the sources of turbulence eddies associated with thermal convection, pressure waves, and wind shears, which result from surface heating, winds, weather systems, orographic forcing, and regions of reduced atmospheric stability. Examples of these processes were selected to show the influence of turbulence on profiles of atmospheric properties.

Fluorescence is one of physical phenomena successfully used in the today's laser radars for remote detection of oil spills (Sato *et al.*, 1978), for remote sensing of vegetation (Matvienko *et al.*, 2006), for fluorescence imaging of the ocean bottom (Sitter and Gelbart, 2001). High-repetition-rate three-dimensional OH imaging using scanned planar laser-induced fluorescence system allows for multiphase combustion control (Cho *et al.*, 2014). Combinations of several physical phenomena (elastic scattering, fluorescent scattering, and differential absorption) were the basics of the developed airborne lidar for observations of atmospheric tracers (Utke, 1991). Some examples of acquiring the media parameters and data 3D presentation are given hereinafter.

Wind and Flow Velocity

The need for remote sensing of atmospheric winds is well established, be it from ground-based sites, air-, or space-borne equipment. Lidar systems observe large volumes of atmosphere with high spatial and temporal resolution, making these data important for many applications. Doppler lidars can use incoherent or coherent techniques. In the incoherent system, to isolate

the laser line from the day light influence, a Fabry-Pérot étalons (interference narrow-band filters) are used. The laser is typically Nd-doubled 532 nm, 3 W average power (Fischer *et al.*, 1995). The field of view of the receiving telescope is of the order of 0.5 mrad that is somewhat larger than the laser divergence of 0.2 mrad to collect all of the laser light, which may be outside the divergence angle because of different factors like pointing jitter, scanning instability, or similar. The collected light is focused by the telescope onto an optical fiber with the diameter of 3.5 mm. The spatial scrambling of the collected light is necessary to remove the effects of inhomogeneity in the scattering aerosol volumes and of changes in normalization degree in detectors, and to avoid systematic offset errors in measured wind velocity. Fig. 23 shows a time series of the horizontal wind field at the altitude of 100 m. The velocity vectors are scaled as shown in the diagram. Complete profiles were generated every 5 min.

Low altitude wind profile measurements with a laser rangefinder were demonstrated using a simple balloon tracking system, with small (0.25 m diameter) lightweight balloons (Wilkerson *et al.*, 2009). Experiments on balloon trajectories demonstrate that laser range detection (± 0.5 m) combined with azimuth and elevation measurements is a simple, accurate, and inexpensive alternative to other wind profiling methods. To increase the maximum detection range to 2200 m, a retroreflector tape was attached to the balloons. Nighttime tracking was facilitated by low power LEDs.

Wind speed and direction results are compared with simultaneous sodar measurements (Fig. 24). The profiling resolution is greatly improved using a laser rangefinder with automatic coordinate and time recording. However, balloon tracking is still man-in-the-loop. Improvements should include automation of the tracking system to collect trajectory points automatically at 1 Hz or faster.

Windshear poses the greatest danger to aircraft during takeoff and landing due to its abrupt change of direction (Fig. 25) (Targ *et al.*, 1991), when the plane is close to the ground and has little time or room to maneuver (Molebny and Steinvall, 2013) (Fig. 26). During landing, the pilot has already reduced engine power and may not have time to increase speed enough to escape the downdraft. During takeoff, an aircraft is near stall speed and thus is very vulnerable to windshear. Microburst windshear often occurs during thunderstorms. But it can also arise in the absence of rain near the ground. Pilots need 10–40 s of warning to avoid windshear.

The first airborne measurements of winds used a pulsed CO₂ laser (Bilbro and Vaughan, 1978; Bilbro, 1984). The lidar was oriented in the fore and aft directions in order to obtain the horizontal vector wind field. Another coherent lidar system Coherent Lidar Airborne Shear Sensor (CLASS) was developed in two versions: with a 10.6 μm CO₂ laser (CLASS-10) and with a solid state 2.02 μm Tm: YAG laser (CLASS-2). Both lidars showed a wind measurement accuracy better than 1 m/s (Targ *et al.*, 1996). Coherent airborne wind lidar system Wind INfrared Doppler lidar (WIND) was developed later in French-German cooperation (Werner *et al.*, 2001).

NASA's Langley Research Center tested three airborne predictive windshear sensor systems (NASA, 2017): microwave radar detecting the raindrop scattering, Doppler laser radar detecting the aerosol scattering, and infrared detector measuring the temperature changes ahead of the airplane. The system was manufactured by Lockheed Corp.'s Missiles and Space Co.; United Technologies Optical Systems Inc., and Lassen Research.

In the wind infrared Doppler lidar (WIND) project, airborne conically scanning CO₂ Doppler lidar was used. Spatial resolution requirements were 250 m in height with a grid size of 10×10 km. The radiation of the local oscillator is mixed with outgoing pulse and with the Doppler-shifted backscattered signal. A locking loop connects both lasers. The laser pulse with a pulse duration

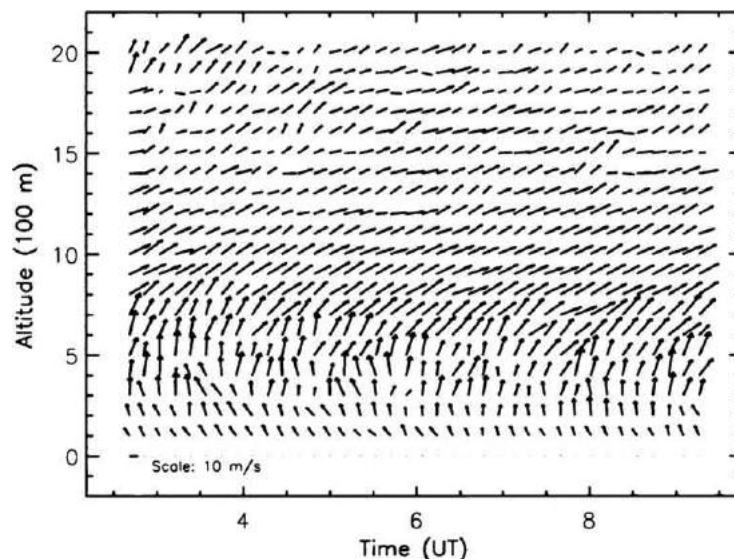


Fig. 23 Examples of displaying the information on wind measurement in the form of vectors vs time, right – vertical distribution in comparison with sonar data. Reproduced from Fischer, K.W., Abreu, V.J., Skinner, W.R., *et al.*, 1995. Visible wavelength Doppler lidar for measurement of wind and aerosol profiles during day and night. Opt. Eng. 34 (2), 499–511.

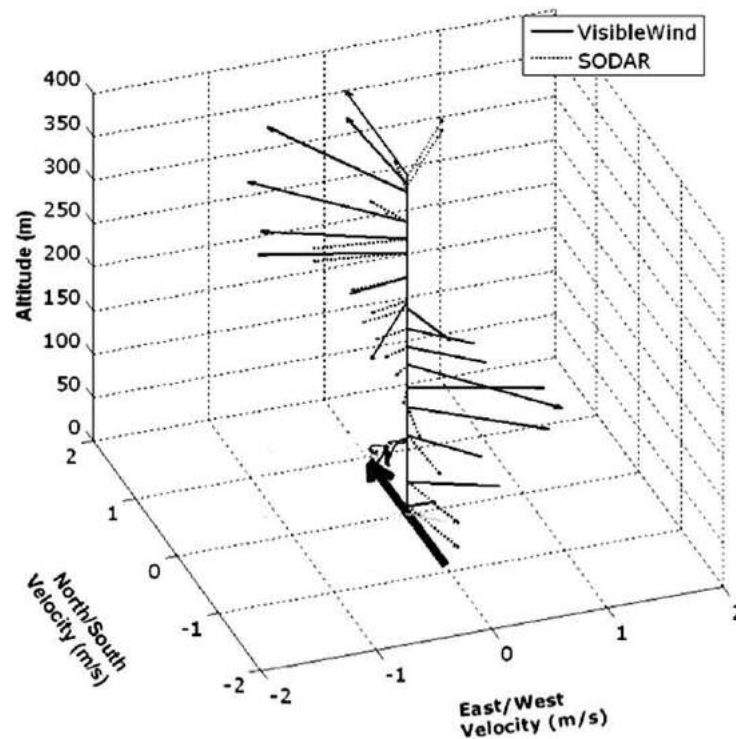


Fig. 24 Wind changing vertically in comparison with sonar data. Reproduced from Wilkerson, T., Bradford, B., Marchant, A., *et al.*, 2009. VisibleWindTM: A rapid-response system for high-resolution wind profiling. Proc. SPIE 7460, 746009.

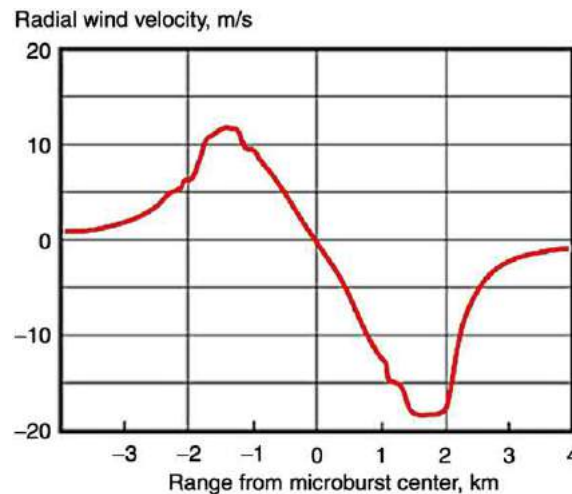


Fig. 25 Velocity distribution within the windshear. Reproduced from Targ, R., Kavaya, M.J., Huffaker, R.M., Bowles, R.L., 1991. Coherent lidar airborne windshear sensor: Performance evaluation. Appl. Opt. 30, 2013–2026.

of 2.5 μs and 100–300 mW of per-pulse power is sent out via the transceiver telescope into the region of investigation with a repetition frequency 10 Hz. The wind-shifted Doppler frequency directly determines the line-of-sight component of the wind vector. At CO_2 laser wavelength 10.6 μm , a velocity component of 1 m/s corresponds to a frequency shift of 189 kHz. Ground-based tests, airborne tests and a validation flight were performed. An example of polar wind map is shown in Fig. 27.

Lockheed Martin Coherent Technologies developed a system called WindTracer (Lockheed Martin, 2015), whose specialty is remote sensing of winds in the critical 40–200 m height regime. With no sidelobes, WindTracer can scan the space near the ground, even adjacent to obstacles. It provides a 30–60 min warning of changes in the winds approaching a wind farm. This information is important for grid electric power operators and for turbine operators to optimize the operations. The wavelength is the eye-safe 1.6 μm , the maximum operating range is up to 12 km, range resolution is about 50 m, the coverage area

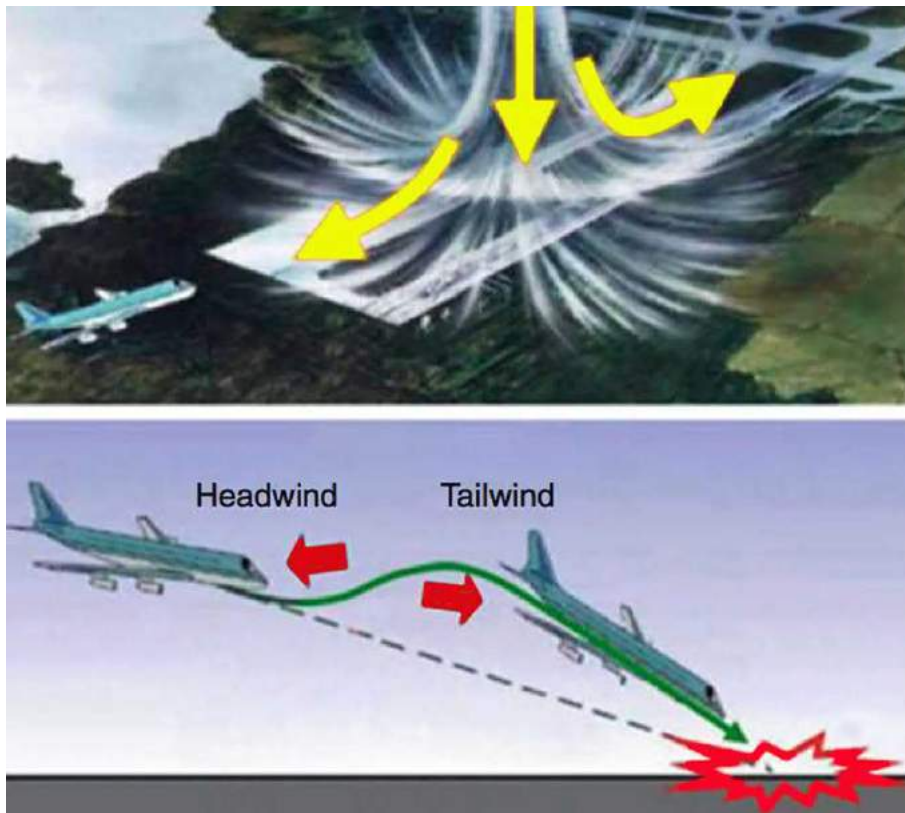


Fig. 26 Windshear and its effect. Reproduced from Molebny, V., Steinvall, O., 2013. Laser remote sensing. Velocimetry based techniques. In: Tuchin, V.V. (Ed.), *Handbook of Coherent-Domain Optical Methods*. New York: Springer, pp. 363–395.

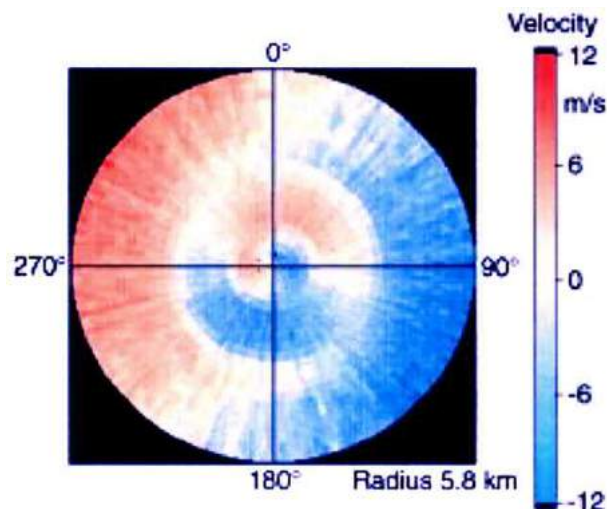


Fig. 27 Polar plot intensities of the wind fields. Reproduced from Werner, C., Flamant, P.H., Reitebuch, O., *et al.*, 2001. Wind infrared Doppler lidar instrument. *Opt. Eng.* 40 (1), 115–125.

is 100–200 km², and a typical vector velocity accuracy is about 1 m/s. The combination of lidar and radar, called WindTracer Terminal Doppler Solution (WTDS) enables wind hazard detection under wider range of weather conditions.

A general view of WindTracer is demonstrated in [Fig. 28](#). The screenshot from its display ([Fig. 29](#)) shows the radial velocity estimates over a 12-km-diameter circular area.

European companies [Leosphere \(2004\)](#) and [Halo Photonics](#) specialize in building compact Doppler lidars with emphasis on wind farming applications. Leosphere developed several models of Windcube (100S/200S/400S) of the Doppler lidar. Radial



Fig. 28 General view of WindTracer. Reproduced from Lockheed Martin. Available at: <http://www.lockheedmartin.com/products/WindTracer/index.html>.

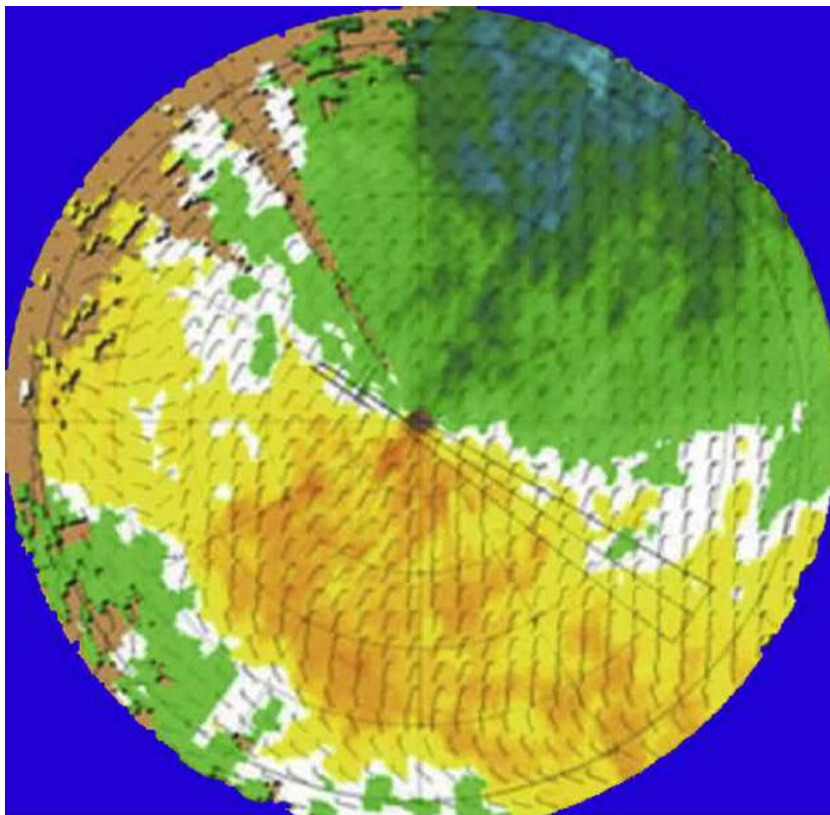


Fig. 29 WindTracer's radial velocity estimates. Reproduced from Lockheed Martin. Available at: <http://www.lockheedmartin.com/products/WindTracer/index.html>.

velocity display from 200S model overlaid on the land map is shown in Fig. 30. An example of velocity-height dependency on time acquired with Halo Photonics Doppler velocimeter is demonstrated in Fig. 31. As the wind energy industry continues to grow, the need for more accurate and sophisticated wind data grows as well. Such lidars are used to explore the locations like ridgelines, oceans, lakes, and forested areas. Typically, the lidars provide 100–200 m vertical wind profiles with an accuracy down to 0.1 m/s. Maximum range of the Halo Photonics streamline instrument using 1500 nm laser wavelength is 9.0 km, range resolution 15 m, maximum measured velocity 19 m/s, velocity resolution 0.0384 m/s.

The New European Wind Atlas (NEWA) project has a goal to create a freely accessible wind atlas covering Europe and Turkey. Common to all experiments is the use of Doppler lidar systems. Many European and American companies and universities will participate in the project. Two pilot experiments, one in Portugal and one in Germany, showed the value of using multiple synchronized, scanning lidars (Mann *et al.*, 2017). Example of the scans at three different elevation angles are shown in Fig. 32 received from the WindScanner developed at Technical University of Denmark for the purposes of NEWA, within the

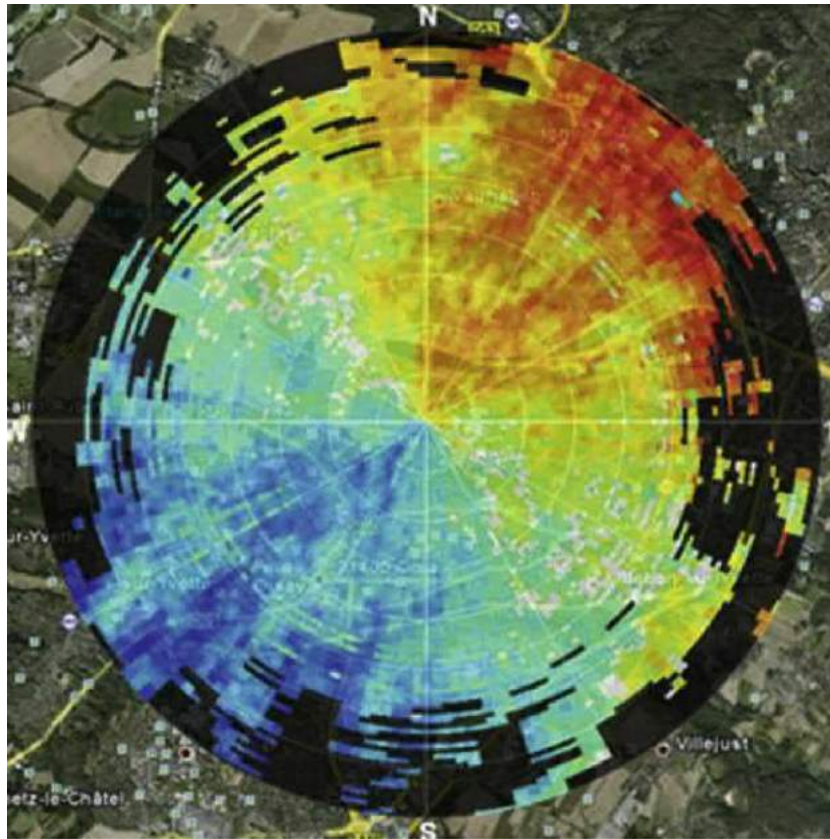


Fig. 30 Radial velocity display from Windcube 200S. Reproduced from Leosphere. Available at: <http://www.leosphere.com/8/wind-energy>.

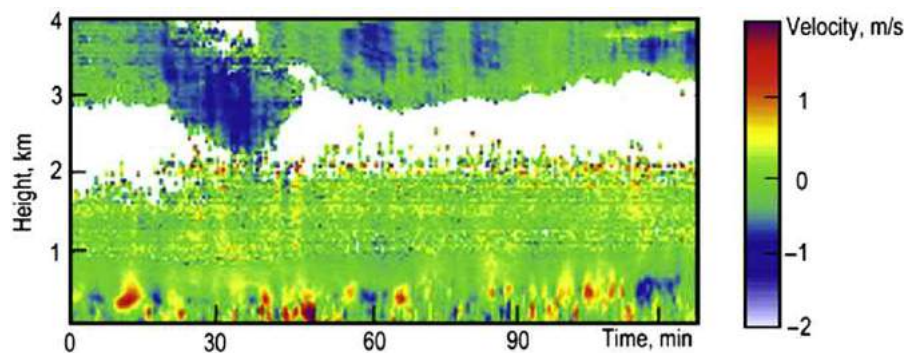


Fig. 31 Velocity (color-coded)-height vs. time acquired with Halo Photonics Doppler velocimeter. Reproduced from Halo Photonics. Available at: <http://halo-photonics.com/index.htm>.

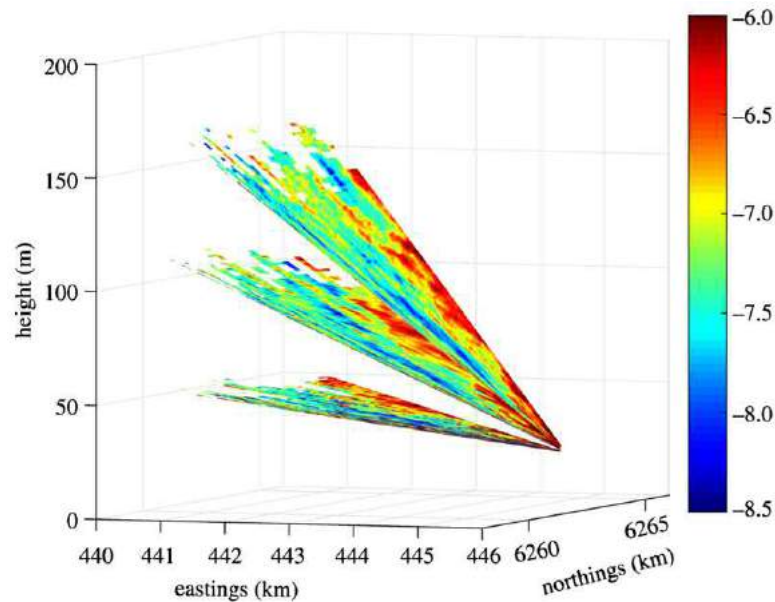


Fig. 32 Wind scans at three different elevation angles. Reproduced from Mann, J., Angelou, N., Arnqvist, J., *et al.*, 2017. Complex terrain experiments in the New European Wind Atlas. *Phil. Trans. R. Soc. A* 375, 20160101.

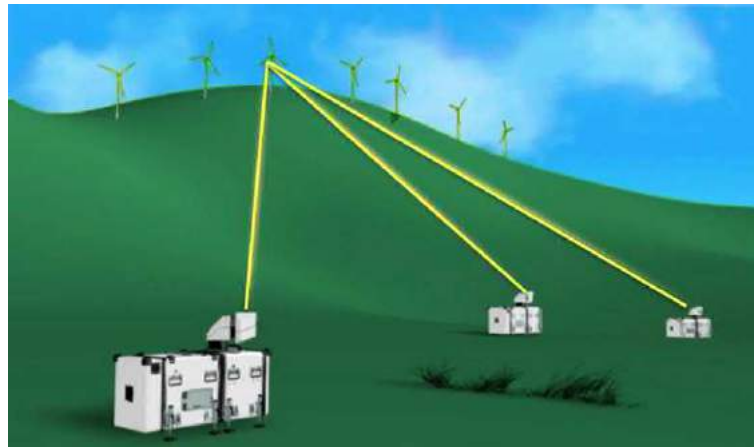


Fig. 33 Three simultaneously operating wind lidars equipped with synchronized steerable-beam scanners. Reproduced from Mikkelsen, T., 2014. Lidar-based research and innovation at DTU wind energy – A review. *J. Phys.: Conf. Ser.* 524, 012 007.

WindScanner.dk project in long- and short-range versions (Mikkelsen, 2014) with the following main parameters (Vasiljevic *et al.*, 2017). Long-range version: pulse mode, range 50–8000 m, 500 range gates, probe length is 25, 35 or 70 m (fixed with range). Short-range version: CW, range 10–50 m, a single range gate, probe length is 0.2–40 m (evolving with range). These instruments confirmed the potentials not only to economically measure the wind field at a single point located within the rotor swept area, but also to map entire wind fields within the volume occupied by modern wind turbine and wind farms.

By combining space and time synchronized measurements along the line of sight from three simultaneously operated wind lidars equipped with synchronized steerable-beam scanners, true 3D wind velocity and turbulence measurements have become attainable for wind energy industry applications (Fig. 33).

WindScanner system consists practically of two or more spatially separated scanning lidars (long- or short-range WindScanners), that are coordinated or controlled by a master computer. The first generation of the long-range WindScanner (LRWS) and short-range WindScanner (SRWS) originate from commercially available vertical profiling lidars Windcube 200 (Werner *et al.*, 2001) and ZephIR, respectively. These profiling lidars were converted into the scanning lidars. The WindScanners have been specifically tailored to perform user-defined and time-controlled scanning trajectories, either independently or in a synchronized mode. A single LRWS can acquire radial velocities simultaneously at a maximum 500 different distances along each line-of-sight at a maximum rate of 10 Hz, whereas a single SRWS can measure radial velocity from one single range at a time but with a maximum rate of 400 Hz.

The lidars in the long-range WindScanner system are usually connected to the master computer using a 3G network. The short-range WindScanner system is formed by connecting the master computer with the lidars via a 300-meter long optic fiber.

The long-range WindScanner system is intended for measurements of mean flow field within a large volume of the atmosphere, while the short-range WindScanner system is typically applied to perform small-scale measurements of turbulent flows around a single wind turbine rotor.

The NEWA project aims to improve wind resource modeling for different site conditions. Areas with steep ridges and forested terrain are of particular interests, since the current engineering (linear) flow models are unable to correctly predict behavior of the flow over the sites with these features. The accompanying projects were focused on optimization of wind generators. To fulfill the project's ambition, it was necessary to investigate how the incoming flow is modified by a wind turbine. Still another objective was to develop numerical tools for wind farm layout optimization in complex terrain.

The site for the tests was chosen in Perdigão, Portugal with the opportunity to measure wind resources along a ridge, occurrences of flow separations on lee sides of hills (i.e., recirculation zone) and valley flows. According to the scientific objectives, the site contains a steep ridge where an isolated wind turbine is operating (Fig. 34, top photo).

A quasi two-dimensional, or a long ridge hill was a logical choice. Also, to assure a two-dimensional flow, dominant winds should be perpendicular to the ridge. Land cover, particularly forests would add to the flow complexity, and is considered desirable, since many wind farms are installed near or within forested regions. The presence of a wind turbine at the site gave the opportunity for wake and inflow measurements.

The two WindScanner systems, comprised a total of six scanning lidars: three LRWSs and three SRWSs. The lidar locations were selected according to the aim to sample the flow field along the South ridge and within the transect perpendicular to the ridges entailing the wind turbine.

The simulation of wind flow was done using steady state formulation for the equations, and the model's wall boundary parameters were calibrated to yield wind speeds of 5–6 m/s and other parameters were taken into account with suggested values. The simulations of the flow from the Northeast and Southwest predict large recirculation zone enclosed in the valley (Fig. 34) and a high complexity of the flow (Fig. 35).

Experimental studies should follow to confirm the simulation with the correctives shown up in the studies. It will be the six-dimensional (6D) maps of wind space, three dimensions of which will be the (x, y, z) coordinates of the space and other three dimensions (V_x, V_y, V_z) will be the velocity components of the wind vectors within that space. If these 6D maps would be supplemented with the temperature in each (x, y, z) point, the information would become seven-dimensional (7D: $x, y, z, V_x, V_y, V_z, T$). The humidity can be the following candidate for the eight-dimensional (8D: $x, y, z, V_x, V_y, V_z, T, h$) space. Reconstructing this 8D space of information as a time sequence, one should get the nine-dimensional (9D: $x, y, z, V_x, V_y, V_z, T, h, t$) space of information. The list of dimension candidates can be changed or extended in correspondence with the tasks of the study.

Non-Doppler, ground-based, scanning aerosol backscatter lidars can measure the wind speed and direction by cross correlation of coherent aerosol structures (Eloranta *et al.*, 1975). The primary assumptions are that the inhomogeneities remain coherent, and that these spatially-coherent turbulent features persist for a time that is long compared to the acquisition time of the lidar system during the advective process.

Non-Doppler incoherent lidars do not require coherent detection, which removes limitations of telescope diameter and relaxes requirements for high-spectral-purity lasers. This lack of restriction becomes more attractive as large lightweight mirrors and small powerful lasers become more affordable and commercially available. In addition to the wind field, these measurements can also be used to compute mean dimensions and lifetimes of turbulence. Unfortunately, the cross-correlation technique does not measure the wind velocity in all conditions. Coherent structures may not be advected with the wind in the scan plane or scan volume. Three-dimensional scan strategies are able to eliminate this possibility but are still susceptible to false velocities from wavelike motions. It is fortunate that irregularities in the aerosol scattering field within turbulent regions are usually oriented in all directions, and, by averaging in space and time, the motion of the coherent structures advecting with the wind dominates the correlation functions.

Based on the aerosol backscatter correlation, a three-beam lidar was developed (Prasad and Mylapore, 2017) to measure wind characteristics for wake vortex and plume tracking applications. This is a direct detection elastic lidar that uses three laser transceivers (Fig. 36), operating at 1030-nm wavelength with ~ 10 -kHz pulse repetition frequency and nanosecond class pulse widths, to directly obtain three components of wind velocities. By tracking the motion of aerosol structures along and between three near-parallel laser beams, three-component wind speed profiles along the field-of-view of laser beams can be obtained. With three 8-in. transceiver modules, placed in a near-parallel configuration on a two-axis pan-tilt scanner (Fig. 37), the lidar measures wind speeds up to 2 km away. Aerosol density fluctuations are cross-correlated between successive scans to obtain the displacements of the aerosol features along the three axes. Using the range resolved elastic backscatter data from each laser beam, which is scanned over the volume of interest, a three-dimensional map of aerosol density can be generated.

Atmosphere Components and Contaminants

3D structure of the water vapor field was observed using a scanning water vapor differential absorption lidar (DIAL) (Behrendt *et al.*, 2009). The instrument is mobile and was applied successfully in two field campaigns. The data products of the DIAL are profiles of absolute humidity with typical range resolutions of 15–300 m and temporal resolution of 1–10 s. Maximum range is several kilometers at both day and night. Spatial and temporal resolution can be traded off against each other.

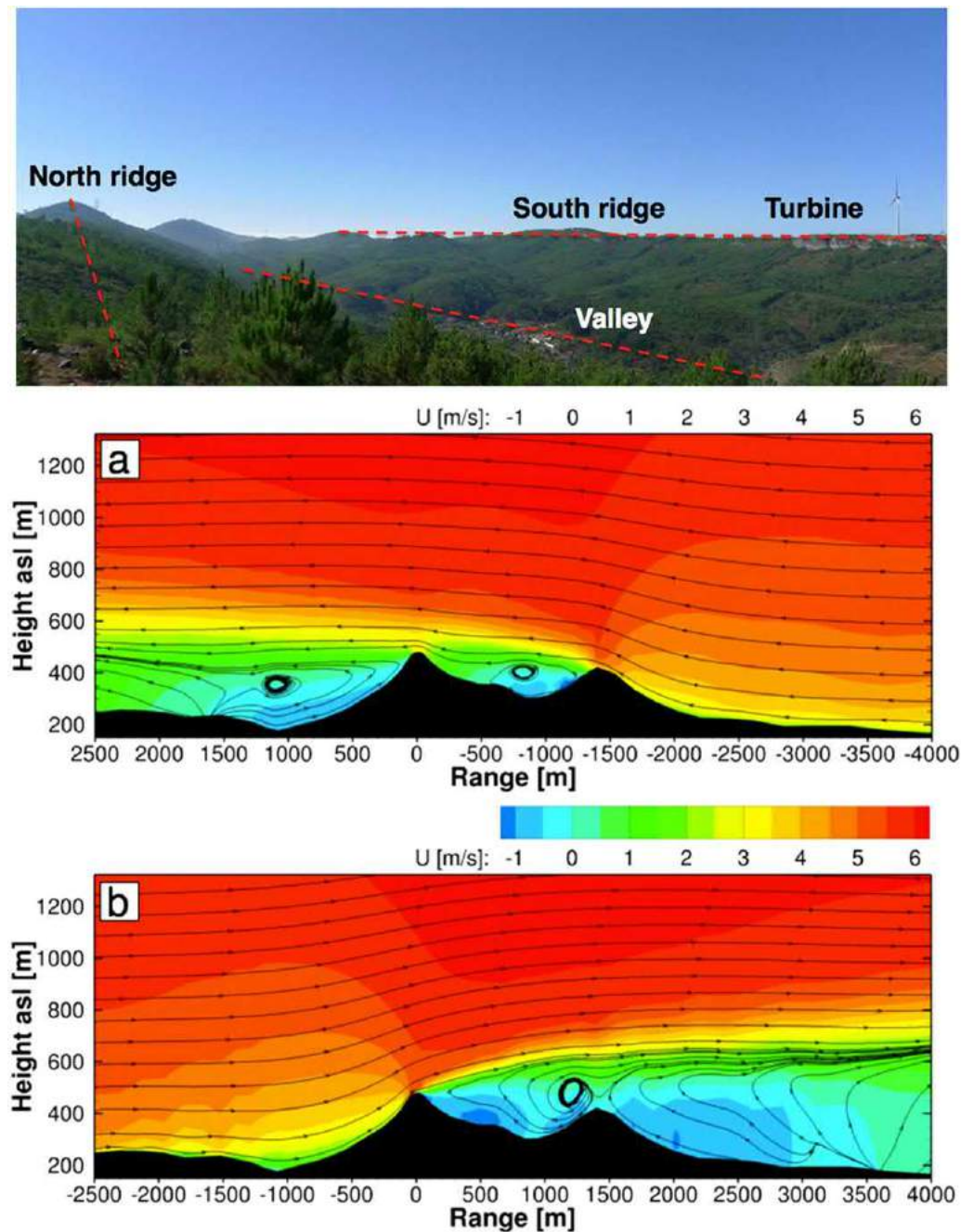


Fig. 34 Profile of the test site (top). Wind flow in a vertical plane: (a) Northeast winds; (b) Southwest winds. Coordinate system origin corresponds to wind turbine base. Reproduced from Vasiljevic, N., Palma, J., Angelou, N., *et al.*, 2017. Perdigão 2015: methodology for atmospheric multi-Doppler lidar experiments. *Atmos. Meas. Tech. Discuss.* doi:10.5194/amt-2017-18.

Beside humidity, the backscatter field and thus aerosols and clouds can be observed simultaneously. The DIAL transmitter is based on an injection-seeded Ti:Sapphire laser operated at 820 nm which is pumped with a Nd:YAG laser.

A measurement example in vertical pointing mode during the first field deployment of the instrument is shown in Fig. 38. Even with a resolution of only 15 m and 1 s, the noise of the data is low; the high variability of the water vapor field in this case is revealed. Comparisons of the data acquired with the instrument and with the radiosondes launched at the lidar site showed a close agreement.

Range-height-intensity scanning measurement examples are presented in Fig. 39. The complexity of the humidity field is striking. Several horizontal layers can be seen in addition to turbulent structures with high moisture value from close to the ground up to a height of about 300 m.

To measure gaseous and liquid water in the atmosphere, a spectrally resolved Raman lidar based on a tripled Nd:YAG laser (100 mJ per pulse at 354.7 nm with a repetition rate of 20 Hz and a pulse width of 6 ns) was developed at the Wuhan University

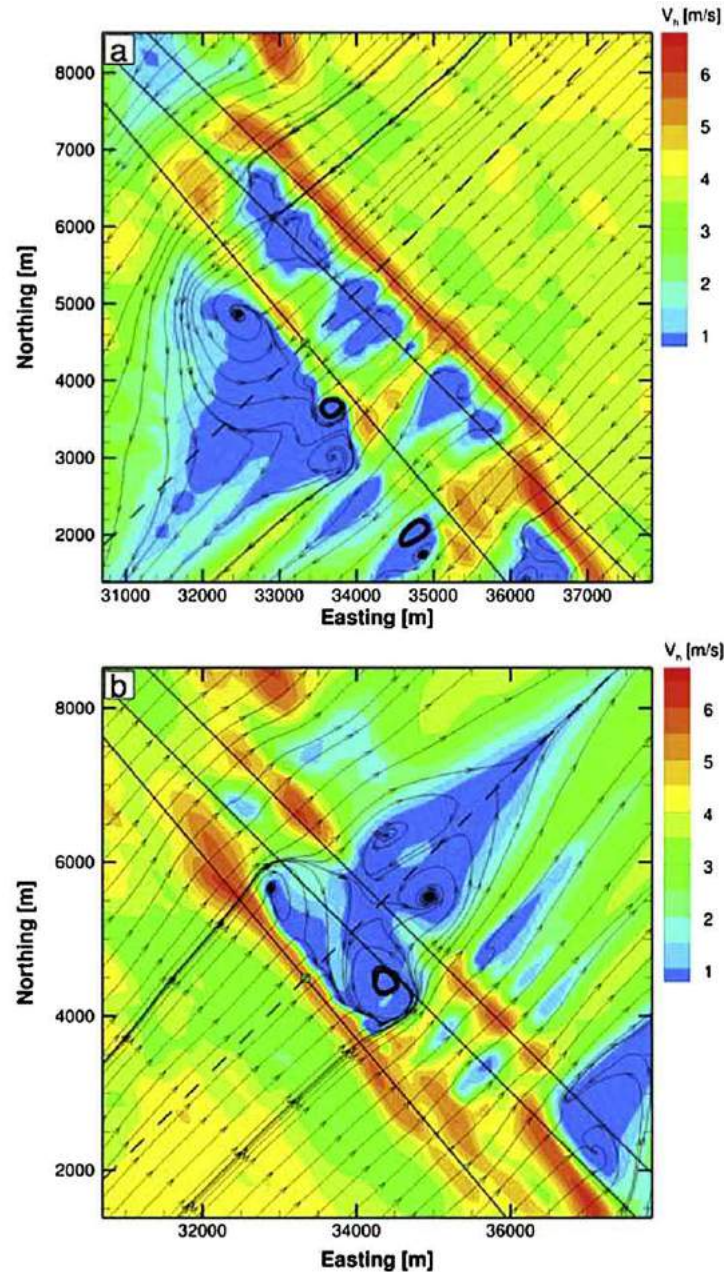


Fig. 35 Wind flow in surface 80 m agl: (a) Northeast winds; (b) Southwest winds. Reproduced from Vasiljevic, N., Palma, J., Angelou, N., *et al.*, 2017. Perdigo 2015: Methodology for atmospheric multi-Doppler lidar experiments. Atmos. Meas. Tech. Discuss. doi:10.5194/amt-2017-18.

(Liu and Yi, 2013). At the receiver side, the backscattered photons are collected by a 450 mm Cassegrain telescope. After the bandpass filtering, the light is directed into a double-grating polychromator to separate the wavelengths and to suppress elastic backscatter. A 32-channel linear-array photomultiplier tube is employed to sample atmospheric Raman water spectrum between 401.65 and 408.99 nm. The 354.7 nm elastic Mie/Rayleigh scattering light in each channel is suppressed by more than 13 orders of magnitude (6 orders by the bandpass filter and 7 orders by the polychromator).

The output photons from each channel are counted with a bin width of 200 ns (corresponding to a range bin length of 30 m). The photon counts from 11,000 laser shots are accumulated to produce one photon count profile at each channel. This yields a time resolution of about 10 min for the raw data. The 32-channel photon count data constitute a spectrum- and altitude-resolved atmospheric Raman water signal. Altitude-resolved atmospheric Raman water spectrum signal is demonstrated in Fig. 40. Note that the prominent Raman water vapor signal peak at about 407.57 nm is visible at the altitudes up to 8.0 km.

The lidar-observed Raman water spectrum in the very clear atmosphere is nearly invariable in shape. It is dominated by water vapor, and can serve as a background reference for Raman lidar identification of the phase state of atmospheric water under



Fig. 36 Direct detection elastic lidar with three laser transceivers. Reproduced from Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. Opt. Eng. 56 (3), 031222.

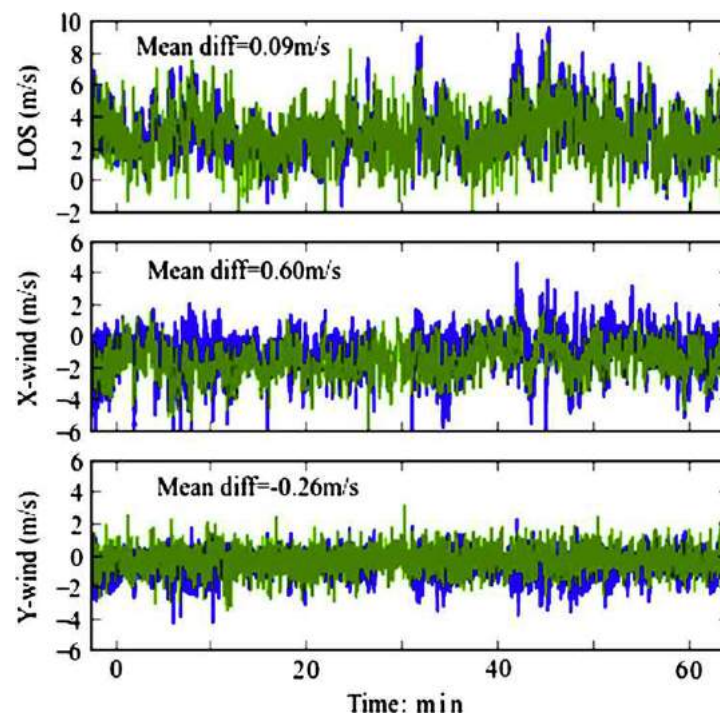


Fig. 37 Three components of lidar wind speed (green line) compared with the sonic anemometer (blue line) from over 2 h of measurements. Reproduced from Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. Opt. Eng. 56 (3), 031222.

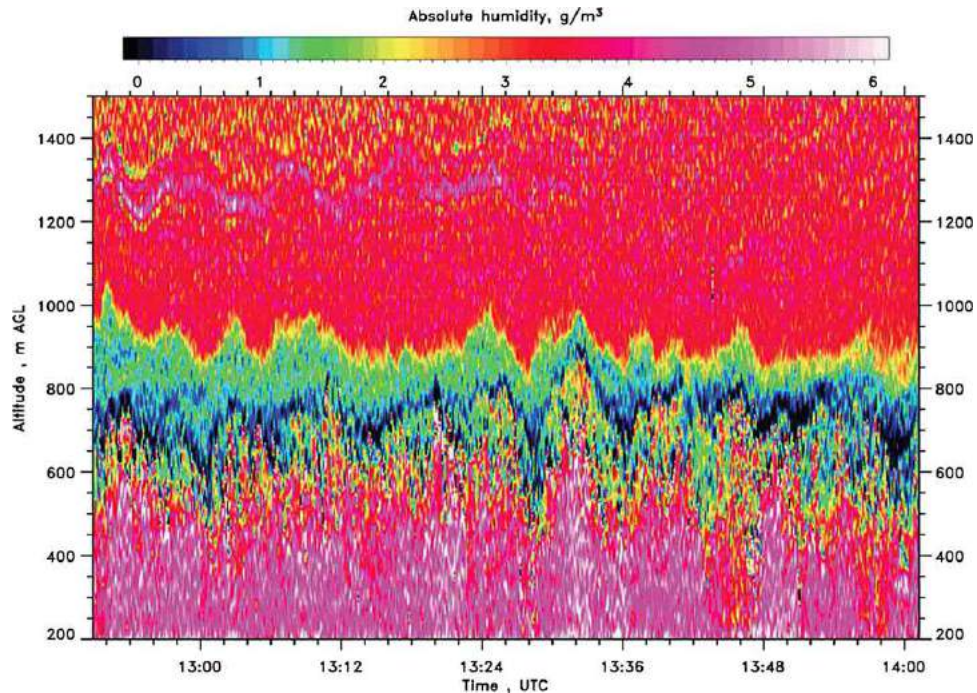


Fig. 38 Vertical-pointing measurement with DIAL. Reproduced from Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. Proc. SPIE 7475, 74750L.

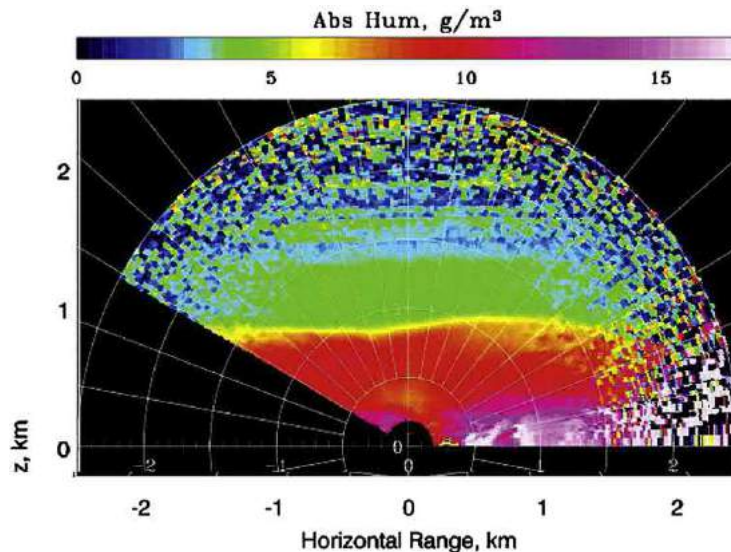


Fig. 39 Range-height indicator scan of the absolute humidity field. Reproduced from Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. Proc. SPIE 7475, 74750L.

various weather conditions. The lidar has measured also the Raman water spectrum of an aerosol/liquid water layer. It was noted that under clear weather conditions the Raman water spectrum intensity in the 401.6–404.7 nm range is permanently at a very low level.

The authors used also a 532-nm polarization/Raman lidar to observe the aerosol/liquid water layers in an apparently cloudless atmosphere (Fig. 41). These layers are characterized by elevated backscattering ratio R and low depolarization ratio δ (generally $\delta < 0.03$) in a narrow altitude range (with a width generally less than 1 km). The layers usually persist for a few to tens of hours.

It was noted by Mukherjee *et al.* (2010) that a Raman-scattering based instrument is expected to be less sensitive than an instrument based on IR absorption. One of the principal reasons for that is a 1012 times weaker Raman scattering cross-section than an IR absorption cross-section. The wavelength dependence of a Raman-scattering cross-section necessitates the use of short

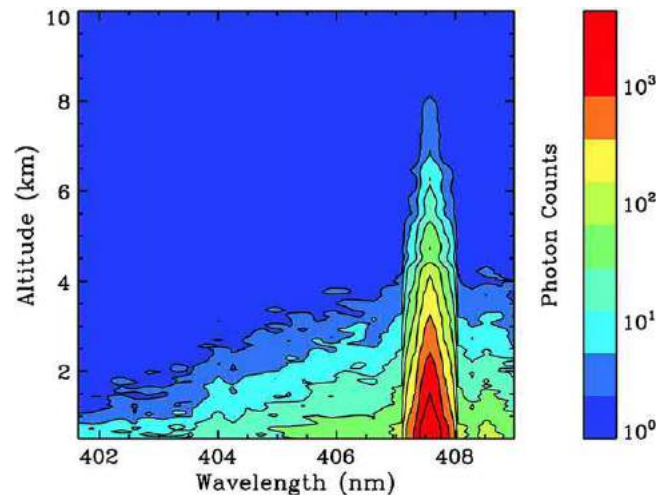


Fig. 40 Altitude-resolved atmospheric Raman water spectrum signal. Reproduced from Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. *Appl. Opt.* 52, 6884–6895.

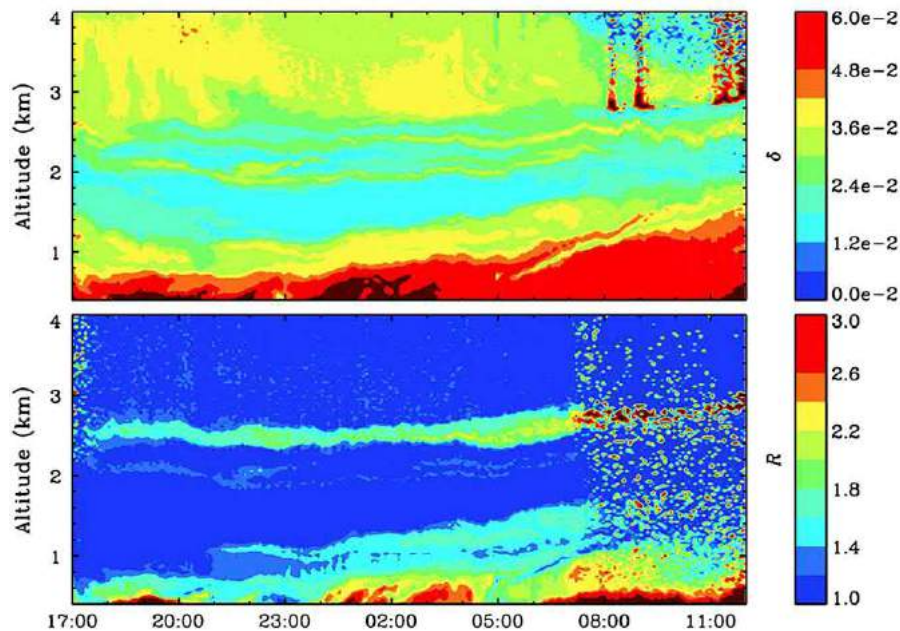


Fig. 41 Time-altitude diagrams of lidar backscatter ratio R and depolarization ratio δ measured by a Raman polarization lidar. Reproduced from Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. *Appl. Opt.* 52, 6884–6895.

wavelengths, often non-eye-safe, that are not appropriate for atmospheric propagation over long distances. On the other hand, strong optical absorption features of many explosives occur in the long wavelength atmospheric window in 8–12 μm .

For a target sample with absorption characteristics at about 10.6 μm , a wavelength of 10.653 μm (10P26 line from the CO_2 laser) was chosen, which is at the peak of the absorption feature. A spatial x-y scan of the local temperature was undertaken. A temperature rise was recorded only where the sample is located. Thus, the spatial dependence of the temperature rise shows the location of the target residue and its wavelength dependence allows identification of the target substance, even in the presence of other absorbers. Detection and identification of trace amounts of TNT at distances of 150 m were demonstrated. The technique should eventually permit real-time detection of explosives at standoff distances approaching a kilometer with high degrees of specificity and confidence.

To enhance the detection sensitivity, Kamerman (2012) proposed to measure the spectral correlation, i.e., the correlation between the sets of spectral lines, not a single one thus defining the degree of similarity between the received spectrum of unknown contaminants and the spectrum of known components which are to be detected. This technique implemented on the pre-detection stage of optical processing overcomes the problem of small cross-section of target chemicals and makes Raman lidar practical for

many important applications. In particular, the short-range stand-off detection of explosive residue or drugs on the exterior of packages, luggage or vehicles now appear to be possible.

Temperature

Temperature remote sensing using the laser radar techniques is based on measurement either of the Raman scattering or of the differential absorption called differential absorption lidar (DIAL). Murray *et al.* (1980) applied a 10 μm DIAL system that utilizes absorption by naturally occurring CO_2 to make path-averaged measurements of atmospheric temperature. Two widely spaced vibrational-rotational lines, the P(38) and P(20) transitions of the 10 μm band, were chosen for their relative high sensitivity and relative freedom from spectral interference. The transmission of one wavelength relative to the other was shown to be an effective measure of the average temperature over a 5 km path. Using CO_2 lidar and scattering from the foothills, relative transmission over that path was measured. A good correlation was obtained between the lidar-measured temperatures and thermometer-measured values.

The first instruments for rotational Raman temperature measurements were proposed by Cooney (1972). Philbrick (1994) measured the temperature structure of the atmosphere based on the rotational Raman scattering in the region between 526 and 532 nm. Arshinov *et al.* (1983) exploited the double-grating spectrometers to clean the signals from the elastic backscatter. Behrendt and Reichardt (2000) presented a filter polychromator in which the interference filters are mounted sequentially under small angles of incidence. High suppression of the elastic backscatter signal in the rotational Raman detection channels allowed temperature measurements independent of the presence of thin clouds or aerosol layers. Behrendt and Reichardt measured the atmospheric temperature in the northern Sweden up to 40 km heights.

Based on a similar double-grating polychromator approach, Jia and Yi (2014) designed a pure rotational Raman lidar in which the wanted pure Raman rotational signals were extracted, and elastically backscattered light was suppressed. A second-harmonic generation Nd:YAG laser was built with the optical head presented in Fig. 42 (D – dichroic mirror; RM – reflecting mirror; BP – bandpass filter; L – lens; FMH – fiber mode homogenizer; F – fiber; FA – fiber bundle array; G – grating; PMT – photo-multiplier tube). The instrument allowed to measure the atmospheric temperature at altitudes of 5–30 km. Fig. 43 compares the results of Behrendt and Reichardt (2000) above acquired in January 1998 in Northern Sweden (Kiruna) with the results of Jia and Yi (2014), acquired in August 2000 in Southern China (Wuhan). Pay attention that at the heights of about 16–17 km the temperatures are equal for both cases.

Fluctuation as a Dimension (Fluctuation Vision)

Atmospheric turbulence provokes much trouble in many applications. Nonetheless, the specificity of fluctuations of the signals propagating in the fluctuating atmosphere and reflected back from specular targets (like optical devices) and from diffuse targets originated the idea of selection of specular targets in the diffuse background, since specular and diffuse targets fluctuate differently when illuminated by the coherent laser light (Molebny, 1993).

It was shown, that these spectra are functions of spatial coherence of laser illumination and relative amplitude of random scanning due to atmosphere turbulence (Molebny *et al.*, 1989). For the diffuse background, the spectra depend also on relative

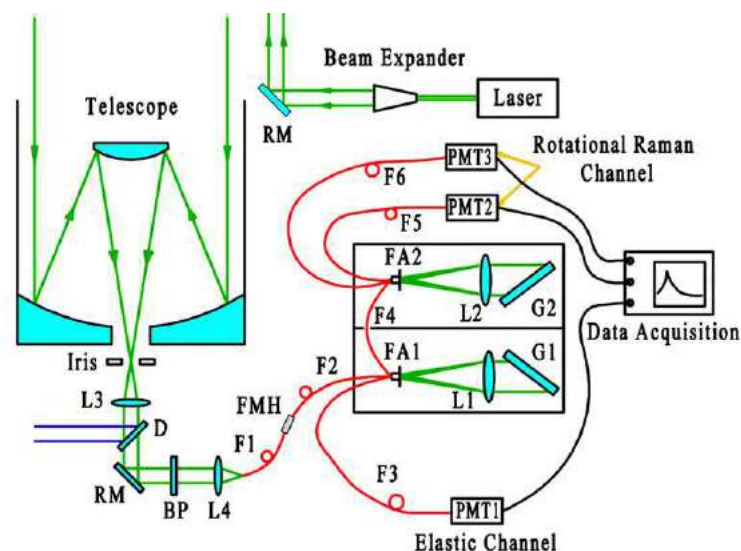


Fig. 42 Optical layout of the rotational Raman lidar (Wuhan University) for atmospheric temperature measurement. Reproduced from Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. *Appl. Opt.* 53, 5330–5343.

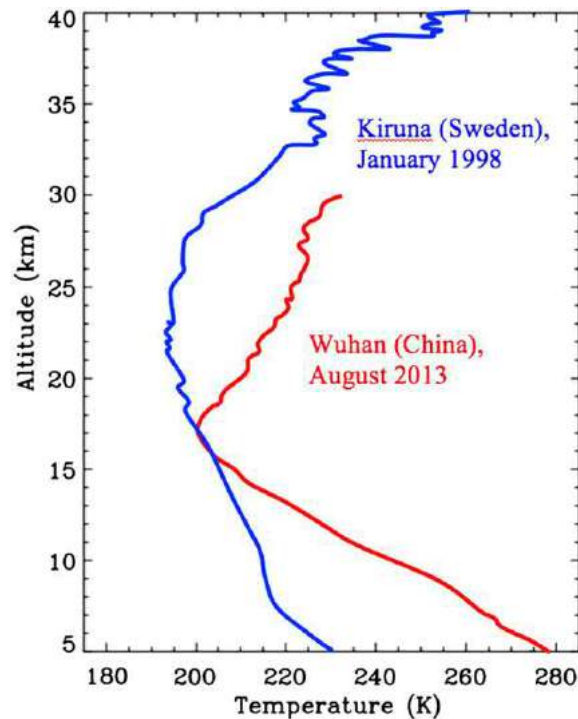


Fig. 43 Comparison of temperature structure for different sites and different seasons. Reproduced from Behrendt, A., Reichardt, J., 2000. Atmospheric temperature profiling in the presence of clouds with a pure rotational Raman lidar by use of an interference-filter-based polychromator. *Appl. Opt.* 39, 1372–1378 and Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. *Appl. Opt.* 53, 5330–5343.

scale of background non-homogeneity. Recursive filtration is one of the ways to implement the fluctuation vision using the atmospheric turbulence. The digitized image of the scene is accumulated at the output of the processing unit in such a way that earlier pixels have less weight. For each pixel in the current frame, the difference is calculated between the current value and the value accumulated from the previous frame. Varying the value of the weight factor, the weight of the preceding history can be adjusted to optimize the signal to noise ratio.

The dependence of spectra on random scanning due to atmosphere turbulence can be avoided due to angular micro-scanning of the beam, or by a scanning of the speckle structure inside the beam. This angular scanning or any other kind of sweeping of the internal speckle structure can be implemented at high frequencies thus simplifying the processing of lidar signals.

Dimension Combinations

Velocity and Temperature

Simultaneous measurements of velocity and scalar fields like temperature involve the combination of different techniques, such as particle image velocimetry (PIV) and planar laser induced fluorescence (PLIF). Tsurikov *et al.* (1999) reported measurements of velocity and conserved scalar fields combining PIV and acetone PLIF imaging in a gaseous turbulent flow. In liquids, there are reports of simultaneous velocity and temperature measurements using PIV and PLIF in a water impinging jet (Sakakibara *et al.*, 1997). Laser induced fluorescence of specially involved (seeded) molecules-tracers (tagged velocimetry) allows for simultaneous velocity and temperature measurements in gaseous flow fields (Hsu *et al.*, 2009; Sánchez-González *et al.*, 2012). A transient NO grid in the flow is “written” using the 355 nm photolysis of NO₂, which is subsequently probed by planar laser induced fluorescence imaging to reconstruct velocity maps.

Experimental setup is shown schematically in Fig. 44. It contains a photodissociation 355 nm Nd:YAG laser operated at 10 Hz producing a total power of 150 mJ/pulse. The 9 mm diameter laser beam is expanded with a $2.5 \times$ beam expander and directed into the chamber perpendicular to the flow axis through sheeting optics. A 50:50 beam splitter is used to optionally split the beam to direct a second “write” laser sheet parallel to the flow axis to obtain the two-component velocity measurements of the flow field. The photo dissociation laser beams are directed through an aluminum mesh or (in another experiment) a micro-cylindrical lens array to produce a periodic modulated pattern of NO photoproducts in the flow.

Each of the two identical fluorescence (probe) laser systems consists of an injection (seeding) Nd:YAG laser operated at 10 Hz. The 532 nm output is used to pump a pulsed dye laser to produce a tunable output beam in a range from 600 to 630 nm using a

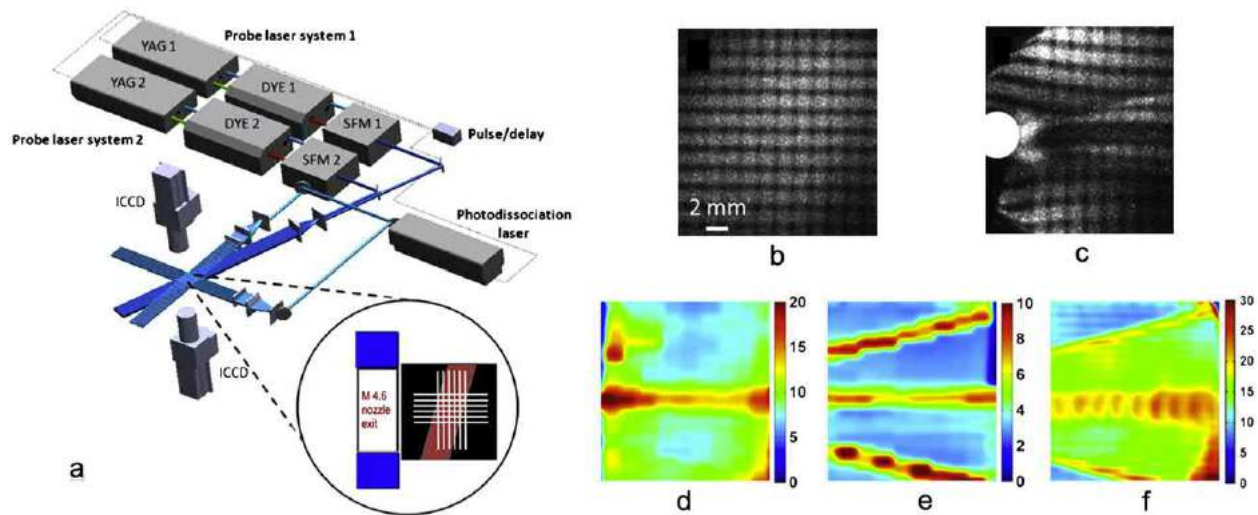


Fig. 44 Schematic of the experiments in the repetitively pulsed hypersonic flow (a). Instantaneous images probing the transition of the vibrationally excited NO in the free stream (b) and in the flow over a sphere (c). Components of velocity u/u_0 and v/v_0 , (d, e) and temperature T/T_0 (f) fluctuations normalized to the free stream values (in percentage). Reconstructed from Sánchez-González, R., Bowersox, R.D.W., North, S.W., 2012. Simultaneous velocity and temperature measurements in gaseous flow fields using the vibrationally excited nitric oxide monitoring technique: A comprehensive study. *Appl. Opt.* 51, 1216–1228.

solution of Rhodamine 610 and Rhodamine 640 in methanol. The dye laser output is mixed with the residual 355 nm beam in a frequency mixing unit to produce approximately 10 mJ/pulse in a range from 223 to 227 nm. This output wavelength range permits the probing of fluorescence transitions. For the experiments, the probe beams are directed into the chamber at an angle of 70° from the streamwise flow direction to avoid the aluminum mesh or the microlens array used for the “write” beam. The fluorescence images were acquired with two ICCD cameras mounted on either side of the chamber perpendicular to the laser sheets.

The measurements are performed during the steady flow phase of the pulsed flow. The basic timing sequence consists of an initial laser pulse, the “write” pulse, that photodissociates NO_2 using 355 nm. After a time delay, a first “read” dye laser excites a specific rotational state of the NO photoproduct, and an associated ICCD camera captures a fluorescence image of the flow. After a second time delay, the second “read” system excites a different NO rotational state to capture a second fluorescence image to get a 2D distribution.

A true simultaneous velocity/temperature measurement is possible when the velocity is determined using the instantaneous images generated by sequential “read” pulses. In Fig. 44, diagrams to the right demonstrate measurements in both the Mach 4.6 free stream (b) and in the wake of the Mach 4.6 flow over a cylinder (c). The images (b) and (c) are instantaneous fluorescence shots. The flow movement is from left to right. The cylinder position is shown as a white circle.

Conclusion

Every new dimension of the acquired information can be compared to a new window in the house. The more windows, the brighter the room is. We outlined the main stages of information extension: from a simple single-dimension range measurement to determining the three-coordinate object location in the space, adding velocity and vibrations as dimensions, determining the parameters of the media in which the laser radiation propagates, combining several dimensions from the same laser radar or from several laser radar systems. Velocity and vibrations can be described as single-dimensional (radial component), two-dimensional (lateral x-y directions), or full three-dimensional. Vibrations can have a fourth (spectral) dimension. Fluctuations of the signal reflected from the target can provide still another dimension of the information describing the target. Polarization is a source of information in laser radars that can be used as a dimension in different applications, e.g., remote measurement of object orientation (Molebny, 1975) or its shape (García-Arellano *et al.*, 2017), clutter suppression (Sassen, 2003), object or its material identification (Raghavan *et al.*, 2008), etc. We sketched a hypothetical multi-dimensional laser radar system which is the combination of 3D space coordinates plus 3D velocity in each point of the space + time dimension. This dimensionality can be even broader if temperature or another medium parameter is added. It is very productive to combine the information from laser radar with information from other sources: acoustic (Donzier and Cadavid, 2005; Lindgren *et al.*, 2011), microwave (Duckworth *et al.*, 1995; Yang and Qianqian, 2009), infrared (Tong *et al.*, 1987; Bartlett *et al.*, 2017), etc.

References

AGM-129 ACM. Directory of U.S. Military Rockets and Missiles. Raytheon (General Dynamics) AGM-129 ACM. Available at: <http://www.designation-systems.net/dusrm/m-129.html>.

- Albota, M.A., Aull, B.F., Fouche, D.G., *et al.*, 2002b. Three-dimensional imaging laser radars with Geiger-mode avalanche photodiode arrays. *Lincoln Lab. J.* 13 (2), 351–370.
- Albota, M.A., Heinrichs, R.M., Kocher, D.G., *et al.*, 2002a. Three-dimensional imaging laser radar with a photon-counting avalanche photodiode array and microchip laser. *Appl. Opt.* 41, 7671–7678.
- Amzajerdian, F., Petway, L., Hines, G., *et al.*, 2012. Fiber Doppler lidar for precision navigation of space vehicles. *Lasers, Sources, and Related Photonic Devices Tech. Digest*. doi:10.1364/FILAS.2012.FTh3.1A.2.
- Arshinov, Y.F., Bobrovnikov, S.M., Zuev, V.E., Mitev, V.M., 1983. Atmospheric temperature measurements using a pure rotational Raman lidar. *Appl. Opt.* 22, 2984–2990.
- Baker, I., Owtion, D., Trundle, K., *et al.*, 2008. Advanced infrared detectors for multimode active and passive imaging applications. *Proc. SPIE* 6940, 69402L.
- Bartlett, P.W., Coblenza, L., Sherwina, G., *et al.*, 2017. A custom, multi-modal sensor suite and data analysis pipeline for aerial field phenotyping. *Proc. SPIE* 10218, 1021804.
- Behrendt, A., Reichardt, J., 2000. Atmospheric temperature profiling in the presence of clouds with a pure rotational Raman lidar by use of an interference-filter-based polychromator. *Appl. Opt.* 39, 1372–1378.
- Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. *Proc. SPIE* 7475, 74750L.
- Bilbro, J., 1984. Airborne Doppler lidar wind field measurements. *Bull. Am. Meteorol. Soc.* 65, 348–359.
- Bilbro, J.W., Vaughan, W.W., 1978. Wind field measurement in the non-precipitous regions surrounding severe storms by an airborne pulsed Doppler lidar system. *Bull. Am. Meteorol. Soc.* 59, 1095–1100.
- Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. *Proc. SPIE* 7684, 768435.
- Bradbury, S.M., Mirzoyan, R., Gebauer, J., *et al.*, 1997. Test of the new hybrid INTEVAC intensified photocell for the use in air Cherenkov telescopes. *Nucl. Instrum. Methods Phys. Res. A* 387, 45–49.
- Brigety, R.E., 2007. *Ethics, Technology, and the American Way of War. Cruise Missiles and US Security Policy*. London, New York: Routledge.
- Busck, J., Heiselberg, H., 2004. Gated viewing and high-accuracy three-dimensional laser radar. *Appl. Opt.* 43, 4705–4710.
- Cho, K.Y., Satiya, A., Pourpoint, T.L., Son, S.F., Lucht, R.P., 2014. High-repetition-rate three-dimensional OH imaging using scanned planar laser-induced fluorescence system for multiphase combustion. *Appl. Opt.* 53, 316–326.
- Cho, P., Anderson, H., Hatch, R., Ramaswami, P., 2006. Real-time 3D lidar imaging. *Lincoln Lab. J.* 16 (1), 147–164.
- Cooney, J., 1972. Measurement of atmospheric temperature profiles by Raman backscatter. *J. Appl. Meteorol.* 11, 108–112.
- Donovan, D.P., Apituley, A., 2013. Practical depolarization-ratio-based inversion procedure: Lidar measurements of the Eyjafjallajökull ash cloud over the Netherlands. *Appl. Opt.* 52, 2394–2415.
- Donzier, A., Cadavid, S., 2005. Small arm fire acoustic detection and localization systems: Gunfire detection system. *Proc. SPIE* 5778, 245–253.
- Duckworth, G.L., Frey, M.L., Remer, C.E., *et al.*, 1995. Comparative study of nonintrusive traffic monitoring sensors. *Proc. SPIE* 2344, 16–29.
- Electron Bombarded Active Pixel Sensor (EBAPS). Available at: <http://www.intevac.com>.
- Eloranta, E.W., King, J.M., Weinman, J.A., 1975. The determination of wind speeds in the boundary layer by monostatic lidar. *J. Appl. Meteorol.* 14, 1485–1489.
- Fischer, K.W., Abreu, V.J., Skinner, W.R., *et al.*, 1995. Visible wavelength Doppler lidar for measurement of wind and aerosol profiles during day and night. *Opt. Eng.* 34 (2), 499–511.
- Flom, T., 1972. Spaceborne laser radar. *Appl. Opt.* 11, 291–299.
- Fraczek, M., Behrendt, A., Schmitt, N., 2012. Laser-based air data system for aircraft control using Raman and elastic backscatter for the measurement of temperature, density, pressure, moisture, and particle backscatter coefficient. *Appl. Opt.* 51, 148–166.
- García-Arellano, A., Cruz-Santos, W., García-Arellano, G., Juvenal Rueda-Paz, J., 2017. One-shot shape measurement of small objects with a pulsed laser and modulation of polarization. *Opt. Eng.* 56 (6), 064102.
- Hair, J.W., Hostetler, C.A., Cook, A.L., *et al.*, 2008. Airborne high spectral resolution lidar for profiling aerosol optical properties. *Appl. Opt.* 47, 6734–6753.
- Halo Photonics. Available at: <http://halo-photonics.com/index.htm>.
- Henderson, S.W., Hale, C.P., Huffaker, R.M., *et al.*, 1993. Eyesafe coherent laser radar for velocity and position measurements, US Pat. 5,237,331 (Aug. 17, 1993).
- Herfurth, T., Schröder, S., Trost, M., *et al.*, 2013. Comprehensive nanostructure and defect analysis using a simple 3D light-scatter sensor. *Appl. Opt.* 52, 3279–3287.
- Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. *Proc. SPIE* 6950, 69500V.
- Hsu, A., Srinivasan, R., Bowersox, R., North, S., 2009. Two-component molecular tagging velocimetry utilizing NO fluorescence lifetime and NO₂ photodissociation techniques in an under expanded jet flow field. *Appl. Opt.* 48, 4414–4423.
- Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. *Appl. Opt.* 53, 5330–5343.
- Kammerman, G.W., 2012. Optical correlation spectroscopy for remote contaminant detection. *Bul. (Visnyk) Natl. Tech. Univ. Ukraine "KPI", Instum. Making, Kiev* 43, 61–71.
- Lockheed Martin. Available at: <http://www.lockheedmartin.com/products/WindTracer/index.html>.
- Leosphere. Available at: <http://www.leosphere.com/8/wind-energy>.
- Lindgren, D., Bank, D., Carlsson, L., *et al.*, 2011. Multisensor configurations for early sniper detection. *Proc. SPIE* 8186, 81860D.
- Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. *Appl. Opt.* 52, 6884–6895.
- Lutzmann, P., Frank, R., Ebert, R., 2000. Laser radar based vibration imaging of remote objects. *Proc. SPIE* 4035, 436–443.
- Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. *Opt. Eng.* 56 (3), 031215.
- Lutzmann, P., Göhler, B., Putten, F., Hill, C.A., 2011. Laser vibration sensing: Overview and applications. *Proc. SPIE* 8186, 818602.
- Mann, J., Angelou, N., Arnqvist, J., *et al.*, 2017. Complex terrain experiments in the New European Wind Atlas. *Phil. Trans. R. Soc. A* 375, 20160101.
- Matvienko, G.G., Grishin, A.I., Kharchenko, O.V., Romanovskii, O.A., 2006. Application of laser-induced fluorescence for remote sensing of vegetation. *Opt. Eng.* 45, 056201.
- Morpheus lander. Available at: <https://morpheuslander.jsc.nasa.gov>.
- Mikkelsen, T., 2014. Lidar-based research and innovation at DTU wind energy – A review. *J. Phys.: Conf. Ser.* 524, 012 007.
- Molebny, V., Steinvall, O., 2013. Laser remote sensing. Velocimetry based techniques. In: Tuchin, V.V. (Ed.), *Handbook of Coherent-Domain Optical Methods*. New York: Springer, pp. 363–395.
- Molebny, V., Steinvall, O., 2014. Multi-dimensional laser radars. *Proc. SPIE* 9080, 908002.
- Molebny, V., 2013. Wavefront measurement in ophthalmology. Aberrometry through the eyes of an engineer. [Chapter 9] In: Tuchin, V.V. (Ed.), *Handbook of Coherent-Domain Optical Methods*. New York: Springer Science + Business Media, pp. 315–361.
- Molebny, V., Zarubin, P., Kamerman, G., 2010. The dawn of optical radar: A story from another side of the globe. *Proc. SPIE* 7684, 76840B.
- Molebny, V.V., 1975. Remote sensing of the orientation of the facets of sea surface, USSR Patent 433339, Bulletin # 23, 20.06.1975.
- Molebny, V.V., 1993. Image synthesis from fluctuations: Fluctuation vision. *Proc. SPIE* 2029, 68–76.
- Molebny, V.V., Protasov, V.G., Steba, A.M., 1989. Fluctuation spectra of light reflected from specular and diffuse surfaces. *Herald (Visnyk) Kiev Univ., Physics* 30, 92–100.
- Mukherjee, A., Von der Porten, S., Patel, C.K.N., 2010. Standoff detection of explosive substances at distances of up to 150 m. *Appl. Opt.* 49, 2072–2078.
- Murray, E.R., Powell, D.D., van der Laan, J.E., 1980. Measurement of average atmospheric temperature using a CO₂ laser radar. *Appl. Opt.* 19, 1794–1797.
- NASA. Making the skies safe from windshear: Langley-developed sensors will help improve air safety. Available at: <https://www.nasa.gov/centers/langley/news/factsheets/Windshear.htm>.
- Palm, S.P., Melfi, S.H., Carter, D.L., 1994. New airborne scanning lidar system: Applications for atmospheric remote sensing. *Appl. Opt.* 33, 5674–5681.
- Perea, J., Libbey, B., 2016. Development of a heterodyne speckle imager to measure 3 degrees of vibrational freedom. *Opt. Express* 24, 8253–8265.
- Philbrick, C.R., Hallen, H.D., 2017. Signatures of dynamical processes in Raman lidar profiles of the atmosphere. *Proc. SPIE* 10191, 101910E.
- Philbrick, C.R., 1994. Raman lidar measurements of atmospheric properties. *Proc. SPIE* 2222, 922–931.

- Philbrick, R., Hallen, H., Wyant, A., *et al.*, 2010. Optical remote sensing techniques characterize the properties of atmospheric aerosols. *Proc. SPIE* 7684, 76840J.
- Pierrotet, D., Amzajerdian, F., Petway, L., *et al.*, 2009. Flight test performance of a high precision navigation Doppler lidar. *Proc. SPIE* 7323, 732311.
- Polytec. 2006. Polytec's vibrometers are indispensable tools to optimize parts and goods and to investigate natural dynamic processes. Available at: <http://www.polytec.com>.
- Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. *Opt. Eng.* 56 (3), 031222.
- Raghavan, M., Morris, M.D., Sahar, N.D., Kohn, D.H., 2008. Polarized Raman spectroscopy: Application to bone biomechanics. *Proc. SPIE* 6853, 68530W.
- Reichardt, J., Wandinger, U., Klein, V., *et al.*, 2012. RAMSES: German meteorological service autonomous Raman lidar for water vapor, temperature, aerosol, and cloud measurements. *Appl. Opt.* 51, 8111–8131.
- Revel, G.M., Castellini, P., Chiarotti, P., *et al.*, 2011. Laser vibrometry vibration measurements on vehicle cabins in running conditions: Helicopter mock-up application. *Opt. Eng.* 50, 101502.
- Sakakibara, J., Hishida, K., Maeda, M., 1997. Vortex structure and heat transfer in the stagnation region of an impinging plane jet (simultaneous measurement of velocity and temperature fields by digital particle image velocimetry and laser-induced fluorescence). *Int. J. Heat Mass Transfer* 40, 3163–3176.
- Sánchez-González, R., Bowersox, R.D.W., North, S.W., 2012. Simultaneous velocity and temperature measurements in gaseous flow fields using the vibrationally excited nitric oxide monitoring technique: A comprehensive study. *Appl. Opt.* 51, 1216–1228.
- Sassen, K., 2003. Polarization in lidar: A review. *Proc. SPIE* 5158, 151–160.
- Sato, T., Suzuki, Y., Kashiwagi, H., Nanjo, M., Kakui, Y., 1978. Laser radar for remote detection of oil spills. *Appl. Opt.* 17, 3798–3803.
- Schmidt, J., Wandinger, U., Malinka, A., 2013. Dual-field-of-view Raman lidar measurements for the retrieval of cloud microphysical properties. *Appl. Opt.* 52, 2235–2247.
- Sitter, D.N., Gelbart, A., 2001. Laser-induced fluorescence imaging of the ocean bottom. *Opt. Eng.* 40 (8), 1545–1553.
- Spinhrne, J.D., Hansen, M.Z., Caudill, L.O., 1982. Cloud top remote sensing by airborne lidar. *Appl. Opt.* 21, 1564–1571.
- Steinvall, O., Tuldahl, M., 2017. Laser range profiling for small target recognition. *Opt. Eng.* 56 (3), 031206.
- Steinvall, O., Persson, R., Berglund, F., Gustafsson, O., Gustafsson, F., 2014. Using an eyesafe military laser range finder for atmospheric sensing. *Proc. SPIE* 9080. [9080–32].
- Stettner, R., 2010. Compact 3D flash lidar video cameras and applications. *Proc. SPIE* 7684, 768405.
- Takeda, M., Motoh, K., 1983. Fourier transform profilometry for the automatic measurement of 3D object shapes. *Appl. Opt.* 22, 3977–3982.
- Targ, R., Kavaya, M.J., Huffaker, R.M., Bowles, R.L., 1991. Coherent lidar airborne windshear sensor: Performance evaluation. *Appl. Opt.* 30, 2013–2026.
- Targ, R., Steakley, B.C., Hawley, J.G., *et al.*, 1996. Coherent lidar airborne wind sensor II: Flight-test results at 2 and 10 μm . *Appl. Opt.* 35, 7117–7127.
- Tong, C.W., Rogers, S.K., Mils, J.P., Kabrisk, M.K., 1987. Multisensor data fusion of laser radar and forward looking infrared (FLIR) for target segmentation and enhancement. *Proc. SPIE* 782, 10–19.
- Tsurikov, M.S., Rehm, J.E., Clemens, N.T., 1999. High-resolution PIV/PLIF measurements of a gas-phase turbulent jet. In: *Proceedings of the 37th Aerospace Sciences Meeting*, AIAA 99-0930.
- Uthe, E.E., 1991. Elastic scattering, fluorescent scattering, and differential absorption airborne lidar observations of atmospheric tracers. *Opt. Eng.* 3 (1), 66–71.
- Vasiljevic, N., Palma, J., Angelou, N., *et al.*, 2017. Perdigo 2015: Methodology for atmospheric multi-Doppler lidar experiments. *Atmos. Meas. Tech. Discuss.* 1–28. doi:10.5194/amt-2017-18.
- Verghese, S., McIntosh, K.A., Liao, Z.L., *et al.*, 2009. Arrays of 128×32 InP-based Geiger-mode avalanche photodiodes. *Proc. SPIE* 7320, 73200M.
- Wang, J.Y., Bartholomew, B.J., Streiff, M.L., Starr, E.F., 1984. Imaging CO₂ laser radar field tests. *Appl. Opt.* 23, 2565–2571.
- Werner, C., Flamant, P.H., Reitebuch, O., *et al.*, 2001. Wind infrared Doppler lidar instrument. *Opt. Eng.* 40 (1), 115–125.
- Wilkerson, T., Bradford, B., Marchant, A., *et al.*, 2009. VisibleWindTM: A rapid-response system for high-resolution wind profiling. *Proc. SPIE* 7460, 746009.
- Wong, C.M., Logan, J.E., Bracikowski, C., Baldauf, B.K., 2010. Automated in-track and cross-track airborne flash LADAR image registration for wide-area mapping. *Proc. SPIE* 7684, 768427.
- Yang, W., Qianqian, W., 2009. The application of lidar in detecting the space debris. *Proc. SPIE* 7160, 71601S.
- ZephIR. Available at: <https://www.zephirlidar.com>.
- Zhang, Q., Su, X., 2005. High-speed optical measurement for the drumhead vibration. *Opt. Express* 13, 3110–3116.

A Review of Laser Range Profiling for Target Recognition

Ove Steinvall, Swedish Defense Research Agency (FOI), Linköping, Sweden

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The classification and identification of small target, and/or targets at long range is a prime need for defense and security. Examples involve targets such as aircraft, unmanned aerial vehicles (UAVs) and missiles as well as small boats at sea and in littoral waters which have become common in illegal immigration, piracy, drug trafficking, and asymmetric threats.

Radar is the dominant sensor for long-range target detection and also offers model- and feature-based classification techniques using high resolution range profiles (HRRPs), target micro-Doppler spectra, range-Doppler imaging, and inverse synthetic aperture radar (ISAR) imaging. Techniques based on target tracking and target behavior are also studied for recognition.

Laser radar (or ladar) can also accomplish small target, or long range, target classification. Laser radar can provide narrow beams resolving multiple targets and can also provide recognition capabilities based on vibration, 1D range profiling, as well as 2D or 3D imaging. In this paper, we will concentrate on range profiling. Compared with its radar counterpart (HRRP) laser radars can offer more stable signals where fluctuations of the range-profile amplitude due to speckle often causes problems in classification.

The speckle problem is much less pronounced for a direct detection optical system, and very high range resolution in the centimeter range can be obtained. This makes a profiling laser radar an interesting candidate for long-range target, or small target recognition, where the transverse resolution for imaging sensors is limited to a single or a few pixels. A laser range finder is today a standard sensor in most EO systems and can have a profiling capability, or be modified to be more suitable for this task.

1D laser techniques for target recognition include vibrometry (Lutzmann *et al.*, 2011) and laser range profiling (LRP) including its extension to tomography. LRP is attractive because the maximum range can be substantial, especially for a small beam width while still preserving a high range resolution. LRP will usually work together with an IR and radar sensor which acquires the target. A laser profiler can complement a jammed radar, or a silent radar wanting to limit the effectiveness of anti-radiation missiles. It can also extend the capability of an IR-sensor to include small and long-range target classification. LRP can be used in a scanning mode to detect targets within a certain sector. The same laser can be used for illumination in active imaging when the target comes closer and is angular resolved.

LRP has been investigated for both search and ID of small surface targets (Kunz *et al.*, 2005; van den Heuvel *et al.*, 2009a; Schoemaker and Benoist, 2011) as well as for aircraft ID (van den Heuvel *et al.*, 2008; Dierking *et al.*, 1998). Recently, we showed by simulations and laboratory measurements that LRP is a promising method of identification of small sea surface targets at long ranges (Steinvall *et al.*, 2012b, 2014; Steinvall and Tuldahl, 2016). A series of laser radar range profiles collected at various aspect angles allows for a tomographic reflective reconstruction of the target (Knight *et al.*, 1989b). Many examples of this are found from simulations and laboratory measurements (Knight *et al.*, 1989a,b; Jin and Levine, 2009; Parker *et al.*, 1988; Henriksson *et al.*, 2012a,b) and other demonstrate field results (Murray *et al.*, 2010). Many of the applications seem to be devoted to satellite imaging (Lasche *et al.*, 2009; Matson and Mosley, 2001). Using time correlated single-photon counting, very high range resolution (millimeter–centimeter) profiling and tomography can be demonstrated (Sjöqvist *et al.*, 2014).

This review article starts with reviewing the theoretical basis of LRP, and relating this theoretical basis to the corresponding techniques for microwave radar. Examples of simulations, and related uncertainty in building a realistic library, to compare measured with stored data, will be discussed. Experimental results will be given for different types of targets and compared with simulated data. Furthermore, examples in signal processing techniques will be discussed. Finally, conclusion and future prospects for the technique will be presented.

Background for LRP

LRP is based on sending out a short pulse, or modulating the laser emission, so that high range resolution is obtained. Depending on the target depth structure the resolution may be high enough to profile this structure, typically meaning at least 10 resolution cells over a target, such as an aircraft.

Direct detection systems, such as present military range finders, offer target maximum ranges in the 10–20 km range and a range resolution in the region 0.2–1 m. The number of speckle lobes N_{sp} falling within the receiver aperture D_{ap} from the target plane from a certain range cell is at a first approximation

$$N_{sp} \sim \left[\frac{D_{ap} D_{target}}{\lambda R} \right]^2 \quad (1)$$

where D_{target} is the illuminated target area, λ the laser wavelength, and R the target range. For $D_{ap}=0.1$ and $D_{target}=1$ m the range for which $N_{sp}=1$ is about 100 km and for $D_{target}=0.1$ m, $R=10$ km. Thus, for very small targets, and long ranges, the speckle noise might be present from a diffuse target. On the other hand, for a glint target turbulence fluctuations will dominate. However, pulse

averaging over 1–3 s will average out most of the turbulence and target speckle fluctuations. This has been demonstrated in both measurements and simulation when range profiling of small surface vessels (Steinvall *et al.*, 2012b; Steinvall and Tulldahl, 2016).

In order to estimate the range performance of a laser profiling system indicative of a laser range finder type we will start by expressing the detected laser power (P_{det}) written as

$$P_{\text{det}} = T_{\text{optics}} \cdot P_0 \cdot \frac{\rho_{\text{target}}}{\pi} \cdot \frac{A_{\text{rec}}}{R^2} \cdot \frac{4 \cdot A_{\text{target}}}{\pi(\phi R)^2} \cdot T_{\text{atm.laser}}^2 \quad (2)$$

where P_0 is the transmitted peak power, R the target range, ϕ the laser divergence, ρ_{target} the target reflectance, and $T_{\text{atm.laser}} = \exp(-\sigma R)$ the one-way atmospheric laser transmission. The signal-to-noise ratio (SNR) is given by $\text{SNR} = P_{\text{det}}/\text{NEP}$. The maximum range for the system as a function of atmospheric extinction can easily be obtained with the help of the Lambert-W function (Steinvall, 2009).

We will choose some sensor data indicative of an eye-safe 1550 nm laser range finder. The laser pulse energy is set to 30 mJ and the pulse length 6 ns, giving a peak power of about 10 MW. The receiver area is assumed to be 0.02 m² corresponding to a diameter of 16 cm, and the receiver noise equivalent power (NEP) was 3 nW. The optical background is in general lower than this figure. The relation between the atmospheric extinction coefficient σ at 1550 nm wavelength and the visibility V was assumed to be $\sigma = 0.996(3/V)^{1.2}$ according to Hutt (Hutt *et al.*, 1994).

Fig. 1 shows the maximum range vs. visibility for different combinations of target strength $A_{\text{target}} \times \rho_{\text{target}}$ m² and the case of an unresolved target. As seen the range can be substantial during a wide range of visibilities. The threshold for the SNR was set to 7. Averaging over 10–100 pulses will improve the SNR approximately as the square root of the number of pulses provided that the individual pulses are summed up with the same “start” bin.

Time correlated single photon counting (TCSPC) is a special form of a direct detection laser radar where a single detected photon will cause the detector to saturate. This technique enables millimeter–centimeter range resolution. In TCSPC, this is accomplished by carefully measuring the time between a laser pulse sync signal and registration of a single-photon event of photons reflected from a target. The measurement is performed multiple times and a histogram of arrival times is computed to gain information about surfaces at different distances within the field of view (FOV) and to exclude spurious detections from detector and background noise. Systems using moderate pulse repetition rates and a limited number of acquisitions are usually called Geiger-mode APD (GMAPD) laser radar, whereas systems using laser pulse repetition rates in the megahertz range, that is, more acquisitions and lower detection probabilities, are termed TCSPC laser radar with single-photon avalanche photodiode (SPAD) detectors. We refer to Sjöqvist *et al.* (2014) for more detailed discussion of these techniques for range profiling and reflectance tomographic imaging.

For a range profile using GMAPDs the number of detected photons per pulse must be kept low, typically 0.1 photon. Otherwise the first photon will block pulses from close lying structures due to the dead time in this form of detection (Geiger mode).

To estimate the maximum range performance the SNR expression is not quite useful. Instead, the detection and false alarm probability ($P_{\text{det}}/P_{\text{fa}}$) should be used. An example of such an analysis is made in Steinvall *et al.* (2012a) where $P_{\text{det}}/P_{\text{fa}}$ is plotted versus range for a visibility $V = 10$ km and a 1 m² diffuse 10% reflectivity target. The average power was 1 mW for the 200 kHz laser with 6 ps pulse length and a 10 cm receiver aperture. The dwelltime T_{dwell} was assumed to be 1 ms which is enough to get a good detection probability. On the other hand, 10–100 ms may be more suitable to collect the range waveform details. We can see from Fig. 2 that both 1060 and 1550 nm laser radars have a good range capability using photon counting with only 1 mW of average power and an integration time $T_{\text{dwell}} = 1$ ms.

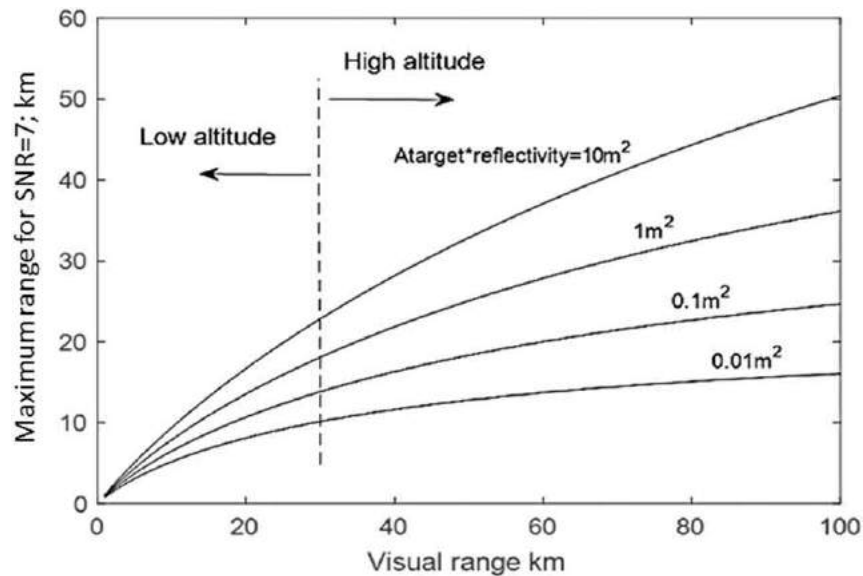


Fig. 1 Range performance for a typical profiling system corresponding to a laser range finder at 1550 nm.

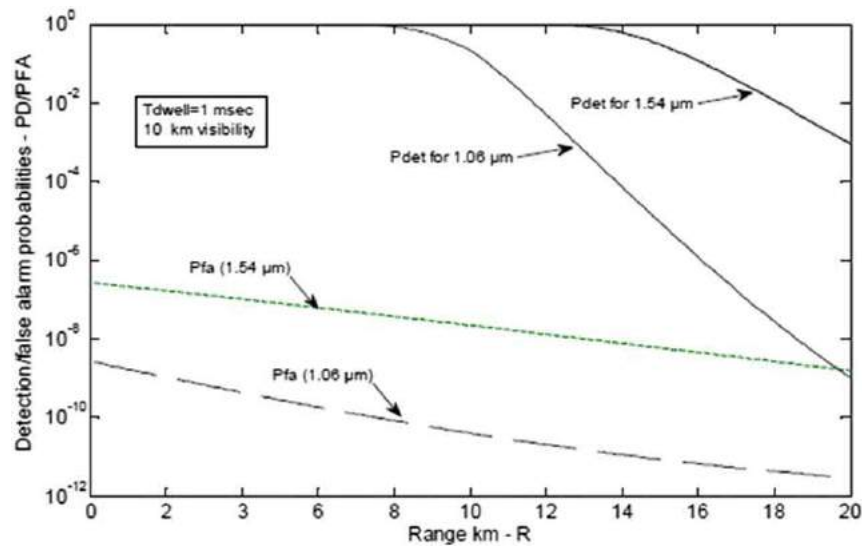


Fig. 2 Comparison between 1.06 and 1.5 μm lidar performance using the parameters according to the text above and in Steinvall *et al.* (2012) for a diffuse 10% reflectivity 1 m^2 target. Reproduced from Steinvall, O., Sjöqvist, L., Henriksson, M., 2012. Photon counting lidar work at FOI, Sweden. Proceedings of SPIE 8375, 83750C.

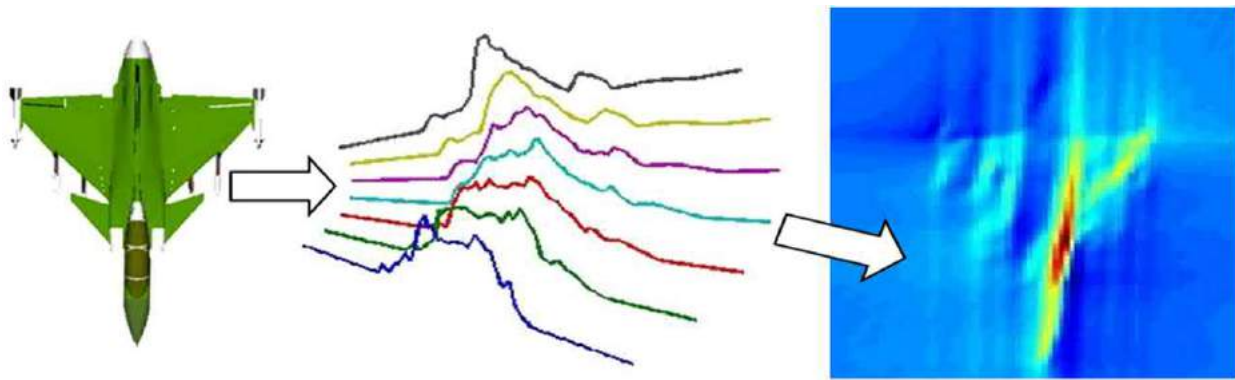


Fig. 3 Simulated laser radar waveforms from an aircraft and reconstruction of the tomographic image using these waveforms. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering Vol. 56 (3), 031206.

An attractive feature of LRP is the possibility to fuse several range profiles taken at different angles of incidence to reconstruct an image of the object (Murray *et al.*, 2010; Parker *et al.*, 1988). A tomographic image can be obtained by using the Radon transform to reconstruct the shape of the target. Satellite imaging by tomography has been demonstrated with a pulsed, mode locked, coherent system (Matson and Mosley, 2001). Frequency chirp techniques for obtaining high range resolution are also a well known method in coherent laser radar. Other techniques like range correlation methods (pulse compression) have been exploited by others using pseudo-random pulse codes. This was demonstrated in Murray *et al.* (2010) for image reconstruction of a meter sized target at 22.4 km range with 15 cm resolution.

Coherent detection also enables the use of the target rotation, and the generated Doppler spectrum, to determine the object size and rotation speed from 1D measurements related to range profiling.

Fig. 3 shows an example from simulations using several laser range rofiles from different aspects to form a tomographic image. However, the tomographic technique has its limitations in practical systems and is probably most useful for a fixed sensor and a rotation target around its own axis (Henriksson *et al.*, 2012a,b).

Simulating Laser Profiling Data

In order to enable target recognition from range profiling the measured range waveform has to be compared with that from a library containing a number of potential targets. The stored waveforms should contain all possible aspect angles of interest. This fact will widen the library content. Coupled to this data storage is also the question if the whole waveform,



Fig. 4 Left: a point cloud model of aircraft SK37 Viggen using the Riegl VZ-400 laser scanner. Right: the polygon model built with the point cloud as a reference. Images FOI.

or only its features, like individual peak positions and the relative amplitude of peaks should be stored. The best is of course if the whole waveform in addition to the characteristic features can be stored because that will enable a more robust signal processing.

As direct measurements of the target profiles over all aspects are unrealistic for many reasons, the modeling of different target waveforms is inevitable. This can be accomplished if a good 3D CAD model of the target is available. The surface angular reflectivity characteristics at the laser wavelength of interest is given from the bi-directional reflection distribution function (BRDF). This function is hard to measure in practice and often has to be modeled based on known or assumed target surface properties. Examples of studies of the BRDF influence on range pulse waveforms are discussed by [Steinvall \(2000\)](#) and the BRDF effects on ladar-based reflection tomography by [Jin and Levine \(2009\)](#). In general, the more glossy the surface is, the more accentuated will those parts be which are normal to the angle of incidence while other surfaces with a slope toward the beam will be reduced in amplitude. Strong glints from a target can mask the following range information if the dynamic range of the receiver is limited. In order to simulate range profiling data a CAD model of the target is desired. 3D CAD models can in many cases be found on the internet to a varying degree of details. These CAD models seldom contain any information about reflectivity and material properties so this has to be assumed. The most straightforward assumption may be to suppose that the surface is diffuse with a 10–20% reflectivity. If the real target is available for measurements this can be scanned using a commercial laser scanner, for example, the Riegl VZ-400 which has the ability to scan a position within an elevation angle of 160 degrees in a full revolution and with a pitch of 0.005 degrees using a 1550 nm laser. Using calibration charts the correct reflectivity in each scan point will automatically be obtained.

Fig. 4 shows a point cloud of an aircraft generated from a scan using the RIEGL scanner and also the model built with the point cloud as a reference. This polygon model including each surface reflectivity will be the base for simulating the laser radar profiling waveforms. Then a sensor model (in our case called the FOI-LadarSim ([Chevalier, 2011](#))) is applied to simulate waveforms from each pixel at the target (pixel density is selectable) and in that way 1D, 2D, and 3D data can be simulated. Beside sensor parameters and their effects, the atmospheric and target speckle are included in the signal generation.

Fig. 5 shows an example of simulated and measured range profiles. The measurements are made against a physical scaled down model of the aircraft JAS Gripen (length about 1.5 m). The measurements were made at short ranges with a 1.5 ns pulsed low power laser mimicking a typical laser range finder with about 10 ns pulse length against a full-sized aircraft ([Pace et al., 2008](#)).

Fig. 6 illustrate range profiling of a small boat complemented with a tube and a box to alter the profile. The dashed line is the simulated pulse return. Deconvolution of the second echo with the transmitter pulse reveals two peaks interpreted as indicated in the figure.

There are many uncertainties that could result in differences between the measured and the simulated data. Examples include incorrect geometry in the CAD models and unknown reflectance of the surface material. Roughly, the outer shape of the models may be correct but because the simulations take into account even small features there might be errors. One example is the inside of an aircraft engine or air inlet which might be a source of error. To solve this, further validation with access to the physical models would be required.

Other error sources include incorrect sensor and environmental effects modeling. For example, the exact beam distribution and beam pointing error at the target are hard to predict in detail. There are also some uncertainties in the system parameters (noise, receiver electronics, laser power, etc.).

Examples of Range Profiles Studies

Naval Targets

TNO Defense, Security, and Safety in the Netherlands have shown results on laser profiling mainly from naval ships ([Schoemaker and Benoist, 2011](#); [van den Heuvel et al., 2009a,b](#)). One example is illustrated in **Fig. 7**.

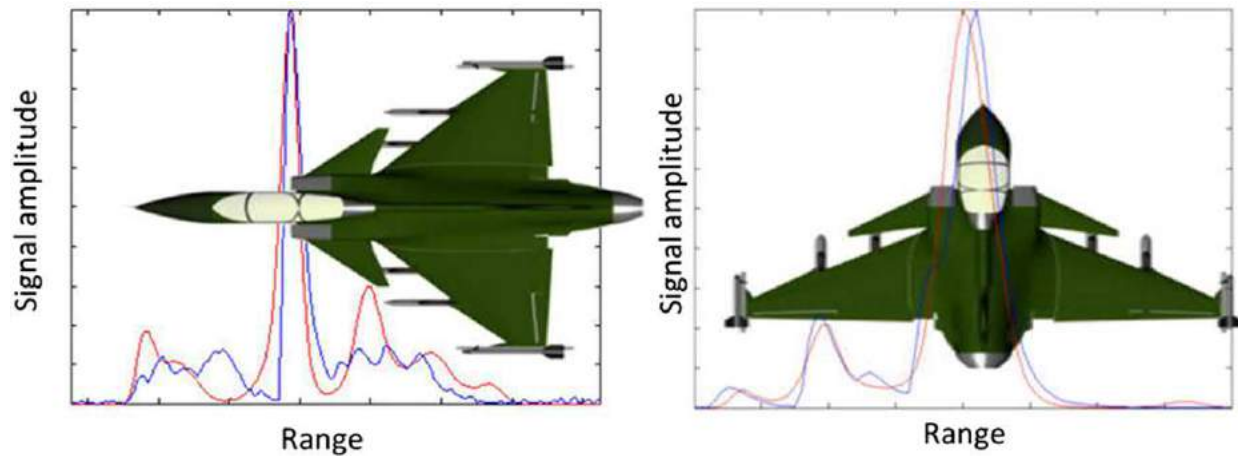


Fig. 5 An example of simulated and measured data. Left: a model of the JAS 39 Gripen seen from the front. Measurement data from a scaled down physical model of the aircraft is given by the red curve, while the simulated of the blue. The superimposed planes' picture shows the structures that correspond to every part of the signal. Right: shows the same comparison seen straight at 90 degrees from the main direction. Reproduced from Pace, P., Steinvall, O., Chevalier, T., 2008. Automatic target classification using one dimensional laser profiling. Technical Memorandum DRDC, Ottawa, TN 2007-000.

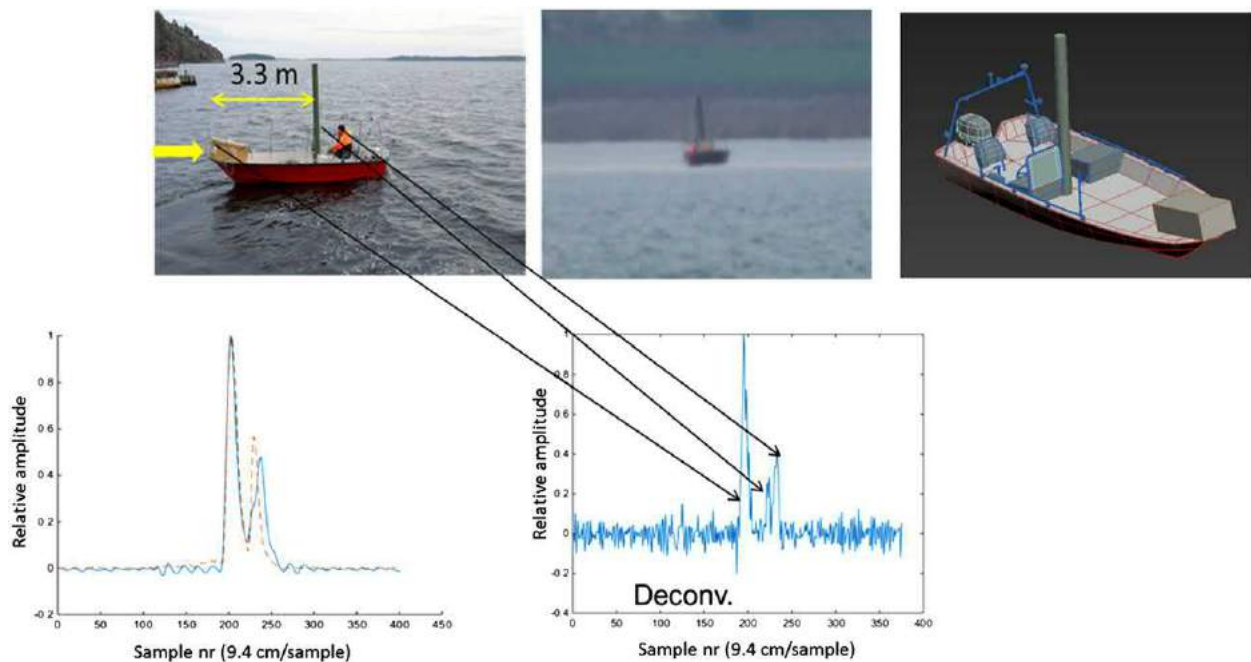


Fig. 6 Upper image shows the small boat complemented with a tube and a box to alter the range profile together with the model for waveform simulation. The first profile (dashed) is the model generated pulse. Deconvolution of the second echo with the transmitter pulse reveals two peaks interpreted as indicated in the figure. Reproduced from Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

Fig. 7 shows the measured and simulated range waveforms from a frigate at 10 km range measured in the head on direction. As seen the correspondence is quite good with the exception from two extra peaks present in the measured waveform. These can be associated with the two extra antenna spheres. In the 3D model these structures above the bridge are absent. The system used in these experiments had a transmitter output at 1570 nm around 30 mJ. The receiver used a 750 mm telescope, with an optical aperture of 125 mm diameter. The receiver bandwidth was 200 MHz (van den Heuvel *et al.*, 2009a,b).

A NATO trial in San Diego under the name "NATO Maritime Target ID" was carried out outside San Diego. Several laser and IR systems were used to investigate small surface classification/identification using both 1D, 2D, and 3D laser radars

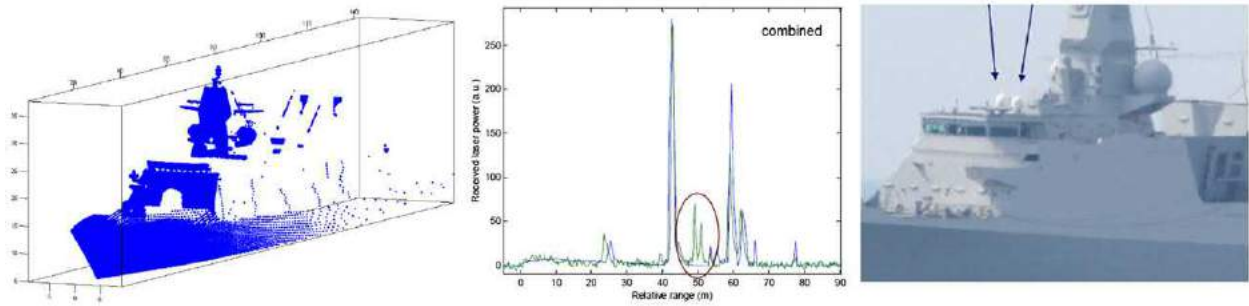


Fig. 7 Left: the point cloud of a ship is used to generate simulated range waveforms. Middle: the simulated (blue) and measured range waveforms from 10 km range. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

(Schoemaker and Benoist, 2011; Steinvall *et al.*, 2010; Armbruster and Hammer, 2012). The authors in Schoemaker and Benoist (2011) examine the discrimination between three different small boats all between 5–6 m in length. A simple and rather crude model of the boats was used to generate the simulated waveforms to compare with the measured. The result shows that the desired accuracy (95% or better) which was achieved in investigating different simulated waveforms for different noise conditions was not achieved for the experimental data.

In a later experiment in Sweden, FOI demonstrated rather good results with a profiling system which had almost the same characteristics as the TNO system (laser pulse width 6 ns, 10 cm optical aperture, 15 mJ /pulse, and 1570 nm laser wavelength at 40 Hz pulse repetition frequency – PRF). The paper Steinvall *et al.* (2015) shows both experimental and simulated results for LRP of small boats out to 6–7 km. Fig. 8 shows the small boats used in the test. These are more different from each other than the boats studied in the San Diego trial mentioned above. In the figure the CAD models of the boats are also shown based on laser scanning.

Fig. 9 shows the structure and the corresponding distances as well as an example of a range waveform. Some interpretation is formulated in the figure text. Fig. 10 shows the echo amplitude variation for the three observed peak during a 10 s data collection at a PRF of 40 Hz. Manual tracking was used to follow the small target at 6 km range which was a demanding task. Also note the high resolution sensors in the SWIR and mid-IR gave hardly any possibility for target classification without a support from the range profile. The SWIR camera had a FOV = 0.7 degree $\times$ 0.6 degree fully zoomed. The illuminator laser had a maximum power of 1.6 W and the beam divergence could be varied between 10 and 50 mrad. The camera had a spectral response between 0.4 and 1.7 μ m and the detector array had 640 $\times$ 512 pixels. The mid-IR camera had a 1280 $\times$ 720 array of 12 μ m pixels responding in the spectral range between 3.7 and 4.95 μ m with a typical sensitivity of 17 mK and a minimum narrow FOV of 0.9 degree $\times$ 0.5 degree.

Airborne Targets

Long-range laser profiling measurements of airborne targets are hard to find in the open literature. Measurements on static targets and results from simulation are reported in open sources. Heuvel *et al.* (van den Heuvel *et al.*, 2009a,b) describes an identification algorithm used to distinguish three aircraft from their simulated range profile with good results. The authors use simulated profiles obtained using 3D computer models for three aircraft: an F15, an F16, and an F117. They assume a lidar with 30 cm range resolution.

Heuvel *et al.* investigate the relation between identification and aspect angle. They assume that all aircraft have the same diffuse reflectivity. The identification range depends both on range and target aspect angle. An analysis as described in van den Heuvel *et al.* (2009a,b) will yield a (Bayes) identification probability for the various aspect angles. The SNR for a given identification probability can be associated with a range for a given system and target. Fig. 11 shows the radial plots of the required range (in arbitrary units) against the aspect angle for an identification probability of 90%. According to the authors it is clear from these plots that the required range is shorter for front view (aspect angle of zero degrees). This may be understood if we consider the aerodynamic shape of the aircraft that is responsible for a low laser return. We see a similar effect at 180-degree aspect angle.

It is also informative to compare the radial plots of the three aircraft. The F16 has a smaller maximum identification range than the F15 which has again a smaller range than the F117. Thus, the F117 can be identified at considerably longer ranges than the F16 on average. This is most likely due to the characteristic shape of the F117 (van den Heuvel *et al.*, 2009a,b).

Example of full scale laser profiling measurements on a real aircraft (SK-60 veteran aircraft) is shown in Fig. 12. Note that the range profiles change shape for rather small changes in azimuth. The measurements were made at 1 km range using a 1570 nm laser.

Fig. 13 shows a profile map from a horizontal scan for two different aircraft. The signature is quite different as can be seen. In Fig. 14 the measured profile covering the whole, aircraft mimicking a long-range profiling is compared using a simple 3D cad



Fig. 8 Test vehicles for the trial in Sweden. The upper boats were slightly modified by placing a stove pipe and a box to change the signature. Boat A (upper): length 4.73 m, width 1.92 m. Boat B (middle): length 6.18 m, width 2.15 m (total length incl. the rear flat plate=6.57 m). Boat C (lower): length 8.64 m, width 3.15 m (total length incl. the rear flat plate=9.55 m). The last row shows the 3D models based on laser scanning using the Riegl scanner. Reproduced from Steinvall, O., *et al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 964901.

model from the internet and assuming a complete diffuse surface for the whole aircraft. Many main features compare well but there are also some difference due to the incomplete aircraft model.

Small targets of interest for detection and classification include UAVs and missiles. Laser profiling of these targets may be done but need a much higher range resolution probably 10 cm or less. TCSPC offers a range resolution in the sub-centimeter level and we have demonstrated this against a UAV mockup according to Fig. 15 (Steinvall and Tuldahl, 2016).

Fig. 16 shows the UAV at the 1.3 km distance and the scanning the pattern from the single detector device color-coded in reflected laser intensity. The size of the pixels was about 15 cm in square. Since we had a weak laser we integrated the counting during 5 s/pixel. The laser PRF was 4 MHz and with a pulse duration of 25 ± 5 ps. The laser's average power was 20 mW. The laser divergence was $280 \mu\text{rad}$ and the detector FOV= $150 \mu\text{rad}$ (which gave some "spillover"). Fig. 17 shows the sum of the range waveforms over the entire scan area (corresponding to a fictitious beam divergence of about 1.7 mrad). We can clearly see echoes of the various structures including that from the cone-like front (provide a linear amplitude change with time). The distances derived from the pulse response are consistent with the actual distances within 1 cm.

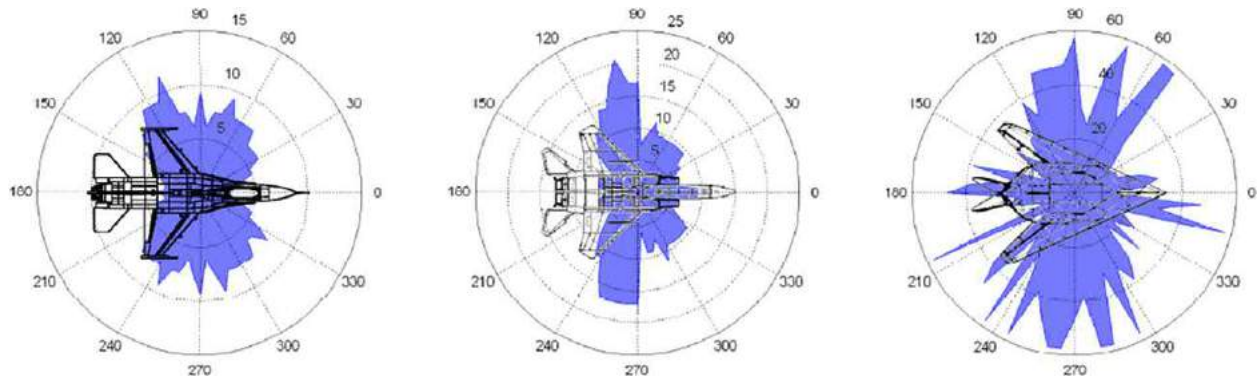


Fig. 11 Radial plot corresponding to an identification probability of 90% for an F16, F15, and F117, left to right respectively. After van den Heuvel, J.C., Schoemaker, R.M., and Schleijsen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

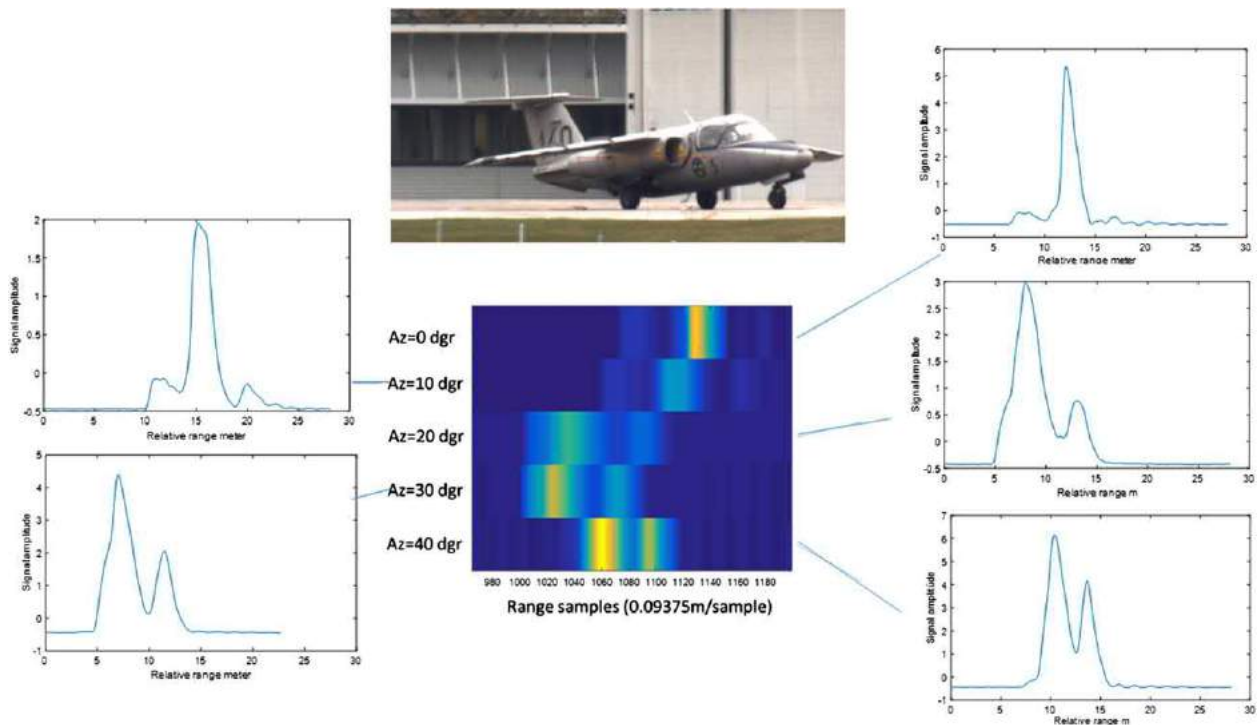


Fig. 12 Range profiles for the different azimuth positions of the SK-60 aircraft. The range profiles are summed over all measurement points on the aircraft for a specific azimuth ($=0$ for head on) and this mimicking a long-range profile when the laser lobe is covering the whole aircraft. Image FOI.

completely disappears, shadowed by the UAV body, and the first wing echo is spread out over time and combined with reflections from the UAV body.

Land Targets and Mapping

Results on LRP of land targets are limited concerning horizontal paths. On the other hand, vertical laser scanning from an airborne platform is well established especially for terrain and underwater sensing.

The civilian market has been leading this field for the last 20 years. Both space and airborne laser radar systems, as well as terrestrial laser radars, have been developed by several different vendors. The large number of publications and applications, as well as the rapid development of hardware, shows the large interest in this technology. A book giving a good overview of topographic laser scanning techniques was published by [Shan and Toth \(2009\)](#). [Fig. 19](#) illustrate some developments of airborne laser scanning system for both terrain and depth sounding applications.

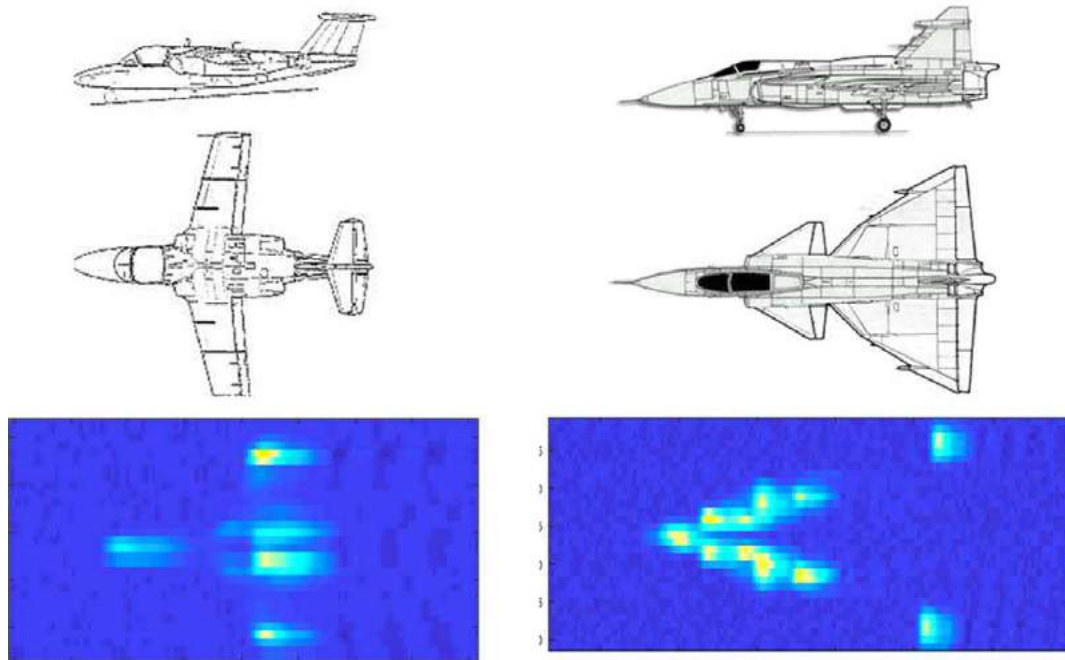


Fig. 13 Laser scan using a beam with 1 mrad beam divergence with the aircraft at 1 km range. Two veteran aircraft SK-60 (left) and Viggen (right) from the Swedish Air Force. Images FOI.

Laser scanners are used on UGVs for navigation and obstacle avoidance. Systems from Teledyne Optech and Velodyne Lidar are the examples of commercial road mapping laser radars with many potential military applications (e.g., generation of synthetic environments, detection of improvised explosive devices, and so on). The accuracy is said to be on the order of centimeters at a maximum range of 100–250 m. Measurement rates are found between 1 and 2 million points per second. This allows a 6-cm point spacing at 10-m range for a vehicle velocity of 43 km/h.

Vehicle borne laser radar is becoming a very important sensor for self-driving cars. This may be the first widespread commercial application of laser radar technology. It will force the technology to be small and compact at a cost of a few hundred dollars per system.

Scanning laser radars have been intensely studied in the United States for missile applications (Gustavson and Davis, 1992). A civilian counterpart is traffic monitoring classifying different vehicles (SICK-Sensor Intelligence, 2017). Target detection and homing were also studied for air to air seekers, and in space for the SDI program. The largest investments in laser radar seeker technology have been made for air to ground seekers. The Low Cost Autonomous Attack System (LOCAAS) was a demonstrator program (Ladar Seeker/ATA Algorithm Captive Flight Test Results, 2017) driven by the Air Force (Eglin Florida). The AFRL's LOCAAS program terminated in mid-2006 without going to production. The seeker was based on scanning laser radar generating 3D imagery of the targets with ATR and aim point selection built into the system. Loitering Attack Missile (Signal, 2017) was a project relying on the laser radar technology developed for LOCAAS. The paper from Andressen *et al.* (2005) gives some insight to this technology.

Laser scanning systems with a profiling capability can also be used for both target search and classification. If the return has a specific signature this could be indicative of a target that differs from the surrounding terrain. In principle, this could include the profiles from land vehicles which probably differs a lot from those from the terrain background. One example of a specific strong and deviating signature is the glint return from optics which can be detected from a scanning system placed on a combat vehicle looking for optical threats (Sjöqvist *et al.*, 2016). One of the problems in these types of systems is to discriminate between real targets like optical sights from other strong reflectors like traffic signs, etc. One technique to overcome this might be to use high resolution range (HRR) profiling based on photon counting (Sjöqvist *et al.*, 2013). See Figs. 20 and 21.

Atmospheric Profiling

One property for a profiling laser may also be to probe the atmospheric transmission, for example, along slant paths. Probing of the density of aerosols and thus the attenuation can be done using the backscatter signal from the atmosphere. Cloud mapping including their height distribution and density (up to a certain level) are also relevant for this type of operation. As an example it is interesting to probe the atmosphere in the dark (e.g., from a ship) when visual references disappear. Atmospheric laser radar or lidar is well investigated in the research community (Molebny *et al.* 2016; Fujii and Fukuchi, 2005; Steinvall *et al.*, 2015b).

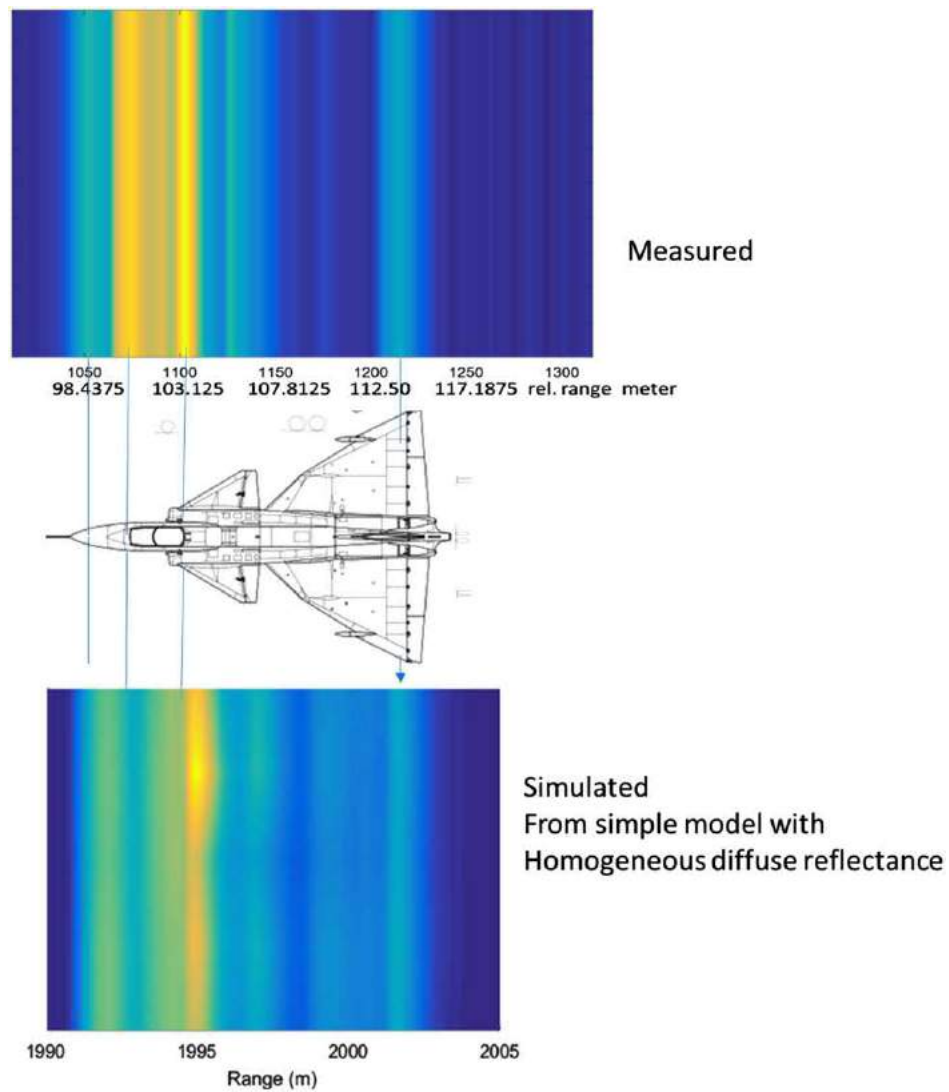


Fig. 14 Comparison between the measured and simulated range profiles in the head on direction for the veteran Viggen aircraft. Images FOI.

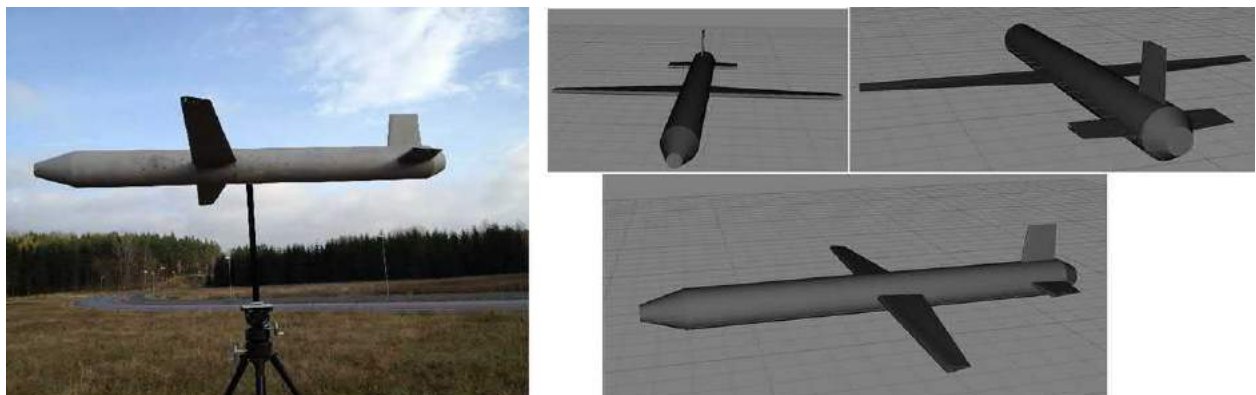


Fig. 15 Left: unmanned aerial vehicle (UAV) mockup made from sandblasted aluminum. Length 3450 mm, diameter 300 mm wing span 3000 mm. The distance from nose to the large wings is 1300 mm. Right: a CAD model for laser radar modeling. Reproduced from Steinvall, O., Tuldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

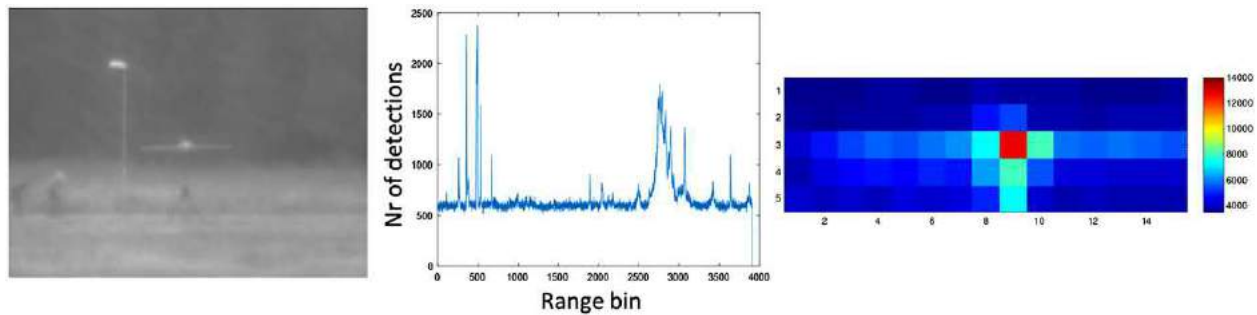


Fig. 16 Measurement with the photon-counting system on the UAV mock-up at 1.3 km distance. In the middle a waveform contains the target and terrain background. Right shows the resulting intensity of each pixel (approximately 15 cm square) in the photon-counting system when it was scanned over the target. Reproduced from Steinvall, O., Tuldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

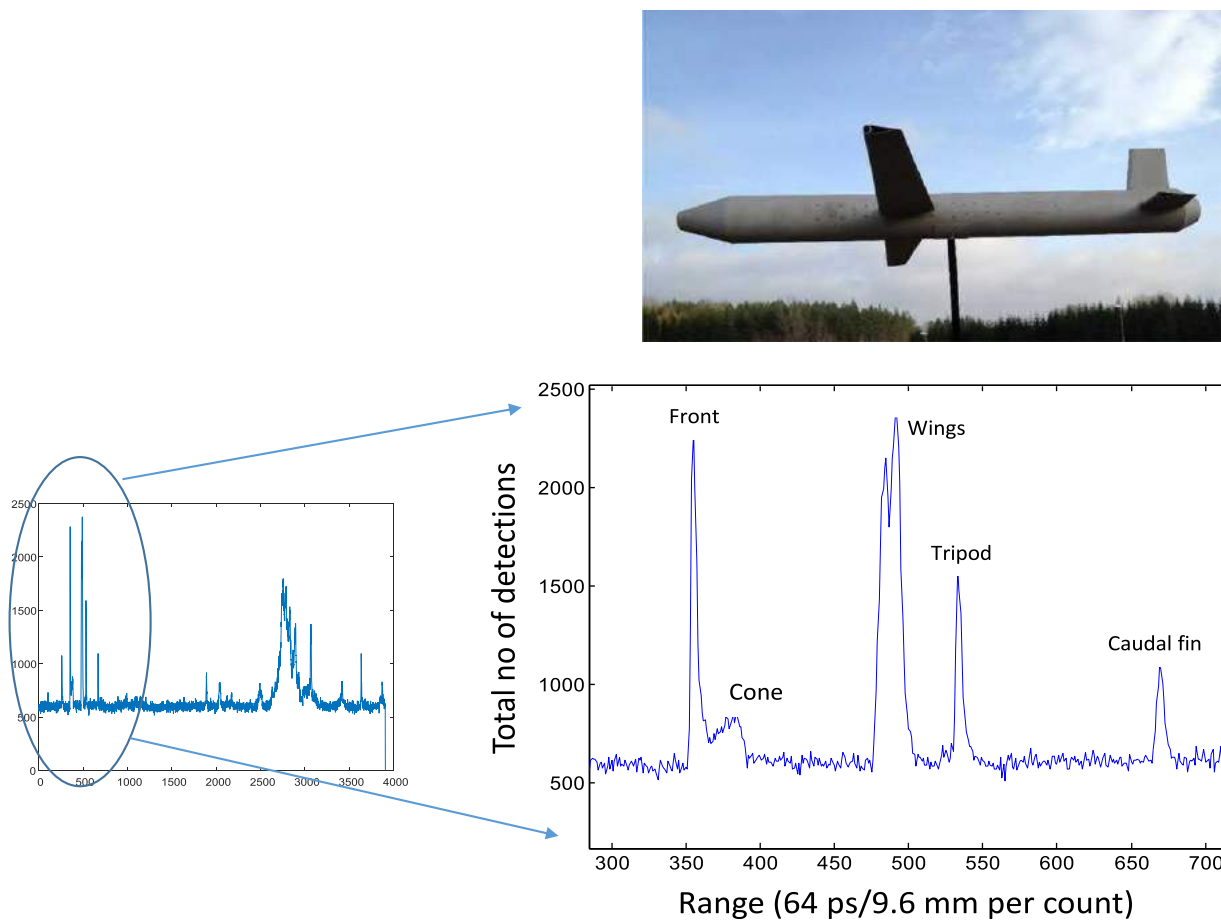


Fig. 17 Summed range waveforms from all 3×15 pixels above the target (UAV). The rear pulses belong to the terrain while the first part is enlarged and shows the range structure of the UAV with very high precision. The wings provide two close echoes because the UAV was not exactly perpendicular to the beam. Reproduced from Steinvall, O., Tuldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

Examples of Tomography Based on Laser Profiling

The aim of reflective tomography is to estimate object surface features based on a set of reflective projections that are measured in angular increments around the object either by rotating the sensor or the object. The basics of laser based reflective tomography are well described by [Knight et al. \(1989b\)](#).

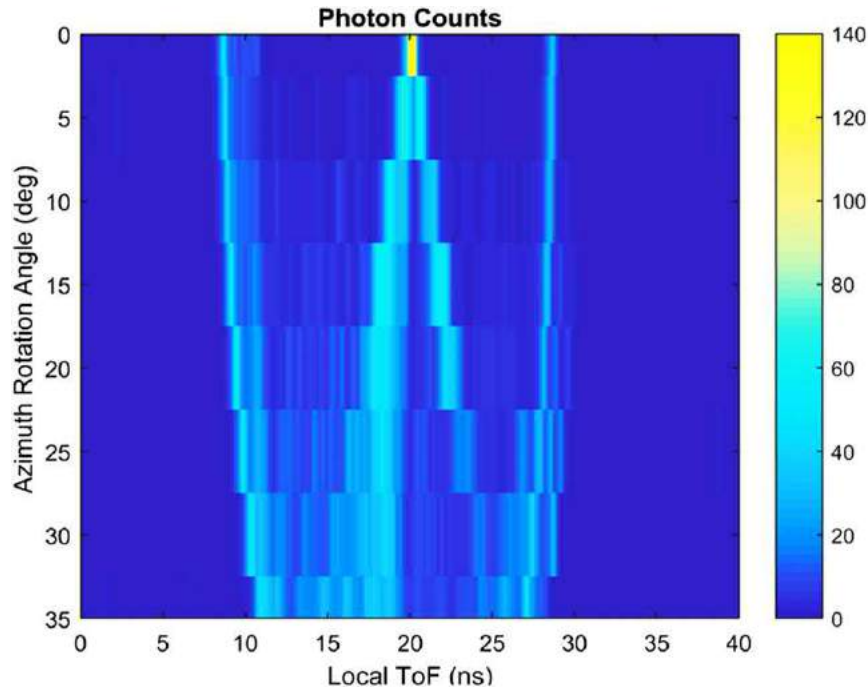


Fig. 18 Illustration of simulated waveforms over the target. The number of detected photons are shown in color scale, and the simulated target aspect angles are presented for the interval from 0 to 35 degrees in azimuth (in steps of 5 degrees). Reproduced from Steinvall, O., Tulldahl, M. 2016. Laser range profiling for small target recognition. *Optical Engineering* 56 (3), 031206.

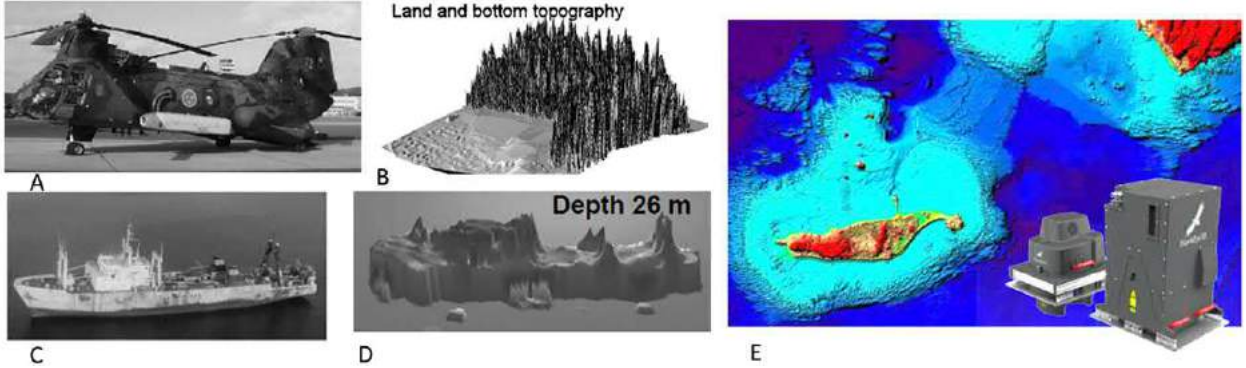


Fig. 19 (A) The original Hawk Eye – a system which developed further can combine topographic mapping on land and water depth sounding (now made by AHAB Leica Geosystems). The figure (B) shows one example of real data. Below (C and D): a ship before and right after it sunk at 26 m depth. Images AHAB, FOI and the Swedish Maritime Administration. (E) Example of a Hawk Eye III, an airborne multi-sensor deep water bathymetric and topographic LiDAR from Leica Geosystems. The green to blue color is indicative of the water depth.

Laser tomography using high range resolution profiling based on TCSPC enables small models mimicking real targets to be studied in detail. Examples are given in Steinvall *et al.* (2014), Henriksson *et al.* (2012a,b), Sjöqvist *et al.* (2014), and Steinvall *et al.* (2012b). The principle is depicted in Fig. 21.

Based on the scheme for filtered backprojection (FBP) using the inverse Radon transform, the tomographic image can be reconstructed. Mathematically, the 2D object can be expressed as (Henriksson *et al.*, 2012a,b):

$$g(x, y) = \sum_{i=1}^m q(z \cos \theta_i + x \sin \theta_i) \Delta \theta \quad (3)$$

where

$$q(r, \theta) = \mathcal{F}^{-1}(f \mathcal{F}(p(r, \theta))) \quad (4)$$

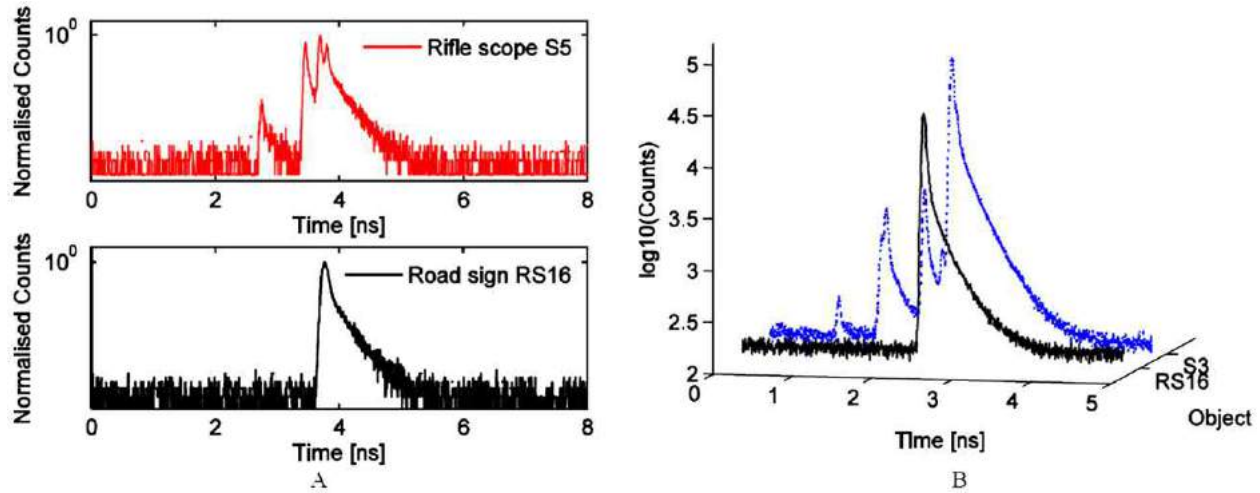


Fig. 20 (A) Example showing range-profile signal from a rifle scope S5 (upper) and a road sign RS19 (lower) at 750 m distance and 100 ms integration time. (B) Comparison between lineshape of a rifle scope (S3) and a road sign (RS16). Reproduced from Sjöqvist, L., Allard, L., Henriksson, M., Jonsson, P., Pettersson, M., 2013. Target discrimination strategies in optics detection. Proceedings of SPIE 8898, 88980K.

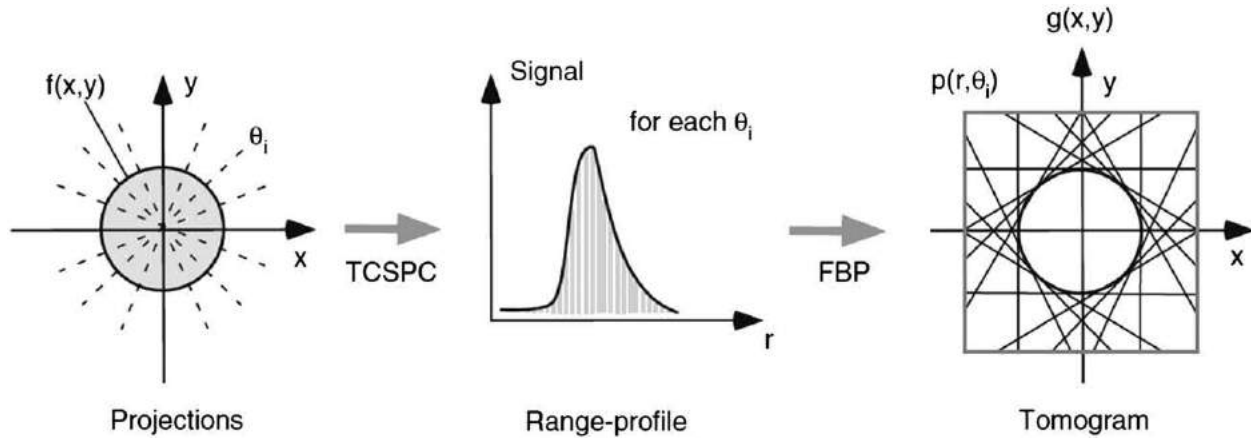


Fig. 21 Schematic showing the principle of reflectance TCSPC tomography with projections and range profiles for each angle of incidence θ_i , and the corresponding tomogram after the filtered backprojection (FBP) transform. Reproduced from Sjöqvist, L., Henriksson, M., Jonsson, P., Steinvall, O., 2014. Time-correlated single-photon counting range profiling and reflectance tomographic imaging. Advanced Optical Technologies 3, 187–197.

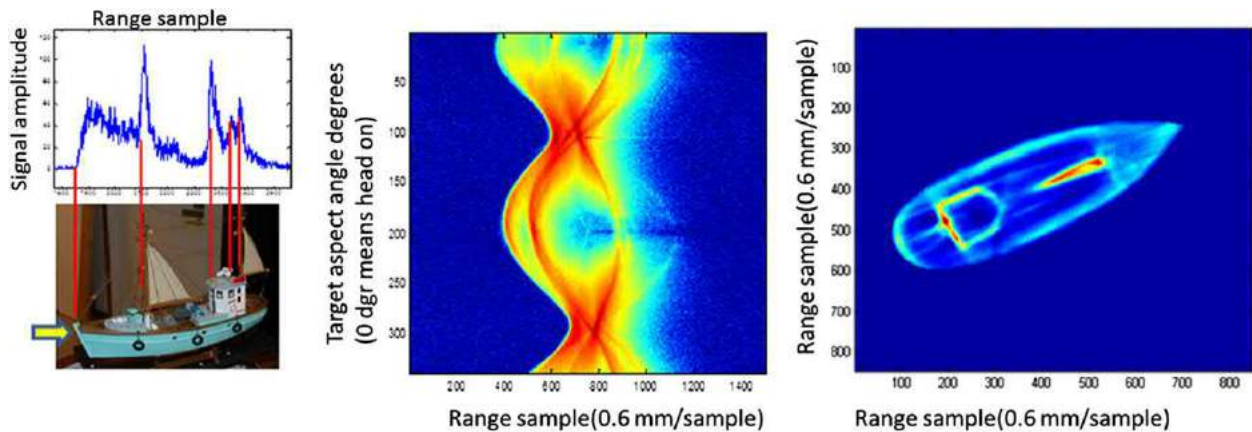


Fig. 22 Example of tomography of a small model boat, 43 cm in length. Our laboratory TCSPC lidar system had a time response of 54 ps corresponding to a range resolution of 0.8 cm. Reproduced from Steinvall, O., Chevalier, T., Grönwall, C., 2014. Simulation and modeling of laser range profiling and imaging of small surface vessels. Optical Engineering 53 (2), 029801.

and $\mathcal{F}$, $\mathcal{F}^{-1}$ denote the forward and inverse Fourier transform, respectively, r is the distance along the measurement direction with zero at the rotation center of the studied object (not a polar coordinate), $p(r, \theta_i)$ is the range profile for each rotation angle θ_i , and f is a filter function. Here it was assumed that the rotation axis is parallel to the y -axis. The filter function, f , was chosen to be a generalized ramp function according to:

$$f(\omega) = |\omega| \cdot e^{-\xi|\omega|^a} \quad 0 \leq |\omega| \leq \pi \quad (5)$$

and $\xi = |1/\omega_c|^a$ with ω_c defining the cut-off frequency and a is an adjustable parameter. Typical values used in our work with a photon counting systems (Sjöqvist *et al.*, 2014) to reduce the high-frequency components were $a=3.4$ and $\omega_c=\pi/4$. The values need to be optimized based on the IRF of the measurement system.

FOI laboratory TCSPC ladar system had a time response of 54 ps corresponding to a range resolution of 0.8 cm. Realistic range resolution from conventional high resolution direct detection range finders of laser radars is in the region 0.2–1 m, which correspond to scale factors between 25 and 125. The extreme range resolution of TCSPC ladar makes it possible to perform experiments using laboratory scale models relevant for scaling to data expected with linear-mode laser radar systems for the identification of large-scale objects, allowing testing of possible system performance in a simple and inexpensive way (Steinvall *et al.*, 2014).

Fig. 22 shows examples from the measurements using the 43-cm long fishing boat as a target. Left is an example of the observed waveform looking straight at the fore of the boat, middle, and range profile for 341 degree in steps of 1 degree. Right is a tomographic image obtained by using the Radon transform to reconstruct the shape of the boat. The measurement range in the laboratory was about 40 m and the beam covered the whole FOV of the receiver, which was 10.2 mrad ($1/e^2$) corresponding to 40 cm at the target.

There are many factors affecting the quality of the tomographic image, for example, strong reflections which cause linear features stretching through the reconstructed image. Jin *et al.* has investigated the BRDF effects in ladar-based reflection tomography (Jin and Levine, 2009). The influence from these artefacts can be partly removed by using the “convex hull” method where the first response (opaque surface) at each angle from the object is used to create an image mask outside which the response is known to be zero (Sjöqvist *et al.*, 2014). Critical issues regarding the quality of the reconstructed tomogram beside SNR include defining the center of rotation, applied angular resolution, and the used angular projection sector. The way of presenting the tomogram also has a large influence on the image quality. Often the log intensity including a threshold give a good quality (see Fig. 23) paired with a type of filtering like the “convex hull” mentioned above.

So far only laboratory measurements examples have been presented. There are, however, also several long-range demonstrations reported in the literature. The use of 1D Doppler-resolved projections to form a 2D tomographic image of a rotating object was demonstrated using the MIT Lincoln Labs 10.6 μm Firepond laser radar. With a model of a Thor-Delta rocket target at 5.4 km ground range, Doppler-time-intensity was demonstrated (Knight *et al.*, 1989b). Images of satellites at ranges up to 1500 km were also collected by the Firepond (Gschwendtner and Keicher, 2000) (Fig. 24).

Matson and Mosley (2001) report on the first satellite feature reconstruction, by use of range-resolved reflective tomography techniques collected on an orbiting satellite. The reconstructed features were two retroreflectors mounted on a satellite. The data were collected with a coherent laser radar system located at the Maui Space Surveillance Site in Maui, Hawaii. The laser system called HI-CLASS was a pulse coherent laser radar emitting 30 J/pulse at 30 Hz repetition rate. The acquisition waveform was a gain-switched pulse with a duration of the order of 10 ms, whereas the imaging waveform is a mode-locked version of the acquisition waveform. The mode-locked waveform was a series of micropulses modulated by the gain-switched pulse waveform in which the micropulses are less than 1.5 ns in width and are separated from each other by 40 ns. Fig. 25 shows some results. A key step in the projection creation process was to align the projections to the center of rotation of the satellite.

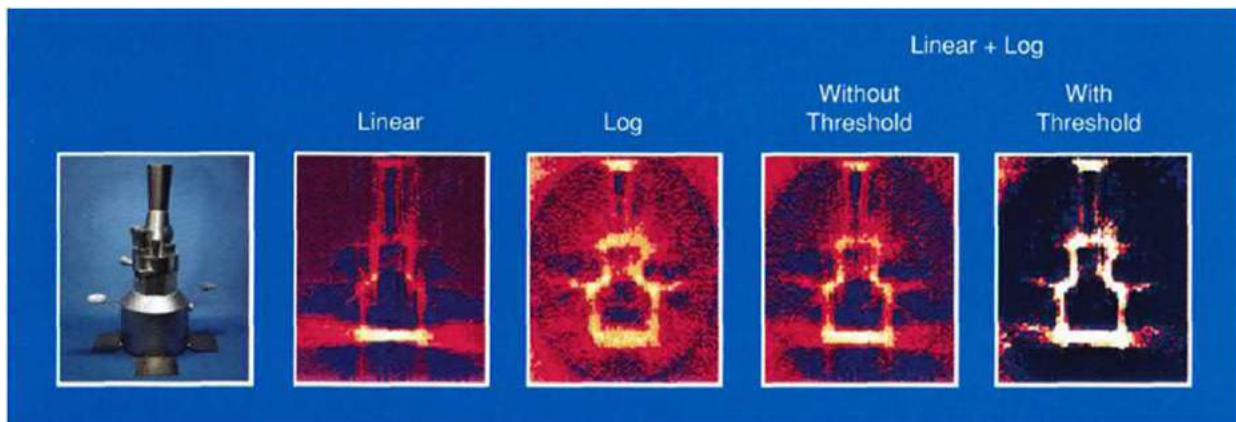


Fig. 23 Photograph and reconstructions of a model of a DSP satellite. Note the better image quality using log intensity and thresholding. Reproduced from Knight, F.K., Kulkarni, S.R. Marino, R.M., Parker, J.K., 1989. Tomographic techniques applied to laser radar reflective measurements. Lincoln Laboratory Journal 2 (2), 143–158.

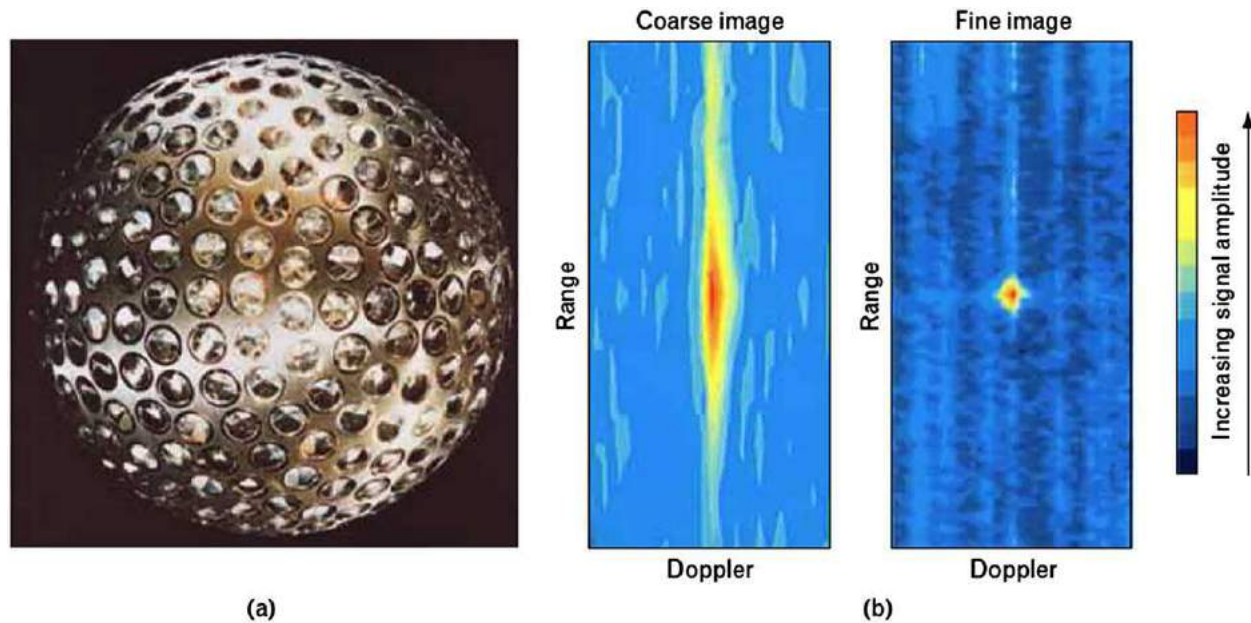


Fig. 24 (a) The Laser Geodynamics Satellite (LAGEOS), a 60-cm aluminum sphere with a brass core. The satellite has a total mass of 406 kg and is in a near circular orbit at an altitude of approximately 5900 km. LAGEOS has 426 silica retroreflectors and 4 germanium retroreflectors to serve as a laser-radar target. (b) Range-Doppler image of the LAGEOS satellite collected by the wideband CO_2 laser radar at Firepond. The coarse image was made with a signal bandwidth of 150 MHz, while the fine image was made with a bandwidth of 1 GHz. Doppler velocity resolution is approximately 30 cm/s. Color in the image represents relative signal amplitude. The range and Doppler sidelobes are better than 20 dB below the main lobe. The range-Doppler images are the point-target responses of the imaging laser radar produced by different bandwidths. Figure text and figure from Gschwendtner, A.B., Keicher, W.E., 2000. Development of coherent laser radar at Lincoln Laboratory. *Lincoln Laboratory Journal* 12 (2), 383–396.

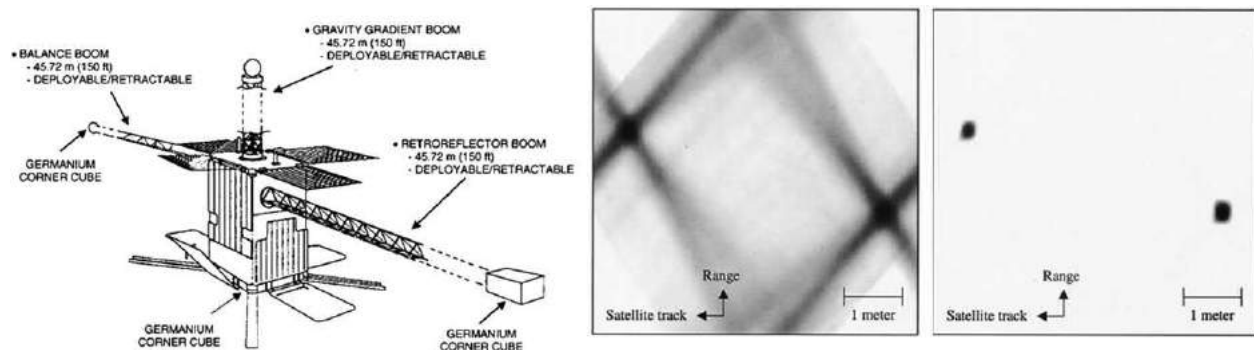


Fig. 25 Left: representation of the LACE satellite. Middle: tomographic reconstruction of the two retroreflectors at the satellites arms. Right: thresholded image to remove limited-view artifacts. Reproduced from Matson, C.L., Mosley, D.E., 2001. Reflective tomography reconstruction of satellite features – field results. *Applied Optics* 40 (14), 2290–2296.

Another long-range example is given by Murray *et al.* (2010). They use pulse coded lidar waveforms that have desirable side-lobe-free autocorrelation properties. In opposite to short high pulse energy techniques this will enable high bandwidth ($> \text{GHz}$), low-peak power, compact and efficient fiber laser transmitters and photonic receivers to be used in compact systems. See Fig. 26 for an example of the result.

A theoretical analysis of including turbulence effects for lidar reflective tomography imaging for space objects is presented by Qu *et al.* (2011). The development of laser range profile, Doppler spectra, and range-resolved Doppler imaging technologies are reviewed in the paper (Wu *et al.*, 2014).

Reflective tomography can also include making 3D images from a series of 2D images. Different techniques have been published (Xinwei *et al.*, 2016; Andersson, 2006; Laurenzis and Woiselle, 2014) for this reconstruction but they all rely on evaluating the range from the pixel intensity increase or decrease with time. The convolution of the camera gate function and the laser pulse shape form this time dependence to be used in the 3D reconstruction. The 3D quality depends on the single frame SNR and the number of frames used for the reconstruction. In Fig. 27 we illustrated how thermal and laser imaging can be combined for target detection and classification using a 2D to 3D tomographic technique.

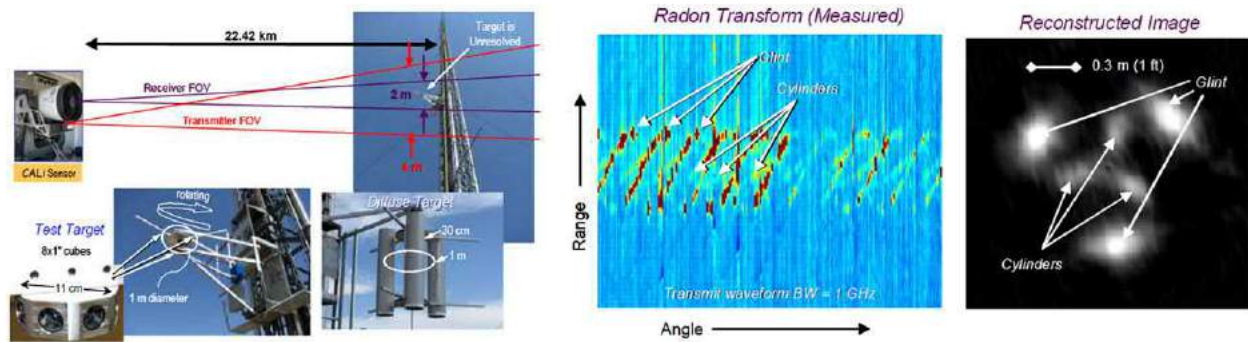


Fig. 26 Left: long-range field test configuration. An unresolved rotating 1 m diameter target consisting of both retro-reflective and diffuse scattering structures is placed on a tower 22 km from the laser platform. The ladar field of view (FOV) is 2 m and the transmitter FOV is 4 m at range. Right: first tomographic image reconstruction of an unresolved target located 22 km from the ladar sensor. Both the retro-reflective (glint) and diffuse cylindrical targets appear in the image. Reproduced from Murray, J., Triscari, J., Fetzer, G., *et al.*, 2010. Tomographic lidar. In: Applications of Lasers for Sensing and Free Space Communications, San Diego, CA, OSA Conference Paper.

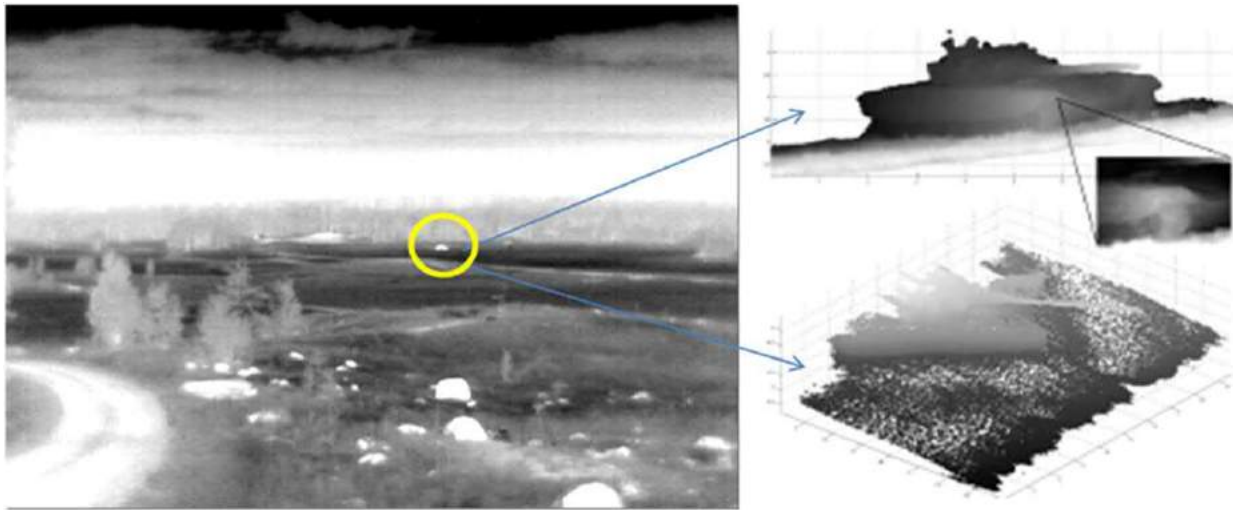


Fig. 27 Illustration of target detection with a thermal imaging camera and the subsequent target classification based on 3D target generation from a series of sliding gates over the target. Reproduced from Andersson, P., 2006. Long range 3D imaging using range gated laser radar images. Optical Engineering 45 (3), 034301–034310.

Signal Processing Techniques for Range Profiling and Tomography

In the microwave radar field signal processing techniques for target classification have been studied more extensively than for laser radar. This should motivate a closer look at the radar signal processing methods to see if they are applicable to laser radar while noting that there are difference in the type of signals they generate.

In the radar field techniques based on sequential classification tracking as well as model and feature-based methods have been developed. The sequential technique is based on successive target measurements performed over long time intervals, typically several seconds or minutes. During this period target maneuvering will be indicative of the type of target. This information from radar (or an optical sensor) will limit the search target search space and allow for other model or feature-based methods to be added for a more robust recognition.

Model-based methods employ physical and/or empirical target models to predict the measured target signals. The measured and predicted (model) signals are compared and the model which gives the smallest difference to the measured signal is chosen.

Feature-based methods do not use a model of the measured signals but rely on extracting a limited number of target features from the measurements. Subsequently, the extracted feature vectors are compared with feature vectors in a training set extracted from measurements.

It seems that model-based methods are most applicable to LRP because real measurements may be hard to achieve at least for military targets. On the other hand, real measured profiles may be collected over time using targets of opportunity.

HRR profiling for microwave radar has been investigated rather extensively (see e.g., Shaw *et al.* (2013), Li and Yang (1993) and references therein). One of the most critical issues about the range profile signature in the microwave domain is its extreme variability as a function of aspect angle. As little as a 0.1-degree aspect angle change can cause such a severe change that the signature is no longer recognizable. This is understood as the radar signal is composed of a number of strong scatter center contribution and a small change of the scattering center positions to a quarter of a wavelength will completely change the phase of the return. This fact motivated that radar signal processing of HRR often contains statistical modeling to capture these variations.

The most direct processing technique for range profiling may be correlation. In this method, the profile is convolved with a reference profile from the model library. It is convenient to normalize the waveforms to the highest peak for example. The method has been applied to real radar HRR data from 24 aircraft with (Hudson and Psaltis, 1993) success provided aspect information is

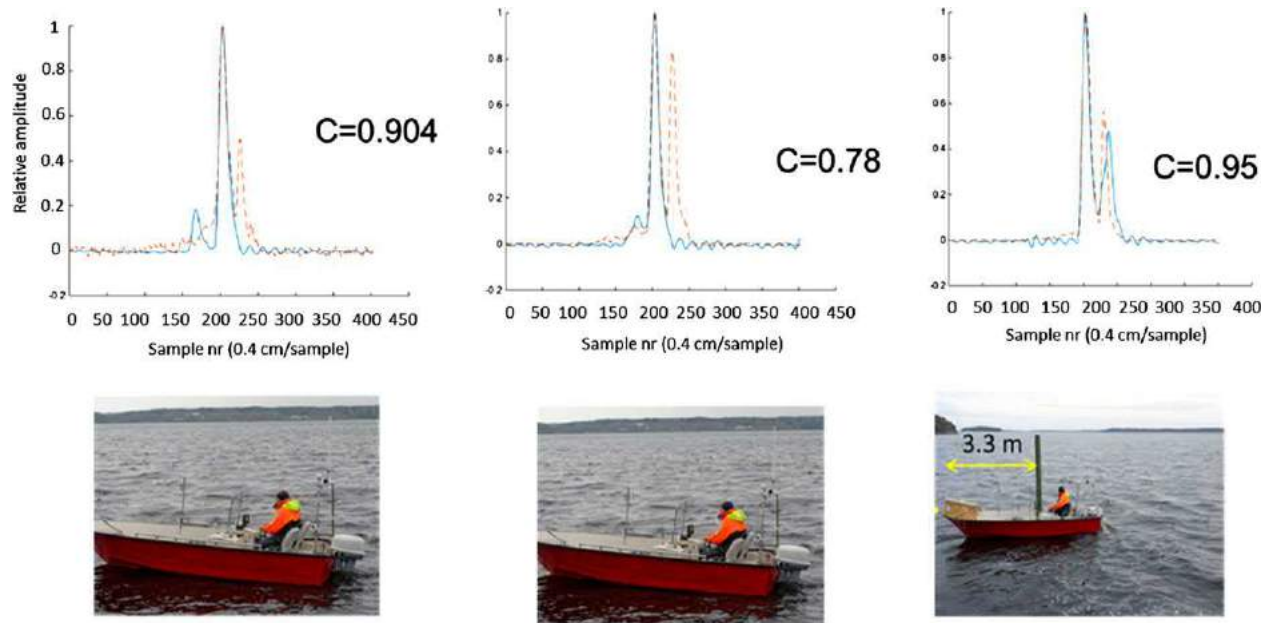


Fig. 28 Example of measured (full line) and simulated (dotted) waveforms together with the correlation coefficients obtained after matching the largest peak. One can observe how the waveforms both match and mismatch probably due to different angular boat positions and beam jitter. Reproduced from Steinvall, O., *et al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 964901.

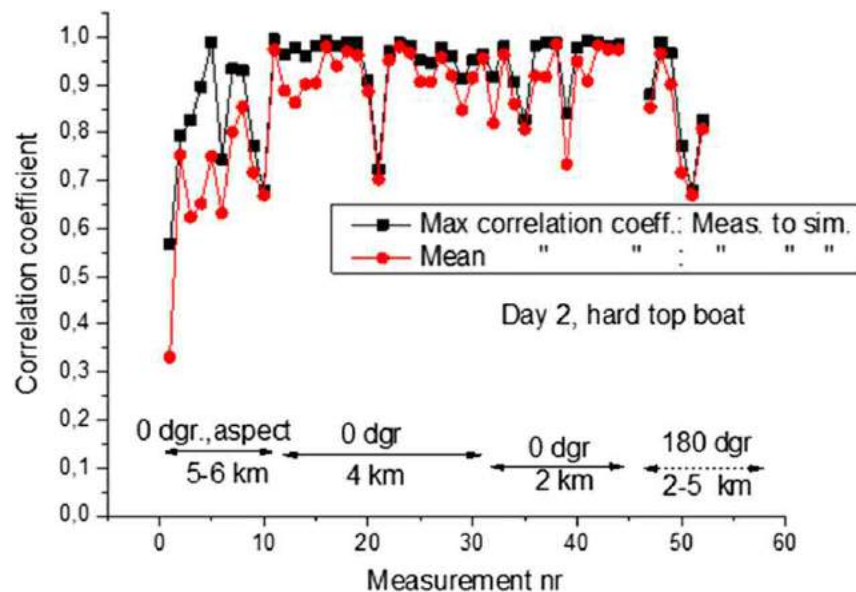


Fig. 29 Correlation coefficient between measured and simulated waveforms from boat B at 0- and 180-degree aspect. Different target ranges are indicated. Reproduced from Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

used, and provided identifications are based on sufficient number of profiles so as to decrease errors due to speckle fluctuations. Li and Yang (1993) used correlation to investigate target classification using experimental data from physical aircraft model with sizes in the 50–80 cm range measured in the frequency range 6–16 GHz.

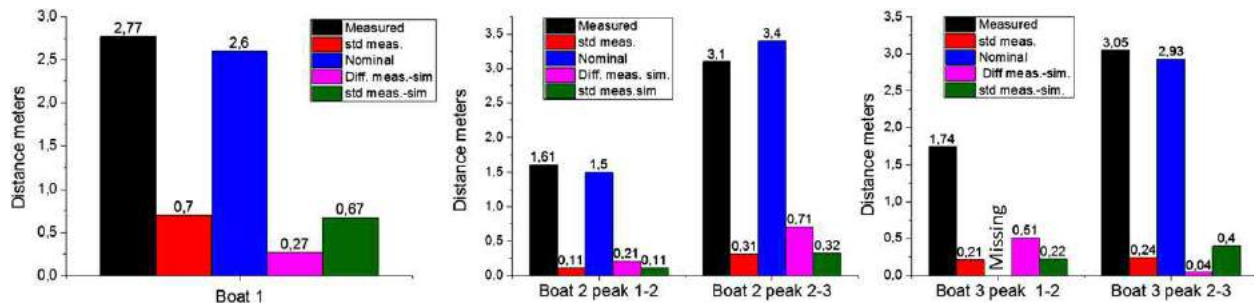


Fig. 30 The measured distances vs. the nominal together with the standard deviations of the measurements and the difference between measurements and simulations. The aspect angle was 0 degree (boats approaching toward the sensor). Reproduced from Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. *Optical Engineering* 56 (3), 031206.

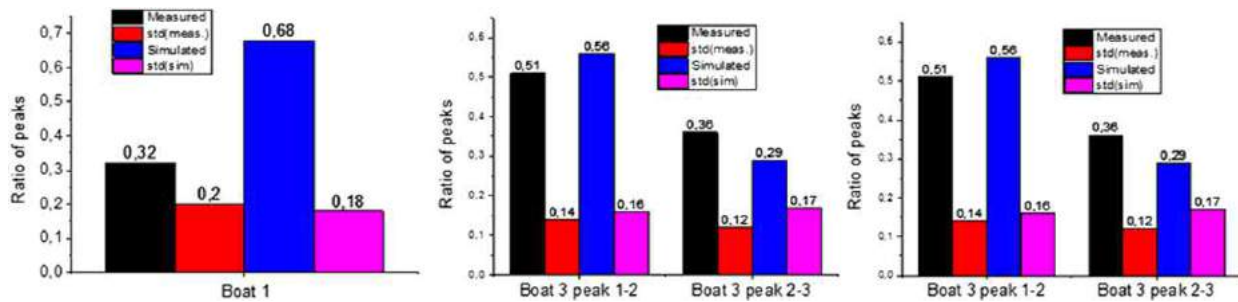


Fig. 31 The ratio of peaks in the waveform for the unmodified boats. Boat 1=A, boat 2=B and boat 3=C. Reproduced from Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. *Optical Engineering* 56 (3), 031206.

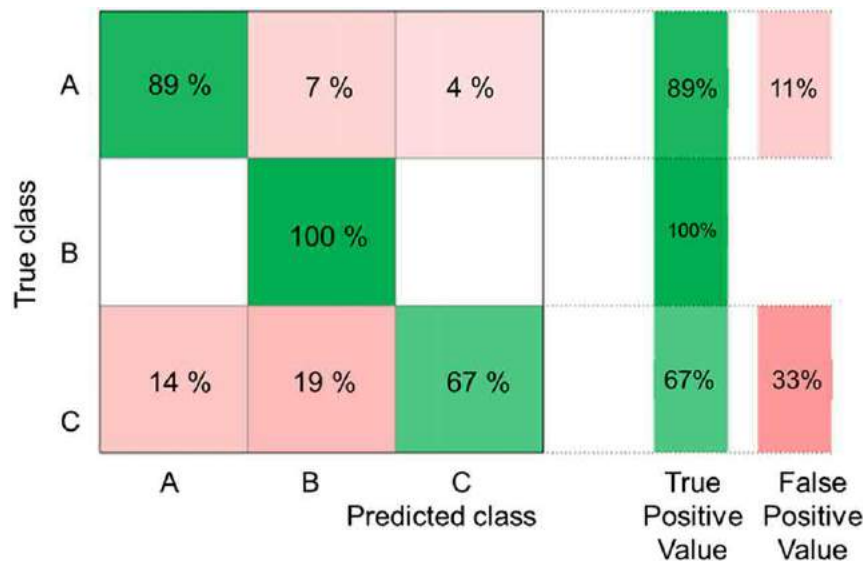


Fig. 32 Confusion matrix for classification using distance and ratio of the three peaks in each range waveform. Right: results for true positive and false negative. The classification accuracy for a decision tree classifier was 89%. Reproduced from Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. *Optical Engineering* 56 (3), 031206.

For radar it has been argued that the magnitude of the HRR profiles is too uncertain to use for target recognition, and one should instead use the positions of peaks in the HRR profile. In Zwart (2004), Zwart applies this approach using the same in-flight measured radar data of five aircraft with promising results. Furthermore, Eigen template based recognition have been investigated in a number of publications, see Xinwei *et al.* (2016) and references therein.

The laser range waveforms are less variable with respect to small aspect changes as the reflection is more or less diffuse in character. Thus, a more parameter-based processing technique seems appropriate to investigate beside correlation.

The correlation for LRP methods has also been tested on small boats (Steinvall *et al.*, 2015a). When the main peak matches a correlation coefficient as exemplified in Fig. 28 was achieved. Note that the correlation value can be rather high even if other the peaks mismatches. In Fig. 29 we exemplify the correlation coefficient for all measurements against the boat B of Fig. 8 and at 0-degree aspect (approaching course) and a few at 180-degree aspect. Most of the 0-degree aspect correlation data show high values, above 0.9 for both maximum and mean values. Lower values can most probably be explained by poor pointing and beam jitter.

As seen from the correlation examples (also see Steinvall *et al.* (2015a)) the discrimination capability only based on correlation analysis may not be sufficient for a robust recognition. It is therefore of interest to add other parameters into the decision process. These can be related to the distance between peaks and the ratio of peak amplitudes.

The absolute distances as shown in Fig. 30 are derived from the peak separation for both simulated and collected range profiles for 0-degree aspect. In general, we can see that the correspondence between the observed and simulated distances is rather good.

The mean peak amplitude ratios for the first and second pair of peaks for every waveforms are shown in Fig. 31, estimated from all measurements at 0 degree and all target types without modifications with the extra box and the pipe. From this we can see that the difference between measured and simulated ratios is rather good irrespective of the lack of detailed reflectivity information and information on how the beam was actually pointed.

Using all available measures that is two distances between peaks and the corresponding ratios of peaks both from measurements and simulations we arrive at a classification accuracy of 89% (Fig. 32). The true positive classification probabilities are now even better, namely 89% for boat A, 100 for boat B, and 67% for boat C with false rates as 11, 0 and 33%, respectively, for A, B, and C. The corresponding values for the positive prediction and false discovery rates are 89, 88, 83% and 11, 12, and 7% respectively.

Heuvel *et al.* (van den Heuvel *et al.*, 2009a) describes a full scenario processing LRP for target classification using a likelihood test. Two aircraft, of which the one to classify is assumed to be a F16, are approaching each other with a speed of about 0.9 Mach at high altitude. The simulated aircraft profiles belong to F117, F15, and F16 obtained from 3D models. An LRP sensor with 30 cm range resolution is mounted on the observation aircraft. Noise has been added so that the signal at distances above 10 km is dominated by noise. The two aircraft are approaching at an angle of about 90 degrees. Atmospheric effects are neglected which is motivated that the scenario is at high altitude. Detection and tracking of the aircraft is based on radar or IRST. Fig. 33 shows the scenario and the result of the Bayesian identification. The F16 is identified at shorter ranges with a maximum probability of 95%. In Fig. 34 the time axis corresponds to absolute ranges between the aircraft and the relative range within the profile. In this figure 250 s correspond to 30 km range and at 320 s the range between the two aircraft is at its minimum, about 5.6 km. This

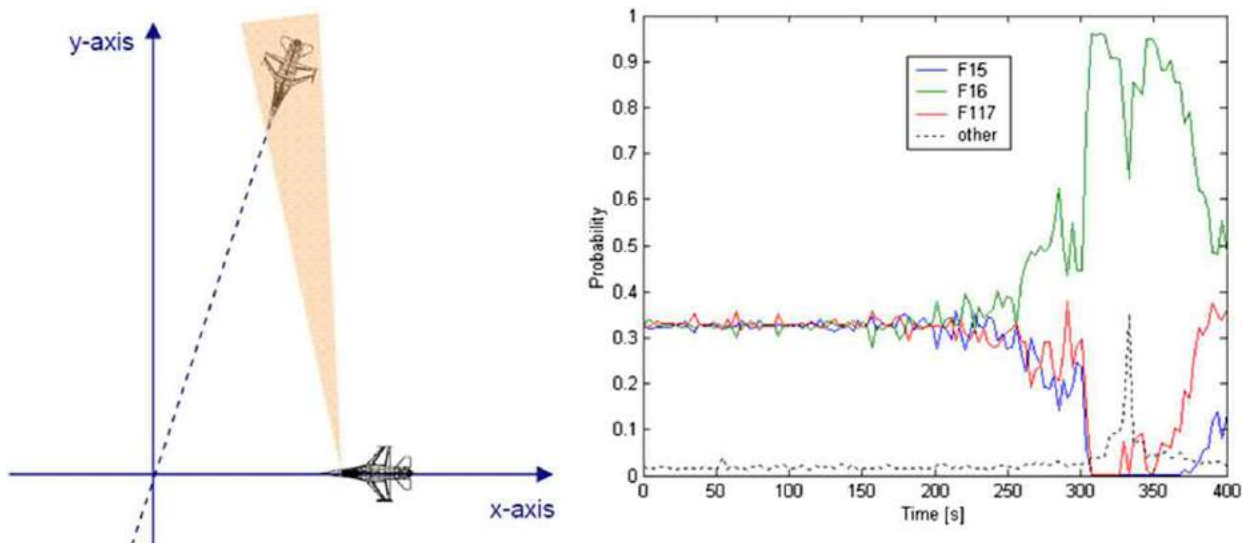


Fig. 33 Left: the scenario with two aircraft, of which the one to classify is assumed to be a F16, are approaching each other with a speed of about 0.9 Mach at high altitude. Right: the Bayes probabilities for an approaching F16 in the scenario. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

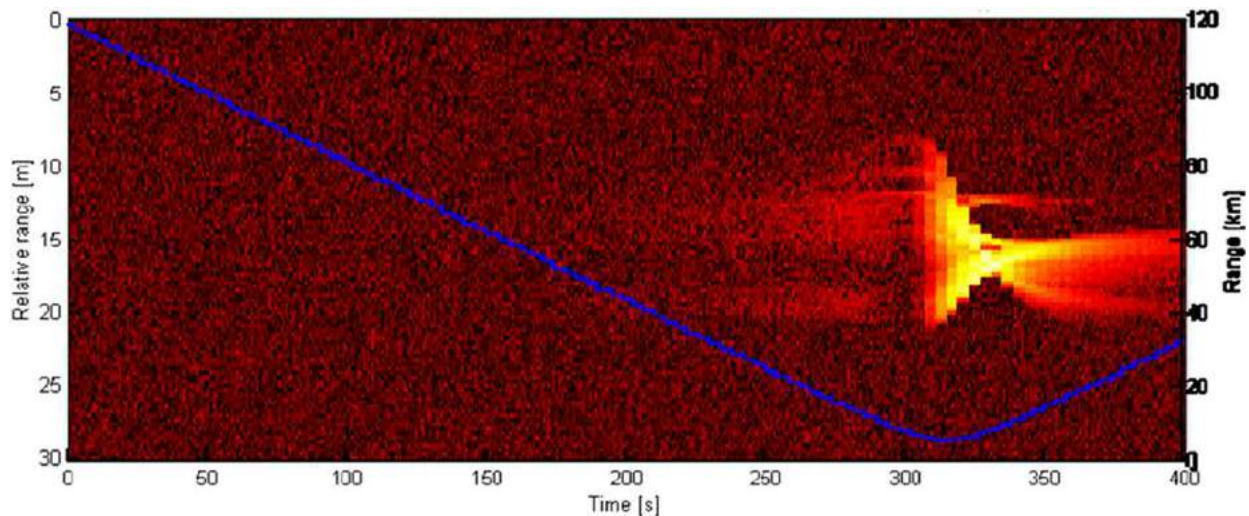


Fig. 34 Simulated laser range profiles for an approaching F16 according to the scenario. Every image column represents a single profile on a linear intensity scale. The blue line indicates the distance corresponding to the right axis. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. *Proceedings of SPIE* 7323, 73230Y.

scenario is of high interest as it contains a dynamic change of aspect angle as well as range (and thereby a changing SNR of the waveforms).

Finally, there are a lot of other signal processing techniques which are suitable for target recognition using range profiles. These include machine learning and deep learning involving neural networks, support vector machines and others.

Conclusions

Although not by far as investigated as its microwave counterpart, LRP has some clear advantages to act as a complement or substitute to radar or IR imaging for target classification.

Compared to traditional angle-angle optical sensors LRP can offer range information along the target down to the sub-centimeter level, if needed. While traditional optical imaging needs large apertures, and often turbulence compensation, for long-range target classification, LRP is relatively insensitive to deblurring turbulence and scattering. It is also relative insensitive to first-order pointing error as long some of laser lobe covers the target. LRP can be a natural extension of present laser range finders. The maximum range at which laser range finder type of systems can produce target classification can extend tens of kilometers.

A laser profiling system can complement a microwave radar where it offers additional target classification properties, for example, for small targets using high resolution techniques. It may also back up a radar that is being jammed. The laser in a profiling system may serve as an illuminator for an imaging gated sensor and thus provide a robust recognition containing both angle-angle imaging and range profiling. Lasers can also provide accurate tracking using both imaging and quadrant detector receivers.

Tomography has been shown to be useful to image rotating objects. In general, however, the application of tomography to defense applications is probably limited due to a number of reasons. Among these we note alignment of the profiles may be difficult as well as obtaining many profiles over a sufficiently large angular sector to give a good image quality (Henriksson *et al.*, 2012a,b).

See also: Very High Range Resolution Lidars

References

- Andressen, C., Anthony, D., DaMommio, T., *et al.*, 2005. Tower test results for an imaging LADAR seeker. *Proceedings of SPIE* 5791, 70–82.
- Andersson, P., 2006. Long range 3D imaging using range gated laser radar images. *Optical Engineering* 45 (3), 034301–034310.
- Armbruster, W., Hammer, M., 2012. Maritime target identification in flash-ladar imagery. *Proceedings of SPIE* 8391, 83910C.
- Chevalier, T., 2011. A ray tracing based model for 3D ladar systems. In: *Proceedings of the International Conference on Computer Graphics Theory and Applications*, pp. 39–48.
- Dierking, M.P., Heitkamp, F., Barnes L., 1998. High temporal resolution laser radar tomography for long range target identification. In: *OSA Signal Synthesis & Reconstruction Conference*.
- Fujii, T., Fukuchi, T., 2005. *Laser Remote Sensing*. Boca Raton, FL: CRC Press.
- Gschwendtner, A.B., Keicher, W.E., 2000. Development of Coherent Laser Radar at Lincoln Laboratory. *Lincoln Laboratory Journal* 12 (2), 383–396.

- Gustavson, R.L., Davis, T.E., 1992. Diode laser radar for low-cost weapon guidance. *Proceedings of SPIE* 1633, 21–32.
- Henriksson, M., Olofsson, T., Grönwall, C., Brännlund, C., Sjöqvist, L., 2012a. Optical reflectance tomography using TCSPC laser radar. *Proceedings of SPIE* 8542, 85420E.
- Henriksson, M., Olofsson, T., Grönwall, C., Brännlund, C., Sjöqvist, L., 2012b. Optical reflectance tomography using TCSPC laser radar. *Proceedings of SPIE* 8542, 85420 E.
- Hudson, S., Psaltis, D., 1993. Correlation filters for aircraft identification from radar range profiles. *IEEE Transactions on Aerospace and Electronic Systems* 29 (3), 741–748.
- Hutt, D.L., Thériault, J.-M., Laroche, V.G., Mathieu, P., Bonnier, D., 1994. Estimating atmospheric extinction for eyesafe laser rangefinders. *Optical Engineering* 33, 3762–3773.
- Jin, X., Levine, R., 2009. Bidirectional reflectance distribution function effects in ladar-based reflection tomography. *Applied Optics* 48, 4191–4200.
- Jonsson, P., Hedborg, J., Henriksson, M., Sjöqvist, L., 2015. *Proceedings of SPIE* 9649, 964905.
- Knight, F.K., Click, D.I., Ryan-Howard, D.P., *et al.*, 1989a. Laser radar reflective tomography utilizing a streak camera for precise range resolution. *Applied Optics* 28, 2196–2198.
- Knight, F.K., Kulkarni, S.R., Marino, R.M., Parker, J.K., 1989b. Tomographic techniques applied to laser radar reflective measurements. *Lincoln Laboratory Journal* 2 (2), 143–158.
- Kunz, G.J., Bekman, H.P.T., Benoist, K.W., *et al.*, 2005. Detection of small targets in a marine environment using laser radar. *Proceedings of SPIE* 5885, 5885F1.
- Ladar Seeker/ATA Algorithm Captive Flight Test Results, 2017. Available at: http://www.fas.org/man/dod101/sys/smart/docs/locas_Industry_Day/sld013.htm (accessed 08.02.17).
- Lasche, J.B., Matson, C.L., Ford, S.D., *et al.*, 2009. Reflective tomography for imaging satellites: Experimental results. In: *Proceedings of SPIE*, 3815. pp. 178–188.
- Laurenzis, M., Woiselle, A., 2014. Laser gated-viewing advanced range imaging methods using compressed sensing and coding of range-gates. *Optical Engineering* 53 (5), 053106.
- Li, H.J., Yang, S.H., 1993. Using range profiles as feature vectors to identify aerospace objects. *IEEE Transactions on Antennas and Propagation* 41 (3), 261–268.
- Lutzmann, P., Göhr, B., van Putten, F., Hill, C.A., 2011. Laser vibration sensing: overview and applications. *Proceedings of SPIE* 8186, 818602.
- Matson, C.L., Mosley, D.E., 2001. Reflective tomography reconstruction of satellite features – field results. *Applied Optics* 40 (14), 2290–2296.
- Molebny, V., McManamon, P., Steinvall, O., Kobayashi, T., Chen, W., 2016. Laser radar: Historical perspective – from the East to the West. *Optical Engineering* 56 (3), 031220.
- Murray, J., Triscari, J., Fetzer, G., *et al.*, 2010. Tomographic lidar. In: *OSA Conference Paper. Applications of Lasers for Sensing and Free Space Communications*, San Diego, CA.
- Pace, P., Steinvall, O., Chevalier, T., 2008. Automatic target classification using one dimensional laser profiling. Technical Memorandum DRDC, Ottawa, TN 2007-000.
- Parker, J.K., Craig, E.B., Klick, D.I., *et al.*, 1988. Reflective tomography: Images from range-resolved laser radar measurements. *Applied Optics* 27, 2642–2643.
- Qu, F., Hu, Y., Wang, D., 2011. Lidar reflective tomography imaging for space object. *Proceedings of SPIE* 8200, 820015.
- Schoemaker, R., Benoist, K., 2001. Characterisation of small targets in a maritime environment by means of laser range profiling. *Proceedings of SPIE* 8037, 803705.
- Shan, J., Toth, C.K., 2009. *Topographic Laser Ranging and Scanning – Principles and Processing*. Boca Raton, FL: CRC Press.
- Shaw, A.K., Paul, A.S., Williams, R., 2013. Eigen-template-based HRR-ATR with multi-look and time-recursion. *IEEE Transactions on Aerospace and Electronic Systems* 49 (4), 2369–2385.
- SICK-Sensor Intelligence, 2017. Designing transportation routes to be more efficient, safer and more environmentally friendly. Available at: <https://www.sick.com/de/en/industries/traffic/c/g285087> (accessed 08.02.17).
- Signal, 2006. Available at: <http://www.afcea.org/signal/articles/anmviewer.asp?a=1099&print=yes> (accessed 08.02.17).
- Sjöqvist, L., Allard, L., Henriksson, M., Jonsson, P., Pettersson, M., 2013. Target discrimination strategies in optics detection. *Proceedings of SPIE* 8898, 88980K.
- Sjöqvist, L., Allard, L., Pettersson, M., *et al.*, 2016. Optics detection and laser countermeasures on a combat vehicle. *Proceedings of SPIE* 9989, 998908.
- Sjöqvist, L., Henriksson, M., Jonsson, P., Steinvall, O., 2014. Time-correlated single-photon counting range profiling and reflectance tomographic imaging. *Advanced Optical Technologies* 3, 187–197.
- Steinvall, O., 2000. Effects of target shape and reflection on laser radar cross sections. *Applied Optics* 39, 4381–4391.
- Steinvall, O., 2009. Laser system range calculations and the Lambert W function. *Applied Optics* 48, B1–B7.
- Steinvall, O., Berglund, D., Allard, L., *et al.*, 2015a. Passive and active EO sensing of small surface vessels. *Proceedings of SPIE*, 964901.
- Steinvall, O., Chevalier, T., Grönwall, C., 2014. Simulation and modeling of laser range profiling and imaging of small surface vessels. *Optical Engineering* 53 (2), 029801.
- Steinvall, O., Elmqvist, M., Chevalier, T., Brännlund, C., 2012a. Measurement and modeling of laser range profiling of small maritime targets. *Proceedings of SPIE* 8542, 85420I.
- Steinvall, O., Elmqvist, M., Karlsson, K., Larsson, H., Axelsson, M., 2010. Laser imaging of small surface vessels and people at sea. *Proceedings of SPIE* 7684, 768417.
- Steinvall, O., Persson, R., Berglund, F., Gustafsson, F., Öhgren, J., 2015b. Using an eye-safe laser rangefinder to assist active and passive electrooptical sensor performance prediction in low visibility conditions. *Optical Engineering* 54 (7), 074103.
- Steinvall, O., Sjöqvist, L., Henriksson, M., 2012b. Photon counting ladar work at FOI, Sweden. *Proceedings of SPIE* 8375, 83750C.
- Steinvall, O., Tuldahl, M., 2016. Laser range profiling for small target recognition. *Optical Engineering* 56 (3), 031206.
- van den Heuvel, J.C., Pace, P., Steinvall, O., 2008. Laser opportunities in new naval missions. *Naval Forces* 29 (5), 46–50.
- van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. *Proceedings of SPIE* 7323, 73230Y.
- van den Heuvel, V., van Putten, F.J.M., Cohen, F.H., Kemp, R.A.W., Franssen, G.C., 2009b. Results from the Search-Lidar Demonstrator Project for detection of small sea-surface targets. *Proceedings of SPIE* 7482, 748205.
- Wu, P., Ming-Jun, W., Ke, X., Yan-jun, G., Yang, T., 2014. Laser radar range profile, Doppler spectra and range resolved Doppler imaging technologies for the target recognition. *Proceedings of SPIE* 9299, 929909.
- Xinwei, W., Liang, S., Pingshun, L., *et al.*, 2016. Range-intensity coding under triangular and trapezoidal correlation algorithms for 3D super-resolution range-gated imaging. *Proceedings of SPIE* 9988, 99880U.
- Zwart, J.P., 2004. Aircraft recognition from features extracted from measured and simulated radar range profiles. PhD Thesis, Universiteit van Amsterdam.

Micro-Lidars for Short Range Detection and Measurement

Vasyl V Molebny, Academy of Technological Sciences of Ukraine, Kiev, Ukraine

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The term micro-lidar has its origin from the Greek word *μικρό* (small) and the abbreviation of “light detection and ranging” meaning “detection and ranging with light.” Often, a similar abbreviation “ladar” is in use where the word “light” is substituted by “laser” (light amplification by stimulated emission of radiation). Another form is “laser micro-radar” (“radar” here is for “detection and ranging with radio waves”). We shall define the laser radar, or lidar, with a prefix “micro” when the distance between the instrument and the object is of the order of a meter or less. Of course, the subdivisions can be involved, creating a series of prefixes from “mini” to “femto” and smaller. The physical principles of detection and ranging can be the same not depending on distance (range) to the object, but can be modified and fitted only for small distances.

Scope of the Chapter

Actually, the “detection and ranging” in the term “lidar” should be understood wider than only “detection” and “measuring the range.” It means acquiring the information about the objects or media due to receiving the optical radiation reflected or scattered in backward direction. Usually, the source of radiation is a laser generating visible or invisible light. There is no severe restriction on the direction of the radiation scattering strongly backward.

The most often used principle of ranging with lidars is based on time-of-flight measurement. This principle is well known when determining the distance from a lightning by taking the count of the time-of-flight of the sound from the moment of light outburst. The idea to use short pulses of light to measure distance was brought out in 1930s by [Lebedev \(2014\)](#). His prototype measured the range up to 3.5 km with the accuracy about 2 m. The principle is applicable for a wide range of distances, from thousands of kilometers to micrometers – for measuring the distance to the Moon, tracking the player position by a play station, or checking the topography of the surface of microchips. It can be a single-dimension measurement (the distance between the lidar and the object), two-dimensional (2D) measurement (e.g., the profile of the surface), and three-dimensional (3D) measurement (e.g., a tomogram of the object).

The light propagates about 0.3 m for 1 ns. If to use bursts of light for measuring a small distance, their duration should be at least an order shorter. There would not be a problem to generate short laser pulses, but when designing a lidar for short distances, the problem is shifted to the detectors and processing. That is why, the time-of-flight technique at small distances uses a phase shift between the sent and the received signals. The micro-lidar can be designed to measure the phase difference between the light carrier frequencies, or between the modulation frequencies.

Continuous-wave (CW) light radiation is used for surface profile measurement, length measurement of small objects, like intraocular distances, small thickness layers or films. To get higher sensitivity, synthetic aperture is added by moving the instrument relatively to the investigated object. Specific field structure can be created by an array of light emitting diodes, enabling to acquire 3D image.

Interferometric methods initiated a new technology of the optical low-coherent tomography (OCT) with multiple modifications and wide practical use, the most exciting of which is a 3D imaging of the structure of the eye. Impressive are also results of the OCT in cerebrovascular and cardiovascular studies and diagnosing. Low-cost nomenclature of OCT using a smartphone is represented by multi-reference technique with its implementation in dermatology.

With micro-lidar, a series of media parameters can be measured. Analysis of Raman spectra allows differentiation between malignant and benign tissues. Analysis of fluorescence delivers information on blood circulation and can help in diagnosing the diseases of the eye. Differences in absorption of light in the anterior chamber of the eye can serve for measurement of the glucose level. Spectral differences in light scattering and absorption for different wavelengths are the basis for the method of measurement of blood oxygenation, including the oxygenation of the vessels of the retina.

In long-range laser radars, it is very important to keep the direction on the target. Similar, but not less sophisticated is the micro-lidar of the CD/DVD pickup head that should keep the head respectively to the fast running track not only along its run, but also at a certain distance in focus of the photodetector. Computer mouse is another example of micro-lidar. Massive production of computers and their periphery made their components so perfect, that they become the basics for other devices, like, for example, the pickup head in the role of the driver of the reference mirror in the smartphone-based OCT instrument.

With the advent of new technologies of vision correction, a problem arose of measurement of the refraction non-homogeneity of the optical system of the eye. Several solutions were found, some of them being typical laser radars. Ray tracing consisted in sounding the eye with narrow laser beam in different points of the eye aperture and measuring the coordinates of the light spot of the laser projection on the eye bottom. The inversion of this method was projecting a single laser beam on the retina, that created there a secondary source of light, and measuring the distortion of the image of this spot of light by the optical system of the eye at the exit from the eye.

The potential of the micro-lidar to measure the velocity of small particles in gas or liquid flow was one of the earliest implementations. Laser velocimeter was called sometimes the anemometer. Among its applications was measurement of missile plume flow, turbulence statistics of the submarine wake, velocity of blood flow in retinal vessels, etc. Several methods were developed using noncoherent and coherent radiation. Translational and axial components of the flow were measured. Doppler principle was incorporated into the OCT instruments enabling the measuring and visualization of the 3D distribution of the velocity in the flows.

Distance Measurement Techniques

Pulse Mode

Since the middle of the 19th century, short light pulses assisted in many discoveries in nature studies (Shapiro, 1984). It was natural that the development of picosecond lasers initiated new experiments in designing the micro-lidars. Park *et al.* (1981) demonstrated an optical ranging system which can see through opaque material. It uses picosecond pulses and a very fast gate with a variable delay at the detector. The system could detect a reflection from a mirror inside an opaque mixture with 1.7 mm range resolution and 0.1 mm transverse resolution. The depth of penetration corresponded to 0.67 cm in human tissue.

A possibility of gated viewing at a small-pathlength difference between object beam and reference beam was demonstrated on recording a hologram with the use of laser light of short pulse duration or short coherence length (Abramson, 1978). Only those parts of the object surface were holographically recorded that corresponded to a small-duration gate.

Fujimoto *et al.* (1986) described the application of femtosecond laser pulses to perform optical ranging using nonlinear cross-correlation gating. The laser pulses are focused onto the sample under study, and the remitted reflections and echoes are measured by nonlinear-optical cross correlation with the incident pulse. The experimental setup (Fig. 1) uses pulses of 65 fs duration at a repetition rate of 125 MHz, laser wavelength 625 nm, and an average power of up to 20 mW. The laser pulses are focused onto the sample under study, and the remitted reflections and echoes are measured by nonlinear-optical cross correlation with the incident pulse. The reference pulse travels through an optical delay line controlled by a stepping motor. The signal and reference are correlated by focusing with a crossed-beam geometry into a 0.5-mm length of angle-phase-matched potassium dihydrogen phosphate (KDP) second-harmonic crystal. Measurement of the integrated second-harmonic energy as a function of the temporal delay between the signal and the reference yields the cross correlation. Spatial resolutions of less than 15 μm was achieved. An example of signals is shown in Fig. 2.

CW Phase Difference Mode

CW modulated signals

In a phase-shift rangefinder, the optical power is modulated with a constant frequency. After reflection from the target, a photodetector collects a part of the laser beam. Measurement of the distance D is deduced from the phase shift $\Delta(\varphi)$ between the

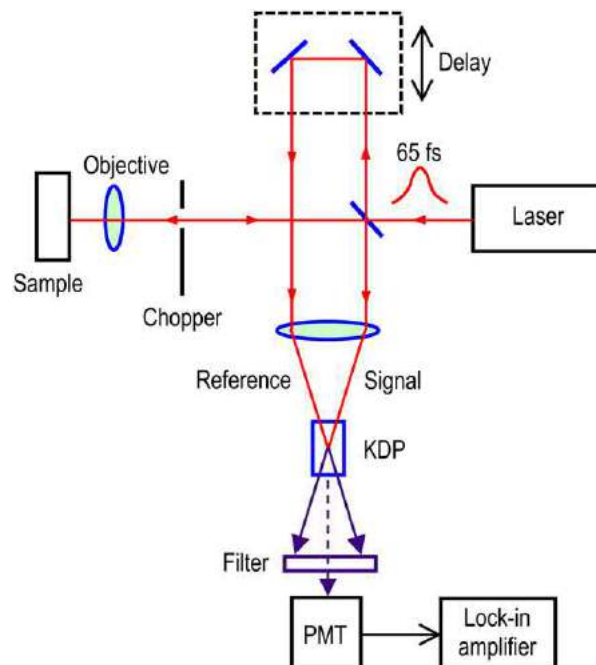


Fig. 1 Schematic of femtosecond optical ranging experiment. BS, beam splitter; PMT, photomultiplier tube; XTAL, nonlinear potassium dihydrogen phosphate (KDP) crystal. Reproduced from Fujimoto, J.G., De Silvestri, S., Ippen, E.P., *et al.*, 1986. Femtosecond optical ranging in biological systems. *Optics Letters* 11, 150–152.

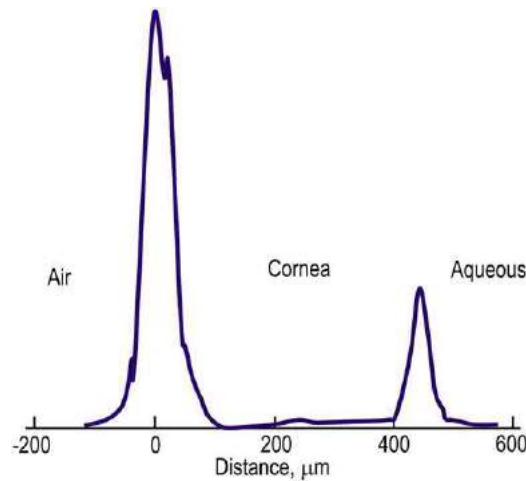


Fig. 2 Optical ranging measurement of the cornea of the rabbit eye performed in vivo. Reproduced from Fujimoto, J.G., De Silvestri, S., Ippen, E. P., *et al.*, 1986. Femtosecond optical ranging in biological systems. *Optics Letters* 11, 150–152.

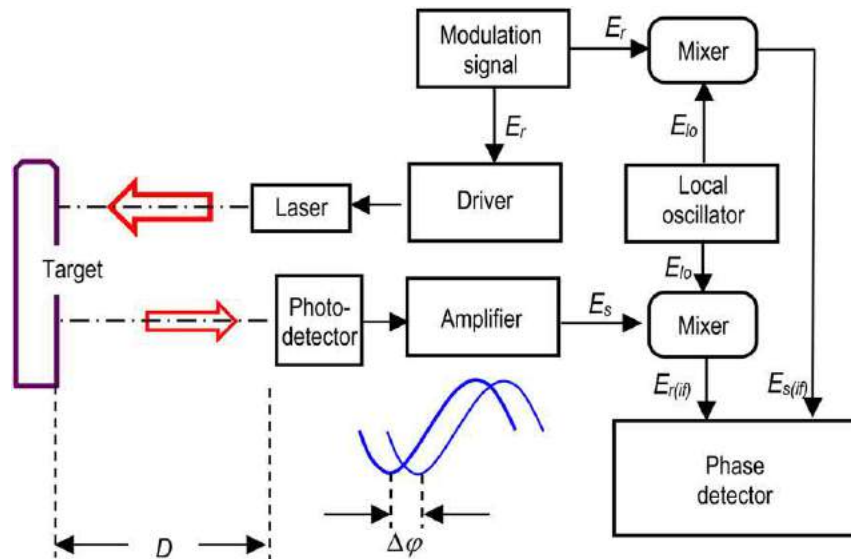


Fig. 3 Block diagram of a phase-shift laser rangefinder using a heterodyne technique. Reproduced from Hu, P., Tan, J., Yang, H., *et al.*, 2011. Phase-shift laser range finder based on high speed and high precision phase-measuring techniques. In: 10th International Symposium on Measurement Technology and Intelligent Instruments. June 29–July 2, 2011, KAIST, Daejeon, Korea.

received and modulated emitted signal: $D = (cF/2)((\varphi)/2\pi)$. The unambiguous distance measurement is limited to $(\varphi) = 2\pi$. This principle is often used with conversion of the modulation frequency to a lower intermediate frequency (Amann *et al.*, 2001; Hu *et al.*, 2011).

In the example of Fig. 3, the laser beam is modulated by a sine voltage E_r . After the detection and filtering, the voltage E_m is selected. Both voltages E_r and E_m are mixed with the signals E_{lo} from the local oscillator. The phase difference (φ) is measured between the signals E_{ifm} and E_{ifr} from the outputs of the mixers. This phase difference is then recalculated into the distance D . To get the measurement resolution 2 mm, at the measurement range $D = 15$ m and with the modulation frequency 10 MHz, the phase-shift should be measured with the resolution 0.05 degree.

Journet and Poujouly (1998), measured the phase shift at an intermediate frequency of 125 kHz. The emitter was a laser diode with 12 mW average output power, the receiver used an avalanche photodiode. The phase shift was measured by direct counting with more than 10 bits resolution, that provided about 0.4 mm for a 60 cm range without phase ambiguity. A computer controls the range-finder through a specially developed interface.

In a later version, Poujouly and Journet (2000) used a digital phase-locked loop to reduce the phase noise (Staszewski and Balsara, 2005). They also described the technique based on under-sampling, applied to digital synchronous detection. Its main advantage is a global simplification of the electronic system, leading to a quite simple development of a twofold

modulation system. This new technique is also very interesting to move toward a kind of smart rangefinder able to adapt different parameters to the different steps of the measurement.

Journet and Lourme (2000) proposed to measure the correlation between the transmitted and the received modulation signals this type of modification is not restricted by the sine modulation, because the distance is determined by the delay of the reference signal, at which the correlation has maximum.

Sugimoto *et al.* (2013) proposed to use two sine modulation signals of different frequencies forming a sinc-similar beating signal (Fig. 4), specific points of which are used as references to measure the delay of the received signal relatively to the transmitted signal.

In a wide use in the long-range laser radars (Agishev, 2009), the continuous-wave frequency modulation (CWFM) with a linearly varying frequency is used (Fig. 5), often called chirp modulation due to its similarity to bird's chirping. For this technique,

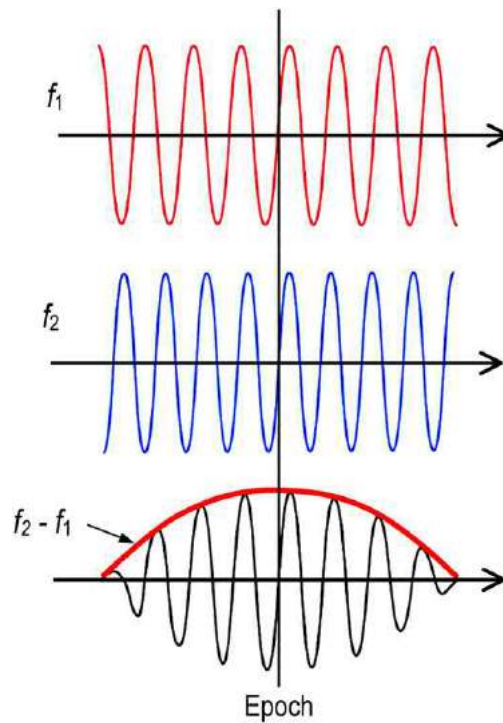


Fig. 4 Two modulation frequencies and their sum (an epoch is a specific point of a central zero crossing in the beating signal). Reproduced from Sugimoto, M., Nakamura, S., Inoue, Y., *et al.*, 2013. LT-PAM: A ranging method using dual frequency optical signals. *International Journal on Smart Sensing and Intelligent Systems* 6 (3), 791–809.

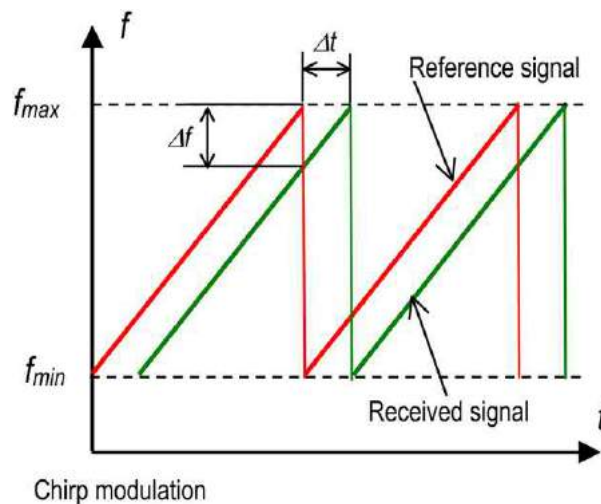


Fig. 5 Time-frequency relationship in a chirp signal. Reproduced from Agishev, R.R., 2009. *Lidar Monitoring of the Atmosphere*. Moscow: Fizmatgiz (in Russian).

the frequency difference between the transmitted and received signals is proportional to the time delay and correspondingly, to the measured distance.

Frequency difference techniques

Axial eye length measurement

Sekine *et al.* (1993) described a noncontact technique for axial eye length measurement. According to this method, the wavelength shift of a single-mode laser-diode beam that is irradiated onto the eyeball causes a phase shift in the interference fringes of reflections from the retina and the cornea. Then the optical distance between the cornea and the retina is obtained from the phase-shift measurement.

The schematic arrangement of the optical system is shown in Fig. 6 with the following notations: BS – beam splitter; C – cornea; CL – collimator lens; OA – optical path compensator; OI – optical isolator; R – retina; SF – spatial filter. The light source is a single-mode laser diode (LD). Illumination light coming from the LD is first split to a measuring interferometer and a reference light path. Light that enters the interferometer is again split by a beam splitter and converged separately at the cornea and retina. Reflections from the cornea and retina return to each optical system and interfere with each other. They are then received by a photodetector PD₁. On the other hand, light that enters the reference light path is reflected by mirrors M₁ and M₂, and these reflections cause interference before they are received by photodetector PD₂.

Smoothly varying the light frequency in a given range, the number of phase zeros is calculated in the measuring and in the reference arms (N_m and N_r). It is shown that the eye length L_{eye} can be calculated as $L_{eye} = L_{ref} (N_m/N_r)$, where L_{eye} is the eye axial optical length, $L_{ref} = L_1 - L_2$. $L_{opt} = L_{opt1} - L_{opt2}$ is adjusted to be zero.

Compared to the technique that uses partially coherent light, this technique is inferior in terms of measurement accuracy but superior in its wide, measurable range of 16–32 mm. The results of measurements showed that double standard deviation of measurement was $2\sigma = \pm 0.11$ mm.

Wavelength tuning interferometry

The idea of measuring the distance using the interferometer with changing frequency was implemented by Lexer *et al.* (1997) as the laser wavelength-tuning interferometry of intraocular distances (Fig. 7). The notations are as follows: OS – object signal; RS – reference signal; SP – scattering potential; MI – reference Michelson interferometer; FT – Fourier transform; NC – numerical correction; PD₁, PD₂ – photodetectors; TL – tunable laser. Shifting the wavelength of a tunable laser diode causes intensity oscillations in the interference pattern of light beams remitted from the intraocular structure. A Fourier transform of the photodetector signal yields the distribution of the scattering level along the light beam illuminating the eye (Fig. 8). The authors demonstrated simultaneous measurement of the anterior segment length, the vitreous chamber depth, and the axial eye length in human eyes in vivo with data-acquisition times in the millisecond range.

Dual-beam double frequency configuration

See *et al.* (1985) applied the differential phase-contrast scanning microscope for the surface studies. Molebny *et al.* (1996a,b) described similar configuration to measure the ablation profiles after the Lasik surgeries for the correction of the human vision (Fig. 9). The laser beam having its carrier frequency f_0 is split by the acousto-optical modulator in two components ($f_0 + f_1$) and ($f_0 + f_2$). The frequencies f_1 and f_2 are generated by the driving generator. After being reflected from the specimen, these components are fed to the photodetector whose output is connected to the phase detector. The phase difference $\Delta(\varphi)$ at the frequency $\Delta f = (f_2 - f_1)$ is the measure of the height difference Δh of the surface along the line connecting the beam projections on the surface. To reconstruct the profile, the surface is scanned along the line in the direction of the line connecting

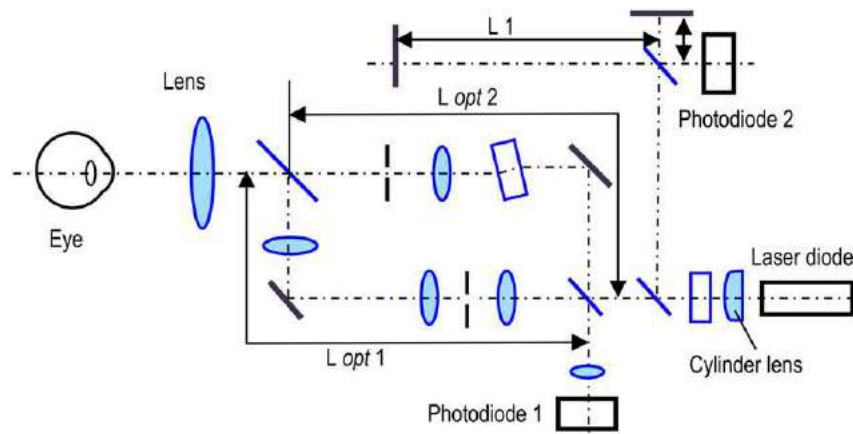


Fig. 6 Scheme of the interferometric system for axial eye-length measurement. Reproduced from Sekine, A., Minegishi, I., Koizumi, H., 1993. Axial eye-length measurement by wavelength-shift interferometry. *Journal of the Optical Society of America A* 10, 1651–1655.

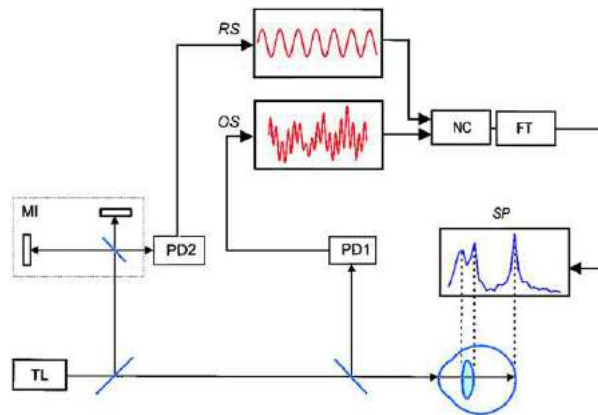


Fig. 7 Scheme of the wavelength tuning interferometer. Reproduced from Lexer, F., Hitzengerger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. *Applied Optics* 36, 6548–6553.

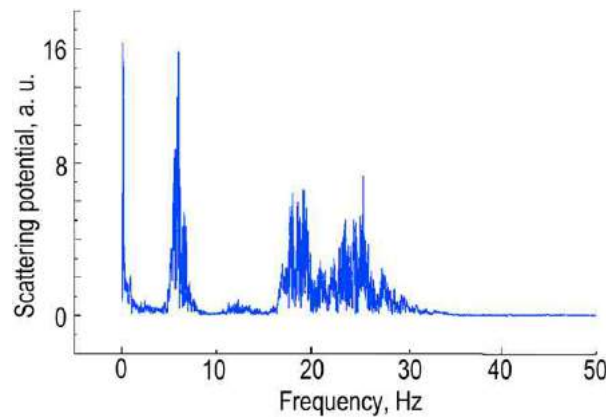


Fig. 8 Interferometry scan of an eye model. Reproduced from Lexer, F., Hitzengerger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. *Applied Optics* 36, 6548–6553.

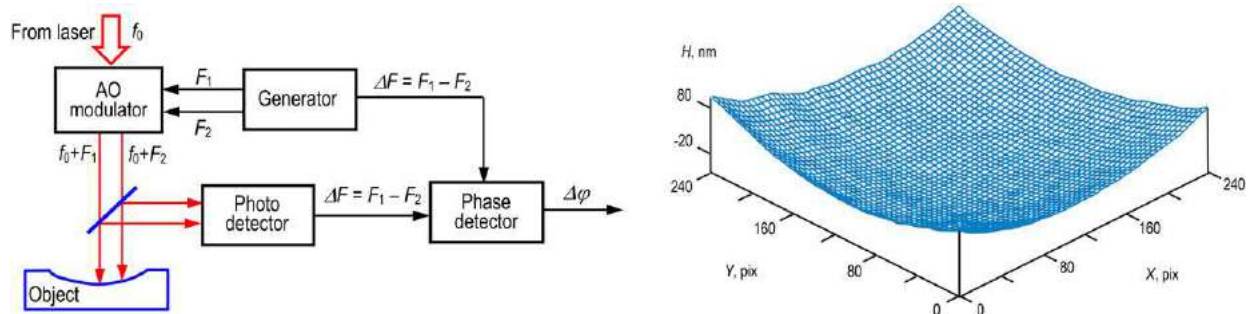


Fig. 9 Dual-beam double frequency configuration for measurement of the ablation profiles after Lasik surgery of vision correction. To the right: an example of the reconstructed profile using a triple-beam three-frequency design. Reproduced from Molebny, V.V., Pallikaris, I.G., Naoumidis, L. P., *et al.*, 1996. Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles. *Proceedings of SPIE* 2748, 68–75.

the beam projections (e.g., along the x direction). The next line of scanning is one-step shifted in y -direction. With laser wavelength 630 nm, $f_0 = 80$ MHz and $\Delta f = 1$ MHz, the sensitivity to profile measurement was achieved better than 10 nm. To make the data for the next line of scan more defined, a triple beam configuration was used with beams split in x and y directions (Molebny *et al.*, 1998).

Instead of splitting the frequencies with high-cost acousto-optic modulators (AOMs), a gas laser (Bourdet and Orszag, 1979) or modern two-wavelength laser diodes can be used (Ishii and Onodera, 1991). High relative wavelength stability for large variations in optical path length can be achieved with two or more single-mode lasers locked to a Fabry–Pérot etalon

(de Groot and Kishner, 1991; Gerstner and Tschudi, 1994). A simpler configuration for short path differences can be designed using multimode laser diodes (Rovati *et al.*, 1998; de Groot, 1991).

3D laser microvision

The 3D laser microvision developed by Shimotahira *et al.* (2001) is based on the principle of a step frequency radar. Optically, it is an interferometer that derives the distance having the information of the phase delay of the scattered signal (Fig. 10 with the notations: NBS – non-polarizing beam splitter; AOM – acousto-optic modulator; PZT – piezoelectric transducer). The carrier frequency from the superstructure grating (SSG) laser source is swept stepwise: $f_n = (f_0 + n\Delta f)$. The wavelength of the SSG laser diode is centered at $1.545 \mu\text{m}$. In each frequency step, the following procedure is repeated. The output laser light is split into the probe and reference beams by means of an AOM. The non-refracted beam (the zero-order beam) from the AOM is used as a probe beam, and the refracted beam (the first-order beam) from the AOM whose carrier frequency is shifted from that of the probe beam by the frequency of excitation of the AOM is used as a local oscillator beam. The probe beam reflected from the target is mixed with the reference beam from the AOM in a coherent detector. The heterodyne-detected signal, both amplitude and phase, is stored. After all frequency stepping is completed, the stored data are processed and a 3D image of the target is displayed.

The lateral resolution (x - y plane resolution) depends on the focusing parameters of the objective lens, and it is best only at one focused depth. To remove this difficulty, the synthetic aperture approach is used coming from that of the microwave side-looking airborne radar (SAR) which makes high-resolution maps from scanned microwave echoes. A linear flight path of an airplane is used as a linear scan path of the probing antenna. With the 3D laser microvision, the light scattered from the target is measured while the probe is moving away from the target along a path perpendicular to the surface of the target. Thus the scanned probe collects information about the scatterer and significantly improves the lateral resolution as well as the depth resolution of the 3D laser microvision.

In this instrument, the laser wavelengths cover a range as wide as 30 nm centered at $1.545 \mu\text{m}$ with an output power of 1 mW. A specially developed digital phase meter can measure the phase with an accuracy of 0.03 degree as the time delay between the rising edge of both, returned and reference signals. To obtain 3D information, the probe beam has to be scanned over a 2D surface, 128×128 pix large, and at each location of it, the depth information is probed to construct a 3D image.

Fig. 11 demonstrates the effectiveness of the synthetic aperture method using a target of a micro-strip line array on a Gallium arsenide (GaAs) wafer (a – without the synthetic aperture scanning, b – with the synthetic aperture scanning). The micro-strip lines are shown at the bottom of the figure. The width of each micro-strip line is $15 \mu\text{m}$, the thickness is less than $1 \mu\text{m}$, and the spacing between the adjacent edges of the micro-strip line is $8 \mu\text{m}$. The vertical synthetic aperture was made in the stepwise z direction. Curve *b* in Fig. 11 shows the result of the synthetic aperture scanning. The image clearly separates the edge spacing of the micro-strip lines. Fig. 12 shows a microscope photograph of a tunneling photodiode that is deposited on the GaAs substrate. The overall deposit area is $500 \mu\text{m} \times 500 \mu\text{m}$.

Interferometric Methods

Optical Interferometric Micro-Lidar

Time-of-flight measurement can be combined with interferometry using a Michelson interferometer, where one of the mirrors is replaced by a sample under test (Dresel *et al.*, 1992; Fig. 13 left). A light source with short coherence length is used. The reference

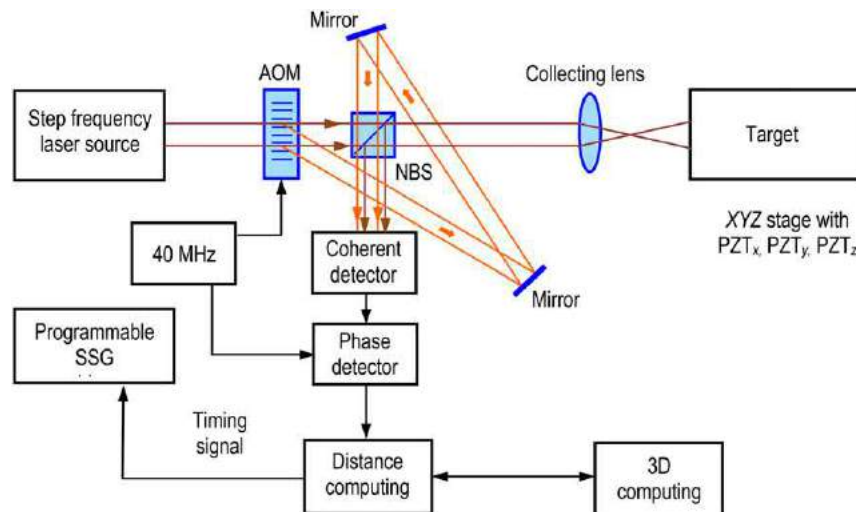


Fig. 10 Block diagram of the 3D laser microvision. Reproduced from Shimotahira, H., Iizuka, K., Chu, S.C., *et al.*, 2001. Three-dimensional laser microvision. *Applied Optics* 40, 1784–1794.

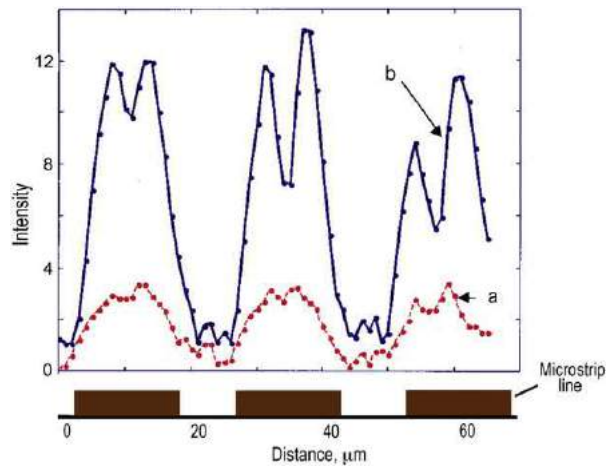


Fig. 11 The target of the three-dimensional (3D) laser microvision – gallium arsenide (GaAs) substrate with thin metal electrodes. Reproduced from Shimotahira, H., Iizuka, K., Chu, S.C., *et al.*, 2001. Three-dimensional laser microvision. *Applied Optics* 40, 1784–1794.

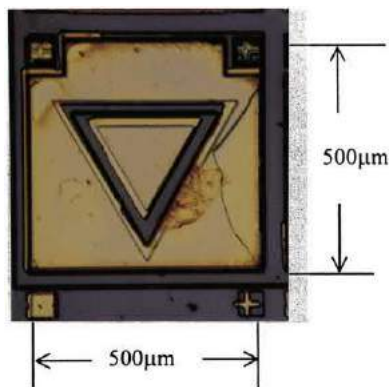


Fig. 12 Measured results of a micro-strip line array on a gallium arsenide (GaAs) wafer. Reproduced from Shimotahira, H., Iizuka, K., Chu, S.C., *et al.*, 2001. Three-dimensional laser microvision. *Applied Optics* 40, 1784–1794.

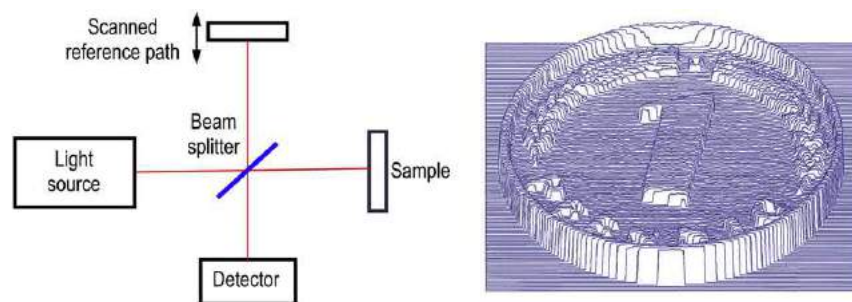


Fig. 13 Basic setup of an optical microradar (left); three-dimensional (3D) reconstruction of a coin image (right). Reproduced from Dresel, T., Häusler, G., Venzke, H., 1992. Three-dimensional sensing of rough surfaces by coherence radar. *Applied Optics* 31, 919–925.

mirror plane and the sample plane are imaged onto a detector. The image of the sample is superimposed with a reference wave. Interference takes place when the light paths from the reference and the sample are equal and only within those speckles that correspond to the surface elements close to the sample plane. These regions are detected and stored, while the reference mirror is scanned along the z axis.

The profile of the sample can be reconstructed by scanning in (x, y) directions. An example of a reconstructed profile is presented in [Fig. 13](#) right.

The first biological application of low-coherence interferometry for the measurement of axial eye length was reported by [Fercher *et al.* \(1988\)](#). Different versions of low-coherence interferometry were developed for noninvasive measurement in

biological tissues. Taking an eye as an object, there will be several positions of the reference mirror along the z axis, where the interference gives maximal signal (Fig. 14 left). They correspond to the cornea, lens, and retina. Scanning in (x, y) directions yields the profiles of these surfaces. Fig. 14, right, shows an example of the structure of the human retina reconstructed from the data of measured intensity of light scattered from the eye bottom, the beam being shifted laterally scan-by-scan.

Optical Low-Coherence Tomography

Time-domain optical low-coherence tomography

This interferometric lidar technique was soon baptized by Huang *et al.* (1991) as the optical coherence tomography (OCT). To be more accurate, it should be named the optical low-coherence tomography. OCT imaging system based on a fiber-optic Michelson interferometer is shown in Fig. 15. The advantage of the fiber optics technology is that the setup is robust and easier in alignment. In the schematic, a Michelson interferometer contains a circulator for dual balanced detection. Dual balanced detection adds the signal from the interference of the sample and reference arms and subtracts excess noise from the light source. The sample/probe arm may be interfaced to a variety of imaging devices.

Early OCT instruments used a low-coherence light source and interferometer with a scanning reference delay arm. This method is known as a traditional or time-domain OCT (TD-OCT). OCT found its first application in ophthalmology for diagnosing ocular diseases such as glaucoma, age-related macular degeneration, and diabetic retinopathy.

Spectral-domain optical low-coherence tomography

Later on, the spectral/Fourier domain OCT (spectral-domain optical low-coherence tomography (SD-OCT)) and the swept source/Fourier domain OCT (SS-OCT) were developed, also termed optical frequency domain imaging (OFDI). The SD-OCT has a powerful sensitivity advantage over the time domain detection, since it measures all of the echoes of light simultaneously. This discovery drove a boom in OCT research and development. The sensitivity is enhanced by 50–100 times, enabling a corresponding increase in imaging speeds. The second type of Fourier domain detection, SS-OCT, uses an interferometer with a narrow-bandwidth, frequency swept light source and detectors which measure the interference output as a function of time. It has the advantage that it does not require a spectrometer and line scan camera. Therefore, it can operate at longer wavelengths where

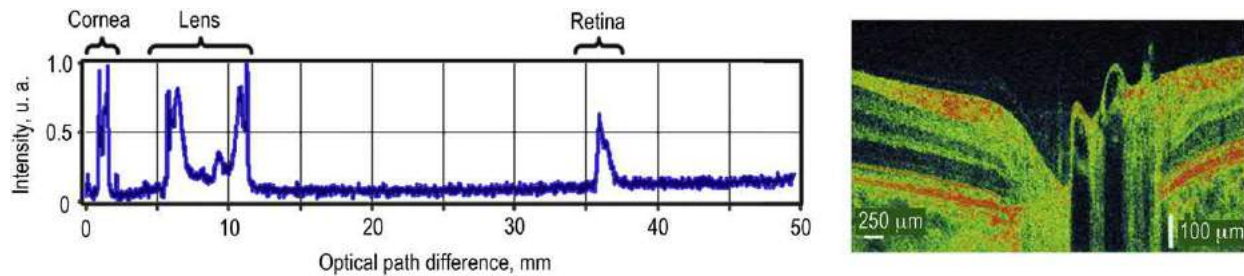


Fig. 14 Intensity distribution along the path of the beam in the eye (left), and the color-coded reconstruction of the intensity of the light scattered from the retina when the beam is scanned laterally (right). Reproduced from Fercher, A.F., Mengedoh, K., Werner, W., 1988. Eye-length measurement by interferometry with partially coherent light. *Optics Letters* 13, 186–188.

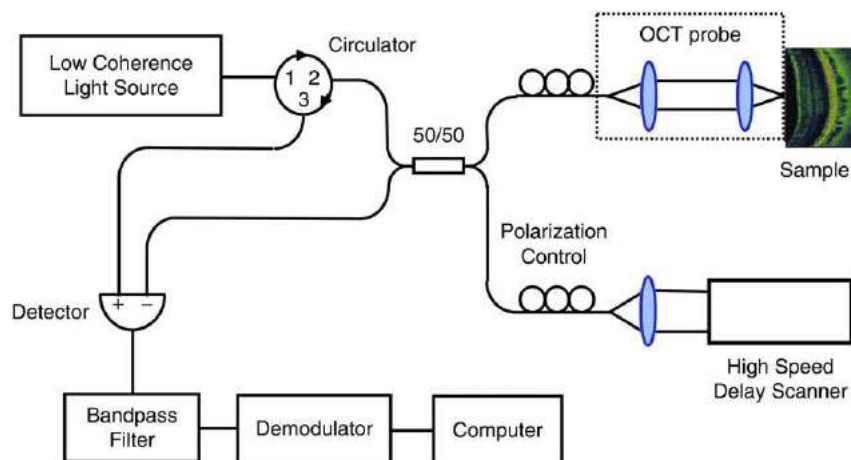


Fig. 15 Schematic of the time-domain optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. *Optical Coherence Tomography*, second ed. Switzerland: Springer.

camera technology is less developed and it can achieve imaging speeds which are much faster than spectral/Fourier domain OCT which is limited by the camera speed. The primary challenge in swept source/Fourier domain OCT is that it requires a high-speed, swept narrow line width light source.

Schematic of a typical spectral domain OCT instrument is presented in Fig. 16. The reference arm has a fixed delay and is not scanned. Interference is detected with a spectrometer and a high-speed, line scan camera. A computer reads the spectrum, rescales it from wavelength to frequency, and Fourier transforms to generate axial scans.

Swept-source optical low-coherence tomography

Schematic of a typical swept source OCT instrument is shown in Fig. 17. It contains a frequency swept light source. The reference arm has a fixed delay and is not scanned. The example shows a sample arm with catheter/endoscope interface and the system is assumed to operate at $1.3\ \mu\text{m}$ wavelengths. This geometry enables dual balanced detection, canceling excess noise in the laser. A portion of the frequency swept light is directed into a Mach-Zehnder interferometer which acts as a periodic frequency filter. This interferometer configuration is more efficient than the classic Michelson interferometer because all of the light is detected. It avoids the spectral resolution and pixel limitations which are inherent in spectrometers and line scan cameras.

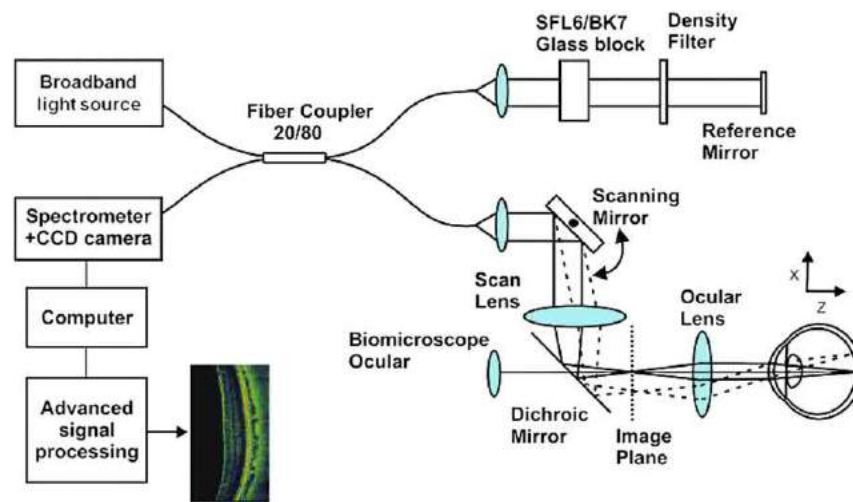


Fig. 16 Schematic of a typical spectral domain optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

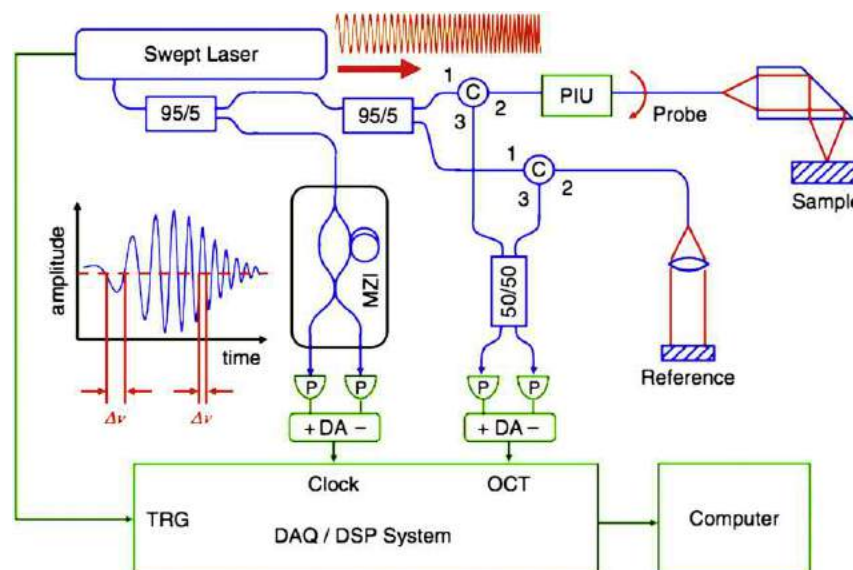


Fig. 17 Schematic of a typical swept source optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

Commercial OCT Instruments for Ophthalmology

Fig. 18 shows (from left to right): (1) an early retinal imaging prototype instrument. The OCT system is interfaced with a slit lamp biomicroscope. The OCT beam is scanned using a pair of galvanometer-actuated mirrors. This system was used at the New England Eye Center during the mid-1990s; (2) Zeiss Cirrus spectral OCT – one of the latest, cutting-edge ophthalmic instruments available today; (3) Copernicus REVO spectral OCT.

Color encoding of the OCT data

Display techniques were developed to display OCT data in the form of thickness maps. An early example of a time-domain OCT topographic color-encoded map of retinal thickness is shown in **Fig. 19** left (Hee *et al.*, 1998). An example display of the spectral-domain Zeiss Cirrus instrument is shown in **Fig. 19** right (Mazzarella and Cole, 2015).



Fig. 18 Commercial optical coherence tomography (OCT) instruments for ophthalmology. Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

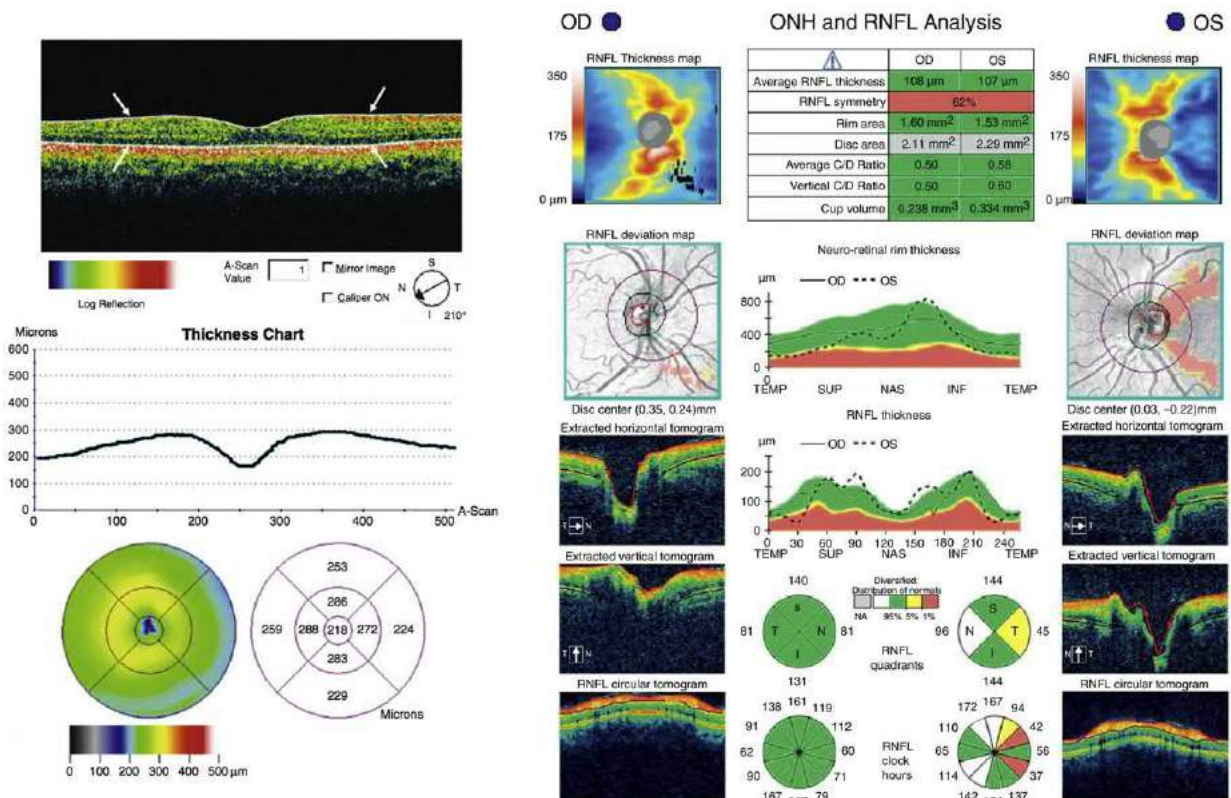


Fig. 19 Display of an early optical coherence tomography (OCT) instrument with color-encoded topographic map of retinal thickness (left). Cirrus OCT display in the glaucoma diagnosing mode (right). Source: internet

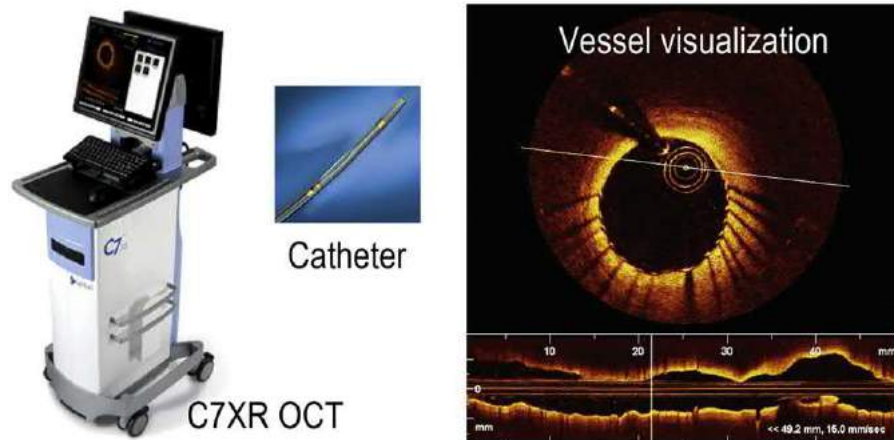


Fig. 20 High-speed intravascular C7XR optical coherence tomography (OCT) system (left); vessel visualization (right). *Source:* internet

OCT in Cerebrovascular and Cardiovascular Research

Fig. 20 left demonstrates the C7XR OCT system (Guiagliumi *et al.*, 2016), that was introduced as the first commercially available FD-OCT system for coronary imaging in 2009. **Fig. 20** right shows an example of intravascular OCT imaging (Adler, 2012). The mechanical actuation necessary to sweep the source wavelength is on the nanometer scale and therefore is accomplished in a compact MEMS-based packaging.

The C7XR system has a spatial resolution of 15 μm and acquires 500 axial lines per frame at a rate of 100 frames/s. At the tip of the catheter, there is a 125- μm -diameter optical fiber assembly that consists of an integral side-looking lens. The optical assembly is encapsulated in a hollow torque wire that translates motion from a drive motor located outside the patient's body. The C7XR system can volumetrically image a 5-cm vessel segment in approximately 3 s.

The latest generation of this type OCT system, the Ilumien Optis, has a higher image acquisition rate (180 frames/s) and can acquire a 75-mm pullback image in only 2.1 s (see Section Relevant Websites). This system has instant 3D reconstruction capability and is used for optimizing stent planning.

Another OCT system developed at the Massachusetts General Hospital (MGH, Boston, MA) has a high-speed wavelength swept laser that uses a polygon-mirror/grating filter to tune over a broad (120-nm-wide) bandwidth centered at 1320 nm. Its axial resolution is 7 μm , and the ranging depth is 4.6 mm in tissue.

The same MGH academic group constructed an OCT system that uses a very broad bandwidth light source and common-path SD-OCT technology, termed microoptical coherence tomography (μOCT), whose resolution is improved by an order (axial resolution $\leq 1 \mu\text{m}$ and a lateral resolution $\leq 2 \mu\text{m}$ in tissue). It provides clear pictures of cellular and subcellular features (Liu *et al.*, 2012).

The Terumo (Tokyo, Japan) OFDI system uses a 1.3- μm scanning laser as a light source. The imaging range is 5 mm with an axial resolution of less than 20 μm in water.

Multiple Reference OCT for Dermoscopy

Development of the newest modalities of OCT is oriented not only on super-fast techniques or having super-high resolution, but also on making the instrument as cheap as possible. One of the examples is a multiple reference OCT (MR-OCT) system (Hogan and Wilson, 2009; Dsouza *et al.*, 2014). The idea is to use a simple along- z scanner made, for example, of a voice coil actuator (VCA). Similar actuator can be taken from a typical CD/DVD pickup head (Subhash *et al.*, 2015) whose role in the pickup head was to keep the CD/DVD track in focus at the photodetector (**Fig. 21** left).

In a prototype, the depth scanning was achieved by using a miniature recirculating optical delay based on a voice coil motor actuator and a partial mirror. The actuator was extracted from a CD optical pickup head and can provide an A-scan rate of 50–600 Hz with an axial displacement of about 60 μm . The prototype MR-OCT sensor setup (Subhash, 2014) was based on a smartphone platform using the Android operating system (**Fig. 21** right). The notations in **Fig. 21** are as follows: PUH – pickup head, SLD – superluminescent diode, L_1 – L_3 – lenses 1–3, BS – beam splitter, PM – partial mirror, VCA – voice coil actuator, RM – reference mirror, D – distance between PM and RM, S – scan range, TD – total depth scan range. Recirculation is provided by adding the PM in front of the RM at a distance D , recirculating in this way the scan range S several times and extending the TD.

Devices With Photonic Arrays

Focus Tracking Micro-Lidars

For some purposes, critical is not the absolute value of the distance but the deviation from a given value. Typical example is a CD/DVD pickup head whose application in the MR-OCT system was described in the previous paragraph.

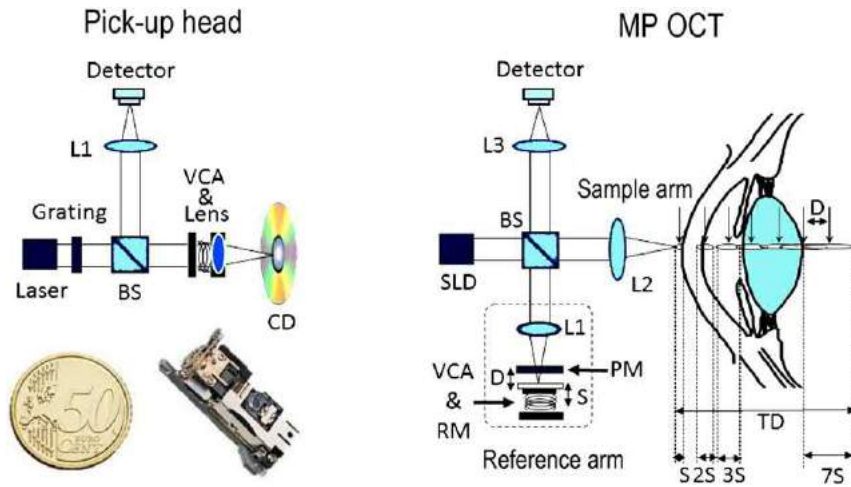


Fig. 21 Pickup head with a voice coil actuator (VCA) and the lens (at the left); experimental setup of a multiple reference-optical coherence tomography (MR-OCT) system using the same VCA with the reference mirror instead of the lens. Reproduced from Subhash, H.M., Hogan, J.N., Leahy, M.J., 2015. Multiple reference optical coherence tomography for smartphone applications. SPIE Newsroom. doi:10.1117/2.1201503.005807.

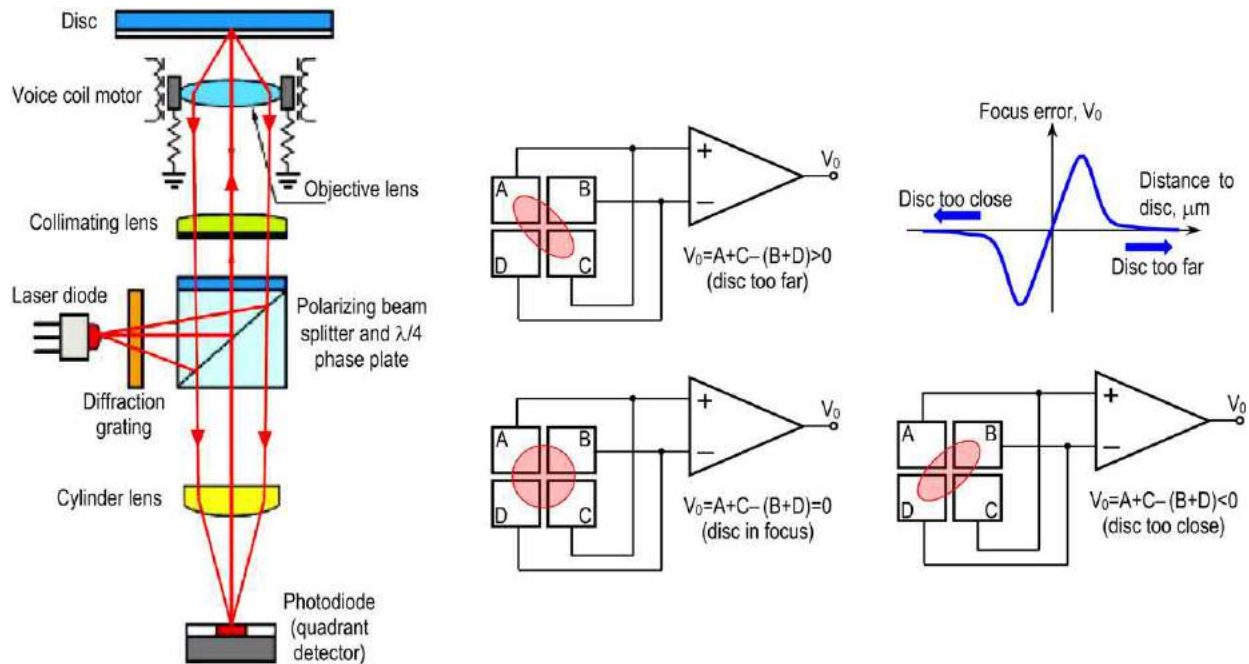


Fig. 22 Pickup head and the principle of focus tracking. Reproduced from Pohlmann, K.C., 2011. Principles of Digital Audio, sixth ed. New York, NY: McGraw-Hill.

A pickup head consists of a transmitting and of a receiving channels (Pohlmann, 2011). The transmitting channel contains a laser diode irradiating the polarized light at an infrared wavelength, a diffraction grating, a polarization beam splitter with a quarter-wave plate, a collimator lens, and an objective lens (Fig. 22). Laser radiation reflected from the CD passes through the objective lens, the collimating lens, the polarization beam splitter, and then is focused by a cylinder lens on the four-quadrant photodetector with four photodetectors A, B, C, and D. The quarter-wave polarization plate converts the linear polarization of the light into the circular one. On the way back, the circular polarization changes the direction of vector rotation, and after the quarter-wave plate, the linear polarization becomes orthogonal to the initial polarization. This conversion enables transporting the light to and from the disk without losses in the beam splitter.

After the cylinder lens, the beam shape at the detector array is either elliptic or circular. The circular shape occurs only if the reflective surface of the specimen is placed exactly at the focal plane of the optical path. The laser beam in this case has a diameter of less than 1 μm . The elliptic shape occurs when the beam reflection takes place out of focus. The orientation and size of the elliptic beam pattern respectively depend on the location and distance from the focal plane. The detector array has four areas,

denoted as A , B , C , and D . They deliver photocurrents, depending on the illuminance integrated over their area, which are subsequently linearly converted into voltages. From these voltages (denoted also as A , B , C , and D) the focus error V_0 can be derived as $V_0 = (A + C) - (B + D)$. When $V_0 = 0$, the laser beam is focused on the disk, if $V_0 > 0$, the disk is too far, and when $V_0 < 0$, the disk is too close. This voltage is applied to the VCA (motor) that adjusts the distance to the position until V_0 equals 0. The mentioned diffraction grating, installed at the exit from the laser, produces a pair of diffracted beams in $+1$ and -1 orders symmetrically situated relatively to the main beam. Projected on the track in the way that the side beams are clinging to the track, one at the left side, another at the right side as shown in Fig. 23.

Besides the four areas A , B , C , and D in the form of quadrants, the photodetector array contains also two additional photodiodes E and F . The filtered difference TR of the signals F and E creates the tracking signal keeping the central laser spot on the track.

Range Imaging With Matricial Light-Emitting Diode Array

Reconstruction of the shape of the surface from the interferometric data has a long history, it was summarized by Creath (1988) with many examples. When a fringe pattern is recorded by a detector array, there is an output of voltages representing the average intensity incident upon the detector element over the integration time. As the relative phase between the object and reference beams is shifted, the intensities measured by point detectors will change.

Some techniques change the phase stepwise by a known amount between intensity measurements (Surrel, 1993), whereas others integrate the intensity while the phase is being shifted (Sasaki and Okazaki, 1986). The first is usually referred to as a phase-stepping technique, and the second as an integrating-bucket technique.

An example implementing the four-bucket algorithm was described by Wang *et al.* (2010). The instrument measures 3D surface profiles and is based on an light-emitting diode (LED) array phase-shift rangefinder (Fig. 24). It employs the fast electronic scanning of a 2D array of LED light sources instead of mechanical scanning, which means no moving or rotating parts are needed in the system. By using high-sensitivity photodiodes (PD) and high-accuracy analog-to-digital converter (ADC) to sample and demodulate the incoming light signal, a high range resolution can be obtained. Fig. 25 illustrates four-bucket principle. A sine wave signal is sampled four times (A_0 , A_1 , A_2 , and A_3) within a modulation period, and each sample point is delayed by a quarter

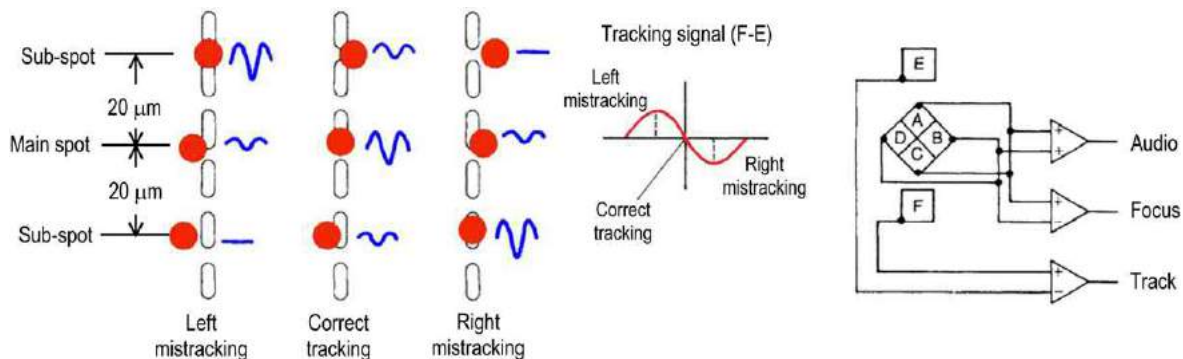


Fig. 23 Principle of keeping the central laser beam on the track. Reproduced from Pohlmann, K.C., 2011. Principles of Digital Audio, sixth ed. New York, NY: McGraw-Hill.

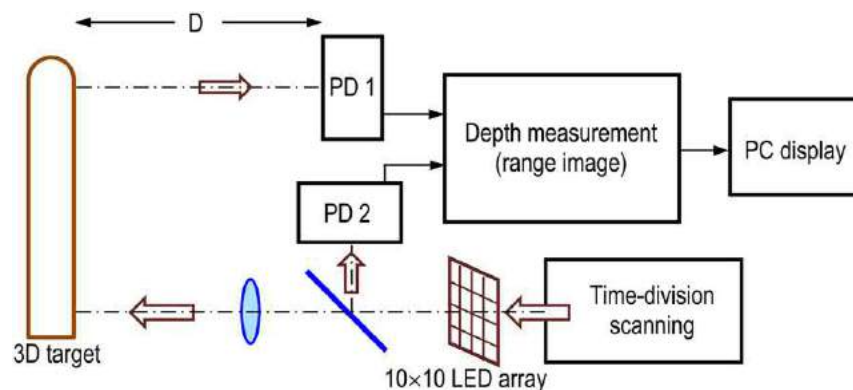


Fig. 24 Block diagram of range imaging system. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

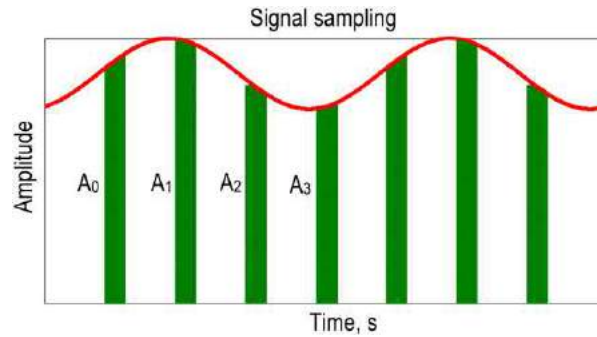


Fig. 25 Principle of phase measurement using four-bucket algorithm. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

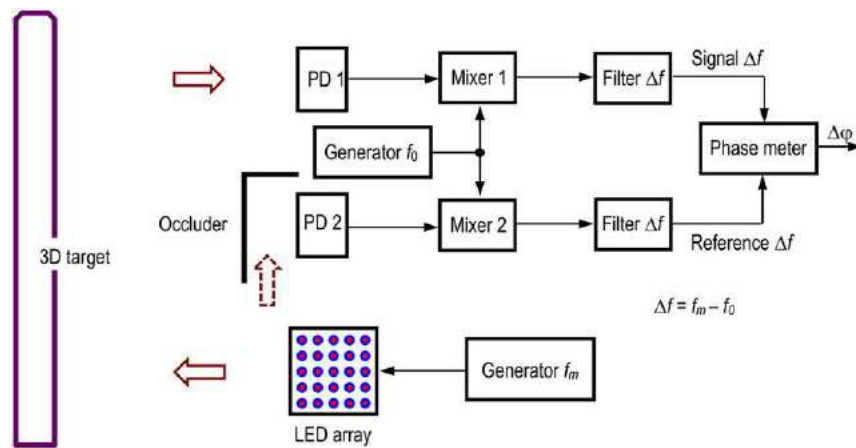


Fig. 26 Block diagram of phase-shift rangefinder using heterodyne technique. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

period of the modulation frequency. From this sampling, based on Nyquist theorem, the phase delay of the signal can be determined and recalculated into the measured distance.

Fig. 26 shows the block diagram of the phase-shift rangefinder using a heterodyne technique. Since the modulated LEDs have temperature drift, two receiving channels are used in the system, of which one serves as a reference channel. The signals from the outputs of two mixers are filtered by band-pass circuits tuned on f_1 with a bandwidth Δf_1 . Then, the two signals are sent to the phasemeter, where the phase measurement is implemented.

The instrument is insensitive to ambient light, since the dc-offset of the received signal is removed. It is no problem to use a large pixel array size and high spatial resolution by utilizing a low-cost close packaged LEDs. The system can be very robust and can be used to perform inexpensive and real-time range scanning for many robotic and industrial automation applications, including those in hazardous environmental conditions. By combining with the mature and low-cost 2D charge-coupled device/complementary metal-oxide semiconductor (CCD/CMOS) imaging technique, it also can achieve real-time, high-quality 3D imaging, which is another important application prospect for this range imaging system. The authors found that the range images can be acquired at a rate of 10 frames/s with a depth resolution better than ± 5 mm in the range of 50–1000 mm.

Time-of-Flight Principle With Matricial Detectors

To get the information in the 3D space, not only the matrices of emitters can be used, but also the matrices of photodetectors as well (Mutto *et al.*, 2012). Such micro-lidar (sensor) may be conceptually interpreted as a set of a multitude of single devices. It does not mean that the matrix of emitters should be manufactured in a single chip with the matrix of receivers. It only means that a single emitter may provide an irradiation that is reflected back by the scene and collected by a multitude of receivers close to each other.

Once the receivers are separated from the emitters, the former can be implemented as CCD/CMOS pixels integrated in a matrix. The lock-in pixels matrix is commonly called time-of-flight (ToF) camera sensor, and, for example, in the case of the MESA

Imaging SR4000 it is made by 176×144 lock-in pixels (see Section Relevant Websites). Such matricial ToF sensor IR emitters are common LEDs and they can be positioned in a configuration mimicking a single emitter co-positioned with the center of the receivers matrix, as shown in Fig. 27. Indeed the sum of all the IR signals emitted by this configuration can be considered as a spherical wave emitted by a single emitter. Fig. 28 shows the actual emitters distribution of the MESA Imaging SR4000.

The operation of a ToF camera as imaging system can be summarized as follows. Each ToF camera sensor pixel, at each period of the modulation sinusoid, collects four samples of the IR signal reflected by the scene. This information is enough to reconstruct the depth values for each pixel and to build the map of distances in the scene. Accurate description of this kind of ToF camera is given by Lange (2000) in his PhD work. Hansard *et al.* (2012) and DAGM Workshop Proceedings (Kolb and Koch, 2009) can be recommended for further reading.

Pattern-Projecting Matricial Sensors

In 2012, the Microsoft Kinect version of the play station was released which included the depth mapping with a light-coded range camera capable to estimate the 3D geometry of the acquired scene at 30 fps with VGA (640×480) spatial resolution (Fig. 29). From the functional point of view, the Kinect range camera is similar to the ToF camera described previously, since they both estimate the 3D geometry of dynamic scenes, are similar also in that they use matricial detectors, but they differ in the principles: the Kinect technology is based on pattern-projection analysis, several patents were issued to PrimeSense (now Apple) (Freedman *et al.*, 2012; Sali and Avraham, 2014; Cohen *et al.*, 2016). Some commenting authors (Mutto *et al.*, 2012) call this technique “matricial active triangulation.” Light patterns can be generated by a light modulator as random distributions of light and dark spots. It can be also a diffraction or speckle pattern. García *et al.* (2008) projected coherent light through ground glass generating

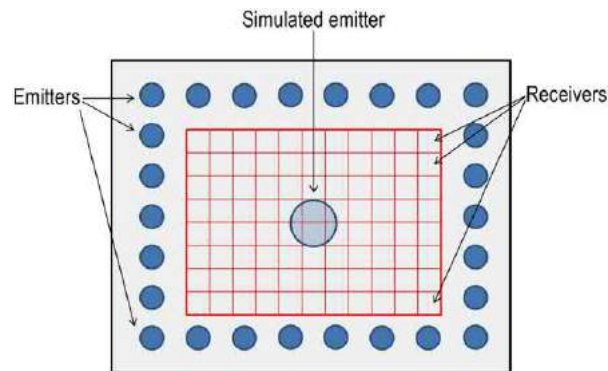


Fig. 27 Scheme of a matricial time-of-flight (ToF) camera sensor. The emitters are distributed around the matrix. Reproduced from Heptagon. SR4000/SR4500 User manual. Available at: <http://mesa-imaging.ch>.

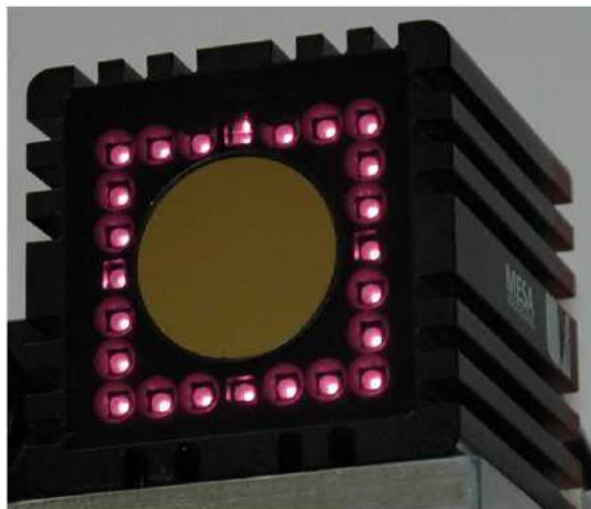


Fig. 28 The emitters of the MESA Imaging SR4000 are the red light emitting diodes (LEDs). Reproduced from Heptagon. SR4000/SR4500 User manual. Available at: <http://mesa-imaging.ch>.



Fig. 29 Microsoft Kinect play station. *Source:* Internet

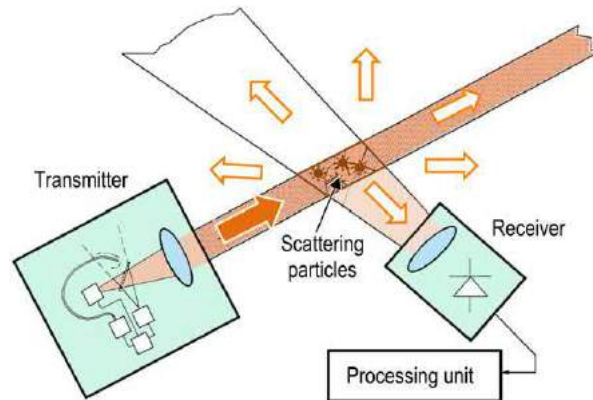


Fig. 30 Typical configuration of the lidar measuring the scatterance of the media. Reproduced from Lonnqvist, J., 1999. Apparatus and Method for Measuring Visibility and Present Weather. US Patent 5,880,836.

random speckle patterns. The spatially random patterns are seen by the sensor. Low correlation between different patterns is used for both 3D mapping of objects and range finding.

Asus X-tion Pro and X-tion Pro Live (see Section Relevant Websites) are other products with a range camera based on the same chip initially manufactured by PrimeSense. All these range cameras, support an IR video-camera and projecting the IR light coded patterns. In this technology, each pixel needs to be associated to a code-word, i.e., a specific local configuration of the projected pattern. A correspondence estimation algorithm analyzes the received code-words in the acquired images in order to compute the conjugate of each pixel of the projected pattern.

Further modifications of the pattern-projection technology are proposed involving combinations of cameras (Sali and Avraham, 2016), or adding a laser illuminator for TOF range measurement (Devaux *et al.*, 2013; Nehmadi and Guterman, 2016).

Micro-Lidars Measuring the Media Parameters

Nephelometers

The typical lidar configuration is shown in Fig. 30 which is the same for hundred-kilometers space lidars and centimeter- and even micrometer-scale micro-lidars (Lonnqvist, 1999). The angle θ between the direction of the laser beam and the optical axis of the receiver can vary from 0 (or near zero) to 180 degrees. The transmitter and the receiver can use a common optics (monostatic configuration) or, they can use separate optical systems (bistatic configuration). The second one is often preferable to exclude the scatter of the transmitted laser light in the optical system. The volume, from where the information is got, is limited in the first approximation by the section of the laser beam encircled by the receiver's field of view.

Jerlov (1976) considers scattering in the concepts of large-particle optics as the result of three physical phenomena: (1) through the action of the particle, light is deviated from rectilinear propagation (diffraction); (2) light penetrates the particle and emerges with or without one or more internal reflections (refraction); (3) light is reflected externally. The amount of scatter (scatterance β) depends on the angle θ . The function $\beta(\theta)$ is called indicatrix. The indicatrix describes the amount of light that is scattered at every possible combination of angles of incoming and outgoing light (influx and efflux angles). The character of scatter depends on how polydispersed system is. Monodispense sytem has oscillatory β curve. The most striking and distinctive feature of a polydispersed system is the pronounced forward scattering. Fig. 31 represents the function $\beta(\theta)$ of the oceanic water at different depths. It is clearly seen that at small depths the indicatrix has more oblong shape in the forward direction. Deeper oceanic waters are clearer and their indicatrix is nearer to that of the pure water.

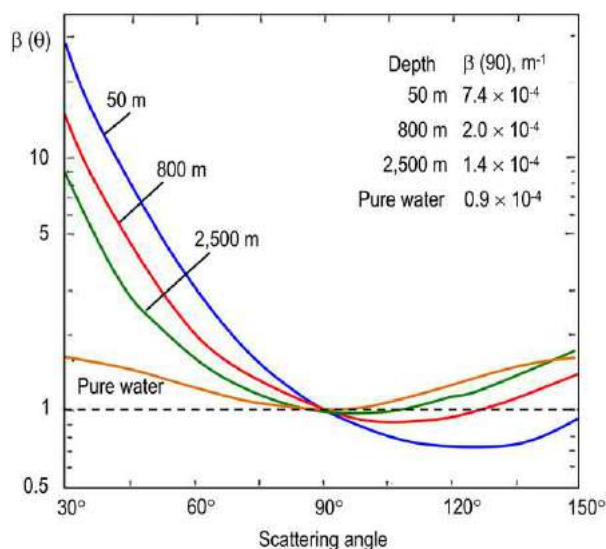


Fig. 31 Indicatrices of the oceanic water at different depths (normalized to β at 90 degree). Reproduced from Jerlov, N.G., 1976. *Marine Optics*. Amsterdam: Elsevier.

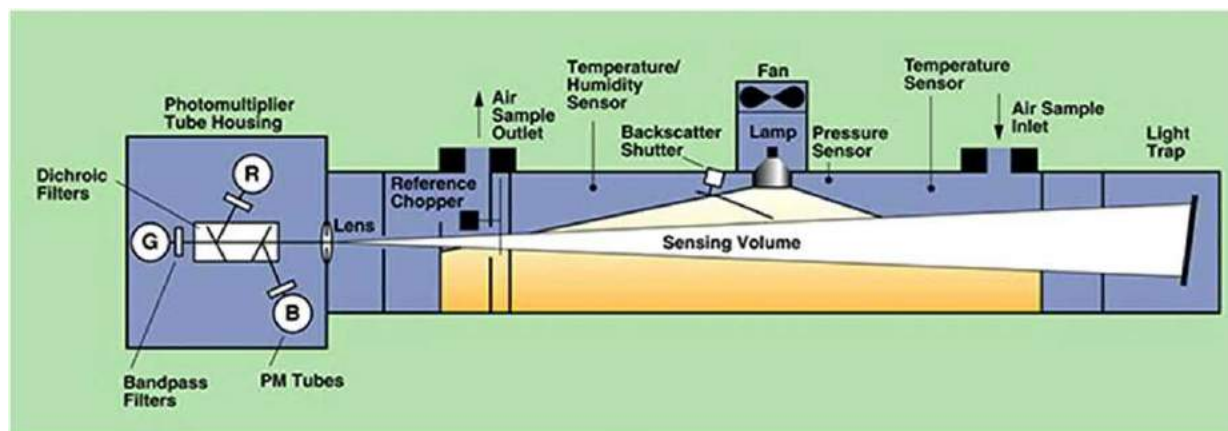


Fig. 32 TSI 3563 nephelometer with additional temperature, humidity, and pressure sensors. Reproduced from Mizrahi, A., Russell, L., Three wavelength integrating nephelometer. Available at: https://www.esrl.noaa.gov/gmd/aero/instrumentation/neph_desc.html.

The instrument for measurement of scatterance (scatterance meter) is called nephelometer. Jerlov (1976) described the nephelometer for measurement of the function $\beta(\theta)$. Angular dependence of scatterance is not always required. In some case, it is enough to have only value of scatter at a certain angle, for example, for turbidity measurement (Mitchell, 2010). For pollution measurement, or for determination of the complex refractive indices of aerosol, use only an integrated value of scatter (Han et al., 2009). At the same time, additional information can be derived when measuring the scatter at different wavelength (Anderson et al., 1996). The three-wavelength model TSI 3563 (Fig. 32) splits the scattered light into red (700 nm), green (550 nm), and blue (450 nm) wavelengths. The TSI nephelometer measures back-scattered light at these wavelengths as well. The one wavelength Radiance Research nephelometer measures forward scattering at 550 nm only.

The main body of the TSI 3563 nephelometer is a 10-cm-diameter aluminum tube, 90 cm long (Mizrahi and Russell). Along the axis is an 8-cm-diameter tube set with aperture plates. The plates range from 7 to 170 degrees on the horizontal range of the lamp. The backscatter shutter allows blocking of the angles from 7 to 90 degrees so that only backscattering is measured. The light trap provides a dark reference against which to measure the scattered light.

The receiving optics is located on the other side of the tube from the trap. The light that passes through the lens is separated by dichroic filters into three wavelengths. The first is a color splitter that passes 500–800 nm light and reflects 400–500 nm light. The reflected light passes through a filter centered at 450 nm (blue) into a photomultiplier tube (PMT). The light that passes through the first splitter goes to a second splitter which passes 500–600 nm light and reflects 600–800 nm light. The reflected light passes through a filter centered at 700 nm (red) into a second PMT. The light that passes through both splitters passes through a filter centered at 550 nm (green) into a third PMT. For further reading we recommend the review of integrating nephelometers of Heintzenberg and Charlson (1996).

Smoke Detectors

Lidar principles are widely used in office and home appliances. A smoke detector is an example. Dohl (2015) patented the smoke detector with a plurality of light emitting devices of different wavelengths and with another plurality of scattered light receiving sections. Fig. 33 demonstrates the design of the smoke detector with one such section in which the optical axes of light emitting diode and of photodetector meet in the scattering volume outside the detector. There can be several LEDs in one transmitting position (Fig. 34), for example, blue and infrared. Such coupling of the light wavelengths enables the differentiation of the type of the smoke: if the signal is higher in the longer wavelength channel, the device decides that the larger particles are contained in the smoke, and vice versa: if the signal is higher in the shorter wavelength channel, the instrument decides on the prevailing number of smaller particles. Positioning the receivers at different angles relatively to the optical axis of the light emitting channel can give the information on the complexity of the scattering particles: in the suggestion that the smoke particles have more complicated shape than the particles of the vapor, the device can suggest whether the scattering particles are of the smoke origin or, they are the water vaporization particles.

Flow Cytometry

When analyzing the smoke detector, we noted that the complexity of particle shape results in higher scatter in the lateral direction relatively to the laser beam. This feature of particle scatter is used in flow cytometry that is performed on particles in suspension (e.g., cells or nuclei). The cell suspension is injected into a flow chamber of the fluidic system, where cells are forced into single file and directed into the path of a laser beam (Craig, 2014). Laser light is focused on these single particles/cells and can be measured

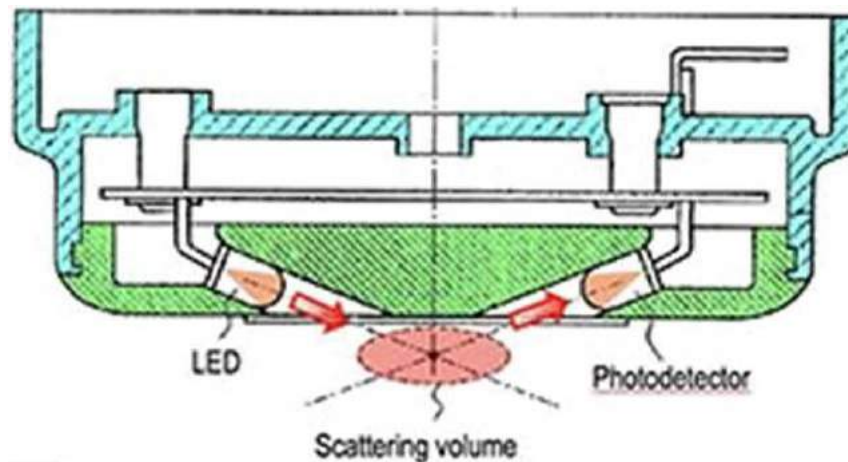


Fig. 33 One channel of the smoke detector. Reproduced from Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.

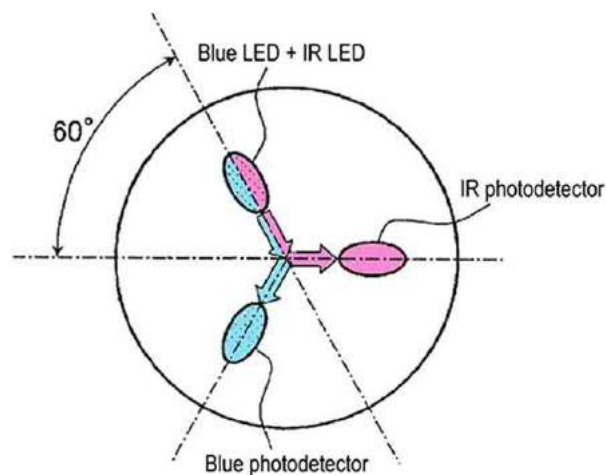


Fig. 34 Two-channel positioning in the multichannel smoke detector. Reproduced from Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.

by photodetectors as it is scattered off particles/cells when they pass in front of the beam. Fig. 35 demonstrates the principle of the Sysmex XT-2000i and XT-1800i Automated Hematology Analyzers.

When the laser beam interacts with a single particle/cell, light is scattered but its wavelength is not altered (elastic scattering). The amount of scattered light can be used to identify the particle/cell because it is related to the physical properties of the particle (size, granularity, and nuclear complexity). Light scattered at a 90 degree angle (side light scatter) is related to the internal complexity and granularity of the particle/cell. Neutrophils produce much side scatter because of their numerous cytoplasmic granules. Light that proceeds in a forward direction (forward light scatter) is related to the particle size. Large cells produce more forward scatter than small cells. As illustrated in Fig. 36, differences in light scattering are used to distinguish the particles: the

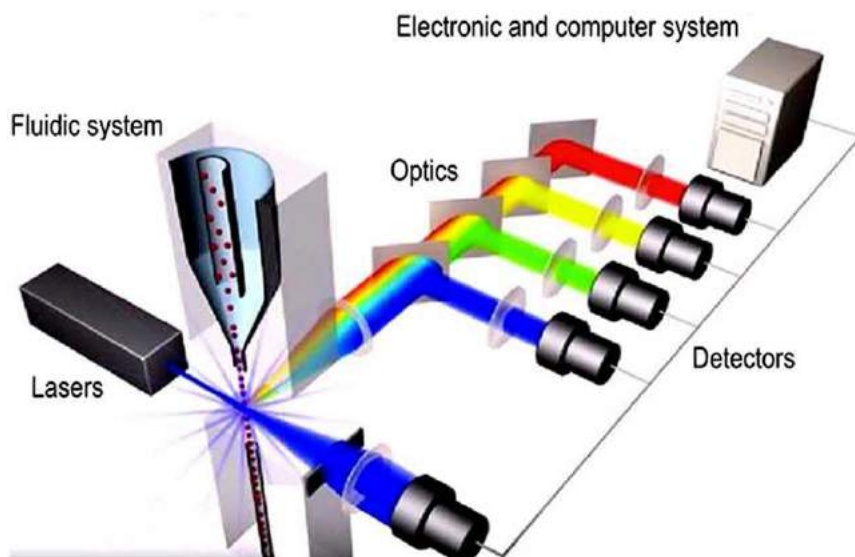


Fig. 35 Configuration of the Sysmex XT-2000i and XT-1800i automated hematology analyzers. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), *Clinical Laboratory Hematology*, second ed. Essex: Pearson, pp. 956–974.

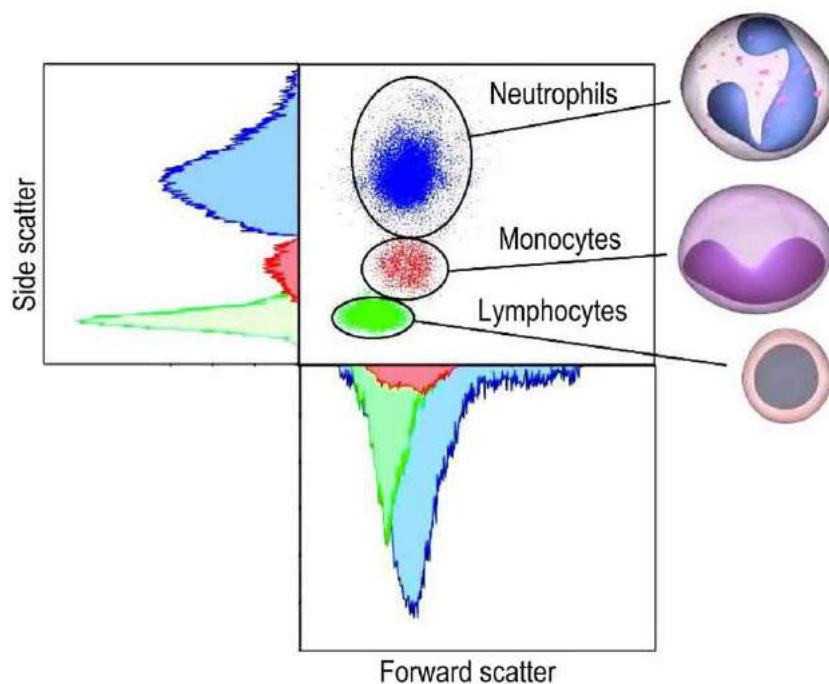


Fig. 36 Distinguishing the particles. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), *Clinical Laboratory Hematology*, second ed. Essex: Pearson, pp. 956–974.

lymphocytes as the smallest ones are in the low side-signal domain, and the neutrophils, being the biggest ones having the most complicated shape are in the high side-signal domain.

In addition to elastic light scattering, flow cytometers are used to detect bound fluorescent markers (fluorochromes), which are molecules that are excited by light of one wavelength and emit light of a different wavelength (fluorescent light). Fluorochromes can label the antibodies or specific changes in DNA or RNA (**Figs. 37** and **38**). Flow cytometers use light of a single wavelength generated by a laser to excite fluorochromes bound to the particle of interest. Light emitted from the fluorochrome is separated from the incident laser light using a combination of filters and mirrors. A PMT then detects and quantifies the emitted light.

Clinical flow cytometers usually contain an argon laser that generates light at 488 nm. This single wavelength is often used to excite three different fluorochromes, each emitting light at different wavelengths. Using three different fluorochromes allows detection of three different antigens on the cell. Excitation of additional fluorochromes to detect further antigens usually requires a second laser light source (e.g., helium neon, emission 633 nm). Two laser light sources allow to identify up to six antigens on each cell analyzed.

Raman Micro-Lidars

In an inelastic scattering of a photon, some of its energy is lost or increased. Correspondingly, the frequency of the photon is shifted to red or blue. A red shift means that part of the photon energy is transferred to the interacting matter (Stokes Raman scattering). In the blue shift, internal energy of the matter is transferred to the photon (anti-Stokes Raman scattering). Detecting the Stokes or anti-Stokes lines in the scattered spectrum is related to the spectroscopy. Still, in practice, when applied to the techniques of studying the properties of atmosphere, the terms “lidar,” “Raman lidar” are used. When coming to the microworld, we may use the term “Raman micro-lidar.” Detecting the inelastic scattering and its usage for measuring the parameters of the medium has a wide field of applications. We recommend for further reading the book edited by [Baudalet \(2013\)](#).

The role of the Raman micro-lidar is to make an “optical biopsy,” i.e., to diagnose disease such as cancer and atherosclerosis without removing tissue from the body in the way by revealing the differences between cancerous and normal tissues of various organs due to morphological and molecular changes in the tissue. We shall refer here to key optical methods of Raman inelastic

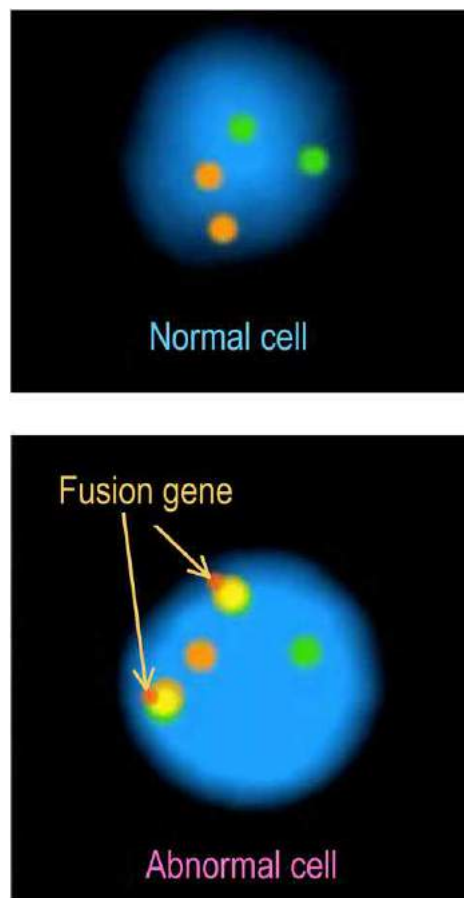


Fig. 37 Normal cell and the abnormal cell with fused genes. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), *Clinical Laboratory Hematology*, second ed. Essex: Pearson, pp. 956–974.

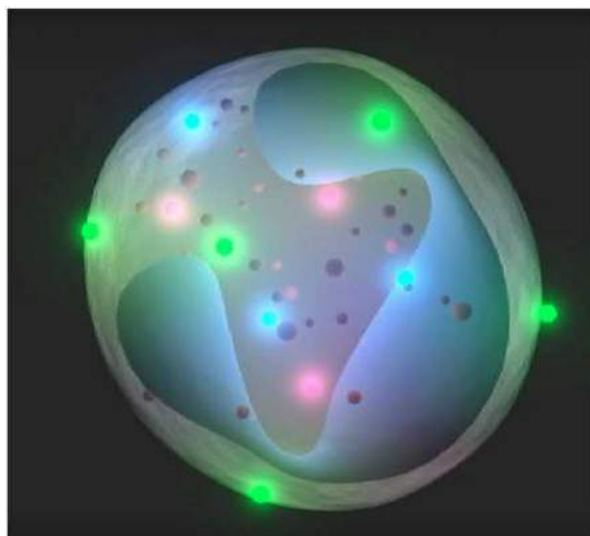


Fig. 38 Neutrophil with fluorescent markers. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. Essex: Pearson, pp. 956–974.

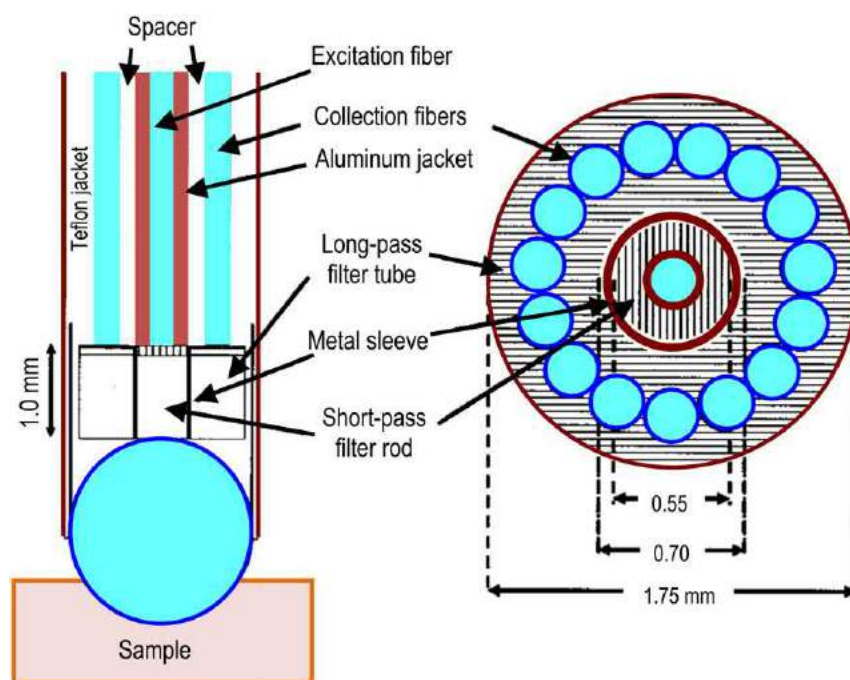


Fig. 39 Schematic of the Raman probe tip. Ball lens B is in contact with the filter module that couples to the fiber bundle. Reproduced from Motz, J.T., Hunter, M., Galindo, L.H., *et al.*, 2004. Optical fiber probe for biomedical Raman spectroscopy. Applied Optics 43, 542–554.

backscattering reflectance and time-resolved spectroscopy. Various human tissue types (prostate, breast, lung, colon, and gastrointestinal) have been studied using optical biopsy (Jelínková, 2013).

The first work on using the Raman scattering to detect cancer was published by Alfano *et al.* (1991). Alfano and Pu (2013) summarized the achievements in the mentioned book of Jelínková. The peak height ratios for Raman data analysis were able to differentiate normal tissue and malignant tumor in breast, gynecological, and cervical tissue. These basis spectra represent the epithelial cell cytoplasm, cell nucleus, fat, β -carotene, collagen, calcium hydroxyapatite, calcium oxalate dihydrate, cholesterol-like lipid deposits, and water.

Motz *et al.* (2004) and Haka *et al.* (2006), both from MIT, developed an optical fiber Raman probe and collected Raman spectra of breast tissue. Their portable Raman system with optical fiber probe is schematically shown in Figs. 39 and 40. An 830-nm-diode laser is focused into the excitation fiber of the Raman probe through a bandpass filter. Consisting of a single central excitation fiber

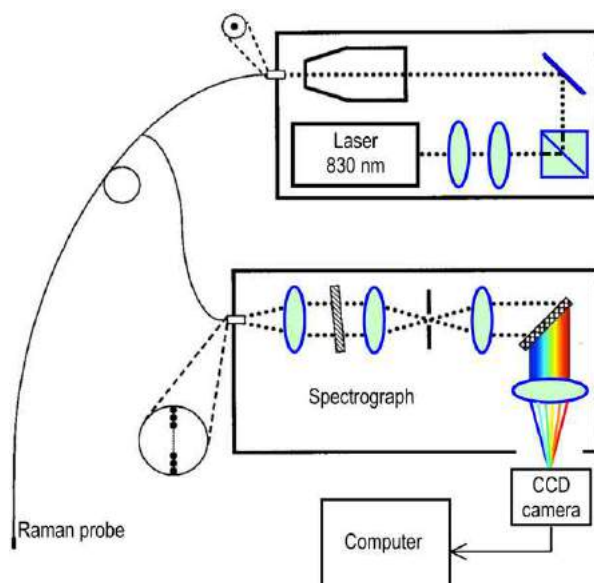


Fig. 40 Schematic of the Raman spectroscopy system used for experimental testing of the Raman probe tip. Reproduced from Haka, A.S., Volynskaya, Z., Gardecki, J.A., *et al.*, 2006. In vivo margin assessment during partial mastectomy breast surgery using Raman spectroscopy. *Cancer Research* 66 (6), 3317–3322.

surrounded by 15 collection fibers of 200 μm core diameter, the diameter of the probe is less than 2 mm. The linear array of collection fibers at the proximal end is coupled to a spectrograph.

Light delivery and collection can be accomplished by using optical fibers that can be incorporated into a biopsy needle (Haka *et al.*, 2005). As opposed to biopsy, a spectroscopic needle measurement has the advantage of providing immediate diagnosis. With the development of minimally invasive breast cancer therapies, such as radiofrequency ablation (Fornage *et al.*, 2004), there is the potential for diagnosis and treatment to be performed in a single procedure.

Working on even less invasive instrument, Komachi *et al.* (2005, 2009) designed a micro-Raman probe with a 600 μm diameter, that can be inserted into coronary arteries. Custom-made filters attached to the front end of a probe eliminate the background Raman signals of the optical fiber itself. Measurement of the Raman spectra of an atherosclerotic lesion of a rabbit artery in vitro demonstrated its excellent performance.

Further studies resulted in a Raman spectroscopy technique to simultaneously identify micro-calcification status and to diagnose the underlying breast lesion, in real time, during stereotactic core needle biopsy procedures (Barman *et al.*, 2013). An approach was introduced based on diffuse reflectance spectroscopy for detection of micro-calcifications that focuses on variations in optical absorption stemming from the calcified clusters and the associated cross-linking molecules. A new powerful decision algorithm was derived from correlating the diffuse reflectance spectra with the corresponding radiographic and histologic assessment.

Time-Resolved Fluorescence Spectroscopy

Fluorescence spectroscopy is one of the optical techniques used to investigate biomolecules whose photoexcited emission is characterized by spectral distribution, photon yield, lifetime of the excited state, as well as polarization. The first measurements were all performed by methods that involved the human eye. Significant advances in instrumentation, particularly in regard to time-resolved measurements, were achieved with the advent of lasers. Time-resolved fluorescence spectroscopy derived short-pulse lidar principles and can provide information on not only the location and intensity of the key biomolecules but also their local biophysical microenvironments (Alfano and Pu, 2013). It also provides temporal information on the underlying dynamics of molecular processes concerning the chromophores and the environment responsible for the fluorescence in tissues. Picosecond techniques enable the study of dynamics of relaxation in macromolecules.

The experimental arrangement for the time-resolved fluorescence measurements of the human breast tissues is schematically shown in Fig. 41. Ultrafast laser pulses of 100 fs duration, 0.1 nJ per pulse, at 620 ± 7 nm from a colliding pulse mode-locked dye laser system at a repetition rate of 82 MHz is used to pump the samples. These laser pulses are amplified and their frequency is doubled in KDP crystal resulting in 310 nm wavelength. The resulting fluorescence emission is collected and directed onto the slit of a synchro-scan streak camera with a temporal resolution of 16 ps. A narrow band filter of 340 ± 5 nm is used to collect the tissue fluorescence and a 310 nm notch filter – to cut off the excitation wavelength. The recorded temporal profiles are analyzed to obtain temporal and polarization information.

Typical time-resolved fluorescence profiles excited by 310 nm from malignant tumor and from normal tissue at 340 nm emission band are shown in the inset of Fig. 41. There is a marked difference between the curve profiles of malignant and

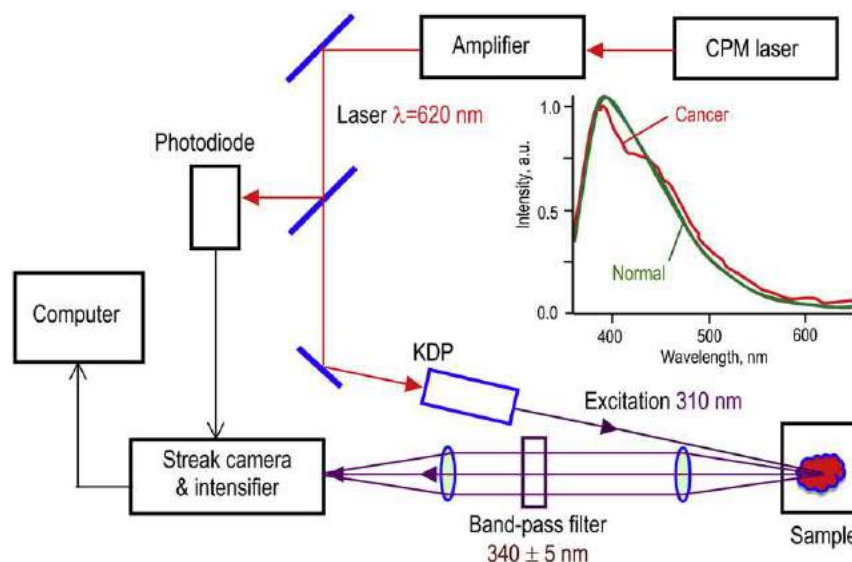


Fig. 41 Time-resolved fluorescence experimental setup and fluorescence profiles with 310 nm excitation and 340 nm emission: malignant tumor (red), normal breast tissue (blue). Reproduced from Alfano, R., Pu, Y., 2013. Optical biopsy for cancer detection. In: Jelíková, H. (Ed.), *Lasers for Medical Applications*, Cambridge: Woodhead, pp. 325–367.

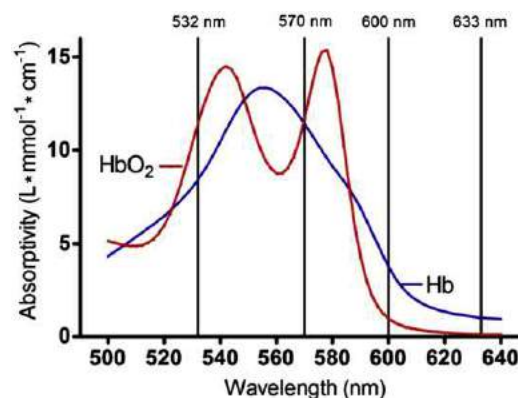


Fig. 42 Absorptivity of oxygenated and deoxygenated hemoglobin. Reproduced from Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. *Applied Optics* 27, 1113–1125.

nonmalignant sample. All temporal curves are analyzed to compare the slow and fast components of fluorescence decay. Also, the amplitude ratios of fast to slow components are analyzed to find the differences.

Oxymetry Lidars

The degree of blood oxygen saturation is one of the diagnostic parameters. Different absorptivity of light propagating in the blood with oxygenated and deoxygenated hemoglobin depending on the light wavelength (Fig. 42) is the basis for pulse oximetry. The technique of pulse oximetry using the cost-effective and patient-friendly sources of light is described in the book of Moyle (2002).

The oxygen supply of the retina is no less important for the human vision. The supply is provided by both the choroidal and retinal circulation. The choroid serves as the oxygen source for the photoreceptors in the avascular outer retina, whereas the retinal circulation plays a critical role, mediated by its autoregulatory capacity, in maintaining the oxygen supply to the neural elements and nerve fibers in the inner retina. Because of the high oxygen needs of the retina (cerebral tissue), any alteration in circulation such as seen in diabetic retinopathy or vascular occlusive diseases results in functional impairment and extensive retinal tissue damage.

Measuring oxygenation of blood could help to assess the circulatory and respiratory condition of patients. The retina is highly metabolically active tissue and information on retinal and choroidal oxygenation would give vital clinical information on the metabolic state of the retina, which may improve the understanding of retinal metabolism in health and disease. Malfunction of the retinal vasculature can result in serious eye diseases, which can affect vision considerably, for example, central retinal arterial occlusion or central retinal vein occlusion.

But how to get information on oxygenation of blood vessels from the bottom of the eye? The principle of pulse oximetry cannot be applied directly since it operates in the transmissive mode. Hickham *et al.* (1963) proposed to measure blood oxygenation in retinal vessels crossing the optic disk by using a two-wavelength (510 nm and 640 nm) photographic technique. This was the first time the oxygenation of the human retina was measured noninvasively, and following these studies many techniques have been developed. Laing *et al.* (1975) modified the instrument having used two other wavelengths (470 nm and 515 nm).

These methods involved laborious microdensitometric analysis of photographic negatives obtained under controlled conditions. To take into account the light scattered by the red blood cells (RBCs) in whole blood, the introduction of a calibration procedure was needed. To overcome the problem, Pittman and Duling (1975) introduced three closely spaced wavelengths. The contribution of scattering to the optical density at each wavelength was determined from optical density values at two isosbestic wavelengths (546 and 520 nm) and the optical density at the third wavelength (555 nm).

Delori *et al.* (1983) implemented this idea using a fundus camera adding there a photoelectric system with control and processing electronics (Fig. 43). Actually, it was the transformation of the fundus camera into a micro-lidar (Delori, 1988).

Most of the knowledge on retinal oxygenation comes from animal studies. Hardarson (2012) describes in his PhD thesis the retinal oximeter Oxymap designed at the University of Iceland which is based on a fundus camera to which an image splitter is attached to simultaneously capture four images of the same area of the fundus (Fig. 44). The images are captured at 542 nm, 558 nm, 586 nm, and 605 nm. All analyses were performed with the 586 nm (isosbestic) and the 605 nm (non-isosbestic) images. The 542 nm and 558 nm images were not used. Discarding of 542 nm and 558 nm was based on earlier testing, which revealed that optical density ratios, using these wavelengths gave less reliable results.

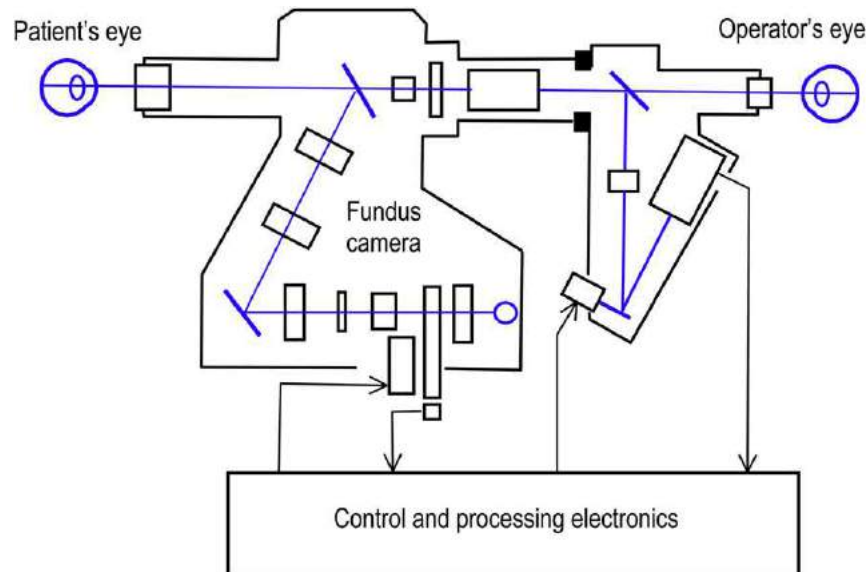


Fig. 43 Fundus camera with incorporated photoelectric system. Reproduced from Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. *Applied Optics* 27, 1113–1125.

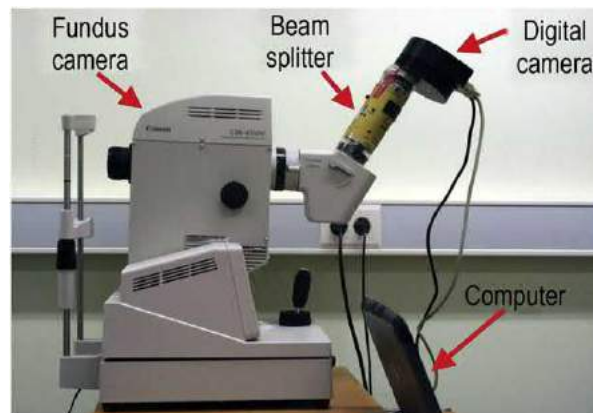


Fig. 44 Retinal oximeter on the platform of Canon CR 6-45NM. Reproduced from Hardarson, S.H., 2012. Retinal oximetry. PhD Thesis, University of Iceland, Reykjavik.

Kristjansdottir (2014) described the Oxymap T1 instrument developed at the University of Iceland (Fig. 45 left) based on the Topcon TRC-50DX mydriatic retinal camera. Two highly sensitive digital cameras, image splitter, optical adapter, and light filters are added. A fundus image captured by camera is split into two by the image splitter. One camera captures the image at 570 nm while the other captures the same area of the fundus simultaneously, at 600 nm (Fig. 46). This is done by inserting two narrow 5 nm band pass filters into the light path to each camera of the oximeter. Another light filter (80 nm band pass filter, 585 nm

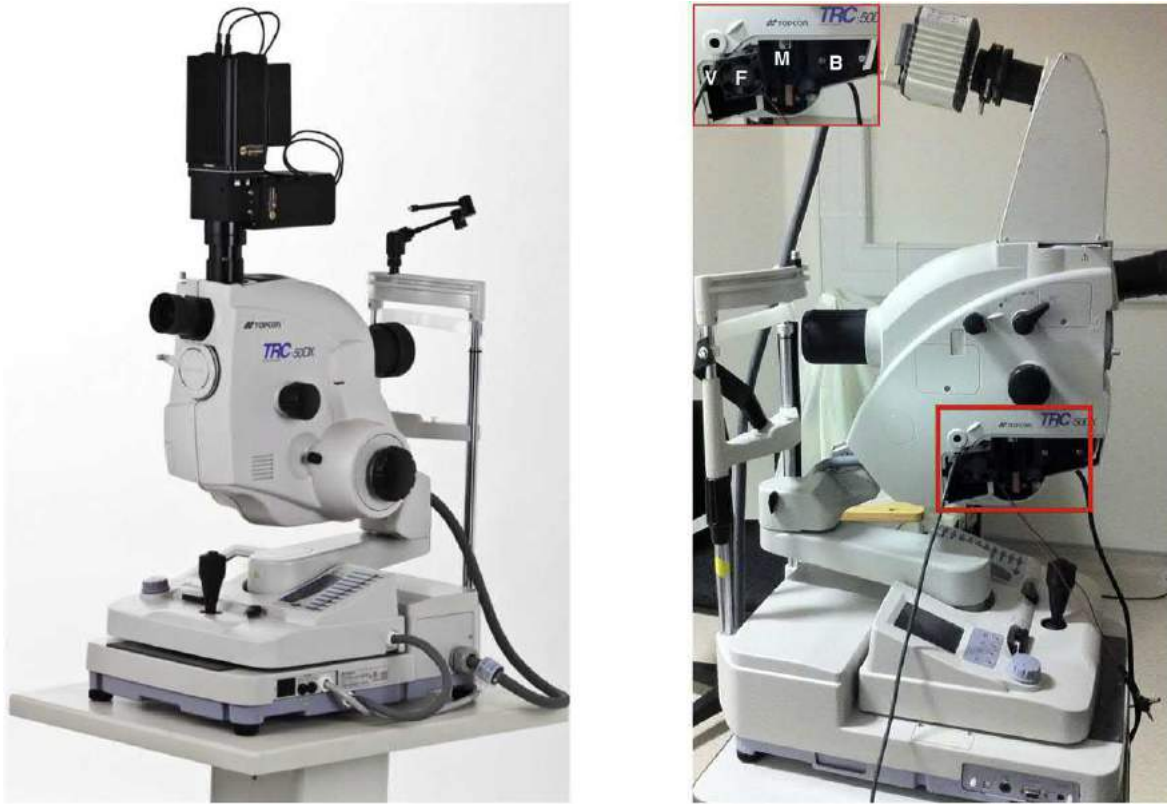


Fig. 45 Two variants of redesigning the Topcon TRC-50DX mydriatic retinal camera: left – (Kristjansdottir, 2014); right – white light was emitted by the standard 100W bulb (B), which was filtered to select a single wavelength of light by an optical filter. Reproduced from Holm, S.P., 2014. Optical imaging of retinal blood flow: studies in automatic vessel extraction, alignment, and driven changes in vessel oxymetry. PhD Thesis, University of Manchester.

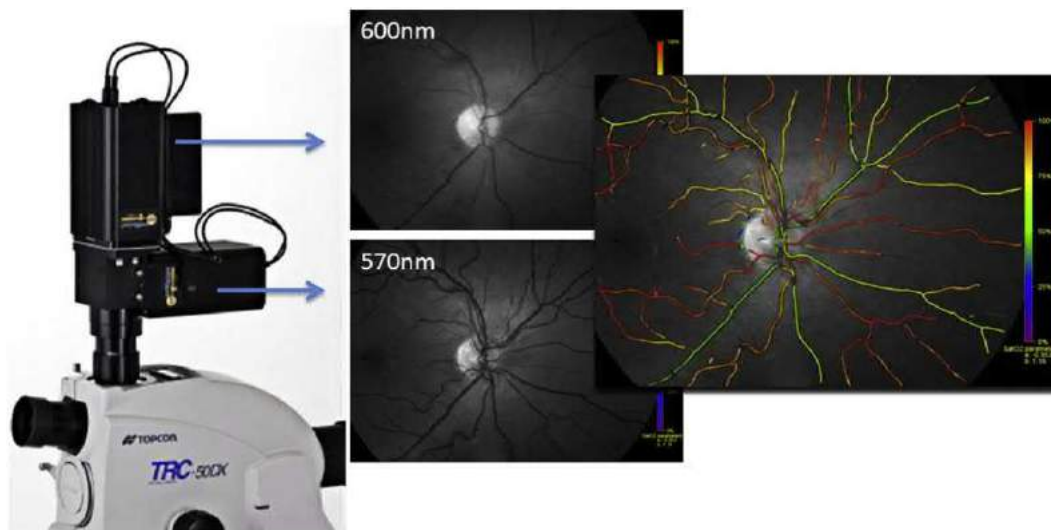


Fig. 46 Reconstruction of the retinal oxygenation map with the Oxymap T1 instrument. Reproduced from Kristjansdottir, J.V., 2014. Choroidal and retinal oximetry. MSc Thesis, University of Iceland Reykjavik.

center wavelength) is also inserted into the fundus camera itself to exclude unnecessary light exposure to subjects' eyes, only allowing light between 545 and 625 nm to exit the camera lens.

A specialized software, Oxymap Analyzer, was developed for calculation of blood vessel oxygen saturation. It analyses two spectral images (570 and 600 nm, the isosbestic is 570 nm wavelength), automatically detects blood vessels, selects measurement points and measures the brightness on vessels at each wavelength. From their ratio, the optical density is calculated. Examples of oxygenation maps of healthy eye and the eye with central retinal vein occlusion are given in Fig. 47. A clear difference is seen in venular oxygen saturation (red color indicates 100% and violet indicates 0% oxygen saturation).

Another approach was introduced by Holm (2014). He also used the Topcon TRC-50DX mydriatic retinal camera (Fig. 45 right) but modified the processing just subtracting one image from another. He developed the intensity-based algorithm using the approach of Saleh and Eswaran (2012) aimed at reducing background variations within the images to improve the contrast between the retinal vasculature and the background tissue, and to reduce the amount of segmentation noise. In the intensity-based algorithm only the green channel is used since it has the maximal contrast of the vasculature. The weighted sum of all three color-channels is used for Gabor filtering. A review of recent achievements in the field of oxymetry is given in the publication of Li *et al.* (2015).

Glucose Lidars

A wide range of optical technologies have been designed in attempts to develop robust noninvasive methods for glucose sensing in biological fluids, especially in human blood (Tuchin, 2009). The methods include infrared absorption; near-infrared scattering; Raman, fluorescent, and thermal gradient spectroscopies; as well as polarimetric, polarization heterodyning, photonic crystal, optoacoustic, optothermal, and OCT techniques. None of them is in a wide clinical use. We shall restrict our interest with two of them that could be easier qualified as lidar techniques.

Differential absorption glucose content measurement

Zhou *et al.* (2011) presented a method of glucose concentration detection in the anterior chamber with a differential absorption optical low-coherence interferometry technique. The anterior chamber is located between the cornea and the lens, with the thickness of 3.13 ± 0.50 mm. It is filled with transparent liquid-aqueous humor, with a total volume about 250 μ L. In the experiments, the cornea and aqueous humor can be treated as nearly non-scattering substance. The difference in the absorption coefficient is much larger than the difference in the scattering coefficient, thus the influence of scattering can be neglected. The light directed on the iris being back-reflected passes the anterior chamber twice (Fig. 48).

Two light sources, one centered within (1625 nm) and the other centered outside (1310 nm) the glucose absorption band, are used for differential absorption measurement. Schematic diagram of the differential absorption system is shown in Fig. 49. The notations are as follows: FC – fiber coupler; OF – optical filter; WDM₁ – wavelength division multiplexer; WDM₂ – wavelength demultiplexer; M – vibrating mirror; L – lens. The light emerging from the eye carries the information of glucose concentration.

In the eye model and pig eye experiments, the authors obtained a resolution glucose level of 26.8 mg/dL and 69.6 mg/dL, respectively. This method has a potential application for noninvasive detection of glucose concentration in aqueous humor, which is related to the glucose concentration in blood.

Polarization-based glucose content measurement

Malik and Cote (2010) described their experiments with the idea to use the polarization parameters of anterior chamber aqueous humor to measure the glucose content. Glucose sensing using optical polarimetry is based on the phenomenon of optical activity,

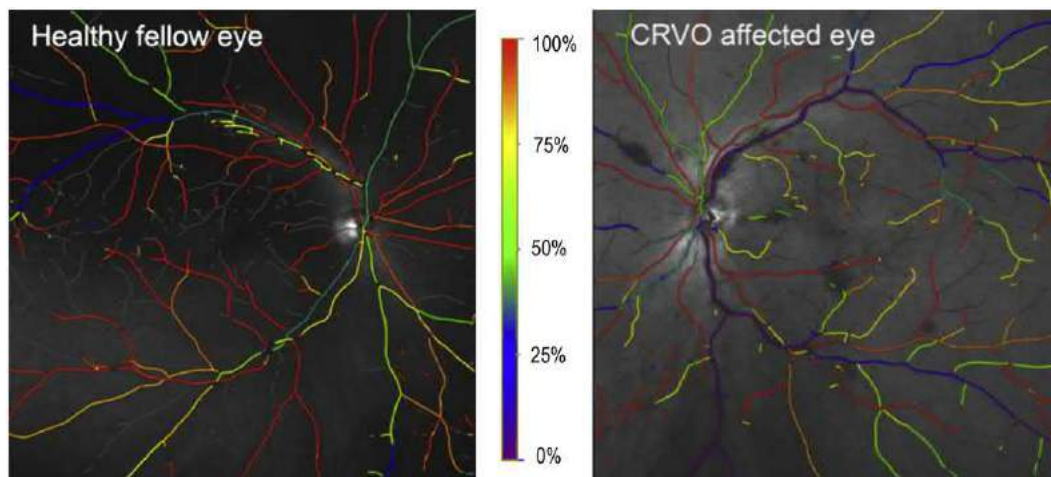


Fig. 47 Oxymetry images from patient eyes: healthy eye (left) and the eye with central retinal vein occlusion (right). Reproduced from Kristjansdottir, J.V., 2014. Choroidal and retinal oxymetry. MSc Thesis, University of Iceland Reykjavik.

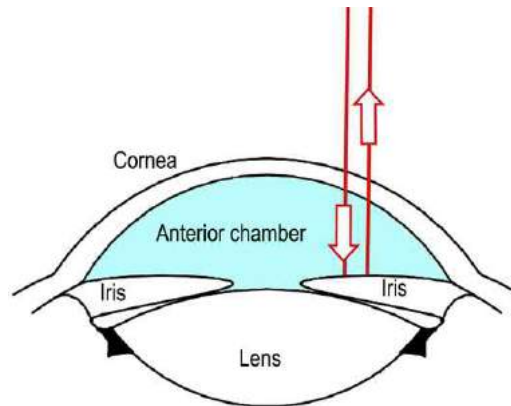


Fig. 48 Optical path through the anterior chamber with back reflection from iris. Reproduced from Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. *Journal of Biomedical Optics* 16 (1), 015004.

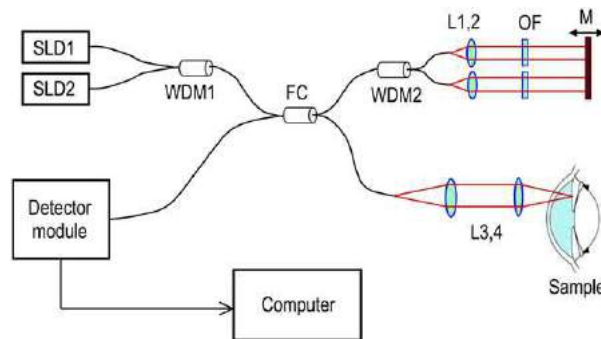


Fig. 49 Schematic diagram of the differential absorption system. Reproduced from Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. *Journal of Biomedical Optics* 16 (1), 015004.

which is the rotation of the orientation of plane polarized light passing through a solution of optically active molecules. The idea of the experiment was to measure the polarization parameters of the light that has passed through the cornea across the anterior chamber of the eye (inset of Fig. 50). The light has to be incident at a relatively glancing angle with respect to the posterior corneal surface in order for the beam to exit the anterior chamber through the cornea.

The problem, the authors met, is the birefringence of the cornea. In polarimetric glucose sensing through the anterior chamber of the eye, this corneal birefringence masks the optical rotation due to glucose in the aqueous humor, and therefore acts as a noise source. When the cornea moves due to eye motion, its birefringence changes the polarization of the light and this is a significant source of polarization noise. In addition, the index mismatch between the cornea and air causes the light beam to bend as it propagates and this complicates the ability to couple the light across the anterior chamber of the eye. It was demonstrated that change in the polarization vector due to corneal birefringence is at least an order of magnitude larger than that due to the change in optical path length and glucose concentration.

In his PhD work Malik (2011) made an effort to understand whether corneal birefringence is wavelength independent. This is important to know since the assumption is that by using multiple wavelengths this noise source may be canceled out. He designed a prototype and used two wavelengths: 635 nm and 532 nm. Experimental setup of the dual-wavelength optical polarimeter is presented in Fig. 50. The notations are as follows: P – polarizer; FC – Faraday compensator; BS – beam splitter; FM – Faraday modulator; PD – photodetector; LIA – lock-in amplifiers. The conclusion of this study was that using a dual-wavelength polarimetric approach has a potential to minimize the corneal birefringence noise and thus quantify blood glucose concentration.

Micro-Lidars for Velocity Measurement

Laser Velocimetry

Velocity measurement was another option that was tried as one of the first applications for lasers as transmitters in laser radars. The earliest publication was made by Yeh and Cummins (1964) from Columbia University who used a microscale lidar to study flow patterns in fluids by injecting dyes into the flow stream, illuminating the stream with monochromatic laser light, and measuring

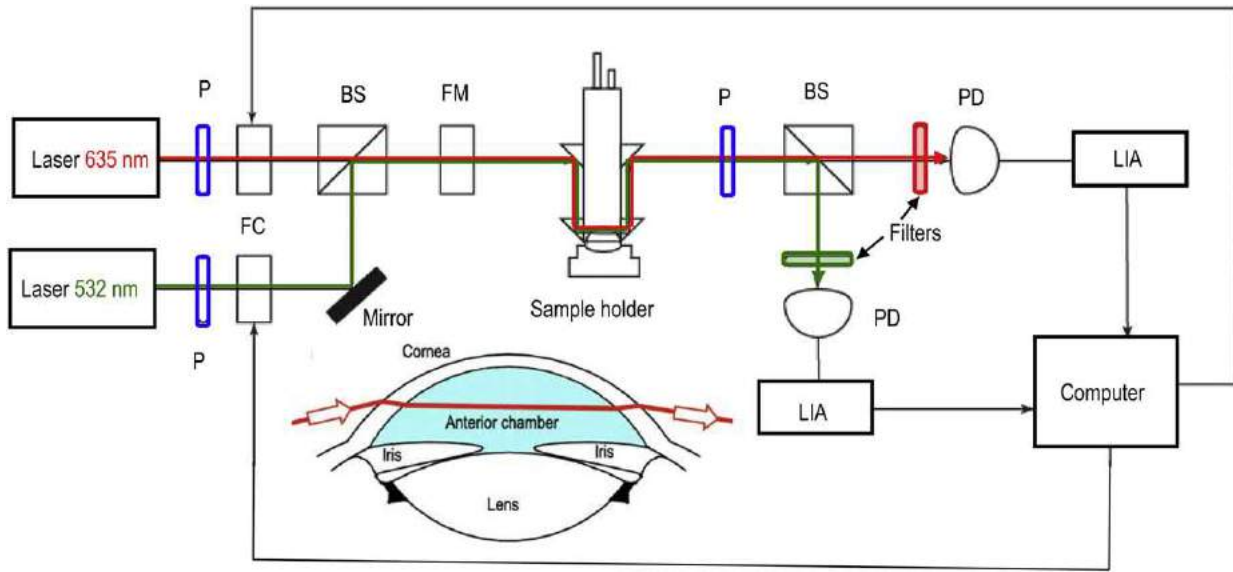


Fig. 50 Experimental setup of the dual-wavelength optical polarimeter. Reproduced from Malik, B.H., 2011. Dual wavelength polarimetry for glucose sensing in the anterior chamber of the eye. PhD Dissertation, Texas A&M University, College Station.

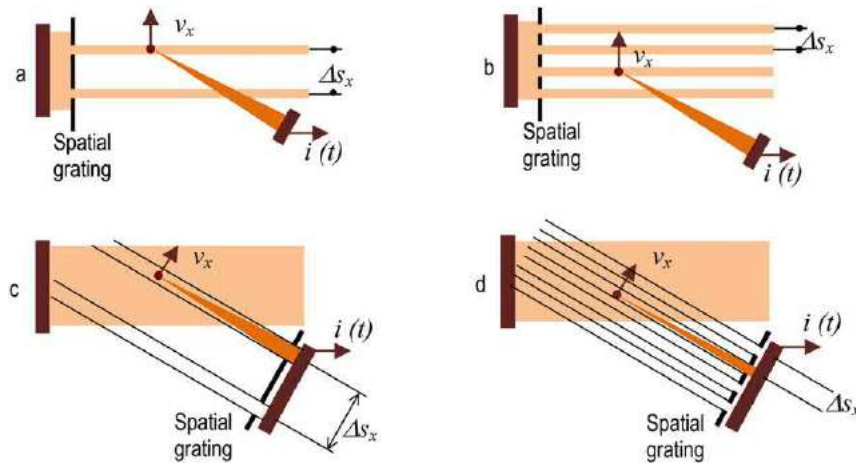


Fig. 51 Flow measurement techniques using an optically fixed measurement volume. ((a),(b)) Show the versions of the illumination by strips of light and ((c),(d)) show formation of detected zones at the receivers. Reproduced from Albrecht, H.-E., Borys, M., Damaschke, N., Tropea, C., 2003. Laser Doppler and Phase Doppler Measurement Techniques. Berlin: Springer.

the Doppler shifts in the Rayleigh scattered light. Twelve years later, the so-called “laser dual beam method” was patented (Schodl, 1976), according to which, two adjacent remote focal points are formed, where the light beams are focused. The particles in the gas stream passing through the focal points reflect light, generating start and stop pulses in the detectors, their time difference being used to calculate the velocity of the stream. Rotating the plane of the pair of the beams, the velocity can be determined in two dimensions. The first of these methods that measured the radial component of the velocity required monochromatic light, whereas the second one measuring the transversal components of the velocity is not limited by the degree of the coherence of light.

Analyzing different approaches to modeling and classification of the methods of velocity measurement, Molebny (1981) paid attention that the model of any of the known methods can be designed to include a field of readout points or readout strips, which would serve as references to measure the velocity. The readout points can be created by focusing the beams of light in adjacent points of space, or creating interference strips in the volume under study. These readout points can be immovable, fixed in space, like in the second of the above mentioned methods, or they can move (run), like in the field of interference of two beams of light with differing wavelengths. The interference can take place in the volume under study, or in the detector plane. Following this classification, we can get the information on the velocity from the field of fixed points, or from the field of moving points.

Fig. 51 left (a) and (c) illustrates the principle of measuring the velocity when the scattering particle crosses two zones. In the case (a) they are the zones (strips) illuminated by light, and the receiver sees both of them. In the case (b) the zones are created by

the fields of view of the receivers. For a single scattering particle, the interval Δt is measured. For multiple particles, the correlation interval Δt is measured.

Fig. 51 right (c) and (d) illustrates multiple zones created by a spatial grating at the transmitter side (a) or at the receiver side (b). In both cases the frequency is measured that is created by bursts of scattered light when the moving particle crosses the illuminated zone.

The book summarizing the first decade of intensive studies of the new instrument – laser Doppler Anemometer (LDA) was published by Durst *et al.* in 1976, its second edition appeared in 1981 (Durst *et al.*, 1981). For further reading we recommend later publications of Albrecht *et al.* (2003), Zhang (2010), and specifically for ocular blood flow – the book of Schmetterer and Kiel (2012).

Blood flow measurement in retinal vessels

Successful application of laser velocimetry for flow measurements was described with application to blood vessels. The feasibility of using laser Doppler velocimetry (LDV) to measure blood flow in individual retinal vessels was first demonstrated by Riva *et al.* (1972), patented later (Riva, 1979). In accordance to the patent, the laser beam was directed into the eye (Fig. 52), the light scattered from the retina was received by the detector. After the detector, the signal was amplified, a loudspeaker was used as a measuring device. The type recorder and other devices was also foreseen for analysis. The laser light scattered from the retina contained the components scattered by RBCs flowing in the retinal vessels and the light scattered from the walls of the vessels and other structures that did not move (Fig. 53). At the output of the detector, the difference frequency was filtered and amplified. It was the Doppler-shift frequency spectrum that was further analyzed. In the experiments, the blood flow in a retinal artery of an anesthetized rabbit was studied. The maximum Doppler frequency shift arising from the light scattered by the RBCs flowing at the maximum speed was estimated from the spectrum and from the intraocular scattering geometry.

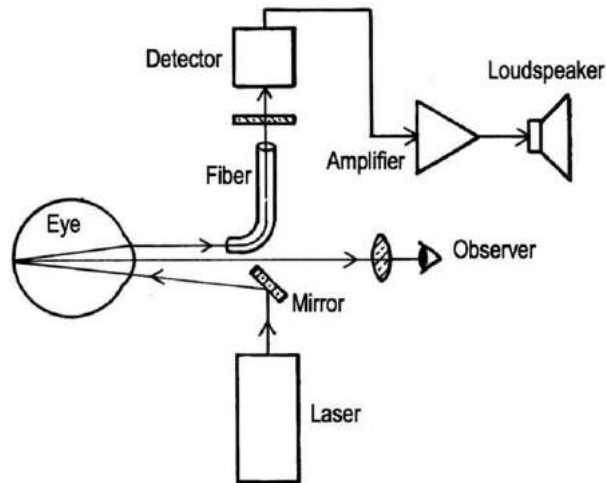


Fig. 52 Schematic of laser Doppler velocimeter patented by Riva in 1979. Reproduced from Riva, C.E., 1979. Blood Flow Measurement. US Patent 4,142,796. March 6, 1979.

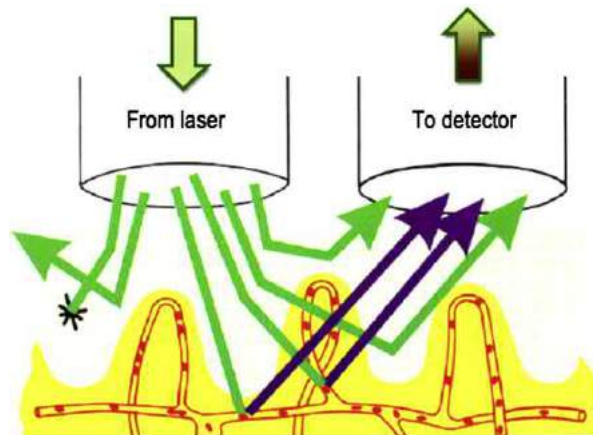


Fig. 53 Light scattered from moving (violet) and nonmoving (green) structures creating the Doppler-shift frequency spectrum. Source: Internet

Two-directional velocity measurement

Several years later, a new approach for determining blood velocity was proposed (Riva *et al.*, 1979, 1983). The procedure involves collecting the light scattered by the RBCs in two distinct directions separated by a known angle. Subsequent analysis yields an absolute measure of the velocity that is independent of the exact orientation of the vessel and of the relative angular orientation of the incident and scattered light beams with respect to the flow direction.

Velocity Measurement With Optical Doppler Tomography

Conventional LDV found its application to measure mean blood perfusion in the peripheral microcirculation. However, strong optical scattering in biological tissue limits the spatial resolution of flow measurements by LDV. Chen *et al.* (1997) described the OCT technique that provides not only in vivo blood flow measurement at discrete locations but also the tissue structure surrounding the vessel. Blood flow in a vein is imaged in Fig. 54, displaying a color-coded velocity image, and a velocity profile along the vein cross-section. The magnitude of blood flow velocity is maximal at the vessel center and decreases monotonically toward the peripheral wall. In a horizontal cross-section of the velocity image near the vessel center, an excellent fit of the velocity profile to a parabolic function indicates that blood flow in the vein is laminar.

Time dimension in velocity measurement

Ma *et al.* (2010) introduced a novel method to measure absolute blood flow velocity in vivo. It measures the velocities of blood plasma across the heart outflow tract. In experiments involving a chicken, the authors acquired four-dimensional (4D) $[(x, y, z) + t]$ images of absolute velocity distributions of the blood plasma with high spatial and temporal resolution in vivo. They reconstructed 4D microstructural images and obtained the orientation of the heart outflow tract at its maximum expansion. Assuming flow is parallel to the vessel orientation, the obtained centerline indicated the flow direction. Using this method, the flow velocity profiles were compared at various positions along the heart outflow tract of the chicken embryo.

OCT capillary velocimetry

OCT found its applications also in cardiovascular diagnostics and studies of cranial vascularization, etc. To demonstrate applications of OCT capillary velocimetry in cerebrovascular research, a 1310 nm spectral/Fourier domain OCT microscope was used for in vivo imaging of the rat cortex (Srinivasan *et al.*, 2012). The light source consisted of two SLDs yielding a spectral bandwidth of 170 nm. The axial (depth) and transverse resolution was 3.6 μm in tissue, full-width-at-half-maximum, and the imaging speed was 47,000 axial scans per second (17.3 μs exposure time), achieved by an In GaAs line scan camera. Capillary velocimetry and its imaging was performed during a hypercapnic challenge in a rat. An example of such imaging is given in Fig. 55.

Speckle contrast tomography

A novel tomographic method based on laser speckle contrast was introduced by Varma *et al.* (2014) that allows the reconstruction of a 3D distribution of blood flow in deep tissues. The experimental setup uses a temperature controlled continuous laser diode (785 nm, 90 mW) to probe the sample. A pair of galvo controlled mirrors are used to scan the laser point source. The light source is focused on the bottom of the sample and the produced speckle patterns are imaged from the top with a monochrome scientific CMOS camera, with exposure time 1 ms. The horizontal field of view is about 4 cm, a pixel diameter is 3×10^{-4} cm. The laser was set in every position during 0.5 s to acquire 35 intensity images. An example of a 3D slice plot of the reconstructed flow velocity for original velocity of 3.18 cm/s is shown in Fig. 56.

Triple-beam spectral-domain OCT

The three beam spectral-domain SD-OCT system (Fig. 57) consists of three independent subsystems. The SLD sources are operated at a central wavelength of 840 nm with a bandwidth of 50 nm and coupled into three miniature fiber collimators to reconstruct the 3D velocity vector. The three beams are split into reference and sample beam via a beam splitter. The notations in Fig. 57 are as follows: C – miniature fiber collimator, FC – fiber collimator, L – lens, G – grating, BS – beam splitter, M – mirror, PP – dispersion-compensating prism pairs, NF – neutral density filter, FPT – facet prism telescope, T – telescope, MEMS – two-axis gimbal-less scanning mirror, LS – linear stage, SLD – superluminescent light emitting diode, CAM – line scan camera.

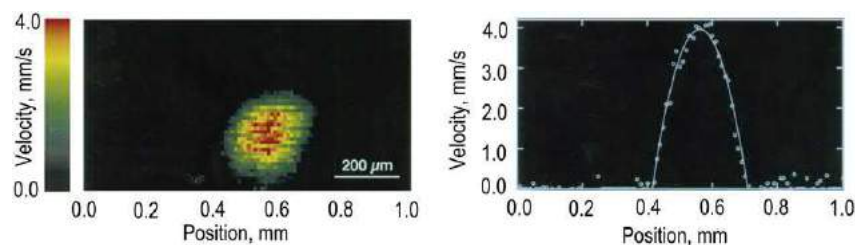


Fig. 54 Optical Doppler tomography images of blood flow in an in vivo biological model. Reproduced from Chen, Z., Milner, T.E., Srinivas, S., *et al.*, 1997. Noninvasive imaging of in vivo blood flow velocity using optical Doppler tomography. *Optics Letters* 22, 1119–1121.

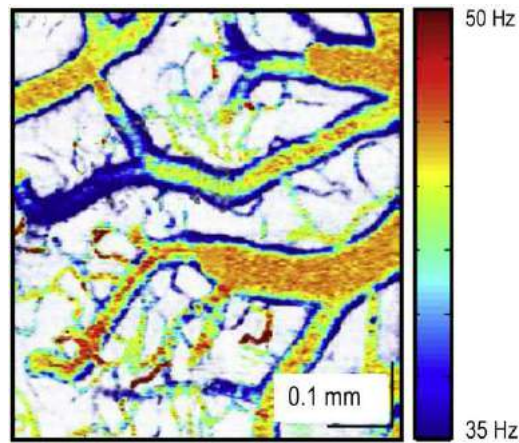


Fig. 55 Cortex velocity map during a hypercapnic challenge in a rat. Reproduced from Srinivasan, V.J., Radhakrishnan, H., Lo, E.H., *et al.*, 2012. OCT methods for capillary velocimetry, *Biomedical Optics Express* 3, 612–629.

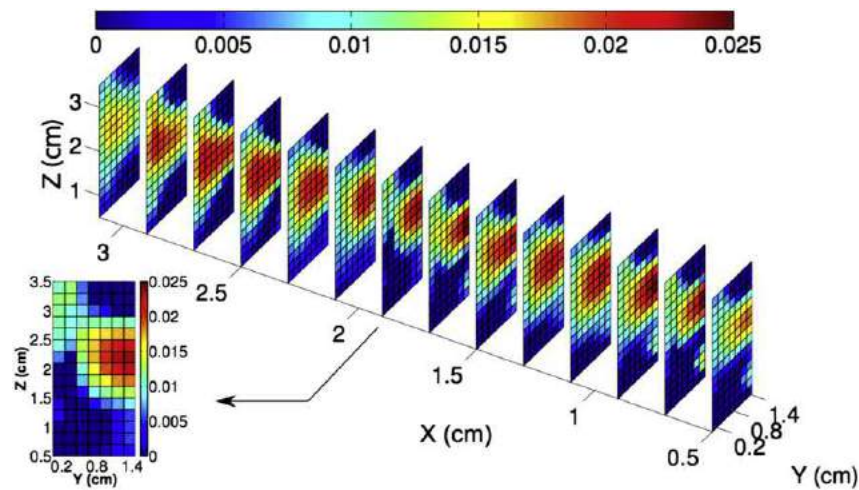


Fig. 56 Reconstructed 3D slice plot of blood flow in deep tissues. Reproduced from Varma, H.M., Valdes, C.P., Kristoffersen A.K., *et al.*, 2014. Speckle contrast optical tomography: A new method for deep tissue three-dimensional tomography of blood flow. *Biomedical Optics Express* 5, 1275–1289.

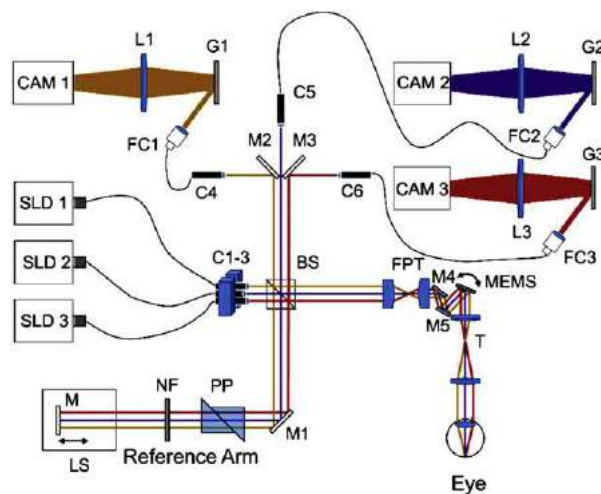


Fig. 57 Three beam spectral-domain optical low-coherence tomography (SD-OCT). Reproduced from Haindl, R., Trasischker, W., Wartak, A., *et al.*, 2015. Total retinal blood flow measurement by three beam Doppler optical coherence tomography. *Biomedical Optics Express* 7, 287–301.

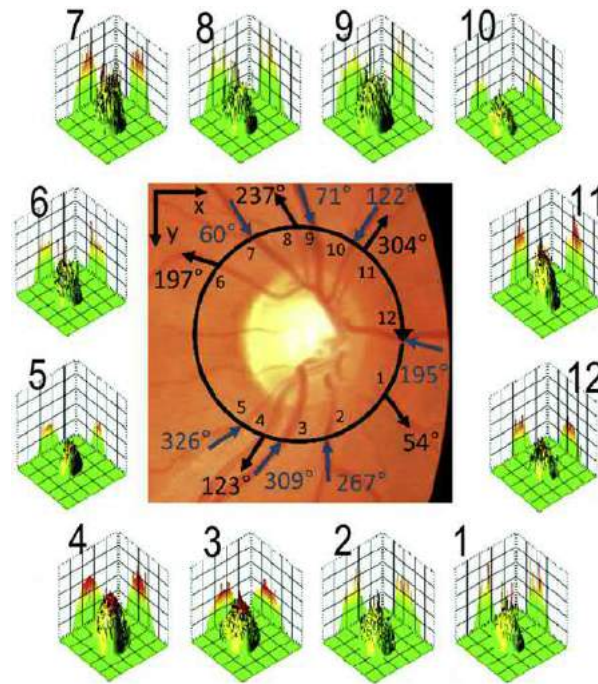


Fig. 58 Fundus photo with reconstructed velocity profiles. Reproduced from Haindl, R., Trasischker, W., Wartak, A., *et al.*, 2015. Total retinal blood flow measurement by three beam Doppler optical coherence tomography. *Biomedical Optics Express* 7, 287–301.

In the sample arm the two-axis gimbal-less MEMS mirror is used which is advantageous compared to a 2D galvo scanner, because both scanning axes are already in the same imaging plane. The MEMS mirror is driven with voltages ranging from 0 to 136 V. Any scan patterns can be realized (linear, raster, circular, and resonant patterns). Maximum optical scan angles of ± 11.5 degree can be achieved.

Fig. 58 displays the 3D velocity profile evaluation in the numbered vessels in the points on the circle. Arrows with numbers show the calculated vessel orientation clockwise relatively to the x axis. Reconstructed velocity profiles for the corresponding vessels around the optical nerve hypoplasia range from 0 to 40 mm/s.

Acoustical Studies and Vibrometry

Laser Doppler vibrometry

Doppler effect is used also to measure vibrations where they can exist. Laser Doppler vibrometry can provide advanced measurements in multiple physiological systems relevant to laboratory and field assessment of human stress and emotion (Tabatabai *et al.*, 2013). It can provide advanced recordings of myocardial and vascular performance, of respiratory efforts and sounds, and of tremor activity. It is responsive to laboratory stressors; muscle vibratory activity can be sensed from multiple muscles, including facial muscles, and the comparison is favorable with electromyography. Laser velocimetry can reliably assess facial muscle activity, associated with emotion and stress at low levels – below the threshold for visible facial deformations.

Wang *et al.* (2011) demonstrated a high-sensitivity pulsed laser vibrometer used to remotely determine the detailed, time-phased mechanical workings of various parts of the human heart. Results reported were validated by electrocardiography and accelerometer readings.

Successive pulse illumination in acoustical studies

In acoustical studies, Degroot (2009) used a method of successive pulse illumination of the investigated field with injected scattering particles (**Fig. 59**). She made videoregistration synchronously with pulse illumination, and measured the shifts of particles in successive frames in different phases of the acoustic process. An example of a 2D distribution of particle velocities is given in **Fig. 60**.

Vibration measurement of the vocal folds

An original technique was developed and used for clinical studies by George *et al.* (2008). To measure vibration dynamics of the vocal folds, they used a high-speed TV camera registration of a laser line projected onto the object in the triangulation mode and calculated the amplitude distribution for each frame (**Fig. 61**). The endoscopic laser projection system stretches a laser line beam in one direction and projects it as a thin laser sheet at an angle onto the surface of the vocal folds. The laser projection channel consists of a semiconductor diode laser and a cylindrical optical system.

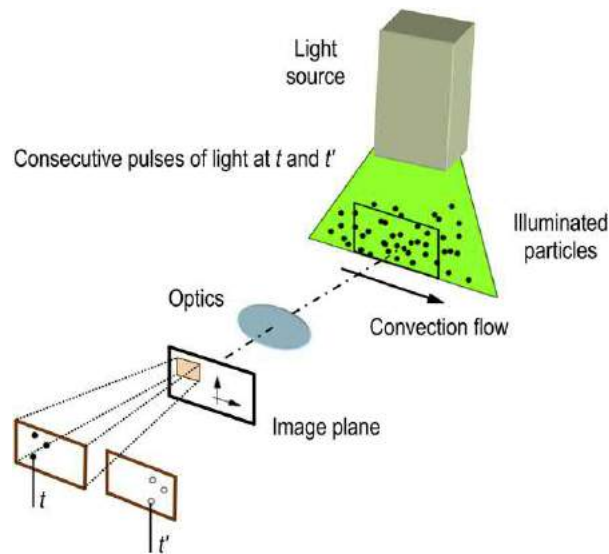


Fig. 59 Method of successive pulse illumination. Reproduced from Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.

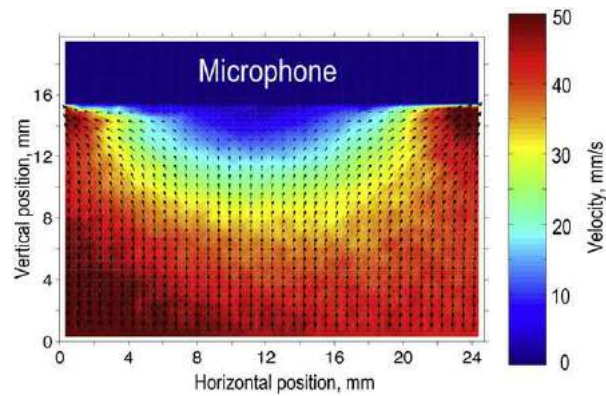


Fig. 60 2D distribution of particle velocities. Reproduced from Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.

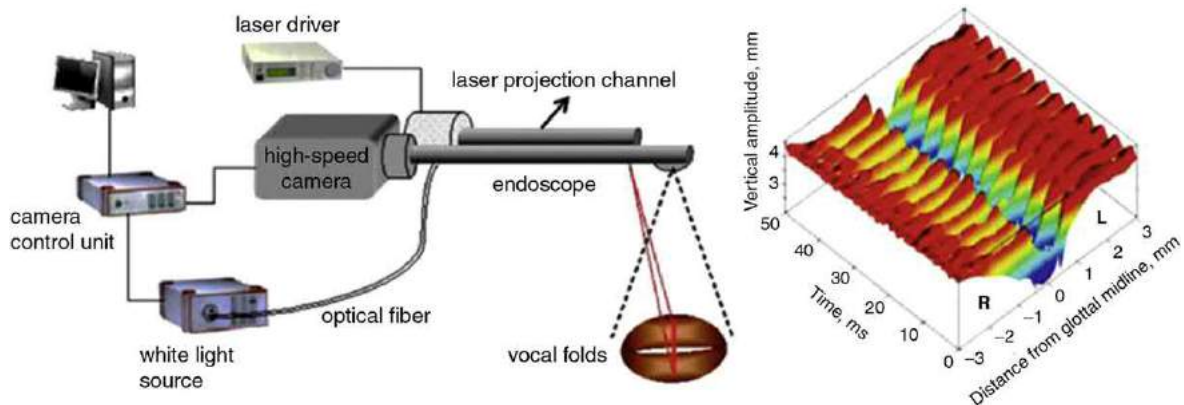


Fig. 61 Schematic view of the laser triangulation laryngoscope and color coded 3D vibration pattern of the vocal folds. Reproduced from George, N.A., de Mul, F.F.M., Qiu, Q., Rakhurst, G.R., Schutte, H.K., 2008. New laryngoscope for quantitative high-speed imaging of human vocal folds vibration in the horizontal and vertical direction. Journal of Biomedical Optics 13, 064024.

The laser projection channel is firmly attached to a 90 degree rigid endoscope, which acts as the receiving channel. The optical axes of the two channels are separated by a distance of 9 mm at the tip of the system. Within a normal working distance of 60–70 mm from the tip of the endoscope, the laser line is 18–20 mm long and 0.4 mm wide. A semiconductor laser emits at 653 nm, delivering an effective laser power density of 1.8 mW/mm², keeping below the exposure limit of 2 mW/mm². The red laser is used because it gives minimum absorption and satisfactory reflectance by the tissue in the visible spectral region.

A compact high-speed digital color camera is used for recording the images. It can record images continuously for a maximum of 2 s at the rate of 4000 fps with an image resolution of 256 × 256 pixels. Temporarily stored data in the camera memory is downloaded to the computer for further analysis. Vibration profiles in both horizontal and vertical directions were calibrated and measured with a resolution of ± 50 mm.

Wavefront Sensing Micro-Lidars

In laser weapon, to deliver the maximum density of energy to the target, atmospheric turbulence should be compensated by means of irradiating the target with a wavefront conjugated laser beam (He, 2002). The wavefront distortion is measured by a laser radar that receives the back reflected laser radiation with a matrix of photodetectors and provides the data for conjugation.

A similar need to measure the wavefront aberrations flared up in ophthalmology with the advent of vision correction using the ablation procedure after the successful wavefront corrections made on live human eye by Pallikaris *et al.* (1990), Seiler *et al.* (1990), McDonald *et al.* (1990), and others. Taboada and Archibald (1981) were the first engineers who paid attention to the potential of ultraviolet excimer laser to ablate the corneal epithelium. Trockel *et al.* (1983) used this effect for making the incisions correcting the ametropia.

Liang *et al.* (1994) used the matricial sensor, manufactured in Arizona by Platt, to measure the wave front aberrations in the human eye using the laser light reflected in outward direction from retina. The task to measure the inward, physiologically genuine, wavefront aberrations was discussed between Molebny and Pallikaris in 1995 (Molebny, 2013). The first clinical ray tracing aberrometer was tested on the live human eyes in the hospital of Crete University in 1998.

Hartmann–Shack Ocular Wavefront Sensing

The idea of light projection through “a perforated screen” was described by Hartmann (1904), developed further by Shack and Platt (1971) who replaced the perforated screen with “an array of contiguous lenticular elements.” A complete system designed and fabricated for astronomy was delivered to Cloudcroft, NM in the early 1970s but was never installed, and the facility was decommissioned (Platt and Shack, 2001).

Liang *et al.* (1994) with lenslet array from Platt, and Molebny *et al.* (1996a,b) with home-made holographic Fresnel micro-lenses built their wavefront sensors in which the laser light was projected on the retina, and back reflection captured by video cameras through the multielement optics.

A simplified schematic is shown in Fig. 62. A thin laser beam is directed into the eye, and being reflected from the retina, exits the eye. Usually, the microlenses are arranged in the form of matrix. Each microlens of the lenslet array projects its part of the beam cross-section on the CCD (CMOS, or other type) camera. Computer reconstructs the wavefront, and can calculate any of the derivative parameters of the optical system of the eye. The combination of a set (a matrix) of lenses (a lenslet array) and a 2D photosensitive detector are usually called Hartmann–Shack sensor. The image in the plane of the photosensitive detector is called hartmanngram. The synonyms of Hartmann–Shack sensor are Shack–Hartmann or Hartmann–Platt sensor.

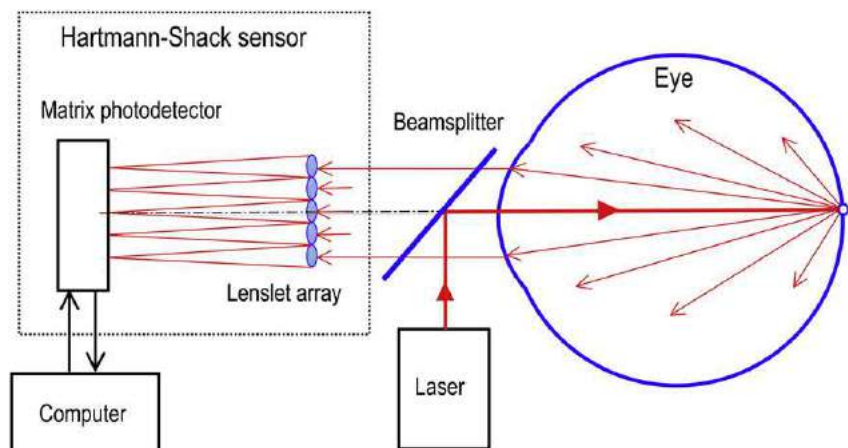


Fig. 62 Principle of measuring the wavefront errors in the eye with Hartmann–Shack sensor. Source: Molebny.

Commercial Hartmann–Shack aberrometers

Let us look at some examples of commercial aberrometers based on Hartmann–Shack wavefront sensing. One of the oldest companies on the market was VISX, later integrated with Abbott medical optics (AMO). iDesign is one of the latest instruments (see Section Relevant Websites). The maximal pupil diameter is 7 mm, the range of measured aberrations from -16 D to $+12$ D of sphere and astigmatism up to 8 D, with the dynamic range for higher order aberrations $8\text{ }\mu\text{m}$ root-mean square (RMS). The number of analyzed points is 1250. The previous WaveScan model captured 240 data points, and had the range of higher order aberrations $1.3\text{ }\mu\text{m}$ RMS. Besides wavefront measurement, the instrument combines functions of corneal topography, autorefractometry, pupillometry, and keratometry. The Hartmann–Shack approach is used in this instrument not only for the wavefront sensing but also to collect information on corneal topography.

Topcon having come later to the market, used a lot of experience from the earlier models of other manufacturers, especially for the interface. Like iDesign machine, Topcon's KR-1W has also five functions: wavefront measurement, corneal topography, autorefractometry, pupillometry, and keratometry (see Section Relevant Websites).

iProfiler (Fig. 63 left) from Carl Zeiss Vision GmbH also keeps the tendency of combining several functions in one instrument: ocular wavefront aberrometer, autorefractometer, corneal topographer, and keratometer (see Section Relevant Websites). Ocular wavefront is reconstructed up to seventh-order Zernike decomposition. The instrument is provided with touch screen. The instrument measures both eyes automatically in approximately 30 s. An example of displayed data with a map of an ocular wavefront errors is shown in Fig. 63 right. The map is color-coded (the scale is shown to the left of the wavefront map).

Wavefront Reconstruction

Presentation of measured wavefront data and the procedure of the wavefront error reconstruction for ophthalmologic applications were described by Southwell (1980), applied by Lane and Tallon (1992).

In general, aberrometers measure the gradient of the wavefront error function or, in other words, the deflection of rays from an un-aberrated direction. From each measured ray or location in the wavefront, the data contain four numbers. The first two are the horizontal (x) and vertical (y) coordinates of the location given. The second two numbers are the measured horizontal and vertical component values of the gradient. When reconstructing, the value of wavefront error should be calculated in each (x, y) point. 2D distribution (a map) of this error can be found as an approximation using the values of wavefront error in each point, or as an interpolation.

Analytically, the approximated surface of the wavefront error contains a set of 2D polynomials that is similar to a stack of multiple layer transparencies. This way of presentation is called a global one because each member of the polynomial is a component describing the whole map, not a part of it. Interpolation describing local specificities (in each zone of the map) is called a zonal presentation.

The approximation based on Zernike polynomial presentation is not the only possibility. Other types of polynomials can be used: Bhatia–Zernike, Fourier, Fourier–Mellin, Taylor, Bessel, Legendre, Didon, Karhunen–Loeve, etc. In most cases, preference is given to Zernike polynomials, because their components, describing the lower-order aberrations, coincide with the conventional ophthalmologic description. When the wavefront error is reconstructed in Zernike polynomials, it can be transformed into the presentation as a Fourier series, and vice versa (Dai, 2006). The same transformation can be done from Zernike to Taylor polynomials, and vice versa. The simplest way of interpolation is linear, just connecting two points. More sophisticated is connecting by a curve, satisfying certain requirements.

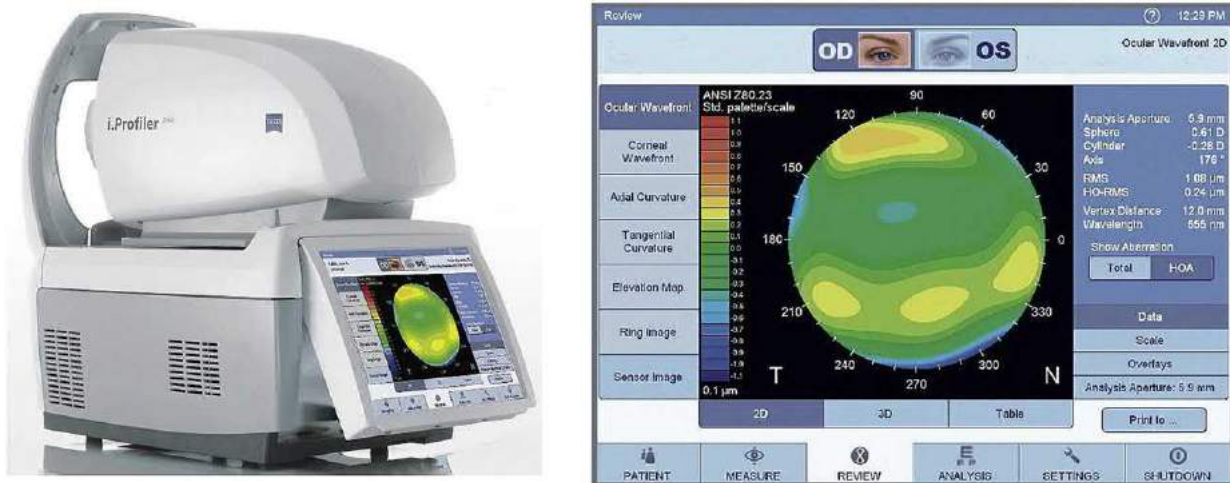


Fig. 63 iProfiler – a version of the aberrometer from Carl Zeiss Vision GmbH (left); a map of ocular wavefront errors (right). *Source:* Molebny.

Ray Tracing Aberrometry

The first ophthalmology-related publications baptizing the technique as ray tracing were made by Molebny *et al.* (1997) and by Navarro and Losada (1997). The ray tracing technique uses measurement of the position of a thin laser beam projected onto the retina. A beam of light is directed into the eye parallel to the visual axis having passed a two-directional (x, y) acousto-optic deflector and a collimating lens (Fig. 64).

The front focal point of the collimating lens is positioned in the center of scanning C of the acousto-optic deflector. Each entrance point, one at a time, provides its own projection on the retina. The position sensing detector measures the transverse displacement δx , δy of the laser spot on retina. An objective lens is used to optically conjugate the retina and the detector plane.

Commercial iTrace visual function analyzer

Fig. 65 shows a version of iTrace visual function analyzer, one of its functions is the wavefront sensing. A set of entrance points overlaid on the image of the eye is encircled by a green outer contour corresponding to the shape of the pupil, which is reconstructed from the eye image. In the process of beam displacement over the eye aperture, data on the transverse aberration for each point of the beam entrance into the eye are collected. They represent a 2D distribution known as a retinal spot diagram (Fig. 65 right bottom). The instrument uses a narrow (0.30 mm) diode laser beam with the wavelength 785 nm being displaced over the entrance pupil of an eye while kept parallel to the visual axis. To measure the positions of laser spots on retina, x and y linear arrays (each of 512 photodetectors) are used.

In the process of measurement, an image of the pupil is captured, its size and position are automatically measured in each TV frame. The permission on firing the probing radiation is given by the software only when the positions of the pupil and visual axis are within a predetermined correspondence. The data, captured with a position-sensing detector, are processed and transferred to

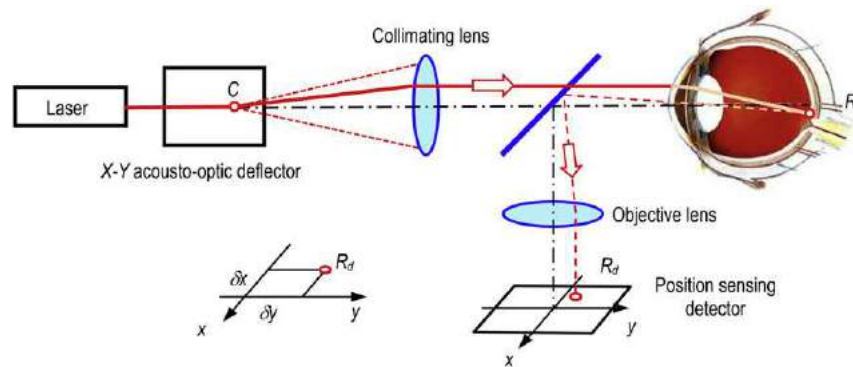


Fig. 64 Ocular ray tracing with a fast x - y acousto-optic deflector. Source: Molebny.

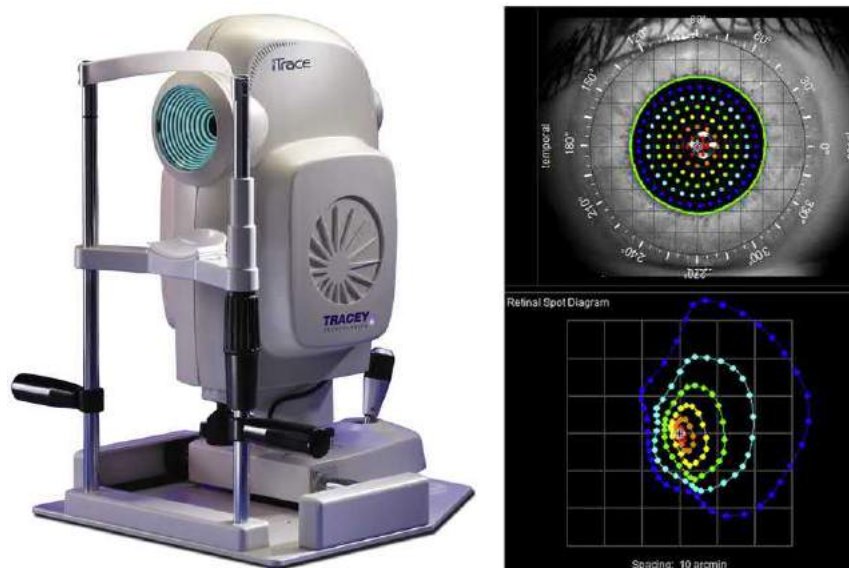


Fig. 65 iTrace visual function analyzer (left); laser beam positions in the entrance aperture of the eye (right top); collocation of laser beam projections on the retina – retinal spot diagram (right bottom). Source: Molebny.

the computer. The total time of scanning for the entire aperture of the eye is within 100 ms. The duration depends on the number of test points at the eye entrance pupil (from 64 to 256). Acousto-optic deflectors are very fast devices: transition time for switching a position is about 10 μ s, i.e., for 256 points, less than 3 ms is necessary. This is a great advantage of acousto-optics as compared to much slower scanners (no more than 20 points/s), used in the laboratory studies of Navarro and Moreno-Barriuso (1999).

Reconstructed maps and functions

Examples of reconstructed wavefront maps are given in Fig. 66, both in 2D and 3D versions. The acquired data are used also to calculate the derivative parameters of the optical system of the eye like point spread function (PSF) and modulation transfer function (MTF). Examples of the PSF and MTF are given in Fig. 67. The instrument allows both 2D and 3D display of both of them, as well as including in calculations any combination of Zernike modes. Any of these functions can be calculated for any size of the pupil.

Locating the visual axis

To optimally correct refractive errors, the eye must be properly positioned for measurements before and during surgery. Manzanera *et al.* (2015) showed that the optimal positioning is in favor of the visual axis. According to Reinstein (2009), corneal refractive ablations centered on the visual axis result in high quality of vision. Based on three-beam scanning micro-lidar, a technique was proposed to objectively locate the visual axis of the eye (Molebny, 2017) which is defined as the line connecting the point of fixation and the center of the foveola. The laser beam consists of three components (f_0 , $f_0 + F_x$ and $f_0 + F_y$), two of them ($f_0 + F_x$ and $f_0 + F_y$) being shifted along x and y axes relatively to the central one (f_0). The triple-beam is scanned along a closed trajectory (Fig. 68), the difference in depth between two pairs of beams Δh_x and Δh_y is determined in the process of scanning. The center of the foveola is defined as the crossing of steepest inclinations (Fig. 69).

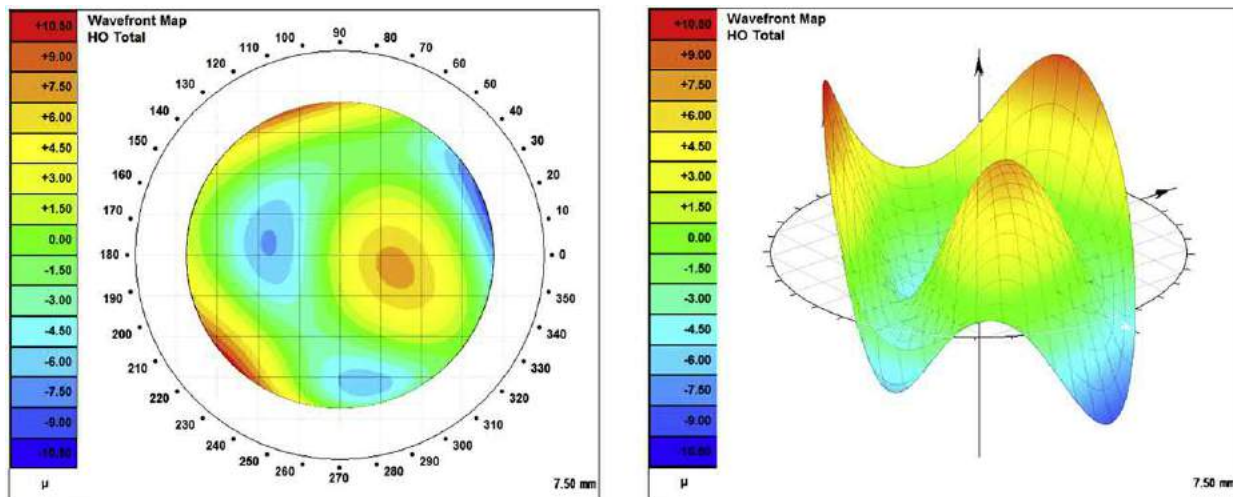


Fig. 66 Wavefront map displayed in Trace in 2D and 3D versions. Source: Molebny.

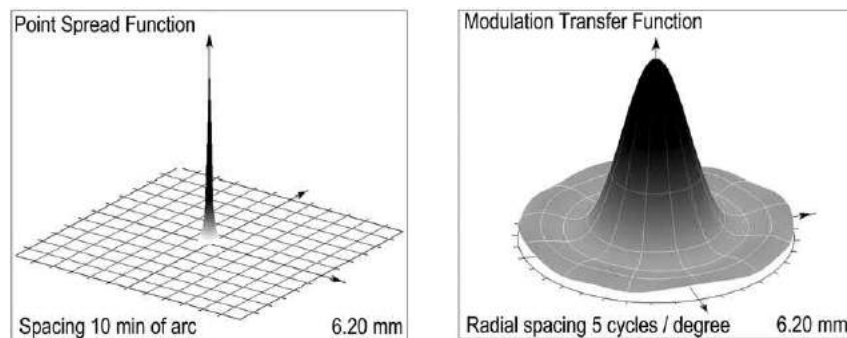


Fig. 67 Reconstructed 3D point spread function (PSF) (left) and 3D modulation transfer function (MTF) (right) of the human eye. Source: Molebny.

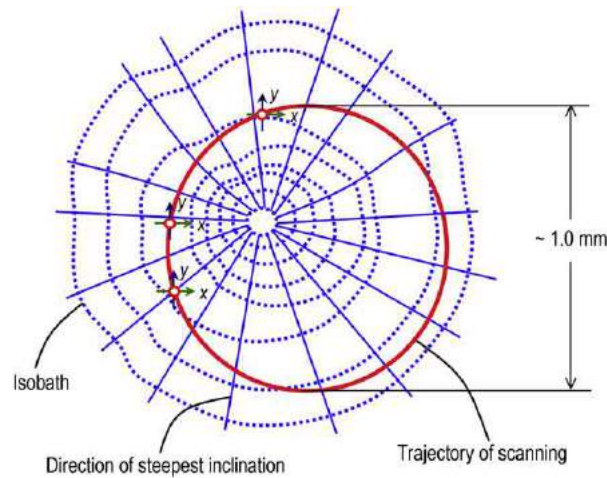


Fig. 68 Trajectory of scanning when determining the center of the foveola. *Source:* Molebny.

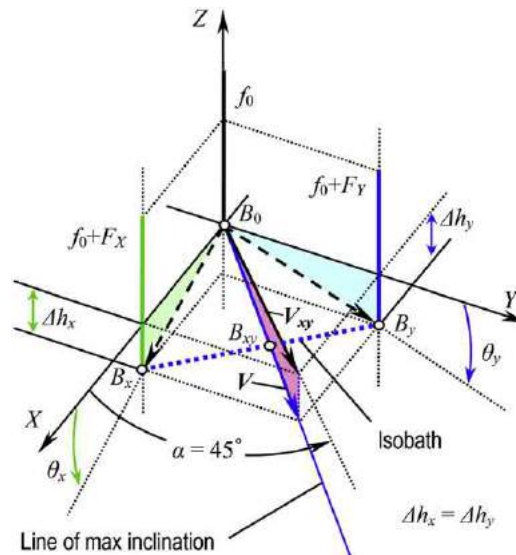


Fig. 69 Triple-beam configuration enabling the measurement of the depth difference. *Source:* Molebny.

Conclusion

When I started working on this chapter, it seemed to be a short walk across a small section of the big and mature field of laser radars, from which this younger sister section borrowed the principal ideas. The reality brought me to a shoreless ocean of new ideas and new opportunities. Technology solutions in this younger sister field appeared even more sophisticated: think about the clever CD/DVD head-up multiplied in millions of computers and other cutting-edge gadgets. Or recollect the flow cytometry taking the responsibility not only to quantitatively evaluate the number of any kind of cells in the human blood, but also to interrogate the signaling from the most dangerous tumor qualifying cells.

The intelligence of military rangefinders look like a children toy in comparison with the interactive play stations analyzing a huge amount of information and evaluating not only the distances to several players but also their positions and movements. Micro-lidars opened new perspectives in studying the human vision and measuring its parameters that enabled achieving the eagle acuity. Micro-lidars make feasible what could not be done with other means, for example, to objectively locate the visual axis of the eye that can't be anatomically even defined.

OCT whose birthday was just several years ago having achieved the highest level of both, information values and financial costs, is attacked today to have replaced the expensive distance-modulation parts with the cheap voice coils from the CD/DVD head-ups.

Sophisticated play stations are only the introduction to what is the dream of the mankind – to robotics, where the micro-lidars will play an important role of the robotic eyes communicating with the robot's brain.

Acknowledgment

The author thanks Dr. Paul McManamon for his idea to inspire the writing of this chapter.

See also: Very High Range Resolution Lidars

References

- Abramson, A., 1978. Light-in-flight recording by holography. *Optics Letters* 3, 121–123.
- Adler, D., 2012. Bench-to-bedside success: Intravascular optical coherence tomography. *SPIE Newsroom*. doi:10.1117/2.1201208.004382.
- Agishev, R.R., 2009. Lidar Monitoring of the Atmosphere (in Russian). Moscow: Fizmatgiz.
- Albrecht, H.-E., Borys, M., Damaschke, N., Tropea, C., 2003. *Laser Doppler and Phase Doppler Measurement Techniques*. Berlin: Springer.
- Alfano, R., Pu, Y., 2013. Optical biopsy for cancer detection. In: Jelínková, H. (Ed.), *Lasers for Medical Applications*. Cambridge: Woodhead, pp. 325–367.
- Alfano, R.R., Tang, G.C., Pradhan, A., *et al.*, 1991. Light sheds light on cancer. *Bulletin of the New York Academy of Medicine* 67, 143–150.
- Amann, M.C., Bosch, T., Lescure, M., *et al.*, 2001. Laser ranging: A critical review of usual techniques for distance measurement. *Optical Engineering* 40 (1), 10–19.
- Anderson, T.L., Covert, D.S., Marshall, S.F., *et al.*, 1996. Performance characteristics of a high-sensitivity, three-wavelength total scatter/backscatter nephelometer. *Journal of Atmospheric and Oceanic Technology* 13, 967–986.
- Barman, I., Dingari, N.C., Saha, A., *et al.*, 2013. Application of Raman spectroscopy to identify microcalcifications and underlying breast lesions at stereotactic core needle biopsy. *Cancer Research* 73 (11), 3206–3215.
- Baudelet, M. (Ed.), 2013. *Laser Spectroscopy for Sensing: Fundamentals, Techniques and Applications*. Amsterdam: Elsevier and Woodhead Publishing.
- Bourdét, G.L., Orszag, A.G., 1979. Absolute distance measurements by CO₂ laser multiwavelength interferometry. *Applied Optics* 18, 225–227.
- Chen, Z., Milner, T.E., Srinivas, S., *et al.*, 1997. Noninvasive imaging of in vivo blood flow velocity using optical Doppler tomography. *Optics Letters* 22, 1119–1121.
- Cohen, D., Rais, D., Sali, E., *et al.*, 2016. Error Compensation in Three-Dimensional Mapping. US Patent 9,330,324. May 3, 2016.
- Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), *Clinical Laboratory Hematology*, second ed. Essex: Pearson, pp. 956–974.
- Creath, K., 1988. Phase-measurements interferometry techniques. In: Wolf, E. (Ed.), *Progress in Optics*. New York, NY: Elsevier Science, pp. 349–393.
- Dai, G., 2006. Zernike aberration coefficients transformed to and from Fourier series coefficients for wavefront representation. *Optics Letters* 31, 501–503.
- Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.
- de Groot, P., 1991. Interferometric laser profilometer for rough surfaces. *Optics Letters* 16, 357–359.
- de Groot, P., Kishner, S., 1991. Synthetic wavelength stabilization of a two-color laser diode interferometer. *Applied Optics* 30, 4026–4033.
- Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. *Applied Optics* 27, 1113–1125.
- Delori, F.C., Weiter, J.J., Mainster, M.A., Flook, V.A., 1983. Oxygen saturation measurements in retinal vessels. *Investigative Ophthalmology & Visual Science* 24 (Suppl.), 13. ARVO.
- Deviaux, J.C., Abdelkader, H.H., Colle, E., 2013. A multi-sensor calibration toolbox for kinect: application to kinect and laser range finder fusion. In: 16th International Conference Advance Robotics (ICAR 2013), November 2013, Montevideo, Uruguay.
- Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.
- Dresel, T., Häusler, G., Venzke, H., 1992. Three-dimensional sensing of rough surfaces by coherence radar. *Applied Optics* 31, 919–925.
- Dsouza, R., Subhash, H., Neuhaus, K., *et al.*, 2014. Dermoscope guided multiple reference optical coherence tomography. *Biomedical Optics Express* 5, 2870–2882.
- Durst, F., Melling, A., Whitelaw, J.H., 1981. *Principles and Practice of Laser-Doppler Anemometry*, second ed. London: Academic Press.
- Dynamic 3D Imaging: Dyn3D Proceedings, DAGM 2009 Workshop. Kolb, A., Koch, R. (Eds.), Berlin: Springer.
- Fercher, A.F., Mengedocht, K., Werner, W., 1988. Eye-length measurement by interferometry with partially coherent light. *Optics Letters* 13, 186–188.
- Fornage, B.D., Sneige, N., Ross, M.I., *et al.*, 2004. Small (less 2 cm) breast cancer treated with ultrasonographically (US) guided radiofrequency ablation: Feasibility study. *Radiology* 231, 215–224.
- Freedman, B., Shpunt, A., Machline, M., Arieli, Y., 2012. Depth Mapping Using Projected Patterns. US Patent 8,150,142. April 3, 2012.
- Fujimoto, J.G., De Silvestri, S., Ippen, E.P., *et al.*, 1986. Femtosecond optical ranging in biological systems. *Optics Letters* 11, 150–152.
- García, J., Zalevsky, Z., García-Martínez, P., *et al.*, 2008. Three-dimensional mapping and range measurement by means of projected speckle patterns. *Applied Optics* 47, 3032–3040.
- George, N.A., de Mul, F.F.M., Qiu, Q., Rakhorst, G.R., Schutte, H.K., 2008. New laryngoscope for quantitative high-speed imaging of human vocal folds vibration in the horizontal and vertical direction. *Journal of Biomedical Optics* 13, 064024.
- Gerstner, K., Tschudi, T., 1994. New diode laser light source for absolute ranging two-wavelength interferometry. *Optical Engineering* 33 (8), 2692–2696.
- Guiagliumi, G., Akasaka, T., Sirbu, V., Kubo, T., 2016. Optical coherence tomography. In: Topol, E.G., Teirstein, P.S. (Eds.), *Textbook of Interventional Cardiology*, seventh ed. Philadelphia: Elsevier, pp. 990–1012.
- Haka, A.S., Shafer-Peltier, K.E., Fitzmaurice, M., *et al.*, 2005. Diagnosing breast cancer by using Raman spectroscopy. *Proceedings of the National Academy of Sciences* 102 (35), 12371–12376.
- Haka, A.S., Volynskaya, Z., Gardecki, J.A., *et al.*, 2006. In vivo margin assessment during partial mastectomy breast surgery using Raman spectroscopy. *Cancer Research* 66 (6), 3317–3322.
- Han, Y., Lu, D., Rao, R., Wang, Y., 2009. Determination of the complex refractive indices of aerosol from aerodynamic particle size spectrometer and integrating nephelometer measurements. *Applied Optics* 48, 4108–4117.
- Hansard, M., Lee, S., Choi, O., Horaud, R., 2012. Time of flight cameras: Principles, methods, and applications. *SpringerBriefs in Computer Science*, Springer.
- Hardarson, S.H., 2012. Retinal oximetry. PhD Thesis, University of Iceland, Reykjavik.
- Hartmann, J., 1904. Objektivuntersuchungen. *Z. Instrumentenke* 24, 1–21.
- He, G.S., 2002. Optical phase conjugation: Principles, techniques, and applications. *Progress in Quantum Electronic* 26, 131–191.
- Hee, M.R., Puliafito, C.A., Duker, J.S., *et al.*, 1998. Topography of diabetic macular edema with optical coherence tomography. *Ophthalmology* 105, 360–370.
- Heintzenberg, J., Charlson, R.J., 1996. Design and applications of the integrating nephelometer: A review. *Journal of Atmospheric and Oceanic Technology* 13, 987–1000.
- Hickham, J.B., Frayser, R., Ross, J.C., 1963. A study of retinal venous blood oxygen saturation in human subjects by photographic means. *Circulation* 27, 375–385.
- Hogan, J.N., Wilson, C.J., 2009. Multiple Reference Non-Invasive Analysis System. US Patent 7,526,329. April 29, 2009.
- Holm, S.P., 2014. Optical imaging of retinal blood flow: Studies in automatic vessel extraction, alignment, and driven changes in vessel oxymetry. PhD Thesis, University of Manchester.

- Huang, D., Swanson, E.A., Lin, C.P., *et al.*, 1991. Optical coherence tomography. *Science* 254, 1178–1181.
- Hu, P., Tan, J., Yang, H., *et al.*, 2011. Phase-shift laser range finder based on high speed and high precision phase-measuring techniques. In: 10th International Symposium on Measurement Technology and Intelligent Instruments. June 29–July 2, 2011, KAIST, Daejeon, Korea.
- Ishii, Y., Onodera, R., 1991. Two-wavelength laser-diode interferometry that uses phase-shifting techniques. *Optics Letters* 16, 1523–1525.
- Jelínková, H. (Ed.), 2013. *Lasers for Medical Applications*. Cambridge: Woodhead Publishing.
- Jerlov, N.G., 1976. *Marine Optics*. Amsterdam: Elsevier.
- Journet, B., Lourme, J.C., 2000. Laser range finding based on correlation method. In: 16th IMEKO World Congress, Vienna, Austria, September 25–28, 2000.
- Journet, B.A., Poujouly, S., 1998. High-resolution laser rangefinder based on a phase-shift measurement method. *Proceedings of SPIE* 3520, 123–132.
- Komachi, Y., Katagiri, T., Sato, H., Tashiro, H., 2009. Improvement and analysis of a micro Raman probe. *Applied Optics* 48, 1683–1696.
- Komachi, Y., Sato, H., Aizawa, K., Tashiro, H., 2005. Micro-optical fiber probe for use in an intravascular Raman endoscope. *Applied Optics* 44, 4722–4732.
- Kristjansdottir, J.V., 2014. Choroidal and retinal oximetry. MSc Thesis, University of Iceland. Reykjavik.
- Laing, R.A., Cohen, A.J., Friedman, E., 1975. Photographic measurements of retinal blood oxygen saturation: Falling saturation rabbit experiments. *Investigative Ophthalmology* 14 (8), 606–610.
- Lane, R.G., Tallon, M., 1992. Wave-front reconstruction using a Shack–Hartmann sensor. *Applied Optics* 31, 6902–6908.
- Lange, R., 2000. 3D Time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. PhD Thesis, University of Siegen, Germany.
- Lebedev, A.A., 2014. *Optics herald*. Rozhdestvensky Optical Society Bulletin 145, 11–15. St-Petersburg.
- Lexer, F., Hitzengerger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. *Applied Optics* 36, 6548–6553.
- Li, J., Ma, J., Wang, N., 2015. Application of retinal oximeter in ophthalmology. *Chinese Journal of Ophthalmology* 51 (11), 864–868.
- Liang, J., Grimm, B., Goelz, S., Bille, J.F., 1994. Objective measurement of wave aberrations of the human eye with the use of a Hartmann–Shack wave-front sensor. *Journal of the Optical Society of America* 11, 1949–1957.
- Liu, L., Gardecki, J.A., Nadkarni, S.K., *et al.*, 2012. Imaging the subcellular structure of human coronary atherosclerosis using 1- μ m resolution optical coherence tomography (μ OCT). *Nature Medicine* 17 (8), 1010–1014.
- Lonnqvist, J., 1999. Apparatus and Method for Measuring Visibility and Present Weather. US Patent 5,880,836. March 9, 1999.
- Ma, Z., Liu, A., Yin, X., *et al.*, 2010. Measurement of absolute blood flow velocity in outflow tract of HH18 chicken embryo based on 4D reconstruction using spectral domain optical coherence tomography. *Biomedical Optics Express* 1, 798–811.
- Malik, B.H., 2011. Dual wavelength polarimetry for glucose sensing in the anterior chamber of the eye. PhD Dissertation, Texas A&M University, College Station.
- Malik, B.H., Cote, G.L., 2010. Characterizing dual wavelength polarimetry through the eye for monitoring glucose corneal birefringence and ultimately enable the design and development of a polarization-based glucose monitoring system. *Biomedical Optics Express* 1, 1247–1258.
- Manzanera, S., Prieto, P.M., Benito, A., *et al.*, 2015. Location of achromatizing pupil position and first Purkinje reflection in a normal population. *Investigative Ophthalmology & Visual Science* 56, 962–966.
- Mazzarella, J., Cole, J., 2015. The anatomy of an OCT scan. *Review of Optometry*. <https://www.reviewofoptometry.com/article/the-anatomy-of-an-oct-scan>.
- McDonald, M.B., Kaufman, H.E., Frantz, J.M., *et al.*, 1990. Excimer laser ablation in a human eye. Case report. *Archives of Ophthalmology* 107 (5), 641–642.
- Mitchell, H.L., 2010. Nephelometer Instrument for Measuring Turbidity of Water. US Patent 7,663,751. February 16, 2010.
- Mizrahi, A., Russell, L., Three wavelength integrating nephelometer. Available at: https://www.esrl.noaa.gov/gmd/aero/instrumentation/neph_desc.html.
- Molebny, V., 2013. Wavefront measurement in ophthalmology. Aberrometry through the eyes of an engineer. In: Tuchin, V.V. (Ed.), *Handbook of Coherent-Domain Optical Methods*, vol. 2. New York, NY: Springer, pp. 315–361.
- Molebny, V., 2017. Method of locating the visual axis objectively. *Ophthalmology Visual Optics* 37 (3), doi:10.1111/opo.12376.
- Molebny, V.V., 1981. *Optical Radar Systems*. Moscow: Mashinostroenie, in Russian.
- Molebny, V.V., Kamerman, G.W., Smirnov, E.M., *et al.*, 1998. Three beam scanning laser radar microprofilometer. *Proceedings of SPIE* 3380, 280–283.
- Molebny, V.V., Kurashov, V.N., Pallikaris, I.G., Naoumidis, L.P., 1996a. Adaptive optics technique for measuring eye refraction distribution. *Proceedings of SPIE* 2930, 147–157.
- Molebny, V.V., Pallikaris, I.G., Naoumidis, L.P., *et al.*, 1996b. Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles. *Proceedings of SPIE* 2748, 68–75.
- Molebny, V.V., Pallikaris, I.G., Naoumidis, L.P., *et al.*, 1997. Retina ray-tracing technique for eye-refraction mapping. *Proceedings of SPIE* 2971, 175–183.
- Motz, J.T., Hunter, M., Galindo, L.H., *et al.*, 2004. Optical fiber probe for biomedical Raman spectroscopy. *Applied Optics* 43, 542–554.
- Moyle, J., 2002. *Pulse Oximetry*, second ed. London: BMJ Books.
- Mutto, C.D., Zanuttigh, P., Cortelazzo, G.M., 2012. *Time-of-flight cameras and Microsoft Kinect™*. New York, NY: Springer.
- Navarro, R., Losada, M.A., 1997. Aberrations and relative efficiency of light pencils in the living human eye. *Optometry and Vision Science* 74, 540–547.
- Navarro, R., Moreno-Barriuso, E., 1999. Laser ray-tracing method for optical testing. *Optics Letters* 24, 951–953.
- Nehmadi, Y., Guterman, Y., 2016. System and Method for Providing 3D Imaging. US Patent 9,303,989. April 5, 2016.
- Pallikaris, I., Papatsanaki, M., Stathi, E., *et al.*, 1990. Laser in situ keratomileusis. *Lasers in Surgery and Medicine* 10 (5), 463–468.
- Park, H., Chodorow, M., Kompfner, R., 1981. High resolution optical ranging system. *Applied Optics* 20, 2389–2394.
- Pittman, R.N., Duling, B.R., 1975. A new method for the measurement of percent oxyhemoglobin. *Journal of Applied Physiology* 38 (2), 315–320.
- Platt, B.C., Shack, R.S., 2001. History and principles of Shack–Hartmann wavefront sensing. *Journal of Refractive Surgery* 17, S573–S577.
- Pohlmann, K.C., 2011. *Principles of Digital Audio*, sixth ed. New York, NY: McGraw-Hill.
- Poujouly, S., Journet, B., 2000. Laser range finding by phase-shift measurement: Moving towards smart systems. *Proceedings of SPIE* 4189, 152–160.
- Reinstein, D., 2009. Ablations centred on visual axis may be more reliable. *Eurotimes* 14 (4), 21.
- Riva, C.E., 1979. Blood Flow Measurement. US Patent 4,142,796. March 6, 1979.
- Riva, C.E., 1983. Fundus Camera-Based Retinal Laser Doppler Velocimeter. US Patent. 4,402,601. September 6, 1983.
- Riva, C.E., Fekke, G.T., Eberli, B., Benary, V., 1979. Bidirectional LDV system for absolute measurement of blood speed in retinal vessels. *Applied Optics* 18, 2301–2306.
- Riva, C.E., Ross, B., Benedek, G.B., 1972. Laser Doppler measurements of blood flow in capillary tubes and retinal arteries. *Investigative Ophthalmology & Visual Science* 11, 936–944.
- Rovati, L., Minoni, U., Bonardi, M., F. Docchio, F., 1998. Absolute distance measurement using comb-spectrum interferometry. *Journal of Optics* 29 (3), J121–J127.
- Saleh, M.D., Eswaran, C., 2012. An efficient algorithm for retinal blood vessel segmentation using h-maxima transform and multilevel thresholding. *Computer Methods in Biomechanics and Biomedical Engineering* 15 (5), 517–525.
- Sali, E., Avraham, A., 2014. Three-Dimensional Mapping and Imaging. US Patent 8,717,417. May 6, 2014.
- Sali, E., Avraham, A., 2016. Three-Dimensional Mapping and Imaging. US Patent 9,350,973. May 24, 2016.
- Sasaki, O., Okazaki, H., 1986. Sinusoidal phase modulating interferometry for surface profile measurement. *Applied Optics* 25, 3137–3140.
- Schmetterer, L., Kiel, J.W. (Eds.), 2012. *Ocular Blood Flow*. Heidelberg: Springer.
- Schodl, R., 1976. Measuring Device for the Measurement of Fluid Flow Rates. US Patent 3,941,477.
- See, C.W., V. Iravani, V., Wickramasinghe, H.K., 1985. Scanning differential phase contrast optical microscope: Application to surface studies. *Applied Optics* 24, 2373–2379.
- Seiler, T., Kahle, G., Kriegerowski, M., 1990. Excimer laser (193 nm) myopic keratomileusis in sighted and blind human eyes. *Refractive and Corneal Surgery* 6 (3), 165–173.
- Sekine, A., Minegishi, I., Koizumi, H., 1993. Axial eye-length measurement by wavelength-shift interferometry. *Journal of the Optical Society of America A* 10, 1651–1655.

- Shack, R.V., Platt, B.C., 1971. Production and use of a lenticular Hartmann screen. *Journal of the Optical Society of America A* 61, 656.
- Shapiro, S.L. (Ed.), 1984. *Ultrashort Light Pulses. Picosecond Techniques and Applications*, second ed. Berlin: Springer.
- Shimotahira, H., Iizuka, K., Chu, S.C., *et al.*, 2001. Three-dimensional laser microvision. *Applied Optics* 40, 1784–1794.
- Southwell, W.H., 1980. Wave-front estimation from wave-front slope measurements. *Journal of the Optical Society of America* 70, 998–1006.
- Srinivasan, V.J., Radhakrishnan, H., Lo, E.H., *et al.*, 2012. OCT methods for capillary velocimetry. *Biomedical Optics Express* 3, 612–629.
- Staszewski, R.B., Balsara, P.T., 2005. Phase-domain all-digital phase-locked loop. *IEEE Transactions on Circuits and Systems II: Express Briefs* 52 (3), 159–163.
- Subhash, H.M., 2014. Smartphone based multiple reference optical coherence tomography (MRO™) system. *Biomedical Optics*, paper BT3A.72.
- Subhash, H.M., Hogan, J.N., Leahy, M.J., 2015. Multiple reference optical coherence tomography for smartphone applications. *SPIE Newsroom*. doi:10.1117/2.1201503.005807.
- Sugimoto, M., Nakamura, S., Inoue, Y., *et al.*, 2013. LT-PAM: A ranging method using dual frequency optical signals. *International Journal on Smart Sensing and Intelligent Systems* 6 (3), 791–809.
- Surrel, Y., 1993. Phase stepping: A new self-calibrating algorithm. *Applied Optics* 39, 3598–3600.
- Tabatabai, H., Oliver, D.E., Rohrbaugh, J.W., Papadopoulos, C., 2013. Novel applications of laser Doppler vibration measurements to medical imaging. *Sensing and Imaging: An International Journal* 14 (1), 13–28.
- Taboada, J., Archibald, C.J., 1981. An extreme sensitivity in the corneal epithelium to far UV ArF excimer laser pulses. *Proceedings of the Scientific Program of the Aerospace Medical Association*. May 4, 1981. San Antonio, TX.
- Trockel, S.L., Srinivasan, R., Braren, B., 1983. Excimer laser surgery of the cornea. *American Journal of Ophthalmology* 96 (6), 710–715.
- Tuchin, V.V. (Ed.), 2009. *Handbook of Optical Sensing of Glucose in Biological Fluids and Tissues*. Boca Raton, FL: CRC Press – Taylor & Francis Group.
- Varma, H.M., Valdes, C.P., Kristoffersen, A.K., *et al.*, 2014. Speckle contrast optical tomography: A new method for deep tissue three-dimensional tomography of blood flow. *Biomedical Optics Express* 5, 1275–1289.
- Wang, C.C., Trivedi, S., Kutcher, S., *et al.* (2011). Non-contact human cardiac activity monitoring using a high sensitivity pulsed laser vibrometer. In: *Proceedings of Conference on Lasers and Electro-Optics (CLEO)*, paper CWB6.
- Wang, H., Jun, X., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. *Optical Engineering* 49 (7), 073201.
- Yeh, Y., Cummins, H.Z., 1964. Localized fluid flow measurements with He–Ne laser spectrometer. *Applied Physics Letters* 4 (10), 176–178.
- Zhang, Zh., 2010. *LDA Application Methods. Laser Doppler Anemometry for Fluidic Dynamics*. Heidelberg: Springer.
- Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. *Journal of Biomedical Optics* 16 (1), 015004.

Relevant Websites

<https://www.vision.abbott/us/homepage.html>
Abbott.

https://www.asus.com/3D-Sensor/Xtion_PRO/
Asus Xtion Pro.

<http://mesa-imaging.ch>
HEPTAGON.

<http://sante.ro/wp-content/uploads/2016/04/Illumien-OPTIS-ENG-Brochure.pdf>
Illumien™ Optis™ PCI Optimization System.

http://www.icsightsound.com/pdf/i.scription_factsheet.pdf
iProfiler by Zeiss.

<http://www.topcon-medical.eu/eu/products/40-kr-1w-wavefront-analyzer.html>
TOPCON.



Encyclopedia of Modern Optics

Second Edition

Editors in Chief

Bob Guenther
Duncan Steel

**ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION**

ENCYCLOPEDIA OF MODERN OPTICS SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 5

Laser Radar ■ Optical Processing ■ Coherent Control
■ Biomedical: Human Eye ■ BioOptics ■ Microscopy
■ Quantum Optics ■ Organic Materials



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 5

Editorial Board	v
List of Contributors for Volume 5	ix
Editors in Chief	xi
Introduction	xiii

VOLUME 5

RF Coherent Detection on Top of Direct Detection Lidar	<i>Barry L. Stann, John F Dammann, Mark M Giza, Brian C Redman, and William C Ruff</i>	1
Optical Masks Using Walsh Functions	<i>Lakshminarayan Hazra and Pubali Mukherjee</i>	14
Coherent Control: Theory	<i>H Rabitz</i>	30
Coherent Control: Experimental	<i>RJ Levis</i>	39
Optics of the Human Eye	<i>David A Atchison</i>	43
Measuring Retinal Blood Flow	<i>Alberto de Castro and Stephen A Burns</i>	64
Adaptive Optics Retinal Imaging Techniques and Clinical Applications	<i>Jessica IW Morgan</i>	72
Femtosecond Lasers in Retinal Imaging	<i>Christina Schwarz and Jennifer J Hunter</i>	85
Confocal and Multiphoton Imaging of Cornea	<i>Gopal S Jayabalan and Josef F Bille</i>	97
Refractive Error and Wavefront Sensing	<i>Larry N Thibos</i>	108
Methods of Vision Correction	<i>Len Zheleznyak, Ramkumar Sabesan, and Geunyoung Yoon</i>	116
Laser and Light in Ophthalmology	<i>Michael Mrochen, Nicole Lemanski, and Bojan Pajic</i>	130
Probing Ocular Mechanics With Light	<i>Susana Marcos, Giuliano Scarcelli, Antoine Ramier, and Seok H Yun</i>	140
Optical Coherence Tomography and Its Application to Imaging of Skin and Retina	<i>Michael Pircher</i>	155
TIRF Microscopy and Variants	<i>Daniel Axelrod</i>	168
Super-Resolution Depth Measurements: Variable Angle TIRF, Super-Critical Angle Fluorescence, MIET	<i>Alexey Chizhik</i>	175
Overview: Microscopy	<i>CJR Sheppard</i>	185
Confocal Microscopy	<i>T Wilson</i>	194
Atom Optics	<i>AD Cronin and DE Pritchard</i>	203
Organic Semiconductors	<i>Fang-Chung Chen</i>	220
OLED/Introduction	<i>Dae-Gyu Moon</i>	232
Harvesting Triplet Excitons in OLED	<i>Gufeng He</i>	240
Organic Light-Emitting Diodes/Light Extraction	<i>Jiun-Haw Lee</i>	247
Conjugated Polymer-Based Solar Cells	<i>Gang Li, Widhya Budiawan, Pen-Cheng Wang, and Chih Wei Chu</i>	256

Dye-Sensitized Solar Cells	<i>Lu-Yin Lin and Kuo-Chuan Ho</i>	270
Organic Inorganic Hybrid Perovskite Materials and Devices	<i>Yihua Chen, Ning Zhou, and Huanping Zhou</i>	282
Light Harvesting in Organic Solar cells	<i>Supriya Pillai, Matthew Wright, and Kah H Chan</i>	292
Organic Lasers	<i>Graham A Turnbull</i>	309
Organic Photodetectors	<i>Jaehoon Jeong, Hwajeong Kim, and Youngkyoo Kim</i>	317
Index		331

LIST OF CONTRIBUTORS FOR VOLUME 5

David A Atchison
*Queensland University of Technology, Brisbane, QLD,
Australia*

Daniel Axelrod
University of Michigan, Ann Arbor, MI, United States

Josef F Bille
University of Heidelberg, Heidelberg, Germany

Widhya Budiawan
*Research Center for Applied Sciences, Academia Sinica,
Taipei, Taiwan (ROC); National Tsing-Hua University,
Hsinchu, Taiwan (ROC); and Academia Sinica and
National Tsing-Hua University, Hsinchu, Taiwan
(ROC)*

Stephen A Burns
*Indiana University School of Optometry, Bloomington,
IN, United States*

Kah H Chan
*University of New South Wales, Sydney, NSW,
Australia*

Fang-Chung Chen
National Chiao Tung University, Hsinchu, Taiwan

Yihua Chen
Peking University, Beijing, P.R. China

Alexey Chizhik
University of Göttingen, Göttingen, Germany

AD Cronin
*Massachusetts Institute of Technology, Cambridge, MA,
USA*

John F Dammann
Army Research Laboratory, Adelphi, MD, United States

Alberto de Castro
*Indiana University School of Optometry, Bloomington,
IN, United States*

Mark M Giza
Army Research Laboratory, Adelphi, MD, United States

Lakshminarayan Hazra
University of Calcutta, Kolkata, India

Gufeng He
Shanghai Jiao Tong University, Shanghai, China

Kuo-Chuan Ho
National Taiwan University, Taipei, Taiwan

Jennifer J Hunter
University of Rochester, Rochester, NY, United States

Gopal S Jayabalan
University of Heidelberg, Heidelberg, Germany

Jaehoon Jeong
*Kyungpook National University, Daegu, Republic of
Korea*

Hwajeong Kim
*Kyungpook National University, Daegu, Republic of
Korea*

Youngkyoo Kim
*Kyungpook National University, Daegu, Republic of
Korea*

Jiun-Haw Lee
National Taiwan University, Taipei City, Taiwan

Nicole Lemanski
Mabel Cheng MD PLLC, Albany, NY, United States

RJ Levis
Temple University, Philadelphia, PA, USA

Gang Li
*The Hong Kong Polytechnic University, Hong Kong,
China*

Lu-Yin Lin
*National Taipei University of Technology (Taipei Tech),
Taipei, Taiwan*

Susana Marcos
Instituto de Optica, CSIC, Madrid, Spain

Dae-Gyu Moon
*Soonchunhyang University, Asan-si, Chungcheongnam-
do, Republic of Korea*

Jessica IW Morgan
*University of Pennsylvania, Philadelphia, PA, United
States*

Michael Mrochen
*IROC Science AG, Zurich, Switzerland and Swiss Eye
Research Foundation, Reinach, Switzerland*

Pubali Mukherjee
MCKV Institute of Engineering, Howrah, India

Bojan Pajic
*Orasis Ag, Reinach, Switzerland and Swiss Eye Research
Foundation, Reinach, Switzerland*

Supriya Pillai
*University of New South Wales, Sydney, NSW,
Australia*

Michael Pircher
Medical University of Vienna, Vienna, Austria

DE Pritchard
*Massachusetts Institute of Technology, Cambridge,
MA, USA*

H Rabitz
Princeton University, Princeton, NJ, USA

Antoine Ramier
*Massachusetts General Hospital, Boston, MA, United
States*

Brian C Redman
Army Research Laboratory, Adelphi, MD, United States

William C Ruff
Army Research Laboratory, Adelphi, MD, United States

Ramkumar Sabesan
*University of Washington School of Medicine, Seattle,
WA, United States*

Giuliano Scarcelli
*University of Maryland, College Park, MD, United
States*

Christina Schwarz
University of Rochester, Rochester, NY, United States

CJR Sheppard
National University of Singapore, Singapore

Barry L Stann
Army Research Laboratory, Adelphi, MD, United States

Larry N Thibos
Indiana University, Bloomington, IN, United States

Graham A Turnbull
University of St Andrews, St Andrews, United Kingdom

Pen-Cheng Wang
*National Tsing-Hua University, Hsinchu, Taiwan
(ROC)*

Chih Wei Chu
*Research Center for Applied Sciences, Academia Sinica,
Taipei, Taiwan (ROC)*

T Wilson
University of Oxford, Oxford, UK

Matthew Wright
*University of New South Wales, Sydney, NSW,
Australia*

Geunyoung Yoon
University of Rochester, Rochester, NY, United States

Seok H Yun
*Massachusetts General Hospital, Boston, MA, United
States*

Len Zheleznyak
University of Rochester, Rochester, NY, United States

Huanping Zhou
Peking University, Beijing, P.R. China

Ning Zhou
Peking University, Beijing, P.R. China

EDITORS IN CHIEF



Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.



Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

RF Coherent Detection on Top of Direct Detection Lidar

Barry L Stann, John F Dammann, Mark M Giza, Brian C Redman, and William C Ruff, Army Research Laboratory, Adelphi, MD, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

This article presents research focused on technologies that will lead to low cost, small size, low weight, and low power laser radars (lidar) for applications such as surveillance, and autonomous vehicle navigation. The motivation is to exploit the very high angular and range resolution of lidar and the robust and predictable nature of lidar signatures, which provide true 3-D images that are critical for the aforementioned applications.

The lidar architecture treated here combines the merits of low-cost, rugged, reliable continuous wave (CW) semiconductor lasers and detectors with modern frequency modulation (FM) radar signal processing technology (Stann *et al.*, 1996). In contrast to previous FM/CW lidar concepts, in the architecture to be presented here the FM is applied to a radio-frequency (RF) subcarrier, not the optical carrier. Such an architecture is referred to in this article as incoherent FM/CW lidar. This architecture focuses on the low-altitude, relatively short-range (< 1 km) applications where the adverse effects of weather on propagation are minimal.

Presently there are several approaches to semiconductor laser lidar. A classic technique using CW semiconductor lasers or even light-emitting diodes (LEDs) involves measuring the range-dependent phase shift of a constant frequency RF subcarrier; two or more frequencies may be used to eliminate range ambiguities. An innovative embodiment of this technique was recently described by Anthes *et al.* (1993). Another technique investigated by a number of researchers is to modulate the CW source with a pseudorandom code and to cross-correlate the transmitted and returned waveform to determine range. Lee and Ramaswami (1992) recently described an implementation of this architecture with an added resolution-enhancement technique. The incoherent FM/CW lidar architecture discussed here has advantages over these techniques, especially for the low-cost applications. In the following text we discuss the important system aspects and advantages of the incoherent FM/CW lidar architecture and present a short mathematical analysis of the ranging theory. We then show some of the recent experimental results obtained with a breadboard version of the lidar coupled with a two-axis galvanometer-type beam scanner. This includes images of a truck obtained outdoors with a higher-power, improved version of the lidar. This follows previous work we had reported that included images taken in the laboratory with a preliminary version of the lidar. Following this we discuss an adaptation of the incoherent FM/CW ranging technique to a 32×32 pixel focal plane array (FPA) breadboard lidar. Images collected in the laboratory are shown and discussed (Stann *et al.*, 2004).

Incoherent FM/CW Lidar Architecture

Lidar Block Diagram

This section describes the lidar architecture and discusses some of the engineering issues associated with an actual design. Referring to the block diagram of the lidar (Fig. 1), a trigger circuit initiates the generation of a linear frequency-modulated (chirp) signal, that is, a signal whose frequency increases linearly as a function of time over period T . Other modulation waveforms are appropriate for this architecture; this one is chosen because the signal processing required to form range gates is readily understood, and the processor has low bandwidth, which makes it easy to implement. For the architecture discussed here, the start frequency of the chirp signal is typically in the tens to low hundreds of megahertz and the stop frequency in the hundreds of megahertz. The difference between the start and stop frequencies, ΔF , is chosen to establish the lidar resolution ΔR according to the usual equation from FM radar theory, $\Delta R = c/2\Delta F$, where c is the velocity of light (Stann *et al.*, 2004). Various commercial devices are available for generating the chirp signal. For short-range, high-resolution applications, commercial oscillators using monolithic microwave integrated circuits (MMICs) or hybrid technologies can be used. For long-range, high-resolution applications, precision linear tuning characteristics are required, which means that YIG-tuned oscillators or direct digital frequency synthesis devices are necessary. For applications where precision oscillators are not affordable, but tuning precision is still required, additional hardware may be built into the lidar that continuously measures the chirp-tuning characteristics. With this information, appropriate compensation techniques can be applied that effectively back out the tuning nonlinearities.

The chirp signal is fed into a wideband power amplifier that modulates the current driving a semiconductor laser diode. A circuit between the power amplifier and the laser diode matches the driving impedance of the amplifier to the laser diode impedance over ΔF . This causes the light beam intensity to remain highly modulated over the chirp bandwidth. Presently, commercial GaAlAs laser diodes are sold that have 2-GHz modulation bandwidths and output powers to 4 W. These specifications support designs with range resolutions less than 0.3 m and ranges to the low kilometers if the processing bandwidth is narrow.

The laser diode converts the chirp current waveform into a light waveform where the power is proportional to the driving current. This light is collected by a lens, collimated, and directed through a hole in a mirror toward a target. The transmitted light that is reflected by the target and propagated back to the lidar is reflected by the mirror and collected by a second lens that focuses

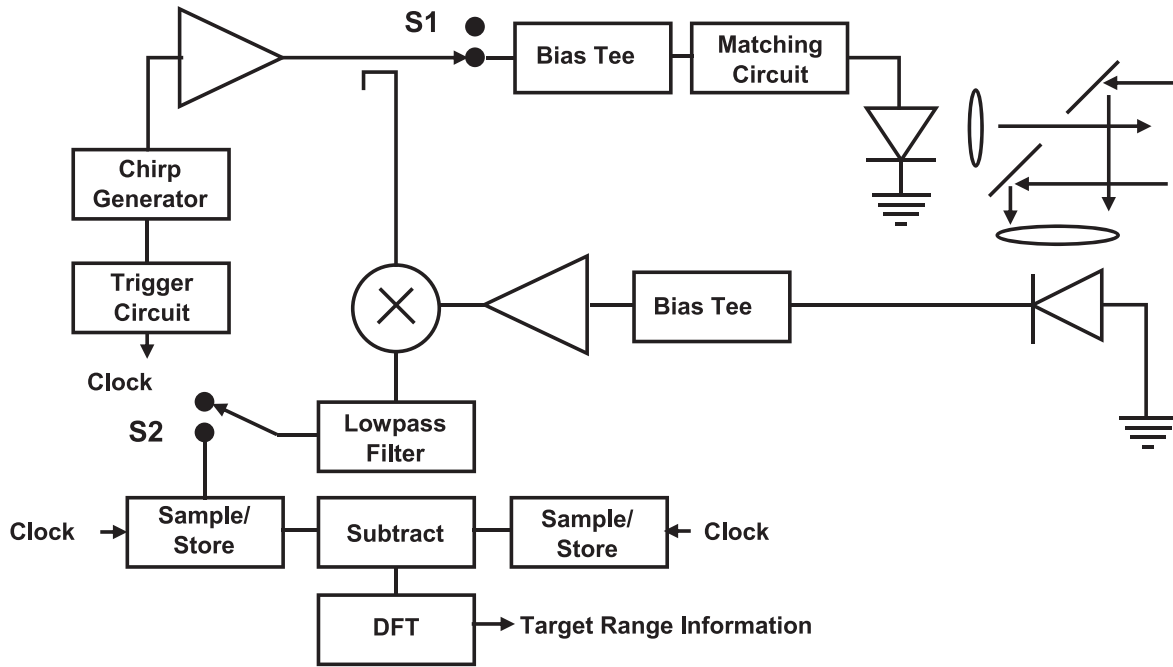


Fig. 1 Block diagram of the incoherent FM/CW lidar. Source: SPIE.

the light onto the active region of a photodiode. This optics design is used in our present breadboard; other optics designs are suitable with this lidar architecture. (Not shown in Fig. 1 is a two-axis beam scanner placed in the transmit/receive path of the light to facilitate image formation.)

The photodiode effectively converts the received light into a current proportional to the light power; thus it recovers a current waveform identical to the original modulating waveform, except for a reduction in amplitude and a time delay equal to the propagation time from the lidar to the target and return. Commercial photodiodes, readily available for this purpose, have bandwidths to 2 GHz. Current from the photodiode is converted into a voltage with a wideband amplifier. The output of the amplifier is mixed with an undelayed sample of the transmitted chirp waveform, and the output of the mixer is fed into a low-pass filter to recover an intermediate frequency (IF) signal with mean frequency f_{if} given by

$$f_{if} = \Delta F \frac{\tau}{T} \quad (1)$$

where τ is the round-trip time delay for the transmitted signal. For moving targets there is a Doppler shift that is made small relative to f_{if} by proper selection of T . Eq. (1) follows from standard FM radar ranging theory and can be found in any radar textbook. Using $\tau = 2D/c$, where D is the distance from the sensor to the target, and substituting this expression into Eq. (1) yields

$$f_{if} = 2\Delta F \frac{D}{cT} \quad (2)$$

which shows that the frequency of the IF waveform is directly proportional to the range to the target. It follows that the range to the target can be determined by measurement of the frequency of the IF waveform using the Fourier transform or some other technique.

Unfortunately, this straightforward method to obtain target range is not normally effective because the mixer creates a “self-clutter signal” at its IF port that may be larger than the target signal by several orders of magnitude. The self-clutter signal is usually caused by imperfections in the construction of the mixer that allow a detected portion of the local oscillator (LO) to appear at the IF port. Self-clutter signals can also be generated when a small fraction of the LO power leaks over to the receive port, travels toward the first microwave device in the receive path, and is reflected back toward the mixer. The self-clutter signal is an arbitrarily shaped function that is essentially unchanged over many chirp periods. This fact is employed to subtract the clutter signal from the IF signal.

To achieve this, switch S1 is opened to interrupt the transmission of light to the target, therefore leaving only the self-clutter signal present at the low-pass filter output. Switch S2 is closed, thereby directing the clutter signal into a circuit that samples and stores the waveform. For subsequent chirp periods, S1 is closed and S2 is open. The total IF signal is now sampled and stored and then fed into a stage that subtracts the self-clutter signal on a chirp-by-chirp basis. In the last stage, the discrete Fourier transform (DFT) of the subtracted data is performed to map the received time-domain IF signal into the frequency domain, or equivalently, the range domain to indicate distance to the target. The updating rate of the stored self-clutter signal is based on the rate of change of the self-clutter signal and the degree of cancellation required to meet lidar sensitivity requirements.

Advantages

The presented approach to FM/CW lidar contrasts with the traditional approaches where the optical carrier is frequency-modulated and the optical signal is coherently detected. These approaches require highly coherent lasers to detect targets at substantial ranges. The short coherence length of the CW semiconductor lasers used in such designs preclude target detection beyond a few tens of meters or less (Dieckmann, 1994). Even if the coherence length were significantly increased through the use of an external resonator or other techniques, problems resulting from the required precise alignment of return and LO signals, speckle, and large optical Doppler shifts would remain. For these reasons, the incoherent FM/CW lidar can be much less complex and costly, and yet more rugged and reliable.

The incoherent FM/CW ranging approach is a linear system, that is, any number of target and clutter returns will be separated in the range image if they are physically separated by more than ΔR . This is a very important feature if, for instance, the target is partially obscured by foliage or smoke, as one may often expect on battlefields. The phase measurement scheme of Anthes or the superresolution technique of Lee cannot resolve more than one target at a time in the lidar beam. A pulsed system can resolve multiple targets only if its pulse width is less than $2\Delta R/c$ and its receiver bandwidth is sufficient to fully resolve the return pulse, which degrades the Signal-to-noise ratio (SNR). Practical pulsed systems use much longer pulses and corresponding lower bandwidths, and resort to superresolution techniques such as leading-edge detection based on a threshold set at a constant fraction of the peak of the return (Stann *et al.*, 1995). This achieves a high range measurement precision, but requires a high SNR and still cannot resolve closely spaced multiple targets. Thus a target within a smoke cloud may not be discernible even if the lidar power is sufficient to penetrate the smoke. The incoherent FM/CW approach can resolve targets separated by ΔR without requiring a high SNR and with low signal processor bandwidths.

Another advantage of the incoherent FM/CW approach is inherent in its use of CW diode lasers. Diode lasers are limited in peak power by catastrophic damage mechanisms, and, like other semiconductor devices, provide more average power efficiently when operated CW. Most commercial applications of diode lasers, such as for compact disks, barcode scanners, and communications, are for CW devices. Therefore we may expect these devices to improve in performance and especially cost without any Department of Defense (DoD) investment. Finally, the incoherent FM/CW technique can provide Doppler information commensurate with the mean frequency of the AM that gives target velocity and may aid in target detection in a clutter environment.

Mathematical Analysis

This section presents an idealized mathematical analysis to give the reader insights into the formation of range gates by the incoherent FM/CW lidar and the relationship of system parameters to range resolution. This analysis is also instructive in that it shows that the theory is valid when ΔF is as large as or larger than f_0 as in our case, but unlike the case in radar, where ΔF is much smaller than f_0 . The discussion here treats only the core theory of the lidar and does not include any analysis of the many other signal processing techniques used in FM radar that can also be exploited to handle problems unique to any real lidar application. Some of the common techniques include the use of alternative modulation waveforms (triangle waves, pseudorandom codes, etc.) to simplify the formation of range cells or inhibit the detection of jamming signals or other operating sensors, the application of weighting on the Intermediate-frequency (IF) waveform to reduce range sidelobes, IF amplifier limiting to effectively normalize the range response to any target return amplitude, the addition of gates to inhibit the detection of clutter and aerosols, Doppler detection to determine target velocity, and methods to determine whether a target is approaching or receding.

FM ranging theory

An analysis of FM ranging theory begins with an expression for the frequency of the transmitted chirp waveform

$$f(t) = f_0 + \frac{\Delta F t}{T}, \quad -\frac{T}{2} \leq t \leq \frac{T}{2} \quad (3)$$

where f_0 is the center or carrier frequency of the chirp waveform and t is time. The phase of the transmitted chirp waveform is

$$\phi_x(t) = \int_{-\infty}^t 2\pi \left(f_0 + \frac{\Delta F}{T} t' \right) dt' \quad (4)$$

While the phase of the received signal is

$$\phi_r(t) = \int_{-\infty}^{t-\tau} 2\pi \left(f_0 + \frac{\Delta F}{T} t' \right) dt' \quad (5)$$

The phase difference between the transmitted and the received is

$$\Delta\phi(t) = \phi_x(t) - \phi_r(t) \quad (6)$$

or

$$\Delta\phi(t) = \int_{t-\tau}^t 2\pi \left(f_0 + \frac{\Delta F}{T} t' \right) dt' \quad (7)$$

Performing the integration and inserting the limits yields

$$\Delta\phi(t) = 2\pi\left(f_0 t + \frac{\Delta F}{2T} t^2\right) - 2\pi\left(f_0(t - \tau) + \frac{\Delta F}{2T}(t^2 - 2t\tau + \tau^2)\right) \quad (8)$$

Canceling the terms and dropping the τ^2 term because τ is set much less than T leaves

$$\Delta\phi(t) = 2\pi f_0 \tau + 2\pi \frac{\Delta F}{T} \tau t \quad (9)$$

The mixer and following filtering process yields the IF waveform,

$$V_m(t) = a_r \cos(\Delta\phi(t)) \quad (10)$$

or

$$V_m(t) = a_r \cos\left(2\pi\left(f_0 \tau + \frac{\Delta F}{T} \tau t\right)\right) \quad (11)$$

where a_r represents the amplitude of the received signal after considering all propagation, scattering, detection, and mixing losses and amplifier gains.

An examination of Eq. (11) shows that the IF waveform is a cosine function whose argument contains two terms. The first is recognized as a fixed Doppler phase term that is proportional to the carrier frequency and time delay or distance to the target. The second term is a function of time and contains the frequency of $V_m(t)$, $\left(\frac{\Delta F}{T}\right)\tau$, which is proportional to τ or the distance to the target. A plot of this function (Fig. 2) as a function of the normalized range $\tau\Delta F$ and normalized t illustrates this behavior. Here the dark regions represent the minimum of $V_m(t)$, and the bright regions represent the maximums. The important feature to notice is that the number of periods observed over normalized time or one sweep of the modulation increases proportionally with the normalized range.

From this, it follows that the target distance can be determined simply by measurement of the frequency of $V_m(t)$ over T . The usual method for determining the frequency content of a waveform requires computing the Fourier transform of the waveform

$$F(f) = a_r/T \int_{-\frac{T}{2}}^{\frac{T}{2}} \cos\left(2\pi\left(f_0 \tau + \frac{\Delta F}{T} \tau t\right)\right) e^{-2\pi f t} dt \quad (12)$$

Performing the integration yields

$$F(f) = e^{j2\pi f_0 \tau} \left(\frac{\sin\left(\pi\left(\Delta F \frac{\tau}{T} - f\right)T\right)}{\pi\left(\Delta F \frac{\tau}{T} - f\right)T} \right) + e^{-j2\pi f_0 \tau} \left(\frac{\sin\left(\pi\left(\Delta F \frac{\tau}{T} + f\right)T\right)}{\pi\left(\Delta F \frac{\tau}{T} + f\right)T} \right) \quad (13)$$

where the factor a_r which modifies the entire equation, has been dropped to effectively normalize the amplitude of the lidar response to a target. For a lidar architecture employing the *DFT* to establish range cells, each complex output sample of the transform represents the magnitude and phase of the respective harmonic frequency components of the input waveform that occur

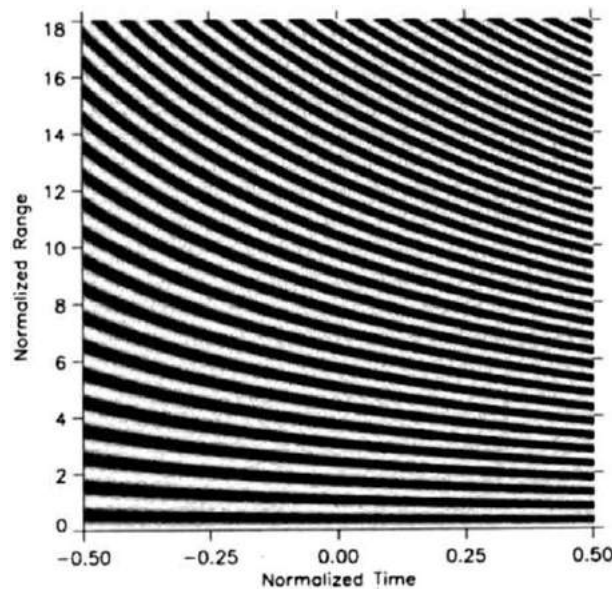


Fig. 2 Intermediate frequency (IF) output waveform as a function of time and normalized range. Source: SPIE.

at integer multiples of $1/T$. This allows f to be replaced in Eq. (13) with n/T , yielding

$$F(n/T) = e^{j2\pi f_0 \tau} \left(\frac{\sin(\pi(\Delta F \tau - n))}{\pi(\Delta F \tau - n)} \right) + e^{-j2\pi f_0 \tau} \left(\frac{\sin(\pi(\Delta F \tau + n))}{\pi(\Delta F \tau + n)} \right) \quad (14)$$

The exponential factors in Eq. (14) represent the Doppler component present in each harmonic line, while the $\sin(x)/x$ factors represent the amplitude of the Doppler components. For n greater than three or four, the second term can be ignored because it is much smaller than the first. Examining the first term shows that each harmonic component of $F(\frac{n}{T})$ has a maximum for

$$\Delta F \tau = n \quad (15)$$

which means that the output samples of the discrete Fourier transform have been mapped to normalized range according to Eq. (15). Equivalently, each output sample corresponds to a range gate.

Fig. 3 shows how the amplitude of individual harmonic lines or range gates of Eq. (14) behave as the normalized target range moves from 0 to 18. These plots are called range responses. For instance, the top plot shows the magnitude of the first range gate (harmonic number 0) as a function of target range. As expected, the magnitude of this range gate is maximum at zero range and decreases with increasing range. The rapid modulation in the basic $\sin(x)/x$ shape of the function is caused by the second term in Eq. (14), which is still strong at small n .

The notion of range resolution can now be established from one of the higher-harmonic-number range responses. From these responses, the normalized distance between the 0.64 levels of the mainlobe of the range response is equal to 1, which means that

$$\Delta F \Delta \tau = 1 \quad (16)$$

Using the substitution, $\Delta R = c\Delta \tau/2$, one finds that the range resolution equals

$$\Delta R = \frac{c}{2\Delta F} \quad (17)$$

In the design of an FM radar, ΔF is a small percentage of f_0 . Under these conditions, the previous closed-form theory is known to be valid. For a high-resolution design of this lidar, ΔF is a high percentage of f_0 , raising the question of whether the theory still holds. To investigate this issue, the lidar was simulated with a brute-force computer code that more closely modeled the processes in an actual system. Comparisons of the results obtained from this model and the closed-form theory showed no significant

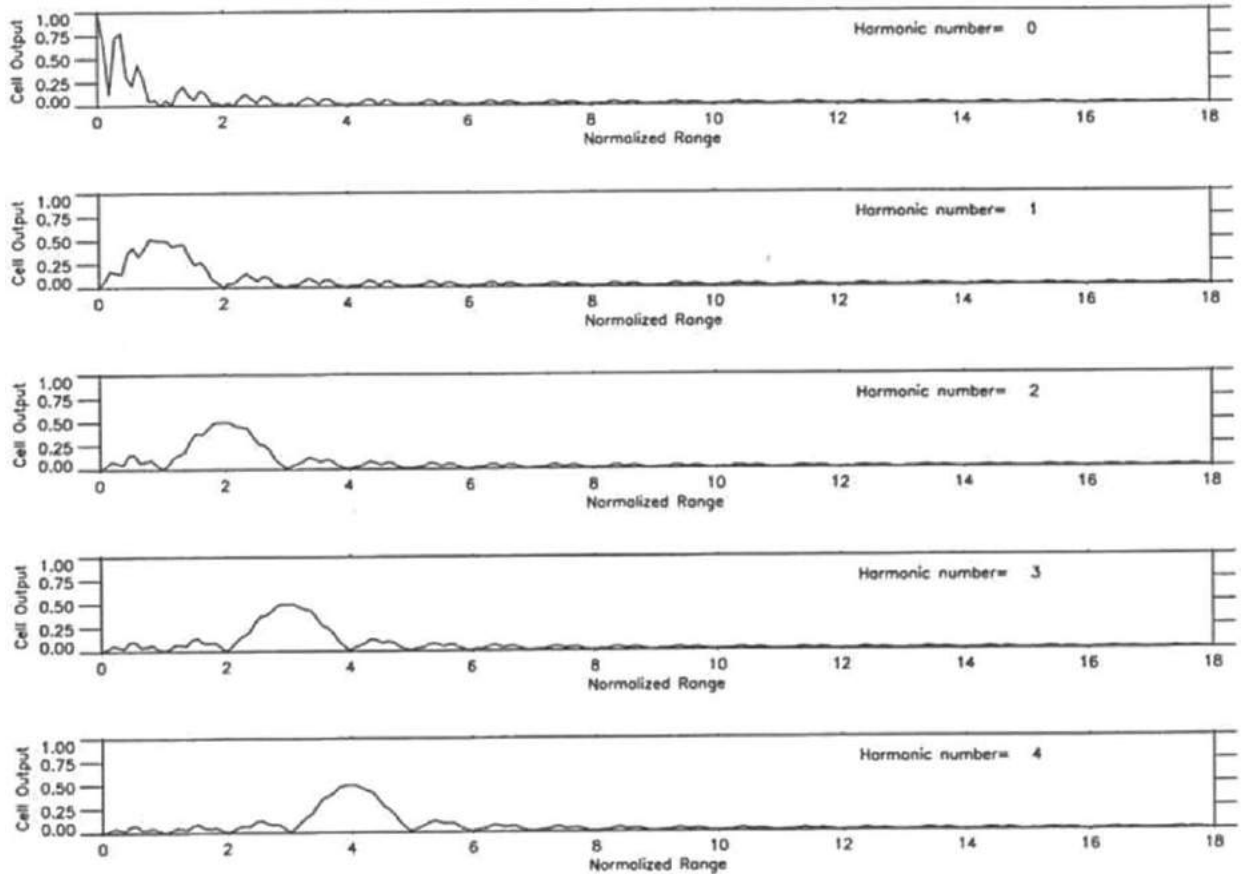


Fig. 3 Range-gate output as a function of normalized target range. Source: SPIE.

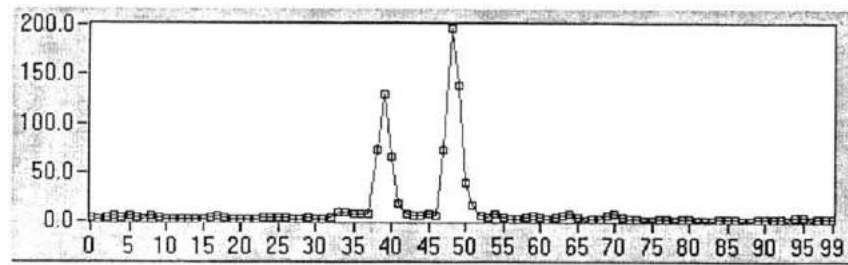


Fig. 4 Discrete Fourier transform (DFT) of intermediate frequency (IF) signal—two targets. Source: SPIE.

differences. Additional verification of the validity of the closed-form theory is evident in the experimental results of the following sections.

Ladar Hardware

A breadboard of the incoherent FM/CW was built to validate the theory and collect high-resolution imagery of targets. The optics and microwave elements of the imaging ladar are organized as shown in the block diagram in [Fig. 1](#), except for the placement of a two-axis galvanometer-type beam scanner in the transmit/receive path. Presently, a 100-mW laser diode at 817 nm is employed as the transmitter, and an avalanche photodiode is used for the receiver. The usable bandwidth of the laser and photodiode combination is about 1.3 GHz, and the AM index is approximately 86%. A personal computer (PC) was configured to set the beam pointing direction, trigger chirps, sample the IF data, cancel the self-clutter, form the range gates, and display the data. This system can collect 80×80 pixel image files and display the collected and processed data in several formats.

Point-target measurements

Point-target measurements recovered from tests with the incoherent FM/CW ladar illustrate the theory and some of its performance characteristics. To accomplish this we recorded the IF signal for targets at an approximately 9-m range for a chirp that starts at 600 MHz and ends at 1200 MHz. The DFT is used to map the IF signal into range gates. [Fig. 4](#) shows when the transform result is used on an IF signal comprised of returns from targets at two different ranges. As discussed earlier, the ladar is a linear system; thus multiple targets can be ranged simultaneously as evidenced by the two impulses present at range-gate 39 and 48. For this test $\Delta F = 600$ MHz thus the distance between range-gate is 0.25 m. This corresponds to 9.75 m for the first target and 12 m for the second target. The first target was deliberately set at 9 m which should correspond to range-gate 36. This error is caused by additional delay in the microwave lines internal to the ladar. When a target is in the middle of a range gate and a Hanning window is used, DFT samples on either side of the maximum should theoretically be 50% of the maximum value, and all others should be substantially suppressed. This is roughly true in the figure. Discrepancies from the theoretical distribution are usually caused by nonlinear FM characteristics in the chirp signal. Other causes may include amplitude modulation of the IF signal due to nonflat amplitude-versus-frequency characteristics of the microwave components, laser diode, and/or detector diode. Nonlinear phase-versus-frequency characteristics of these same devices can also cause spreading of the signal distribution. Additional hardware can be added to the system to measure component errors, and appropriate compensation can be applied to the IF signal to minimize spreading effects. Little effort has been expended so far in reducing these errors in the breadboard ladar.

Imaging measurements

The imaging capabilities of the ladar were tested at an outdoor site on the ARL Adelphi, MD, campus. [Fig. 5](#) shows the site with an Army pickup truck oriented perpendicular to the ladar at about a 52-m range. All specularly reflective surfaces on the truck were masked with black tape for laser safety reasons. A calibrated standard reflectivity target (18%) is positioned in the lower right of the photograph, and a grassy hill is located about 15 m beyond the truck.

An 80×80 pixel image of the truck was collected with the imaging ladar for aspect angles of approximately 0, 45, and 90 degrees. For brevity, [Fig. 6](#) shows only the 0 degree result. The uppermost image is formed pixel by pixel by summing outputs from all of the range cells from the front of the truck to beyond the hill. The bottom image sum outputs from the range cells from the front through the rear bumper of the truck only, thereby eliminating the returns from the hill and enhancing the contrast of the truck from the background. The pixel density is sufficient to allow one to determine that the target is indeed a pickup truck. The calibration target appears in each image in the lower right.

32 × 32 Pixel FPA Ladar Breadboard Using Incoherent Chirped Amplitude Modulation

Introduction

This section presents a FPA ladar architecture that is useful for a variety of applications such as reconnaissance and robotic navigation and is based on the optically incoherent FM/CW ranging architecture reported in the previous section. The incoherent FM/CW ranging architecture is a natural fit to a FPA design because the final system bandwidth is low enough to provide sufficient



Fig. 5 Ladar image site. Source: SPIE.

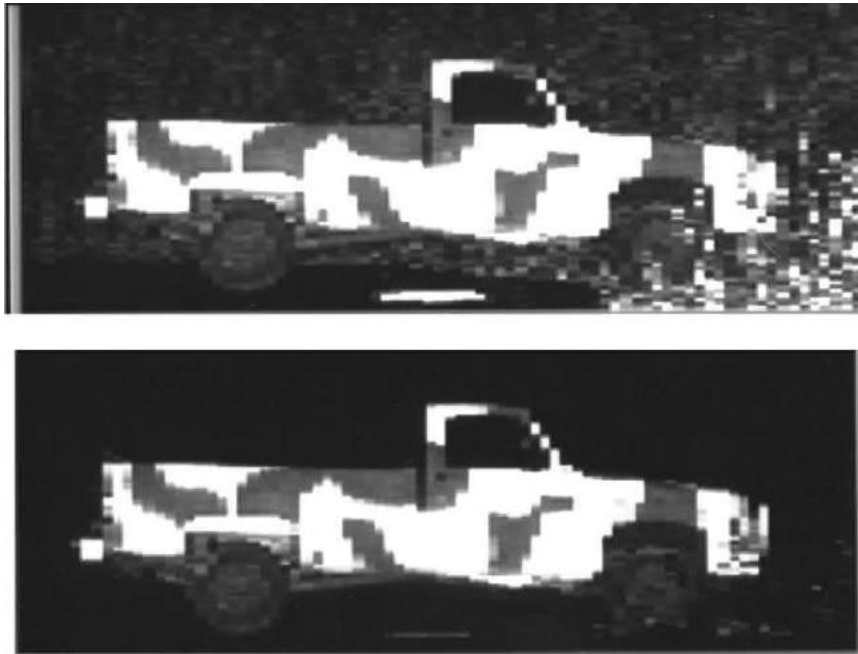


Fig. 6 Ladar images of a pickup truck (All range gates, top, partial). Source: SPIE.

signal-to-noise while using low-cost and low power laser diodes. In the following text we describe the FPA architecture, optoelectronic (OE) detector design, read-out based on code division multiple access (CDMA) techniques, and signal processing. We then summarize work to build a brassboard and related hardware modifications to reduce pixel-to-pixel cross-talk, and the addition of a 32-channel data acquisition system (DAS) and new signal processing and display software. Imagery collected in the lab is displayed along with a discussion of the improvements to image quality achieved with the reduction of pixel-to-pixel cross-talk.

Incoherent FM/CW FPA Breadboard Architecture

In the following text, we briefly describe the operating principles of the 32×32 pixel FPA lidar breadboard using the block diagram shown in Fig. 7. A clock circuit controls the generation of a chirp signal that serves as the laser intensity modulation and RF LO signal. As in the previous section the chirp signal is simply a sinusoidal waveform whose frequency linearly increases over a period T and then flies back to the original start frequency (i.e., sawtooth FM). The chirp signal for our breadboard has a start frequency of 200 MHz and a stop frequency of 800 MHz. FM/CW radar ranging theory supports a rich variety of improvisations to the basic sawtooth modulation format. However, to simplify our present work with the lidar breadboard, we use only the pure sawtooth modulation format, which yields high range resolution with the least amount of lidar signal processing.

To modulate the laser illumination, the chirp signal is fed into a wideband RF power amplifier with a low output driving impedance. Output from the amplifier and a dc current are summed in a bias tee to provide a modulated current drive for a

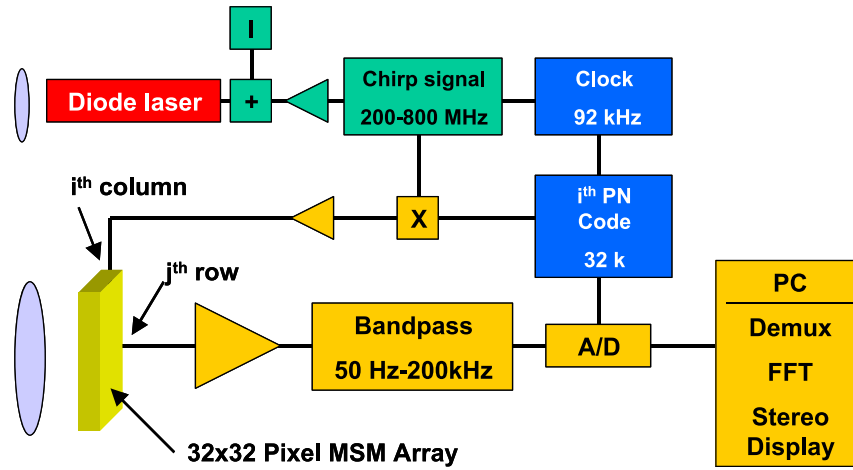


Fig. 7 Laboratory lidar breadboard block diagram. Source: SPIE.

broad-area semiconductor diode laser at 808 nm with a bandwidth at least equal to the bandwidth of the chirp waveform. A high-percentage intensity modulation of the light beam is desirable. The divergent laser beam from the semiconductor laser is collected with a collimator and focused to project a beam sufficiently wide to encompass or floodlight the target scene of interest.

A small portion of the transmitted beam is reflected from the target toward the lidar and collected by the receiver optics. An array of OE-mixing metal-semiconductor-metal (MSM) detectors is located at the focal plane of the receiver optics (Ruff *et al.*, 2000; Shen *et al.*, 2000). When the transmitter RF modulation waveform (LO voltage) is applied across the OE-mixing detectors, a photocurrent response is recovered at each detector that is the product or mixing of the RF LO and the intensity modulated light waveforms. (Note that in contrast to optically coherent detection with an FM-CW waveform, for the chirped AM system the LO is an RF voltage waveform not a laser beam, and the frequency of the intensity modulation of the transmitted laser light is chirped, not the frequency of the laser light's electromagnetic wave.) With sawtooth modulation, the instantaneous transmitted and received chirp waveforms differ in frequency (by the IF) because the round-trip light propagation time delays the received chirp waveform with respect to the LO chirp waveform. Thus, mixing of the LO and return signal in the detectors produces a sinusoidal photocurrent (IF waveform) whose frequency is proportional to range.

To recover the photocurrents from the individual pixels, we employ a unique read-out based on CDMA techniques discussed in more detail in a previous publication (Stann *et al.*, 2003). To implement this technique, we multiply the LO signal for each column of the array by a unique binary code that induces a phase shift on the LO of either 0 or 180 degrees. The net effect is to nearly orthogonalize the photocurrents for all pixels in a row, thus allowing all currents to be combined into a single wire that connects to a transimpedance amplifier (TIA) to the side of the detector array. The TIA converts the photocurrents into a voltage that is then sampled at the code clock rate by an analog-to-digital converter (A/D). The CDMA FPA read-out technique was originally conceived because it allowed us to obtain read-out function with discrete electronic components without having to build a custom read-out integrated circuit (ROIC), which requires substantial development time and is usually expensive. As mentioned, the CDMA technique multiplexes all of the pixel photocurrents in a row onto a single row readout line. This reduces the number of readout lines from 1024 to 32 for the 32×32 pixel FPA. The CDMA readout also requires that different codes modulate the LO drives for each column, so that 32 code modulated LO drive lines are required for the 32×32 pixel array. Thus, the total number of drive and readout lines is 64 for the 32×32 pixel array. In general, an $N \times N$ pixel array will require N readout and N LO lines (a total of $2N$) for the CDMA read-out. Demuxing of the data is done digitally in software on a PC. Here the data from each row's A/D is successively multiplied by the codes associated with each pixel, which are generated in software, to recover the IF signal from the respective pixels. The fast Fourier transform (FFT) is taken for each of the demuxed photocurrents, i.e., each of the IF signals, to map the data space into a three-dimensional image space.

32 × 32 Pixel Laboratory Breadboard Hardware

Fig. 8 shows the front view of our laboratory breadboard. We illuminate with a 3.2-W GaAs multi-stripe diode laser where some of the power is coupled into a fiber cable and directed to allow flood illumination of the receiver field of view (FOV). The receiver lens is 5 cm in diameter, and its position is adjustable to center the FOV on the FPA and to adjust the focus.

A fan-in/fan-out board feeds-in code-modulated LO signals to the FPA columns and feeds-out the photocurrents from each FPA row. Circuitry to apply the codes to the FPA columns are on separate boards that plug-into the rear of the fan-in/fan-out board. Each column driver board supports circuitry to drive two detector array columns. The RF LO signal enters the board from a single input and is immediately fed into a two-way power splitter. Each two-way splitter output is fed into the LO port of a balanced mixer where they are bi-phase modulated by the PN code applied to the IF port. These modulated outputs are then amplified by a pair of MMICs amplifiers that are connected to microstrip lines on the fan-in/fan-out board. CAT5 cables carry the code signals and

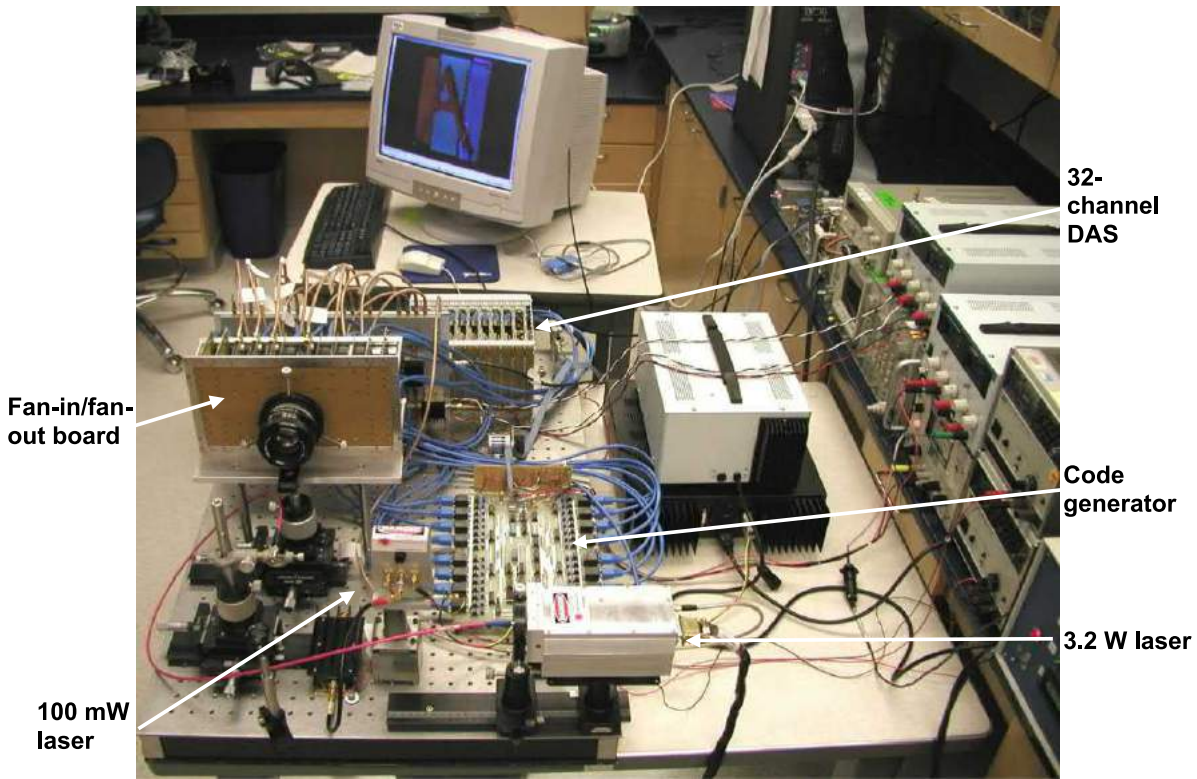


Fig. 8 Front (top) view of 32×32 pixel laboratory breadboard. Source: SPIE.

power to the driver board. The RF bandwidth extends from 200 to 800 MHz and each channel can deliver up to 100 mW into a $50\text{-}\Omega$ load. This signal level is more than adequate to drive the GaAs detector array and is likely to be adequate to drive an InGaAs array.

The FPA row output currents are fed through the fan-out board to the TIA boards plugged into the back of the fan-in/fan-out board. Each TIA board supports four amplifier channels. Each channel includes a TIA stage based on a new wideband, FET input, operational amplifier that allows high transimpedance gains with wide bandwidths. This capability is especially important for achieving low noise amplification which in-turn increases the lidar's maximum range. The current operating parameters of each amplifier chain are as follows: a transimpedance gain of $300\text{ M}\Omega$, a low-pass cutoff of 200 kHz, and a highpass cutoff of 50 Hz. CAT5 computer patch cables connect the power, grounds, and output signals to a 32-channel DAS that acquires the TIA amplifier outputs. The 32-channel DAS can sample (with 8 bits) each channel at up to a 400 kHz rate; for the current experiments we sample at a 93 kHz rate to facilitate debugging of the lidar.

The code generator appears in the lower middle of the picture. The code generator is based on a classical linear feedback shift register (LFSR) design where the outputs of the fourteenth and fifteenth stages of the LFSR are fed into an exclusive OR, and the resulting output is then fed into the first stage of the LFSR. This process yields a fifteen-bit maximal length binary PN code that repeats every 32,767 ($2^{15} - 1$) clock cycles. The code sequence is clocked out at 93 kHz into a series of parallel-out shift registers that provide 32 replicas of the code delayed by two clock periods (i.e., 2 chips). Additional circuitry is included on the board to restart the code sequence at the beginning of each FM chirp, and to inject a seed word into the code generator to insure that the sequence is initialized identically for every chirp. To mitigate transient effects in the driver and TIA boards, the code is run continuously except for a $10\text{-}\mu\text{s}$ period at the beginning of each chirp in which the registers are cleared, and the seed word is injected.

The chirp generator is built around a low-cost, commercial, direct digital synthesizer (DDS) chip (not shown in Fig. 8). Output from the DDS is multiplied-up in frequency by phase-locking to a voltage-controlled microwave oscillator. The basic operating parameters for the chirp generator, including FM deviation, center frequency, and chirp rate, are entered on the lidar control and processing PC which then feeds the data to the chirp generator. The nominal parameter settings are 600 MHz for the FM deviation, 500 MHz for the center frequency, and 3 Hz for the chirp rate. Once the operating parameters are loaded, the chirp generator runs continuously, and sends a digital low-to-high transition at the beginning of each chirp to initialize and start the code generator, and to start the data acquisition process.

Lidar control, signal processing, and display is performed on a PC. A Labview code performs the lidar control and data acquisition functions along with providing display of some intermediate signal processing steps to aid in monitoring system performance. The Labview code captures data from the 32-channel DAS at a 3-Hz frame-rate and stores it on a hard disk. A "C"

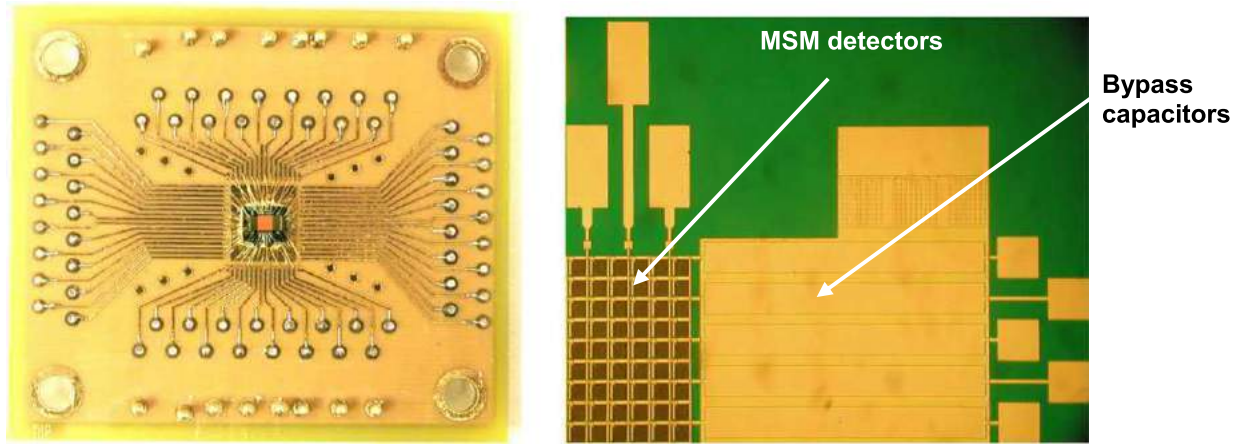


Fig. 9 Metal-semiconductor-metal (MSM) detector array on PC board mount (left), upper right corner of FPA chip showing bypass capacitors (right). Source: SPIE.

code is running concurrently that reads the data from the hard disk, performs the demuxing operations required for the FPA read-out, computes the Fourier transform of the resulting IF signals to range-gate the data, and displays the data in stereo. This process occurs continuously requiring a period of one to two seconds. The interactive 3-D display (stereo display), written at ARL, thresholds the intensity-versus-range signal for each pixel and draws a rectangle at each range where the threshold is exceeded. The size of the rectangle matches the area subtended by the pixel at that range, so a large flat object would be precisely tiled by the rectangles, but rough or very small objects would appear like a point cloud. The display allows a user to zoom and pan around the data and also allows stereoscopic viewing. The depth perception afforded by the stereoscopic viewing and the capability to change perspective enables the user to interpret the data in much the same way as he or she views complex scenes in real life.

MSM Detector Design and Mounting Scheme

The mounting technique for the FPA is shown in the left of [Fig. 9](#). Here LO signals are input from the fan-in/fan-out board to IC socket pins that are soldered into the FPA mount. Microstrip lines (i.e., conductors routed over a ground plane) then carry the LO signals to pads in close proximity to the FPA mounting position; standard wirebonds connect these pads to bonding pads on the FPA chip. The width of the microstrip conductors and the dielectric thickness were made as small as possible but consistent with commercial printed circuit board fabrication techniques to allow maximum distance between lines and therefore achieve minimum cross-talk. This new board also featured IC socket pins that mated with the pins soldered into the FPA mounting board to facilitate changing of the FPA. S_{12} measurements on the new fan-in/fan-out board and FPA mounting board showed that the cross coupling was usually less than -40 db over the lidar's RF bandwidth.

The FPA is 32×32 -pixel interdigitated-finger MSM photodetector array with integrated capacitors at the row outputs shown on right of [Fig. 9](#). The integrated capacitors will replace the discrete capacitors mounted on the fan-in/fan-out board. These capacitors bypass the LO signal currents to ground thereby suppressing LO signals that may enter the TIAs and create noise. They additionally prevent the row busses from being modulated by the LO voltages which would create pixel-to-pixel cross-talk. The integrated capacitors perform the bypassing function more effectively than the discrete capacitors because the lead inductance between the FPA row output and ground will be much lower.

The 32×32 FPA of $60 \mu\text{m}$ pixels is fabricated on semi-insulating GaAs. The thickness of the second silicon nitride layer is such that the capacitance at the end of each row is approximately 15 pf.

Results of Laboratory Breadboard Tests

In this section we briefly discuss our most recent imaging experiments with the breadboard lidar to give the reader some measure of its performance. For these tests, the bandwidth of the modulation is 600 MHz; this establishes a distance of 0.25 m between range cells. All data for a single lidar frame is collected in approximately 0.33 s; this yields a lidar frame rate of 3 Hz that is also the final signal processing bandwidth. The receiver lens diameter is 5.0 cm. As mentioned, some light power from the 3.2-W laser illumination module is coupled into a fiber optic patch cord and light at the output of the cable is then passed through a lens system to form a circular flood illumination field that overlaps the FOV of the receiver. This technique yields a much more uniform illumination field than can be obtained by using the light that is emitted directly out of the laser. The fiber collects 300 mW of laser light; the portion of this light that is in the FOV of the receiver is about 200 mW thus each pixel is illuminated with approximately 0.2 mW of optical power.

The first imaging test examined the lidar performance for a single pixel at one range. Here the collimator of the 100-mW laser illuminator was adjusted so that a small spot (less than the size of a projected pixel) was projected on the wall approximately 11-m

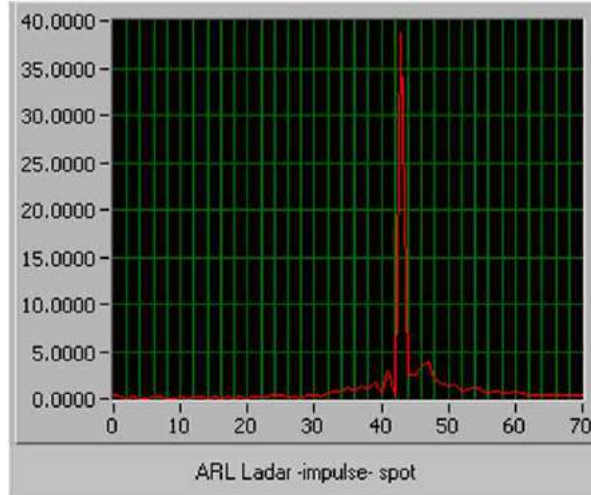


Fig. 10 Ladar impulse response (range dimension). Source: SPIE.

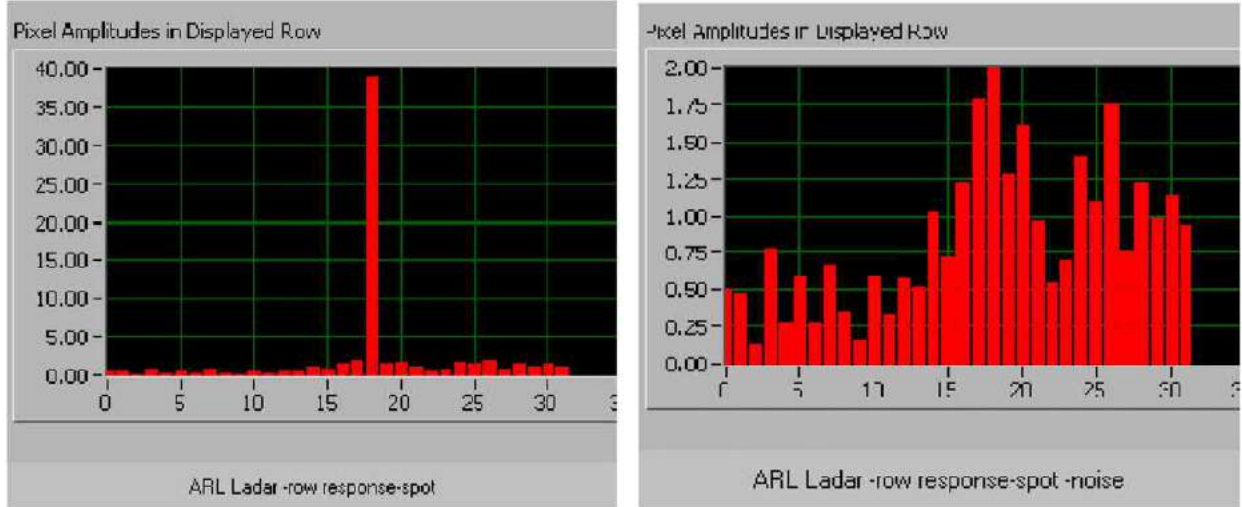


Fig. 11 Impulse response (cross-range dimension; expanded vertical scale on right). Source: SPIE.

distance from the breadboard. The range performance or impulse response in the range dimension is shown in Fig. 10. The horizontal axis is displayed in units of resolution cells which for the amount of FM deviation used (600 MHz) corresponds to 0.25 m; the vertical axis is simply amplitude in arbitrary units. The response peaks at 43 units (11.75 m), as expected, with an amplitude of 38 units; low responses before and after the peak (2.5 and 4 units of amplitude, respectively) are caused by reflections on the microwave transmission lines feeding various components in the breadboard and can be eliminated by improving impedance matching at those points. Overall, the performance in range is acceptable and can be easily improved by more careful attention to the microwave design.

The impulse response in the cross-range or azimuthal dimension is shown in left of Fig. 11. In this figure, the horizontal axis is expressed as pixel number (0 corresponds to pixel number 1) and the vertical axis is again in arbitrary units; the light is positioned on row 21, column 19. As expected, largest amount of signal energy appears at the correct pixel position, however lower values appear at all the other pixel positions. In actuality this will happen because the code read-out concept as implemented here with maximal length pseudorandom sequences will have some non-zero signal level in the other pixel positions. This level of cross-talk is for the most part established by the length of the sequence; other factors such as bandlimiting of any amplifier or filter in the code application or IF signal chain will also affect the cross-talk level (Stann *et al.*, 2003). Considering the situation where the code length (32,767 bits is used in the breadboard) is the only factor, the cross-talk-to-signal ratio should be about -45 db. To better examine the cross-talk level we expanded the vertical scale and displayed the results in the right half of Fig. 11. Here we found that the cross-talk level in pixels to the left of pixel 19 is about -43 db after correcting for the Johnson noise present when the illumination was shut-off. This result is satisfactory considering that any filtering effects in the breadboard will degrade cross-talk

level over the ideal situation. For pixels to the right of pixel 19, the cross-talk level is approximately 10 db higher; this level is expected to degrade image quality particularly in situations where large portions of a target are at a same range. We believe that the cause for this elevated cross-talk level on the right side of the signal is that the LO signals that are routed on the right hand side of the fan-in/fan-out board are coupling to the IF output lines that are also routed on the right hand side of the board and then feeding back to the FPA pixel output bus. This causes a small amount of LO signal from all of the LO lines on the right hand side of the fan-in/fan-out board to be applied to all the detectors in a row. When this happens, the photocurrent in the illuminated pixel becomes modulated at a low level by the codes for other columns. The demuxing step then reports that some signal energy is present in those codes or pixel positions. As mentioned we believe that the problem can be mitigated by improving the RF bypassing of the FPA row output busses by building-in capacitors at the end of each row of the FPA chip. In summary the azimuthal cross-talk is reduced about 20 db over the results achieved in the prior year with the FPA mounted in the leadless chip carrier.

For the imaging test, we set up a scene about 10 m from the breadboard in the back of our lab bay as shown in Fig. 12. Here a cardboard "A" is supported on a pedestal followed by a poster board to the left side of the "A" and 0.25 m further in range. A brown cardboard box is positioned roughly behind the poster board about 0.75 m further in range. A white cardboard box is alongside the brown cardboard box and 0.25 m further in range.

As mentioned earlier, the lidar breadboard can run continuously where a Labview code performs the lidar control and data acquisition functions while a "C" code performs pixel demuxing, range-gate formation, and stereo display. The 3-D image produced by the "C" code updates about every 1–2 s depending on what intermediate data is displayed in the Labview program. This capability is particularly interesting because successive frames are sequentially shown while the viewer can rotate the image to

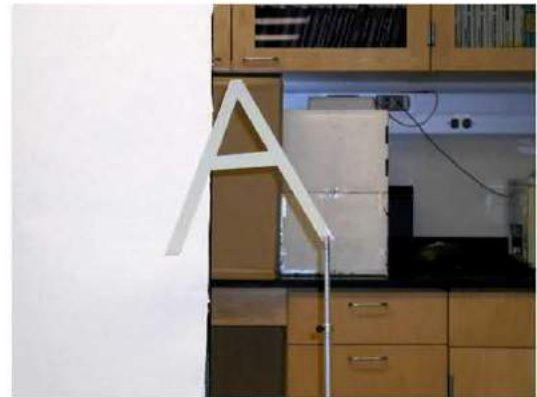


Fig. 12 Laboratory scene. Source: SPIE.

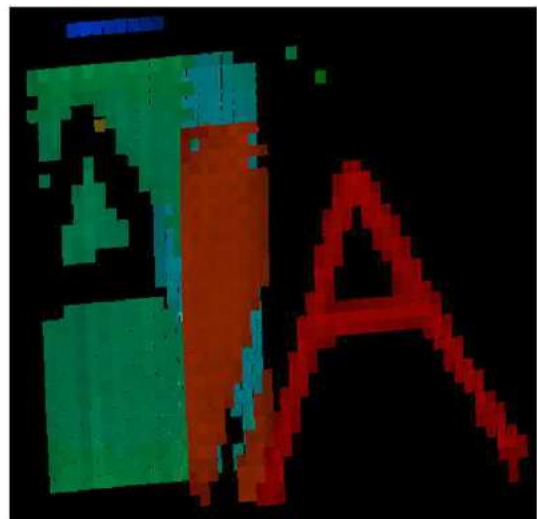
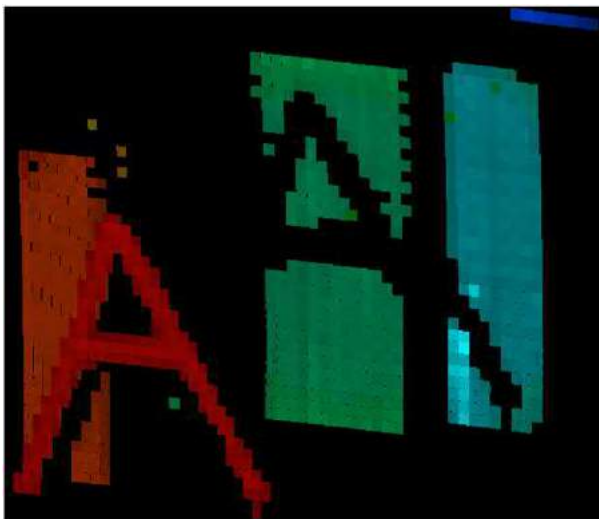


Fig. 13 Breadboard lidar image of laboratory scene. Source: SPIE.

other aspects. In Fig. 13 we show the laboratory scene as imaged from one viewpoint, but displayed at two aspects. The “A” is very well-formed and shows a clear staircase function along the legs as one would expect for an image with pixel-limited resolution. It appears even clearer with the stereo display on the lidar’s PC where degradation caused by the screen capture and other processes needed for presentation in a written report are absent. The poster board one resolution cell behind the “A” is clearly separated from the “A” and exhibits the shadow of one leg of the “A.” About six artifacts are present in the space between the poster board and the two cardboard boxes. These are the result of somewhat less than theoretical performance of the code-multiplexed read-out. As alluded in the previous section, diagnostic measurements indicate that the cause of this degraded performance is LO signals coupling into the IF output lines and then backing-up into the FPA photocurrent busses. We are reasonably confident that the new FPA design with built-in bypass capacitors at the end of each photocurrent buss is expected to attenuate the LO signals and therefore eliminate the image artifacts. The last two boxes are clearly imaged and show the shadow of the “A” and their respective range displacements. The farthest box has a bright region on its left side; this is caused by a glint return from scotch tape on the box.

Conclusions

We have demonstrated the range measurement capability of an incoherent FM/CW lidar where the FM is applied to an RF subcarrier that AMs the light output, instead of to the optical carrier. We outlined the theory of this approach and proved that it is applicable even when the FM bandwidth is greater than the center frequency of the AM. To mitigate the problems of self-clutter created by mixer imbalances or other processes, we proposed and demonstrated in software a simple subtraction technique that can effectively remove false targets caused by such self-clutter. Measurements made with the breadboard imaging lidar have shown that the range resolution is commensurate with the FM bandwidth. An implementation of the self-clutter cancellation technique in the lidar indeed shows that near-range false targets can be eliminated. Measurements have also shown that the lidar can simultaneously resolve and display multiple targets in the same image pixel that are displaced in range by more than the intrinsic lidar resolution. Finally, images collected of a truck demonstrate the potential of this lidar to enhance the contrast of a target in a clutter background.

Overall, this lidar approach should be advantageous for a number of low-cost short to mid-range commercial and military applications, for a variety of reasons. The use of incoherent light detection should inherently produce a lidar that is less costly and complex, and more rugged and reliable, than lidars that use coherent detection. FM ranging techniques have very low post-mixing signal-processing bandwidths, thus substantially simplifying and reducing the cost of the signal processor. Additionally, the ability to detect multiple targets in one pixel may facilitate signal processing for lidars that may have to function in environments where clutter and aerosols are present. Finally, semiconductor laser diodes provide more average power when they are operated CW; thus acceptable lidar SNRs may be easier to achieve using low-power inexpensive lasers.

We built and successfully tested a 32×32 pixel lidar breadboard using the incoherent FM/CW ranging principles. Also demonstrated was the use of a unique code multiplexing scheme to read-out the FPA post-mixing signals that was built using discrete components. The 32×32 pixel lidar breadboard image quality is substantially improved over image quality obtained with prior work when the FPA was mounted in a standard leadless chip carrier. Now, the images contain clear shadows and completely filled-in backdrops. Currently the system is amplifier noise limited where the noise performance is a factor of roughly 10 worse (in a voltage sense) than the level theoretically achievable. For the future we will need to investigate techniques to improve the signal-to-noise of the receiver. The most desirable method is to develop an OE-mixing detector with a modest amount of gain ($10 \times$). We also believe that the OE mixing detector should be designed using InGaAs that will allow the breadboard to operate at eye-safe wavelengths.

References

- Anthes, J.P., Garcia, P., Pierce, J.T., Dressendorfer, P.V., 1993. Nonscanned LADAR imaging and applications. *Proceedings of SPIE: Applied Laser Radar Technology* 1936, 11–21.
- Dieckmann, A., 1994. FMCW-LIDAR with tunable twin-guide laser diode. *Proceedings of SPIE* 2271, 134–142.
- Lee, H.S., Ramaswami, R., 1992. Study of pseudo noise CW diode laser radar for ranging applications. *Proceedings of SPIE: Cooperative Intelligent Robotics in Space III* 1829, 36–45.
- Ruff, W.C., Bruno, J.D., Kennerly, S.W., *et al.*, 2000. Self-mixing detector candidates for an FM/CW lidar architecture. In: *Proceedings of SPIE: Laser Radar Technology and Applications V*, presented at SPIE AeroSense, vol. 4035.
- Shen, P., Stead, M.R., Taysing-Tara, M.A., *et al.*, 2000. Interdigitated finger semiconductor photodetector for optoelectronic mixing. *SPIE AeroSense* 4028, 426.
- Stann, B., Ruff, W., Sztankay, Z., 1995. A practical low-cost high-range-resolution lidar. *Proceedings of SPIE: Applied Laser Radar Technology II* 2472, 118–129.
- Stann, B.L., Abou-Auf, A., Aliberti, K., *et al.*, 2003. Research progress on a focal plane array lidar system using chirped amplitude modulation. *Proceedings of SPIE: Laser Radar Technology and Applications VIII* 5086, 47.
- Stann, B.L., Aliberti, K., Carothers, D., *et al.*, 2004. A 32×32 pixel focal plane array lidar system using chirped amplitude modulation. *Proceedings of SPIE: Laser Radar Technology and Applications IX* 5412, 264.
- Stann, B.L., Ruff, W.C., Sztankay, Z.G., 1996. Intensity-modulated diode laser radar using frequency-modulation/continuous-wave ranging techniques. *Optical Engineering* 35, 3270–3278.

Further Reading

- Skolnik, M.I., 1995. CW and frequency modulated radar. In: *Introduction to Radar Systems*. New York, NY: McGraw-Hill, pp. 72–112 (Chapter 3).

Optical Masks Using Walsh Functions

Lakshminarayan Hazra, University of Calcutta, Kolkata, India
Pubali Mukherjee, MCKV Institute of Engineering, Howrah, India

© 2018 Elsevier Ltd. All rights reserved.

Introduction

For the sake of developing unambiguous specifications for an optical imaging system, it becomes imperative to define a unique image of an object, and so a one-to-one correspondence between points on the object and image is often assumed. This is an approximate model, and except for a few trivial cases, image formation by real optics can never render a point image corresponding to a point object. This is so, primarily for two reasons. First, the presence of residual aberrations in real optical systems, inhibits transformation of the homocentric bundle of rays, originating from a point object and admitted by the imaging system in the object space, into a homocentric bundle of rays converging to the desired point in the image space. Next, even a completely aberration free system does not form a point image corresponding to a point object on account of the inevitable effects of diffraction arising out of finite size of aperture of any real imaging system.

Fig. 1 shows the schematic of an axially symmetric imaging system forming image of the point object O at the image plane located at O' . The entrance pupil and the exit pupil are located at the points E and E' , respectively.

Let (X, Y) and (ξ, η) be the rectangular Cartesian coordinates of a point on the exit pupil plane and the image plane, respectively. The corresponding origins are taken at E' and O' . α is the image space semi-angular aperture, and $(n \sin \alpha)$ is the image space numerical aperture, where n is the refractive index of the image space. Normalized coordinates for points on the exit pupil plane are (x, y) . They are related to the geometrical coordinates by

$$x = \left(\frac{X}{h}\right) \quad y = \left(\frac{Y}{h}\right) \quad (1)$$

where h is the radius of the exit pupil. The normalized coordinates (u, v) for points on the image plane are given by

$$u = \left(\frac{n \sin \alpha}{\lambda}\right) \xi \quad v = \left(\frac{n \sin \alpha}{\lambda}\right) \eta \quad (2)$$

On account of axial symmetry, the normalized distance from the axis of any point with coordinates (X, Y) on the exit pupil plane can be expressed as

$$r = \sqrt{x^2 + y^2} = \frac{\sqrt{X^2 + Y^2}}{h} \quad (3)$$

Similarly the normalized distance p of any point on the image plane with geometrical distance χ from O' is given by

$$q = \left(\frac{n \sin \alpha}{\lambda}\right) \chi = \left(\frac{n \sin \alpha}{\lambda}\right) \sqrt{\xi^2 + \eta^2} \quad (4)$$

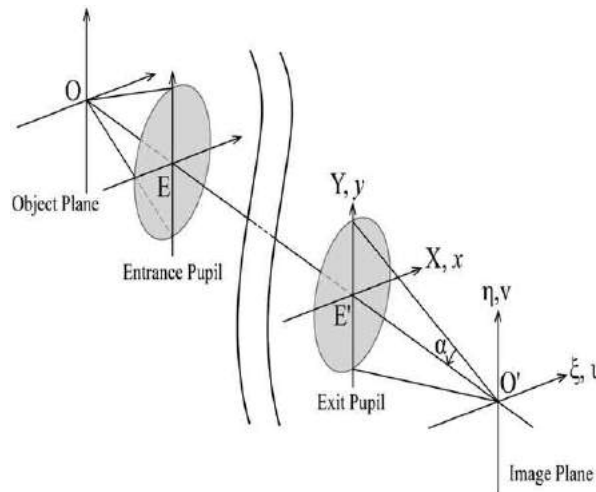


Fig. 1 Object/image planes and entrance/exit pupils of an axially symmetric imaging system.

The complex amplitude $F(u, v)$ on the transverse image plane located at O' is given by the finite Fourier transform of the complex amplitude $f(x, y)$ on the exit pupil at E' , and is given by

$$F(u, v) = \iint f(x, y) e^{i2\pi(ux+vy)} dx dy \quad (5)$$

The limits of integration are set by the finite contour of the pupil. The intensity pattern $I(u, v)$ is equal to the squared modulus of the complex amplitude

$$I(u, v) = |F(u, v)|^2 \quad (6)$$

For axially symmetric systems, the intensity distribution on the blur patch is same along any azimuth on the transverse plane at O' , and the two dimensional Fourier transform of Eq. (5) reduces to a Hankel transform of the circularly symmetric pupil function $f(r)$.

$$F(p) = m \int f(r) J_0(pr) r dr \quad (7)$$

where

$$p = 2\pi q = \frac{2\pi}{\lambda} (n \sin \alpha) \chi \quad (8)$$

and m is a multiplicative constant.

The intensity distribution $I(p)$ is given by

$$I(p) = |F(p)|^2 \quad (9)$$

For aberration free uniform amplitude imaging systems with axially symmetric circular aperture the pupil function $f(r)$ is given by

$$f(r) = \begin{cases} 1, & 0 \leq r \leq 1 \\ 0, & r > 1 \end{cases} \quad (10)$$

The amplitude distribution $\hat{F}(p)$ in the blur patch on the transverse image plane along any azimuth from the optical axis is given by

$$\hat{F}(p) = m \int_0^1 J_0(pr) r dr \quad (11)$$

Defining the normalized amplitude $\hat{F}_N(p)$ as $\hat{F}_N(p) = \hat{F}(p)/\hat{F}(0)$, we get

$$\hat{F}_N(p) = 2 \frac{J_1(p)}{p} \quad (12)$$

The normalized intensity is given by

$$\hat{I}_N(p) = |\hat{F}_N(p)|^2 = \left[\frac{2J_1(p)}{p} \right]^2 \quad (13)$$

where $J_1(p)$ indicates value of the Bessel function (of the first kind) of order 1 for argument p .

This blur patch consists of a bright central lobe surrounded by bright and dark rings with decreasing intensity, and is known as the Airy pattern. The intensity on the transverse plane along any azimuth from O' is shown in Fig. 2. This pattern is known as the Airy pattern. The uniform amplitude pupil of the imaging system is often called the Airy pupil.

Abbe/Rayleigh limit of two-point resolution on the transverse image plane is based on the intensity distribution in the Airy pattern, and the least resolvable distance $\delta\chi$ on the transverse image plane is proportional to $\{\lambda/n(\sin \alpha)\}$, where $n(\sin \alpha)$ is the numerical aperture of the imaging system and λ is the working wavelength. For three dimensional objects, in the two-dimensional image of an axial section of the object the least resolvable distance becomes larger than $\delta\chi$ on account of overlap of defocussed images of neighboring axial sections of the object. The axial distance beyond which this effect becomes insignificant is related with

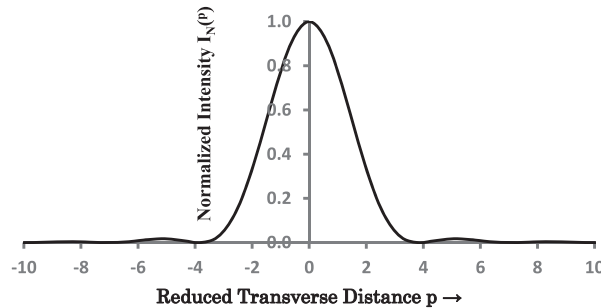


Fig. 2 Airy pattern: the point spread function (PSF) for an aberration-free diffraction limited imaging system.

the location of first zero in the axial distribution of intensity, and this axial resolution $\delta\zeta$ is proportional to $\{\lambda/[n \sin \alpha]^2\}$. It should be noted that both these results refer to aberration-free diffraction limited systems. In presence of aberrations, usually the least resolvable distance obtained in practice is much larger than $\delta\chi$ obtained for aberration free systems.

In order to cater to the needs of diverse applications, it becomes necessary to develop different types of imaging systems with prespecified requirements for $\delta\chi$ and/or $\delta\zeta$; in general, for real time whole field imaging systems, this calls for tailoring of the point spread function (PSF) on the transverse image plane and/or of the distribution of intensity along the axis in the neighborhood of the image plane. In some related applications, sometimes it becomes necessary to tailor the three-dimensional PSF.

Optical Masks

A straightforward approach to tailor the PSF is to use masks over the pupil of the imaging system. From the early days of development of the theories of image formation different types of mask, for example, amplitude, phase and mixed masks are being explored to accomplish the desired goals. Since these three types of mask are lossy, lossless and leaky, respectively, the lossless phase masks are often preferred in practical applications.

Some of the well-known applications of optical masks are:

- apodization
- high frequency enhancement in images
- superresolution
- tailoring the depth of focus
- optical traps.

Apodization is a technique widely used in astronomy for imaging of a faint star in the vicinity of another star that is brighter by several degrees of magnitude, and in spectroscopy for imaging a faint line neighboring a line that is brighter by several degrees of magnitude. This is accomplished with the help of an optical mask on the pupil of the imaging system for lowering down the intensity of the sidelobes in the PSF. It is interesting to note that the name “apodization” is derived from Greek “ α ” implying “to take away,” and “ $\pi\omicron\delta\omicron\sigma$ ” meaning “feet.” In terms of Fourier optics, apodization leads to increase in modulation transfer function (MTF) of the lower spatial frequencies in the image and a concomitant decrease in MTF of the higher spatial frequencies in the image. Optical masks can also be synthesized to accomplish the opposite effect, and this phenomenon is called “High frequency enhancement in images.” However, in both cases, the cut-off frequency in the image is not affected by the use of optical masks. As mentioned earlier, the cut-off frequency in an imaging setup is determined by diffraction effects corresponding to the image space numerical aperture and the working wavelength.

Superresolution refers to imaging techniques that can lead to resolution in the image beyond the diffraction limit. Obviously it calls for narrowing down the central lobe of the diffraction limited PSF; but, any attempt to achieve the same is accompanied with increase in intensity of the sidelobes, so much so that the advantages accruing from narrowing of the central lobe cannot be effectively harnessed for increase in resolution. Attempts for superresolution have, therefore, been directed to post-processing of data obtained from scanning techniques for imaging. The narrow central lobe obtained by optical masks can be used for narrow scanning spot, if the noisy sidelobes are eliminated with the help of a suitable spatial filter. On the other hand, if the PSF is tailored by an optical mask in a manner so that at least several sidelobes surrounding the narrow central lobe have low intensity, the imaging set up can be used to obtain resolution beyond the diffraction limit, albeit over a restricted field of view.

Applications of optical imaging can be classified on the basis of requirements for large depth of focus, or small depth of focus. Note that the depth of focus for diffraction limited aberration free systems is determined by the image space numerical aperture, and the working wavelength; also residual aberrations affect the depth of focus considerably. Optical masks can be conveniently used for tailoring the depth of focus.

Single and multiple optical traps have opened up new frontiers in the field of micro and nanotechnology for manipulation of minute particles. Suitable optical masks for tailoring the number, shape and location of optical traps constitute a major challenge.

For the sake of systematic analysis and synthesis of optical masks to be used on the pupil, the pupil function $f(r)$ describing transmission of the mask is expressed in terms of a set of base functions $b_n(r)$ as

$$f(r) = \sum_{n=0}^{N-1} a_n b_n(r) \quad (14)$$

Typically, Straubel functions, trigonometric functions, Bessel functions, Chebyshev functions, Zernike polynomials, prolate spheroidal wave functions, etc. have been experimented within search for suitable base functions. Often the convenient amenability of obtaining the finite Fourier or Hankel transforms has weighed in favor of choice of base functions. If the members of the set of base functions are mutually orthogonal, the number of terms to be used in specific applications need not be determined heuristically, for the addition or removal of any term of the series does not affect other terms. With this backdrop in view we explore the use of Walsh functions in analysis and synthesis of optical masks.

Walsh Functions

Walsh functions were introduced by Walsh in 1923 as “a closed set of normal orthogonal functions” defined over the interval $(0, 1)$ and having values $+1$ or -1 within the interval. They are a complete set of normal orthogonal functions over a given finite interval and take on values $+1$ or -1 , except at a finite number of points of discontinuity, where they take the value 0. The order of the Walsh function is equal to the number of zero crossings within the specified domain over which the set is defined. The complete set of Walsh functions not only provides a viable alternative but also introduces a significant simplification in the approach of using suitable base functions. They have the interesting property that an approximation of a continuous function over a finite interval by a finite number of base functions of this set leads to a piecewise-constant approximation to the function. Also, the system of Walsh functions possesses some unique properties among all the systems orthonormal in Hilbert space $L_2(0, 1)$. In fact, the system of Walsh functions may be derived from the character group of the dyadic group which is a topologic group derived from the set of binary representations of the real numbers. This topology of the dyadic group makes properties of the system of Walsh functions strikingly different from those of other systems. Walsh functions have been utilized to a great advantage in the field of signal coding and transmission, in allied problems of information processing and in digital signal processing. Two dimensional Walsh functions in the usual rectangular coordinates have been extensively used in digital image processing applications. For the treatment of problems of image formation where one usually deals with some form of axial symmetry, it is useful to develop Walsh functions in polar coordinates. For systems with rotational symmetry about the axis, it is sufficient to deal only with radial Walsh functions. The latter have been developed as a special case of Walsh functions in polar coordinates and have been found to be quite useful in the treatment of problems of pupil plane filtering and the studies on aberrated optical imagery.

One Dimensional Walsh Functions

A Walsh function of a single variable x is represented as $W_k(x)$ where the subscript k refers to the order of the function. The function takes on values either $+1$ or -1 over a specified domain and the set of Walsh functions is orthogonal and complete in the interval. The same implies the following relation for members of the set of Walsh functions defined over the interval $(0, 1)$

$$\int_0^1 W_k(x)W_{\ell}(x)dx = \delta_{k\ell} \quad (15)$$

where $\delta_{k\ell}$ is the Kronecker delta defined as

$$\delta_{k\ell} = \begin{cases} 0, & k \neq \ell \\ 1, & k = \ell \end{cases} \quad (16)$$

The set has a further specification through the unique definition of the zeroth element of the set as

$$W_0(x) = 1, \quad 0 \leq x \leq 1 \quad (17)$$

Fig. 3 shows the first four one-dimensional Walsh functions $W_k(x)$, $k=0, 1, 2$, and 3.

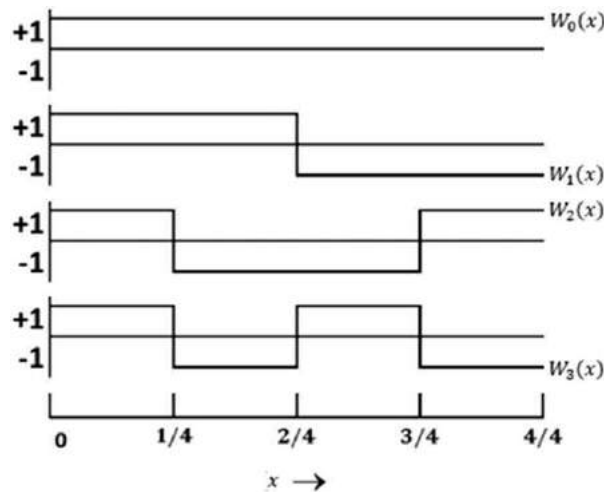


Fig. 3 One dimensional Walsh functions, $W_k(x)$, $k=0,1,\dots,3$.

In order to obtain an analytical representation for the unidimensional Walsh functions $W_k(x)$, let the integer k be expressed as

$$k = \sum_{i=0}^{(v)-1} k_i 2^i \quad (18)$$

where k_i are the bits, 0 or 1 of the binary numeral for k , and 2^v is the integral power of 2, whose value just exceeds k . Thus for $k=0$, 1 , $v=1$; for $k=0, 1, 2, 3$, $v=2$; for $k=0, 1, 2, 3, 4, 5, 6, 7$, $v=3$, and so on.

For all x in the interval $(0, 1)$ we define

$$W_k(x) = \prod_{i=0}^{(v)-1} \text{sgn}[\cos(2^i \pi x k_i)] \quad (19)$$

where

$$\begin{aligned} \text{sgn}(x) &= +1, & x > 0 \\ &= 0, & x = 0 \\ &= -1, & x < 0 \end{aligned} \quad (20)$$

It may be noted from Fig. 3 that the order of the Walsh function is given by the number of zero crossings, i.e., the sign changes of the function in the interval $(0, 1)$.

Two Dimensional Walsh Functions in Rectangular Coordinates

In two dimensions, the orthogonality condition takes the form

$$\int_0^1 \int_0^1 W_k(x) W_\epsilon(y) W_\epsilon(x) W_\epsilon(y) dx dy = \delta_{k\epsilon} \delta_{\epsilon\epsilon} \quad (21)$$

Fig. 4 depicts a few two dimensional rectangular Walsh functions $W_k(x)W_\epsilon(y)$, $k=0,1,2,3$, $\epsilon=0,1,2,3$.

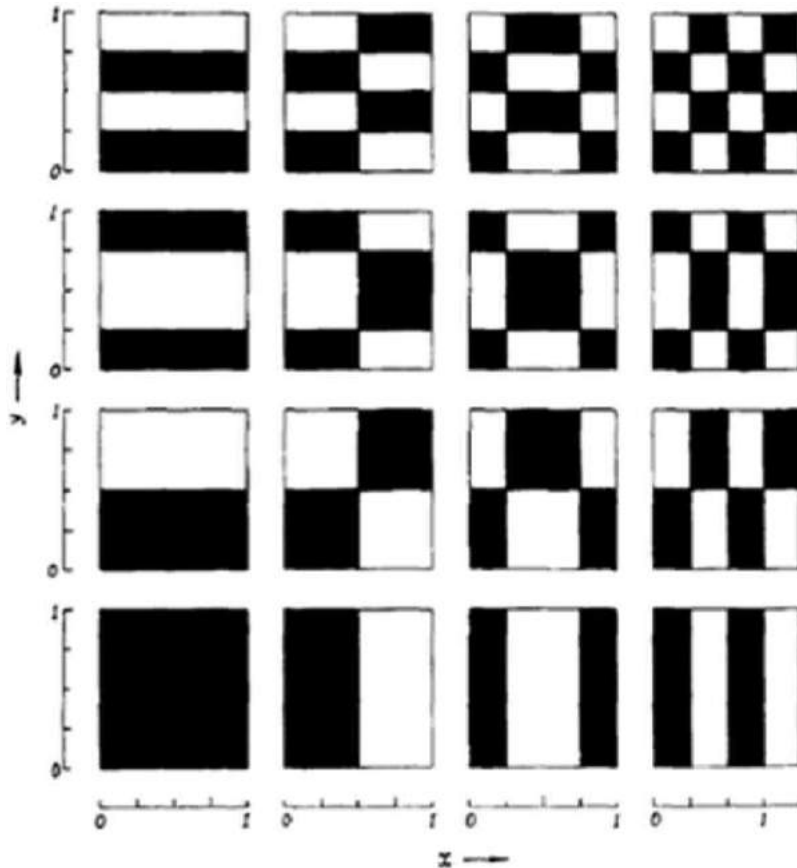


Fig. 4 Two dimensional rectangular Walsh functions $W_k(x)W_\epsilon(y)$, $k=0,1,2,3$, $\epsilon=0,1,2,3$.

Two Dimensional Walsh Functions in Polar Coordinates

In polar coordinates the orthogonality condition reduces to

$$\frac{1}{\pi} \int_0^1 \int_0^{2\pi} \varphi_k(r) \mathcal{A}_\epsilon(\theta) \varphi_\epsilon(r) \mathcal{A}_\epsilon(\theta) r dr d\theta = \delta_{k\epsilon} \delta_{\epsilon\epsilon} \quad (22)$$

where $\varphi_k(r)$ and $\mathcal{A}_\epsilon(\theta)$ are the radial Walsh function and azimuthal Walsh function, respectively.

Fig. 5 depicts a few two-dimensional polar Walsh functions $\varphi_k(r) \mathcal{A}_\epsilon(\theta)$, $k=0,1,2,3$, $\epsilon=0,1,2,3$.

Azimuth Invariant Case: Radial Walsh Functions

Alike the case of unidimensional Walsh functions, $W_k(x)$, radial Walsh functions $\varphi_k(r)$ of index $k \geq 0$ and argument r over the interval $(0,1)$ require integer k to be expressed in the form

$$k = \sum_{i=0}^{\gamma-1} k_i 2^i \quad (23)$$

where k_i are the bits, 0 or 1 of the binary numeral for k and 2^γ is the integral power of 2 that just exceeds k . Thus for all r in $(0,1)$ we define

$$\varphi_k(r) = \prod_{i=0}^{\gamma-1} \text{sgn}[\cos(2^i \pi r^2 k_i)] = \prod_{i=0}^{\gamma-1} \text{sgn}[\cos(2^i \pi t k_i)] = \varphi_k(t) \quad (24)$$

where $t=r^2$.

The first four radial Walsh functions $\varphi_k(r)$, $k=0,1,2,3$ are shown in (r, θ) space in Fig. 6(a). The corresponding radial variation of $\varphi_k(r)$ along any azimuth is shown in Fig. 6(b).

For azimuth invariant case the orthogonality condition takes the form

$$\int_0^1 \varphi_k(r) \varphi_\epsilon(r) r dr = \frac{1}{2} \delta_{k\epsilon} \quad (25)$$

Note that in $t(=r^2)$ space, the variation of $\varphi_k(t)$ with t is identical as the variation of one dimensional Walsh functions $W_k(x)$ with x .

Walsh Filters as Optical Masks on Pupil

Walsh filters of various orders may be obtained from the corresponding Walsh functions by realizing values of $+1$ or -1 with 0 or π phase, respectively. Note: $e^{i0} = +1$, and $e^{i\pi} = -1$. Radial Walsh filters, derived from radial Walsh functions, form a set of orthogonal phase filters that take on values either 0 or π phase, corresponding to $+1$ or -1 values of the radial Walsh functions over prespecified annular regions of the circular filter. Order of these filters is given by the number of zero-crossings, or equivalently phase transitions within the domain over which the set is defined.

Radial Walsh filters may be used as optical masks on the pupil plane of a rotationally symmetric imaging system for tailoring of the distribution of intensity on and around the focal/image plane of optical imaging systems.

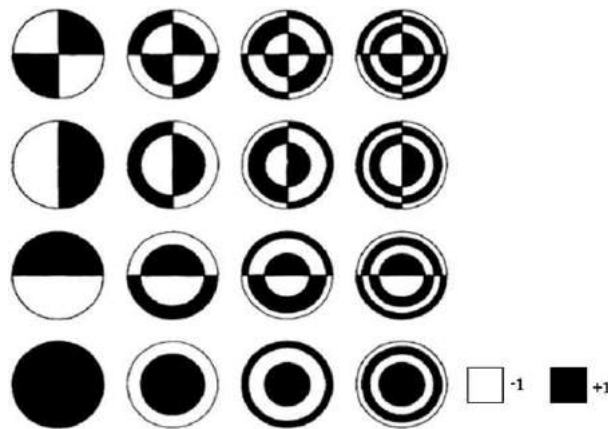


Fig. 5 Two dimensional polar Walsh functions $\varphi_k(r) \mathcal{A}_\epsilon(\theta)$, $k=0,1,2,3$, $\epsilon=0,1,2,3$.

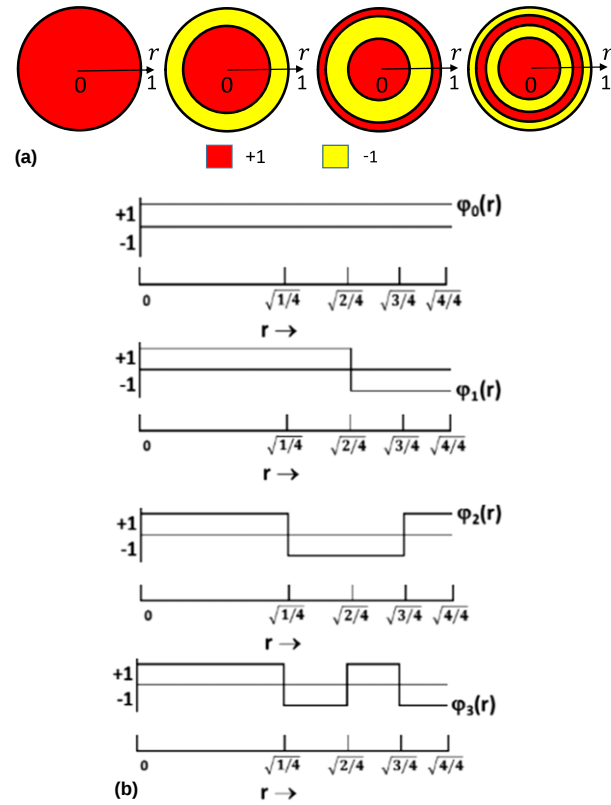


Fig. 6 (a) Radial Walsh functions $\varphi_k(r)$, $k=0,1,\dots,3$ in (r,θ) space. (b) Radial Walsh functions $\varphi_k(r)$ vs. r along any azimuth θ , $k=0,1,\dots,3$.

Intensity Distribution on the Transverse Image/Focal Plane With Walsh filters as Optical Masks

For the k th order radial Walsh filter on the pupil, the pupil function may be expressed as

$$f(r) = \varphi_k(r) = \sum_{m=1}^M e^{ik\psi_m} B_m(r) \quad (26)$$

where $B_m(r)$ are zero-one functions defined as

$$\begin{aligned} B_m(r) &= 1; \quad r_{m-1} \leq r \leq r_m \\ &= 0; \quad \text{otherwise} \end{aligned} \quad (27)$$

where

$$r_m = \left[\frac{m}{M} \right]^{\frac{1}{2}} \text{ and } r_{m-1} = \left[\frac{m-1}{M} \right]^{\frac{1}{2}} \quad (28)$$

and $M=2^J$ is the largest power of 2 that just exceeds k . Over the m th zone of $\varphi_k(r)$, value of $k\psi_m$ is 0 or π , if $\varphi_k(r) = +1$ or -1 , respectively, as per the governing relation Eq. (24).

For better accuracy of results each of the M zones of the pupil is divided into J number of equal area subzones such that the whole exit pupil has $T=MJ$ number of equal area concentric zones (Fig. 7). The outer and inner radii of the j th subzone of the m th zone are given by

$$r_m^j = \left[\frac{(m-1)J + j}{T} \right]^{\frac{1}{2}} \text{ and } r_m^{j-1} = \left[\frac{(m-1)J + (j-1)}{T} \right]^{\frac{1}{2}} \quad (29)$$

respectively.

The normalized complex amplitude distributions on transverse plane in the farfield for Walsh filters $\varphi_k(r)$ is given by

$$F_N(p) = 2 \sum_{m=1}^M \exp[ik\psi_m] \sum_{j=1}^J \int_{r_m^{j-1}}^{r_m^j} J_0(pr) r dr = 2 \sum_{m=1}^M \exp[ik\psi_m] \sum_{j=1}^J \mathcal{J}_m^j(p) \quad (30)$$

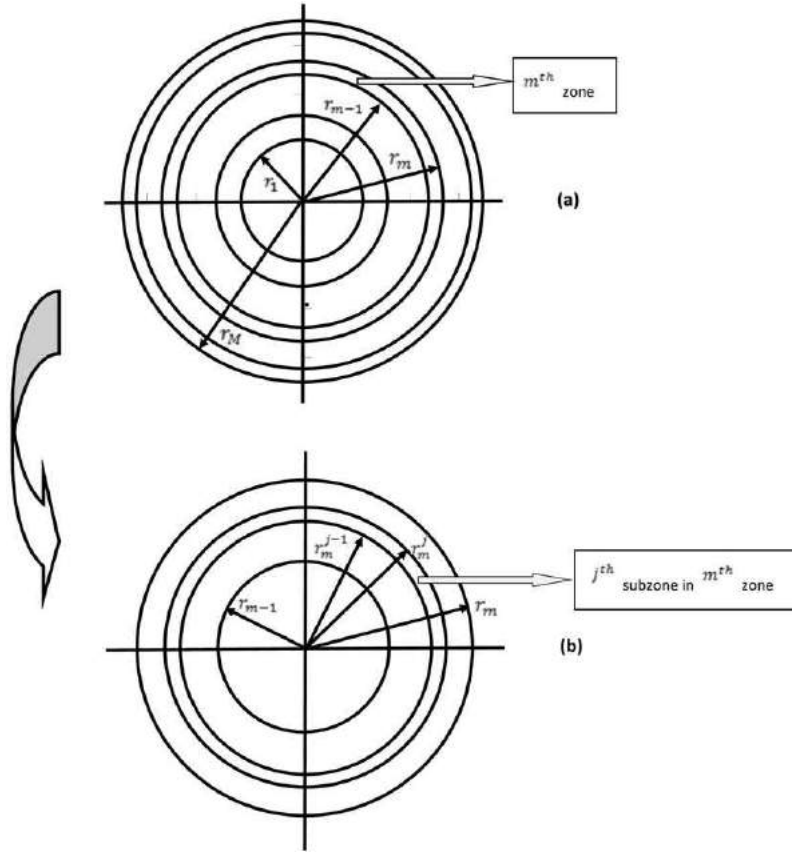


Fig. 7 (a) m th zone in a radial Walsh filter $\varphi_k(r)$ consisting of M concentric equal area zones. (b) j th subzone in the m th zone of the phase filter consisting of J equal area subzones in the m th zone.

where

$$\mathcal{J}_m^j(p) = \left[\frac{r_m^j J_1(pr_m^j) - r_m^{j-1} J_1(pr_m^{j-1})}{p} \right] \quad (31)$$

The normalized intensity distribution on the transverse image plane with radial Walsh filters on $\varphi_k(r)$ the exit pupil is given by

$$I_N(p) = |F_N(p)|^2 = 4 \sum_{m=1}^M \sum_{u=1}^M \cos[k\psi_m - k\psi_u] \sum_{j=1}^J \sum_{v=1}^J \mathcal{J}_m^j(p) \mathcal{J}_u^v(p) \quad (32)$$

For a Walsh filter of zero order on the exit pupil, the transverse intensity distribution on the image/focal plane is the Airy pattern where the normalized central intensity at the origin has magnitude one as in Fig. 4. For any Walsh filter of nonzero order on the exit pupil, the central intensity is zero. There is no central lobe in the transverse intensity distribution on the focal plane in presence of a radial Walsh filter of any order, except for the order zero. A central dip is surrounded by bright and dark rings for all nonzero orders of radial Walsh filters. Fig. 8 presents the intensity distribution on the transverse image/focal plane in presence of selected radial Walsh filters $\varphi_k(r)$, $k=1, \dots, 7$ on the pupil. For the sake of comparison, a part of the Airy pattern is also shown superimposed on the intensity distribution for the filter $\varphi_k(r)$.

Distinct characteristics of the PSF corresponding to individual Walsh filters deserve attention, and may be harnessed in emerging applications like optical encryption. On the other hand, a finite set of orthogonal Walsh filters has already been utilized in the design of optimum amplitude filters for Luneburg's third apodization problem. This problem involves determination of the optimum amplitude filter maximizing the fraction of encircled energy within a circle of prespecified radius w with respect to the total energy in the diffraction pattern. The circle is taken as concentric with the center of the PSF. The amplitude transmission of the optimum filter was taken as consisting of a weighted combination of the first weight radial Walsh functions, so that the pupil function is expressed by

$$f(r) = \sum_{k=0}^7 a_k \varphi_k(r) \quad (33)$$

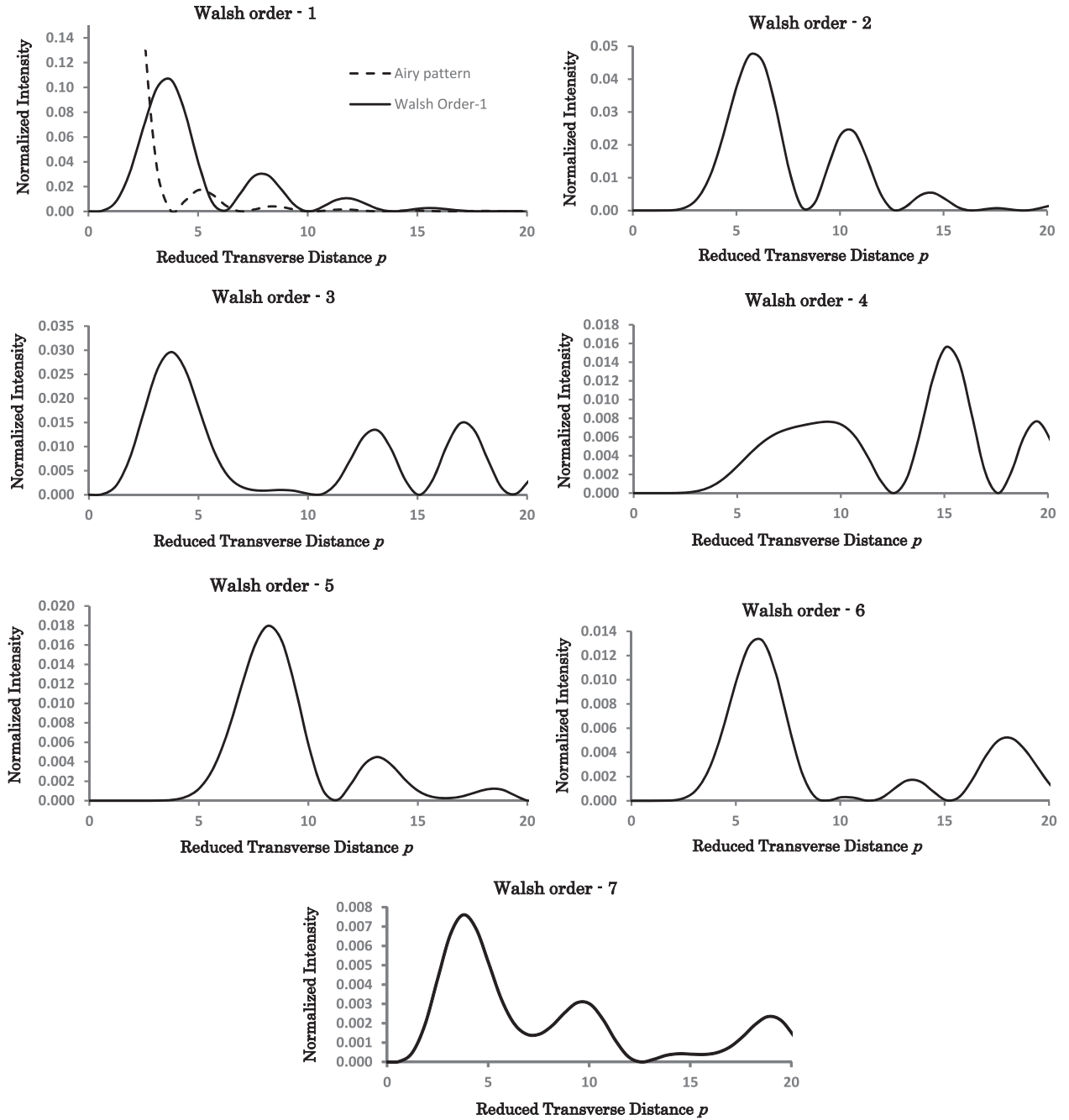


Fig. 8 Normalized intensity distribution on the transverse image/focal plane with radial Walsh filters $\varphi_k(r)$, $k=1, \dots, 7$.

and the apodizing amplitude filter consists of eight concentric equal area annular zones, amplitude over the whole region of each zone being same. Fig. 9 shows five such filters corresponding to five prespecified values for the radius w . w is expressed in the same dimensionless unit as p . Corresponding PSFs on the image plane are shown in Fig. 10. Note the discerning effects of apodization by the five different filters.

Intensity Distribution Along the Axis Around the Image/Focal Plane

The complex amplitude distribution $F(\chi, \Delta\zeta)$ on the transverse plane located at O'' on the axis needs to be calculated by using Fresnel diffraction formula. With reference to Fig. 11 and Fig. 12, let $E'O' = R'_0$, $E'O'' = R'$ and $O'O'' = R' - R'_0 = \Delta\zeta$, where $\Delta\zeta$ refers to the location of the transverse plane with reference to the focal/image plane, and χ refers to the perpendicular distance of a point on the transverse plane from the axis. Detailed derivation of the Fresnel diffraction formula shows that the complex amplitude

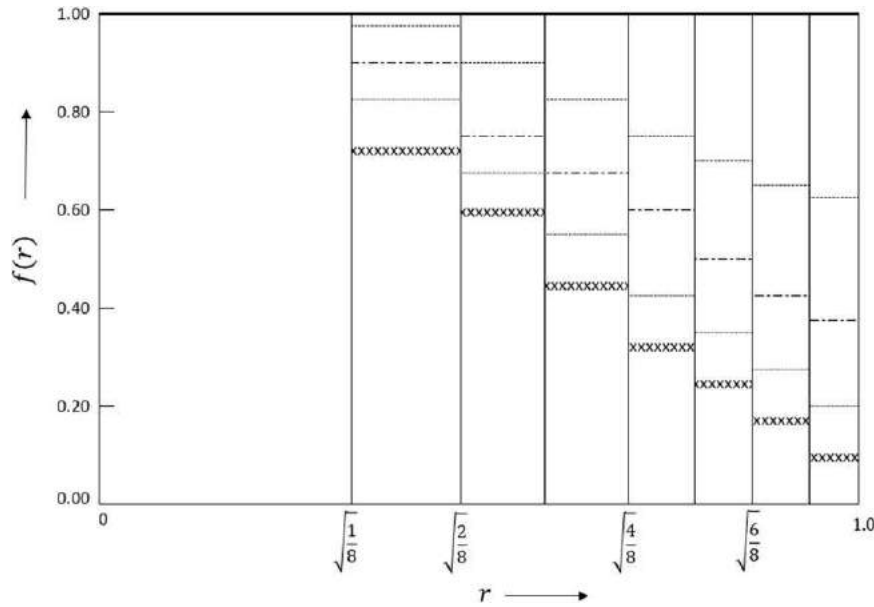


Fig. 9 Optimum eight step apodizing filters corresponding to different values of w . —, $w=0$; — —, $w=2.0$; - · - ·, $w=3.0$; · · · ·, $w=4.0$; xxxx, $w=5.0$. Reprinted with permission from Hazra, L.N., 1977. A new class of optimum amplitude filters. Optics Communications 21, 232–236.

distribution on the transverse plane at O'' is given by a two dimensional Fourier transform of the defocused pupil function of the imaging system. For an axially symmetric pupil function, the 3D distribution of the complex amplitude near the region of the image/focal plane is given by, to a multiplicative factor C ,

$$F(\chi, \Delta\zeta) = C \int_0^1 f(r) \exp[ik W_{20} r^2] J_0(pr) r dr \quad (34)$$

In Eq. (36) below, the axial shift $\Delta\zeta$ of the transverse plane is expressed in terms of “defocus” aberration W_{20} , where the wave aberration function $W(r)$ is given by

$$W(r) = W_{20} r^2 \quad (35)$$

W_{20} is related to $\Delta\zeta$ by

$$W_{20} = \frac{1}{2} n h^2 \left(\frac{1}{R'_0} - \frac{1}{R'} \right) = \frac{1}{2n} (n \sin \alpha)^2 \Delta\zeta \quad (36)$$

where h is the maximum semidiameter of the exit pupil.

Instead of actual distances χ and $\Delta\zeta$, expressions for complex amplitude $F(p, a)$ are derived for normalized dimensionless parameters p and a , respectively. p has been already defined in Eq. (8), and the reduced axial coordinate a is related to $\Delta\zeta$ by

$$a = \frac{1}{n\lambda} (n \sin \alpha)^2 \Delta\zeta \quad (37)$$

where $\Delta\zeta$ is the actual longitudinal shift representing the axial distance of the transverse plane from the paraxial focal plane, as shown in Fig. 11. n represents the refractive index of the image space, and $(n \sin \alpha)$ is the image space numerical aperture.

W_{20} is related to a by

$$kW_{20} = a\pi \quad (38)$$

Notwithstanding the use of Hankel transform in evaluation of the complex amplitude on a defocused plane, it should be noted that the numerical results for intensity remain valid even for larger values of a if the oscillatory integral in Eq. (34) is evaluated correctly. Special measures were undertaken in our numerical quadrature algorithm as noted below.

In presence of phase filter $Q(r)$, the defocused pupil function $f(r)$ is given by

$$f(r) = \exp[ik\{Q(r) + W_{20}r^2\}] \quad (39)$$

For better accuracy of results each of the M zones of the pupil is divided into J number of equal area subzones as in Fig. 7 such that the whole exit pupil has $T=MJ$ number of equal area concentric zones. For each of these subzones the outer and inner radii r_m^j

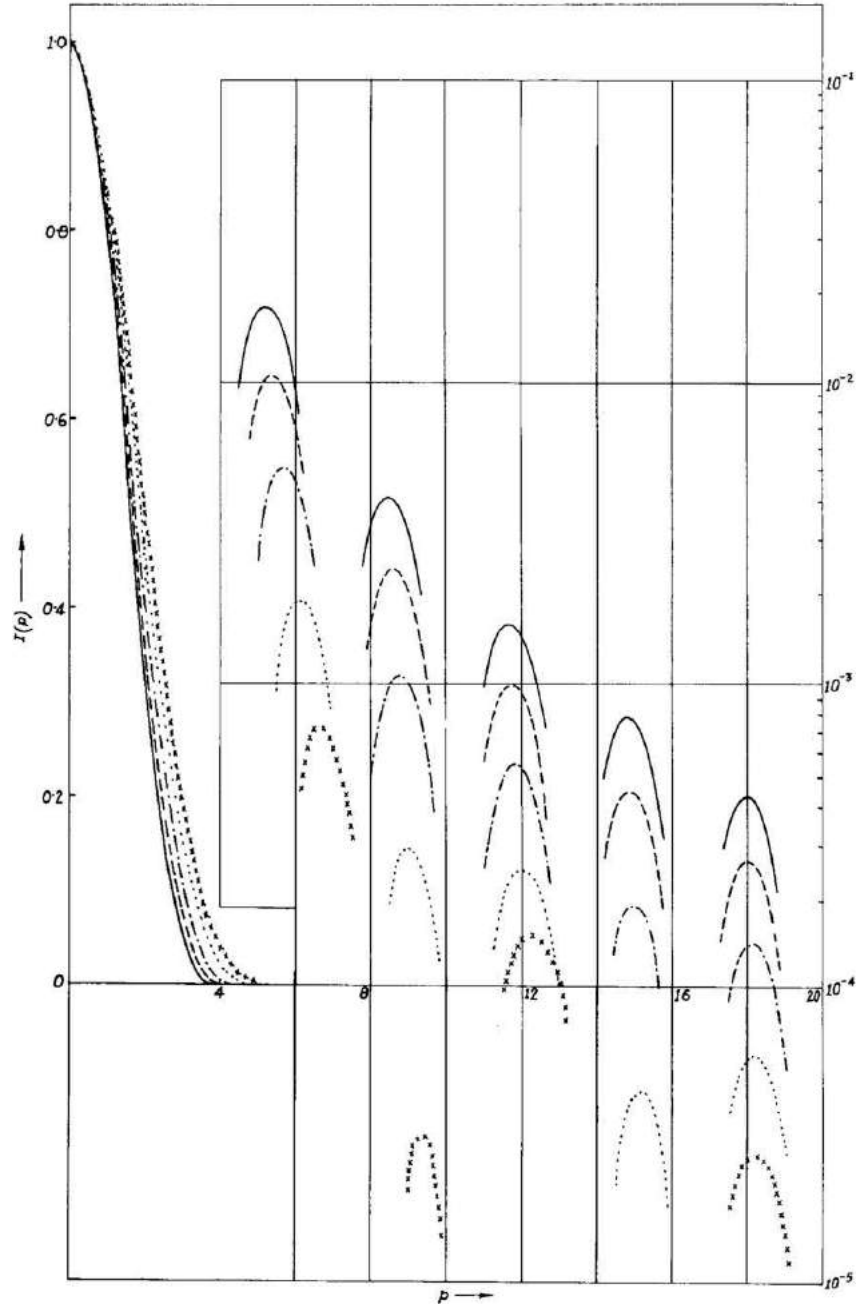


Fig. 10 Point spread functions (PSFs) corresponding to the optimum filters of Fig. 9. The intensity scale is linear up to $p=6.0$; a logarithmic scale is used for $p \geq 4.0$ for better display of the sidelobes. —, $w=0$; — —, $w=2.0$; - - - , $w=3.0$; , $w=4.0$; xxx, $w=5.0$. Reprinted with permission from Hazra, L.N., 1977. A new class of optimum amplitude filters. *Optics Communications* 21, 232–236.

and r_m^{j-1} are given by

$$r_m^j = \left[\frac{(m-1)J + j}{T} \right]^{\frac{1}{2}}$$

$$r_m^{j-1} = \left[\frac{(m-1)J + (j-1)}{T} \right]^{\frac{1}{2}} \quad (40)$$

The normalized complex amplitude at a point O'' on the axis is

$$F_N(a) = F_N(\Delta\zeta) = 2 \sum_{m=1}^M \exp[ik\eta/m] \sum_{j=1}^J \exp[ik\bar{W}_m^j] \int_{r_m^{j-1}}^{r_m^j} r dr \quad (41)$$

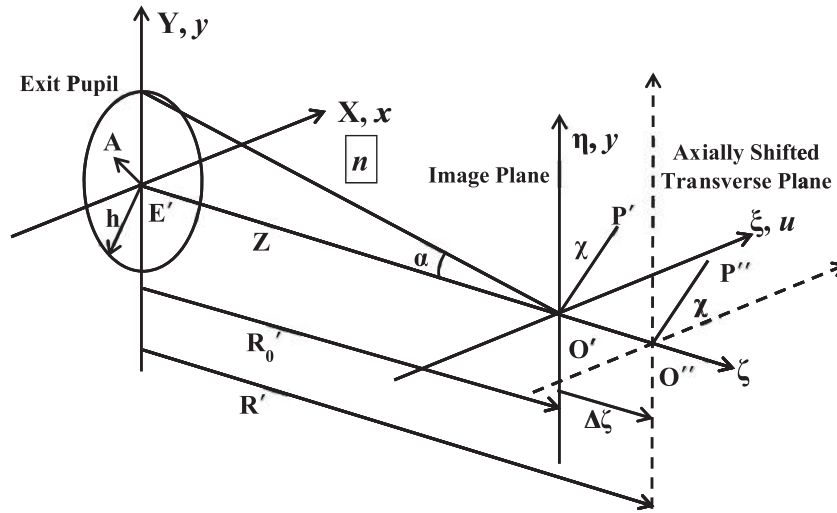


Fig. 11 Exit pupil, image plane and an axially shifted transverse plane in the image space of an axially symmetric imaging system.

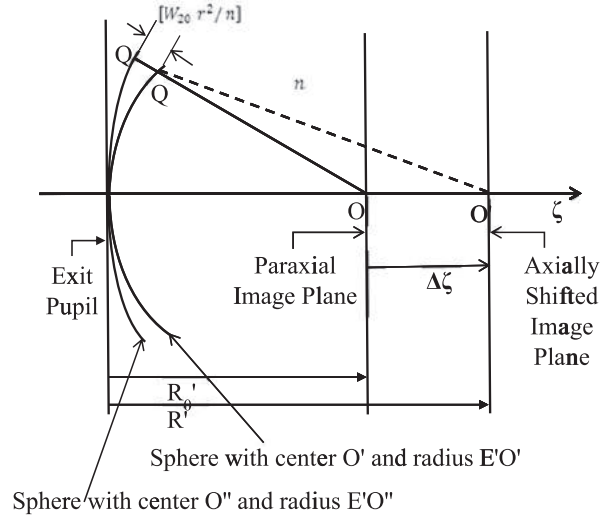


Fig. 12 Shift of image plane by $\Delta\zeta$ gives rise to defocus or “Defect of Focus” aberration [Q”Q’].

The axial shift $O'O'' = \Delta\zeta$ (as in Fig. 12). $\bar{W}_m^j$ is the average value of $W(r)$ over the j th subzone of the m th zone and is given by

$$\overline{W}_m^j = \frac{\int_{r_m^{j-1}}^{r_m^j} W(r) r dr}{\int_{r_m^{j-1}}^{r_m^j} r dr} = \frac{W_{20}}{2T} [2(m-1)J + (2j-1)] \quad (42)$$

and

$$\int_{r_m^{j-1}}^{r_m^j} r dr = \Re_j = \left[\frac{(r_m^j)^2}{2} - \frac{(r_m^{j-1})^2}{2} \right] = \left[\frac{1}{2T} \right] \quad (43)$$

In Eq. (26) it is seen that the radial Walsh filter $\varphi_k(r)$ can be expressed as

$$\varphi_k(r) = \sum_{m=1}^M e^{ik\psi_m} B_m(r) \quad (44)$$

From the defining relation Eq. (24) the value of $\varphi_k(r)$ over the m -th annular zone with inner and outer radii given by Eq. (28) is either $+1$ or -1 ; value of $e^{ik\psi_m}$ on the m -th annular zone of the corresponding radial Walsh filter is either $+1$ or -1 so that the phase $k\psi_m$ on the m -th zone is either 0 or π , respectively.

From Eqs. (41–44)

$$F_N(a) = \frac{1}{T} \sum_{m=1}^M \exp[ik\psi_m] \sum_{j=1}^J \exp[ik\bar{W}_m^j] \quad (45)$$

Therefore, when the radial Walsh filter $\phi_k(r)$ is used as an optical mask over the exit pupil, the normalized intensity distribution along the axis may be expressed as

$$I_N(a) = |F_N(\Delta\zeta)|^2 = \frac{1}{T^2} \sum_{m=1}^M \sum_{u=1}^M \sum_{j=1}^J \sum_{v=1}^J \cos\left[\{k\psi_m - k\psi_u\} + \{k\bar{W}_m^j - k\bar{W}_u^v\}\right] \quad (46)$$

Illustrative results for variation of normalized axial intensity with the distance from the focal/image plane are presented in Fig. 13 when radial Walsh filters $\phi_k(r)$ are used as optical masks on the exit pupil. The positive and negative values of a represent axial shifts on either side of the focal/image plane. For the radial Walsh filter $\phi_0(r)$, which is same as the Airy pupil, the axial intensity on the image plane is maximum. But for all radial Walsh filters of integral order other than zero, the axial intensity on the

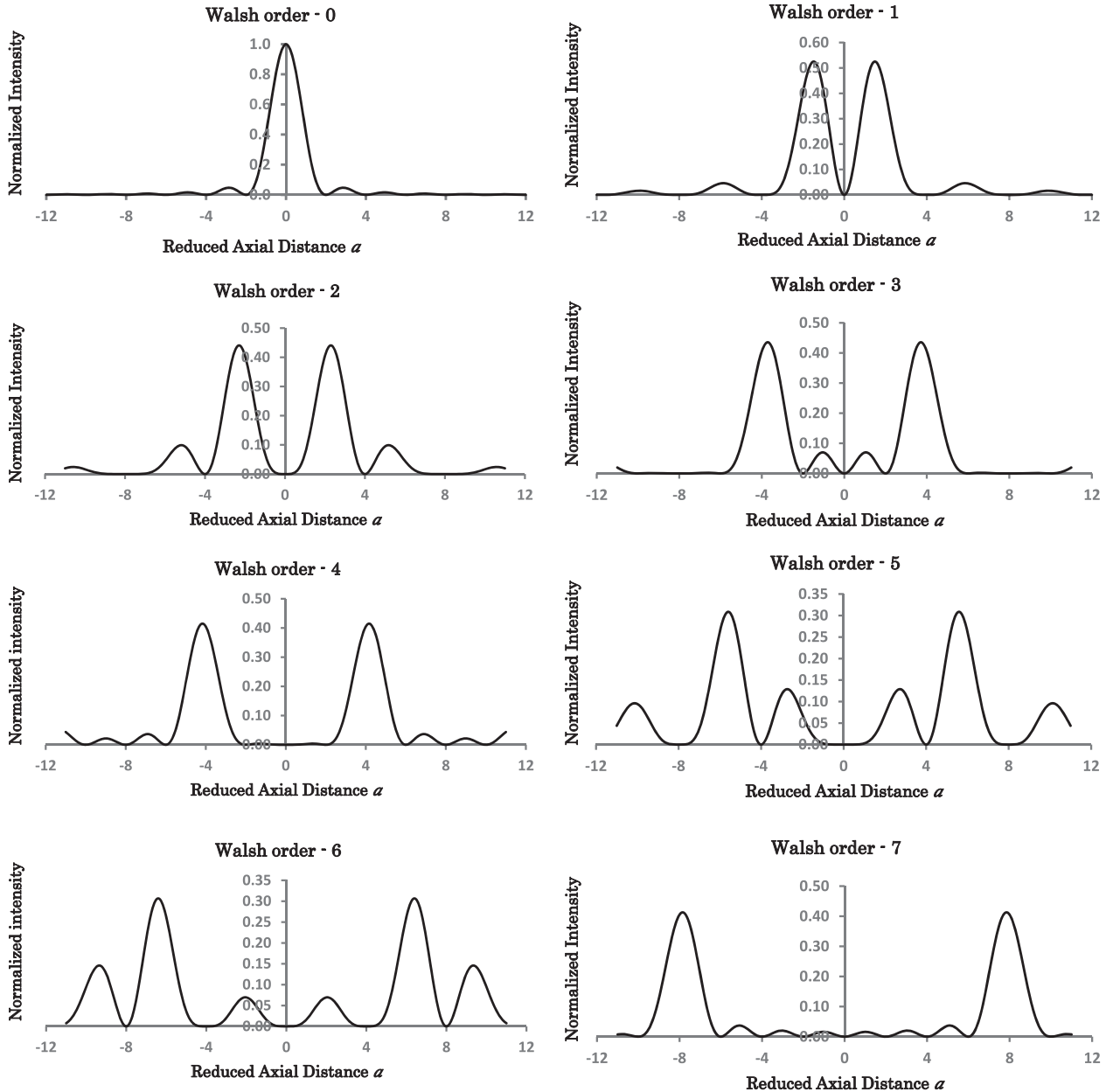


Fig. 13 Normalized intensity distribution along the axis around the transverse image/focal plane with radial Walsh filters $\phi_k(r)$, $k=0,\dots,7$.

image plane is zero. However, for all filters, the intensity along the axis is same for both $+\Delta\zeta$ and $-\Delta\zeta$, where $\Delta\zeta$ represents the distance of the point O'' from O' (Fig. 11).

For Walsh filters of order 0, there is one peak in the axial intensity distribution. In case of radial Walsh filters of lower orders, two distinct peaks are present, and locations and local variation of intensity surrounding the peaks depend on the order of the function. As the order of the radial Walsh filter increases, more peaks of varying intensities appear, but all peaks are located symmetrically with respect to the image plane. Number, location and intensity level in the peaks can be tailored by suitable combination of radial Walsh filters.

Intensity Distribution on Neighboring Transverse Planes Shifted Axially From the Image Plane

Following the analyses presented above, the normalized complex amplitude distribution on a transverse plane located at an axial distance of $\Delta\zeta$ from the image/focal plane (Fig. 11) is given by

$$F_N(p, a) = F_N(\chi, \Delta\zeta) = 2 \sum_{m=1}^M \exp[ik\psi_m] \sum_{j=1}^J \exp[ik\bar{W}_m^j] \int_{r_m^{j-1}}^{r_m^j} I_0(pr) r dr \quad (47)$$

where $\bar{W}_m^j$ is the average value of the wave aberration function over the j th subzone of the m th zone and is given by

$$\bar{W}_m^j = \frac{\int_{r_m^{j-1}}^{r_m^j} W(r) r dr}{\int_{r_m^{j-1}}^{r_m^j} r dr} = \frac{W_{20}}{2T} [2(m-1)J + (2j-1)] \quad (48)$$

and

$$\int_{r_m^{j-1}}^{r_m^j} I_0(pr) r dr = \mathcal{J}_m^j(p) = \left[\frac{r_m^j J_1(pr_m^j) - r_m^{j-1} J_1(pr_m^{j-1})}{p} \right] \quad (49)$$

From Eqs. (47)–(49) it follows

$$F_N(p, a) = 2 \sum_{m=1}^M \exp[ik\psi_m] \sum_{j=1}^J \exp[ik\bar{W}_m^j] \mathcal{J}_m^j(p) \quad (50)$$

When a radial Walsh filter $\phi_k(r)$ is used as an optical mask on the exit pupil, the normalized intensity distribution $I_N(p, a)$ on a neighboring transverse plane located at an axial distance of $\Delta\zeta$ from the image/focal plane is given by squared modulus of $F_N(p, a)$. It follows

$$I_N(p, a) = 4 \sum_{m=1}^M \sum_{u=1}^M \sum_{j=1}^J \sum_{v=1}^J \cos[\{k\psi_m - k\psi_u\} + \{k\bar{W}_m^j - k\bar{W}_u^v\}] \mathcal{J}_m^j(p) \mathcal{J}_u^v(p) \quad (51)$$

A cross section of the three dimensional light distributions around the focal plane produced by some of the radial Walsh filters are presented in Fig. 12. On account of axial symmetry of the imaging system, the intensity distribution is the same along any plane containing the ζ axis. The intensity distribution along (χ, ζ) plane in the focal/image region is presented in Fig. 14 for the radial Walsh filters $\phi_k(r)$, $k=0, 1, 2, \dots, 7$. The elliptical blobs or patches of focussed light seen in the figures are the cross sections of the 3D intensity distributions produced near the focus. Except for the case of $\phi_0(r)$, where a single structure is observed, in all other cases multiple structures are observed. The number and locations of the regions of high light concentration, and the local variation of light intensity in these regions can be tailored by using suitable weighted combination of radial Walsh filters. These regions of high light concentration are effectively multiple optical traps.

Walsh Filters and Self-Similarity

The set of radial Walsh filters can be classified into distinct self-similar groups and subgroups where members of each subgroup possess self-similar structures or phase sequences. The axial and transverse intensity distributions, and also the three dimensional light distribution around the image plane exhibited by one radial filter are similar to the same exhibited by other members of the self-similar group to which the filter belongs. This characteristic of Walsh filters is quite similar to that exhibited by fractal zone plates and other fractal diffracting apertures.

In case of higher order Walsh functions, one or more of the annular zones may become very narrow, so much so that analysis based on scalar diffraction theory may not be valid. Similarly, predictions based on scalar theory are incorrect when optical masks

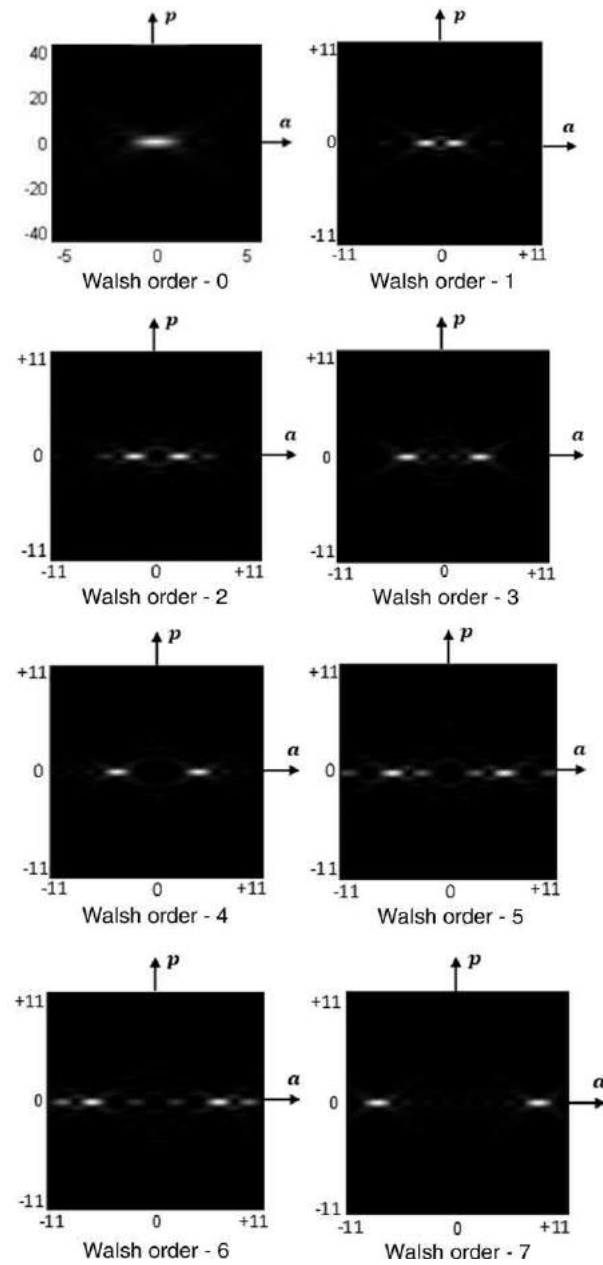


Fig. 14 Intensity distribution along (χ, ζ) plane in the focal/image region with the radial Walsh filters $\varphi_k(r)$, $k=0,1,2,\dots,7$ on the exit pupil.

using Walsh functions are used in imaging systems of high numerical aperture. Both these challenges can be tackled by using vector diffraction theory.

See also: Fraunhofer Diffraction

Further Reading

- Beauchamp, K.G., 1985. Walsh Functions and Their Applications. London: Academic Press.
 Boivin, A., 1964. Théorie et calcul des figures de diffraction de révolution. Paris: Gauthier-Villars.
 Gu, M., 2000. Advanced Optical Imaging Theory. Berlin: Springer.
 Harmuth, H.F., 1972. Transmission of Information by Orthogonal Functions. Berlin: Springer-Verlag.
 Hazra, L.N., 1977. A new class of optimum amplitude filters. Optics Communications 21, 232–236.
 Hazra, L.N., 2007. Walsh filters in tailoring of resolution in microscopic imagery. Micron 38, 129–135.

- Hazra, L.N., Banerjee, A., 1976. Application of Walsh functions in generation of optimum apodizers. *Journal of Optics* 5, 19–26.
- Hazra, L.N., Guha, A., 1986. Far field diffraction properties of radial Walsh filters. *Journal of the Optical Society of America A* 3, 843–846.
- Hazra, L.N., Mukherjee, P., 2015. Self-similarity in radial Walsh filters and axial intensity distributions in the far-field diffraction pattern. *Journal of the Optical Society of America A* 31, 139–147.
- Jacquinot, P., Roizen-Dossier, B., 1964. Apodization. In: Wolf, E. (Ed.), *Progress in Optics*, vol. III. Amsterdam: North-Holland.
- Martinez-Corral, M., Andrés, P., Ojeda-Castañeda, J., Saavedra, G., 1995. Tunable axial superresolution by annular binary filters: Application to confocal microscopy. *Optics Communications* 119, 491–498.
- Neumann, K.C., Block, S.M., 2004. Optical trapping. *Review of Scientific Instruments* 75, 2787–2809.
- Toraldo di Francia, G., 1958. *La diffrazione della luce*. Torino: Edizioni Scientifiche Einaudi.
- Walsh, J.L., 1923. A closed set of normal orthogonal functions. *American Journal of Mathematics* 45, 5–24.

Introduction

Since the inception of quantum mechanics almost a century ago, a prime activity has been the observation of quantum phenomena in virtually all areas of chemistry and physics. However, the natural evolution of science leads to the desire to go beyond passive observation to active manipulation of quantum mechanical processes. Achieving control over quantum phenomena could be viewed as engineering at the atomic scale guided by the principles of quantum mechanics, for the alteration of system properties or dynamic behavior. From this perspective, the construction of quantum mechanically operating solid-state devices through selective material growth would fall into this category. The focus of this article is principally on the manipulation of quantum phenomena through tailored laser pulses. The suggestion of using coherent radiation for the active alteration of microworld processes may be traced to the early 1960s, almost immediately after the discovery of lasers. Since then, the subject has grown enormously to encompass the manipulation of (1) chemical reactions, (2) quantum electron transport in semiconductors, (3) excitons in solids, (4) quantum information systems, (5) atom lasers, and (6) high harmonic generation, amongst other topics. Perhaps the most significant use of these techniques may be their provision of refined tools to ultimately better understand the basic physical interactions operative at the atomic scale.

Regardless of the particular application of laser control over quantum phenomena, there is one basic operating principle involved: active manipulation of constructive and destructive quantum wave interferences. This process is depicted in Fig. 1, showing the evolution of a quantum system from the initial state $|\psi_i\rangle$ to the desired final state $|\psi_f\rangle$ along three of possibly many interfering pathways. In general, there may be many possible final accessible states $|\psi_f\rangle$, and often the goal is to achieve a high amplitude in one of these states and low amplitude in all the others. The target state might actually be a superposition of states, and an analogous picture to that in Fig. 1 would also apply to steering about the density matrix. The process depicted in Fig. 1 may be thought of as a microscale analog of the classic double-slit experiment for light waves. As the goal is often high-finesse focusing into a particular target quantum state, success may call for the manipulation of many quantum pathways (i.e., the notion of a 'many-slit' experiment). Most applications are concerned with achieving a desirable outcome for the expectation value $\langle\psi_f|O|\psi_f\rangle$ associated with some observable Hermitian operator O .

The practical realization of the task above becomes a control problem when the system is expressed through its Hamiltonian $H=H_0+V_c$ where H_0 is the free Hamiltonian describing the dynamics without explicit control; it is assumed that the free evolutionary dynamics under H_0 will not satisfactorily achieve the physical objective. Thus, a laboratory-accessible control term V_c is introduced in the Hamiltonian to achieve the desired manipulation. Even just considering radiative interactions, the form of V_c could be quite diverse depending on the nature of the system (e.g., nuclear spins, electrons, atoms, molecules, etc.) and the intensity of the radiation field. Many problems may be treated through an electric dipole interaction $V_c=-\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ where $\boldsymbol{\mu}$ is a system dipole moment and $\boldsymbol{\varepsilon}(t)$ is the laser electric field. Where appropriate, this article will consider the interaction in this form, but other suitable radiative control interactions may be equally well treated. Thus, the laser field $\boldsymbol{\varepsilon}(t)$ as a function of time (or frequency) is at our disposal for attempted manipulation of quantum systems.

Before considering any practical issues associated with the identification of shaped laser control fields, a fundamental question concerns whether it is, in principle, possible to steer about any particular quantum system from an arbitrary initial state $|\psi_i\rangle$ to an arbitrary final state $|\psi_f\rangle$. Questions of this sort are addressed by a controllability analysis of Schrödinger's equation

$$i\hbar\frac{\partial|\psi(t)\rangle}{\partial t}=[H_0-\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\psi(t)\rangle, \quad |\psi(0)\rangle=|\psi_i\rangle \quad (1)$$

Controllability concerns whether, in principle, some field $\boldsymbol{\varepsilon}(t)$ exists, such that the quantum system described by $H_0=H_0-\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ permits arbitrary degrees of control. For finite-dimensional quantum systems (i.e., those described by evolution amongst a discrete

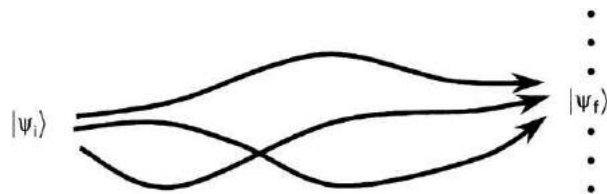


Fig. 1 The evolution of a quantum system under laser control from the initial state $|\psi_i\rangle$ to the final state $|\psi_f\rangle$. The latter state is chosen to have desirable physical properties, and three of possibly many pathways between the states are depicted. Successful control of the process $|\psi_i\rangle \rightarrow |\psi_f\rangle$ by a tailored laser pulse generally requires creating constructive quantum wave interferences in the state $|\psi_f\rangle$ from many pathways, and destructive interferences in all other accessible final states $|\psi_f\rangle \neq |\psi_f\rangle$.

set of quantum states), the formal tools for such a controllability analysis exist both for evaluating the controllability of the wave function, as well as the more general time evolution operator $U(t)$, which satisfies

$$i\hbar \frac{\partial U}{\partial t} = [H_0 - \boldsymbol{\mu} \cdot \boldsymbol{\varepsilon}(t)]U, \quad U(0) = 1 \quad (2)$$

Analyses of this type can be quite insightful, but they require detailed knowledge about H_0 and $\boldsymbol{\mu}$. Cases involving quantum information science applications are perhaps the most demanding with regard to achieving total control. Most other physical applications would likely accept much more modest levels of control and still be categorized as excellent achievements.

Theoretical tools and concepts have a number of roles in considering the control of quantum systems, including (1) an exploration of physical/chemical phenomena under active control, (2) the design of viable control fields, (3) the development of algorithms to actively guide laboratory control experiments towards achieving their dynamical objectives, and (4) the introduction of special algorithms to reveal the physical mechanisms operative in the control of quantum phenomena. Activities (1)–(3) have thus far been the primary focus of theoretical studies in this area, and it is anticipated that item (4) will grow in importance in the future.

A few comments on the history of laser control over quantum systems are relevant, as they speak to the special nature of the currently employed successful closed-loop quantum control experiments. Starting in the 1960s and spanning roughly 20 years, it was thought that the design of lasers to manipulate molecular motion could be achieved by the application of simple physical logic and intuition. In particular, the thinking at the time focused on using cw laser fields resonant with one or more local modes of the molecule, as state or energy localization was believed to be the key to successful control. Since quantum dynamics phenomena typically occur on ultrafast time-scales, possibly involving spatially dispersed wave packets, expecting to achieve high quality control with a light source operating at one or two resonant frequencies is generally wishful thinking. Quantum dynamics phenomena occur in a multifrequency domain, and controls with a few frequencies will not suffice. Over approximately the last decade, it became clear that successful control often calls for manipulating multiple interfering quantum pathways (cf., Fig. 1). In turn, this recognition led to the need for broad-bandwidth laser sources (i.e., tailored laser pulses). Fortunately, the necessary laser pulse-shaping technologies have become available, and these sources continue to expand into new frequency ranges with enhanced bandwidth capabilities. Many chemical and physical applications of control over quantum phenomena involve performing a significant amount of work (i.e., quantum mechanical action), and sufficient laser field intensity is required. A commonly quoted adage is ‘no field, no yield’, and at the other extreme, there was much speculation that operating nonlinearly at high field intensities would lead to a loss of control because the quantum system would effectively act as an amplifier of even weak laser field noise. Fortunately, the latter outcome has not occurred, as is evident from a number of successful high field laser control experiments.

Quantum control may be expressed as an inverse problem with a prescribed chemical or physical objective, and the task being the discovery of a suitable control field $\boldsymbol{\varepsilon}(t)$ to meet the objective. Normally, Schrödinger’s Eq. (1) is viewed as linear, and this perspective is valid if the Hamiltonian is known a priori, leaving the equation to be solved for the wave function. However, in the case of quantum control, neither the wave function nor the control field is known a priori, and the two enter bilinearly on the right-hand side of Eq. (1). Thus, quantum control is mathematically a nonlinear inverse problem. As such, one can anticipate possibly diverse behavior in the process of seeking successful controls, as well as in the ensuing quantum dynamical control behavior. The control must take into account the evolving system dynamics, thereby resulting in the control field being a function (a) of the current state of the system $\boldsymbol{\varepsilon}(\psi)$ in the case of quantum control design, or dependent on the system observations $\boldsymbol{\varepsilon}(\langle O \rangle)$ in the case of laboratory closed-loop field discovery efforts where $\langle O \rangle = \langle \psi | O | \psi \rangle$. Furthermore, the control field at the current time t will depend in some implicit fashion on the future state of the evolving system at the final target time T . It is evident that control field design and laboratory discovery present highly nontrivial tasks.

Although the emphasis in this review is on the theoretical aspects of quantum control theory, the subject exists for its laboratory realizations, and those realizations, in turn, intimately depend on the capabilities of theoretical and algorithmic techniques for their successful implementations. Accordingly, theoretical laser control field design techniques will be discussed, along with algorithmic aspects of current laboratory practice. This material respectively will focus on optimal control theory (OCT) and optimal control experiments (OCEs), as seeking optimality provides the best means to achieve any posed objective. Finally, the last part of the article will present some general conclusions on the state of the quantum control field.

Quantum Optimal Control Theory for Designing Laser Fields

When considering control in any domain of application, a reasonable approach is to computationally design the control for subsequent implementation in the laboratory. Quantum mechanical laser field design has taken on a number of realizations. At one limit is the application of simple intuition for design purposes, and in some special cases (e.g., the weak field perturbation theory regime), this approach may be applicable. Physical insights will always play a central role in laser field design, but to be especially useful, they need to be channeled into the proper mathematical framework. Many of the interesting applications operate in the strong-field nonperturbative regime. In this domain, serious questions arise regarding whether sufficient information is available about the Hamiltonian to execute reliable designs. Regardless of whether the Hamiltonian is known accurately or is an

acknowledged model, the theoretical study of quantum control can provide physical insight into the phenomena involved, as well as possibly yielding trial laser pulses for further refinement in the laboratory.

Achieving control over quantum mechanical phenomena often involves a balance of competing dynamical processes. For example, in the case of aiming to create a particular excitation in a molecule or material, there will always be the concomitant need to minimize other unwanted excitations. Often, there are also limitations on the form, intensity, or other characteristics of the laser controls that must be adhered to. Another goal is for the control outcome to be as robust as possible to the laser field fluctuations and Hamiltonian uncertainties. Overriding all of these issues is merely the desire to achieve the best possible physical result for the posed quantum mechanical objective. In summary, all of these desired extrema conditions translate over to posing the design of control laser fields as an optimization problem. Thus, optimal control theory forms a basic foundation for quantum control field design.

Control field design starts by specification of the Hamiltonian components H_0 and μ in Eq. (1), with the goal of finding the best control electric field $\varepsilon(t)$ to balance the competing objectives. Schrödinger's equation must be solved as part of this process, but this effort is not merely a forward propagation task, as the control field is not known a priori. As an optimal design is the goal, a cost functional $J = J(\text{objectives, penalties, } \varepsilon(t))$ is prescribed, which contains the information on the physical objectives, competing penalties, and any costs or constraints associated with the structure of the laser field. The functional J could contain many terms if the competing physical goals are highly complex. As a specific simple illustration, consider the common goal of steering the system to achieve an expectation value $\langle \psi(T) | O | \psi(T) \rangle$ as close as possible to the target value O_{target} associated with the observable operator O at time T . An additional cost is imposed to minimize the laser fluence. These criteria may be embodied in a cost functional of the form

$$J = \left(\langle \psi(T) | O | \psi(T) \rangle - O_{\text{target}} \right)^2 + \omega \int_0^T \varepsilon^2(t) dt - 2 \Im \int_0^T dt \left\langle \lambda(t) \left| i\hbar \frac{\partial}{\partial t} - H_0 + \mu \cdot \varepsilon(t) \right| \psi(t) \right\rangle \quad (3)$$

Here, the parameter $\omega \geq 0$ is a weight to balance the significance of the fluence term relative to achieving the target expectation value, and $|\lambda(t)\rangle$ serves as a Lagrange multiplier, assuring that Schrödinger's equation is satisfied during the variational minimization of J with respect to the control field. Carrying out the latter minimization will lead to Eq. (1), along with the additional relations

$$i\hbar \frac{\partial}{\partial t} |\lambda(t)\rangle = [H_0 - \mu \cdot \varepsilon(t)] |\lambda(t)\rangle, \quad |\lambda(T)\rangle = 2\sigma O |\lambda(T)\rangle \quad (4)$$

$$\varepsilon(t) = \frac{1}{\omega} \Im \langle \lambda(t) | \mu | \psi(t) \rangle \quad (5)$$

$$\sigma = \langle \psi(T) | O | \psi(T) \rangle - O_{\text{target}} \quad (6)$$

Eqs. (1) and (4)–(6) embody the OCT design equations that must be solved to yield the optimal field $\varepsilon(t)$ given in Eq. (5). These design equations have an unusual mathematical structure within quantum mechanics. First, insertion of the field expression in Eq. (5) into Eqs. (1) and (4) produces two cubically nonlinear coupled Schrödinger-like equations. These equations are in the same family as the standard nonlinear Schrödinger equation, and as such, one may expect that unusual dynamical behavior could arise. Although Eq. (4) for $|\lambda(t)\rangle$ is identical in form to the Schrödinger Eq. (1), only $|\psi(t)\rangle$ is the true wavefunction, with $|\lambda(t)\rangle$ serving to guide the controlled quantum evolution towards the physical objective. Importantly, Eq. (1) is an initial value problem, while Eq. (4) is a final value problem. Thus, the two equations together form a two-point boundary value problem in time, which is an inherent feature of temporal engineering optimal control as well. The boundary value nature of these equations typically leads to the existence of multiple solutions, where σ in Eq. (6) plays the role of a discrete eigenvalue, specifying the quality of the particular achieved control field design. The fact that there are generally multiple solutions to the laser design equations can be attractive from a physical perspective, as the designs may be sorted through to identify those being most attractive for laboratory implementation.

The overall mathematical structure of Eqs. (1) and (4)–(6) can be melded together into a single design equation

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = [H_0 - \mu \cdot \varepsilon(\psi, \sigma, t)] |\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_i\rangle \quad (7)$$

The structure of this equation embodies the comments in the introduction that control field design is inherently a nonlinear process. The main source of complexity arising in Eq. (7) is through the control field $\varepsilon(\psi, \sigma, t)$ depending not only on the wave function at the present time t , but also on the future value of the wavefunction at the target time T contained in σ . This structure again reflects the two-point boundary value nature of the control equations.

Given the typical multiplicity of solutions to Eqs. (1) and (4)–(6), or equivalently, Eq. (7), it is attractive to consider approximations to these equations (except perhaps for the inviolate Schrödinger Eq. (1), and much work continues to be done along these lines. One case involves what is referred to as tracking control, whereby a path is specified for $\langle \psi(t) | O | \psi(t) \rangle$ evolving from $t=0$ out to $t=T$. Tracking control eliminates σ to produce a field of the form $\varepsilon(\psi, t)$, thereby permitting the design equations to be explicitly integrated as a forward-marching problem toward the target; the tracking equations are still nonlinear with respect to the evolving state. Many variations on these concepts can be envisioned, and other approximations may also be introduced to deal with special circumstances. One technique appropriate for at least few-level systems is stimulated Raman adiabatic passage (STIRAP), which seeks robust adiabatic passage from an initial state to a particular final state, typically with nearly total destructive interference

occurring in the intermediate states. This design technique can have special attractive robustness characteristics with respect to field errors. STIRAP, tracking, and various perturbation theory-based control design techniques can be expressed as special cases of OCT, with suitable cost functionals and constraints. Further approximation methods will surely be developed, with the rigorous OCT concepts forming the general foundation for control field design.

As the OCT equations are inherently nonlinear, their solution typically requires numerical iteration, and a variety of procedures may be utilized for this purpose. Almost all of the methods employ local search techniques (e.g., gradient methods), and some also show monotonic convergence with respect to each new iteration step. Local techniques typically evolve to the nearest solution in the space of possible control fields. Such optimizations can be numerically efficient, although the quality of the attained result can depend on the initial trial for the control field and the details of the algorithm involved. In contrast, global search techniques, such as simulated annealing or genetic algorithms, can search more broadly for the best solution in the control field space. These more expansive searches are attractive, but at the price of typically requiring more intensive computations. Effort has also gone into seeking global or semiglobal input $\rightarrow$ output maps relating the electric field $\epsilon(t)$ structure to the observable $O[\epsilon(t)]$. Such maps may be learned, hopefully, from a modest number of computations with a selected set of fields, and then utilized as high-speed interpolators over the control field space to permit the efficient use of global search algorithms to attain the best control solution possible. At the foundation of all OCT design procedures is the need to solve the Schrödinger equation. Thus, computational technology to improve this basic task is of fundamental importance for designing laser fields.

Many computations have been carried out with OCT to attain control field designs for manipulating a host of phenomena, including rotational, vibrational, electronic, and reactive dynamics of molecules, as well as electron motion in semiconductors. Every system has its own rich details, which in turn, are compounded by the fact that multiple control designs will typically exist in many applications producing comparable physical outcomes. Collectively, these control design studies confirm the manipulation of constructive and destructive quantum wave interferences as the general mechanism for achieving successful control over quantum phenomena. This conclusion may be expressed as the following principle:

Control field–system cooperativity principle: Successful quantum control requires that the field must have the proper structure to take full advantage of all of the dynamical opportunities offered by the system, to best satisfy the physical objective.

This simple statement of system–field cooperativity embodies the richness, as well as the complexity, of seeking control over quantum phenomena, and also speaks to why simple intuition alone has not proved to be a generally viable design technique. Quantum systems can often exhibit highly complex dynamical behavior, including broad dispersion of the wave packet over spatial domains or multitudes of quantum states. Handling such complexity can require fields with subtle structure to interact with the quantum system in a global fashion to manage all of the motions involved. Thus, we may expect that successful control pulses will often have broad bandwidth, including amplitude and phase modulation. Until relatively recently, laser sources with these characteristics were not available, but the technology is now in hand and rapidly evolving (see the discussion later).

The cooperativity principle above is of fundamental importance in the control of all quantum phenomena, and a simple illustration of this principle is shown in Fig. 2 for the control of wave packet motion on an excited state of the NO molecule.

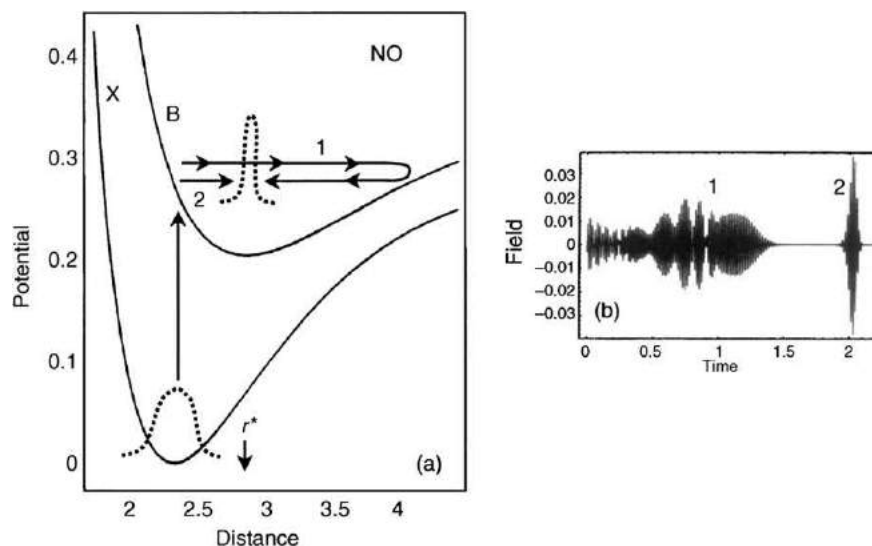


Fig. 2 Control of wave packet evolution on the electronically excited B state of NO, with the goal of creating a narrow final packet over the excited state well. (a) indicates that the ground-state packet is brought up in two pieces using the coordinated dual pulse in (b). The piece of the packet from the first pulse actually passes through the target region, then bounces off the right side of the potential, to finally meet the second piece of the packet at just the right time over the target location r^* for successful control. This behavior is an illustration of the control field–system cooperativity principle stated in the text.

The target for the control is a narrow wave packet located over the well of the excited Bstate. The optimal control field consists of two coordinated pulses, at early and late times, with both features having internal structure. The figure indicates that the ensuing excited state wave packet has two components, one of which actually passes through the target region during the initial evolution, only to return again and meet the second component at just the right place and time, to achieve the target objective as best as possible. Similar cooperativity interpretations can be found in virtually all implementations with OCT.

The main reason for performing quantum field design is to ultimately implement the designs in the laboratory. The design procedures must face laboratory realities, which include the fact that most Hamiltonians are not known to high accuracy (especially for polyatomic molecules and complex solid-state structures), and secondly, a variety of laboratory field imperfections may unwittingly be present. Notwithstanding these comments, OCT has been fundamental to the development of quantum control, including laying out the logic for how to perform the analogous optimal control experiments. Design implementations and further OCT development will continue to play a basic role in the quantum control field. At present, perhaps the most important contribution of OCT has been to (1) highlight the basic control cooperativity principle above, and (2) provide the basis for developing algorithms to successfully guide optimal control experiments, as discussed below.

Algorithms for Implementing Optimal Control Experiments

The ultimate purpose of considering control theory within quantum mechanics is to take the matter into the laboratory for implementation and exploitation of its capabilities. Ideally, theoretical control field designs may be attained using the techniques discussed above, followed by the achievement of successful control upon execution of the designs in the laboratory. This appealing approach is burdened with three difficulties: (1) Hamiltonians are often imprecisely known, (2) accurately solving the design equations can be a significant task, and (3) realization of any given design will likely be imperfect due to laboratory noise or other unaccounted-for systematic errors. Perhaps the most serious of these difficulties is point (1), especially considering that the best quality control will be achieved by maximally drawing on subtle constructive and destructive quantum wave interference effects. Exploiting such subtleties will generally require high-quality control designs that, in turn, depend on having reliable Hamiltonians. Although various designs have been carried out seeking robustness with respect to Hamiltonian uncertainties, the issue in point (1) should remain of significance in the foreseeable future, especially for the most complex (and often, the most interesting!) chemical/physical applications. Mitigating this serious problem is the ability to create shaped laser pulse controls and apply them to a quantum system, followed by a probe of their effects at an unprecedented rate of thousands or more independent trials per minute. This unique capability led to the suggestion of partially, if not totally, sidestepping the design process by performing closed-loop experiments to let the quantum system teach the laser how to achieve its control in the laboratory. Fig. 3 schematically shows this closed loop process, drawing on the following logic:

1. **The molecular view.** Although there may be theoretical uncertainty about the system Hamiltonian, the actual chemical/physical system under study 'knows' its own Hamiltonian precisely! This knowledge would also include any unusual exigencies, perhaps associated with structural or other defects in the particular sample. Furthermore, upon exposure to a control field, the system 'solves' its own Schrödinger equation impeccably accurately and as fast as possible in real time. Considering these points, the aim is to replace the offline arduous digital computer control field design effort with the actual quantum system under study acting as a precise analog computer, solving the true equations of motion.
2. **Control laser technology.** Pulse shaping under full computer control may be carried out using even hundreds of discrete elements in the frequency domain controlling the phase and amplitude structure of the pulse. This technology is readily available and expanding in terms of pulse center frequency flexibility and bandwidth capabilities.
3. **Quantum mechanical objectives.** Many chemical/physical objectives may be simply expressed as the desire to steer a quantum mechanical flux out one clearly defined channel versus another.
4. **Detection of the control action.** Detection of the control outcome can often be carried out by a second laser pulse or any other suitable high duty cycle detection means, such as laser-induced mass spectrometry. The typical circumstance in point (3) implies that little, if any, time-consuming offline data analysis will be necessary beyond that of simple signal averaging.
5. **Fast learning algorithms.** All of the four points above may be slaved together using pattern recognition learning algorithms to identify those control fields which are producing better results, and bias the next sequence of experiments in their favor for further improvement. Although many algorithms may be employed for this purpose, the global search capabilities of genetic algorithms are quite attractive, as they may take full advantage of the high throughput nature of the experiments.

In linking together all of these components into a closed-loop OCE, it is important that no single operational step significantly lags behind any other, for efficiency reasons. Fortunately, this economy is coincident with all the tasks and technologies involved. For achieving control alone, there is no requirement that the actual laser pulse structure be identified on each cycle of the loop. Rather, the laser control 'knob' settings are adequate information for the learning algorithm, as the only criterion is to suitably adjust the knobs to achieve an acceptable value for the chemical/physical objective. The learning algorithm in point (5) operates with a cost functional J , analogous to that used for computational design of fields, but now the cost functional can only depend on those quantities directly observable in the laboratory (i.e., minimally, the current achieved target expectation value). In cases where a physical understanding is sought about the control mechanism, the laser field structure must also be measured. This additional

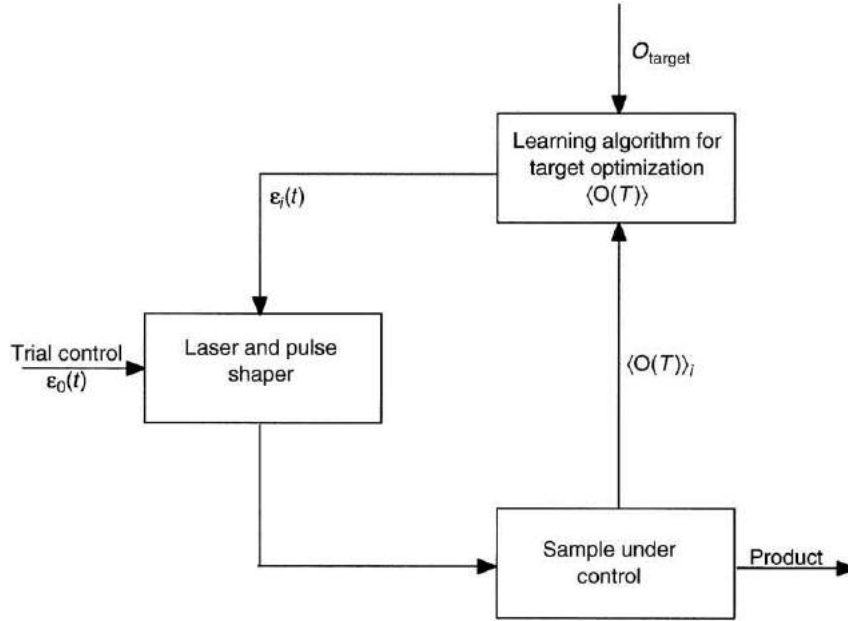


Fig. 3 A schematic of the closed-loop concept for allowing the quantum system to teach the laser how to achieve its control. The actual quantum system under control is in a loop with a laser and pulse shaper all slaved together with a pattern recognition algorithm to guide the excursions around the loop. The success of this concept relies on the ability to perform very large numbers of control experiments in a short period of time. In principle, no knowledge of the system Hamiltonian is required to steer the system to the desired final objective, although a good trial design $\varepsilon_0(t)$ may accelerate the process. The i th cycle around the loop attempts to find a better control field $\varepsilon_i(t)$ such that the system response $\langle O(T) \rangle_i$ for the observable operator O comes closer to the desired value O_{target} .

information is typically also not a burden to attain, as it often is only required for the single best control field at the end of the closed-loop learning experiments.

The OCE process in Fig. 3 and steps (1)–(5) above constitute the laboratory process for achieving optimal quantum control, and exactly the same reasons put forth for OCT motivate the desire for attaining optimal performance: principally, seeking the best control result that can possibly be obtained. Seeking optimal performance also ensures some degree of inherent robustness to field noise, as fleeting observations would not likely survive the signal averaging carried out in the laboratory. Additionally, robustness as a specific criterion may also be included in the laboratory cost functional in item (5). When a detailed physical interpretation of the controlled dynamics is desired, it is essential to remove any extraneous control field features that have little impact on the system manipulations. Without the latter clean-up carried out during the experiments, the final field may be contaminated by structures that have little physical impact. Although quantum dynamics typically occurs on femtosecond or picosecond time-scales, the loop excursions in Fig. 3 need not be carried out on the latter time-scales. Rather, the loop may be traversed as rapidly as convenient, consistent with the capabilities of the apparatus. A new system sample is introduced on each cycle of the loop. This process is referred learning control, to distinguish it from real-time feedback control.

The closed-loop quantum learning control algorithm can be self-starting, requiring no prior control field designs, under favorable circumstances. Virtually all of the current experiments were carried out in this fashion. It remains to be seen if this procedure of ‘going in blind’ will be generally applicable in highly complex situations where the initial state is far from the target state. Learning control can only proceed if at least a minimal signal is observed in the target state. Presently, this requirement has been met by drawing on the overwhelmingly large number of exploratory experiments that can be carried out, even in a brief few minutes, under full computer control. However, in some situations, the performance of intermediate observations along the way to the target goal may be necessary in order to at least partially guide the quantum dynamics towards the ultimate objective. Beneficial use could also be made from prior theoretical laser control designs $\varepsilon_0(t)$ capable of at least yielding a minimal target signal.

The OCE closed-loop procedure was based on the growing number of successful theoretical OCT design calculations, even with all of their foibles, especially including less than perfect Hamiltonians. The OCT and OCE processes are analogous, with their primary distinction involving precisely what appears in their corresponding cost functionals. Theoretical guidance can aid in identifying the appropriate cost functionals and learning algorithms for OCE, especially for addressing the needs associated with attaining a physical understanding about the mechanisms governing controlled quantum dynamics phenomena.

A central issue is the rate that learning control occurs upon excursions around the loop in Fig. 3. A number of factors control this rate of learning, including the nature of the physical system, the choice of objective, the presence of field uncertainties and measurement errors, the number of control variables, and the capabilities of the learning algorithm employed. Achieving control is a matter of discrimination, and some of these factors, whether inherent to the system or introduced by choice, may work against

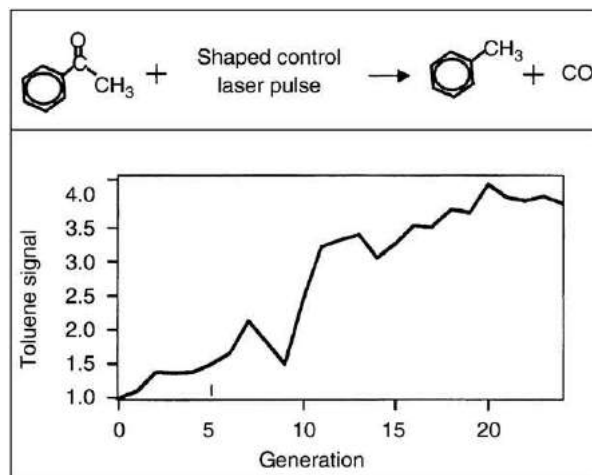


Fig. 4 An illustration of dissociative rearrangement achieved by closed-loop learning control in Fig. 3. The optimally deduced laser pulse broke two bonds in the parent acetophenone molecule and formed a new one to yield the toluene product. The laboratory learning curve is shown for the toluene product signal as a function of the generations in the genetic algorithm guiding the experiments. The fluctuations in the learning curve for optimizing the toluene yield corresponds to the algorithm searching for the optimal yield as the experiments proceed.

attaining good-quality discrimination. At this juncture, little is known quantitatively about the limitations associated with any of these factors.

The latter issues have not prevented the successful performance of a broad variety of quantum control experiments, and at present the ability to carry out massive numbers of closed-loop excursions has overcome any evident difficulties. The number of examples of successful closed-loop quantum control is rapidly growing, with illustrations involving the manipulation of laser dye fluorescence, chemical reactivity, high harmonic generation, semiconductor optical switching, fiber optic pulse transmission, and dynamical discrimination of similar chemical species, amongst others. Perhaps the most interesting cases are those carried out at high field intensities, where prior speculation suggested that such experiments would fail due to even modest laser field noise being amplified by the quantum dynamics. Fortunately, this outcome did not occur, and theoretical studies have indicated that surprising degrees of robustness to field noise may accompany the manipulation of quantum dynamics phenomena. Although the presence of field noise may diminish the beneficial influence of constructive and destructive interferences, evidence shows that field noise does not appear to kill the actual control process, at least in terms of obtaining reasonable values for the control objectives. One illustration of control in the high-field regime is shown in Fig. 4, demonstrating the learning process for the dissociative rearrangement of acetophenone to form toluene. Interestingly, this case involves breaking two bonds, with the formation of a third, indicating that complex dynamics can be managed with suitably tailored laser pulses.

Operating in the strong-field regime, especially for chemical manipulations, also has the important benefit of producing a generic laser tool to control a broad variety of systems by overcoming the long-standing problem of having sufficient bandwidth. For example, the result in Fig. 4 uses a Ti:sapphire laser in the near-infrared regime. Starting with a bandwidth-limited pulse of intensity near $\sim 10^{14}$ W cm $^{-2}$, in this regime, the dynamic power broadening can easily be on the order of ~ 1 eV or more, thereby effectively converting the otherwise discrete spectrum of the molecule into an effective continuum for ready multiphoton matching by the control laser pulse. Operating under these physical conditions is very attractive, as the apparatus is generic, permitting a single shaped-pulse laser source (e.g., a Ti:sapphire system with phase and amplitude modulation) to be utilized with virtually any chemical system where manipulations are desired. Although every desired physical/chemical goal may not be satisfactorily met (i.e., the goals must be physically attainable!), the means are now available to explore large numbers of systems. Operating under closed loop in the strong-field tailored-pulse regime eliminates the prior serious limitation of first finding a physical system to meet the laser capabilities; now, the structure of the laser pulse can be shaped to meet the molecule's characteristics. Strong field operations may be attractive for these reasons, but other broadband laser sources, possibly working at weaker field intensities, might be essential in some applications (e.g., controlling electron dynamics in semiconductors, where material damage is to be avoided). It is anticipated that additional broadband laser sources will become available for these purposes.

The coming years should be a period of rapid expansion, including a thorough exploration of closed-loop quantum control capabilities. As with the use of any laboratory tool, certain applications may be more amenable than others to attaining successful control. The particular physical/chemical questions need to be well posed, and controls need to have sufficient flexibility to meet the objectives. The experiments ahead should be able to reveal the degree to which drawing on optimality in OCE, combined with the performance of massive numbers of experiments can lead to broad-scale successful control of quantum phenomena. One issue of concern is the richness associated with the large numbers of phase and amplitude control knobs that may be adjusted in the laboratory. Some experiments have already operated with hundreds of knobs, while others have restricted their number in a variety of ways, to simplify the search process. Additional technological and algorithmic advances may be required to manage the

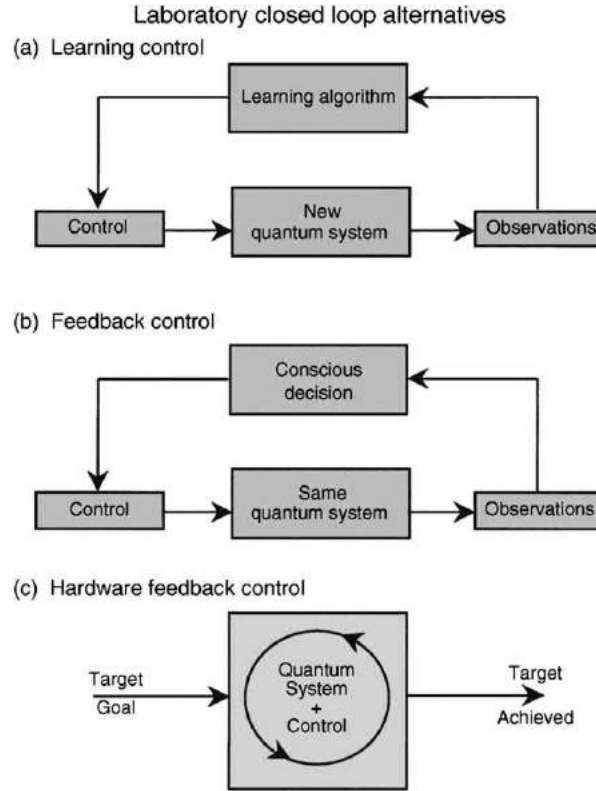


Fig. 5 Three possible closed-loop formulations for attaining laser control over quantum phenomena. (a) is a learning control process where a new system is introduced on each excursion around the loop. (b) utilizes feedback control with the loop being traversed with the same sample, generally calling for an accurate model of the system to adjust the controls on each loop excursion. (c) is, at present, a dream machine where the laser and the sample under control are one unit operating without external algorithmic guidance. The closed-loop learning control procedure in (a) appears to form the most practical means to achieve laser control over quantum phenomena, especially for complex systems.

high-dimensional control space searches. Fortunately, for typical applications, the search does not reduce to seeking a needle in a haystack, as generally there are multiple control solutions, possibly all of very good quality.

As a final comment on OCE, it is useful to appreciate the subtle distinction of the procedure from closed-loop feedback experiments. This distinction is illustrated in Fig. 5, pointing to three types of closed-loop quantum control laboratory experiments. The OCE procedure in Fig. 5(a) produces a generic laser tool capable of controlling a broad variety of systems with an emphasis in the figure placed on the point that each cycle around the loop starts with a new sample for control. The replacement of the sample on each cycle eliminates a number of difficulties, principally including concerns about sample damage, avoidance of operating the loop at the ultrafast speed of the quantum mechanical processes, and elimination of the effect that ‘to observe is to disturb’ in quantum mechanics. Thus, learning control provides a generic practical procedure, regardless of the nature of the quantum system. Fig. 5(b) has the same structure as that of Fig. 5(a), except the loop is now closed around the same single quantum system, which is followed throughout its evolution. All of the issues mentioned above that are circumvented by operating through laboratory learning control in Fig. 5(a) must now be directly faced in the setup of Fig. 5(b). The procedure in Fig. 5(b) will likely only be applicable in special circumstances, at least for the reason that many quantum systems operate on time-scales far too fast for opto-electronics and computers to keep up with. However, there are certain quantum mechanical processes that are sufficiently slow to meet the criteria. Furthermore, a period of free evolutionary quantum dynamics may be permitted to occur between one control pulse and the next, to make the time issue manageable. While the learning control process in Figs. 3 and 5(a) can be performed model-free, the feedback algorithm in Fig. 5(b) generally must operate with a sufficiently reliable system Hamiltonian to carry out fast design corrections to the control field, based on the previous control outcome. These are very severe demands, but they may be met under certain circumstances. One reason for considering closed-loop feedback control in Fig. 5(b) is to explore the basic physical issue of quantum mechanical limitations inherent in the statement ‘to observe is to disturb’. It is suggestive that control over many types of quantum phenomena may fall into the semiclassical regime, lying somewhere between classical engineering behavior and the hard limitations of quantum mechanics. Experiments of the type in Fig. 5(b) will be most interesting for exploring this matter. Finally, Fig. 5(c) introduces a *gedanken* experiment in the sense that the closed-loop process is literally built into the hardware. That is, the laser control and quantum system act as a single functioning unit operating in a stable fashion, so as to automatically steer the system to the desired target. This process may involve engineering the quantum mechanical system prior to control in cases where that freedom exists, as well as engineering the laser components involved.

The meaning of such a device in [Fig. 5\(c\)](#) can be understood by considering an analogy with airplane flight, where the aircraft is constructed to have an inherent degree of aerodynamic stability and will essentially fly (glide) on its own accord when pushed forward. It is an open question whether closing the loop in the hardware can be attained for quantum control, and an exploration of this concept may require additional laser technologies.

Conclusions

This article presented an overview of theoretical concepts and algorithmic considerations associated with the control of quantum phenomena. Theory has played a central role in this area by revealing the fundamental principles underlying quantum control, as well as by providing algorithms for designing controls and guiding experiments to discover successful controls in the laboratory. The challenge of controlling quantum phenomena, in one sense, is an old subject, going back at least 40 years. However, roughly the first 30 of these years would best be described as a period of frustration, due to a lack of full understanding of the principles involved and the nature of the lasers needed to achieve success. Thus, from another perspective, the subject is quite young, with perhaps the most notable development being the introduction of closed-loop laboratory learning procedures. At the time of writing this article, these procedures are just beginning to be explored for their full capabilities. To appreciate the young nature of this subject, it is useful to note that the analogous engineering control disciplines presently occupy many thousands of engineers worldwide, both theoreticians and practitioners, and have done so for many years. Yet, engineering control is far from considered a mature subject. Armed now with the basic concepts and proper laboratory tools, one may anticipate a thorough exploration of control over quantum phenomena, including its many possible applications.

Acknowledgment

Support for this work has come from the National Science Foundation and the Department of Defense.

See also: Coherent Control: Experimental. Overview: Coherence

Further Reading

Brixner, T., Damrauer, N.H., Gerber, G., 2001. Femtosecond quantum control. *Advances in Atomic, Molecular, and Optical Physics* 44, 1–54.
Rabitz, H., de Vivie-Riedle, R., Motzkus, M., Kompa, K., 2000. Whither the future of controlling quantum phenomena? *Science* 288, 824–828.
Rabitz, H., Zhu, W., 2000. Optimal control of molecular motion: Design, implementation and inversion. *Accounts of Chemical Research* 33, 572–578.
Rice, S.A., Zhao, M., 2000. *Optical Control of Molecular Processes*. New York: John Wiley.

Coherent Control: Experimental

RJ Lewis, Temple University, Philadelphia, PA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The experimental control of quantum phenomena in atoms, molecules, and condensed phase materials has been a long-standing challenge in the physical sciences. For example, the idea that a laser beam could be used to selectively cleave a chemical bond has been pursued since the demonstration of the laser in the early 1960s. However, it was not until very recent times that one could realize selective bond cleavage. One key to the rapid acceleration of such experiments is the capability of manipulating the relative phases between two or more paths leading to the same final state. This is done in practice by controlling the phase of two (or two thousand!) laser frequencies coupling the initial state to some desired final state. The present interest in phase-related control stems in part from research focusing on quantum information sciences, controlling chemical reactivity, and developing new optical technologies, such as in biophotonics for imaging cellular materials. Ultimately, one would like to identify coherent control applications having impact similar to linear regime applications like compact disk reading (and writing), integrated circuit fabrication, and photodynamic therapy. Note that in each of these the phase of the electromagnetic field is not an important parameter and only the brightness of the laser at a particular frequency is used. While phase is of negligible importance in any one photon process, for excitations involving two or more photons, phase not only becomes a useful characteristic, it can modulate the yield of a desired process between zero and one hundred percent.

The use of optical phase to manipulate atomic and molecular systems rose to the forefront of laser science in the 1980s. Prior to this there was considerable effort directed toward using the intensity and high spectral resolution features of the laser for various applications. For example, the high spectral resolution character was employed in purifying nuclear isotopes. Highly intense beams were also used in the unsuccessful attempt to pump enough energy into a single vibrational resonance to selectively dissociate a molecule. The current revolution in coherent control has been driven in part by the realization that controlling energy deposition into a single degree of freedom (such as selectively breaking a bond in a polyatomic molecule) can not be accomplished by simply driving a resonance with an ever more intense laser source. In the case of laser-induced photo dissociation, such excitation ultimately results in statistical dissociation throughout the molecule due to vibrational mode coupling. In complex systems, controlling energy deposition into a desired quantum state requires a more sophisticated approach and one scheme involves coherent control.

The two distinct regimes for excitation in coherent control, the time and frequency domain experiments, are formally equivalent. While the experimental approach and level of success for each domain is vastly different, both rely on precisely specifying the phase relations for a series of excitation frequencies. The basic idea behind phase control is simple. One emulates the interference effects that can be easily visualized in a water tank experiment with a wave propagating through two slits. In the coherent control experiment, two or more indistinguishable pathways must be excited that connect one quantum state to another. In either water waves, or coherent control, the probability for observing constructive or destructive interference at a target state depends on the phase difference between the paths connecting the two locations in the water tank, or the initial and final states in the quantum system. In the case of coherent control, the final state may have some technological value, for instance, a desirable chemical reaction product or an intermediate in the implementation of quantum computation.

Frequency domain methods concentrate on interfering two excitation routes in a system by controlling the phase difference between two distinct frequencies coupling an initial state to a final state. Perhaps the first mention of the concept of phase control in the optical regime can be traced to investigations in Russia in the early 1960s, where the interference between a one- and three-photon path was theoretically proposed as a means to modulate a two-photon absorption. This two-path interference idea lay dormant until the reality of controlling the relative phase of two distinct frequencies was achieved in 1990 using gas phase pressure modulation for frequency domain experiments. The idea was theoretically extended to the control of nuclear motion in 1988. The first experimental example of two-color phase control over nuclear motion was demonstrated in 1991. It is interesting that the first demonstrations of coherent control did not, however, originate in the frequency domain approach, but came from time domain investigations in the mid-1980s. This was work that grew directly out of the chemical dynamics community and will be described below.

While frequency domain methods were developed exclusively for the observation of quantum interference, the experimental implementation of time-domain methods was first developed for observing nuclear motion in real time. In the time domain, a superposition of multiple states is prepared in a system using a short duration (~ 50 fs) laser pulse. The superposition state will subsequently evolve according to the phase relationships of the prepared states and the Hamiltonian of the system. At some later time the evolution must be probed using a second, short-time-duration laser pulse. The time-domain methods for coherent control were first proposed in 1985 and were based on wavepacket propagation methods developed in the 1970s. The first experiments were reported in the mid-1980s.

To determine whether the time or frequency domain implementation is more applicable for a particular control context, one should compare the relevant energy level spacing of the quantum system to the bandwidth of the excitation laser. For instance, control of atomic systems is more suited to frequency domain methods because the characteristic energy level spacing in an atom

(several electron volts) is large compared to the bandwidth of femtosecond laser pulses (millielectron volts). In this case, preparation of a superposition of two or more states is impractical with a single laser pulse at the present time, dictating that the control scheme must employ the initial state and a single excited eigenstate. In practice, quasi-CW nanosecond duration laser pulses are employed for such experiments because current ultrafast (fs duration) laser sources cannot typically prepare a superposition of electronic states. One instructive exception involves the excitation of Rydberg states in atoms. Here the electronic energy level spacing can be small enough that the bandwidth of a femtosecond excitation laser may span several electronic states and thus can create a superposition state. One should also note that the next generation of laser sources in the attosecond regime may permit the preparation of a wavepacket from low-lying electronic states in an atom. In the case of molecules, avoiding a superposition state is nearly impossible. Such experiments employ pico- to femtosecond duration laser pulses, having a bandwidth sufficient to excite many vibrational levels, either in the ground or excited electronic state manifold of a molecule. To complete the time-dependent measurement, the propagation of the wavepacket to the final state of interest must be observed (probed) using a second ultrafast laser pulse that can produce fluorescence, ionization, or stimulated emission in a time-resolved manner. The multiple paths that link the initial and final state via the pump and the probe pulses are the key to the equivalence of the time and frequency domain methods.

There is another useful way to classify experiments that have been performed recently in coherent control, that is into either open-loop or closed-loop techniques. Open-loop signifies that a calculation may be performed to specify the required pulse shape for control before the experiment is attempted. All of the frequency-based, and most of the time-based experiments fall into this category. Closed-loop, on the other hand, signifies that the results from experimental measurements must be used to assist in determining the correct parameters for the next experiment, this pioneering approach was suggested in 1992. It is interesting that in the context of control experiments performed to date, closed-loop experiments represent the only route to determining the optimal laser pulse shape for controlling even moderately complex systems. In this approach, no input from theoretical models is required for coherent control of complex systems. This is unlike typical control engineering environments, where closed-loop signifies the use of experiments to refine parameters used in a theoretical model.

There is one additional distinction that is of value for understanding the state of coherent control at the present time. Experiments may be classified into systems having Hamiltonians that allow calculation of frequencies required for the desired interference patterns (thus enabling open-loop experiments) and those having ill-defined Hamiltonians (requiring closed-loop experiments). The open-loop experiments have demonstrated the utility of various control schemes, but are not amenable to systems having even moderate complexity. The closed-loop experiments are capable of controlling systems of moderate complexity but are not amenable to precalculation of the required control fields. Systems requiring closed-loop methods include propagation of short pulses in an optical fiber, control of high harmonic generation for soft X-ray generation, control of chemical reactions and control of biological systems. In the experimental regime, the value of learning control for dealing with complex systems has been well documented.

The remainder of this article represents an overview of the experimental implementation of coherent control. The earliest phase control experiments involved time-dependent probing of molecular wavepackets, and were followed by two-path interference in an atomic system. The most recent, and general, implementation of control involves tailoring a time-dependent electromagnetic field for a desired objective using closed-loop techniques. Since the use of tailored laser pulses appears to be rather general, a more complete description of the experiment is provided.

Time-Dependent Methods

The first experiments demonstrating the possibility of coherent control were performed to detect nuclear motion in molecules in real time. These investigations were the experimental realization of the pump-dump or pulse-timing control methods as originally described in 1985. The experimental applications have expanded to numerous systems including diatomic and triatomic molecules, small clusters, and even biological molecules. For such experiments, a vibrational coherence is prepared by exciting a superposition of states. These measurements implicitly used phase control both in the generation of the ultrashort pulses and in the coherent probing of the superposition state by varying the time delay between the pump and probe pulses, in effect modulating the outcome of a quantum transition. Since the earliest optical experiment in the mid-1980s, there have been many hundreds of such experiments reported in the literature. The motivation for the very first experiments was not explicitly phase control, but rather the observation of nuclear motion of molecules in real time. In such later measurements, oscillations in the superposition states were observable for many cycles, up to many tens of picoseconds in favorable cases, i.e., diatomic molecules in the gas phase. The loss of signal in the coherent oscillations is ultimately due to passage of the superposition state in to surrounding bath states. This may be due to collision with another molecule, dissociation of the system or redistribution of the excitation energy into other modes of the system not originally excited in the superposition. It is notable that coherence can be maintained in solution phase systems for tens of picoseconds and for the case of resonances having weak coupling to other modes, coherence can even be modulated in large biological systems on the picosecond time scale.

Decoherence represents loss of phase information about a system. In this context, phase information represents our detailed knowledge of the electronic and nuclear coordinates of the system. In the case of vibrational coherence, decoherence may be represented by intramolecular vibrational energy transfer to other modes of the molecule. This can be thought of as the propensity of vibrational modes of a molecule to couple together in a manner that randomizes deposited energy throughout the molecule.

Decoherence in a condensed phase system includes transfer of energy to the solvent modes surrounding the system of interest. For an electronically excited molecule or atom, decoherence can involve spontaneous emission of radiation (fluorescence), or dissociation in the case of molecules. Certain excited states may have high probability for fluorescence and would represent a significant decoherence pathway and an obstacle for coherent control. These states may be called 'lossy' and are often used in optical pumping schemes as well as for stimulated emission pumping. An interesting question arises as to whether one can employ such states in a pathway leading to a desired final state, or must such lossy states be avoided at all costs in coherent control. This question has been answered to some degree by the method of rapid adiabatic passage. In this experiment, an initial state is coupled to a final state by way of a lossy state. The coupling is a coherent process and the preparation involves dressed states that are eigenstates of the system. In this coherent superposition the lossy state is employed, but no population is allowed to build up and thus decoherence is circumvented.

Two-Path Interference Methods

Interference methods form perhaps the most intuitive example of coherent control. The first experimental demonstration was reported in 1990 and involved the modulation of ionization probability in Hg by interfering a one-photon and three-photon excitation pathway to an excited electronic state (followed by two-photon ionization). The long delay between the introduction of the two-path interference concept in 1960 and the demonstration in 1990 was the difficulty in controlling the relative phase of the one- and three-photon paths. The solution involved generating the third harmonic of the fundamental in a nonlinear crystal, and copropagating the two beams through a low-density gas. Changing the pressure of the gas changes the optical phase retardance at different rates for different frequencies, thus allowing relative phase control between two colors. The ionization yield of Hg was clearly modulated as a function of pressure. Such schemes have been extended to diatomic molecular systems and the concept of determining molecular phase has been investigated. While changing the phase of more than two frequencies has not been demonstrated using pressure tuning (because of experimental difficulties), a more flexible method has been developed to alter relative phase of up to 1000 frequency bands, as will be described next.

Closed-Loop Control Methods

Perhaps the most exciting new aspect of control in recent years concerns the prospect for performing closed-loop experiments. In this approach a more complex level of control in matter is achieved in comparison to the two-path interference and pump-probe control methods. In the closed-loop method, an arbitrarily complex time-dependent electromagnetic field is prepared that may have multiple subpulses and up to thousands of interfering frequency bands. The realization of closed-loop control over complex processes relies on the confluence of three technologies: (i) production of large bandwidth ultrafast laser pulses; (ii) spatial modulation methods for manipulating the relative phases and amplitudes of component frequencies of the ultrafast laser pulse; and (iii) closed-loop learning control methods for sorting through the vast number of pulse shapes available as potential control fields. In the closed-loop method, the result of a laser molecule interaction is used to tailor an optimized pulse.

In almost all of the experimental systems used for closed-loop control today, Kerr lens mode-locking in the Ti:sapphire crystal is used to phase lock the wide bandwidth of available frequencies (750–950 nm) to create the ultrafast pulse. In this phenomenon, the intensity variation in the short laser pulse creates a transient lens in the Ti:sapphire lasing medium that discriminates between free lasing and mode-locked pulses. As a result, all of the population inversion available in the lasing medium may be coerced into enhancing the ultrashort traveling wave in the cavity in this way. The wide bandwidth available may then be amplified and tailored into a shaped electromagnetic field. Spatial light modulation is one such method for modulating the relative phases and amplitudes of the component frequencies in the ultrafast laser pulse to tailor the shaped laser pulse. Pulse shaping first involves dispersing the phase-locked frequencies on an optical grating, the dispersed radiation is then collimated using a cylindrical lens and the individual frequencies are focused to a line forming the Fourier plane. At the Fourier plane, the time-dependent laser pulse is transformed into a series of phase-locked continuous wave (CW) laser frequencies, and thus Fourier transformed from time to frequency space. The relative phases and amplitudes may be modulated in the Fourier plane using an array of liquid crystal pixels. After altering the spectral phase and amplitude profile, the frequencies are recombined on a second grating to form the shaped laser pulse. With one degree of phase control and 10 pixels there are an astronomic number (360^{10}) of pulse shapes that may be generated in this manner. (At the present time, pulse shapers have up to 2000 independent elements.) Evolutionary algorithms are employed to manage the available phase space. Realizing that the pulse shape is nothing more than the summation of a series of sine waves, each having an associated phase and amplitude, we find that a certain pulse shape can be represented by a genome consisting of an array of these frequency-dependent phases and amplitudes. This immediately suggests that the methods of evolutionary search strategies may be useful for determining the optimal pulse shape for a desired photoexcitation process.

The method of closed-loop optimal control for experiments was first proposed in 1992 and the first experiments were reported in 1997 regarding optimization of the efficiency of a laser dye molecule in solution. Since that time a number of experiments have been performed in both the weak and strong field regime. In the weak field regime, the intensity of the laser does not alter the field free structure of the excited states of the system under investigation. In the strong field, the field free states of the system are altered by the electric field of the laser. Whether one is in the weak or strong field regime depends on parameters including the

characteristic level spacing of the states of the system to be controlled and the intensity of the laser coupling into those states. Examples of weak field closed-loop control include optimization of laser dye efficiency, compression of an ultrashort laser pulse, dissociation of alkali clusters, optimization of coherent anti-Stokes Raman excitation, and optical pulse propagation in fibers. In the strong field, the laser field is used to manipulate the excited states of the molecule as well as induce population transfer among these states. In this sense, the laser is both creating and exciting resonances, in effect allowing universal excitation with a limited frequency bandwidth of 750–850 nm. For chemical applications, it is worth noting that most molecules do not absorb in this wavelength region in the weak field. Strong field control has been used for controlling bond dissociation, chemical rearrangement, photonic reagents, X-ray generation, molecular centrifuges and the manipulation of mass spectral fragmentation intensities. In the latter case, a new sensing technology has emerged wherein each individual pulse shape represents an independent sensor for a molecule. Given the fact that thousands of pulse shapes can be tested per minute, this represents a new paradigm for molecular analysis.

At the time of this writing, the closed-loop method for adaptively tailoring control fields, has far outstripped our theoretical understanding of the process (particularly in the strong field regime). Adaptive control is presently an active field of investigation, encompassing the fields of physics, engineering, optics, and chemistry. In the coming years, there will be as much work done on the mechanism of control as in the application of the methods. Due to small energy level spacings and complexity of the systems, we anticipate the majority of applications to be in the chemical sciences.

See also: Coherent Control: Theory. Overview: Coherence

Further Reading

- Brumer, P.W., Shapiro, M., 2003. *Principles of the Quantum Control of Molecular Processes*. New York: Wiley Interscience.
- Chen, C., Elliott, D.S., 1990. Measurement of optical phase variations using interfering multiphoton ionization processes. *Physics Review Letters* 65, 1737.
- Levis, R.J., Rabitz, H., 2002. Closing the loop on bond selective chemistry using tailored strong field laser pulses. *Journal of Physical Chemistry A* 106, 6427–6444.
- Levis, R.J., Menkir, G.M., Rabitz, H., 2001. Selective bond dissociation and rearrangement with optimally tailored, strong-field laser pulses. *Science* 292 (5517), 709–713.
- Rice, S.A., Zhao, M., 2000. *Optical Control of Molecular Dynamics*. New York: Wiley Interscience.
- Stolow, A., 2003. Femtosecond time-resolved photoelectron spectroscopy of polyatomic molecules. *Annual Review of Physical Chemistry* 54, 89–119.
- Zewail, A.H., Casati, G., Rice, S.A., *et al.*, 1997. Femtochemistry: Chemical reaction dynamics and their control. *Chemical Reactions and Their Control on the Femtosecond Time Scale*, Solvay Conference on Chemistry 101, 3–45.

Optics of the Human Eye

David A Atchison, Queensland University of Technology, Brisbane, QLD, Australia

© 2018 Elsevier Ltd. All rights reserved.

Basic Optical Structure

The human eye is shown in Fig. 1. The outer layer consists of the cornea at the front and the sclera at the back. The transparent cornea is approximately spherical with a radius of curvature of about 8 mm. The white, opaque sclera is approximately spherical with a radius of curvature of approximately 12 mm. The middle layer of the eye is the uveal tract, which consists of the iris anteriorly, the choroid posteriorly (not shown), and the intermediate ciliary body. Inside the eye, about 3 mm from the cornea, is the transparent lens. The anterior chamber between cornea and iris, and the posterior chamber between iris, ciliary body and lens, contain fluid called the aqueous humor. The vitreous chamber between the lens and the back of the eye contains a transparent, colorless, gelatinous mass called the vitreous humor.

The inner layer of the eye is the light-sensitive tissue called the retina, which is connected to the brain via the optic nerve. Light passes through the inner retinal layers to reach the photoreceptors, where light is absorbed to initiate the electrical and chemical processes involved in vision. Photoreceptors are of two types: rods and cones. Cones are associated with vision in bright (photopic) lighting conditions; there are three types that absorb at different wavelength ranges and initiate color vision. Rods are associated with vision in low (scotopic) light levels, and there is an intermediate (mesopic) light level range where both cones and rods operate. The cones predominate in the fovea, which is 1.5 mm (5 degree field) wide. The fovea is free of rods in its central 1 degree field. When the eye fixates, or “looks at,” a small object of interest, the retinal image forms at the foveal center. The inner retinal layers contain ganglion cells, whose axons leave the retina at the optic disk. This is functionally blind because there are no cones or rods here. It is approximately 5 degree horizontally and 7 degree vertically, and its center is approximately 15 degree nasally and 1.5 degree upwards from the foveal center.

Light enters the eye through the cornea, and is refracted by the cornea and lens to focus at the retina. The cornea has the greater power, but the lens power in young eyes can change to allow the eye to focus at different distances – the process of accommodation. The hole in the iris is the aperture stop. When we look at the aperture stop, we see it as the pupil, which is closer to the front of the eye and slightly bigger than the aperture stop. The aperture stop controls how much light reaches the retina and the retinal image quality.

In good quality man-made optical systems, the optical axis is the line that joins the centers of curvatures of the refracting surfaces. However, the eye is not rotationally symmetric; surfaces and the aperture stop are tilted and decentered relative to each other. The optical axis can be defined as the line of best fit through the centers of curvature, which in practice is determined by the best alignment of the images formed by reflection from the surfaces. To complicate matters, the fovea center is about 5 degree from the optical axis as measured at the posterior nodal point of the eye. This lack of symmetry has led to a number of axes being required for the eye. For example, the line-of-sight is the line joining an object of interest and the fovea, and which passes through the center of the aperture stop (see Section Axes and Angles).

Ignoring the lack of symmetry, eyes can be treated like many other optical systems with cardinal points and with an equivalent power. Of more importance than equivalent power in a clinical context is refractive error, which is the error in the equivalent power due to an imbalance between equivalent power and eye length.

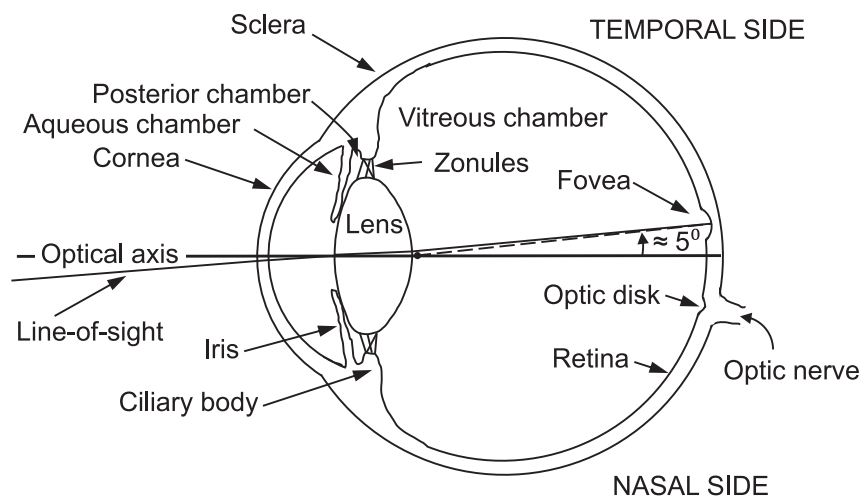


Fig. 1 Horizontal section of a right eye as seen from above.

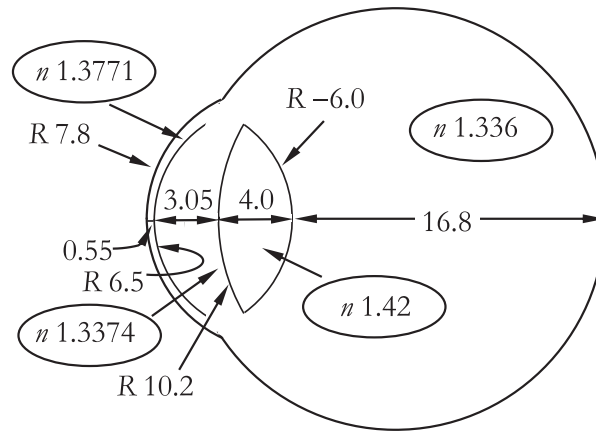


Fig. 2 The Le Grand four-refracting surface schematic eye (Atchison and Smith, 2000), showing dimensions in millimeters and refractive indices. R – surface radius of curvature, n – refractive index.

Dimensions of the eye vary greatly between individuals, and alter with age and accommodation. Average values of populations have been used to construct model (schematic) eyes. These can be made of different complexities, with different number of surfaces. The “paraxial” schematic eyes contain centered, spherical surfaces. These are useful for basic properties, such as component powers, positions of cardinal points, and image size and illuminance calculations, but not for subtle optical effects such as higher-order aberrations. One schematic eye is shown in Fig. 2 as representative of dimensions in relaxed (non-accommodating) adult eyes. The equivalent power of this eye is approximately 60 m^{-1} . When dealing with powers, the term diopter with symbol “D” is usually used instead of inverse meters, a practice followed in the rest of this article.

The range of angles from the fovea at which perception is possible is referred to as the field-of-vision. Its limits are determined by anatomical, optical, and sensory factors. Field limits are about 100 degree temporally, 90 degree inferiorly, and 40 degree superiorly and nasally. Rotating the eye to remove the upper brow and nose restrictions increases the superior field to about 80 degree and the nasal field to about 60 degree; the latter is now limited by extent of the temporal sensitive retina.

Humans have two eyes which are separated by approximately 60–70 mm in adults. Binocular vision provides slightly improved contrast sensitivity and vision acuity over monocular vision. The total field-of-view is about 210 degree with a 120 degree binocular overlap. The slightly different sizes and/or shapes of the two retinal images of a close object allow three-dimensional perception known as stereopsis. Presenting different images to the two eyes is exploited in a number of circumstances such as enabling near vision at a range of distances where this would otherwise not be possible by putting lenses of different power in front of the two eyes, and having aspects of the object which can only be seen by one eye through filters placed in front of the eyes (e.g., 3-D effects in cinemas). When the eyes accommodate to focus for a near object, they rotate inwards (convergence) to fixate the object.

Cornea

The cornea has about 40 D of power, which about two-thirds of the power of the unaccommodated eye. It contains a tear film (about 4–7 μm thick) at the front, an epithelium, Bowman’s membrane, the stroma, Descemet’s membrane and the endothelium. The tear film has oily, aqueous and mucous layers, but they are mixed to variable extents during the blinking cycle. Because the tear film is thin and conforms to the shape of the underlying cornea, it is sometimes treated as being optically inert; however it is important in compensating for the roughness of the outer epithelial cells and there are circumstances in which it cannot be ignored optically such as fitting of rigid contact lenses. The stroma has about 90% of the corneal thickness. It contains about 200–250 layers (lamellae) of cylindrical fibrils. The fibrils within a lamella lie parallel to each other and fibrils within a lamella are inclined at large angles to fibrils in adjacent lamellae. While the different layers of the cornea have different refractive indices, the cornea is usually assigned a constant value which of about 1.376.

Estimates of the anterior and posterior surface radii of curvature of the cornea are about 7.8 and 6.3 mm (Atchison and Smith, 2000), respectively, but there are wide ranges. On average males have flatter corneas than females. Corresponding powers for these radii of curvature are approximately 48 and -6 D ; the much smaller power for the posterior cornea is because of the much lower refractive index (0.04) between aqueous and cornea than between cornea and air (0.376). The two radii of curvature are highly correlated (Dunne *et al.*, 1992; Lowe and Clark, 1973; Patel *et al.*, 1993). Some instruments determine a corneal power based on the anterior radius of curvature and an equivalent index such as 1.3375.

The surfaces are not perfectly spherical; they have both toricity and asphericity. Toricity means that the radius of curvature varies with meridian and produces the optical defect of astigmatism (see Section Refractive anomalies/Astigmatism). Most cornea surfaces are also aspheric and are often represented as conics in two dimensions or as conicoids in three dimensions. One way of

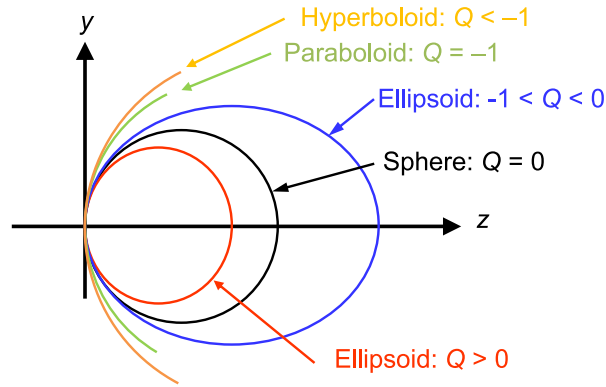


Fig. 3 The Y-Z section of conicoids, showing the influence of asphericity on conicoid shape. All surfaces have the same vertex radius of curvature.

expressing a conicoid is

$$z = \frac{c(x^2 + y^2)}{1 + \sqrt{1 - (1 + Q)(x^2 + y^2)}}$$

where the coordinates at a point on the surface are (x, y, z) with the z -axis corresponding to the optical axis, c is the vertex curvature (the inverse of the vertex radius of curvature), and Q is the surface asphericity giving different types of conicoids (**Fig. 3**):

- $Q < -1$ hyperboloid
- $Q = -1$ paraboloid
- $-1 < Q < 0$ prolate ellipsoid (z -axis is major axis)
- $Q = 0$ sphere
- $0 < Q$ ellipsoid with the major axis in the xy plane

Other forms of representing corneal shape include bi-conicoids, in which there are separate x -direction and y -direction vertex curvatures.

Most anterior corneal surfaces can be describe as prolate ellipsoids, i.e., they flatten away from the vertex, with the mean asphericity being about -0.20 ([Atchison and Smith, 2000](#)). One consequence of this is that the aberration known as spherical aberration is reduced from that which would occur for spherical corneas. There are fewer estimates of asphericity of the posterior corneal surface than of the anterior corneal surface. Mean estimates of the posterior cornea are negative. These are less likely to be accurate than the anterior corneal estimates because of the issue of imaging the posterior cornea through the anterior cornea.

Corneal thickness is about 0.5 mm near the center, and increases into the periphery because the posterior surface is more curved than the anterior surface.

Aperture Stop and Pupils

The iris forms the aperture stop of the eye. Iris color varies considerably, and this depends upon the amount of pigmentation. The stop position varies between individuals, and with age and accommodation. In young adults it is about 3.5 mm behind the anterior cornea. The aperture size is determined by the balance between the sphincter pupillae and dilator pupillae muscles. The former is a ring around the inner margin of the iris, and its contraction causes the stop to constrict. The latter consists of cells that extend radially from the sphincter, and its contraction causes the stop to dilate.

The entrance pupil is the image of the aperture stop formed in object space by the cornea, and the exit pupil is the image of the aperture stop formed by the lens (**Fig. 4**). Paraxial ray tracing through model eyes, in which it is assumed that the aperture stop lies in the anterior vertex plane of the lens, gives approximate positions and sizes of pupils. While varying between different eyes and with accommodation, the entrance pupil is approximately 13% bigger and 0.5 mm closer to the cornea than the aperture stop, while the exit pupil is about 3% bigger and 0.1 mm behind the aperture stop.

Because the entrance pupil is what we see when we look into the eye, the aperture stop is usually referred to loosely as the pupil. Compared with the entrance pupil, the exit pupil has little significance, and for the rest of this article the term “pupil” will refer to the entrance pupil.

Illumination level is the most important factor affecting pupil size, and pupil diameter may vary between 2 mm at high illuminations to 8 mm in the dark. [Watson and Yellott \(2012\)](#) provided a comprehensive description of the dependence of pupil size on illumination and how this is affected by factors of age, size of adapting field, and whether one of both eyes are adapted. The influence of age on pupil size will be discussed later. Other factors that are important are accommodation and convergence, which usually occur together, and state of arousal. Sometimes effects of these are transient, for example, accommodating from distance to a near target will produce constriction that is usually not sustained.

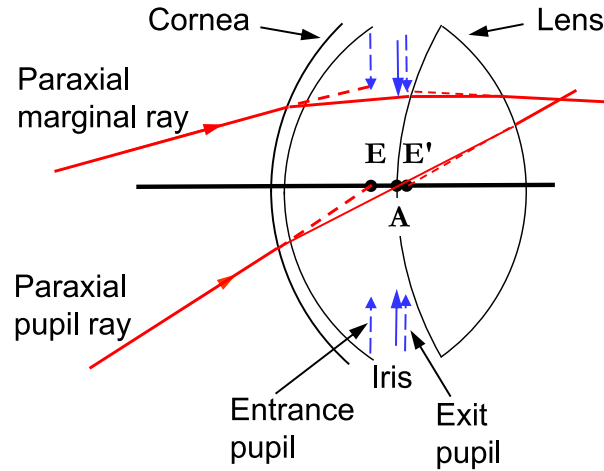


Fig. 4 The aperture stop, the entrance pupil and the exit pupil of the eye. The paraxial pupil ray is directed toward the center of the entrance pupil, by refraction reaches the center of the aperture stop, and is refracted by the lens so that it appears to come from the center of the exit pupil. In a similar fashion, the marginal pupil ray shows the top edges of the aperture stop and the pupils.

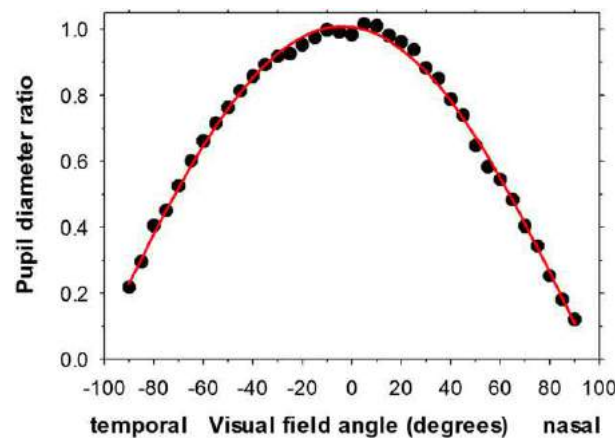


Fig. 5 Pupil diameter ratio as a function of horizontal visual field angle θ for one subject. The data are fitted by $1.009 \cos \left[\frac{(\theta+3.936)}{1.121} \right]$. Unpublished data from the study of Mathur *et al.* (2013).

Drugs can affect pupil size and some are intended for this purpose. Mydriatics cause pupil dilation by either stimulating the sympathetic division of the autonomic nervous system (sympathomimetics) or by blocking its parasympathetic division (parasympatholytics). Miotics cause pupil constriction by either stimulating the parasympathetic division (parasympathomimetics) or by blocking the sympathetic division (sympatholytics).

When the pupil is viewed along the line-of-sight it appears circular in most eyes. When the eye is viewed eccentrically along the horizontal visual field meridian, it becomes elliptical. This is described by a cosine function, flatter by about 12% than the cosine of the viewing angle and decentered by a few degrees into the temporal visual field (Mathur *et al.*, 2013) (Fig. 5).

The pupils of most eyes are decentered relative to the center of the limbus, the line demarking the clear cornea from the surrounding sclera. This decentration is about 0.5 mm nasally, and usually becomes more nasal as the pupil constricts. Mean pupil decentration shift across a large range of illuminances is about 0.12 mm, but with individual shifts of up to 0.5 mm (e.g., Mathur *et al.*, 2014).

The pupil size and decentration, particularly the former, have considerable effects on vision. Pupil diameter affects retinal light level and subsequently light adaptation. However, this is not a large effect considering the pupil area can change by a factor of only 16, which is much smaller than the million times range over which the eye operates. The diameter affects the depth-of-focus, which is the range of distances over which the eye cannot detect any change in focus, with the depth-of-focus usually narrowing as pupil size increases. Retinal image quality and visual performance are influenced by pupil size; for large pupil diameters optical defects of the eye known as aberrations cause deterioration in retinal image quality, while for small pupil diameters diffraction limits retinal image quality. Pupil decentration interacts with diameter to affect image quality.

Lens

The lens is a mass of cellular tissue contained within an elastic capsule. An epithelial cell layer underneath the capsule extends from the anterior vertex to the equator. Throughout life, new epithelial cells form near the equator and elongate forwards and backwards as fibers between the epithelium and capsule. These join other fibers at junctions called sutures.

The capsule is attached to the ciliary body by suspensory ligaments called zonules. Contraction of the ciliary body's muscle causes the ciliary body to move forwards and inwards. This reduces tension on the zonules, which in turn reduce tension on the lens and allow the lens to take up a more curved form under the influence of the capsule. Lens power increases because the surfaces increase in power and because of changes of refractive index distribution within the lens. This allows the eye to change focus from distant to close objects. This process is reversible. The ability of the eye to focus on objects at different distances is known as accommodation. This is usually measured as a change in inverse of distances from the eye, rather than change in lens power. As an example if the eye changes its focus from an object 1 m away to 25 cm away, the change in accommodation is $1/0.25 - 1/1 = 3.0$ D. The magnitude of possible accommodation is called the amplitude of accommodation. Its peak of about 10 D occurs in the second decade of life, with corresponding "unaccommodated" and "accommodated" equivalent lens powers of about 19 and 30 D, respectively. From the peak, amplitude declines steadily to become zero in about the mid-50s. This is described further in the Section Changes with Age/Accommodation and Presbyopia.

The biometric parameters of the lens vary considerably between people, and with age and accommodation. Some values will be given here which are reasonable estimates for young adults in the unaccommodated state. The radii of curvature for the anterior and posterior surfaces are about 10 mm and -6 mm, respectively, and the central thickness of the lens is about 3.6 mm. The lens diameter is 9–10 mm. During accommodation the radii of curvature decrease, more so for the front surface than for the back surface. This is combined with increase in the axial thickness, most of which occurs in the nucleus of the lens. Both the anterior and vitreous chambers are reduced in depth, and the diameter decreases.

One of the most intriguing things about the lens is its gradient refractive index. There is a plateau of high index (about 1.41) in the central (nucleus) region, but the index reduces toward the periphery where it is about 1.38. This index variation produces continuous refraction of rays, so that there are both surface and internal contributions to the power of the lens. Many model eyes have a constant refractive index rather than a gradient index, and to adequately account for the loss of gradient index power this "equivalent" index must be made higher (e.g., 1.42) than the maximum index of the lens. Gullstrand (1924) referred to the contribution of changes in gradient index during accommodation to changes in power as the "intracapsular mechanism of accommodation."

Axes and Angles

Several axes are required to fully describe the optical properties of the eye, because of a lack of symmetry and the departure of the fovea from a best fit optical axis. The validity of some axes is dependent on idealized eye properties. As well as the axes, names are given to the angles between them. A full treatment of these axes is outside the scope of this article (see Atchison (2017a)).

The most important axes are the optical axis, the line-of-sight, the visual axis and the pupillary axis (Fig. 6). The optical axis does not truly exist, but it can be considered to be the line of best fit through the centers of curvature of the intraocular surfaces. The line-of-sight is the line joining the fixation point T and the foveal center T' through the center of the pupil E . This is usually on

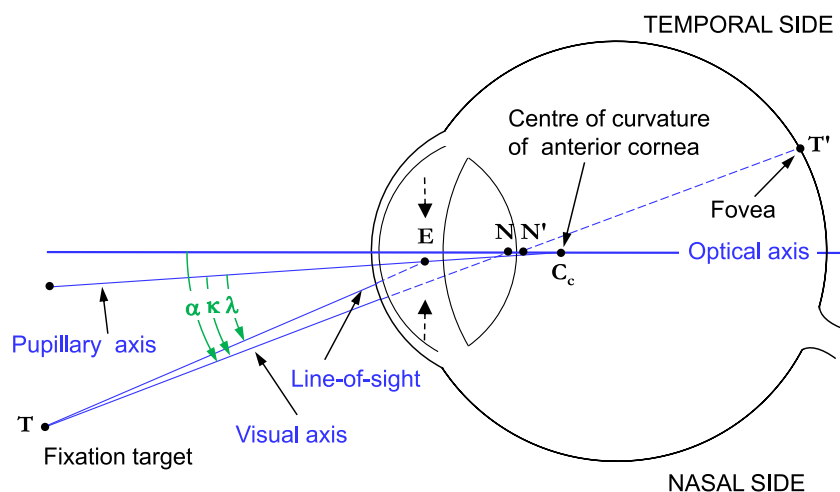


Fig. 6 Some axes and angles of the eye. The object has been shown very close to the eye and the angles are exaggerated for clarity. Adapted from Atchison, D.A., 2017a. Axes and angles of the eye. In: Artal, P. (Ed.), *Handbook of Visual Optics*. vol. I. Boca Raton, FL: Francis & Taylor Group (Chapter 17).

the nasal side of the optical axis in object space. Because the pupil center moves with accommodation and luminance, this quantity is not fixed. The visual axis is the line joining the fixation point and the fovea center through the nodal points N and N' ; it is usually close to the line-of-sight. The pupillary axis is the line passing through the center of the entrance pupil and which is normal to the anterior corneal surface. Usually this lies between the optical axis and the line-of-sight.

The angle between the optical axis and the visual axis is called alpha (or α) and is approximately 5 degree horizontally and 2 degree vertically (visual axis nasal to and above optical axis in object space). The angle between the pupillary axis and line-of-sight is called lambda (or λ) and the angle between the pupillary axis and the visual axis is called kappa (or κ). The angles are of some diagnostic significance, particularly λ which is used to diagnose eccentric fixation and heterotropia (turned eye).

Schematic Eyes

Schematic eyes are models that show optical related biometry of eyes. These have a variety of uses, such as calculating retinal image size, estimating retinal light levels, determining how refractive errors arising from variations in eye dimensions, calculating appropriate powers of intraocular lenses in cataract surgery, estimating aberrations and retinal image quality, and designing ophthalmic instrument and ophthalmic corrections.

Schematic eyes are available in a range of complexity. The paraxial schematic eyes have spherical refractive surfaces that are centered on a common optical axis. Some are not anatomical accurate, for example, the reduced eyes have only one refractive surface. These schematic eyes can describe optics of real eyes accurately only within the paraxial region for which rays are close to the optical axis and subtend small angles to it.

Finite, or wide angle, schematic eyes provide better predictions of aberrations and retinal image quality than the paraxial schematic eyes. They can have various features, such as aspheric surfaces, gradient refractive index, surface tilts and misalignments, media whose refractive indices vary with wavelength, and curved retinas. These eyes require more elaborate calculations than paraxial schematic eyes. Like the paraxial eyes, they are usually based on average values of population values, but can be customized to more accurately describe individual eyes.

An example of a paraxial schematic eye is the Le Grand full theoretical eye (Fig. 7). This contains four-refracting surfaces. It is available in both relaxed and accommodated (7.1 D) forms, but only the former is shown here. It has served as a basis for several paraxial and finite schematic eyes. The figure shows the three pairs of cardinal points. Light leaving the anterior focal point F passes into the eye and is imaged toward infinity. Light passing into the eye from infinity and parallel to the optical axis is imaged at the posterior focal point F' . For light traveling into the eye, the object position can be considered relative to the anterior principal point P and refraction can be treated as if it occurred at the posterior principal point P' . The reverse is true for light traveling from the retina out of the eye. Light traveling into the eye toward the anterior nodal point N appears to pass through the posterior nodal point N' on the image side without the angle of inclination changing.

The constant lens index can be placed by a continuously varying index $N(r)$ described by

$$N(r) = c_0 + c_1 r^2 + c_2 r^4 + c_3 r^6$$

or

$$N(r) = c_0 + c_1 r^p$$

where r is the relative distance from the center of the lens to its edge, and c_0 , c_1 , c_2 , c_3 , and p are coefficients. Different levels of sophistication of such modeling have been proposed (Atchison, 2017b).

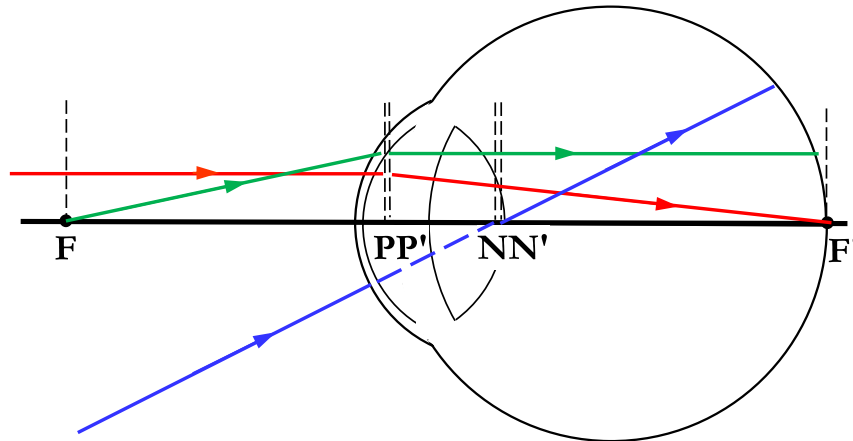


Fig. 7 The unaccommodated version of the Le Grand four-refracting surface schematic eye, showing the cardinal points.

The change in refractive index with wavelength, or chromatic dispersion, of media has been incorporated into some schematic eyes using equations such as Cauchy dispersion equation

$$n(\lambda) = A + B/\lambda^2 + C/\lambda^4 + D/\lambda^6 \dots$$

and the Cornu dispersion equation

$$n(\lambda) = n_\infty + K/(\lambda - \lambda_0)$$

Here λ is wavelength and A, B, C, D, n_∞, K , and λ_0 are constants for a particular medium.

Most of the well-known schematic eyes are representative of average values in particular populations at the times and places the eyes were devised. Many variations of the common schematic eyes have been proposed to account for changes with age in children and adults, with accommodation and with particular ocular conditions. With increased ocular biometry available, personalized eye models are being developed.

When deciding upon which schematic eye might be employed for one of the applications described in the first paragraph of this section, a good principle is to use the simplest model that is adequate, even if this is anatomically inaccurate.

Light and the Eye

Many regions of the electromagnetic spectrum are important to the eye because they may cause damage to various parts of the eye, but the main region in which we are interested is that producing a visual response and to which we refer as light. The wavelength range of light is approximately 390 to 800 nm, for which the spectral colors we see are violet (~390–450 nm), blue (~450–500 nm), green (~500–560 nm), yellow (~560–600 nm), orange (~600–620 nm), and red (~620–800 nm). The eye response within this range depends upon the luminance. Fig. 8 shows the relative luminous efficiencies under photopic and scotopic conditions, with respective peak sensitivities at 555 and 507 nm. The change in peak between these two values, that occurs under the intermediate mesopic conditions of approximately 3–0.03 cd m⁻², is referred to as the Purkinje shift.

Radiometric quantities (radiant flux, radiant intensity, radiance and irradiance) can be converted into the corresponding photometric quantities (luminous flux, luminous intensity, luminance and illuminance) by simple equations that weight the radiometric values at any wavelength by the appropriate relative luminous efficiency.

Passage of Light Into Eye

Not all of light entering the eye reaches the retina. Some is lost by specular reflection at the surfaces of the cornea and lens, by scattering, and by absorption.

The specular reflections are image forming: the virtual images formed by single reflection and by transmission out of the eye are called the Purkinje-Sanson I, II, III and IV images, and correspond to the anterior cornea, posterior cornea, anterior lens and posterior lens surfaces, respectively. These images can be used to estimate lens equivalent refractive index, lens surface radii of

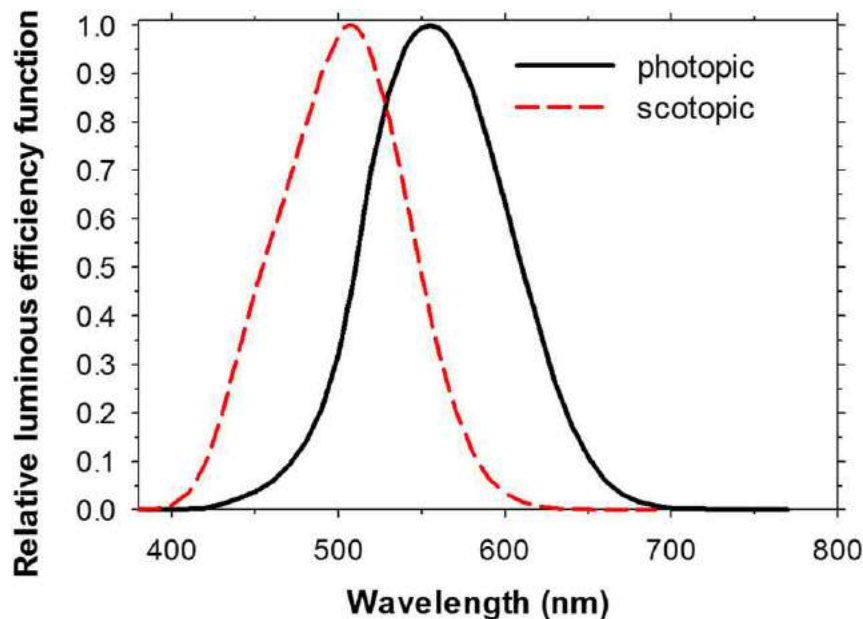


Fig. 8 The relative luminous efficiency functions in photopic and scotopic lighting levels.

curvature, and axes of the eye. Image I is by far the brightest because of the large refractive index between air and the cornea (strictly its tear film) and is located about 3.9 mm inside the eye. Image II is difficult to see because it is in a similar location to image I and is much dimmer because of the low refractive index difference (0.04) between cornea and aqueous. In the unaccommodated eye, Image III is about twice as big as image I and is about 10 mm inside the eye, while image IV is about 4 mm inside the eye, about three-quarters the size of image I and inverted. Both images III and IV change with accommodation, with image III becoming much smaller and moving closer to the cornea.

Transmittance of the eye and its components is highly dependent upon wavelength and age. Of particular interest is the protection given to internal structures from ultraviolet radiation (<380 nm). Considering natural radiation, most ultraviolet below about 288 nm is absorbed by the oxygen and ozone in the atmosphere; 288 nm corresponds approximately to the wavelength at which the cornea starts passing radiation through to the lens. In turn, the lens absorbs nearly all radiation up to 380 nm. Absorption for wavelengths beyond 600 nm is dominated by water, while for shorter wavelengths the spectral absorption is far more absorbing than occurs for water (Atchison and Smith, 2000).

Scatter is caused by spatial variations of refractive index within a medium, usually at the microscopic level, and is a combination of diffraction, reflection, and reflection. The angular distribution of scatter may be complex. Distinction can be made between forward-scattered light which affects retinal imagery and is sometimes called straylight, and back-scattered light directed out of the eye. The forward scatter produces a veiling glare that reduces the contrast of the retinal image and reduces visibility of objects. The backward-scattered light is important for observing intraocular structures in techniques such as slit-lamp biomicroscopy. Given that the cornea and lens contain inhomogeneities on the scale of the wavelength of light, their transparencies in healthy eyes are remarkably high.

An interesting phenomenon related to transmittance is birefringence. This occurs when the velocity of light depends not only on the direction of light travel but also on orientation of the electric field. This results in two refractive indices in a material, and the birefringence is quantified by their difference. Estimates of birefringence have been made for the cornea and retina.

Light at the Retina

Probably about 50–90% of light entering the eye reaches the retina to form images, but there is considerable age and wavelength dependence. Ignoring these factors, a simple measure of retinal illuminance is the troland. The retinal illuminance E_t in trolands is the product of the area A of the pupil in mm^2 and object luminance L in cd m^{-2} :

$$E_t = LA$$

In terms of pupil diameter in millimeter, this is

$$E_t = \pi LD^2/4$$

The luminous efficiency of a beam of light depends upon the entry point of the pupil. This is known as the Stiles–Crawford effect of the first kind (Stiles and Crawford, 1933). This is due to the wave-guide properties of the photoreceptors and can be considered both as an optical and neural phenomenon. A convenient equation for this is

$$\eta/\eta_{\max} = e^{-\rho(r-r_{\max})^2}$$

where $\eta/\eta_{\max}$ is relative luminous efficiency for which η is sensitivity at position r from pupil center and $\eta_{\max}$ is peak sensitivity at $r_{\max}$, and with ρ the retinal directionality parameter (Fig. 9(a)). The fit can be extended to two dimensions to allow for non-rotational symmetry. Mean foveal ρ is $0.12 \pm 0.03 \text{ mm}^{-1}$ with horizontal and vertical components for $r_{\max}$ of 0.5 ± 0.7 mm nasal and 0.2 ± 0.6 mm superior (Atchison and Smith, 2000). The Stiles–Crawford effect is sometimes factored into retinal illuminance determinations and image quality determinations as a photometric efficiency $S(D)$ for a pupil diameter D as it were a filter at the pupil, which from the previous equation and ignoring the offset of the peak from the pupil center is

$$S(D) = \frac{4 \left[1 - e^{-\frac{\rho D^2}{4}} \right]}{\rho D^2}$$

A smaller, but equivalent pupil of diameter D_e without a Stiles–Crawford effect is given by

$$D_e = D \sqrt{[S(D)]}$$

This is shown in Fig. 9(b).

The ρ value of 0.12 mm^{-1} is applicable for subjective measurements not far from the middle of the vision spectrum, and under photopic conditions. It will vary with wavelength and reduce as luminance decreases. Objective methods give much higher values. The Stiles–Crawford effect reduces the effects of aberrations on retinal image quality, although the influence of this is probably small.

Light Interaction With the Back of the Eye

In addition to the absorption of light by the photoreceptors discussed above, there are other interactions with the back of the eye in the retina, choroid and sclera, collectively called the fundus. There are important absorbing pigments: macula pigment in the retina called xanthophyll, the visual pigments in the photoreceptors responsible for initiating the neural process of vision, melanin

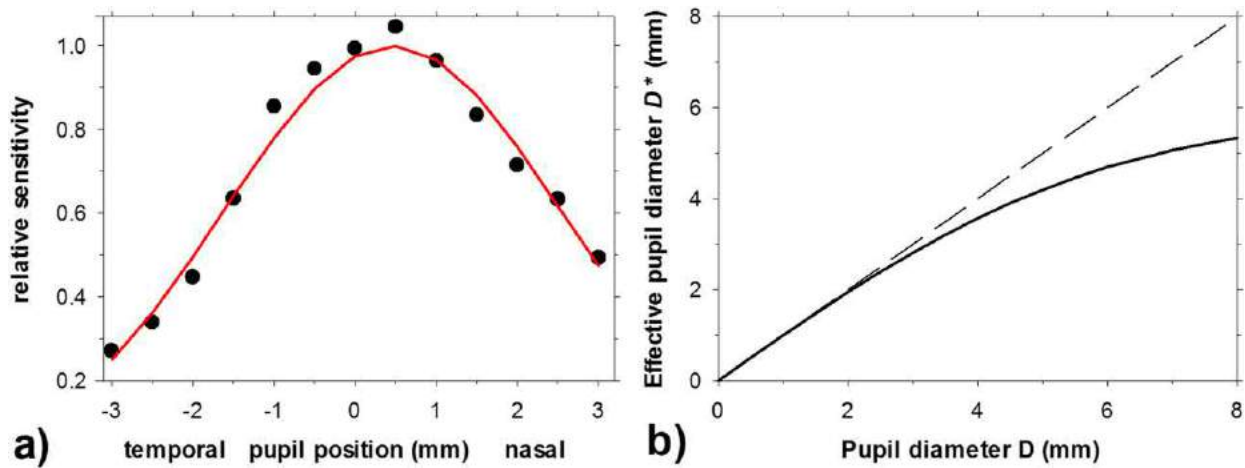


Fig. 9 The Stiles–Crawford effect. (a) Relative sensitivity as a function of position in the pupil, with data fitted by $e^{-0.116(x-0.46)^2}$ and (b) the photometrically equivalent pupil diameter as a function of pupil diameter for a ρ value of 0.12 mm^{-1} .

in the retinal pigment epithelium and choroid, and hemoglobin mainly in the choroid. The sclera backscatters light strongly so that most light reaching it passes back to the retina.

Fundus reflectance, the light reflected and scattered at the fundus and which passes back out of the eye, is used for diagnostic evaluation. Reflectance is low at short wavelengths and gradually increases with increase in wavelength, largely attributable to blood in the choroid. Some of the scattered light will contribute to veiling glare; it is thought that an important role of the Stiles–Crawford effect is to reduce the luminous efficiency of this light.

Refractive Anomalies

When the eye fixates an object of interest, ideally the image is focused at the retina. This may require some accommodation effort of the eye so that the object lies within the far and near points corresponding to minimum and maximum accommodation, respectively.

An eye whose far point is at infinity is referred to as an emmetropic eye, and this is considered the normal eye provided that it has appropriate accommodative amplitude. A refractive anomaly, or refractive error, occurs when the far point is not at infinity. Eyes with far points that are not at infinity are referred to as ametropic eyes. Emmetropia and ametropia might be regarded as opposites, but emmetropia is part of the ametropia distribution.

The departure from emmetropia is often considered to be an error of refraction, and ametropias are referred to also as refractive errors. Refractive anomalies can be corrected with ophthalmic lenses in the form of spectacle, contact, and intraocular lenses. Ophthalmic practitioners do not distinguish between the terms of refractive error and refractive correction, although from a purist perspective they are opposites.

While emmetropia may be defined in terms of the far point being at infinity, within the precision of refractive correction and ophthalmic lens manufacture this is not a practical definition. People with corrections within a range such as -0.25 D to $+0.50 \text{ D}$ may be considered to be emmetropic, but some authorities may give slightly different ranges.

If the range of accommodation is reduced with age so that near objects of interest cannot be seen clearly, it gives a refractive anomaly called presbyopia.

Spherical Refractive Anomalies

In spherical refractive anomalies, the refraction is unaffected by the meridian in the pupil of the eye. These are categorized according to the far point position (the spherical refractive errors) or to the near point position (presbyopia).

For a myopic eye, the far point is in front of the eye (Fig. 10(a)). An object at infinity is focused at the eye's posterior focal point F' , which is in front of the retina and consequently the retinal image is blurred. There is a mismatch between the length of the eye and its power, so the eye can be considered as being too powerful for its length or as being too long for its power, although the latter is the better point of view as most myopic eyes have excessive length compared with emmetropic eyes. Clear focus on a distant object can be achieved by placing an ophthalmic lens of appropriate negative power at the front of the eye (Fig. 11(a)). This lens works by forming a virtual image at its back focal plane, which coincides with the far point of the eye.

Myopia has been classified by magnitude (e.g., low, moderate, high), and age of onset (e.g., early onset or late onset). Uncorrected myopes complain of blurred distance vision, which is most noticeable at night when pupils are large and there is a small depth-of-focus. If the level of myopia is sufficient, they may complain that close objects also appear blurred.

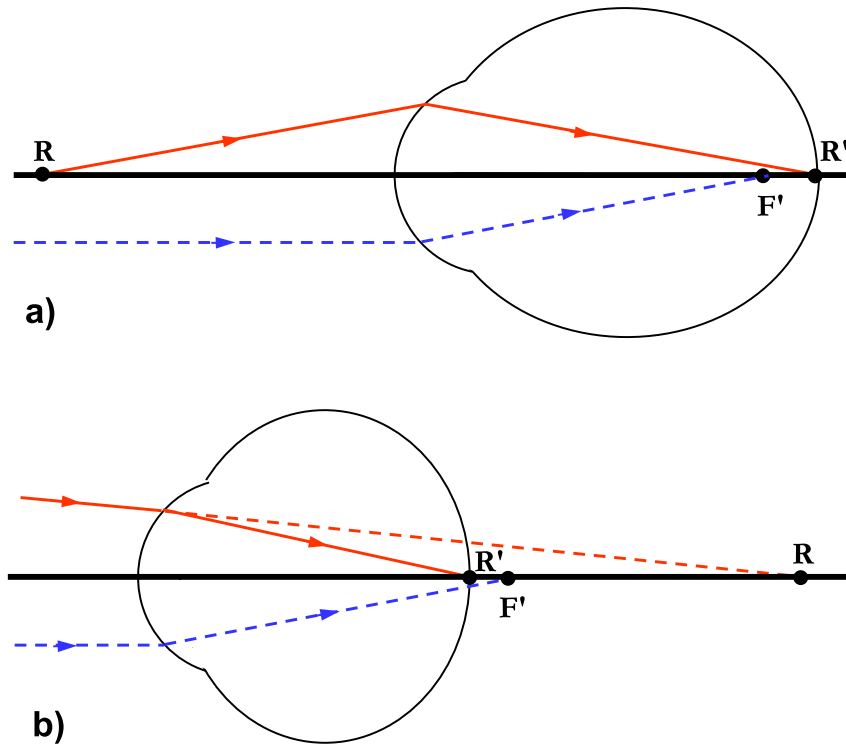


Fig. 10 (a) Myopic and (b) hyperopic eyes, showing the far points at R , retinas at R' , and posterior focal points at F' .

For a hyperopic (also hypermetropic) eye, the far point is behind the eye and the posterior focal point F' is behind the retina (Fig. 10(b)). An object at infinity is focused toward this point, and the retinal image is blurred unless the eye has sufficient accommodation amplitude to bring the image into sharp focus. As for myopia, hyperopia can be considered as a mismatch between the length of the eye and its power, but now with the eye being too weak for its length or too short for its power. Clear focus on a distant object can be achieved by placing an ophthalmic lens of appropriate positive power at the front of the eye (Fig. 11(b)). This lens forms a real image at its back focal plane, coinciding with the far point of the eye.

Young hyperopes may have difficulty relaxing accommodation, with residual ciliary muscle tonus producing latent hyperopia. Total hyperopia consists of manifest and latent components, with the portion of the former that can be compensated by accommodative effort being referred to as facultative hyperopia and the remainder as absolute hyperopia. Taking into account the loss of accommodation with age, this can be expressed as

$$\text{Total} = (\text{Facultative} \downarrow + \text{Absolute} \uparrow) + \text{Latent} \downarrow = \text{Manifest} \uparrow + \text{Latent} \downarrow$$

with the directions of the arrows indicating whether a component increases or decreases with aging.

Because uncorrected hyperopes must make more accommodative effort than emmetropes or myopes to view close objects, they tend to complain of sore eyes and headaches associated with effort in near visual tasks. Depending on the level of hyperopia, they may complain of blurred near vision and blurred distance vision, with the former being worse than the latter.

Presbyopia (from the Greek for “old eye”) is the difficulty people have in performing near tasks because of the age-related decrease in accommodative amplitude. The near point moves away from the eye to be beyond the near task position. Age of onset is related to the sign and magnitude of refraction, with uncorrected hypermetropes having problems earlier in life than uncorrected myopes. Presbyopia is corrected by ophthalmic lenses that are more positively, or less negatively, powered than the distance correction. Of course, uncorrected presbyopes’ main complaint is difficulty performing close tasks.

Astigmatism

Astigmatic refractive errors occur when the refractive error depends upon meridian (Fig. 12). It is usually due to one of more of the refracting surfaces, and particularly the anterior corneal surface, having a toroidal shape. However, it may be due to one or more surfaces being tilted or displaced. The level of astigmatism is the difference being the least and most powerful meridians. Astigmatism is corrected by lenses which have different powers in these meridians, with the power difference being referred to as a cylinder. The cylinder can be given as either positive or negative power, depending upon how the difference is determined. The axis of a positive cylinder will be at 90 degree to its corresponding negative cylinder axis. The axis of a cylinder is specified by an anticlockwise rotation from the right side of an observer looking at the lens in front of a patient’s face, to a maximum of 180 degree. The vertical meridian is usually given as 180 degree rather than 0 degree.

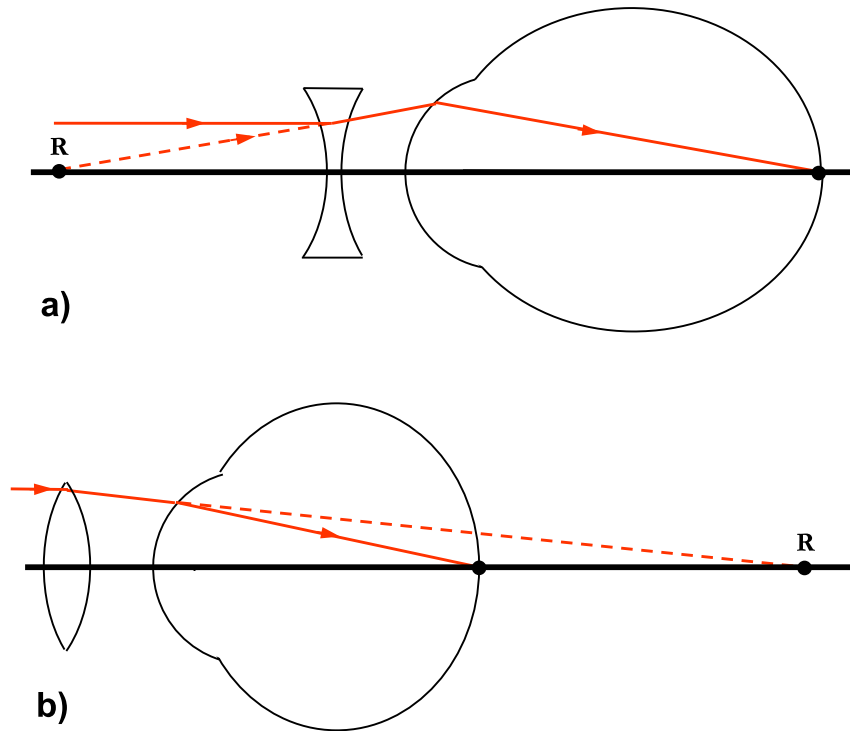


Fig. 11 Ophthalmic corrections for (a) myopia and (b) hyperopia. The posterior focal planes of the correcting lenses coincide with the far points of the eyes at R .

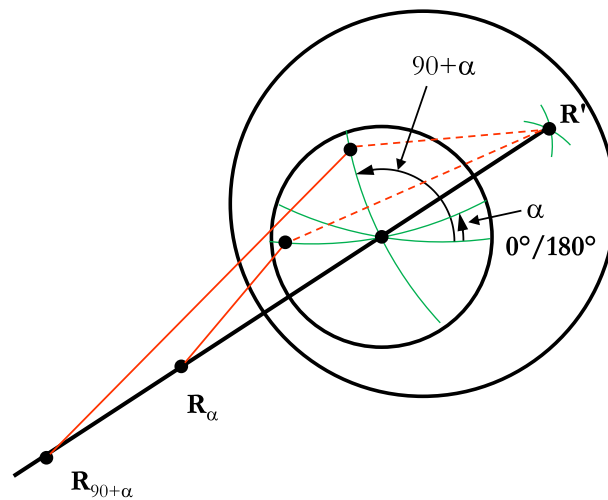


Fig. 12 Astigmatism in the eye. Far points R_α and $R_{\alpha+90}$ are imaged at retinal point R' .

There are different subclassifications of astigmatism. In myopic astigmatism, the eye is too powerful for its length in one (simple myopic astigmatism) or both (compound myopic astigmatism) principal meridians. In hyperopic astigmatism, the eye is too weak for its length in one (simple hyperopic astigmatism) or both (compound hyperopic astigmatism) principal meridians. Mixed astigmatism occurs when one principal meridian is myopic and the other is hyperopic.

Astigmatism may also be classified by axis. With-the-rule astigmatism requires a correcting negative cylinder with an axis within ± 30 degrees of the horizontal meridian; it is usually associated with an anterior cornea that is steeper along the vertical than along the horizontal meridian. Against-the-rule astigmatism requires a correcting negative cylinder lens with an axis within ± 30 degrees of the vertical meridian; it is usually associated with an anterior cornea that is steeper along the horizontal than along the vertical meridian. Oblique astigmatism occurs if the axes are beyond 30 degree from the horizontal and vertical meridians.

Astigmatism causes blurred vision at all distances, although, depending on the type of astigmatism, this may be worse at distance or near. Astigmatism may complain of tired and sore eyes and headaches associated with demanding visual tasks.

Ophthalmic lenses to correct astigmatism usually have one spherical surface and one toroidal surface, with the latter generally being the back surface, but are often referred to as sphero-cylindrical lenses. They have the specification

$$S/C \times \alpha$$

where S and C are spherical and cylindrical powers and α is the cylindrical axis. Another specification that is becoming popular is

$$M, J_{180}, J_{45}$$

where M is the mean sphere (also equivalent sphere) and J_{180} and J_{45} are regular and oblique astigmatism components. The two specifications are related by

$$M = S + C/2, J_{180} = -C \cos(2\alpha)/2, J_{45} = -C \sin(2\alpha)/2$$

and, using negative power cylinders,

$$C = -2\sqrt{(J_{180}^2 + J_{45}^2)}, \alpha = \tan^{-1}(J_{45}/J_{180})/2, S = M - C/2$$

Some rules have to be used to ensure that α is calculable and is correctly within the range of 0–180 degrees. For calculations of α that do not retain information about signs of J_{180} and J_{45} :

$$\text{If } J_{180} = 0, J_{45} < 0, \alpha = 135 \text{ degree}$$

$$\text{If } J_{180} = 0, J_{45} \geq 0, \alpha = 45 \text{ degree}$$

$$\text{If } J_{180} < 0, \alpha = \alpha + 90 \text{ degree}$$

This should give an angle between -90 and 90 degrees. To put the angle between 0 and 180 , 180 degrees must be added to negative angles to place them in the range 90 to 180 degrees. Either add 180 degree to negative angles or apply the rule:

$$\text{If } J_{180} \geq 0 \text{ and } J_{45} \leq 0, \alpha = \alpha + 180 \text{ degree}$$

The M, J_{180}, J_{45} form is useful for averaging a number of readings or for analyzing population data. Such calculations are not valid using the $S/C \times \alpha$ form.

Anisometropia

Here the refractive errors of fellow eyes of a person are different. Subclassifications include anisomyopia when fellow eyes are myopic, anisohyperopia when fellow eyes are hyperopic and antimetropia when one eye is myopic and its fellow eye is hyperopic. Subclassification could extend to consideration of astigmatism. Usually a threshold is set for anisometropia, for example, a person might be considered to have clinically significant anisometropia if it exceeds 1.0 D.

Correction of Refractive Anomalies

As mentioned before, refractive anomalies may be corrected by spectacle, contact and intraocular lenses. Nonsurgical and surgical procedures for the cornea are also available. The former occurs in orthokeratology in which rigid contact lenses are worn overnight to alter the shape of the anterior cornea. Surgical procedures involve removal of tissue from the cornea to alter anterior corneal shape. Future approaches are in development for the noninvasive alteration of the cornea's refractive index, thereby correcting refractive anomalies without the removal of tissue.

Several procedures have been adopted for the correction of presbyopia, where the challenge is to correct vision for different distances. For spectacles, this has meant having different powers in different parts of the lens in the form of multifocal or progressive addition lenses. This approach has also been attempted in contact lenses relying on pressure from the bottom eyelid to move lenses upwards relative to the cornea on downwards eye movement during near tasks. Usually contact lenses and intraocular lenses, and even corneal surgery, use simultaneous vision in which the best compromise between the demands of different distances is attempted; this will reduce image contrast compared with in-focus monofocal corrections as some light must always be out of focus ([Fig. 13](#)). Modalities include aspheric designs and annuli with different powers. Vision at different distances is highly dependent on pupil size. This dependence is overcome to some extent by using diffractive optics so that light from the entire pupil is concentrated into two orders.

Another modality for correcting presbyopia is a form of simultaneous vision, called monovision, in which different corrections are applied to fellow eyes, for example, one eye corrected for distance tasks and the other for near tasks. This relies on the brain being able to pay attention only to the eye with the clearer image for a particular task. This can be used with contact lenses, intraocular lenses, and corneal refractive surgery.

Another approach to presbyopia is not to modify refraction but to increase depth-of-focus by the use of small apertures in contact lenses, intraocular lenses or in disks implanted into the cornea.

All these correction types come with side effects, for example, spectacle lenses correcting myopia will minify the size of the retinal image. As mentioned above simultaneous vision lenses cannot provide crisp vision at any distance. Spectacle lens correction of anisometropia results in different retinal image sizes and different prismatic effects when looking through lens peripheries, and these may compromise binocular vision.

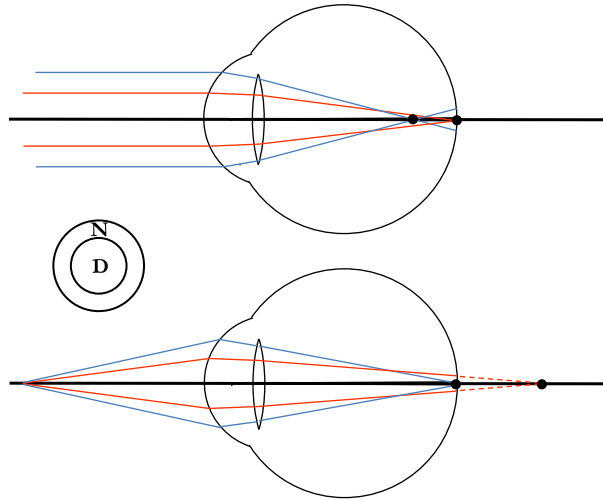


Fig. 13 Imagery with a bifocal contact lens with inner distance and outer near zones. Effects are exaggerated.

Corrections have aberrations that interact with those of the eye. Spectacle lenses do not move with the eye, and their important aberrations are associated with foveal vision and oblique incidence of light on the lens as the eye rotates behind it. Aberrations highly dependent on pupil size, such as spherical aberration, are not important for spectacle lenses as the surface curvatures are too small to affect aberrations. Contact lenses and intraocular lenses move with the eye and have high surface curvatures, and aberrations highly dependent on pupil size are important.

Some ingenious arrangements have been considered for the restoration of accommodation involving moving or altering the shape of intraocular lenses under the influence of the ciliary body during accommodative effort; these rely on the ciliary body's accommodative apparatus being intact. One surgical proposal relies on a maverick theory of accommodation and presbyopia in which the action of the ciliary body is to increase the lens diameter with age, at odds with the argument given here (see Section Changes With Age, Accommodation and Presbyopia), and this is made more difficult by a supposed increase in lens diameter with age. The surgery involves increase in the diameter of the eye so that the ciliary body can once again influence lens shape. At this time, none of these methods has been able to restore more than 2–3 D of accommodation, and usually any improvement is transient.

Aberrations

The refractive anomalies can be considered as optical aberrations of the eye. They reduce retinal image quality, prevent the image from being a perfect replica of an object (disregarding size and luminance effects), and adversely affect visual performance (e.g., visual acuity when reading letter charts). There are other aberrations which can interact with the refractive anomalies. Some of these aberrations become important when the refractive anomalies are corrected. These additional aberrations can be classified as monochromatic aberrations and the chromatic aberrations. Monochromatic aberrations are present at any wavelength and indeed change with wavelength. The chromatic aberrations occur when there is multi-wavelength light, and result from the varying refractive index of a transparent material with wavelength. This latter produces the separation of white light into its colored components known as chromatic dispersion.

Monochromatic Aberrations

There are three common ways of representing monochromatic aberrations (**Fig. 14(a)**). A wave aberration is the departure of the wavefront from the ideal waveform, as measured at the exit pupil. Transverse aberration is the departure of a ray from its ideal location at the image surface. Longitudinal aberration is the departure of the intersection of a ray with a reference axis (the pupil ray) from its ideal intersection; this may not always exist. The monochromatic aberrations must be measured in object space, with the exit pupil becoming the entrance pupil in this description (**Fig. 14(b)**).

In vision science, aberrations are specified as wave aberrations using the ophthalmic optics OSA variant of a Zernike polynomial function series (**ANSI, 2010; ISO, 2008**). The wave aberration $W(\rho, \theta)$ has the polar representation

$$W(\rho, \theta) = \sum_{n=0}^k \sum_{m=2i-n}^n c_n^m Z_n^m(\rho, \theta) \quad (1)$$

where Z_n^m is a particular Zernike polynomial and c_n^m is its coefficient (weighting). ρ is the relative distance from the center of the pupil and ranges from 0 to 1 (**Fig. 15**). θ is a meridian in radians and ranges from 0 to 2π . It is measured from the positive X-axis (to an observer's right when he or she looks at an eye) with an anticlockwise angle, as for the cylindrical axis scheme given earlier.

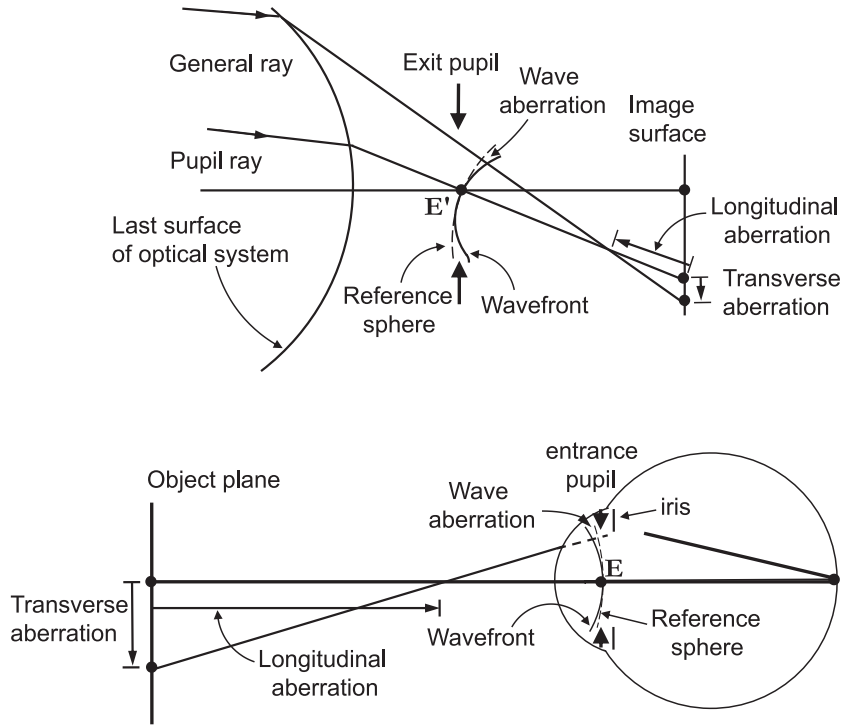


Fig. 14 Wave, transverse and longitudinal aberrations: (a) general optical system and (b) determined in object space as they must be for the eye.

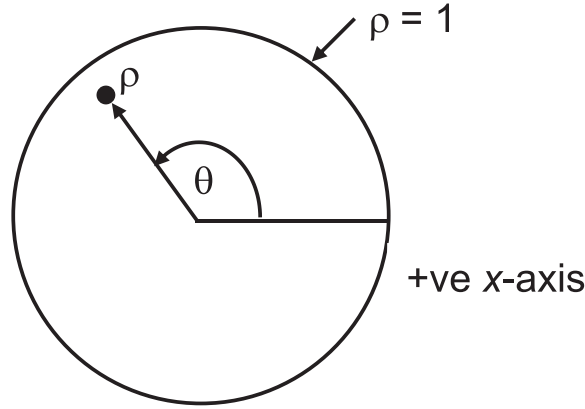


Fig. 15 Pupil coordinate system for specifying ocular aberrations.

i is counted from 0 to n , and k is the maximum order of the polynomial series. The Zernike polynomial function Z_n^m is defined as

$$Z_n^m(\rho, \theta) = N_n^m R_n^{|m|}(\rho) \cos(m\theta), \text{ for } m \geq 0 \text{ and}$$

$$Z_n^m(\rho, \theta) = N_n^m R_n^{|m|}(\rho) \sin(m\theta), \text{ for } m < 0$$

where $R_n^{|m|}$ is a radial polynomial that is a function of ρ and $N_n^{|m|}$ is a normalization term. $R_n^{|m|}$ is given by

$$R_n^{|m|}(\rho) = \sum_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)!}{s! [0.5(n+|m|)-s]! [0.5(n-|m|)-s]! \rho^{n-2s}}$$

where the index n is the highest power of the radial polynomial and the index m describes the meridional frequency of the sinusoidal component of the Zernike polynomial. The normalization term N_n^m is given by

$$N_n^m = \sqrt{n+1}, \text{ for } m = 0 \text{ and } N_n^m = \sqrt{2(n+1)}, \text{ for } m \neq 0$$

The normalization means that the variance of each Zernike polynomial function is 1, and so the square of the coefficient of the Zernike polynomial function becomes the contribution of that polynomial function to the variance of the wave aberration.

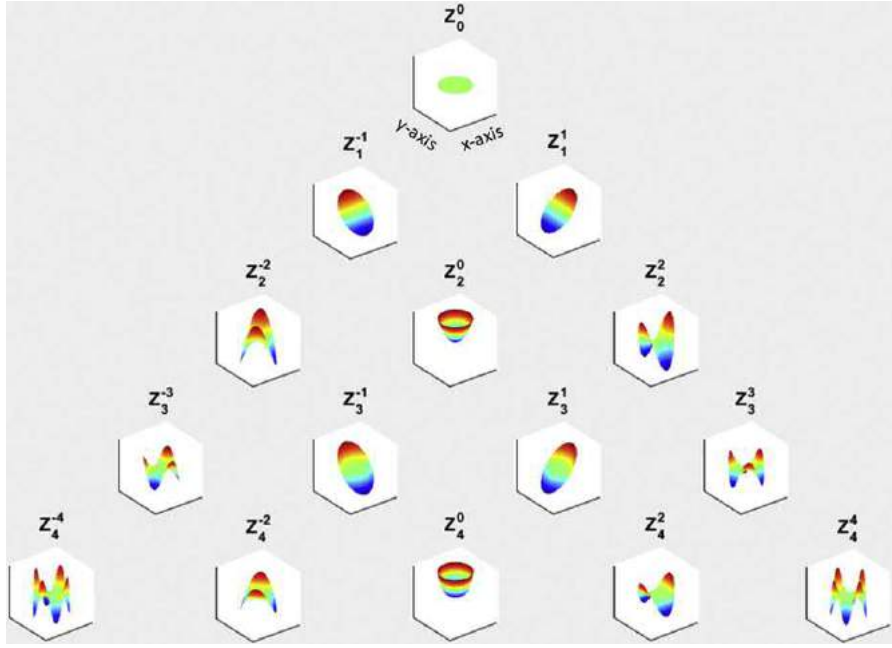


Fig. 16 Three-dimensional representation of the first five orders of the Zernike polynomial functions pyramid.

Summing the squares of the coefficients for Zernike polynomial functions of a particular order n gives the contribution of that order to the wave aberration variance. For the total variance, it is usual to disregard the piston coefficient, whose value is arbitrary, and the tilt coefficients which do not contribute to quality (at least in monochromatic light). The root-mean-square (RMS) of the wave aberration (the square root of its variance) can be used as a measure of image quality. This is given by

$$\text{RMS} = \sqrt{\sum_{n>1, \text{ all } m} (c_n^m)^2}$$

The coefficients c_n^m are relative to the size of the pupil, and must alter with pupil size. The change is not simple because a Zernike polynomial function of a particular order contains lower order terms, for example, the primary spherical aberration polynomial is $Z_4^0 = \sqrt{5}(6\rho^4 - 6\rho^2 + 1)$. There are various ways by which Zernike coefficients can be recalculated at a new pupil size. Comparing aberration of different eyes, or the aberrations of a particular eye at different times, is valid only if this is done at the same pupil size.

The Zernike polynomial function series is sometimes represented as a pyramid in which order n changes vertically and frequency m changes horizontally **Fig. 16** shows Zernike aberration functions against pupil position for terms up to the 4th-order.

The first row order is $n=0$. The single function is called piston. Its coefficient is usually manipulated to make the wave aberration zero at the pupil center.

The second row order is $n=1$. The terms are tilts or prisms, Z_1^{-1} in the (vertical) y -direction and Z_1^{+1} in the (horizontal) x -direction. They are rotated functions of one another. This rotational nature continues down the pyramid, so that in any row a function with a negative value of the index m is a rotated form of the function with the same, but positive number for m .

The third row order is $n=2$. The terms are the refraction terms. The rotationally symmetric center term is called defocus (Z_2^0). The side terms are the astigmatisms. Oblique astigmatism (Z_2^{-1}) with $m=-2$ is on the left; it has a maximum value of $+\sqrt{6}$ along the 45 degree meridian and a minimum value of $-\sqrt{6}$ along the 135 degree meridian. Horizontal/vertical astigmatism (Z_2^{+1}) with $m=+2$ is on the right; it is rotated by -45 degree from oblique astigmatism to have a maximum value along the horizontal meridian of $+\sqrt{6}$ and a minimum value along the vertical meridian of $-\sqrt{6}$.

The fourth row order is $n=3$. The terms are the first of what are called the higher-order polynomial functions. The two middling terms are the comas, one vertical (Z_3^{-1}) and one horizontal (Z_3^{+1}). The fifth-row order is $n=4$, and its rotationally symmetric center term is called spherical aberration (Z_4^0).

Even order Zernike coefficients can be converted into refraction in the M, J_{180}, J_{45} classification given in the Section Refractive Anomalies according to

$$\begin{aligned} M &= -\frac{4\sqrt{3}c_2^0 - 12\sqrt{5}c_4^0 + 24\sqrt{7}c_6^0 \dots}{R^2} \\ J_{180} &= -2\frac{\sqrt{6}c_2^{-2} - 6\sqrt{10}c_4^{-2} + 12\sqrt{14}c_6^{-2} \dots}{R^2} \\ J_{45} &= -2\frac{\sqrt{6}c_2^{+2} - 6\sqrt{10}c_4^{+2} + 12\sqrt{14}c_6^{+2} \dots}{R^2} \end{aligned}$$

where is the pupil semidiameter for which coefficients apply. Converting from an error into a correction requires the initial negative signs on the right hand sides of the equations. When doing these calculations, it is convenient to have the coefficients in micrometers and the pupil size in millimeters so that the corrections are in diopters. A lens power minimizing the RMS of the wavefront is determined by using only the first terms on the right hand sides of the equations. As additional terms are added, we come closer to a paraxial refraction, in which we are determining the refraction based on rays traced into/out of the eye for a central part of the pupil.

Most instruments for measuring aberrations use near infrared wavelengths where fundus reflection is much higher than for visible light. This requires corrections ΔM and Δc_2^0 to the spherical equivalent refraction and defocus wave aberration coefficient, respectively. A schematic eye such as the Indiana Chromatic eye can be used for this purpose (Thibos *et al.*, 1992). For the infrared wavelength 840 nm and visible wavelength 550 nm, ΔM is approximately -0.87 D, and from the previous equation for M ,

$$\Delta c_2^0 = -\Delta M \cdot \frac{R^2}{4\sqrt{3}} = -0.144\Delta M \cdot R^2$$

Several investigations have been conducted on the magnitudes and distributions of the higher-order aberrations including how they vary with pupil size, age and refraction, eye side (right versus left), following refractive surgery, in various disease conditions, and the corneal and lenticular contributions to eye aberrations. Concerning pupil size, aberrations increase as pupil size increases and the relative contribution of the higher-order aberrations increases. There is considerable between-eye symmetry, with many higher-order terms being highly correlated between fellow eyes (Hartwig and Atchison, 2012). For on-axis vision, the most important higher-order aberrations are the comas and spherical aberrations. It cannot be overemphasized that the higher-order aberrations become important only when the lower order aberrations are largely corrected.

Aberrations associated with peripheral vision have become of considerable interest in recent years. The eye has very high peripheral levels of defocus and astigmatism, and there is a shift in the pattern of the former with hyperopia and emmetropia tending toward myopia in the periphery (relative peripheral myopia) and myopes tending toward hyperopia in the periphery (relative peripheral hyperopia), at least along the horizontal vision field (Atchison, 2012). This has led to a consideration that hyperopes and emmetropes with relative peripheral hyperopia might have a tendency to develop myopia, and that compensating for the peripheral refraction might slow the progression of myopia. An alternative view is that the change in peripheral refraction pattern is a consequence rather than a cause of developing myopia.

Accompanying peripheral changes in the second order aberrations are changes in the higher-order aberrations. Most notably, horizontal coma changes approximately linearly with horizontal visual field angle and vertical coma changes approximately linearly with vertical visual field angle.

Chromatic Aberrations

Chromatic dispersion produces two types of aberration, longitudinal and transverse.

Longitudinal chromatic aberration of the eye is shown in Fig. 17(a). A beam of multiwavelength light from an axial target O enters the eye. The refractive indices inside the eye vary with wavelength, so the path followed by a ray depends upon its wavelength. Refractive indices decrease with increase in wavelength, so the eye power decreases as wavelength increases. If the eye is focused at the target for a yellow light, rays of longer wavelength (such as reds) are focused behind the retina and rays of shorter wavelength (such as blues) are focused in front of the retina.

Longitudinal chromatic aberration can be quantified as a “chromatic difference of power” $\Delta F(\lambda)$, which is the difference in power of the eye between a wavelength λ and the reference wavelength $\bar{\lambda}$, i.e.,

$$\Delta F = F(\lambda) - F(\bar{\lambda})$$

It can also be quantified as a “chromatic difference of refraction” $\Delta R_x(\lambda)$, which is the difference between vergences for which a target is focused at the retina for different wavelengths, or more simply is the change in refraction with wavelength; the second way is how the aberration is measured experimentally (Fig. 17(b)). In mathematical terms

$$\Delta R_x(\lambda) = L(\lambda) - L(\bar{\lambda})$$

where $L(\lambda)$ and $L(\bar{\lambda})$ are the object vergences required at λ and $\bar{\lambda}$, respectively. For the situation shown in the figure, these vergences are both negative.

The relationship between chromatic difference of refraction and chromatic difference of power is (Atchison and Smith, 2000)

$$\Delta R_x(\lambda) = \frac{[n'(\lambda) - n'(\bar{\lambda})] [F(\bar{\lambda}) + L(\bar{\lambda})]}{n'(\bar{\lambda})} - \Delta F(\lambda)$$

where $n'(\lambda)$ and $n'(\bar{\lambda})$ are vitreous refractive indices at λ and $\bar{\lambda}$, respectively.

For an emmetropic eye focused at infinity, $L(\bar{\lambda}) = 0$, and the equation reduces to

$$\Delta R_x(\lambda) = \frac{[n'(\lambda) - n'(\bar{\lambda})] F(\bar{\lambda})}{n'(\bar{\lambda})} - \Delta F(\lambda)$$

The range of chromatic difference of refraction is about 2.1 D between 400 and 700 nm, and is slightly higher than that for a reduced eye filled with water. There are minor differences between studies and between participants within studies. Despite the

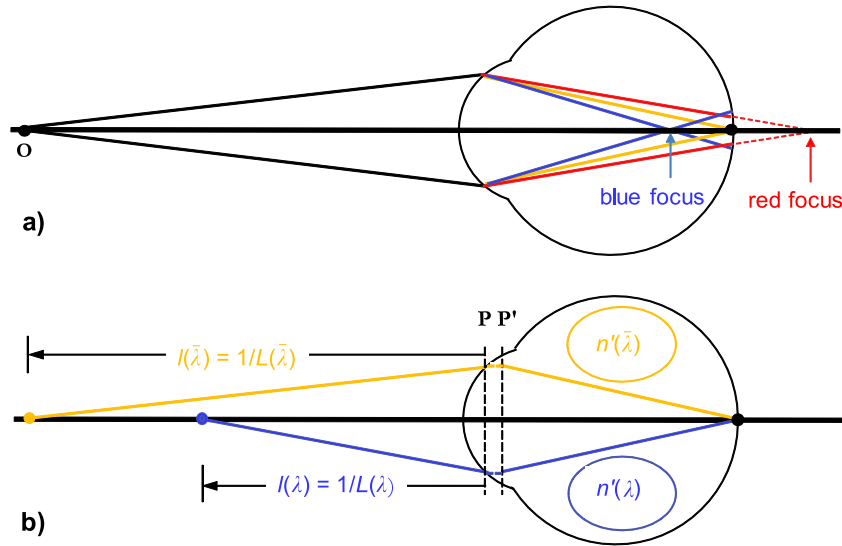


Fig. 17 Longitudinal chromatic aberration (greatly exaggerated). (a) Imagery at the eye for blue, yellow and red lights when the yellow light is focused at the retina and (b) determining chromatic difference of refraction. Object distances are specified relative to the anterior principal plane of the eye.

large magnitude, longitudinal chromatic aberration has little deleterious effect on vision, largely because its effects are attenuated by the variable sensitivity of the eye to different wavelengths (Fig. 8).

Transverse chromatic aberration of the eye is shown in Fig. 18(a) for an eye which is a centered optical system and for off-axis object point Q. The different wavelength images are at different depths because of the longitudinal chromatic aberration. Also, because the power is less for long wavelengths than for short wavelengths, longer wavelength rays are deviated less than shorter wavelength rays to reach the retina further away from the optical axis.

Like longitudinal chromatic aberration, transverse chromatic aberration must be measured outside the eye, where it is quantified as a "chromatic difference of position" (Fig. 18(b)). Rays of wavelength λ and $\bar{\lambda}$ originate from different positions in object space, but pass through the same point in the pupil to intersect at the retina. The chromatic difference of position $t(\lambda)$ for height h of the rays in the pupil, relative to the visual axis ray (ray passing through the nodal points) is

$$t(\lambda) = a_\lambda - a_{\bar{\lambda}}$$

where a_λ and $a_{\bar{\lambda}}$ are angles subtended by the rays with the visual axis in object space.

The transverse chromatic aberration and longitudinal chromatic aberration, or more specifically the chromatic difference of position and the chromatic difference of refraction, are related by the equation (Atchison and Smith, 2000)

$$t(\lambda) \approx h \Delta R_x(\lambda)$$

Because the eye is a decentered optical system, there is transverse chromatic aberration associated with foveal vision, and the foveal chromatic difference of position is

$$t(\lambda) \approx d \Delta R_x(\lambda)$$

where d is the departure of the line-of-sight, the axis joining the fixation target and the center of the pupil, from the visual axis (Fig. 18(b)).

The range of foveal chromatic difference of position across the visible spectrum can be higher than 1 min of arc. Usually the line-of-sight is nasal to the visual axis in object space (Fig. 6). In binocular vision, this can give rise to the phenomenon of chromostereopsis in which reddish objects appear to be closer than bluish objects at the same distance. Potentially transverse chromatic aberration can have deleterious effects on vision because it produces wavelength-dependent changes in spatial phase of images. This may be a problem for highly decentered natural or artificial pupils. Two-thirds loss of resolution may result for 3 mm displacement of small artificial pupils (Green, 1967; Thibos *et al.*, 1991).

Retinal Image Quality

Retinal image quality can be determined in different ways. One way is to use the wave aberrations themselves or a metric based on them such as the root-mean-squared aberrations. Another common quality measure is the point spread function, which is the shape and dimension of the image of a point object. The optical transfer function measures changes, from object to image, in the contrast and phase of sinusoidal varying luminance patterns. The wave aberration is related to the point spread function and optical transfer function by Fourier transforms, but this overestimates the quality of the retinal image as it does not take into account scatter within the media and the fundus. The point spread function can be measured by a double pass method (light

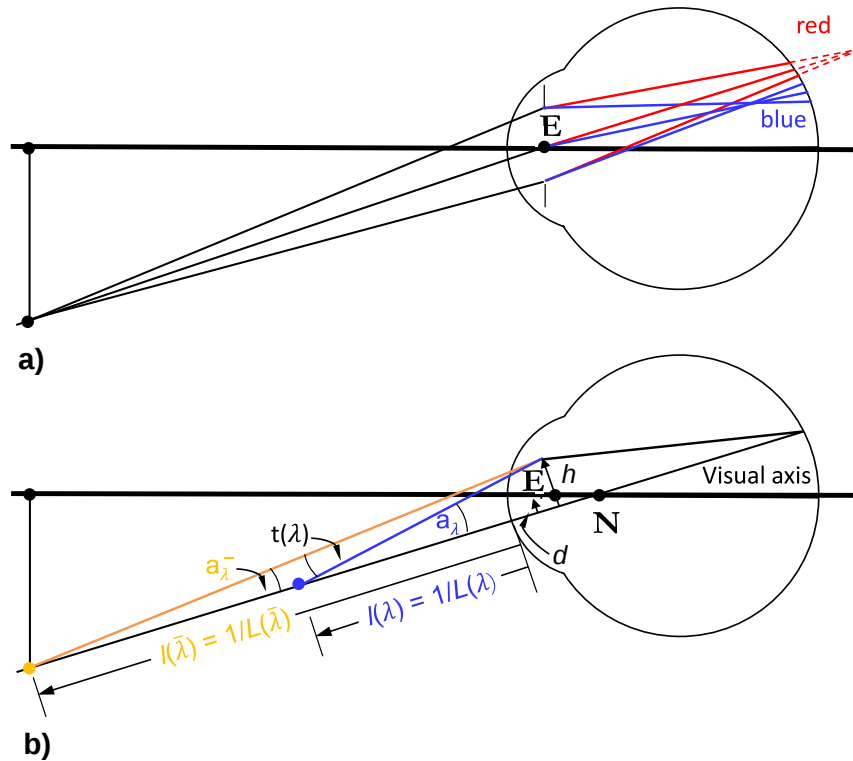


Fig. 18 Transverse chromatic aberration (greatly exaggerated). (a) Centered pupil and an off-axis object and (b) determining chromatic difference of position (for simplicity, the nodal points are assumed to be at a common point).

passes into the eye and is reflected out), and the contrast aspect of the optical transfer function can be estimated by comparing contrast sensitivities of the eye obtained with screen-based equipment and with an interferometric method in which the optics are effectively by-passed (Archison and Smith, 2000).

Several metrics based on wave aberrations have been used to estimate objective refractions. They involve manipulating refraction to minimize or maximize some aspect of the aberrations or of image quality, for example, maximizing the Strehl intensity ratio which is the height of the point spread function relative to its aberration-free height.

Changes With Age

Pathological changes that occur in the optics of the eye with age, such as cataract, will not be considered here.

Cornea

As age increases, there is fiber degeneration, decreased spacing between collagen fibrils of the stroma, and increases in the cross-sectional area of collagen fibers. Descemet's membrane increases in thickness with age and the size of endothelial cells becomes more variable. Endothelial function may become impaired, with the aqueous humor seeping into the cornea, disrupting structural order and increasing light scatter.

The anterior surface radius of curvature is usually greater along the horizontal than along the vertical meridian in young eyes, but this trend reverses with age. There is little variation in transmittance with age, but Boettner and Wolter (1962) found a decrease in the direct (non-scattered) light with increased age.

Aperture Stop and Iris

The natural pupil becomes smaller with increase in age, a process known as senile miosis, and this is accompanied by decreases in speed and magnitude of pupil reactions.

Lens

The human lens grows throughout life with considerable changes in size, shape, mass, and stiffness. The young lens is soft and easy to deform, while the old lens is stiff and unable to be deformed. With age, the lens becomes thicker along its axis with

corresponding decrease in anterior chamber depth. Its surfaces become more curved, with greater change for the anterior than for the posterior surface. Equatorial thickness either does not change or increases slowly with age. There are changes in refractive index distribution such that the refractive index change is restricted to a smaller proportion of the peripheral lens, accompanied by a reduction in the equivalent index (Kasthurirangan *et al.*, 2008).

Light Loss at the Retina

Retinal illumination decreases with age due to senile miosis and decrease in ocular transmittance. Decrease is much more rapid for shorter than for longer wavelengths (e.g., van de Kraats and van Norren, 2007) (Fig. 19). As a consequence, the lens becomes more yellow with increase in age. With aging, there is increase in both forward and backward-scattered light.

Refraction

The distribution of refraction is strongly age-dependent. Newly-born children have a normal distribution whose mean is considerably hyperopic with a distribution range to approximately ± 10 D (Grosvenor, 2002). The growth of ocular components is coordinated so that emmetropia can be achieved by 6 years of age, a process known as emmetropization, but emmetropia is not always subsequently maintained. By 6–8 years the mean refraction is slightly hyperopic and the distribution is steeper than a normal distribution (leptokurtosis). A pronounced tail in the myopic direction develops after this and is observed in data of 20-year olds. The distribution is stable between the ages of 20 and 40 years, after which the distribution becomes less leptokurtic and then there is usually a hyperopic shift (Grosvenor, 2002). The age-related formation of cataracts can alter the refraction.

Most studies of refraction have been cross-sectional, meaning the measurements of people of different ages occurred within a limited time span. These can mask intergenerational (longitudinal) changes. In particular they do not pick up changes in the direction of myopia that have been occurring in the last few decades (e.g., Edwards and Lam, 2004) (Fig. 20). Myopia prevalence is correlated with increasing urbanization (Ip *et al.*, 2008; Zhan *et al.*, 2000) and its prevalence in East Asian countries is now in excess of 80% by early adulthood (He *et al.*, 2004; Lin *et al.*, 2004; Wu *et al.*, 2001). Myopia is a leading cause of blindness in later life, as having severe myopia doubles the risk of developing serious eye problems, such as cataract and retinal detachment. There is no “safe” level of myopia, with even low levels increasing the risk of other ocular diseases (Flitcroft, 2012).

The age-related changes mentioned above for corneal toricity affect astigmatism of the eye. Considerable astigmatism, usually against-the-rule, exists in the first year, but decreases quickly during early infancy. Astigmatism is usually with-the-rule up to 40 years, after which the prevalence of against-the-rule astigmatism increases.

Accommodation and Presbyopia

Presbyopia is the name given to the loss of clear and comfort near vision which occurs because of the age-related loss of accommodation amplitude. Onset for most people is the early to mid-40s, but is affected by the amplitude, the nature and duration of close work, and other refractive conditions. The decline in amplitude of accommodation might be affected by a number of factors including diet, race, exposure to ultraviolet radiation, and temperature. As an example of the influence of refractive conditions, in the absence of a distant correction, a low myope would need a lesser amplitude of accommodation than a hyperope to perform near tasks and would become presbyopic later in life.

Previously it was mentioned that, after peaking early in life, accommodation amplitude gradually reduces to become zero by mid-50s. Many clinically related studies suggest that some accommodation remains into older age (Fig. 21). This is an artefact of subjective measurement (the participants’ judgement is involved) involving depth-of-focus.

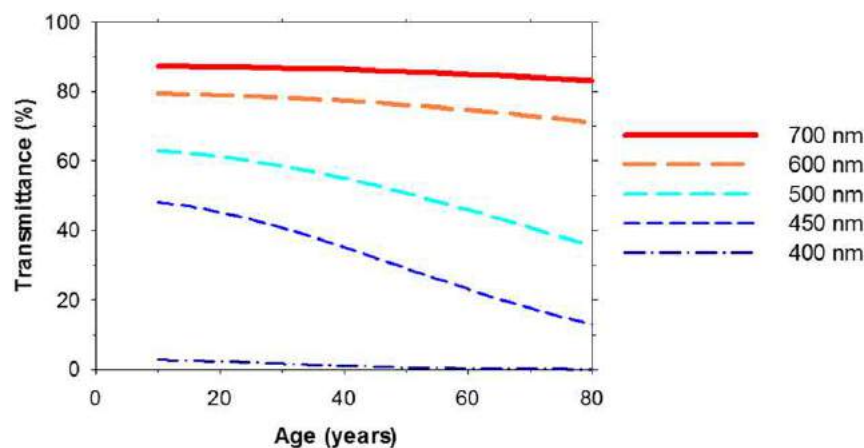


Fig. 19 Transmittance of the eye as a function of age at different wavelengths according to the fit given by Eq. (8) of van de Kraats and van Norren (2007).

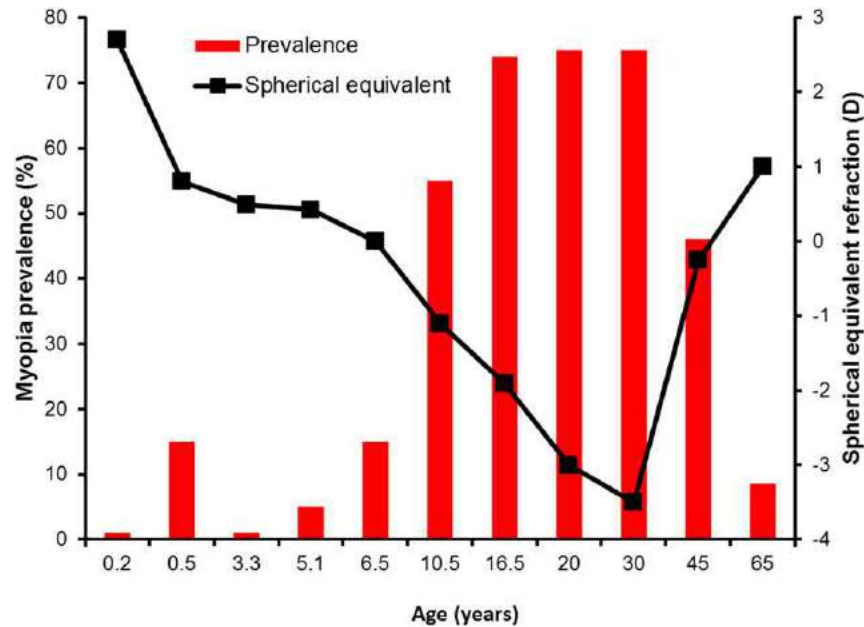


Fig. 20 Mean spherical equivalent refraction (D) as a function of age in a Hong Kong population. Data of Edwards, M.H., Lam, C.S., 2004. The epidemiology of myopia in Hong Kong. *Annals Academy of Medicine, Singapore* 33 (1), 34–38.

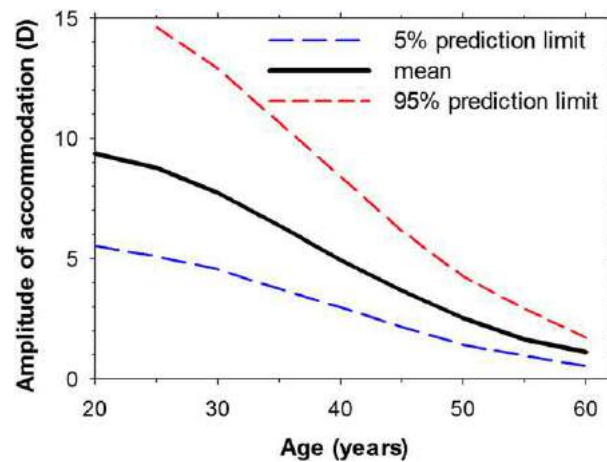


Fig. 21 Subjective amplitude of accommodation as a function of age from the study of Ungerer (1986).

Aberrations

Taking into account normal age-related pupillary miosis, there is reduction in higher-order aberrations with increase in age (Applegate *et al.*, 2007). Considering fixed pupil sizes, total higher-order aberrations increase throughout adulthood and spherical aberration becomes more positive with increase in age (Applegate *et al.*, 2007).

See also: Lenses and Mirrors

References

- American National Standards Institute (ANSI), 2010. American National Standard Z80.28 for Ophthalmics – Methods for Reporting Optical Aberrations of Eyes. Washington, DC: American National Standards Institute.
- Applegate, R.A., Donnelly 3rd, W.J., Marsack, J.D., Koenig, D.E., Pesudovs, K., 2007. Three-dimensional relationship between high-order root-mean-square wavefront error, pupil diameter, and aging. *Journal of the Optical Society of America A* 24 (3), 578–587.

- Atchison, D.A., 2012. The Glenn A. Fry Award Lecture 2011: Peripheral optics of the human eye. *Optometry and Vision Science* 89 (7), 954–966.
- Atchison, D.A., 2017a. Axes and angles of the eye. In: Artal, P. (Ed.), *Handbook of Visual Optics*, vol. I. Boca Raton, FL: Francis & Taylor Group. Chapter 17.
- Atchison, D.A., 2017b. Paraxial schematic eyes. In: Artal, P. (Ed.), *Handbook of Visual Optics*, vol. I. Boca Raton, FL: Francis & Taylor Group. Chapter 16.
- Atchison, D.A., Smith, G., 2000. *Optics of the Human Eye*. Oxford: Butterworth-Heinemann.
- Boettner, E.A., Wolter, J.R., 1962. Transmission of the ocular media. *Investigative Ophthalmology* 1, 776–783.
- Dunne, M.C., Royston, J.M., Barnes, D.A., 1992. Normal variations of the posterior corneal surface. *Acta Ophthalmology* 70 (2), 255–261.
- Edwards, M.H., Lam, C.S., 2004. The epidemiology of myopia in Hong Kong. *Annals, Academy of Medicine, Singapore* 33 (1), 34–38.
- Flitcroft, D.I., 2012. The complex interactions of retinal, optical and environmental factors in myopia aetiology. *Progress in Retinal Eye Research* 31, 622–660.
- Green, D.G., 1967. Effect of ocular chromatic aberration on monocular visual performance. *Journal of Physiology* 190, 583–593.
- Grosvenor, T.P., 2002. *Primary Care Optometry*, fourth ed. Boston: Butterworth-Heinemann.
- Gullstrand, A., 1924. Appendix IV. Mechanism of accommodation. In: Southall, J.P.C. (Ed.), *Helmholtz, H. H. (1909) Handbuch der Physiologischen Optik*. Translated From the Third German Edition, vol. 1. Washington, DC: Optical Society of America, pp. 382–415.
- Hartwig, A., Atchison, D.A., 2012. Analysis of higher-order aberrations in a large clinical population. *Investigative Ophthalmology and Vision Science* 53 (12), 7862–7870.
- He, M., Zeng, J., Liu, Y., *et al.*, 2004. Refractive error and visual impairment in urban children in southern China. *Investigative Ophthalmology and Vision Science* 45 (3), 793–799.
- Ip, J.M., Huynh, S.C., Robaei, D., *et al.*, 2008. Ethnic differences in refraction and ocular biometry in a population-based sample of 11–15-year-old Australian children. *Eye (London)* 22 (5), 649–656.
- International Standards Organisation (ISO), 2008. *Ophthalmic Optics and Instruments – Reporting Aberrations of the Human Eye*. Geneva, Switzerland: ISO, (ISO 24157: 2008).
- Kasthurirangan, S., Markwell, E.L., Atchison, D.A., Pope, J.M., 2008. In vivo study of changes in refractive index distribution in the human crystalline lens with age and accommodation. *Investigative Ophthalmology and Vision Science* 49 (6), 2531–2540.
- Lin, L.L., Shih, Y.F., Hsiao, C.K., Chen, C.J., 2004. Prevalence of myopia in Taiwanese schoolchildren: 1983 to 2000. *Annals, Academy of Medicine, Singapore* 33 (1), 27–33.
- Lowe, R.F., Clark, B.A., 1973. Posterior corneal curvature. Correlations in normal eyes and in eyes involved with primary angle-closure glaucoma. *British Journal of Ophthalmology* 57 (7), 464–470.
- Mathur, A., Gehrmann, J., Atchison, D.A., 2013. Pupil shape as viewed along the horizontal visual field. *Journal of Vision* 3, 1–8.
- Mathur, A., Gehrmann, J., Atchison, D.A., 2014. Influences of luminance and accommodation stimuli on pupil size and pupil center location. *Investigative Ophthalmology and Vision Science* 55 (4), 2166–2172.
- Patel, S., Marshall, J., Fitzke, F.W., 1993. Shape and radius of posterior corneal surface. *Refractive and Corneal Surgery* 9 (3), 173–181.
- Stiles, W.S., Crawford, B.H., 1933. The luminous efficiency of rays entering the eye pupil at different points. *Proceedings of the Royal Society of London Series B: Biological Sciences* 112, 428–450.
- Thibos, L.N., Bradley, A., Zhang, X., 1991. Effect of ocular chromatic aberration on monocular visual performance. *Optometry and Vision Science* 68, 599–607.
- Thibos, L.N., Ye, M., Zhang, X., Bradley, A., 1992. The chromatic eye: A new reduced-eye model of ocular chromatic aberration in humans. *Applied Optics* 31, 3594–3600.
- Ungerer, J., 1986. *The Optometric management of presbyopic airline pilots*. Unpublished PhD Thesis, University of Melbourne.
- van de Kraats, J., van Norren, D., 2007. Optical density of the aging human ocular media in the visible and the UV. *Journal of the Optical Society of America A* 24 (7), 1842–1857.
- Watson, A.B., Yellott, J.I., 2012. A unified formula for light-adapted pupil size. *Journal of Vision* 12, 1–16.
- Wu, H.M., Seet, B., Yap, E.P., *et al.*, 2001. Does education explain ethnic differences in myopia prevalence? A population-based study of young adult males in Singapore. *Optometry and Vision Science* 78 (4), 234–239.
- Zhan, M.Z., Saw, S.M., Hong, R.Z., *et al.*, 2000. Refractive errors in Singapore and Xiamen, China – A comparative study in school children aged 6 to 7 years. *Optometry and Vision Science* 77 (6), 302–308.

Measuring Retinal Blood Flow

Alberto de Castro and Stephen A Burns, Indiana University School of Optometry, Bloomington, IN, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The retina of the eye, in addition to being one of our most vital sensory organs, is the most accessible portion of the central nervous system (CNS). This has led to an array of noninvasive approaches to measuring its structure and function. It is uniquely accessible for optical measurements, since evolution has resulted in an optical window that lets us probe more than 2 cm into the body. While early optical measurements of the retina were primarily structural and based on photography, it was quickly realized that the blood supply of the retina, which enters and exits the eye through the optic nerve head and courses across the surface, provides a unique window into the CNS. The retina has the added advantage that the circulation is laid out in a layered form, with most of the vessels lying in a series of layers parallel to the walls of the eye (Snodderly *et al.*, 1992), an arrangement which parallels the neural layers of the retina. This arrangement allows ready optical access to the vessels and is simpler than the layout in the cortex, which can be quite complex. Thus the eye was an early target for measurements of blood flow and its regulation. The techniques used included measures of the flow of fluorescein dye (Wessing, 1974; Jung *et al.*, 1983; Wolf *et al.*, 1989, 1990), speckle decorrelation (Fercher and Briers, 1981; Suzuki *et al.*, 1991), and laser Doppler velocimetry (Riva *et al.*, 1972, 1979; Feke and Riva, 1978) or Doppler optical coherence tomography (OCT) (Chen *et al.*, 1997; Westphal *et al.*, 2002). Measurements have included blood velocity, total flow, and have incorporated the regulation of blood flow, allowing measures of both the metabolic demands of the retina and body, as well as how the oxygen supply is driven by neuronal tissue. That is, when a visual stimulus is presented to the eye, the blood flow must change to meet the increased metabolic demand of the neural tissue, a process referred to as neurovascular coupling (Riva *et al.*, 2001; Pournaras *et al.*, 2008; Kur *et al.*, 2012). Some retinal diseases, such as diabetic retinopathy (DR), have a severe impact on both the structure of the retinal vasculature and the regulation of blood flow (Grunwald *et al.*, 1986, 1990; Kohner *et al.*, 1995; Garhöfer *et al.*, 2004; Deak and Schmidt-Erfurth, 2013).

Most of the blood flow in the posterior pole of the eye is supplied by the central retinal artery and exits the eye through central retinal vein, both of which traverse the optic nerve. For this reason measurements near the optic nerve can capture a large proportion of the blood supply to the neural retina as it enters and exits the eye. These measurements therefore can provide a tool for understanding the total retinal metabolic demand. These vessels are relatively large, typically on the order of 100 μm or larger, and thus are easily resolved and carry large volumes of blood, providing a high signal:noise ratio (SNR) for measurements. However, vascular control (Villringer and Dirnagl, 1995; Riva *et al.*, 2001, 2005; Pournaras *et al.*, 2008) is performed at a more local level (Zhong *et al.*, 2012), and in fact it is not known how local the control of retinal blood flow is, nor the area over which it is regulated. In recent years there has been increasing interest in making more local measurements of retinal blood flow driven in part by the growing availability of adaptive optics in retinal imaging.

This article provides a brief overview of current optical techniques for measuring retinal blood flow and then concentrates on approaches to measure the cell velocity in the capillaries of the human retina.

Total Retinal Blood Flow

Early measurements of blood flow depended on injecting fluorescein into the body and then using photographic and video techniques to measure the transit time (Jung *et al.*, 1983). These dye methods improved with the development of techniques based on the scanning laser ophthalmoscope (SLO) (Wolf *et al.*, 1989, 1990) and because they measured the transit of blood moving from the central retinal artery to the draining veins, they gave an estimate of the total transit time of the retina. By combining this with the approximate size of the vessels, and making assumptions about the nature of the flow based on parabolic velocity distribution (Pouisselle's law), an estimate of total blood delivered to the eye could be obtained. These techniques are mostly based on plasma velocity, since fluorescein is a small molecule and is present in the plasma fraction of the blood. The advent of the laser allowed development of laser Doppler velocimetry measurements (Feke and Riva, 1978; Riva *et al.*, 1979), and by making these measurements from two points in the pupil, variations due to angle and bulk movements could be accounted for, improving the accuracy. More recently, OCT has been used to make Doppler measurements (White *et al.*, 2003; Leitgeb *et al.*, 2004; Dai *et al.*, 2013; Doblhoff-Dier *et al.*, 2014; Aschinger *et al.*, 2015). In both techniques the Doppler shift arising from red blood cells is used to provide a measurement of the velocity of the cells. Typically the centerline velocity has been measured, since this provides a maximum frequency shift on the impinging light. It is becoming increasingly apparent that the flow in the retina is not parabolic (Bishop *et al.*, 2001; Logean *et al.*, 2003; Zhong *et al.*, 2011; Willerslev *et al.*, 2014) and the advances in both axial and lateral resolution are allowing more accurate estimates of total blood flow (Baumann *et al.*, 2011).

The movement of cells in the vasculature can also be used to generate flowmetry estimates. In these it is the variation over time, rather than the Doppler shift, that is used to estimate perfusion of the tissue. This was first attempted using speckle decorrelation (Fercher *et al.*, 1985) and flowmetry (Riva *et al.*, 1992) and the motion alone has been used to qualitatively separate vessels according to their speed (Ferguson *et al.*, 2004). OCT also can be combined with this technique, and this is generally called OCT

angiography (or OCT-A) although there are a number of variants (Makita *et al.*, 2006; Wang *et al.*, 2007; Tokayer *et al.*, 2013). Finally the motion can be directly tracked by imaging a single line repeatedly over a larger vessel with an adaptive optics scanning laser ophthalmoscope (AOSLO). In this method, by stopping the slow scanner, streaks created by the erythrocytes passing through are visible in the spatiotemporal image and its slope can be related directly to speed (Zhong *et al.*, 2008, 2011).

Capillary Blood Flow

Why Measure Blood Flow in Capillaries

Noninvasive access to capillaries and capillary flow is of increasing interest in the retina. Capillary velocities have mainly been measured in the area surrounding the fovea, where there is one or two capillary layers and one can easily assume that these are perpendicular to the retina, i.e., in the imaging plane. There are a number of purposes for such measurements, ranging from the understanding of the local control of blood flow, to biological parameters governing flow, such as the interaction of blood with the vascular walls, to the impact of disease on both the physical properties of the vascular system and the regulation of flow. Techniques for measuring flow at the capillary level depend on the ability to measure the movement of cells on an individual basis. A number of techniques have been devised for this, including psychophysics, fluorescent detection techniques, and, most recently, adaptive optics (Fig. 1). Some general patterns emerge from these techniques, however, it is important to note that different techniques measure different aspects of the blood, ranging from the relatively large leukocytes (psychophysics and recently, AO imaging), to cell aggregates interspersed with plasma (fluorescein), to red blood cells (direct imaging with adaptive optics). Because of this heterogeneity in the measuring techniques, differences in velocities obtained using different techniques are not surprising. The small diameters of the capillaries and the need for cells to deform as they move through them can cause differences in velocity for different blood fractions. However, understanding the differences between the methods and the velocities being measured remains an important topic of research as will be an understanding of the variability of the normal retina, since this provides the background upon which clinical measurements need to be compared. The next section of this article compares results obtained with six different methods, most having very different assumptions.

Blue Field Entoptic Phenomenon

The retina itself can be used as a detector for measuring retinal blood flow. This is what is done with the blue field entoptoscope (Kato, 1952; Riva and Petrig, 1980). When looking at a relatively bright blue field (including a clear blue sky), bright dots can be observed moving in an erratic manner in the observer's visual field. These dots are believed to be white blood cells moving in the capillaries of the eye. White blood cells do not absorb blue light and thus, when crossing over a region of retina the cells are detected by the photoreceptors giving the appearance of a bright dot. By comparing the perceived motion of these dots to a simulation on a computer monitor the apparent speed can be matched (Riva and Petrig, 1980). The blue field technique gives an average velocity of 0.77 mm s^{-1} and this varies with exercise (Kato, 1952). Velocity estimates are pulsatile, changing with the heart beat (Riva and Petrig, 1980) and the mean velocity decreases with age by 20% (Grunwald *et al.*, 1993). The measured pulsatility, defined as $(V_{\text{max}} - V_{\text{min}})/V_{\text{mean}}$, is 0.98 (Riva and Petrig, 1980). The impact of diabetes on capillary flow is non-monotonic; in diabetics with no retinopathy, Fallon *et al.* (1986) observed average velocities similar to those in non diabetics

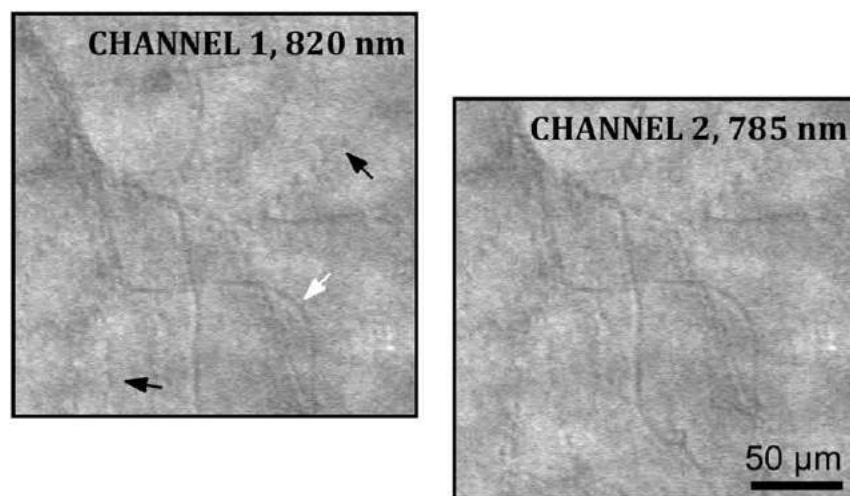


Fig. 1 Adaptive optics scanning laser ophthalmoscope (AOSLO) images of capillaries in the perifoveal area of a healthy subject. The use of multiply scattered light has allowed direct visualization of leukocytes (passing through the capillary marked with the white arrow) and erythrocytes (black arrow) in the capillaries in the retina (individual video frames).

Table 1 Average velocities measured in the blood cells. The different measurement techniques target different cells type. Numerical values are the average value found in each study, the standard deviation of the measurement when this was reported in between parenthesis, the range of velocities measured

Author	Cell type	Velocity (mm mm s ⁻¹)	Measurement method
Kato (1952)	Leukocyte	0.77	Entoptic phenomenon
Riva and Petrig (1980)	Leukocyte	0.74 (0.25–1.33)	Entoptic phenomenon
Fallon <i>et al.</i> (1986)	Leukocyte	0.54 ± 0.19 Non diabetic 0.51 ± 0.24 No retinopathy 0.74 ± 0.32 Background DR 0.37 ± 0.20 Preproliferative DR 0.56 ± 0.27 Proliferative DR 0.42 ± 0.14 Photocoagulated	Entoptic phenomenon
Wolf <i>et al.</i> (1991)	Erythrocyte	3.28 ± 0.45 Non diabetics 2.89 ± 0.57 Diabetics	Fluorescein, scanning laser ophthalmoscope (SLO)
Arend <i>et al.</i> (1991)	Erythrocyte	3.28 ± 0.45 Healthy 2.51 ± 0.88 No retinopathy 2.42 ± 0.40 Background DR 2.28 ± 0.16 Preproliferative DR 2.27 ± 0.58 Proliferative DR	Fluorescein, SLO
Grunwald <i>et al.</i> (1993)	Leukocyte	0.92 ± 0.12 for 20–50 y.o. 0.61 ± 0.21 for 50–70 y.o.	Entoptic phenomenon
Scheiner <i>et al.</i> (1994)	Leukocyte	0.60 ± 0.07 No flickering 1.01 ± 0.11 16 s Flickering	Entoptic phenomenon
Arend <i>et al.</i> (1995)	Erythrocyte	2.68 ± 0.30	Fluorescein, SLO
	Leukocyte	0.89 ± 0.20	Entoptic phenomenon
Yang <i>et al.</i> (1997)	Leukocyte	1.37 ± 0.35 1.41 ± 0.29	Fluorescein, SLO Labeled leukocytes, SLO
Paques <i>et al.</i> (2000)	Leukocyte	1.43 ± 1.30 In macula 1.82 ± 1.40 In peripapillary area	Labeled leukocytes, SLO
Martin and Roorda (2005)	Leukocyte	1.37 ± 0.29 (0.77–2.10)	Adaptive optics scanning laser ophthalmoscope (AOSLO)
Martin and Roorda (2009)	Leukocyte	1.30 ± 0.40 (0.49–2.31)	AOSLO
Tam and Roorda (2011)	Leukocyte	1.80 ± 0.22 (1.1–2.5)	AOSLO
	Plasma gap	1.30 ± 0.55 (0.4–1.6)	
Tam <i>et al.</i> (2011)	Leukocyte	1.82 ± 0.42	AOSLO
Bedgood and Metha (2012)	Erythrocytes	1.30 (0.3–3.3)	AO fundus imaging
de Castro <i>et al.</i> (2016)	Erythrocytes	1.14 (0.3–2.23)	Dual channel AOSLO

Abbreviation: DR, diabetic retinopathy.

(0.51 and 0.54 mm s⁻¹, respectively), but reported an increase of velocity in patients with background retinopathy (0.74 mm s⁻¹), and a decrease in the preproliferative stage (0.37 mm s⁻¹), to increase again after photocoagulation (0.42 mm s⁻¹). Additionally, the velocity of the leukocytes is sensitive to neural activity, increasing in the presence of flickering background lights. This velocity increase (Scheiner *et al.*, 1994) occurs about 15 s after the onset of the flickering stimulus and under the conditions used increased more than 50% (from 0.60 to 1.01 mm s⁻¹). In general, velocities measured with the blue field entoptoscope are some of the lower velocity values obtained (Table 1), but it is important to note that these are almost certainly leukocytes moving through the capillaries near the foveal avascular zone, which are some of the smallest capillaries, and the slower values may reflect the mechanical constraints on fitting large cells through small capillaries.

Fluorescein Angiography

Fluorescein angiography is a common clinical technique for investigating the health of the retinal vasculature. During fluorescein angiography the fluorescein dye is injected (typically in the arm) and arrives as a dye front at the eye. Because the early phases of the fluorescein angiogram have high contrast it is possible to obtain excellent images of the vascular network. If fluorescein angiograms are collected using video techniques, segments of low and high fluorescence can be observed moving through the capillaries in the parafoveal region of the retina. The segments of low fluorescence are thought to correspond to erythrocytes in rouleaux formation (that is strings of red blood cells closely packed to each other). The cells are not stained by the fluorescein and thus are dark and the more fluorescent segments then correspond to cell-free plasma (Arend *et al.*, 1994). The velocity of the hypofluorescent segments, which are sparse enough to be readily identified, can be determined by measuring

the distance traveled between two frames. Using videos captured with a SLO, an average speed in the perifovea of 3.28 mm s^{-1} was obtained in normal subjects and a lower value (2.89 mm s^{-1}) was observed in diabetics (Arend *et al.*, 1991; Wolf *et al.*, 1991).

While the hypofluorescent segments were assumed to be rouleaux, Yang *et al.* (1997) made measurements on two types of hyperfluorescent signals: hyperfluorescent dots and hyperfluorescent segments. He considered the dots to be leukocytes and measured a speed of 1.37 mm s^{-1} and, similar to Wolf *et al.* (1991), hypothesized that the segments might be the plasma gaps located between erythrocyte rouleaux. The segments moved at a lower velocity than the dots, consistent with the size difference of the particles being tracked. The velocities measured with fluorescein angiography are higher than those obtained with the blue field entoptoscope. This could be because they measure slightly different parts of the retina since with an SLO measurements can be performed at higher eccentricity than when using the blue field entoptic phenomenon (Arend *et al.*, 1995), or perhaps because the visual backgrounds are different leading to a different demand on retinal blood flow. Both techniques provided a background, but that of the SLO is very bright and is actually flickering due to the sequential scan with local intensity varying at approximately 30 frames per second.

Fluorescently Stained Leukocytes

A variant of the fluorescein method involves the use of labeled leukocytes. In this method blood is withdrawn from a subject, leukocytes are separated from the plasma and erythrocytes by centrifugation, and they are stained with fluorescein. When reinjected into the same patient from whom they were extracted the externally stained leukocytes can be observed passing through the capillaries of the retina, providing a very high contrast signal. The velocity can be determined without the risk of aliasing due to the sparse number of stained cells. Thus a single cell can be followed as it changes position over several frames of a video. Yang *et al.* (1997) measured leukocyte velocity using an SLO and measured an average value of 1.41 mm s^{-1} . Paques *et al.* (2000) used a similar method and reported a mean velocity of 1.43 mm s^{-1} in the macula and 1.82 mm s^{-1} in the peripapillary area. This technique is very powerful, and since the subject's own cells are used as a low risk of complications, however, it is not possible in some regions to obtain approvals for its use in humans.

Leukocyte Imaging With AOSLO

Adaptive optics has now enabled imaging at cellular scale in the living human retina (Liang *et al.*, 1997). The first applications to measuring blood flow with adaptive optics used a confocal AOSLO focused on the cone photoreceptors. Under these conditions, dark shadows can be visualized moving across the retina. This motion is most evident in the capillaries close to the fovea. Martin and Roorda (2005) argued that they should be leukocytes and calculated their speed by measuring the distance traveled between several frames giving an average velocity of 1.37 mm s^{-1} , similar to previous studies of leukocytes velocity as discussed above. The velocity of these shadows varies with the cardiac pulse (Martin and Roorda, 2009) with an average minimum speed of 1.0 mm s^{-1} and an average maximum speed of 1.58 mm s^{-1} , resulting in a pulsatility of 0.45. The speed of the leukocytes was also determined by computationally straightening the capillaries and calculating the slope of the changes of intensity along a capillary as a function of time (Tam *et al.*, 2011). In this spatiotemporal plot thick, high contrast and sparse traces and thin, low contrast and dense traces were observed. This pattern of results is consistent with the thick traces being leukocytes and the thin traces corresponding to erythrocytes or plasma gaps between erythrocytes. However, contrary to expectations, the velocity of the leukocytes was found to be higher than the velocity found in plasma gaps (1.80 vs. 1.30 mm s^{-1}) suggesting that either these traces are not a good proxy for erythrocyte velocity or that leukocytes and erythrocytes travel at similar velocities when restricted to single-file capillaries (Tam *et al.*, 2011).

Erythrocyte Imaging With AOSLO

Adaptive optics also allows directly imaging cells in the retinal vessels. If a confocal AOSLO is focused on the small vessels and capillaries, bright moving dots are seen. These can move quite quickly. It is possible to sample these moving cells more densely by momentarily halting the slow direction of the scan. The retinal image at this point becomes a position versus time image, and the cells are seen as diagonal streaks, with the angle of the streak representing the component of the cell's velocity in the direction of the scan (Zhong *et al.*, 2008). This approach works well for vessels ranging in size from about $15\text{--}20 \mu\text{m}$ in diameter upwards, but at the small end is limited by eye movements, since to capture the entire cardiac cycle the scanning line should remain on the vessel for about 1 s, and micro-saccades and ocular drift can move it off. In animals under anesthesia this or similar techniques (Guevara-Torres *et al.*, 2016) work relatively well, and the advent of highly stabilized AOSLO's hold promise for capillary measurements (Vogel *et al.*, 2006; Burns *et al.*, 2007; Yang *et al.*, 2014). In addition to the eye movements there are other disadvantages of this technique in that not all of the cells are easily seen. Visibility of the cells in a confocal imaging mode is highly dependent on the orientation of the cells and thus brightness can vary rapidly. In addition, there are so many cells that great care needs to be taken to avoid aliasing. The contrast issue can be dealt with by imaging using multiply scattered light imaging (Elsner *et al.*, 2000). This form of imaging improves contrast of vasculature and allows red blood cells to be directly imaged, while traveling through the capillaries (Chui *et al.*, 2012). The potential for aliasing has recently been solved by using two image wavelengths and a scanning system to collect images from the two channels with a slight offset between them

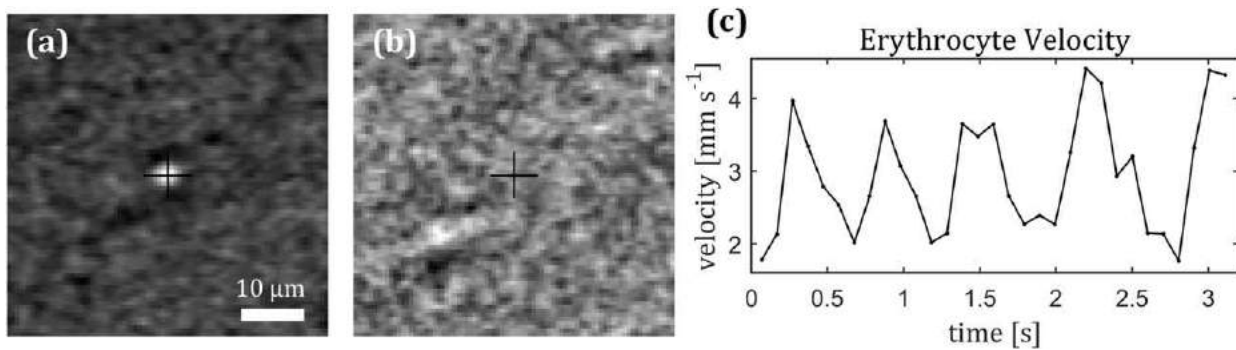


Fig. 2 The erythrocyte position was detected in one of the channels (a) and when studying the same region in the other channel (b) the one with the temporal shift, a displaced image of the same cell was observed. In the figure, the shift of the image of the average cell is used to calculate the velocity at different times (c).

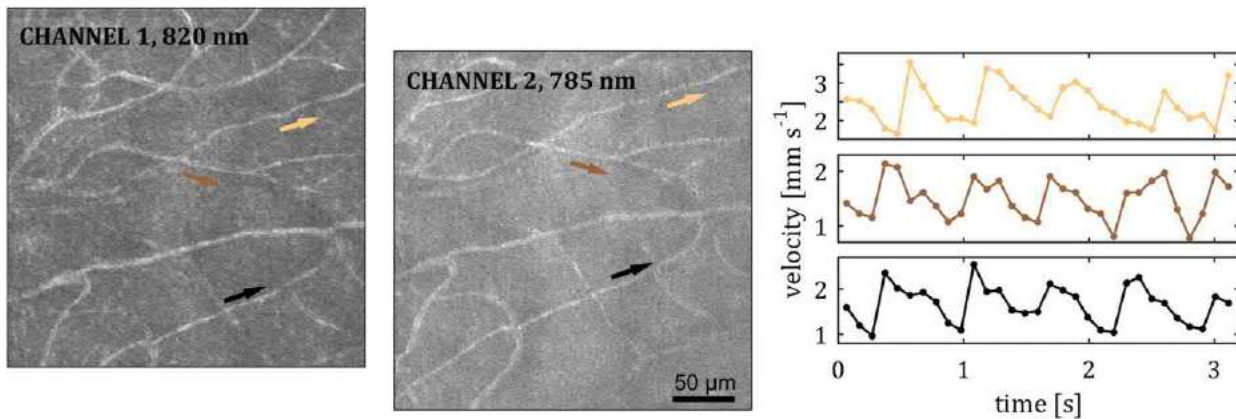


Fig. 3 Using a dual channel scanning laser ophthalmoscope, the velocity of the red blood cells passing through three capillaries was measured simultaneously.

(de Castro *et al.*, 2016). This spatial offset is equivalent to a temporal separation, because the same retinal region is imaged at slightly different times in each channel. By varying the spatial misalignment, the temporal offset can be varied over a wide range of values to match the imaging task. If the displacement is made in a pupil conjugate plane, then it becomes possible to rapidly adjust the temporal separation between channels. Thus by using multiply scattered light imaging and a separation of 71 lines (which corresponds to 4.7 ms, since the frequency of the fast scan was 15.1 kHz), de Castro *et al.* (2016) were able to measure the velocity of the red blood cells in the capillaries of the foveal and parafoveal region (Fig. 2). The velocity measurements were pulsatile, with peak values ranging from 0.70 to 3.98 mm s⁻¹. The average velocities of three different subjects, 10 capillaries each, ranged from 0.30 to 2.23 mm s⁻¹ with a mean value of 1.14 mm s⁻¹, similar to estimates of other methods although lower than the fluorescein methods (Table 1).

To obtain a better signal to noise ratio de Castro *et al.* averaged cell images from three consecutive frames, i.e., one measurement every 107 ms. They were able, under good conditions to determine the velocity at the frame rate of the imaging system, 36 ms. The technique also allowed simultaneous measurements of multiple capillaries within a single field of view, typically between 1.2 and 1.8 degrees. One striking feature of the measurements is that different capillaries within the same video and of the same size have different red blood cell velocities (Fig. 3).

Erythrocyte Imaging With Flood Illuminated Adaptive Optics. Particle Image Velocimetry

Adaptive optics can also be used in non-scanning system. Using a flood illuminated adaptive optics ophthalmoscope with a high temporal resolution camera, Bedggood and Metha (2012) were successful in imaging erythrocytes traveling through the capillaries and velocities could be determined with particle image velocimetry methods. To achieve a frame rate of 460 frames per second, the vertical extent of the imaged regions was reduced to 70 μm and images were captured during sequences of 130 ms (60 frames). They found a wide range of velocities (from 0.3 to 3.3 mm s⁻¹, mean 1.3 mm s⁻¹) and a general increase in velocity during the captured time sequences. They reported that the same capillary segments showed different velocities when

imaged several times and also that, within the same sequence, while some capillary segments increased their velocity, others did not. The authors argued that the reason for the increase in velocity during imaging was not neurovascular coupling, since the reaction time in capillaries in animal models as well as the human eye is typically longer (Scheiner *et al.*, 1994) than their imaging time. The advantage of the flood illuminated approach is that image rates can be very high, and corrections for timing arising from the nature of scanning systems are not needed. However, image quality is not as good as with the AOSLO methods due to the lower contrast, and the applicability of this approach to image blood cells in clinical or older populations is not yet known.

Discussion

The literature on blood velocity and blood flow in the eye is very large, and as we have shown, interpretation requires understanding not only the measurement technology, but the biophysical properties of blood flow within different size blood vessels and indeed under different physiological and pathological conditions. However, because flow measurements can be combined with other optical measurements such as oximetry (Delori, 1988) to determine metabolic activity (Blair *et al.*, 2016) the promise to use these techniques for understanding both the normal and abnormal properties of the CNS is great. It is also important to note that most of the techniques have different advantages and disadvantages, and so one particular method may not be ideal to answer all questions.

The different techniques do seem to give different estimates of velocity, and some of these differences may reflect the different components of the blood being used to measure velocity as well as where in the retina they are measured. However, many of the techniques use leukocytes, and while these can be expected to move more slowly through capillaries, since they are much larger, the values found in erythrocytes (Bedggood and Metha, 2012; de Castro *et al.*, 2016) are comparable to those measured on leukocytes using the entoptic phenomenon (Kato, 1952; Riva and Petrig, 1980; Fallon *et al.*, 1986; Grunwald *et al.*, 1993), labeling with fluorescein (Yang *et al.*, 1997; Paques *et al.*, 2000), or direct imaging with AOSLO (Martin and Roorda, 2005, 2009; Tam and Roorda, 2011). However, the peak velocities found with higher temporal resolution using the dual channel AOSLO (de Castro *et al.*, 2016) are only comparable to those obtained injecting intravenous fluorescein (Arend *et al.*, 1991, 1995; Wolf *et al.*, 1991).

Finally, it is clear that there is a lot of variability in the measurements. Table 1 summarizes measurements of capillary blood velocity. And within any cell class the variability in the velocity measurements is large and this may reflect more than measurement variability, since different capillaries have different speeds and composition of cells (Tam *et al.*, 2011; de Castro *et al.*, 2016; Guevara-Torres *et al.*, 2016).

This natural variation may place a fundamental limit on the simple application of these techniques to clinical measurements, since it will be necessary to obtain sufficient sample sizes to understand the cause of differences. However, it should not be a problem when testing the response of the retina to different conditions within an eye (such as exercise or visual stimulation) since in this case the response of several individual vessels can be compared. Finally, it is important to note that the eye now provides a tissue in which it is possible to measure the entire vascular tree ranging from vessels larger than 100 μm to the smallest capillaries. This will allow improved understanding of the biophysical properties of the entire vascular network and how it responds to stress and disease.

References

- Arend, O., Harris, A., Sponsel, W.E., *et al.*, 1995. Macular capillary particle velocities: A blue field and scanning laser comparison. *Graefes Archive for Clinical and Experimental Ophthalmology* 233, 244–249.
- Arend, O., Harris, A., Wolf, S., 1994. Capillary blood flow velocity measurements in cystoid macular edema with the scanning laser ophthalmoscope. *American Journal of Ophthalmology* 117, 819–820.
- Arend, O., Wolf, S., Jung, F., *et al.*, 1991. Retinal microcirculation in patients with diabetes mellitus: Dynamic and morphological analysis of perifoveal capillary network. *British Journal of Ophthalmology* 75, 514–518.
- Aschinger, G.C., Schmetterer, L., Doblhoff-Dier, V., *et al.*, 2015. Blood flow velocity vector field reconstruction from dual-beam bidirectional doppler oct measurements in retinal veins. *Biomedical Optics Express* 6, 1599–1615.
- Baumann, B., Potsaid, B., Kraus, M.F., *et al.*, 2011. Total retinal blood flow measurement with ultrahigh speed swept source/Fourier domain oct. *Biomedical Optics Express* 2, 1539–1552.
- Bedggood, P., Metha, A., 2012. Direct visualization and characterization of erythrocyte flow in human retinal capillaries. *Biomedical Optics Express* 3, 3264–3277.
- Bishop, J.J., Nance, P.R., Popel, A.S., Intaglietta, M., Johnson, P.C., 2001. Effect of erythrocyte aggregation on velocity profiles in venules. *American Journal of Physiology* 280, H222–H236.
- Blair, N.P., Wanek, J., Felder, A.E., *et al.*, 2016. Inner retinal oxygen delivery, metabolism, and extraction fraction in Ins2akita diabetic mice. *Investigative Ophthalmology & Visual Science* 57, 5903–5909.
- Burns, S.A., Tumber, R., Elsner, A.E., Ferguson, D., Hammer, D.X., 2007. Large-field-of-view, modular, stabilized, adaptive-optics-based scanning laser ophthalmoscope. *Journal of the Optical Society of America A, Optics, Image Science, and Vision* 24, 1313–1326.
- Chen, Z.P., Milner, T.E., Srinivas, S., *et al.*, 1997. Noninvasive imaging of in vivo blood flow velocity using optical doppler tomography. *Optics Letters* 22, 1119–1121.
- Chui, T.Y.P., Vannasdale, D.A., Burns, S.A., 2012. The use of forward scatter to improve retinal vascular imaging with an adaptive optics scanning laser ophthalmoscope. *Biomedical Optics Express* 3, 2537–2549.
- Dai, C.X., Liu, X.J., Zhang, H.F., Puliafito, C.A., Jiao, S.L., 2013. Absolute retinal blood flow measurement with a dual-beam doppler optical coherence tomography. *Investigative Ophthalmology & Visual Science* 54, 7998–8003.
- Deak, G.G., Schmidt-Erfurth, U., 2013. Imaging of the parafoveal capillary network in diabetes. *Current Diabetes Reports* 13, 469–475.

- de Castro, A., Huang, G., Sawides, L., Luo, T., Burns, S.A., 2016. Rapid high resolution imaging with a dual-channel scanning technique. *Optics Letters* 41, 1881–1884.
- Delori, F.C., 1988. Noninvasive technique for oximetry of blood in retinal vessels. *Applied Optics* 27, 1113.
- Doblhoff-Dier, V., Schmetterer, L., Vilser, W., *et al.*, 2014. Measurement of the total retinal blood flow using dual beam Fourier-domain doppler optical coherence tomography with orthogonal detection planes. *Biomedical Optics Express* 5, 630–642.
- Elsner, A.E., Miura, M., Burns, S.A., *et al.*, 2000. Multiply scattered light tomography and confocal imaging: Detecting neovascularization in age-related macular degeneration. *Optics Express* 7, 95–106.
- Fallon, T.J., Chowienczyk, P., Kohner, E.M., 1986. Measurement of retinal blood flow in diabetes by the blue-light entoptic phenomenon. *British Journal of Ophthalmology* 70, 43–46.
- Fekke, G.T., Riva, C.E., 1978. Laser doppler measurements of blood velocity in human retinal-vessels. *Journal of the Optical Society of America* 68, 526–531.
- Fercher, A.F., Briers, J.D., 1981. Flow visualization by means of single-exposure speckle photography. *Optics Communications* 37, 326–330.
- Fercher, A.F., Hu, H.Z., Vry, U., 1985. Rough surface interferometry with a two-wavelength heterodyne speckle interferometer. *Applied Optics* 24, 2181–2188.
- Ferguson, R.D., Hammer, D.X., Elsner, A.E., Webb, R.H., Burns, S.A., 2004. Wide-field retinal hemodynamic imaging with the tracking scanning laser ophthalmoscope. *Optics Express* 12, 5198–5208.
- Garhöfer, G., Zawinka, C., Resch, H., *et al.*, 2004. Reduced response of retinal vessel diameters to flicker stimulation in patients with diabetes. *British Journal of Ophthalmology* 88, 887–891.
- Grunwald, J.E., Brucker, A.J., Schwartz, S.S., *et al.*, 1990. Diabetic glycemic control and retinal blood-flow. *Diabetes* 39, 602–607.
- Grunwald, J.E., Piltz, J., Patel, N., Bose, S., Riva, C.E., 1993. Effect of aging on retinal macular microcirculation: A blue field simulation study. *Investigative Ophthalmology & Visual Science* 34, 3609–3613.
- Grunwald, J.E., Riva, C.E., Sinclair, S.H., Brucker, A.J., Petrig, B.L., 1986. Laser doppler velocimetry study of retinal circulation in diabetes mellitus. *Archives of Ophthalmology* 104, 991–996.
- Guevara-Torres, A., Joseph, A., Schallek, J.B., 2016. Label free measurement of retinal blood cell flux, velocity, hematocrit and capillary width in the living mouse eye. *Biomedical Optics Express* 7, 4228–4249.
- Jung, F., Kiewewetter, H., Körber, N., *et al.*, 1983. Quantification of characteristic blood-flow parameters in the vessels of the retina with a picture analysis system for video-fluorescence angiograms: Initial findings. *Graefes Archive for Clinical and Experimental Ophthalmology* 221, 133–136.
- Kato, K., 1952. Studies on the velocity of the blood stream in the retinal capillaries. *The Keio Journal of Medicine* 1, 251–268.
- Kohner, E.M., Patel, V., Rassam, S.M.B., 1995. Role of blood-flow and impaired autoregulation in the pathogenesis of diabetic-retinopathy. *Diabetes* 44, 603–607.
- Kur, J., Newman, E.A., Chan-Ling, T., 2012. Cellular and physiological mechanisms underlying blood flow regulation in the retina and choroid in health and disease. *Progress in Retinal and Eye Research* 31, 377–406.
- Leitgeb, R.A., Schmetterer, L., Hitzenberger, C.K., *et al.*, 2004. Real-time measurement of in vitro flow by Fourier-domain color doppler optical coherence tomography. *Optics Letters* 29, 171–173.
- Liang, J., Williams, D.R., Miller, D.T., 1997. Supernormal vision and high-resolution retinal imaging through adaptive optics. *Journal of the Optical Society of America A* 14, 2884–2892.
- Logean, E., Schmetterer, L., Riva, C.E., 2003. Velocity profile of red blood cells in human retinal vessels by confocal scanning laser Doppler velocimetry. *Laser Physics* 13, 45–51.
- Makita, S., Hong, Y., Yamanari, M., Yatagai, T., Yasuno, Y., 2006. Optical coherence angiography. *Optics Express* 14, 7821–7840.
- Martin, J.A., Roorda, A., 2005. Direct and noninvasive assessment of parafoveal capillary leukocyte velocity. *Ophthalmology* 112, 2219–2224.
- Martin, J.A., Roorda, A., 2009. Pulsatility of parafoveal capillary leukocytes. *Experimental Eye Research* 88, 356–360.
- Paques, M., Boval, B., Richard, S., *et al.*, 2000. Evaluation of fluorescein-labeled autologous leukocytes for examination of retinal circulation in humans. *Current Eye Research* 21, 560–565.
- Pournaras, C.J., Rungger-Brändle, E., Riva, C.E., Hardarson, S.H., Stefansson, E., 2008. Regulation of retinal blood flow in health and disease. *Progress in Retinal and Eye Research* 27, 284–330.
- Riva, C., Benedek, G.B., Ross, B., 1972. Laser Doppler measurements of blood-flow in capillary tubes and retinal arteries. *Investigative Ophthalmology* 11, 936.
- Riva, C.E., Falsini, B., Logean, E., 2001. Flicker-evoked responses of human optic nerve head blood flow: Luminance versus chromatic modulation. *Investigative Ophthalmology & Visual Science* 42, 756–762.
- Riva, C.E., Fekke, G.T., Eberli, B., Benary, V., 1979. Bidirectional LDV system for absolute measurement of blood speed in retinal vessels. *Applied Optics* 18, 2301–2306.
- Riva, C.E., Harino, S., Petrig, B.L., Shonat, R.D., 1992. Laser Doppler flowmetry in the optic-nerve. *Experimental Eye Research* 55, 499–506.
- Riva, C.E., Logean, E., Falsini, B., 2005. Visually evoked hemodynamical response and assessment of neurovascular coupling in the optic nerve and retina. *Progress in Retinal and Eye Research* 24, 183–215.
- Riva, C.E., Petrig, B., 1980. Blue field entoptic phenomenon and blood velocity in the retinal capillaries. *Journal of the Optical Society of America* 70, 1234–1238.
- Scheiner, A.J., Riva, C.E., Kazahaya, K., Petrig, B.L., 1994. Effect of flicker on macular blood flow assessed by the blue field simulation technique. *Investigative Ophthalmology & Visual Science* 35, 3436–3441.
- Snodderly, D.M., Weinhaus, R.S., Choi, J.C., 1992. Neural vascular relationships in central retina of macaque monkeys (*Macaca fascicularis*). *Journal of Neuroscience* 12, 1169–1193.
- Suzuki, Y., Masuda, K., Ogino, K., *et al.*, 1991. Measurement of blood-flow velocity in retinal-vessels utilizing laser speckle phenomenon. *Japanese Journal of Ophthalmology* 35, 4–15.
- Tam, J., Roorda, A., 2011. Speed quantification and tracking of moving objects in adaptive optics scanning laser ophthalmoscopy. *Journal of Biomedical Optics* 16, 036002.
- Tam, J., Tiruveedhula, P., Roorda, A., 2011. Characterization of single-file flow through human retinal parafoveal capillaries using an adaptive optics scanning laser ophthalmoscope. *Biomedical Optics Express* 2, 781–793.
- Tokayer, J., Jia, Y., Dhalla, A.H., Huang, D., 2013. Blood flow velocity quantification using split-spectrum amplitude-decorrelation angiography with optical coherence tomography. *Biomedical Optics Express* 4, 1909–1924.
- Villringer, A., Dirnagl, U., 1995. Coupling of brain activity and cerebral blood-flow – Basis of functional neuroimaging. *Cerebrovascular and Brain Metabolism Reviews* 7, 240–276.
- Vogel, C.R., Arathorn, D.W., Roorda, A., Parker, A., 2006. Retinal motion estimation in adaptive optics scanning laser ophthalmoscopy. *Optics Express* 14, 487–497.
- Wang, R.K., Jacques, S.L., Ma, Z., *et al.*, 2007. Three dimensional optical angiography. *Optics Express* 15, 4083–4097.
- Wessing, A., 1974. Fluorescence-television-angiography of retina and anterior segments of eye with highly sensitive television photographic valves. *Klinische Monatsblätter Für Augenheilkunde* 165, 817–822.
- Westphal, V., Yazdanfar, S., Rollins, A.M., Izatt, J.A., 2002. Real-time, high velocity-resolution color doppler optical coherence tomography. *Optics Letters* 27, 34–36.
- White, B.R., Pierce, M.C., Nassif, N., *et al.*, 2003. In vivo dynamic human retinal blood flow imaging using ultra-high-speed spectral domain optical doppler tomography. *Optics Express* 11, 3490–3497.
- Willerslev, A., Li, X.Q., Munch, I.C., Larsen, M., 2014. Flow patterns on spectral-domain optical coherence tomography reveal flow directions at retinal vessel bifurcations. *Acta Ophthalmologica* 92, 461–464.
- Wolf, S., Arend, O., Toonen, H., *et al.*, 1991. Retinal capillary blood flow measurement with a scanning laser ophthalmoscope. Preliminary Results. *Ophthalmology* 98, 996–1000.

- Wolf, S., Jung, F., Kieseewetter, H., Korber, N., Reim, M., 1989. Video fluorescein angiography – Method and clinical-application. *Graefes Archive for Clinical and Experimental Ophthalmology* 227, 145–151.
- Wolf, S., Toonen, H., Arend, O., *et al.*, 1990. Retinal capillary bloodflow measurement by means of a scanning laser ophthalmoscope. *Biomedizinische Technik* 35, 131–134.
- Yang, Q., Zhang, J., Nozato, K., *et al.*, 2014. Closed-loop optical stabilization and digital image registration in adaptive optics scanning light ophthalmoscopy. *Biomedical Optics Express* 5, 3174–3191.
- Yang, Y., Kim, S., Kim, J., 1997. Fluorescent dots in fluorescein angiography and fluorescein leukocyte angiography using a scanning laser ophthalmoscope in humans. *Ophthalmology* 104, 1670–1676.
- Zhong, Z., Huang, G., Chui, T.Y.P., Petrig, B.L., Burns, S.A., 2012. Local flicker stimulation evokes local retinal blood velocity changes. *Journal of Vision* 12, 3.
- Zhong, Z., Petrig, B.L., Qi, X., Burns, S.A., 2008. In vivo measurement of erythrocyte velocity and retinal blood flow using adaptive optics scanning laser ophthalmoscopy. *Optics Express* 16, 12746–12756.
- Zhong, Z., Song, H., Chui, T.Y.P., Petrig, B.L., Burns, S.A., 2011. Noninvasive measurements and analysis of blood velocity profiles in human retinal vessels. *Investigative Ophthalmology & Visual Science* 52, 4151–4157.

Adaptive Optics Retinal Imaging Techniques and Clinical Applications

Jessica IW Morgan, University of Pennsylvania, Philadelphia, PA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Ophthalmoscopy is a fundamental part of a fundus exam, because it allows observation of the retina through the natural optics of the eye. Since the first fundus photograph was taken in 1886 (Jackman and Webster, 1886), numerous modalities for ocular imaging have been developed. However, the overarching goal of these imaging techniques remains stationary: to understand the structure and function of the normal visual system and to assist in the diagnosis, prognosis, and treatment of the diseased visual system. The purpose of this current article is to describe state-of-the-art adaptive optics (AO) high-resolution ophthalmic imaging, the detection schemes used in different AO imaging modalities, the current analysis methods utilized for interpreting high-resolution AO images of the outer retina, and a brief discussion of the current and future clinical applications for this technology.

Optical Limitations for Imaging in the Eye

One can think of retinal imaging in the same terms as imaging with a microscope, where the retina is the sample to be imaged and the optical components of the eye provide the final objective lens in front of the sample. All optical systems, including the eye, are limited in resolution by their point spread function (psf), which is defined as the volume of space over which a focused point of light extends. Therefore the resolution of a retinal image, regardless of which imaging modality or detection scheme is employed, will be limited by the size of the eye's psf. The eye's psf in turn is dependent upon the quality of the optical components of the eye (the cornea and lens) as well as the limiting aperture of the eye (the pupil). For an ideal, aberration-free eye, the psf would be determined by diffraction through the pupil. In the diffraction limit, the spatial extent of the psf decreases with increasing pupil diameter, leading to the resolution limit:

$$l_{\text{transverse}} = \frac{1.22f\lambda_0}{D}$$

where f is the focal length of the eye, λ is the central imaging wavelength, and D is the pupil diameter. Unfortunately, the cornea and lens of the real eye are not ideal optical elements; these elements impose optical aberrations on light entering (or reflected out of) the eye, and thereby increase the size of the psf beyond that of the diffraction limit. For real eyes with pupil diameters that are 2 mm or smaller, the size of the psf is dominated mainly by diffraction. However, for pupil diameters larger than 2 mm, ocular aberrations become the dominant factor for determining the extent of the psf (Packer and Williams, 2003). In theory, eliminating the optical aberrations imposed on light entering or exiting the eye by the cornea and lens would reduce the size of the psf to that observed in the diffraction limiting case, even at large pupil sizes. Optically compensating for the eye's optical aberrations thus becomes the main concept behind high-resolution AO ophthalmoscopy.

AO

AO is a technique first proposed by astronomers to measure and correct for optical aberrations imposed by turbulence in earth's atmosphere (Babcock, 1953). Incorporating this technique into ground-based telescopes resulted in improvements in transverse resolution and contrast of astrological images. The same concept was applied to ophthalmoscopy by Liang *et al.* (1997). The goal of AO ophthalmoscopy is straightforward: first measure, then correct for the optical aberrations of the living eye in order to attain diffraction limited resolution through large pupil sizes.

Ophthalmic AO systems typically use a Shack–Hartman wavefront sensor to measure the optical aberrations of the eye. This consists of a lenslet array and a two-dimensional (2D) camera. The lenslet array is placed at a plane optically conjugate with the eye's pupil. Each lenslet of the array focuses a small section of the wavefront to a single spot on the camera. For an ideal, aberration-free optical system, the result is a perfectly spaced 2D grid of focused spots, called the reference. For an aberrated system, such as the real eye, the arrangement of focused spots will be distorted from the perfectly spaced reference grid, and the x and y displacement of each focused spot from its ideal reference location allows a direct calculation of the wave aberration at the location of the lenslet within the pupil plane (Liang *et al.*, 1994).

After the wavefront sensor, the second critical component for an AO system is the wavefront corrector. This optical element also is placed optically conjugate with the eye's pupil and therefore the lenslet array of the Shack–Hartman wavefront sensor. Ophthalmic AO systems typically use a deformable mirror for this purpose, but other techniques, such as liquid crystal spatial light modulators, also can be used. The shape of the deformable mirror is adjusted to be one half of the amplitude of the wave

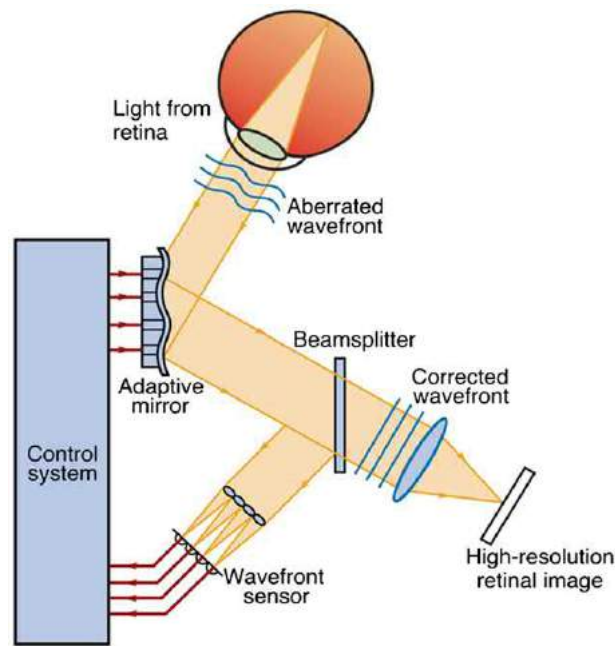


Fig. 1 Schematic diagram of wavefront measurement and adaptive optics (AO) correction. Reprinted from Carroll, J., Gray, D.C., Roorda, A., Williams, D.R., 2005. Recent advances in retinal imaging with adaptive optics. *Optics and Photonics News* 16, 36–42 with permission from Joseph Carroll.

aberration at each point in the pupil, such that the imposed phase delay at each point in the pupil results in a (close-to) aberration-free wavefront that when measured using the Shack–Hartman wavefront sensor approaches the reference grid of focus spots. This AO correction yields an aberration corrected psf for the eye, which is thus close to that of the diffraction limit, even for dilated pupils. A schematic diagram of wavefront measurement and AO correction is shown in [Fig. 1](#) ([Carroll et al., 2005](#)). Most state-of-the-art AO systems run in closed loop with real-time AO correction occurring multiple times per second, as has been shown to provide improved AO quality in comparison with open loop systems ([Hofer et al., 2001](#)). Retinal imaging with a visible or near infrared light source (400–900 nm) incorporating AO correction over a dilated pupil (7 mm) provides sufficient transverse resolution for observing individual cells in the living retina, noninvasively.

AO Retinal Imaging

It is important to note that AO on its own is not a method of retinal imaging. Rather, AO is a tool used to increase the transverse resolution of a given imaging modality by correcting the optical aberrations of the eye. To obtain an AO retinal image, AO must be incorporated with technologies that encompass retinal image acquisition techniques. Here we describe three such retinal imaging modalities, AO flood-illuminated ophthalmoscopy, AO scanning laser/light ophthalmoscopy (AOSLO), and AO-optical coherence tomography (AO-OCT), and discuss applications of these three imaging techniques for studying the structure and function of the normal and diseased retina.

AO Flood-Illuminated Ophthalmoscopy

The first application of AO to ophthalmoscopy occurred when Liang, Williams, and Miller incorporated AO correction into a flood-illuminated fundus camera ([Liang et al., 1997](#)). A fundus camera system involves using a light source to uniformly illuminate a patch of retina and a 2D science camera to record the backscattered light from that same retinal patch. Considerations for imaging include ensuring investigators use only non-phototoxic (ocularly safe) amounts of light to illuminate the retina and a camera sensitive enough to capture the low reflective signal returning back from the retina and through the pupil of the eye. The first report of AO corrected fundus photography showed improvement in both contrast and resolution in images of the cone photoreceptors in comparison to non-AO corrected images of the same retinal patch. Since this early work, improvements in both light sources and the speed of digital cameras have made possible high frame acquisition rates using AO corrected flood-illumination retinal imaging.

An important consideration for flood-illuminated retinal imaging systems is the image exposure time. One property of flood-illuminated systems is that all pixels within the 2D image are acquired at the same time. This is an advantage in that the retinal

image incorporates few distortions and thus the spatial relationship between features in the image represents true retinal anatomy. On the other hand, the eye is continually undergoing fixational eye motion, and thus, any eye motion that occurs during the image exposure/acquisition time will result in a blurring of retinal features. For imaging microscopic features, such as the individual cones in the photoreceptor mosaic, even small eye movements can result in blurring out the contrast between neighboring cones. By limiting the exposure time of a single image to less than 10 msec, [Rha et al. \(2006\)](#) found most retinal images exhibited cone mosaic structure, and thus concluded that blurring from fixational eye motion was minimal when using this short exposure duration. Early applications of AO flood-illumination ophthalmoscopy demonstrated the arrangement and variability of the three cone classes in the trichromatic cone mosaic ([Roorda and Williams, 1999](#)) and loss of waveguiding cones in a dichromat ([Carroll et al., 2004](#)) and in patients with inherited retinal degenerations ([Wolfing et al., 2006](#)). In addition, AO flood-illumination ophthalmoscopy has been used to examine the reflectance of cones over time both in the absence and presence of visual stimuli ([Pallikaris et al., 2003](#); [Jonnal et al., 2007](#)).

AO Scanning Laser/Light Ophthalmoscopy

Scanning laser/light ophthalmoscopy (SLO) was the second retinal imaging modality to incorporate AO correction ([Roorda et al., 2002](#)). SLO operates by focusing a point source of light (a laser or super-luminescent diode) to the retina, and using a point detector, such as a photomultiplier tube (PMT), to detect the backscattered (or emitted) light from that point. The point source on the retina is then repeatedly raster scanned in 2D, and the time-sync between the scanners and the point detector is used to construct a 2D image sequence based on the x, y position of the scanned beam on the retina. A feature of this image collection paradigm is that each pixel of the image sequence is obtained at a different point in time, yielding real-time movies of the retina, but including in the raw data all eye movements. Even when asked to hold the eye perfectly still, subjects experience involuntary eye movements; this involuntary motion exists not only between frames of the image sequence (inter-frame motion), but also within frames of the image sequence (intra-frame motion). Intra-frame motion complicates the registration and averaging of multiple frames together, but results in precise measurements of fixational eye motion ([Sheehy et al., 2015](#)). Again, the transverse resolution of an SLO is limited by the eye's psf, which is improved when using AO aberration correction.

Most AOSLO imaging systems place a pinhole in front of the photodetector that is confocal with the retinal layer of interest in order to reduce the amount of backscattered light detected from retinal layers other than the plane of focus. This results in improved contrast in the retinal image in comparison with flood-illuminated AO systems. Using a confocal pinhole also provides coarse axial sectioning through the retinal layers. While the psf provides the transverse resolution, the size of the confocal pinhole provides the axial resolution for an AOSLO system. Close-to diffraction limited systems with just over 1 Airy disk-sized confocal pinholes have demonstrated axial resolutions of $\sim 50 \mu\text{m}$ ([Venkateswaran et al., 2004](#)). The first application of confocal AOSLO by [Roorda et al. \(2002\)](#) demonstrated the coarse axial sectioning obtained by imaging waveguiding photoreceptor reflectance, blood flow through the retinal capillaries, and reflectance of the nerve fiber layer at the same retinal location. Confocal AOSLO imaging is the most often applied AO imaging modality to date. Among numerous other applications, this imaging paradigm has enabled routine observation of the cone and rod photoreceptor waveguiding reflectance mosaic in normal sighted individuals ([Fig. 2](#)) and

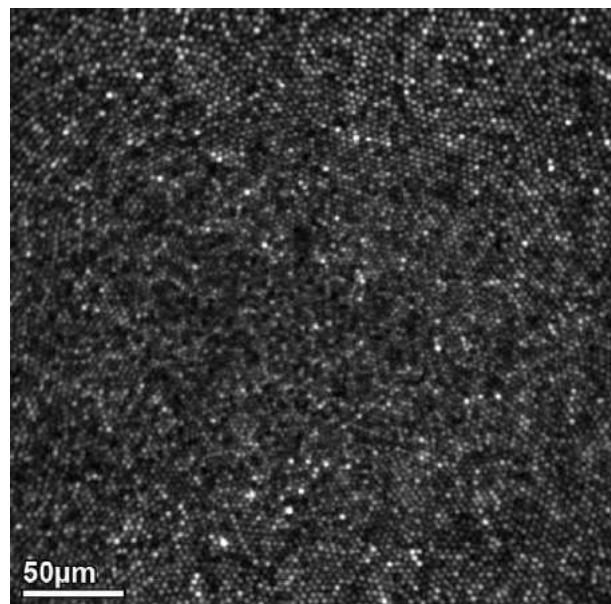


Fig. 2 Confocal adaptive optics scanning laser/light ophthalmoscopy (AOSLO) image of the foveal cone mosaic from a normal sighted control.

in patients with retinal disease, visualization of the parafoveal capillaries, the retinal vascular network and blood flow in normal sighted individuals and in patients with diabetes, and examination of the lamina cribrosa (for a recent review describing AOSLO image findings in disease, see [Morgan, 2016](#)).

Again, the AOSLO is analogous to a scanning microscope, so just as in microscopy, one can manipulate the illumination and the detection beams in the system to attain different views of the retinal structure other than those obtained from confocal reflectance. One demonstration of this came from combining autofluorescence imaging with AOSLO to visualize the retinal pigment epithelium (RPE) in human and nonhuman primates ([Morgan *et al.*, 2009](#)). The RPE cytoplasm contains inherently fluorescent lipofuscin, a byproduct of the natural processes of phototransduction and phagocytosis. Lipofuscin granules consist of a conglomeration of multiple molecules that when excited with a short wavelength in the visible spectrum (blue/green) emit with a longer visible wavelength (red). Unfortunately, the autofluorescence signal from lipofuscin is low – too low to observe cellular structure in an individual autofluorescence AOSLO frame and too low for using the autofluorescence signal to register multiple frames together. Safety limits for ocular light exposures prevent simply increasing the excitation irradiance until a usable signal is obtained. To overcome these problems, [Morgan *et al.* \(2009\)](#) used simultaneous autofluorescence and confocal reflectance imaging to obtain two temporally and spatially synced retinal images, one with the low signal to noise autofluorescence signal and the other with the high signal to noise confocal reflectance signal. The high signal to noise reflectance images were used to determine the fixational eye motion present from frame to frame, and thus the lateral translation required to stabilize the image sequence. Since the autofluorescence image was acquired simultaneously, the fixational eye motion present in the autofluorescence image sequence was the same as the fixational eye motion present in the reflectance image sequence. Morgan *et al.* then applied the same lateral translations needed to stabilize the reflectance image to the autofluorescence image sequence and integrated the autofluorescence signal over the full image series. This novel image processing technique enabled visualization of the RPE cell mosaic in living humans ([Fig. 3](#)) and nonhuman primates and is routinely used to co-register simultaneously captured image sequences regardless of signal strength.

Autofluorescence imaging of the RPE lipofuscin was a natural extension from confocal AOSLO reflectance imaging because of the intrinsic contrast provided by the autofluorescent material in the retina. However, fluorescent imaging in the retina is not limited to autofluorescence of lipofuscin, but rather can be extended to other fluorescent markers as well. The general technique is to use the appropriate excitation and emission wavelengths for the fluorophore of interest, and attain a signal with enough spatial contrast to allow observation of the feature of interest in the final image. Fluorescence imaging with AOSLO has been demonstrated using both fluorescein ([Chui *et al.*, 2014](#)) and indocyanine green (ICG), both of which are fluorescent dyes approved for injection to the bloodstream, and which are generally used in ophthalmic imaging to elucidate abnormalities in the retinal and choroidal blood vessels. ICG has also been shown to provide additional contrast for viewing RPE cells ([Tam *et al.*, 2016](#)). Other fluorophores in the retina can provide the needed signal and contrast; fluorescence of cells tagged with fluorescent markers have been demonstrated in nonhuman primates and rodents ([Geng *et al.*, 2012](#)), two photon fluorescence has been demonstrated in nonhuman primates ([Sharma *et al.*, 2016](#)), and single photon intrinsic infrared autofluorescence of melanin in humans and other species holds promise. Regardless of the fluorophore being excited, successful AOSLO (or AO flood illumination) fluorescence imaging requires: (1) the presence of a fluorophore with spatial contrast; (2) an excitation and emission spectra that will pass

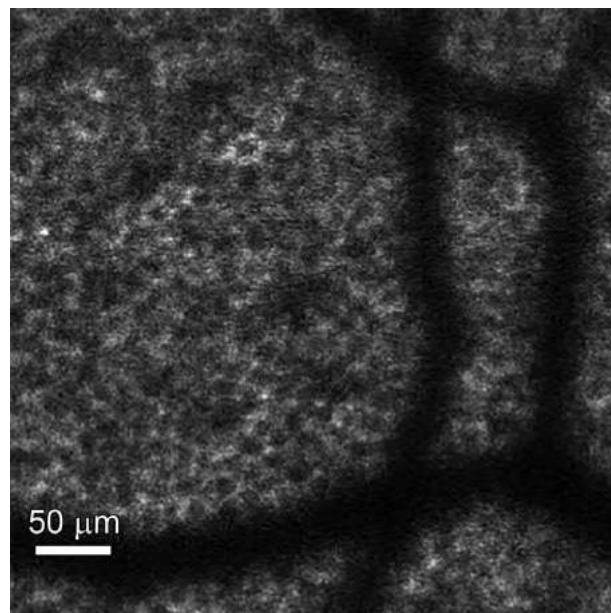


Fig. 3 Autofluorescence image of the retinal pigment epithelium (RPE) mosaic at approximately 20 degrees superior to fixation in a normal sighted control.

through the optics of the eye; (3) safe (non-phototoxic) excitation; and (4) sufficient quantity of emitted signal which is captured by the detector.

For more than the first decade following the invention of AOSLO, both reflectance and fluorescence imaging was done using a confocal pinhole in front of the point detector as described above. Then, in 2012, [Chui et al. \(2012\)](#) published a study in which the authors offset the pinhole from its central position, effectively blocking confocal reflectance and capturing a portion of the non-confocal multiply scattered light from the retina. This study demonstrated the advantage of this method for imaging retinal vasculature, and became the first of a series of imaging innovations using non-confocal AOSLO detection techniques ([Fig. 4](#)), each of which showed distinct advantages for elucidating various retinal structures. The second non-confocal AOSLO detection technique described, termed dark-field ([Scoles et al., 2013](#)), involved excluding the light that would normally pass through the confocal pinhole and detecting the light passing through an annulus surrounding the blocked pinhole. This technique captured the non-confocal multiply scattered light in the retina, and allowed visualization of the RPE cells in normal retina at some locations without using fluorescence techniques ([Fig. 5](#)). Non-confocal “split detection” shortly followed in which the dark-field signal was split in half using a knife edge and two images (one from each side of the knife edge) were captured simultaneously using two separate detectors. The two images were then recombined using the following equation:

$$I_{x,y,\text{split}} = \frac{I_{x,y,\text{det}_1} - I_{x,y,\text{det}_2}}{I_{x,y,\text{det}_1} + I_{x,y,\text{det}_2}} + 128$$

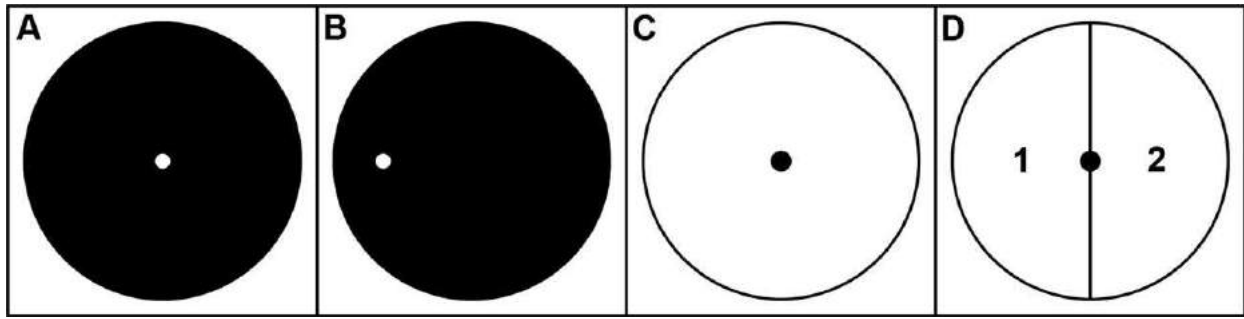


Fig. 4 Adaptive optics scanning laser/light ophthalmoscopy (AOSLO) detection schemes showing the spatial extent of the retinal plane over which light is blocked from (black) or captured by (white) the point detector. (A) Confocal. (B) Offset pinhole. (C) Dark-field. (D) Non-confocal split detection, where two point detectors are used and the two images are recombined using their difference divided by their sum to form the non-confocal split detection image.

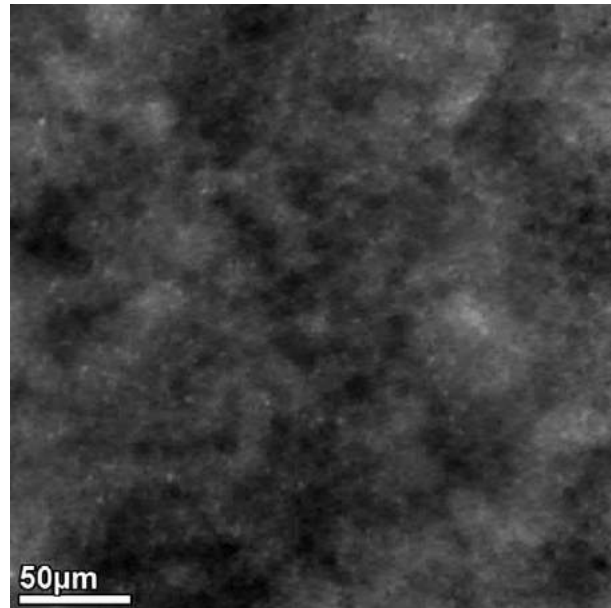


Fig. 5 Dark-field adaptive optics scanning laser/light ophthalmoscopy (AOSLO) image of the retinal pigment epithelium (RPE) mosaic from a normal sighted control.

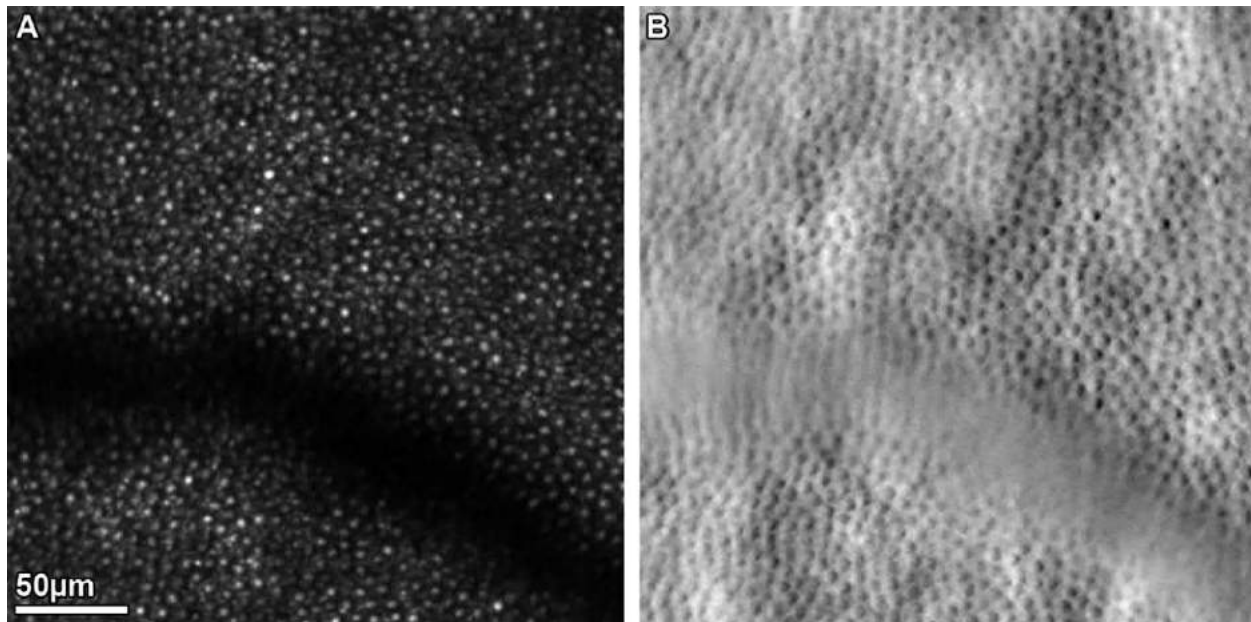


Fig. 6 Confocal (A) and non-confocal split detection (B) adaptive optics scanning laser/light ophthalmoscopy (AOSLO) image of the photoreceptor mosaic from a normal sighted control.

where $I_{x,y}$ is the image intensity at each pixel in the images from the two non-confocal detectors. This detection scheme resulted in the first in vivo images of the photoreceptor inner segment mosaic (Fig. 6; Scoles *et al.*, 2014). The technique also showed improvements in contrast of vasculature, with greater improvements visible in blood vessels running parallel to the direction of the knife edge. Noteworthy is that the denominator of the split detection equation above is simply the dark-field image. Thus one significant advantage of combining non-confocal split detection with confocal AOSLO imaging is that by using three separate point detectors, the confocal, non-confocal split detection, and dark-field images all can be acquired simultaneously. In theory, this allows investigators to observe the photoreceptor waveguided reflectance mosaic, the inner segment mosaic, and the RPE mosaic at the same location at the same time. In practice, slight differences in the optimal focus level for each cell class sometimes requires sequential imaging of the multiple retinal layers. In addition, RPE imaging with dark-field to date has not proven 100% successful, since there can be significant cross-talk between the photoreceptor mosaic images and the RPE (Scoles *et al.*, 2013). Regardless, the technique of simultaneously collecting multiple images under these difference detection configurations holds promise for studying retina structure at the cellular level in normal sighted individuals and in patients with retinal disease. Finally, other configurations of knife edge and offset pinhole arrangements have led to the first in vivo images of inner retinal neurons in mice (Guevara-Torres *et al.*, 2015), nonhuman primates and humans (Rossi *et al.*, 2017). Innovation surrounding AOSLO illumination and detection schemes have clearly provided methods to elucidate contrast between cells that are transparent under standard reflectance and confocal detection techniques. Undoubtedly, investigators will continue to explore novel methods for noninvasive imaging of additional retinal cell types, and work to optimize imaging conditions to allow routine observation of multiple cell types in the presence and absence of disease.

AO-OCT

While AO provides high transverse resolution as described above, optical coherence tomography (OCT) provides high axial resolution. Time-domain OCT (without AO) was first demonstrated by Huang *et al.* (1991). In time-domain OCT, interference signals are measured between the retina and reflections from a reference arm of an interferometer as it is translated in depth. The inference signal generated using this paradigm is used to create a one-dimensional (1D) cross-sectional image (A-scan) showing the relative depth of each backscattering layer in the retina. Following the acquisition of a single A-scan, the OCT beam is laterally translated on the retina so that the location of the subsequent A-scan is adjacent to the previous one, thereby creating a 2D cross-sectional image (B-scan) of the retina, built up over time. Raster scanning the location of the A-scan in 2D creates a three-dimensional (3D) volumetric OCT data set.

Applications for time-domain OCT were limited by the A-scan acquisition rate, which in turn was limited by the time needed to physically move the reference arm through the depth corresponding to the dimension of each A-scan. Keeping the reference arm static during the A-scan acquisition would in part overcome the limitation imposed by acquisition time but required further technological innovation within the OCT imaging technique. Spectral domain OCT and swept source OCT provided this needed innovation by operating in the spectral regime and utilizing a stationary detector (using a spectrometer to detect the OCT signal or

a tunable light source that is rapidly swept through a broad band, respectively). State-of-the-art OCT systems all take advantage of the increased imaging speeds and higher sensitivity afforded by spectral domain OCT methods in comparison to time-domain OCT (Choma *et al.*, 2003).

The properties governing the transverse and axial resolution in OCT are independent, making OCT a perfect complement for AO. As described above, the transverse resolution of OCT is governed by the eye's psf. The axial resolution, however, is dependent only on the central imaging wavelength (λ_0) and the bandwidth of the imaging source ($\Delta\lambda$).

$$l_{\text{axialOCT}} \propto \frac{\lambda_0^2}{\Delta\lambda}$$

Combining AO with OCT yields subcellular resolution in all dimensions, in theory enabling subcellular volumetric imaging in the living retina. The first application of AO-OCT was by Hermann *et al.* (2004). Since that time numerous groups have utilized AO-OCT to acquire volume data sets of the retina with increased resolution, in particular, showing axial reflections arising from the inner/outer segment junction (also termed ellipsoid layer) and cone/rod outer segment tips. More recently, improvements in software to allow intra-scan registration and hardware to use multiple synchronized detectors to further increase scan acquisition speed (Kocaoglu *et al.*, 2014) have led to AO-OCT images showing the inner/outer segment junctions of the cones, the cone outer segment tips, the RPE mosaic, and the choriocapillaris (Kocaoglu *et al.*, 2016; Liu *et al.*, 2016). AO-OCT imaging has also enabled visualization of the individual cells within the retinal volume corresponding to the ganglion cell bodies, showing not only the individual ganglion cells in a monolayer, but also the individual ganglion cells piled on top of each other in depth (Liu *et al.*, 2017). Applications for AO-OCT imaging abound, especially now that the technology has reached a point where the inner retinal cells can be visualized. To date, however, AO-OCT technology has not been as widely adopted as the AOSLO or AO flood-illumination imaging techniques.

Post Processing and AO Image Analysis

Each of the AO imaging modalities described above results in the acquisition of a retinal image with cellular resolution. Regardless of the imaging modality used, the signal to noise ratio of the retinal image can be improved by averaging multiple images together. Other applications gain an increase in contrast (e.g., motion contrast) by comparing retinal features between sequentially acquired images. Both averaging and comparing sequential images require co-registration of the images. Further, the images obtained using AO imaging systems generally span a field of view of two degrees per side or less (with some exceptions, that then sacrifice sampling density by enlarging pixel size). To cover a larger field of view, investigators typically acquire multiple images adjacent to one another, and in post-processing analysis stitch them together to observe a larger portion of the retina. The result of both multi-frame averaging and multi-image location montaging, means that post processing of AO images regularly occurs. While the concept of averaging and montaging is valid for all the imaging modalities discussed, the image processing steps undertaken must take into account the nature and properties of the acquisition modality being used. For example, a series of frames acquired using a flood-illuminated system at nominally the same retinal location may show translation and rotation between frames, though each frame individually should be free from motion artifacts (assuming the involuntary eye motions of the subject are minimal over the acquisition time of the image as was shown for most images acquired in less than 10 ms). Thus co-alignment of multiple images can be accomplished by finding the optimal transformation to account for the translation and rotation between individual frames using cross-correlation or other comparative algorithms, such as the scale invariant features transform (SIFT). Once the individual images are transformed to be co-aligned, a simple average of the aligned frames provides a higher signal to noise retinal image than an individual frame, and this averaged image from flood-illuminated AO systems is mostly free of motion artifacts.

In comparison, the image processing for both AOSLO and AO-OCT images is more involved, since both techniques are acquired over a period of time throughout which the eye is continually undergoing involuntary eye motions. This results not only in the inter-frame/inter-scan translation and rotation motion described in the flood illumination acquisition condition but also motion within a frame/scan, considered intra-frame/intra-scan motion. An added complication for scanning systems comes from any nonlinearities in the scanner velocity. For example, most AOSLO systems use a fast resonant scanner with a sinusoidal velocity. Evenly sampled pixels in time then result in a sinusoidal distortion in the pixel scale across the frame along the direction parallel to the scanner motion, with pixels on the edges of the image elongated in comparison with pixels closer to the middle of the frame, which are more compressed. These same types of lateral distortions are present in OCT images acquired with systems that also use a nonlinear scanner. The first step of image post-processing is to de-warp the images to remove any effect from a nonlinear scan velocity. The second step is to average together multiple dewarped frames. This is routinely done using an intra-frame registration algorithm, which involves first choosing a reference frame (in the ideal case free from intra-frame motion), second dividing the remaining frames into strips (along the fast scan direction), and third, registering each of these strips to the reference frame (Dubra and Harvey, 2010). The product of this intra-frame 2D registration is to allow averaging of multiple frames to increase the signal to noise ratio of the final registered image, while removing the intra-frame motion from frames other than the reference, which would otherwise blur the features in the image. Unfortunately, this technique does not correct intra-frame motion contained within the selected reference image, and thus the final averaged image will necessarily contain the same spatial distortions that exist within the reference image. At present there is no widely accepted technique for correcting the intra-frame motion within the reference frame, though solutions for using multiple reference frames to identify the ground truth for the image have been proposed

(Cooper *et al.*, 2016a,b; Bedggood and Metha, 2017). In comparison with AOSLO, AO-OCT image registration and averaging has the added complication that the retina is moving in 3D, so registering adjacent A-scans to account for axial movements throughout a B-scan and volumetric data set is also required.

Following multi-frame registration and averaging, adjacent images or volumetric data sets must be montaged to expand the field of view from the small field acquired over a single imaging location to a larger field covering a bigger portion of the retina. Most montaging to date has been done manually, however, an automated montaging algorithm was recently described and validated at the cellular level by Chen *et al.* (2016). The automated algorithm was based on feature matching between images using a combination of SIFT and random sample consensus (RANSAC) functions and the resultant automated montage was of comparable quality to manual montages made from the same image sets. While the software was only validated using images acquired from an AOSLO, the authors speculated that the algorithm would also work on *en face* images acquired under other AO imaging modalities.

Metrics to Analyze Cell Mosaics

AO imaging has enabled visualization of many layers and cell types in the retina, and with this visualization comes the desire to quantify metrics within the image to assess retinal health. To date the most studied feature of AO images remains the waveguided reflectance of the cone photoreceptors. As a result, the most quantification of AO images has been done on the cone mosaic. The discussion of AO metrics below describes what has been done to date and therefore is heavily focused on analysis of the cone mosaic. However, the techniques described here could be applied to any cell mosaic visualized, and almost assuredly will be applied to other cell classes beyond the cone mosaic as they become imaged more routinely.

As described above, AO ophthalmoscopy has enabled routine imaging of the cone mosaic both in normal sighted individuals and in patients with retinal disease. With the ability to observe waveguiding cones, comes the ability to identify cone locations and then quantify parameters of the cone mosaic as a function of retinal eccentricity. Numerous metrics have been used to quantify the structure of the cone mosaic including: cone density, average cone spacing, cone size, mosaic regularity, number of neighbors, nearest/farthest neighbor spacing, cone orientation, and others (Cooper *et al.*, 2016a,b). Each metric has its own set of advantages and disadvantages, and thus the best characterization of a mosaic may be a combination of multiple metrics. However, all mosaic metrics share one thing in common: prior to quantification, investigators must first determine the location of each cone in the mosaic. In addition, regardless of which metric is used, all contiguous cones within the analyzed region must be identified in order to obtain the most accurate results.

Determining the cone locations is many times done in a semi-automated fashion. Initially, cone locations are found by automated algorithms, such as the one described by Li and Roorda (2007), which identify local intensity maxima in an image, while also imposing restrictions on the location of the local maxima (such as a minimum spacing between maxima to eliminate false positives from noise and multiple counts for a single cone). Generally, automated cone selections are then manually reviewed to allow an operator to add cone locations (correct false negatives from the automated algorithm), and remove or move cone locations (correct false positives from the automated algorithm). To date, no automated cone identification software has proven itself superior enough to eliminate the need for manual review. In addition, there are many times when automated algorithms currently will fail, in particular, in cases of disease where cones may not exhibit normal reflectance or spacing patterns. Since automated algorithms are just searching for local maxima within an image, the algorithms at present will find “cones” in any image, so operators must inspect images prior to using automated cone selection algorithms and must review the algorithm’s output in order to maintain high quality cone identifications. Automated algorithms have also been presented for identifying the cone inner segments in split detection images and the RPE cells in autofluorescence images. Regardless of the imaging modality or automated algorithm, manual inspection of the results remains necessary. Manually/semi-automated grading of normal cone locations has been shown to be repeatable and reliable between graders (Fig. 7).

Once cone (or other cell) locations are identified, a number of metrics can be applied to determine parameters of the mosaic at a given retinal location, both in normal and diseased eyes. The two most commonly used mosaic parameters are cone density and cone spacing. Cone density simply counts the number of cells identified over a given area. Typically, cell spacing is found by the density recovery profile (DRP) algorithm, which first measures the distance between each pair of identified cones in an image, second determines a histogram of distances by placing the paired distances into bins, and third identifies the distance at which the first peak of the histogram occurs. Distances between cells can also be used to identify the average nearest/farthest neighbor spacing. In a perfectly regular triangularly packed mosaic, the nearest, farthest, and DRP cone spacing would all be the same. Thus the width of the DRP histogram, or the variability in any of the cell spacing metrics gives information regarding the regularity of the cell mosaic.

Regularity can also be assessed through calculation of the Voronoi diagram, which assigns each pixel in the image to its nearest cell center; boundaries of the Voronoi diagram correspond to midpoints between cell locations and vertices correspond to locations equidistant to three cell centers. This type of analysis allows quantification of the average and standard deviation of the number of Voronoi neighbors, where in a perfectly regular triangularly packed mosaic all cells have exactly six neighbors. Each Voronoi cell area is the inverse of cell density at a particular cell location, and the average Voronoi cell area is the inverse of the cell density over the image.

Regardless of which mosaic metric is used (density, spacing, regularity, or other), the Voronoi diagram is useful for identifying the boundaries over which a metric should be calculated (Fig. 8). Cone/cell metrics are currently only calculated over small regions

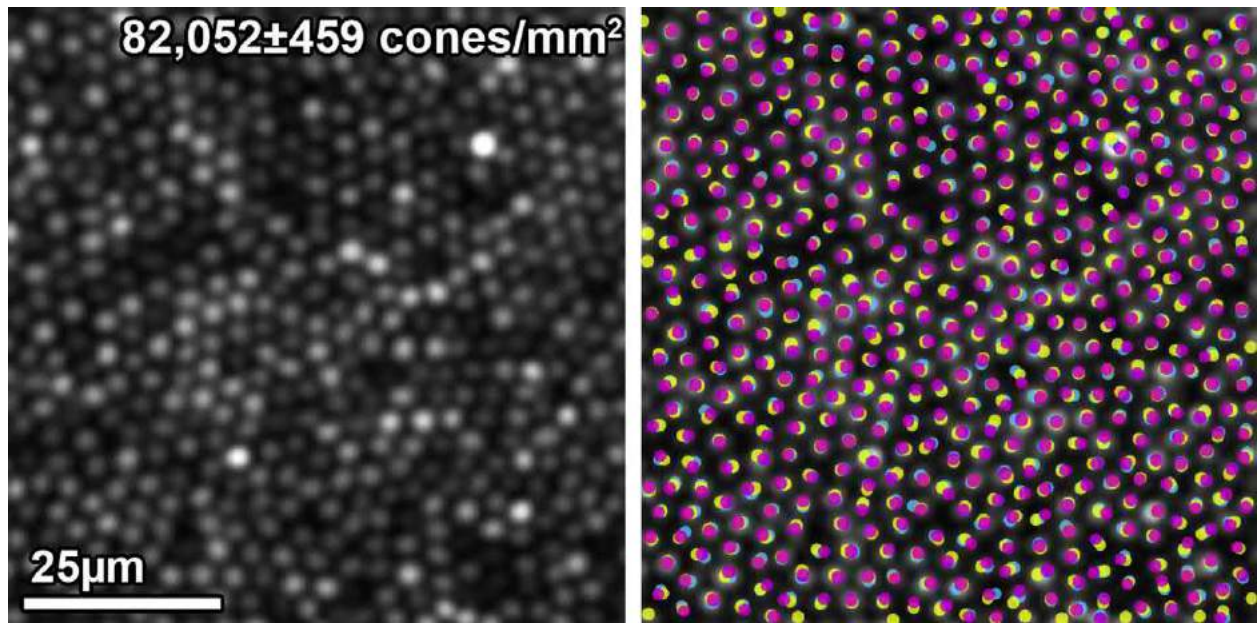


Fig. 7 Confocal adaptive optics scanning laser/light ophthalmoscopy (AOSLO) image of a normal sighted control (left). Three-trained observers manually identified cones (colored points overlaid on the cone mosaic image, right). Mean $\pm$ standard deviation of the cone density measurements from the three observers is shown.

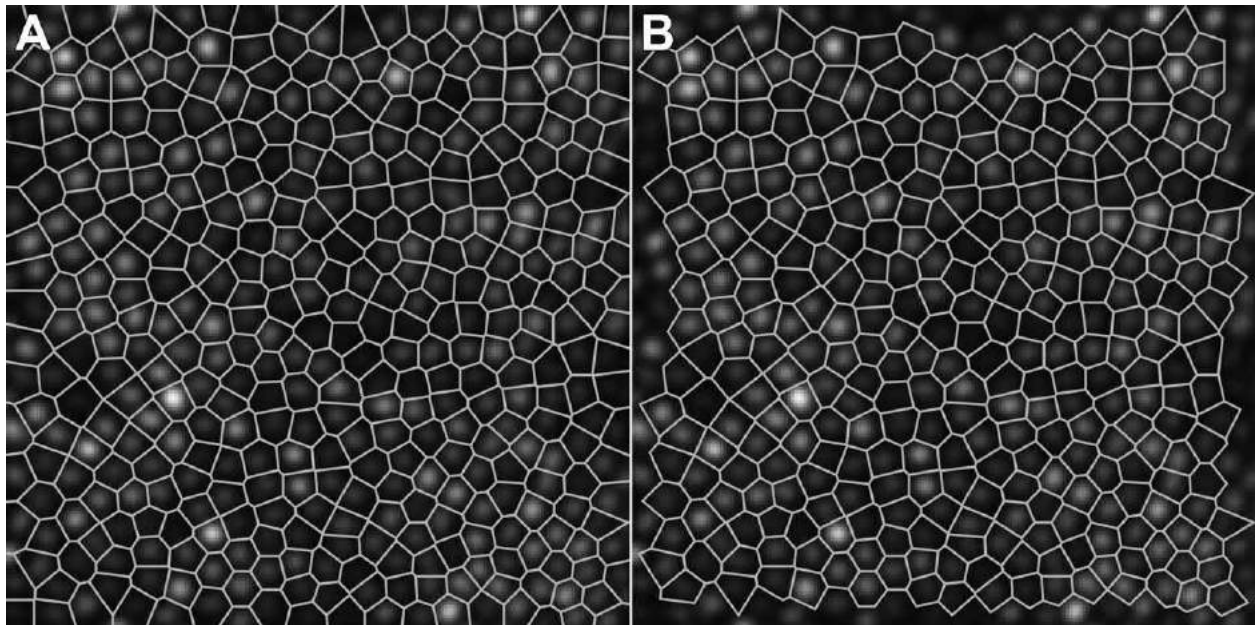


Fig. 8 Voronoi diagram overlaid on a confocal adaptive optics scanning laser/light ophthalmoscopy (AOSLO) image of the cone mosaic (A). The edge of the image shows the cells that are not bounded within the Voronoi diagram. Image metrics should be calculated over only bound Voronoi cells (B) to reduce edge effect errors. Even within the bound Voronoi mosaic, error from edge effects are still present, for example, the bound but pointed Voronoi cells at the top of the image (B).

of interest selected throughout the larger montage, and the borders of those regions of interest can lead to errors in mosaic metrics (As an example, consider a border cell, whose nearest/farthest neighbor is outside of the region of interest and therefore is not identified.). To reduce boundary error, only cells who are fully bound within the Voronoi diagram (i.e., have no open Voronoi edges) should be included in summary metrics from an image.

One of the advantages of *in vivo* imaging is that the same cells can be followed over time. AO imaging has enabled longitudinal tracking of cells, though despite two decades of research, the number of studies that evaluate longitudinal cone mosaic metrics is

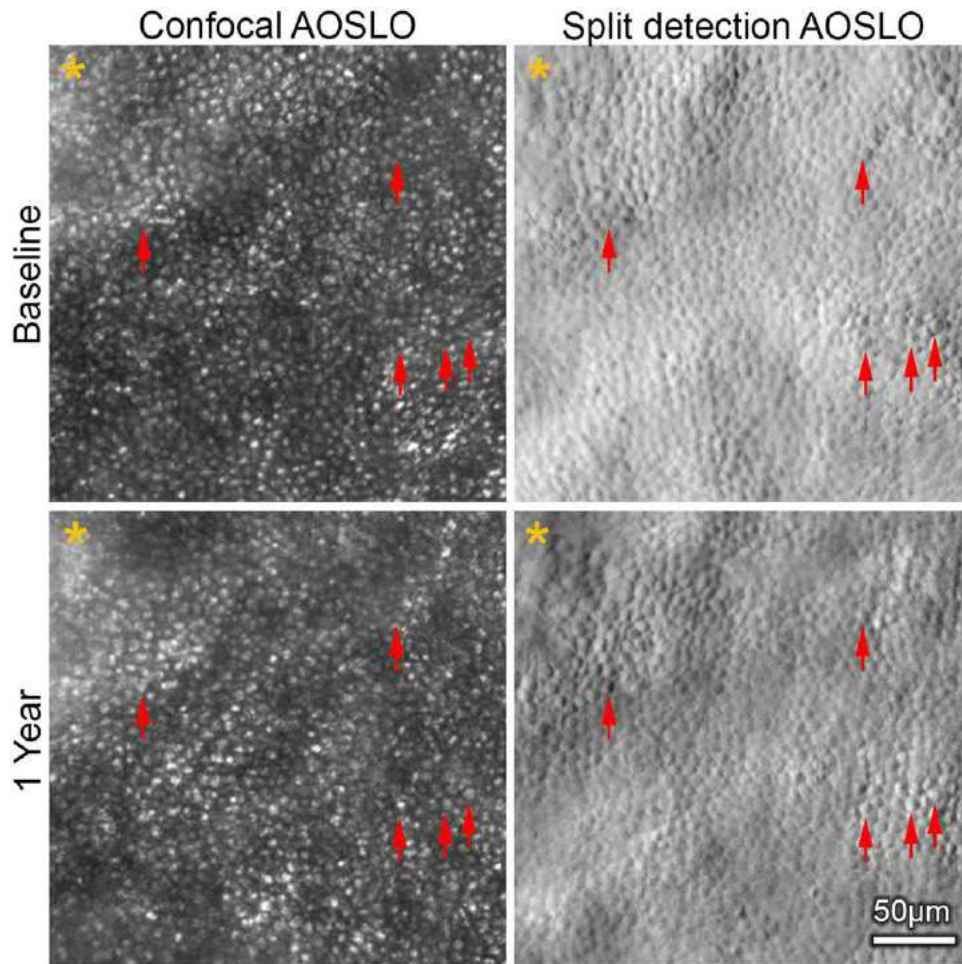


Fig. 9 Longitudinal confocal and non-confocal split detection AOSLO images at baseline and 1 year in a patient with Choroideremia, an X-linked inherited retinal degeneration. The cone photoreceptor mosaic is visible in both the confocal and split detection images, though the confocal reflectance from the cones of the Choroideremia patient appears “smudged” in comparison to the confocal reflectance of normal cones as seen in Figs. 1 and 5(A). Cell-by-cell alignment is demonstrated over the 1 year time period, with the red arrows showing the same cells visible in each imaging modality at each time point. The yellow asterisk marks the sharp border between intact retina and atrophy in this patient. Mild inward progression of the border can be observed over the 1 year time period.

limited. This is in part due to the complicated analysis process of ensuring that the same cells are followed. First, the small field of view of most AO imaging systems requires that longitudinal montaging be performed. For patients, this means any change in the preferred location of fixation will complicate longitudinal alignment. Second, the reflectance profiles of cones changes on time scales of minutes, hours, and days. Thus while cones are expected to remain stationary in the mosaic, the reflectance pattern over the mosaic will change from time point to time point. This means intensity matching will not successfully align images from different time points, and that pattern matching between time points will be more difficult. Third, for scanning systems, any distortions caused by eye motion will be different between images from different time points. This will further complicate alignment between images from different imaging sessions. Finally, in cases of pathology, true changes in retinal structure may make cell-by-cell alignment impossible. Despite these difficulties, longitudinal imaging can be performed and alignment with cellular precision can be achieved (Fig. 9). At present this process is tedious and time-consuming, and more robust and reliable algorithms are needed before the technique will be widely used.

Applications to Clinical Trials

Before any new technology is accepted into routine medical care, it must demonstrate clinical utility by changing currently accepted standards of care. Investigators have long touted the potential for AO to influence patient care, both by enabling earlier detection of disease and by assessing treatment outcomes. However, to date most clinical studies utilizing AO imaging have been limited to cross-sectional studies investigating affected patients’ cone densities (or other metric/AO feature) in comparison to

normal. Nonetheless, AO imaging is now entering a phase where investigators are increasingly including AO imaging as a secondary/exploratory outcome measure in clinical trials testing the safety and efficacy of experimental interventions.

There are at least two conceivable mechanisms through which AO imaging could influence clinical trials. First, AO imaging could identify patients most amenable to a particular treatment. Second, AO could inform assessments of the safety and efficacy of treatment at the cellular level. As just one example, recent advances have led to an era where gene therapy approaches are conceivable for treating inherited retinal degenerations. Gene therapy aims to restore cellular function to dysfunctional or non-functional cells but its success requires that the target cells are present prior to treatment. Therefore, in theory, AO imaging could identify which patients, and which areas of their retinas, structurally retain dysfunctional cells that would be amenable to gene therapy treatment. In addition, longitudinal AO imaging of the treated patients could provide evidence of retinal changes (either positive or negative) following intervention at the same cellular scale. This type of clinical application for AO is not limited to gene therapy, but could also benefit any treatment option that targets cellular structure or function, including stem cell therapies, small molecular therapies and optogenetics.

Despite the high potential for AO to provide cellular level insight to the treatment of retinal disease, there remains only one study utilizing AO imaging to assess an experimental therapy. [Talcott et al. \(2011\)](#) applied AOSLO imaging to assess longitudinal changes in cone density in eyes treated with ciliary neuro-trophic factor (CNTF) implants in comparison to sham-treated fellow eyes in three patients (two with retinitis pigmentosa and one with Usher's syndrome). The authors found that eyes treated with CNTF retained higher cone densities over 2–3 years than sham-treated control eyes. Despite this AO finding, the larger Phase 2 trial investigating the efficacy of CNTF found no visual acuity or visual field benefit for CNTF-treated eyes compared to sham-treated eyes ([Birch et al., 2013](#)). Together these results beg the question: is cone density an earlier biomarker of CNTF treatment efficacy than visual acuity or visual fields? One possibility is that longer trials assessing visual acuity and visual fields might show a benefit from CNTF implants, as better visual acuity and/or larger visual fields would ultimately be expected in eyes that retain more cones. Alternatively, retaining visible cones in AO images might not correspond to a functional benefit in these tested patients.

The study by [Talcott et al. \(2011\)](#) for CNTF implants highlights critical considerations for clinical trials. As a practical matter, the structure and economics of clinical trials for experimental therapies mean that a high value is placed on outcome measures that can reveal safety and efficacy (or lack thereof) reasonably soon after application of a therapy. Such outcome measures are best formulated in the context of the most precise characterizations of disease presentation and progression. Because of its increased resolution, cellular imaging capabilities and quantifiable analysis, AO retinal imaging has high potential to yield this precise characterization. Here, AO imaging is already playing a clinical role, and numerous studies have characterized retinal diseases at the cellular level ([Morgan, 2016](#)). However, before AO metrics can become accepted outcome measures for therapeutic clinical trials, investigators must transition from cross-sectional observational studies to longitudinal studies, which utilize cellular metrics to investigate the effect a given treatment has on the natural progression of a disease. Subsequently, the investigators must relate altered disease progression found from AO imaging to altered disease progression measured clinically. While it seems intuitive that patients who retain more cone structure will experience higher visual function, this correlation remains to be demonstrated in a treated patient population.

Functional Assessment in Conjunction With AO Imaging

As described above, most clinical applications of AO have been cross-sectional observational studies, intended to characterize diseased retinal structure. However, interest abounds and capabilities exist for measuring cellular function at the same resolution with which investigators image the retina. There are two methods for investigating function at the cellular level: psychophysical measures and physiological measures.

Psychophysical measures of vision involve recording a patient's subjective response to the presentation of a visual stimulus, such as is done in clinical microperimetry. AO-assisted microperimetry enables the presentation of visual stimuli to the retina with the same cellular resolution (same spatial extent of the psf) as attained for retinal imaging. This, in combination with retinal tracking through AOSLO imaging, has allowed repeated presentations of a visual stimulus targeted to the same photoreceptor or group of photoreceptors, thereby enabling measurements of single cell visual thresholds ([Harmening et al., 2014](#)). For a full review on AO psychophysical testing, see [Tuten et al. \(2012\)](#).

While psychophysical testing remains the gold standard for assessing visual function, measuring function one cell at a time is not practical for testing either high volumes of cells within a patient or for testing high volumes of patients. Instead, objective, validated, and high-throughput biomarkers of function are needed. Physiological biomarkers of cellular function through imaging could provide this high-throughput alternative to single cell psychophysical measures. As one example, investigators have shown that cone near infrared reflectance ([Jonnal et al., 2007](#); [Grieve and Roorda, 2008](#); [Rha et al., 2009](#)) and optical path length ([Hillmann et al., 2016](#)) can change following a visible stimulus. These stimulus-evoked intrinsic reflectance responses hold promise as biomarkers for cone function, though further study demonstrating that these measures correlate with functional vision are needed.

Conclusions

AO is a tool through which investigators can measure and correct for the optical aberrations of the living eye and thereby improve the resolution of retinal imaging devices. AO in combination with numerous retinal imaging modalities and detection schemes

has enabled observation of single cells in the living retina. Clinical applications abound, especially as investigators transition from cross-sectional observational studies to longitudinal studies of retinal disease and its treatment.

Conflict of Interest

IJWM holds US Patent 8226236 on material related to this article.

Acknowledgments

This article was supported by NIH U01EY025477, NIH U01EY025864, the Foundation Fighting Blindness, Research to Prevent Blindness, the F.M. Kirby Foundation, and the Paul and Evanina Mackall Foundation Trust. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

See also: Confocal Microscopy. Optics of the Human Eye

References

- Babcock, H.W., 1953. The possibility of compensating astronomical seeing. *Publications of the Astronomical Society of the Pacific* 65,229–236.
- Bedggood, P., Metha, A., 2017. De-warping of images and improved eye tracking for the scanning laser ophthalmoscope. *PLOS ONE* 12 (4), e0174617.
- Birch, D.G., Weleber, R.G., Duncan, J.L., Jaffe, G.J., Tao, W., 2013. Randomized trial of ciliary neurotrophic factor delivered by encapsulated cell intraocular implants for retinitis pigmentosa. *American Journal of Ophthalmology* 156,283–292.
- Carroll, J., Gray, D.C., Roorda, A., Williams, D.R., 2005. Recent advances in retinal imaging with adaptive optics. *Optics and Photonics News* 16,36–42.
- Carroll, J., Neitz, M., Hofer, H., Neitz, J., Williams, D.R., 2004. Functional photoreceptor loss revealed with adaptive optics: An alternate cause for color blindness. *Proceedings of the National Academy of Sciences of the United States of America* 101,8461–8466.
- Chen, M., Cooper, R.F., Han, G.K., *et al.*, 2016. Multi-modal automatic montaging of adaptive optics retinal images. *Biomedical Optics Express* 7,4899–4918.
- Choma, M., Sarunic, M., Yang, C., Izatt, J., 2003. Sensitivity advantage of swept source and Fourier domain optical coherence tomography. *Optics Express* 11,2183–2189.
- Chui, T.Y., Dubow, M., Pinhas, A., *et al.*, 2014. Comparison of adaptive optics scanning light ophthalmoscopic fluorescein angiography and offset pinhole imaging. *Biomedical Optics Express* 5,1173–1189.
- Chui, T.Y., Vannasdale, D.A., Burns, S.A., 2012. The use of forward scatter to improve retinal vascular imaging with an adaptive optics scanning laser ophthalmoscope. *Biomedical Optics Express* 3,2537–2549.
- Cooper, R.F., Sulai, Y.N., Dubis, A.M., *et al.*, 2016a. Effects of intraframe distortion on measures of cone mosaic geometry from adaptive optics scanning light ophthalmoscopy. *Translational Vision Science and Technology* 5,10.
- Cooper, R.F., Wilk, M.A., Tarima, S., Carroll, J., 2016b. Evaluating descriptive metrics of the human cone mosaic. *Investigative Ophthalmology & Visual Science* 57,2992–3001.
- Dubra, A., Harvey, Z., 2010. Registration of 2D images from fast scanning ophthalmic instruments. *Lecture Notes in Computer Science* 6204,60–71.
- Geng, Y., Dubra, A., Yin, L., *et al.*, 2012. Adaptive optics retinal imaging in the living mouse eye. *Biomedical Optics Express* 3,715–734.
- Grieve, K., Roorda, A., 2008. Intrinsic signals from human cone photoreceptors. *Investigative Ophthalmology & Visual Science* 49,713–719.
- Guevara-Torres, A., Williams, D.R., Schallek, J.B., 2015. Imaging translucent cell bodies in the living mouse retina without contrast agents. *Biomedical Optics Express* 6,2106–2119.
- Harmening, W.M., Tuten, W.S., Roorda, A., Sincich, L.C., 2014. Mapping the perceptual grain of the human retina. *Journal of Neuroscience* 34,5667–5677.
- Hermann, B., Fernandez, E.J., Unterhuber, A., *et al.*, 2004. Adaptive-optics ultrahigh-resolution optical coherence tomography. *Optics Letters* 29,2142–2144.
- Hillmann, D., Spahr, H., Pfaffle, C., *et al.*, 2016. In vivo optical imaging of physiological responses to photostimulation in human photoreceptors. *Proceedings of the National Academy of Sciences of the United States of America* 113,13138–13143.
- Hofer, H., Chen, L., Yoon, G.Y., *et al.*, 2001. Improvement in retinal image quality with dynamic correction of the eye's aberrations. *Optics Express* 8,631–643.
- Huang, D., Swanson, E.A., Lin, C.P., *et al.*, 1991. Optical coherence tomography. *Science* 254,1178–1181.
- Jackman, W., Webster, J., 1886. Photographing the retina of the living human eye. *Photographic News* 23,340–341.
- Jonnal, R.S., Rha, J., Zhang, Y., *et al.*, 2007. In vivo functional imaging of human cone photoreceptors. *Optics Express* 15,16141–16160.
- Kocaoglu, O.P., Liu, Z., Zhang, F., *et al.*, 2016. Photoreceptor disc shedding in the living human eye. *Biomedical Optics Express* 7,4554–4568.
- Kocaoglu, O.P., Turner, T.L., Liu, Z., Miller, D.T., 2014. Adaptive optics optical coherence tomography at 1 MHz. *Biomedical Optics Express* 5,4186–4200.
- Liang, J., Grimm, B., Goelz, S., Bille, J., 1994. Objective measurement of the wave aberrations of the human eye using a Hartmann–Shack wavefront sensor. *Journal of the Optical Society of America A* 11,1949–1957.
- Liang, J., Williams, D.R., Miller, D.T., 1997. Supernormal vision and high-resolution retinal imaging through adaptive optics. *Journal of the Optical Society of America A* 14,2884–2892.
- Li, K.Y., Roorda, A., 2007. Automated identification of cone photoreceptors in adaptive optics retinal images. *Journal of the Optical Society of America A* 24,1358–1363.
- Liu, Z., Kocaoglu, O.P., Miller, D.T., 2016. 3D imaging of retinal pigment epithelial cells in the living human retina. *Investigative Ophthalmology & Visual Science* 57, OCT533–OCT543.
- Liu, Z., Kurokawa, K., Zhang, F., Miller, D., 2017. In vivo imaging of human retinal ganglion cells with AO-OCT. *IOVS*. 58: ARVO E-Abstract 3430.
- Morgan, J.I., 2016. The fundus photo has met its match: Optical coherence tomography and adaptive optics ophthalmoscopy are here to stay. *Ophthalmic and Physiological Optics* 36,218–239.
- Morgan, J.I.W., Dubra, A., Wolfe, R., Merigan, W.H., Williams, D.R., 2009. In vivo autofluorescence imaging of the human and macaque retinal pigment epithelial cell mosaic. *Investigative Ophthalmology & Visual Science* 50,1350–1359.
- Packer, O., Williams, D.R., 2003. Light, the retinal image, and photoreceptors. In: Shevell, S. (Ed.), *The Science of Color*. Oxford: Elsevier Science Publishing Company.
- Pallikaris, A., Williams, D.R., Hofer, H., 2003. The reflectance of single cones in the living human eye. *Investigative Ophthalmology & Visual Science* 44,4580–4592.
- Rha, J., Jonnal, R.S., Thorn, K.E., *et al.*, 2006. Adaptive optics flood-illumination camera for high speed retinal imaging. *Optics Express* 14,4552–4569.

- Rha, J., Schroeder, B., Godara, P., Carroll, J., 2009. Variable optical activation of human cone photoreceptors visualized using a short coherence light source. *Optics Letters* 34,3782–3784.
- Roorda, A., Romero-Borja, F., Donnelly III, W.J., *et al.*, 2002. Adaptive optics scanning laser ophthalmoscopy. *Optics Express* 10,405–412.
- Roorda, A., Williams, D.R., 1999. The arrangement of the three cone classes in the living human eye. *Nature* 397,520–522.
- Rossi, E.A., Granger, C.E., Sharma, R., *et al.*, 2017. Imaging individual neurons in the retinal ganglion cell layer of the living eye. *Proceedings of the National Academy of Sciences of the United States of America* 114,586–591.
- Scoles, D., Sulai, Y.N., Dubra, A., 2013. In vivo dark-field imaging of the retinal pigment epithelium cell mosaic. *Biomedical Optics Express* 4,1710–1723.
- Scoles, D., Sulai, Y.N., Langlo, C.S., *et al.*, 2014. In vivo imaging of human cone photoreceptor inner segments. *Investigative Ophthalmology & Visual Science* 55,4244–4251.
- Sharma, R., Schwarz, C., Williams, D.R., *et al.*, 2016. In vivo two-photon fluorescence kinetics of primate rods and cones. *Investigative Ophthalmology & Visual Science* 57,647–657.
- Sheehy, C.K., Tiruveedhula, P., Sabesan, R., Roorda, A., 2015. Active eye-tracking for an adaptive optics scanning laser ophthalmoscope. *Biomedical Optics Express* 6,2412–2423.
- Talcott, K.E., Ratnam, K., Sundquist, S.M., *et al.*, 2011. Longitudinal study of cone photoreceptors during retinal degeneration and in response to ciliary neurotrophic factor treatment. *Investigative Ophthalmology & Visual Science* 52,2219–2226.
- Tam, J., Liu, J., Dubra, A., Fariss, R., 2016. In vivo imaging of the human retinal pigment epithelial mosaic using adaptive optics enhanced indocyanine green ophthalmoscopy. *Investigative Ophthalmology & Visual Science* 57,4376–4384.
- Tuten, W.S., Tiruveedhula, P., Roorda, A., 2012. Adaptive optics scanning laser ophthalmoscope-based microperimetry. *Optometry and Vision Science* 89,563–574.
- Venkateswaran, K., Roorda, A., Romero-Borja, F., 2004. Theoretical modeling and evaluation of the axial resolution of the adaptive optics scanning laser ophthalmoscope. *Journal of Biomedical Optics* 9,132–138.
- Wolfing, J.I., Chung, M., Carroll, J., Roorda, A., Williams, D.R., 2006. High-resolution retinal imaging of cone-rod dystrophy. *Ophthalmology* 113,1014–1019.

Femtosecond Lasers in Retinal Imaging

Christina Schwarz and Jennifer J Hunter, University of Rochester, Rochester, NY, United States

© 2018 Elsevier Inc. All rights reserved.

Introduction

Early diagnosis and timely therapeutic intervention is critical to avoid blindness in retinal disease. While clinical symptoms often manifest late in the disease course, structural changes are visible on a microscopic scale long before onset of clinical disease. State-of-the-art adaptive optics scanning light ophthalmoscopes (AOSLO) capture reflected or scattered light from the fundus of the eye and provide exquisite resolution in *en face* dimensions ($<2\ \mu\text{m}$), which is suited to observe changes in cellular structure (Roorda and Duncan, 2015). In unhealthy retina, cellular metabolism and function are altered prior to structural changes and cell death. Detecting these early changes may provide an opportunity for treatment to prevent vision loss. The most promising imaging modalities are those that can interrogate the retina's molecular composition and biochemical pathways, providing thus information on the tissue's anatomy and physiology.

Fluorescence imaging has great potential because it can spectrally target particular fluorophores, either endogenous or exogenous, offering the possibility of high contrast imagery of specific molecular species. A milestone for biological imaging was achieved in 1990, when Denk and colleagues applied two-photon excited fluorescence (TPEF) to microscopy (Denk *et al.*, 1990). TPEF microscopy is based on the fact that two photons arriving at a fluorescent molecule closely in time can behave like a single photon of half the wavelength (Göppert-Mayer, 1931; Göppert, 1929). Great strides have been made since in cell biology and neuroscience thanks to the technique's optical sectioning capability in highly scattering tissue and its ability to provide structural and functional information (Campagnola *et al.*, 1999; Denk and Detwiler, 1999; Denk *et al.*, 1995; Helmchen, 2009; Svoboda *et al.* 1997; Svoboda and Yasuda, 2006; Yuste and Denk, 1995). TPEF offers yet another advantage when applied to the eye: a host of endogenous fluorophores excitable in the ultraviolet (UV) can be studied in the living primate eye for the first time. Fluorophores with valuable diagnostic potential, such as NAD(P)H (Huang *et al.*, 2002; Shuttleworth, 2010; Zipfel *et al.*, 2003) and all-*trans*-retinol (Imanishi *et al.*, 2004; Kaplan, 1985), are normally inaccessible because their excitation requires UV light that is blocked by the cornea and crystalline lens (Boettner and Wolter, 1962). However, as illustrated in Fig. 1, infrared (IR) light is readily transmitted by the eye's optics and can excite these fluorophores via two-photon absorption. Fluorescence is emitted in the visible spectrum, which exits the eye relatively unattenuated and can be collected for image formation.

This article reviews the current state of *in vivo* nonlinear retinal imaging. It treats the theory of nonlinear ophthalmoscopy, which is generally performed with a nonlinear microscope where the anterior optics of the eye is used instead of an optimized microscope objective to investigate the back of the eye. Further, it discusses its practical implementation as well as light safety aspects. Finally, an outlook on promising future research is given. Due to the fast growth of this area and the multidisciplinary nature of it, this chapter is not intended to be self-contained. Other reviews on nonlinear microscopy (Diaspro, 2002; Diaspro and

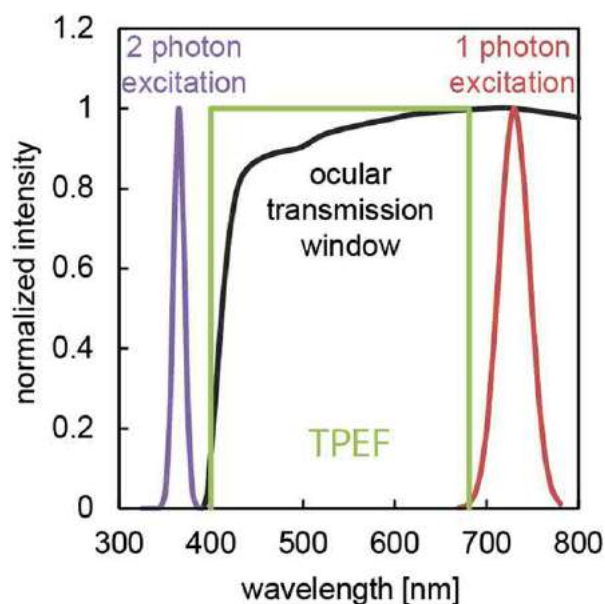


Fig. 1 Many key fluorophores are only excitable in the UV outside the ocular transmission window (Boettner and Wolter, 1962). However, infrared photons can excite these molecules via two-photon absorption.

Chirico, 2003; Diaspro *et al.*, 2005; Masters and So, 2008; Svoboda and Yasuda, 2006; Xu and Webb, 1997) and multiphoton imaging of the eye (Imanishi *et al.*, 2007; Sharma and Hunter, 2017) are highly recommended.

In Vivo Multiphoton Retinal Imaging With Adaptive Optics

By nature the retina is a multilayered tissue that is highly transparent for visible light. While this property is beneficial for vision, it becomes a challenge when the goal is to visualize individual retinal layers. Photoreceptors, nerve fiber layer and retinal vessels can readily be imaged in the visible or IR by capturing reflected light in a confocal detection scheme. Visible or IR light also excites single-photon fluorescence from lipofuscin or melanin, which can be harnessed to reveal the retinal pigment epithelium (Delori *et al.*, 1995; Granger *et al.*, 2017; Morgan *et al.*, 2009). Other retinal layers have so far been impossible to visualize with confocal reflectance or single-photon fluorescence methods, although offset aperture techniques show promise (Scoles *et al.*, 2014; Guevara-Torres *et al.*, 2016; Rossi *et al.*, 2017). Advantageously, autofluorescent molecules, which in the living eye are excitable via multiphoton absorption in the UV, are present throughout the retina. The occurrence and distribution of these molecules may be directly associated with retinal health and physiology. Valuable information can be gained by capitalizing on the fluorescence properties of these molecules, such as the absorption and emission spectra, the fluorescence lifetime, anisotropy, etc. By measuring these characteristics, cells and their organelles can be studied and information about the composition of the cells themselves as well as their surrounding environment can be gained.

Structural In Vivo Multiphoton Ophthalmoscopy

The retina contains many UV-excitabile fluorophores that are abundant in different cell classes or cell compartments: Photoreceptors and retinal pigment epithelial cells contain retinoids (Chen *et al.*, 2005; Imanishi *et al.*, 2004; Kaplan, 1985) that take part in the visual cycle. Lipofuscin accumulates within the retinal pigment epithelium (Bindewald-Wittich *et al.*, 2006), especially at advanced age (Kennedy *et al.*, 1995a). Collagen is present in basement membranes and extracellular matrix and is, along with elastin, an integral component of vessel walls (Zoumi *et al.*, 2002, 2004). As precursor molecules for hemoglobin, porphyrins are present in blood and tissue fluids (Zheng *et al.*, 2011). Additionally, all living cells contain NAD(P)H (Chen *et al.*, 2005) and flavoproteins (Zhao *et al.*, 2007), because these molecules are involved in cellular metabolism. NAD(P)H is highly concentrated in the mitochondria, but is also present in the cytosol in the cell. Flavoproteins are almost exclusively present in the mitochondria.

These fluorophores provide intrinsic optical contrast from cells throughout the retina (Gualda *et al.*, 2010), even though their efficiency is relatively poor compared to specifically designed exogenous fluorophores. Nowadays two-photon excited fluorescence ophthalmoscopy (TPEFO) can be performed routinely in the living eye of mouse (Palczewska *et al.*, 2010, 2014a) and primate (Hunter *et al.*, 2011). Fig. 2 shows near-foveal cones in the living macaque eye imaged in 730 nm confocal reflectance and TPEF modality. While the reflectance of single cones differs considerably between cones, the amount of emitted fluorescence is similar. Using an AOSLO, cellular mosaics in many retinal layers other than the photoreceptor layer have been imaged non-invasively at 730 nm and ~900 nm excitation wavelength (Sharma *et al.*, 2016b). RPE cells appeared dark in the center and bright along the cell wall. At the nerve fiber layer, bright streaks, probably the Müller cells, were visible between the darker nerve fibers. Fluorescence was also captured from vessel walls. A major breakthrough was the visualization of cell layers that are invisible in confocal reflectance imaging methods, such as the nuclear layers (Sharma *et al.*, 2016b). Of particular interest are thereby the ganglion cells because these are the primary output neurons of the retina, which degenerate in glaucoma, the second leading cause of blindness worldwide (Quigley, 2011). Capitalizing on experience gained with two-photon ophthalmoscopy in monkey to find the best focal depth and retinal location for ganglion cell imaging, parameters could be optimized to obtain the first in vivo

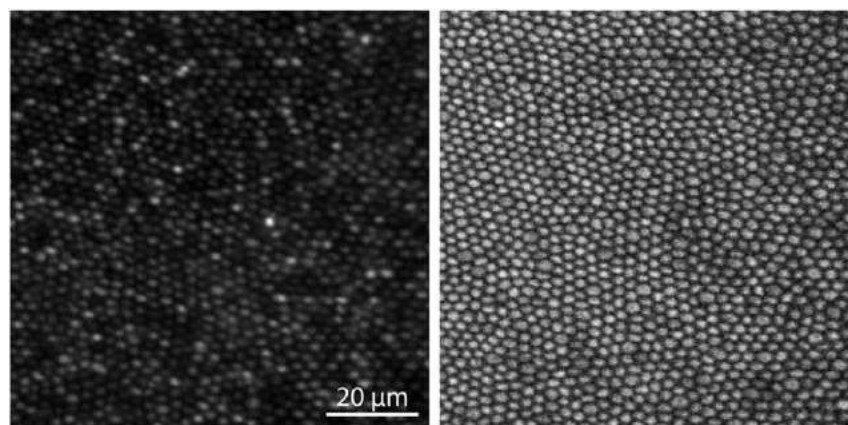


Fig. 2 Reflectance (left) and TPEF (right) image of near-foveal cones in the living macaque eye.

images of ganglion cells with offset aperture reflectance imaging at light levels that are safe for use in the living human eye (Rossi *et al.*, 2017).

Functional In Vivo Multiphoton Ophthalmoscopy

Recording autofluorescence from retinoids and NADH over time has the potential to track functional activity at a cellular scale (Chen *et al.*, 2005; Imanishi *et al.*, 2004; Kaplan, 1985). TPEF from photoreceptors has been shown to increase in response to photopigment bleaching and decline again in the dark (Sharma *et al.*, 2016a), offering the possibility of monitoring photoreceptor function in vivo. The prevailing view is that the dominant source of time-varying TPEF from photoreceptor outer segments is all-*trans*-retinol, an intermediate component of the visual cycle (Ala-Laurila *et al.*, 2006; Chen *et al.*, 2005; Kaplan, 1985; Liebman, 1969, 1973; Tsina *et al.*, 2004). The visual cycle (see Fig. 3 for a schematic of the canonical visual cycle) occurs in several steps involving numerous enzymes and chaperone proteins (Kiser *et al.*, 2012, 2014; Lamb and Pugh, 2004; Saari, 2000). TPEF imaging quantifies an intermediate step or steps inside the visual cycle and might therefore prove to be a valuable tool for optically dissecting the normal visual cycle. The kinetics of the normally functioning visual cycle in nonhuman primates have been characterized in rod photoreceptors using TPEF (Sharma *et al.*, 2017). Fig. 3 shows the time course of TPEF during adaptation to the imaging laser and the response to a brief flash intended to instantaneously bleach the visual pigment with subsequent recovery.

Insufficiencies can occur in any step of the visual cycle due to genetic mutation, dietary factors, or age, which will result in slowed pigment regeneration or even blindness (Lamb and Pugh, 2004). Multiphoton retinal imaging could thus give diagnostic and prognostic insight into retinal diseases such as Stargardt macular dystrophy, retinitis pigmentosa, Leber congenital amaurosis, fundus albipunctatus, Sorsby fundus dystrophy, and Bothnia dystrophy (Lamb and Pugh, 2004; Thompson and Gal, 2003; Travis *et al.*, 2007). Previous techniques to quantify visual cycle kinetics by measuring dark adaptation and/or rhodopsin densitometry have relied on psychophysics, the electroretinogram, or rhodopsin densitometry. There are a number of limitations to using these methods: Psychophysical measurements are subjective and time consuming. Electroretinograms, although objective, are based on global voltage variations generated by many circuits in the retina and are thus highly susceptible to eye movements. While rhodopsin densitometry estimates the kinetics of rhodopsin regeneration directly, it is subject to artifacts such as reflectance changes associated with neural responses or vascular changes (Grieve and Roorda, 2008; Masella *et al.*, 2014a). So far, such contamination from intrinsic optical signals has not been observed in TPEF ophthalmoscopy, and only further experimentation will reveal whether there are important additional components to the TPEF signal.

In addition to endogenous fluorophores, TPEF of extrinsic fluorophores, such as genetically encoded calcium indicators (GECIs) can provide important functional information at the cellular scale. GECIs can be delivered to specific cells or cell compartments within living animals via transgenesis or viral constructs. Fig. 4 shows a retinal ganglion cell labeled with GCaMP5 that was imaged with a two-photon AOSLO in the living macaque eye. For GECIs in retinal cells, the advantage of two-photon over single-photon excitation lies in the use of near-infrared light, which provides reduced visual stimulation and the intrinsic optical sectioning of TPEF. In addition, at these long wavelengths, extrapolating from the trend at 900 nm, axial chromatic aberration and ocular dispersion is expected to be minimal (Atchison and Smith, 2000) which can improve efficiency and further reduce visual interference from the imaging beam. In mouse, in vivo functional responses to visible stimulation have been recorded using TPEF imaging (without adaptive optics correction) of the GECI GCaMP6 (Bar-Noam *et al.*, 2016).

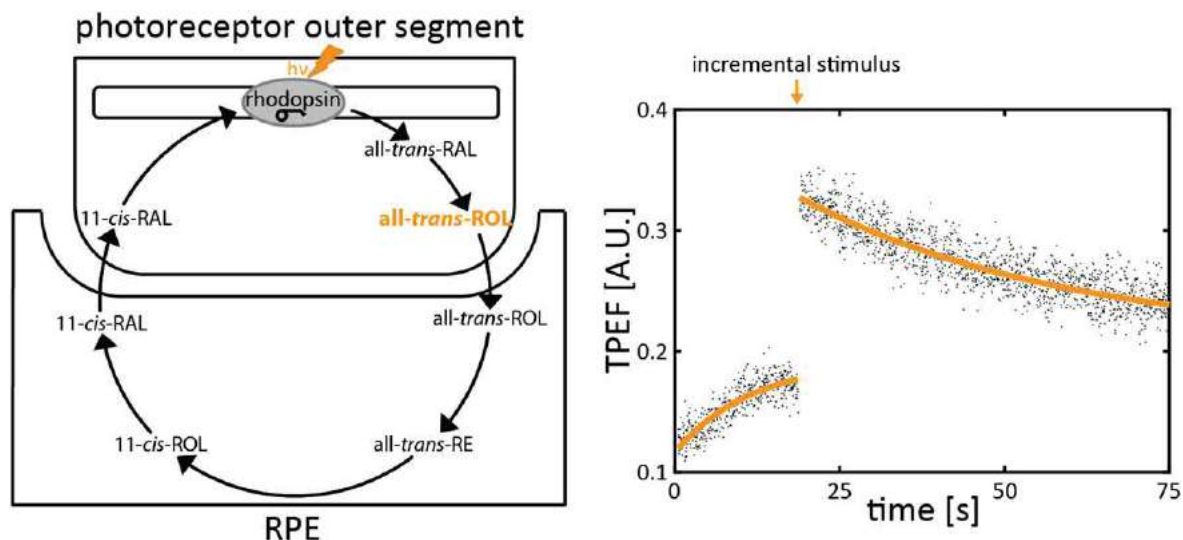


Fig. 3 (left) Schematic of the canonical visual cycle regenerating bleached rhodopsin. All-*trans*-retinol is the strongest candidate for fluorescence changes in response to stimulation. (right) Time course of TPEF. The initial fluorescence increase is due to stimulation by the imaging laser. After an instantaneous full bleach, fluorescence increases abruptly and gradually declines.

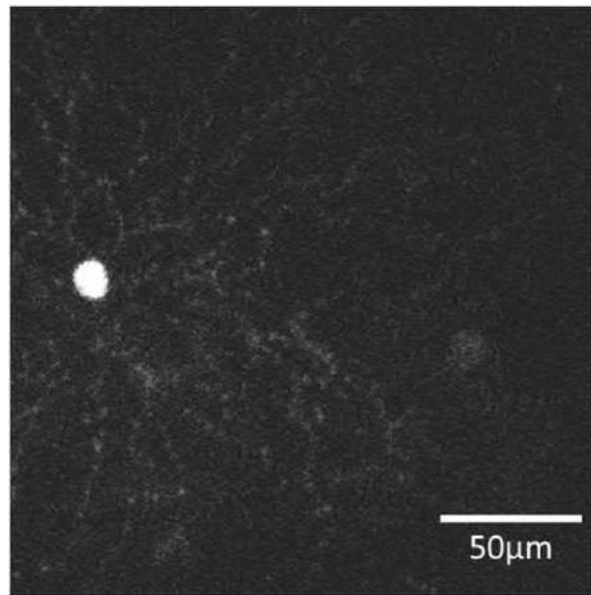


Fig. 4 Retinal ganglion cell labeled with GCaMP5 in the living macaque eye. The highly fluorescent cell nucleus is visible as well as the large dendritic arbor. In collaboration with Robin Sharma, Lu Yin, and William Merigan.

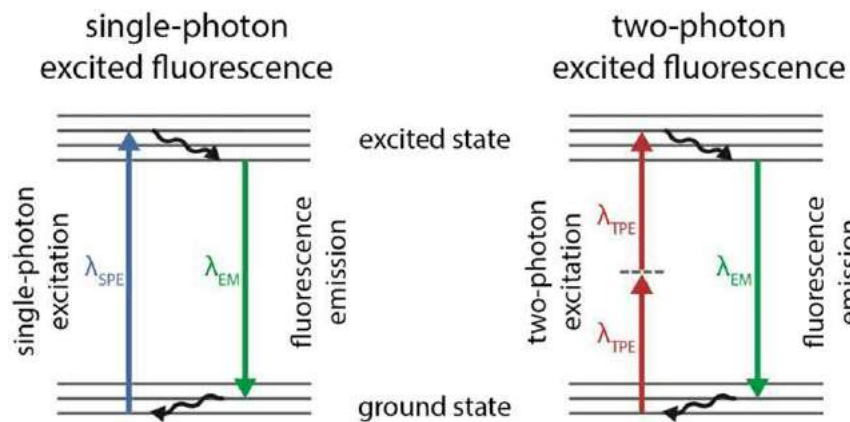


Fig. 5 Jablonski diagram illustrating the excitation of a fluorophore via a single-photon and a two-photon process.

Theory of Multiphoton Retinal Imaging

A number of books and articles on the theory of multiphoton imaging have been published over the years (Denk *et al.*, 2006; Diaspro, 2002; So *et al.*, 2000; Xu and Webb, 1997). Only the most important aspects for retinal imaging are specified here.

The mechanisms of single-photon excitation (SPE) and two-photon excitation (TPE) are illustrated in the Jablonski diagrams in Fig. 5. In single-photon excitation, a fluorescent molecule absorbs a photon and is raised to an excited electronic state. The photon energy must match the energy gap between ground and excited state. As the molecule relaxes back to the ground state, energy can be lost through fluorescence, that is, emission of photons. The emitted photons are of longer wavelength (less energy) than the photons used for excitation due to non-radiative energy loss in the form of heat or vibration within the molecule. The energy difference between fluorescence excitation and emission is referred to as Stokes shift.

Fluorescent molecules can also be excited via a multiphoton process, where two or more photons are absorbed quasi-simultaneously (Göppert-Mayer, 1931; Göppert, 1929). The total energy of the absorbed photons must equal the energy required for conventional single-photon excitation. For two-photon excited fluorescence, the excitation wavelength is approximately twice the wavelength for single-photon excitation.

$$\lambda_{SPE} \approx \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)^{-1} \approx \frac{\lambda_{TPE}}{2}$$

The temporal window for two-photon absorption can be calculated via the energy-time uncertainty (Louisell, 1973), taking into account the photon energy $\Delta E = \frac{h \cdot c}{\lambda_{TPE}}$ at the wavelength for two-photon excitation λ_{TPE} in the UV or visible range

$$\Delta t \approx \frac{\hbar}{2\Delta E} = \frac{\lambda_{TPE}}{4\pi c} \approx 10^{-16} \text{ s}$$

where $\hbar = \frac{h}{2\pi}$ with Planck's constant $h \approx 6.63 \times 10^{-34} \frac{\text{m}^2 \text{kg}}{\text{s}}$ and the speed of light $c \approx 3 \times 10^8 \frac{\text{m}}{\text{s}}$.

For multiphoton excitation, ultrafast lasers emitting very short pulses ($\sim \text{fs}$) are required. Such lasers produce high peak powers, while the average power is low. The peak power can be calculated via

$$P_{\text{peak}} = \frac{\langle P(t) \rangle}{f_{\text{rep}} \cdot \tau_p} = g^{(1)} \cdot \frac{P_{\text{avg}}}{f_{\text{rep}} \cdot \tau_p}$$

where P_{avg} is the average power, f_{rep} is the pulse repetition frequency and τ_p the pulse duration. The factor $g^{(1)}$ depends on the pulse shape and is 0.94 and 0.88 for Gaussian and Hyperbolic-secant-squared pulse shapes respectively.

Because time and frequency domain are related via the Fourier transform, pulse duration and spectral bandwidth are linked. This behavior is expressed in the time-bandwidth product. The time-bandwidth product of a non-chirped, transform-limited Gaussian pulse is $\tau_p \cdot \Delta\lambda = \frac{2\ln 2 \cdot \lambda^2}{\pi \cdot c}$. It can be derived by calculating the FWHM of the Fourier transform of a Gaussian function as $\tau_p \cdot \Delta f = \frac{2\ln 2}{\pi} \approx 0.44$, where $\Delta f = \frac{c \cdot \Delta\lambda}{\lambda^2}$ with λ being the center wavelength. The shorter the pulse durations, the broader is the spectral bandwidth, and the more challenging it becomes to focus the beam in both space and time.

The response of fluorescent molecules to excitation depends on two parameters: The absorption cross section δ_2 quantifies the affinity to absorb photons at a certain wavelength λ and the quantum yield η specifies the efficiency in the conversion of absorbed energy into light emission. The quantum yield is preserved in single-photon and multi-photon excitation; however, absorption cross sections differ. Two-photon absorption cross sections are usually broader and blue shifted from what would be expected from single-photon absorption cross sections (Huang *et al.*, 2002). One advantage of the broader absorption cross sections for two-photon excitation is that a single laser can simultaneously excite multiple fluorophores. In honor of Maria Göppert-Mayer, the widely used yet unofficial unit of the two-photon absorption cross section is the GM, corresponding to $10^{-50} \text{ cm}^4 \text{ s}^{-1} \text{ photon}^{-1}$. Two-photon absorption cross sections are about two orders of magnitude greater for exogenous than endogenous fluorophores. For instance, the two-photon absorption cross section is ~ 6 GM for GFP (Diaspro *et al.*, 2005) and 0.02 GM for NADH (So *et al.*, 2000) at a wavelength of 700 nm.

The fluorescence intensity of molecules that are excited via a two-photon effect is proportional to the absorption cross section δ_2 , the quantum yield η and the intensity I squared of incident light (Diaspro and Chirico, 2003):

$$I_f(t) \propto \delta_2(\lambda_{TPE}) \cdot \eta \cdot I^2(t) = \delta_2(\lambda_{TPE}) \cdot \eta \cdot \frac{P^2(t) \cdot NA^4}{\pi^2 \cdot \lambda_{TPE}^4}$$

taking into account that intensity is defined as power per area $I(t) = \frac{P(t)}{A}$ and the illuminated area at the focal plane of a diffraction-limited system in the paraxial approximation $A \approx \pi \frac{\lambda_{TPE}^2}{NA^2}$ (Born and Wolf, 1999). For lasers with pulse repetition rate f_{rep} and pulse width τ_p , the time-averaged power squared depends on the peak power P_{peak} and the pulse shape $\langle P^2(t) \rangle = g^{(2)} \cdot \langle P(t) \rangle^2 = g^{(2)} \cdot \frac{P_{\text{avg}}^2}{\tau_p \cdot f_{\text{rep}}}$. The factor $g^{(2)}$ is related to the energy contained within one pulse and is 0.66 and 0.59 for Gaussian and Hyperbolic-secant-squared pulses respectively. The average number of photons that the fluorescing medium emits per second is therefore:

$$\langle n_f(t) \rangle \propto \delta_2(\lambda_{TPE}) \cdot \eta \cdot g^{(2)} \cdot \frac{P_{\text{avg}}^2}{\tau_p \cdot f_{\text{rep}}} \left(\frac{NA^2}{\pi \cdot h \cdot c \cdot \lambda_{TPE}} \right)^2$$

where, P_{avg} is the time-averaged laser power, λ_{TPE} is the two-photon excitation wavelength, NA is the numerical aperture of the objective, τ_p is the pulse duration and f_{rep} is the laser repetition rate, h is Planck's constant and c the speed of light. It can be seen that the efficiency of TPEF is dependent on the spatial focusing as well as the temporal pulse shape. These parameters must be optimized for successful TPEF imaging.

If the excitation rate exceeds the de-excitation rate, ground state depletion occurs. In other words, if the fluorescence lifetime of the fluorescent molecule ($\sim 10^{-9} \text{ s}$ for biological molecules) is greater than the fluorescence emission rate $\frac{1}{\langle n_f(t) \rangle}$, fluorescence emission eventually saturates. This means that increasing the imaging laser power is only beneficial up to a certain value. For greater laser powers, it is likely that the targeted molecules are raised to even higher excited states.

Second harmonic generation (SHG) is also a second order nonlinear process but differs from two-photon absorption in several aspects. In contrast to two-photon absorption, SHG involves an electron transition to a virtual (not a real) state, energy is conserved, and the SHG pulses are temporally synchronous with the laser excitation pulses. The emitted photon has twice the energy (half the wavelength) of the absorbed photons. SHG can only arise from media lacking a center of symmetry. In biological samples, SHG imaging has proven valuable for visualization of highly ordered structural proteins such as collagen without the need for exogenous fluorophores (Campagnola *et al.*, 2001). SHG and TPEF imaging can be performed with the very same microscope or ophthalmoscope. Both techniques provide complementary information.

Considerations, Artifacts and Implementation of Two-Photon Excited Fluorescence Ophthalmoscopy

Any imaging system is fundamentally limited in the smallest detail it can resolve. The resolution of an optical system is related to the full width at half maximum (FWHM) of its intensity point spread function (PSF) (Born and Wolf, 1999). The intensity I in the focal region of an aberration-free lens of circular aperture and numerical aperture $NA = n \sin \alpha$ is given as

$$I(u, v) = \left| 2 \int_0^1 J_0(v\rho) \exp\left(\frac{1}{2}iu\rho^2\right) \rho d\rho \right|^2$$

where J_0 is the zero-order Bessel function of the first kind, ρ is the radial coordinate in the pupil plane and u and v are axial and radial normalized coordinates given by

$$u = \frac{8\pi}{\lambda} z n \sin^2\left(\frac{\alpha}{2}\right) \text{ and } v = \frac{2\pi}{\lambda} r n \sin \alpha$$

where λ is the imaging wavelength, and z and r are the axial and lateral distance from the focus.

Intensity PSFs in the paraxial approximation are listed for different types of microscopes in Table 1 (Gu and Sheppard, 1995; Sheppard and Gu, 1990).

For simplicity, we consider that detection in a single-photon microscope occurs at a wavelength that is greater than the excitation wavelength $\lambda_{SPEF} = g \cdot \lambda_{SPE}$ ($g > 1$) and that the illuminating light in a two-photon microscope is of twice the wavelength of that used in a single-photon microscope $\lambda_{TPE} = 2\lambda_{SPE}$ to image the same fluorophore.

In a conventional single-photon microscope, the entrance aperture of the objective lens is uniformly illuminated and the detection is performed pointwise. The imaging PSF is only determined by the illumination PSF: $I_{SP} = I(u, v)$.

In a confocal single-photon microscope, both illumination and detection are performed pointwise. The imaging PSF is determined only by the illumination and the detection PSF: $I_{SP} = I^2(u, v)$.

To increase signal detection efficiency, most two-photon microscopes are built in the conventional, not the confocal configuration. Due to the quadratic dependence of a two-photon event on the excitation intensity, the PSF of such a conventional two-photon microscope is: $I_{TP} = I^2\left(\frac{u}{2}, \frac{v}{2}\right)$.

The lateral resolution for point-like objects can be calculated from the intensity distributions in the focal plane $z=0$, whereas the axial resolution for such objects is calculated from the intensity distribution along the optical axis ($r=0$).

Based on this assessment, the shorter the imaging wavelength and the greater the numerical aperture of the objective, the better is the resolution. For in vivo ophthalmoscopy, the NA is limited by the maximum diameter of the dilated pupil of the eye and therefore small compared to common microscope objectives. When the pupil has been dilated, the human and macaque eye has a numerical aperture of ~ 0.2 , whereas the mouse eye has a numerical aperture of ~ 0.5 (Geng et al., 2011). Therefore, the small angle approximation holds for ophthalmoscopy.

The lateral resolution of a conventional single-photon microscope is:

$$d_{SP,r} = \frac{3 \cdot 83}{2\pi} \frac{\lambda}{NA} \approx 0.61 \frac{\lambda}{NA}$$

The axial resolution is:

$$d_{SP,z} = \frac{\lambda}{2 n \sin^2 \frac{\alpha}{2}} = \frac{\lambda}{n(1 - \cos \alpha)} = \frac{\lambda}{n(1 - \sqrt{1 - \sin^2 \alpha})} \approx \frac{2\lambda n}{NA^2}$$

With the Taylor series for $\sin^2 \alpha < 1$, $\sqrt{1 - \sin^2 \alpha} \approx 1 - \frac{1}{2} \sin^2 \alpha$ and $NA = n \sin \alpha$. From the given PSFs in Table 1, it can be easily seen that spatial resolution when imaging point-like objects in a confocal single-photon microscope is improved by factor of $\frac{1}{\sqrt{2}}$, whereas in a conventional two-photon microscope it is worse by a factor of $\sqrt{2}$ with respect to a conventional single-photon microscope.

For a typical excitation wavelength of 365 nm when targeting retinol and the numerical aperture of the human eye, the spatial resolution limits for an optically-perfect eye would be 1.1 μm in a conventional and 0.8 μm in a confocal single-photon ophthalmoscopy and 1.6 μm in a two-photon ophthalmoscopy. Taking into account the refractive index of the vitreous of

Table 1 Comparison of PSFs for different types of microscopes.

	Conventional single-photon microscope	Confocal single-photon microscope	Conventional two-photon microscope
PSF	$I_{SP} = I(u, v)$	$I_{cSP} = I^2(u, v)$	$I_{TP} = I^2\left(\frac{u}{2}, \frac{v}{2}\right)$
Radial PSF ($z=0$)	$I_{SP,r}(v) = \left[\frac{2J_1(v)}{v}\right]^2$	$I_{cSP,r}(v) = \left[\frac{2J_1(v)}{v}\right]^4$	$I_{TP,r}(v) = \left[\frac{2J_1\left(\frac{v}{2}\right)}{\frac{v}{2}}\right]^4$
Axial PSF ($r=0$)	$I_{SP,a}(u) = \left[\frac{\sin\left(\frac{u}{4}\right)}{\frac{u}{4}}\right]^2$	$I_{cSP,a}(u) = \left[\frac{\sin\left(\frac{u}{4}\right)}{\frac{u}{4}}\right]^4$	$I_{TP,a}(u) = \left[\frac{\sin\left(\frac{u}{8}\right)}{\frac{u}{8}}\right]^2$

$n_0 = 1.336$, the axial resolution is 24 μm in a conventional and 17 μm in confocal single-photon ophthalmoscope and 34 μm in a two-photon ophthalmoscope. Even though lateral and axial resolution when imaging point-like objects in a two-photon microscope are worse compared to other microscope types discussed here, the real advantage is due to the optical sectioning when imaging thick specimens, because the background is greatly reduced.

In general, all optical systems exhibit some optical imperfections. Optical materials such as glass elements can greatly reduce the system efficiency if their optical properties are not adequately taken into account. In addition, the optics of the eye is far from being perfect. The most deleterious imperfections for nonlinear ophthalmoscopy are monochromatic aberrations, chromatic aberrations and material dispersion. Understanding and correcting for these optical aberrations is essential when optimizing the imaging system. Monochromatic aberrations are imperfections that are present even for a single wavelength. These aberrations result from imperfections in surface shape of the optical elements and local refractive index changes. Monochromatic aberrations of lower orders can be corrected for with spherical and cylindrical lenses, whereas the real-time correction of higher order aberrations requires adaptive optics (AO). This technique makes it possible to routinely resolve single photoreceptors with scanning light ophthalmoscopes (Dubra *et al.*, 2011). Excellent reviews on AO for vision science have been published (Porter *et al.*, 2006; Roorda and Duncan, 2015).

If the spectral bandwidth of the imaging light is broad, the overall optical quality of the system is affected by the wavelength-dependent refractive indices of the optical elements. The refractive index of the eye is greater for shorter than for longer wavelengths. Transverse chromatic aberration causes a change in magnification with wavelength. When imaging close to the optical axis or in the temporal near-periphery, transverse chromatic aberration is generally small (Winter *et al.*, 2016). Longitudinal chromatic aberration causes light of different wavelengths to focus at different distances from the lens (Thibos *et al.*, 1992; Vinas *et al.*, 2015) and can be statically corrected with an achromatizing lens, emitting the opposite chromatic aberration of the eye (Fernández *et al.*, 2006). The simplest design is a symmetrical triplet compensating for the chromatic aberration of the eye to be imaged, when placed in a conjugate pupil plane. Proper alignment of the achromatizing lens with the pupil of the eye is important to avoid introducing additional transverse chromatic aberration.

Furthermore, the speed of light in the ocular media changes with wavelength, which results in light from a broadband source not arriving at the same time on the retina, and therefore in an increase in pulse duration. The broadened pulse duration can be estimated via $\tau_{out} = \tau_{in} \sqrt{1 + 7.68 \left(\frac{D \cdot L}{\tau_{in}^2} \right)^2}$, where L is the distance traversed within the optical media and D is the optical dispersion parameter. At 725 nm, the optical dispersion value of the vitreous humor is $32.2 \frac{\text{fs}^2}{\text{mm}}$, whereas that of the cornea-lens complex is $183 \frac{\text{fs}^2}{\text{mm}}$, for instance (Coello *et al.*, 2007).

Several approaches exist for dispersion correction. Second-order dispersion can be corrected by a prism pair, which is sufficient for pulses with a bandwidth of less than 30 nm. For lasers with larger bandwidths additional correction of higher-order dispersion may considerably increase TPEF efficiency. For this purpose, commercial 4 f pulse shapers that optimize dispersion pre-compensation via multiphoton intrapulse interference phase scan (MIIPS) (Lozovoy *et al.*, 2004; Pestov *et al.*, 2010) have been used with microscopes. MIIPS relies on a high-precision grating in combination with a spatial light modulator (SLM). The grating spectrally disperses the beam on the way to the SLM and recombines the spectral components on the way back. At the SLM, every pixel corresponds to a different spectral range and the phase of the pulse can be independently modulated for different wavelengths. In this way, the SLM has the capability to shape the temporal profile of the pulse. Traditionally, the SHG signal after transmission through a thin nonlinear crystal is used for optimization of two-photon microscopes. In the case of a two-photon ophthalmoscope, only second-order dispersion has been corrected using the TPEF signal from the retina as feedback (Hunter *et al.*, 2011).

Effects of Femtosecond Lasers on the Retina

A disadvantage of fluorescence imaging, and especially TPEF, is that the *in vivo* retina must be exposed to much higher light levels than are used in conventional reflectance imaging. Roughly speaking, *in vivo* single photon imaging generally requires exposures that are ~ 1 order of magnitude and TPEF requires ~ 4 orders of magnitude more light than reflectance imaging. The need to address safety in TPEFO is all the more pressing given the indication that diseased retina might be more susceptible to light damage than healthy retina, which is based on a study in dogs suffering a rhodopsin mutation as a model for retinitis pigmentosa (Cideciyan *et al.*, 2005). Efficient two-photon imaging currently requires the use of sub-100 fs pulsed lasers, but due to a lack of data, safety standards for these light exposures are not well established (ANSI Z136.1-2014). The use of a pulsed light source for non-linear retinal imaging generates peak powers that have the potential to cause thermal, mechanical, and photochemical damage (Boulton *et al.*, 2001; Rockwell *et al.*, 2010) or some combination of these mechanisms as well as transient effects.

Thermal Damage

Excess heating of the tissue ($> 20^\circ\text{C}$) leads to immediate thermal damage, seen as local coagulation and an inflammation response. Thermal lesions usually occur with exposure durations on the time scale of μs up to 10 s with a $\text{time}^{2/3}$ dependence of the damage threshold upon exposure duration for visible and near infrared laser exposures (Lund and Edsall, 1999). Cell death may be attributed to protein denaturation (Mellerio, 1994). In the retina, the action spectrum for minimal thermal damage is

consistent with the absorption spectrum of melanin (Lund *et al.*, 2008). The time for heat dissipation from a melanin granule in the retina is treated as 5 μ s (ANSI Z136.1-2014). The ultrashort femtosecond lasers used for TPE have a pulse repetition frequency faster than that cooling time and are often treated as continuous-wave (CW) lasers at the average power for the purposes of safety calculations in addition to considering the safety of an individual pulse corrected by the multiple pulse correction factor. Specific safety standard calculations are outside the scope of this review.

Mechanical Damage

Ultrashort pulse lasers emit peak powers that have the potential to cause mechanical damage. Mechanical (or acoustic) damage is due to compression or tension within the tissue (Youssef *et al.*, 2011) after energy deposition that occurs more rapidly than dissipation is possible. On the one hand, the nonlinear dependence of the refractive index on irradiance can cause self-focusing of the laser beam. In the extreme case, the beam diameter is decreased below the diffraction limit. On the other hand, high peak powers can cause ionization and thus plasma formation. Plasma has absorptive properties different from those of ordinary gases. The critical power for self-focusing and plasma formation can be calculated and compared to ensure they are above the estimated peak power for an ocular exposure at a given wavelength.

Self-focusing

When a short pulse propagates through a nonlinear medium, the Kerr effect leads to a phase delay which is largest on the beam axis (where the optical density is highest) and smaller outside the axis. This is similar to the action of a lens. Self-focusing occurs if the laser peak power approaches the critical power (Boyd *et al.*, 2009; Marburger, 1975) defined by:

$$P_{cr} = \frac{0.148 \cdot \lambda^2}{n_0 \cdot n_2}$$

where the linear and nonlinear refractive indices for vitreous humor are $n_0 = 1.336$ and $n_2 = 43.9 \cdot 10^{-21} \frac{m^2}{W}$ (Rockwell *et al.*, 1993), respectively.

Plasma Formation

Laser-induced breakdown (LIB) is the partial or complete ionization of a solid, liquid or gas through absorption of laser energy. The ionization results in a gas of charged particles, a plasma, which absorbs optical radiation much more strongly than ordinary matter. The plasma can thus be rapidly heated by the laser beam to very high temperatures, producing plasma expansion, an audible acoustic signature and a visible plasma emission. There are two mechanisms which can lead to LIB: direct ionization of the medium by multiphoton absorption or cascade ionization where free electron(s) have to be present at the beginning of the pulse. In the femtosecond regime only “pure” multiphoton ionization is fast enough to allow breakdown to occur over the duration of the pulse. For laser-induced breakdown with subsequent bubble formation, a critical electron density of 10^{24} electrons cm^{-3} is needed, whereas for laser-induced breakdown without subsequent bubble formation, a critical electron density of 10^{22} electrons cm^{-3} is sufficient. Low-density plasmas may also be a mechanism for femtosecond pulse tissue damage (Vogel *et al.*, 2002a,b). The critical electron density associated with significant optical absorption in the plasma is 10^{18} electrons per cm^3 .

The irradiance threshold for plasma formation in the femtosecond laser pulse regime in vitreous humor can be calculated (Schwarz *et al.*, 2016) following the example of Cain *et al.* (2005). Calculations are based on the first order model of Kennedy (Kennedy, 1995; Kennedy *et al.*, 1995b), with special attention to Kennedy (1995) sections IV D and V F. For the ultrashort pulse regime, this model relies on the theory of multiphoton ionization in condensed media (Keldysh, 1965). The irradiance threshold for multiphoton breakdown as derived by Kennedy is:

$$I_{cr} = \frac{\left(\frac{\rho_{cr}}{A \cdot 0.5 \cdot \tau_p} \right)^{1/K}}{B}$$

with ρ_{cr} being the critical free electron density explained below, $K = \langle 1 + \frac{\Delta}{\hbar \cdot \omega} \rangle = 4$ the number of photons required to ionize the medium when $\Delta = 6.5$ eV is the ionization energy of the medium and $\omega = \frac{2 \cdot \pi \cdot c}{\lambda}$ the optical frequency where $c = 3 \cdot 10^8 \frac{m}{s}$ is the vacuum speed of light, λ the wavelength and the parameters

$$A = \left(\frac{2}{9 \cdot \pi} \right) \cdot \omega \cdot \left(\frac{m' \cdot \omega}{\hbar} \right)^{\frac{3}{2}} \cdot e^{2K} \cdot \Phi(z) \cdot \left(\frac{1}{16} \right)^K$$

and

$$B = \frac{q^2}{m' \cdot \Delta \cdot \omega^2 \cdot c \cdot \epsilon_0 \cdot n_0}$$

where ω , K and Δ are the same as above, $\epsilon_0 = 8.85 \cdot 10^{-12} \frac{C}{Vm}$ is the permittivity of free space, $n_0 = 1.336$ the index of refraction of the medium at frequency ω , $m' = 4.55 \cdot 10^{-31}$ kg the exciton reduced mass, $\hbar = 6.58 \cdot 10^{-16}$ eVs the reduced Planck constant, $q = 1.6 \cdot 10^{-19}$ C the electron charge, and $\Phi(z) = e^{-z^2} \cdot \sum_{n=0}^{\infty} \frac{z^{2n+1}}{n!(2n+1)}$ represents Dawson's Integral with $z = \sqrt{2 \cdot K - \frac{2 \cdot \Delta}{\hbar \cdot \omega}}$.

Photochemical Damage

Photochemical damage occurs when the energy in a photon of light is strong enough to break chemical bonds in the irradiated molecules. The bond breakage can result in conformational changes to a molecule or in the formation of free radicals and reactive oxygen species. These molecular changes can eventually result in cell damage or death. Photochemical damage is usually observed with visible light exposures longer than a second. However, recent results suggest photochemical damage to the RPE is also possible using near infrared light (Zhang *et al.*, 2016). The particular molecules involved in photochemical damage are expected to vary with exposure wavelength.

As a consequence of the nonlinear optical process underlying two-photon excitation, light in the UV range of wavelengths that is usually blocked by the ocular media (Boettner and Wolter, 1962) can be generated at the retina. This is observed in the sensitivity to and color appearance of near-infrared pulsed lasers as compared to CW sources of the same wavelength (Palczewska *et al.*, 2014b). It is not known what impact the increased proportion of UV and visible light in combination with IR light might have on retinal structure and function. Recent evidence suggests that threshold damage as a result of sub-100 fs light at 730 nm 5 times above the ANSI standard (ANSI Z136.1-2014) corresponds to a loss of S-cones (Schwarz *et al.*, 2017), possibly via a photochemical mechanism.

Infrared Autofluorescence Reduction

Exposure to infrared illumination required for TPEFO causes a long-lasting reduction in infrared autofluorescence (IRAF) (Masella *et al.*, 2014b). Fundus IRAF is thought to originate mainly from the choroid and RPE with a slightly higher relative contribution of the choroid in older and darkly pigmented subjects. Candidate fluorophores are oxidized melanin and components closely related to melanin. IRAF reduction can occur with pulsed and CW infrared light exposures (Schwarz *et al.*, 2016) and appears to be consistent with a photochemical mechanism. In non-human primate, partial recovery was seen at 1 month, with full recovery within 21 months. This long-lasting effect of infrared illumination in both humans and monkeys occurs at exposure levels four to five times below current safety limits. Possible explanations are therefore the bleaching of melanin or changes in fluorescence efficiency of oxidized melanin secondary to changes in the environment. To date, there is no evidence of IRAF reduction causing a clinically significant impact on retinal health or visual function.

Outlook

In vivo multiphoton fluorescence ophthalmoscopy is a growing field for retinal imaging of intrinsic and extrinsic fluorophores because of its ability to provide insight into both retinal structure and function. The ability to image fluorescence from both NAD(P)H and FAD opens the possibility for reduction–oxidation (redox) measurements (Chance and Thorell, 1959; Skala *et al.*, 2007), an indication as to the state of cellular respiration. Whether a cell is favoring aerobic or anaerobic metabolism can be an indication of its health. Further insight can be gained by tracking of fluorescent properties such as emission spectra and fluorescence lifetime. The composition of fluorophores contributing to the emission signal may be identified by their spectral contribution. Fluorescence lifetime is a measure of the average delay between fluorescence excitation and the time at which photons are emitted (Venetta, 1959). The lifetime varies with the conformational state of the fluorophore and its microenvironment, which may provide a means to distinguish healthy from unhealthy cells. For example, NADH exists in both free and protein-bound forms with identical emission spectra, but different fluorescence lifetimes (0.3 ns and 2.36 ns, respectively) (Skala *et al.*, 2007). Altered lifetimes of NADH and FAD have been demonstrated in precancerous cells, likely related to the increased reliance on glycolysis in cancer cells (Skala *et al.*, 2007). Using non-cellular scale single-photon fluorescence lifetime imaging ophthalmoscopy (Schweitzer *et al.*, 2004), differences from control in ocular fluorescence lifetime have been reported for AMD (Schweitzer *et al.*, 2009), Stargardt disease (Dysli *et al.*, 2016), Alzheimer's disease (Jentsch *et al.*, 2015) and diabetes (Schweitzer *et al.*, 2011). However, measurements of fluorescence lifetime using TPEF ophthalmoscopy of the non-human primate photoreceptor mosaic have only recently been reported (Feeks *et al.*, 2017). Furthermore, fluorescence lifetime of extrinsic fluorescent markers can provide insight into particular cellular properties such as pH or the ratio of cytosolic NADH: NAD⁺ (Mongeon *et al.*, 2016; Yellen and Mongeon, 2015).

TPEF ophthalmoscopy has the potential to be an effective clinical measure of photoreceptor and RPE health and provide early detection of outer retinal dysfunction. This requires the safe translation of this imaging modality for use in humans. Schwarz *et al.* (2016) have demonstrated in macaque that TPEF measurements of photoreceptor function are feasible at light levels within safety standards. Using a variety of clinical and high resolution imaging methods, such as fundus photography, reflectance and fluorescence cSLO, OCT, fluorescein angiography, and AO imaging of photoreceptors and RPE, IRAF reduction was the only significant effect. No other structural or functional changes were detected for any of the exposures, though of course we cannot exclude the possibility that changes could be detected with more sensitive tests or longer follow-up. No effect of repeated exposures on TPEF time course could be detected, suggesting that visual cycle function is maintained. Photoreceptors and RPE appeared normal in adaptive optics images. This safety study suggests that TPEF ophthalmoscopy will soon be feasible in people.

Acknowledgements

The authors wish to thank Robin Sharma, Sarah Walters, James Feeks, Grazyna Palczewska, Krzysztof Palczewski and David R. Williams. This work was funded by grants NIH/NEI P30-EY001319, R01-EY022371, R01-EY004367, U01-EY025497, and EY007125, and an Unrestricted Grant to the University of Rochester, Department of Ophthalmology from Research to Prevent Blindness, New York, NY, the Beckman-Argyros Award and the Alcon Research Award.

See also: Probing Ocular Mechanics With Light

References

- Ala-Laurila, P., Kolesnikov, A.V., Crouch, R.K., *et al.*, 2006. Visual cycle: Dependence of retinol production and removal on photoproduct decay and cell morphology. *The Journal of General Physiology* 128 (2), 153–169. <http://doi.org/10.1085/jgp.200609557>.
- Atchison, D.A., Smith, G., 2000. *Optics of the Human Eye*. Butterworth Heinemann.
- Bar-Noam, A.S., Farah, N., Shoham, S., 2016. Correction-free remotely scanned two-photon in vivo mouse retinal imaging. *Light: Science & Applications* 5 (1), e16007. <http://doi.org/10.1038/lsa.2016.7>.
- Bindewald-Wittich, A., Han, M., Schmitz-Valckenberg, S., *et al.*, 2006. Two-photon-excited fluorescence imaging of human RPE cells with a femtosecond Ti: Sapphire laser. *Investigative Ophthalmology & Visual Science* 47 (10), 4553–4557. <http://doi.org/10.1167/iovs.05-1562>.
- Boettner, E.A., Wolter, J.R., 1962. Transmission of the ocular media. *Investigative Ophthalmology & Visual Science* 1 (6), 776–783.
- Born, M., Wolf, E., 1999. *Principles of Optics*. Cambridge University Press.
- Boulton, M., Rózanowska, M., Rózanowski, B., 2001. Retinal photodamage. *Journal of Photochemistry and Photobiology B: Biology* 64 (2–3), 144–161. [http://doi.org/10.1016/S1011-1344\(01\)00227-5](http://doi.org/10.1016/S1011-1344(01)00227-5).
- Boyd, R.W., Lukishova, S.G., Shen, Y.R., 2009. Self-focusing: Past and Present: Fundamentals and Prospects.
- Cain, C.P., Thomas, R.J., Noojin, G.D., *et al.*, 2005. Sub-50-fs laser retinal damage thresholds in primate eyes with group velocity dispersion, self-focusing and low-density plasmas. *Graefes Archive for Clinical and Experimental Ophthalmology* 243 (2), 101–112. <http://doi.org/10.1007/s00417-004-0924-9>.
- Campagnola, P.J., Clark, H.A., Mohler, W.A., Lewis, A., Loew, L.M., 2001. Second-harmonic imaging microscopy of living cells. *Journal of Biomedical Optics* 6 (3), 277–286. <http://doi.org/10.1117/1.1383294>.
- Campagnola, P.J., Wei, M.D., Lewis, A., Loew, L.M., 1999. High-resolution nonlinear optical imaging of live cells by second harmonic generation. *Biophysical Journal* 77 (6), 3341–3349. [http://doi.org/10.1016/S0006-3495\(99\)77165-1](http://doi.org/10.1016/S0006-3495(99)77165-1).
- Chance, B., Thorell, B., 1959. Localization and kinetics in living of reduced pyridine by microfluorometry. *The Journal of Biological Chemistry* 234 (11), 3044–3050.
- Chen, C., Tsina, E., Cornwall, M.C., *et al.*, 2005. Reduction of all-trans retinal to all-trans retinol in the outer segments of frog and mouse rod photoreceptors. *Biophysical Journal* 88 (3), 2278–2287. <http://doi.org/10.1529/biophysj.104.054254>.
- Cideciyan, A.V., Jacobson, S.G., Aleman, T.S., *et al.*, 2005. In vivo dynamics of retinal injury and repair in the rhodopsin mutant dog model of human retinitis pigmentosa. *Proceedings of the National Academy of Sciences of the United States of America* 102 (14), 5233–5238. <http://doi.org/10.1073/pnas.0408892102>.
- Coello, Y., Xu, B., Miller, T.L., Lozovoy, V.V., Dantus, M., 2007. Group-velocity dispersion measurements of water, seawater, and ocular components using multiphoton intrapulse interference phase scan. *Applied Optics* 46 (35), 8394–8401.
- Delori, F.C., Dorey, C.K., Staurengi, G., *et al.*, 1995. In vivo fluorescence of the ocular fundus exhibits retinal pigment epithelium lipofuscin characteristics. *Investigative Ophthalmology & Visual Science* 36 (3), 718–729.
- Denk, W., Detwiler, P.B., 1999. Optical recording of light-evoked calcium signals in the functionally intact retina. *Proceedings of the National Academy of Sciences of the United States of America* 96 (12), 7035–7040. <http://doi.org/10.1073/pnas.96.12.7035>.
- Denk, W., Piston, D.W., Webb, W.W., 2006. Multi-photon molecular excitation in laser-scanning microscopy. *Handbook of Biological Confocal Microscopy: Third Edition*. 535–549. http://doi.org/10.1007/978-0-387-45524-2_28.
- Denk, W., Strickler, J.H., Webb, W.W., 1990. Two-photon laser scanning fluorescence microscopy. *Science (New York, N. Y.)* 248 (4951), 73–76. <http://doi.org/10.1126/science.2321027>.
- Denk, W., Sugi-mori, M., Llinás, R., 1995. Two types of calcium response limited to single spines in cerebellar Purkinje cells. *Proceedings of the National Academy of Sciences of the United States of America* 92 (18), 8279–8282. <http://doi.org/10.1073/pnas.92.18.8279>.
- Diaspro, A., 2002. *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*.
- Diaspro, A., Chirico, G., 2003. Two-photon excitation microscopy. *Advances in Imaging and Electron Physics* 126, 195–286.
- Diaspro, A., Chirico, G., Collini, M., 2005. Two-photon fluorescence excitation and related techniques in biological microscopy. *Quarterly Reviews of Biophysics* 38 (2), 97–166. <http://doi.org/10.1017/S0033583505004129>.
- Dubra, A., Sulai, Y., Norris, J.L., *et al.*, 2011. Noninvasive imaging of the human rod photoreceptor mosaic using a confocal adaptive optics scanning ophthalmoscope. *Biomedical Optics Express* 2 (7), 1864–1876. <http://doi.org/10.1364/BOE.2.001864>.
- Dysli, C., Wolf, S., Hatz, K., Zinkernagel, M.S., 2016. Fluorescence lifetime imaging in Stargardt disease: Potential marker for disease progression. *Investigative Ophthalmology & Visual Science* 57 (3), 832. <http://doi.org/10.1167/iovs.15-18033>.
- Feeks, J.A., Walters, S., Schwarz, C., Hunter, J.J., 2017. Two-photon fluorescence lifetime ophthalmoscopy of intrinsic fluorophores on a cellular scale in the living macaque. In: *ARVO Meeting* 2017.
- Fernández, E.J., Unterhuber, A., Povazay, B., *et al.*, 2006. Chromatic aberration correction of the human eye for retinal imaging in the near infrared. *Optics Express* 14 (13), 6213–6225.
- Geng, Y., Schery, L.A., Sharma, R., *et al.*, 2011. Optical properties of the mouse eye. *Biomedical Optics Express* 2 (4), 717–738. <http://doi.org/10.1364/BOE.2.000717>.
- Göppert, M., 1929. Über die Wahrscheinlichkeit des Zusammenwirkens zweier Lichtquanten in einem Elementarakt. *Die Naturwissenschaften* 17 (48), 932. <http://doi.org/10.1007/BF01506585>.
- Göppert-Mayer, M., 1931. Über Elementarakte mit zwei Quantensprüngen. *Annalen Der Physik* 114, 273–294.
- Granger, C., Williams, D.R., Rossi, E.A., 2017. Near-infrared autofluorescence imaging reveals the retinal pigment epithelial mosaic in the living human eye. In: *ARVO Meeting* 2017.
- Grieve, K., Roorda, A., 2008. Intrinsic signals from human cone photoreceptors. *Investigative Ophthalmology and Visual Science* 49 (2), 713–719. <http://doi.org/10.1167/iovs.07-0837>.
- Gualda, E.J., Bueno, J.M., Artal, P., 2010. Wavefront optimized nonlinear microscopy of ex vivo human retinas. *Journal of Biomedical Optics* 15 (2), 26007. <http://doi.org/10.1117/1.3369001>.

- Gu, M., Sheppard, C.J.R., 1995. Comparison of three-dimensional imaging properties between two-photon and single-photon fluorescence microscopy. *Journal of Microscopy* 177 (2), 128–137. <http://doi.org/10.1111/j.1365-2818.1995.tb03543.x>.
- Guevara-Torres, A., Joseph, A., Schalliek, J.B., 2016. Label free measurement of retinal blood cell flux, velocity, hematocrit and capillary width in the living mouse eye. *Biomedical Optics Express* (BOE) 7, 4228–4249.
- Helmchen, F., 2009. Two-photon functional imaging of neuronal activity. In: Frostig, R.D. (Ed.), *In Vivo Optical Imaging of Brain Function*. CRC Press, p. 428.
- Huang, S., Heikal, A.A., Webb, W.W., 2002. Two-photon fluorescence spectroscopy and microscopy of NAD(P)H and flavoprotein. *Biophysical Journal* 82 (5), 2811–2825. [http://doi.org/10.1016/S0006-3495\(02\)75621-X](http://doi.org/10.1016/S0006-3495(02)75621-X).
- Hunter, J.J., Masella, B., Dubra, A., *et al.*, 2011. Images of photoreceptors in living primate eyes using adaptive optics two-photon ophthalmoscopy. *Biomedical Optics Express* 2 (1), 139–148. <http://doi.org/10.1364/BOE.2.000139>.
- Imanishi, Y., Batten, M.L., Piston, D.W., Baehr, W., Palczewski, K., 2004. Noninvasive two-photon imaging reveals retinyl ester storage structures in the eye. *The Journal of Cell Biology* 164 (3), 373–383. <http://doi.org/10.1083/jcb.200311079>.
- Imanishi, Y., Lodowski, K., Koutalos, Y., 2007. Two-photon microscopy: Shedding light on the chemistry of vision. *Biochemistry* 46 (34), 9674–9684. <http://doi.org/10.1021/bi701055g>.
- Jentsch, S., Schweitzer, D., Schmidtko, K.U., *et al.*, 2015. Retinal fluorescence lifetime imaging ophthalmoscopy measures depend on the severity of Alzheimer's disease. *Acta Ophthalmologica* 93 (4), e241–e247. <http://doi.org/10.1111/aos.12609>.
- Kaplan, M.W., 1985. Distribution and axial diffusion of retinol in bleached rod outer segments of frogs (*Rana pipiens*). *Experimental Eye Research* 40 (5), 721–729. [http://doi.org/10.1016/0014-4835\(85\)90141-1](http://doi.org/10.1016/0014-4835(85)90141-1).
- Keldysh, L.V., 1965. Ionization in the field of a strong electromagnetic wave. *Soviet Physics JETP* 20 (5), 1307–1314. <http://doi.org/10.1234/12345678>.
- Kennedy, C.J., Rakoczy, P.E., Constable, I.J., 1995a. Lipofuscin of the retinal pigment epithelium: A review. *Eye* (London, England) 9, 763–771. <http://doi.org/10.1038/eye.1995.192>.
- Kennedy, P.K., 1995. A first-order model for computation of laser-induced breakdown thresholds in ocular and aqueous media: Part I – Theory. *IEEE Journal of Quantum Electronics* 31 (12), 2241–2249.
- Kennedy, P.K., Boppart, S.A., Hammer, D.X., *et al.*, 1995b. A first-order model for computation of laser-induced breakdown thresholds in ocular and aqueous media: ii – Comparison to experiment. *IEEE Journal of Quantum Electronics* 31 (12), 2250–2257.
- Kiser, P.D., Golczak, M., Maeda, A., Palczewski, K., 2012. Key enzymes of the retinoid (visual) cycle in vertebrate retina. *Biochimica et Biophysica Acta – Molecular and Cell Biology of Lipids* 1821 (1), 137–151. <http://doi.org/10.1016/j.bbalip.2011.03.005>.
- Kiser, P.D., Golczak, M., Palczewski, K., 2014. Chemistry of the retinoid (visual) cycle. *Chemical Reviews* 114, 194–232. <http://doi.org/10.1021/cr400107q>.
- Lamb, T.D., Pugh, E.N., 2004. Dark adaptation and the retinoid cycle of vision. *Progress in Retinal and Eye Research* 23 (3), 307–380. <http://doi.org/10.1016/j.preteyeres.2004.03.001>.
- Liebman, P.A., 1969. Microspectrophotometry of retinal cells. *Annals of the New York Academy of Sciences* 157, 250–264.
- Liebman, P.A., 1973. Microspectrophotometry of visual receptors. In: Langer, H. (Ed.), *Biochemistry and Physiology of Visual Pigments*. Berlin: Springer-Verlag, pp. 299–305.
- Louisell, W.H., 1973. Self-focusing: Theory. *Progress in Quantum Electronics* 4, 35–110.
- Lozovoy, V.V., Pastirk, I., Dantus, M., 2004. Multiphoton intrapulse interference. IV. Ultrashort laser pulse spectral phase characterization and compensation. *Optics Letters* 29 (7), 775–777.
- Lund, D.J., Edsall, P., 1999. Action spectrum for retinal thermal injury. *SPIE* 3591, 324–334.
- Lund, D.J., Edsall, P., Stuck, B.E., 2008. Spectral dependence of retinal thermal injury. *Journal of Laser Applications* 20 (2), 76. <http://doi.org/10.2351/1.2900534>.
- Marburger, J.H., 1975. Self-focusing: Theory. *Progress in Quantum Electronics* 4, 35–110.
- Masella, B.D., Hunter, J.J., Williams, D.R., 2014a. New wrinkles in retinal densitometry. *Investigative Ophthalmology & Visual Science* 55 (11), 7525–7534. <http://doi.org/10.1167/iovs.13-13795>.
- Masella, B.D., Williams, D.R., Fischer, W.S., Rossi, E.A., Hunter, J.J., 2014b. Long-term reduction in infrared autofluorescence caused by infrared light below the maximum permissible exposure. *Investigative Ophthalmology and Visual Science* 55 (6), 3929–3938. <http://doi.org/10.1167/iovs.13-12562>.
- Masters, B.R., So, P.T.C., 2008. *Handbook of Biomedical Nonlinear Optical Microscopy*.
- Mellerio, J., 1994. Light effects on the retina. *Principles and Practice of Ophthalmology*. 1–23.
- Mongeon, R., Venkatachalam, V., Yellen, G., 2016. Cytosolic NADH-NAD⁺ redox visualized in brain slices by two-photon fluorescence lifetime biosensor imaging. *Antioxidants & Redox Signaling* 25 (10), 553–563. <http://doi.org/10.1089/ars.2015.6593>.
- Morgan, J.I.W., Dubra, A., Wolfe, R., Merigan, W.H., Williams, D.R., 2009. In vivo autofluorescence imaging of the human and macaque retinal pigment epithelial cell mosaic. *Investigative Ophthalmology & Visual Science* 50 (3), 1350–1359. <http://doi.org/10.1167/iovs.08-2618>.
- Palczewska, G., Dong, Z., Golczak, M., *et al.*, 2014a. Noninvasive two-photon microscopy imaging of mouse retina and retinal pigment epithelium through the pupil of the eye. *Nature Medicine* 20 (7), 785–789. <http://doi.org/10.1038/nm.3590>.
- Palczewska, G., Maeda, T., Imanishi, Y., *et al.*, 2010. Noninvasive multiphoton fluorescence microscopy resolves retinol and retinal condensation products in mouse eyes. *Nature Medicine* 16 (12), 1444–1449. <http://doi.org/10.1038/nm.2260>.
- Palczewska, G., Vinberg, F., Stremplewski, P., *et al.*, 2014b. Human infrared vision is triggered by two-photon chromophore isomerization. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1410162111.
- Pestov, D., Lozovoy, V.V., Dantus, M., 2010. Single-beam shaper-based pulse characterization and compression using MIIPS sonogram. *Optics Letters* 35 (9), 1422–1424.
- Porter, J., Queener, H., Lin, J., Thorn, K., Awwal, A., 2006. *Adaptive Optics for Vision Science: Principles, Practices, Design and Applications*.
- Quigley, H.A., 2011. Glaucoma. *The Lancet* 377 (9774), 1367–1377. [http://doi.org/10.1016/S0140-6736\(10\)61423-7](http://doi.org/10.1016/S0140-6736(10)61423-7).
- Rockwell, B.A., Roach, W.P., Rogers, M.E., *et al.*, 1993. Nonlinear refraction in vitreous humor. *Optics Letters* 18 (21), 1792–1794. <http://doi.org/10.1364/OL.18.001792>.
- Rockwell, B.A., Thomas, R.J., Vogel, A., 2010. Ultrashort laser pulse retinal damage mechanisms and their impact on thresholds. *Medical Laser Application* 25 (704), 84–92. <http://doi.org/10.1016/j.mla.2010.02.002>.
- Roorda, A., Duncan, J.L., 2015. Adaptive optics ophthalmoscopy. *Annual Review of Vision Science* 1 (1), 19–50. <http://doi.org/10.1146/annurev-vision-082114-035357>.
- Rossi, E.A., Granger, C.E., Sharma, R., *et al.*, 2017. Imaging individual neurons in the retinal ganglion cell layer of the living eye. *Proceedings of the National Academy of Sciences* 114 (3), 201613445. <http://doi.org/10.1073/pnas.1613445114>.
- Saari, J.C., 2000. Biochemistry of visual pigment regeneration – The Friedenwald Lecture. *Investigative Ophthalmology* 41 (2).
- Schwarz, C., Sharma, R., Fischer, W.S., *et al.*, 2016. Safety assessment in macaques of light exposures for functional two-photon ophthalmoscopy in humans. *Biomedical Optics Express* 7 (12), 5148–5169.
- Schwarz, C., Sharma, R., Keller, M., Williams, D.R., Hunter, J.J., 2017. Intense ultrashort pulsed light in the infrared selectively damages putative S cones. In: *ARVO Meeting* 2017.
- Schweitzer, D., Hammer, M., Schweitzer, F., *et al.*, 2004. In vivo measurement of time-resolved autofluorescence at the human fundus. *Journal of Biomedical Optics* 9 (6), 1214. <http://doi.org/10.1117/1.1806833>.
- Schweitzer, D., Klemm, M., Quick, S., *et al.*, 2011. Detection of early metabolic alterations in the ocular fundus of diabetic patients by time-resolved autofluorescence of endogenous fluorophores. *Proceedings of SPIE* 8087, 1–9. <http://doi.org/10.1117/12.889109>.
- Schweitzer, D., Quick, S., Schenke, S., *et al.*, 2009. Comparison of parameters of time-resolved autofluorescence between healthy subjects and patients suffering from early AMD. *Der Ophthalmologe* 106 (8), 714–722.

- Scoles, D., Sulai, Y.N., Langlo, C.S., *et al.*, 2014. *In vivo* imaging of human cone photoreceptor inner segments. *Investigative Ophthalmology and Visual Science* 55 (7), 4244–4251.
- Sharma, R., Hunter, J.J., 2017. Multiphoton imaging of the retina. *Handbook of Visual Optics, Volume Two: Instrumentation and Vision Correction*. 392.
- Sharma, R., Schwarz, C., Hunter, J.J., *et al.*, 2017. Formation and clearance of all- *trans* -retinol in rods investigated in the living primate eye with two-photon ophthalmoscopy. *Investigative Ophthalmology & Visual Science* 58 (1), 604. <http://doi.org/10.1167/iov.16-20061>.
- Sharma, R., Schwarz, C., Williams, D.R., *et al.*, 2016a. *In vivo* two-photon fluorescence kinetics of primate rods and cones. *Investigative Ophthalmology & Visual Science* 57 (2), 647–657. <http://doi.org/10.1167/iov.15-17946>.
- Sharma, R., Williams, D.R., Palczewska, G., Palczewski, K., Hunter, J.J., 2016b. Two-photon autofluorescence imaging reveals cellular structures throughout the retina of the living primate eye. *Investigative Ophthalmology & Visual Science* 57 (2), 632. <http://doi.org/10.1167/iov.15-17961>.
- Sheppard, C.J.R., Gu, M., 1990. Image formation in two-photon fluorescence microscopy. *Optik* 86 (3), 379–386. <http://doi.org/10.1016/j.neuint.2009.12.015>.
- Shuttleworth, C.W., 2010. Use of NAD(P)H and flavoprotein autofluorescence transients to probe neuron and astrocyte responses to synaptic activation. *Neurochemistry International* 56 (3), 379–386. <http://doi.org/10.1016/j.neuint.2009.12.015>.
- Skala, M.C., Ricking, K.M., Gendron-Fitzpatrick, A., *et al.*, 2007. *In vivo* multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia. *Proceedings of the National Academy of Sciences* 104 (49), 19494–19499. <http://doi.org/10.1073/pnas.0708425104>.
- So, P.T.C., Dong, C.Y., Masters, B.R., Berland, K.M., 2000. Two-photon excitation fluorescence microscopy. *Annual Review of Biomedical Engineering* 2, 399–429.
- Svoboda, K., Denk, W., Kleinfeld, D., Tank, D.W., 1997. *In vivo* dendritic calcium dynamics in neocortical pyramidal neurons. *Nature* 385 (6612), 161–165. <http://doi.org/10.1038/385161a0>.
- Svoboda, K., Yasuda, R., 2006. Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron* 50 (6), 823–839. <http://doi.org/10.1016/j.neuron.2006.05.019>.
- Thibos, L.N., Ye, M., Zhang, X., Bradley, A., 1992. The chromatic eye: A new reduced-eye model of ocular chromatic aberration in humans. *Applied Optics* 31 (19), 3594–3600.
- Thompson, D.A., Gal, A., 2003. Vitamin A metabolism in the retinal pigment epithelium: Genes, mutations, and diseases. *Progress in Retinal and Eye Research* 22 (5), 683–703. [http://doi.org/10.1016/S1350-9462\(03\)00051-X](http://doi.org/10.1016/S1350-9462(03)00051-X).
- Travis, G.H., Golczak, M., Moise, A.R., Palczewski, K., 2007. Diseases caused by defects in the visual cycle: Retinoids as potential therapeutic agents. *Annual Review of Pharmacology and Toxicology* 47, 469–512. <http://doi.org/10.1146/annurev.pharmtox.47.120505.105225>.
- Tsina, E., Chen, C., Koutalos, Y., *et al.*, 2004. Physiological and microfluorometric studies of reduction and clearance of retinal in bleached rod photoreceptors. *The Journal of General Physiology* 124 (4), 429–443. <http://doi.org/10.1085/jgp.200409078>.
- Venetia, B.D., 1959. Microscope phase fluorometer for determining the fluorescence lifetimes of fluorochromes. *Review of Scientific Instruments* 30 (6), 450–457. <http://doi.org/10.1063/1.1716652>.
- Vinas, M., Dorronsoro, C., Cortes, D., Pascual, D., Marcos, S., 2015. Longitudinal chromatic aberration of the human eye in the visible and near infrared from wavefront sensing, double-pass and psychophysics. *Biomedical Optics Express* 23 (4), 513–522. <http://doi.org/10.1364/OE.23.00948>.
- Vogel, A., Huttman, G., Paltauf, G., Noack, J., 2002a. Femtosecond-laser-produced low-density plasmas in transparent biological media: A tool for the creation of chemical, thermal and thermomechanical effects below the optical breakdown threshold. *SPIE Photonics West* 4633, 23–37.
- Vogel, A., Noack, J., Huetmann, G., Paltauf, G., 2002b. Low-density plasmas below the optical breakdown threshold: potential hazard for multiphoton microscopy, and a tool for the manipulation of intracellular events. *SPIE Photonics West* 4620, 202–216. <http://doi.org/10.1117/12.470693>.
- Winter, S., Sabesan, R., Tiruveedhula, P., *et al.*, 2016. Transverse chromatic aberration across the visual field of the human eye. *Journal of Vision* 16 (14), 1–10. <http://doi.org/10.1167/16.14.9>.
- Xu, C., Webb, W.W., 1997. Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. *Topics in Fluorescence Spectroscopy* 5, 471–540. (ST–Multiphoton excitation of molecular).
- Yellen, G., Mongeon, R., 2015. Quantitative two-photon imaging of fluorescent biosensors. *Current Opinion in Chemical Biology* 27, 24–30. <http://doi.org/10.1016/j.cbpa.2015.05.024>.
- Youssef, P.N., Sheibani, N., Albert, D.M., 2011. Retinal light toxicity. *Eye (London, England)* 25 (1), 1–14. [<http://doi.org/10.1038/eye.2010.149>].
- Yuste, R., Denk, W., 1995. Dendritic spines as basic functional units of neuronal integration. *Nature*. <http://doi.org/10.1038/375682a0>.
- Zhang, J., Sabarinathan, R., Bubel, T., Williams, D.R., Hunter, J.J., 2016. Action spectrum for photochemical retinal pigment epithelium (RPE) disruption in an *in vivo* monkey model. *Proceedings of SPIE* 9706, 1–6. <http://doi.org/10.1117/12.2213615>.
- Zhao, L., Qu, J., Niu, H., 2007. Identification of endogenous fluorophores in the photoreceptors using autofluorescence spectroscopy. *Proceedings of SPIE* 6826, 1–7. <http://doi.org/10.1117/12.760139>.
- Zheng, W., Li, D., Zeng, Y., Luo, Y., Qu, J.Y., 2011. Two-photon excited hemoglobin fluorescence. *Biomedical Optics Express* 2 (1), 71–79. <http://doi.org/10.1364/BOE.2.000071>.
- Zipfel, W.R., Williams, R.M., Christie, R., *et al.*, 2003. Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Sciences of the United States of America* 100 (12), 7075–7080. <http://doi.org/10.1073/pnas.0832308100>.
- Zoumi, A., Lu, X., Kassab, G.S., Tromberg, B.J., 2004. Imaging coronary artery microstructure using second-harmonic and two-photon fluorescence microscopy. *Biophysical Journal* 87 (4), 2778–2786. <http://doi.org/10.1529/biophysj.104.042887>.
- Zoumi, A., Yeh, A., Tromberg, B.J., 2002. Imaging cells and extracellular matrix *in vivo* by using second-harmonic generation and two-photon excited fluorescence. *Proceedings of the National Academy of Sciences of the United States of America* 99 (17), 11014–11019. <http://doi.org/10.1073/pnas.172368799>.

Confocal and Multiphoton Imaging of Cornea

Gopal S Jayabalan and Josef F Bille, University of Heidelberg, Heidelberg, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Most of our daily activities depend on a proper vision, and any loss in the visual acuity will strongly influence our quality of life. Preserving vision throughout an individual's lifetime is of a major concern in ophthalmology. The reliable diagnoses in ophthalmology are determined by detecting and monitoring the infinitesimal changes in tissue structures. The ophthalmologists mostly rely on slit-lamp biomicroscopy that was developed over a century ago in the 1890s. Subjective slit-lamp biomicroscopy remains as the backbone of clinical examination; however, objective tools and technologies are needed in order to provide a comprehensive understanding of structural alterations at a cellular level. Corneal disease is one of the leading causes of blindness worldwide, and a prompt and accurate diagnosis of corneal diseases is a milestone for successful treatment and followup.

Since the invention of the first ophthalmoscope by Hermann von Helmholtz in 1851, the quest for quantitative in vivo imaging of ocular surfaces for diagnostic and monitoring purposes was initiated. Recently, a number of imaging technologies were developed for research and clinical purposes. Boosted by the newest developments in microscopy, confocal microscopy has been widely employed in clinics and research for corneal diagnosis. Confocal microscopy has emerged as a powerful technique to study the corneal structure (Fig. 1) with high spatial and contrast images compared to conventional microscopy.

With the introduction of ultrafast pulsed laser sources, multiphoton microscopy practically became feasible and experienced a tremendous success in biological microscopy. Multiphoton microscopy allows three-dimensional (3D) tissue imaging over time, and it is an important asset in investigational imaging. Intrinsic tissue fluorescence, called autofluorescence, which is initiated by collagen, elastin, NAD(P)H etc., can be imaged by multiphoton microscopy and provides structural and functional information of the cornea. In this article, we will discuss the principles and the potential use and limitations of confocal and multiphoton microscopy for corneal imaging.

Principles of Confocal and Multiphoton Microscopy

Confocal Microscopy

The principle of confocal microscopy was patented in 1957 by Marvin Minsky, in order to overcome the limitations of conventional wide-field microscopy. In a conventional wide-field microscope, the entire specimen (i.e., cornea) will be illuminated evenly by a light source, and the specimens in the optical path are all excited at the same time. The resulting fluorescence (i.e., emitted light) from the specimen will be detected by the photodetector or camera including a largely unfocused background part. In contrast, confocal microscopy uses a point illumination source with a pinhole in front of the detector to eliminate the out of

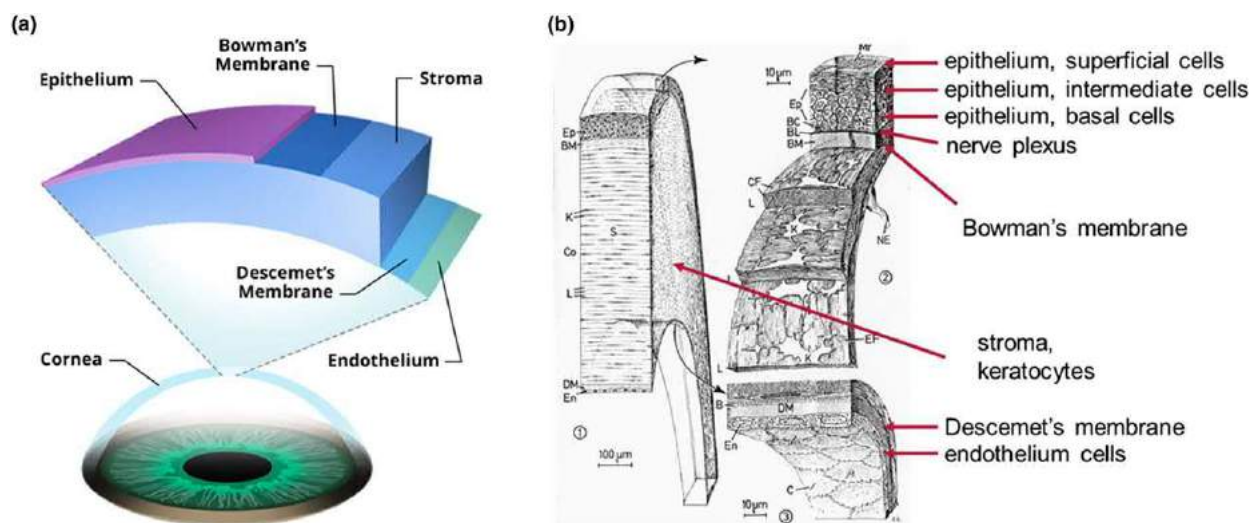


Fig. 1 Anatomical structure of cornea: (a) schematic presentation of a cornea (<http://www.allaboutvision.com/visionsurgery/corneal-inlays-onlays.htm>) and (b) layers of cornea (courtesy from the slide Krstic, R.V., 1991. Human microscopic anatomy. Springer).

focus light or glare. The optical resolution of confocal microscopy images is better than conventional wide-field microscopy, since the fluorescence is produced very close to the focal plane.

Multiphoton Microscopy

Multiphoton microscopy includes two-photon excited fluorescence (TPEF), second harmonic generation (SHG) and third harmonic generation (THG) imaging (Fig. 2). In 1931, the theoretical basis of two-photon excitation was established by Maria Göppert-Mayer, and then the photophysical effect of two-photon excitation was experimentally verified by Kaiser and Garret in 1963. Two-photon excitation is a fluorescence process in which a fluorophore (a molecule that fluoresces) is excited by simultaneous absorption of two photons. The familiar single-photon (i.e., confocal) fluorescence process involves exciting a fluorophore from the electronic ground state to an excited state by a single photon. This single-photon process typically requires photons in the ultraviolet or blue/green spectral range. However, the same excitation process can be generated by the simultaneous absorption of two less energetic photons (typically in the infrared spectral range) under sufficiently intense laser illumination. This nonlinear process occurs if the sum of energies of the two photons is greater than the energy gap between the molecule's ground and excited states. Since this process depends on the simultaneous absorption of two infrared photons, the probability of two-photon absorption by a fluorescent molecule is a quadratic function of the excitation radiance. Under sufficient intense excitation, three-photon and higher photon excitation are also possible. This process is similar to two-photon excitation wherein three photons or higher number of photons must interact with the fluorophore simultaneously to produce the fluorescence.

Almost all multiphoton microscopy is based on ultrashort pulsed lasers, because the multiphoton efficiency of an n -photon process depends on the peak power P according to P^n relation. The use of pulsed lasers enables efficient fluorophore excitation at a fast scanning rate. Moreover, focusing a pulsed laser into a specimen can generate a harmonic upconversion as well. Although the nonlinear optical effect known as SHG has been recognized in early days of laser physics, the use of SHG and its three-photon variant third (THG) is fairly new in research. Harmonic generation is a nonlinear process in which photons with the same frequency interact with a nonlinear material and generate a new photon with twice the energy. SHG is useful for investigating the ordered structural protein assemblies such as collagen fibers in the cornea.

Confocal Laser Scanning Microscopy

Confocal laser scanning microscopy (CLSM) has been widely implemented in clinics for diagnostic purposes. It is based on a conventional optical microscope, instead of a lamp a laser is used as a light source. The laser is focused on the cornea through the objective lens to a single spot in the focal plane. The laser beam is scanned in a raster pattern using the scanning mirrors in x - y direction and an entire image of the cornea at a certain focal plane will be reconstructed. The emitted light generated from the cornea at the focal plane will be collected by the same objective lens and the collected signals will be recorded using a photo-multiplier tube (PMT). This scanning principle does not improve the spatial resolution over conventional optical microscopy, because the excitation is produced throughout the entire beam path of the laser in the cornea, i.e., above and below the focal plane. However, the spatial resolution can be improved by placing a pinhole in front of the detector to eliminate the out of focus light. This leads to an excellent spatial resolution and enables serial optical sections from thick specimens (Fig. 3(a)).

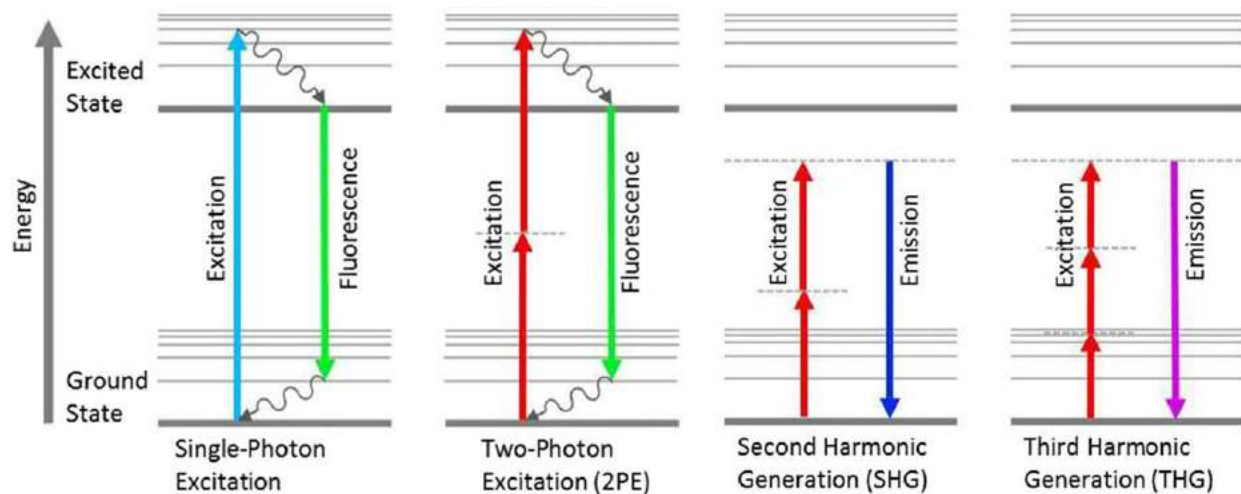


Fig. 2 Jablonski diagram of single-photon excitation, two-photon excitation (2PE), second harmonic generation (SHG), and third harmonic generation (THG). (<http://www.azooptics.com/Article.aspx?ArticleID=1175>)

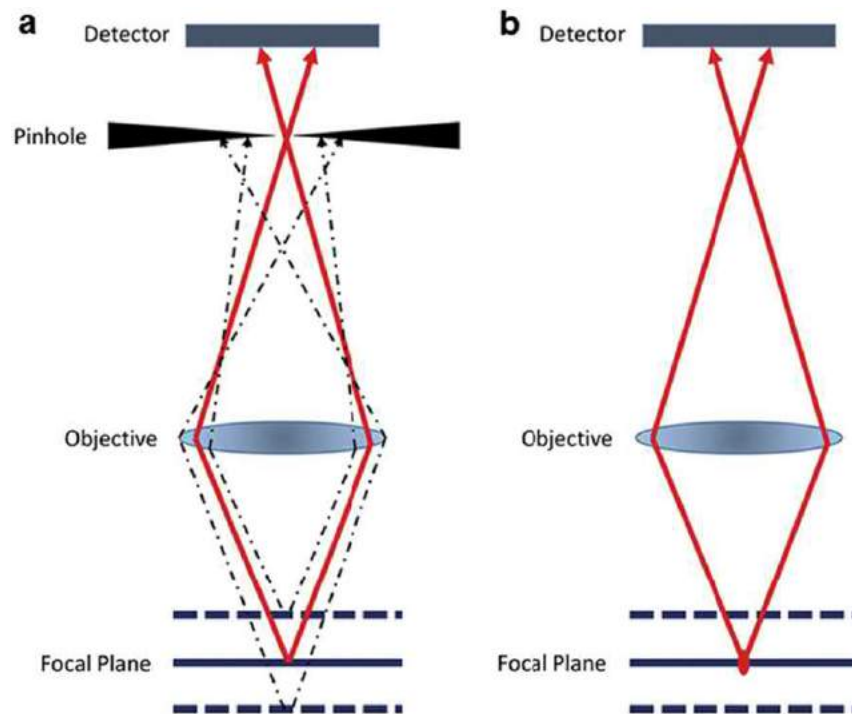


Fig. 3 The principle of confocal and two-photon microscopy. (a) In confocal microscopy, the emitted fluorescent light from the focal plane will pass through the pinhole and the out of focus light will be blocked by the pinhole and (b) in multiphoton microscopy, the two-photon excitation generates fluorescence only at the focal plane, so a pinhole is not required since no background fluorescence is produced. Reproduced from Allocca, G., Kusumbe, A.P., Ramasamy, S.K., Wang, N., 2016. Confocal/two-photon microscopy in studying colonisation of cancer cells in bone using xenograft mouse models. *BoneKey Reports* 5. doi:10.1038/bonekey.2016.84.

Two-Photon Laser Scanning Microscopy

The invention of two-photon fluorescence light microscopy by Denk, Webb, and coworkers revolutionized the *in vivo* 3D cellular imaging. By focusing an ultrashort pulsed laser through the objective lens sufficient excitation intensity can be generated on the focal plane (**Fig. 3(b)**). When the laser beam is focused on the specimen, the photons become crowded and the probability of two photons interacting simultaneously will be increased.

The significant advantage of two-photon excitation over single-photon (confocal) excitation is the narrow localization. In confocal microscopy, although the fluorescence is excited throughout the illuminated volume of the specimen, the signal originating from the focal plane will pass through the pinhole. In two-photon excitation, the fluorescence will be generated only at the focal plane, and no background fluorescence will be produced. Therefore, a pinhole is not required for two-photon microscopy. Due to the localized excitation of two-photon excitation, the 3D resolution of the images can be achieved identically to confocal microscopy. Since there is no absorption in out of focus specimen area, the excitation light can penetrate much deeper through the specimen to the focal plane. This results in greater penetration compared to confocal microscopy. Also, two-photon microscopy minimizes photobleaching and photodamage.

Confocal and Multiphoton Microscopy of Cornea

Three types of confocal microscopes are available for corneal imaging:

- Tandem scanning microscopy.
- Slit scanning microscopy.
- Laser scanning microscopy.

Tandem scanning microscopy sheds the light of high intensity through a pinhole and generates a two-dimensional (2D) image. However, tandem scanning microscopy for confocal imaging is not produced anymore. In contrast to tandem scanning, slit scanning microscopy has been employed in clinics with reduced scanning and examination time with high contrast imaging. Slit scanning has the ability to scan many points in parallel to the axis of the slit. Laser scanning microscopy is the latest addition to corneal imaging, where the laser beam is scanned over the back of the microscope objective through scanning mirrors. The Heidelberg Retina Tomograph (HRT) in combination with the Rostock Cornea Module (RCM) is a commercially available

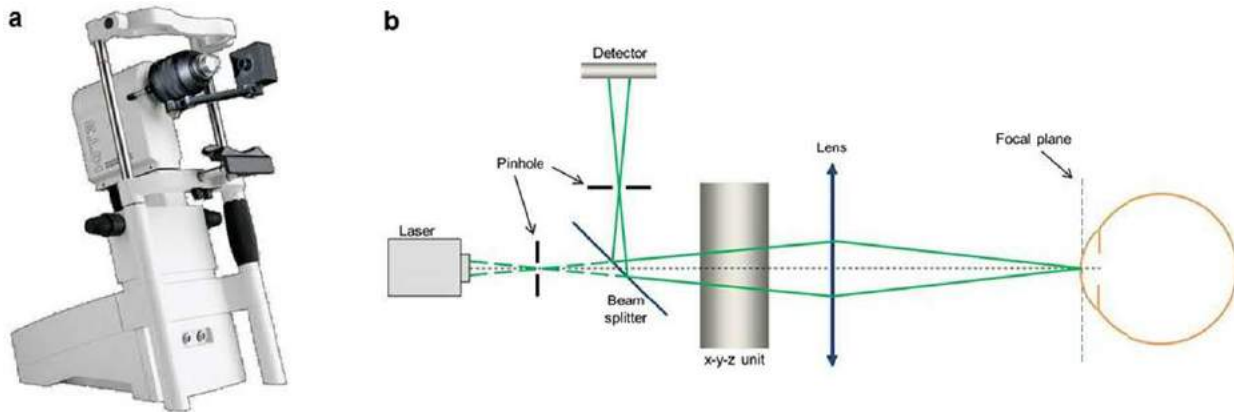


Fig. 4 (a) Heidelberg Engineering Heidelberg retinal tomograph-rostock cornea module (HRT-RCM) confocal microscopy and (b) schematic diagram of the optical set up of the HRT-RCM. Heidelberg Engineering Brochure.

confocal microscope that has been widely used in clinics for corneal imaging. The HRT-RCM uses a class I laser and this laser poses no risk of ocular injury.

Heidelberg Retinal Tomograph-RCM

The Heidelberg Retinal Tomograph-RCM (HRT-RCM) (**Fig. 4(a)**) was developed by Heidelberg Engineering GmbH in cooperation with Rostock Eye clinic for in vivo investigation of the cornea. A 670-nm laser is used as light source and the laser is focused on the cornea with a high numerical aperture objective lens ($63\times/0.95$ NA or $60\times/0.90$ NA; interchangeable). The laser is raster scanned by the scanning mirrors and is focused on the cornea by the objective lens. The emitted fluorescent light from the cornea is captured using a detector. The confocal pinhole is used in front of the detector to eliminate the out of focus light (**Fig. 4(b)**). The confocal imaging of the cornea can be acquired with the field of view $250\ \mu\text{m} \times 250\ \mu\text{m}$, $400\ \mu\text{m} \times 400\ \mu\text{m}$, and $500\ \mu\text{m} \times 500\ \mu\text{m}$ with a $63\times$ objective lens. A viscous sterile gel is interposed between the objective lens and cornea to eliminate the optical interface between two different refractive indices. The depth of the focus can be adjusted manually relative to the cornea in a plane parallel to the corneal surface, to scan the cornea from epithelium to endothelium.

Confocal Imaging of Normal Cornea

The HRT-RCM provides excellent resolution with sufficient contrast at different layers of the cornea, and distinct corneal layers from epithelium to endothelium can be visualized. The corneal epithelium is composed of five to six layers (**Fig. 5(a)–(d)**), as superficial cells, intermediate cells, and basal cells. The superficial cells are flat polygonal cells with well-defined border, prominent nuclei and uniform density of cytoplasm. It is about $40\text{--}50\ \mu\text{m}$ in diameter and approximately $5\ \mu\text{m}$ thick. The intermediate cells also known as wing cells are similar to superficial cells, but the nuclei are not evident. It is typically about $30\text{--}45\ \mu\text{m}$ in size and characterized by the cell borders. Basal cells are approximately $10\text{--}15\ \mu\text{m}$ and smaller in size. The basal cells appear denser than the superficial and wing cells.

The subbasal nerve plexus (**Fig. 5(e)**) is characterized by the presence of hyper-reflective fibers of $4\text{--}8\ \text{mm}$ length, and connected with anastomoses and organized in a vortex pattern in the lower nasal quadrant of the paracentral cornea. The visualization of corneal nerve plexus was not possible until the introduction of confocal microscopy. The nerve fibers appear bright and well contrasted against a dark background.

Corneal stroma accounts for 90% of total corneal thickness and the keratocytes density is much higher in anterior stroma than in posterior stroma. The keratocyte counts are $1000\ \text{cells}/\text{mm}^2$ in the anterior stroma and $700\ \text{cells}/\text{mm}^2$ in the posterior stroma. The confocal image of the stroma exhibits multiple irregular oval, round or bean shaped bright structures representing keratocytes (**Fig. 5(f)** and **(g)**). The collagen fibers cannot be visualized by confocal microscopy in normal cornea.

Corneal endothelium is composed of a single layer of cells that are $4\text{--}6\ \mu\text{m}$ thick and approximately $20\ \mu\text{m}$ in diameter. The endothelial cells progressively decline with age, and the normal endothelial cell count amounts to $3000\ \text{cells}/\text{mm}^2$ in a healthy cornea. The endothelial cells appear as bright hexagonal and polygonal cells with unrecognizable nuclei (**Fig. 5(h)**).

Confocal Microscopy in Corneal Pathologies

Confocal microscopy has been an effective tool in clinical diagnosis, and it is widely used to assess corneal infections, corneal dystrophies, pre and postsurgical assessment of laser-assisted in situ keratomileusis (LASIK), photorefractive keratectomy (PRK), keratoconus and infectious keratitis. Also, it is used to monitor the cornea after injuries, and for differential diagnosis of

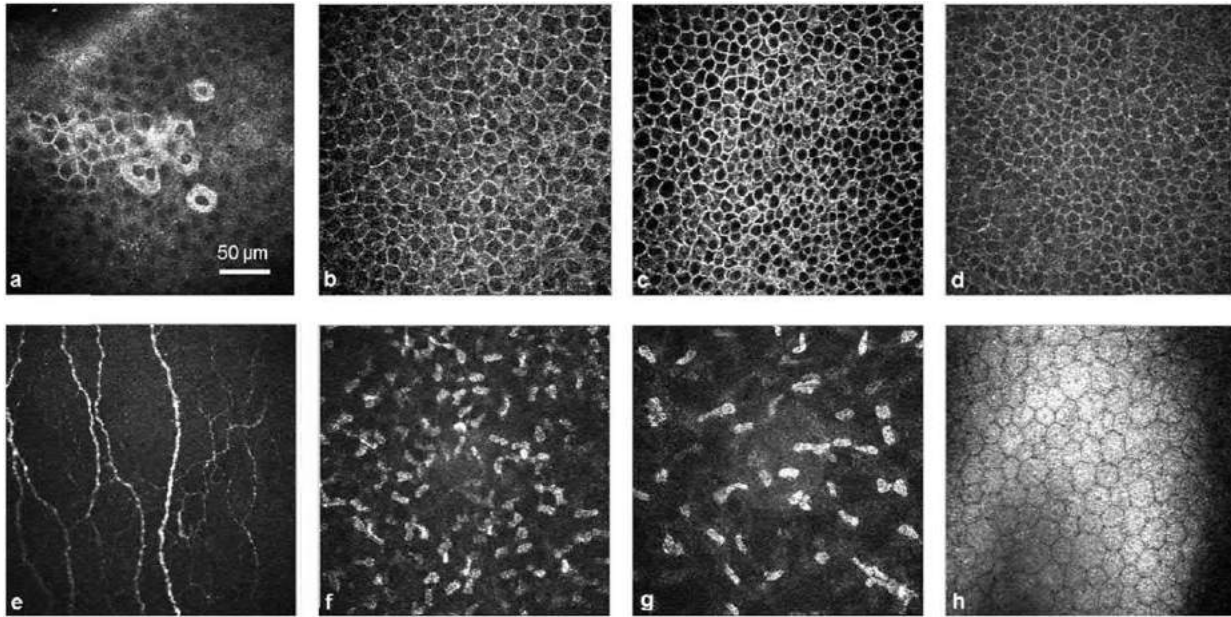


Fig. 5 In vivo confocal microscopy images of the corneal layers in normal human subject from epithelium to endothelium. (a)–(d) Epithelium, (e) subbasal nerve plexus, (f) and (g) anterior and posterior stroma, and (h) endothelium. Courtesy of Prof. R. Guthoff, Prof. J. Stave, Dr. A. Zhivov, University of Rostock (from Heidelberg PowerPoint).

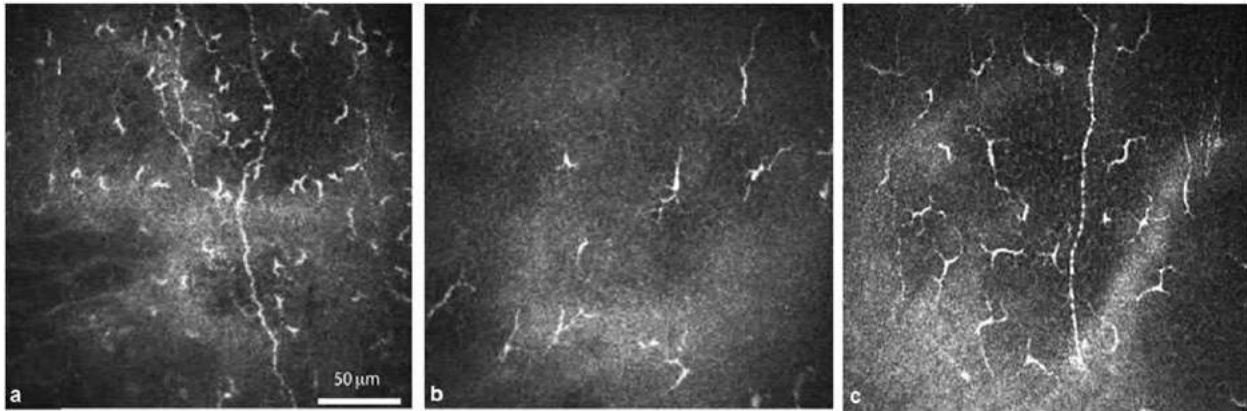


Fig. 6 In vivo confocal microscopy images demonstrating different forms of Langerhans cells. (a) Individual cell bodies without processes; (b) cells bearing dendrites; and (c) cells arranged in a meshwork via long interdigitating dendrites. Reproduced from Guthoff, R.F., Baudouin, C., Stave, J., 2006. Atlas of Confocal Laser Scanning In-Vivo Microscopy in Ophthalmology. Springer: Germany.

conjunctival tumors and conjunctivitis. As an example for clinical confocal imaging of the cornea, the Langerhans cell distribution and postsurgical assessment of LASIK using the HRT-RCM is shown in Figs. 6 and 7.

Langerhans Cells

Langerhans cells can be evaluated using confocal microscopy with emphasis on cell morphology and cell distribution. In confocal microscopy, the Langerhans cells are visualized typically as bright corpuscular particles with dendritic morphology. These cells are distributed in low numbers at the center of the cornea and higher densities at the periphery of the cornea. Confocal microscopy helps in differentiating the Langerhans cells bodies that lack dendrites, Langerhans cells with small dendritic process form a local network, and Langerhans cells forming a meshwork via long interdigitating dendrites (Fig. 6(a)–(c)). The average density of Langerhans cells in normal subjects is 34 ± 3 cells/mm² (range 0–64 cells/mm²) in the center of the cornea, and 98 ± 8 cells/mm² (range 0–208 cells/mm²) in the periphery. Langerhans cells contribute to the immune and inflammatory responses, and determine the cell-mediated immunity.

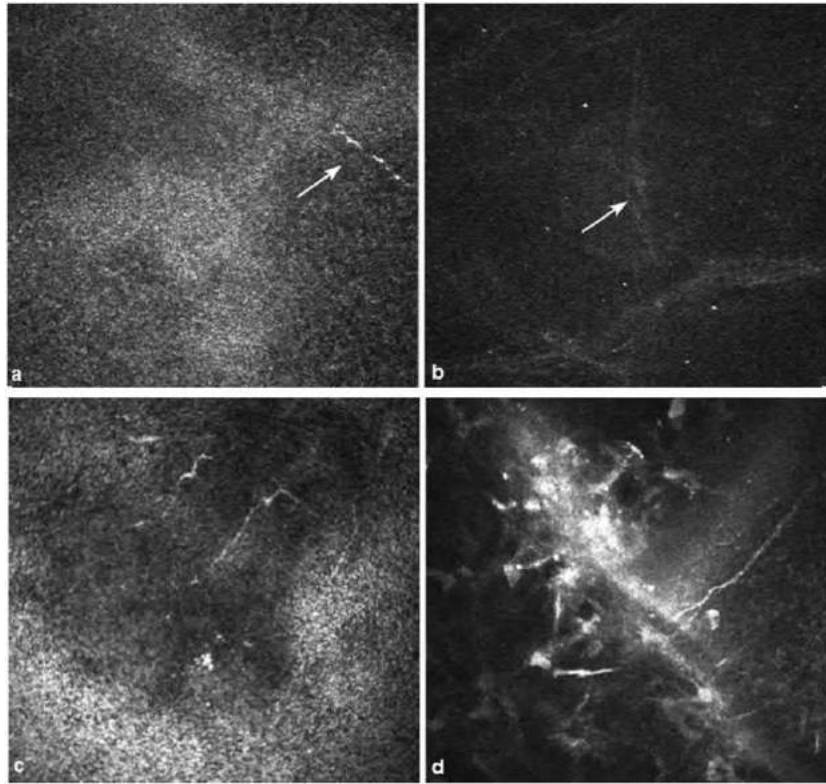


Fig. 7 Innervation after refractive surgery. (a) Nerve fibers near subbasal layer; (b) nerve fibers near Bowman's membrane; (c) two months after LASIK – small regenerating nerve fibers in the central cornea; and (d) nine months after LASIK – new nerve fiber growing from the edge of flap region to center. Reproduced from Guthoff, R.F., Baudouin, C., Stave, J., 2006. *Atlas of Confocal Laser Scanning In-Vivo Microscopy in Ophthalmology*. Springer: Germany.

Postsurgical Assessment of LASIK

LASIK is an excimer laser-based refractive surgical procedure that is currently being successfully used by surgeons to correct refractive errors (myopia, hyperopia etc.). After LASIK surgery, the cornea is typically evaluated by slit-lamp biomicroscopy and computerized corneal topography. The confocal microscopy evaluation of LASIK pre- and postsurgery adds new aspects in corneal diagnosis. The distribution of the corneal nerves and degeneration of the nerve structures can be visualized by confocal microscopy, and thus a direct comparison between innervation and sensitivity or symptom severity of the dry eye can be provided (Fig. 7).

Two-Photon Laser Scanning Microscopy

Two-photon laser scanning microscopy based on the HRT was developed to investigate the cornea (Fig. 8). A compact femto-second laser was employed in the system as the light source and an additional detection path for the two-photon mode was implemented. Both the confocal and two-photon images can be captured using the same instrument. An electronic switch facilitates the switching between the two imaging modes, while both detection modes are synchronized with the scanning unit. The laser beam is deflected in x - and y -direction via two scanning mirrors (scan unit, Fig. 8(b)) to trace out a square raster of $300 \times 300 \mu\text{m}^2$, $200 \times 200 \mu\text{m}^2$, or $150 \times 150 \mu\text{m}^2$ at a frame rate up to 16 Hz. The laser is focused on the cornea with a high numerical aperture objective lens ($40 \times /0.80$ NA or $16 \times /0.80$ NA; interchangeable). The wide field of view can be achieved by using the $16 \times$ objective lens with less distinct cell morphology. Like confocal microscopy (HRT-RCM) the focus can be adjusted manually and different layers of the cornea can be visualized. A viscous sterile gel has to be placed between the objective lens and the cornea to eliminate the optical interface between two different indices.

Two-Photon Imaging of Normal Cornea

The TPEF imaging of cornea has been extensively employed in preclinical studies, but has not been implemented in clinics yet. Although confocal microscopy has been used in clinical diagnosis, the collagen fiber orientation in the stroma cannot be visualized, and it provides only morphological information rather than the tissue's functional state. With two-photon imaging

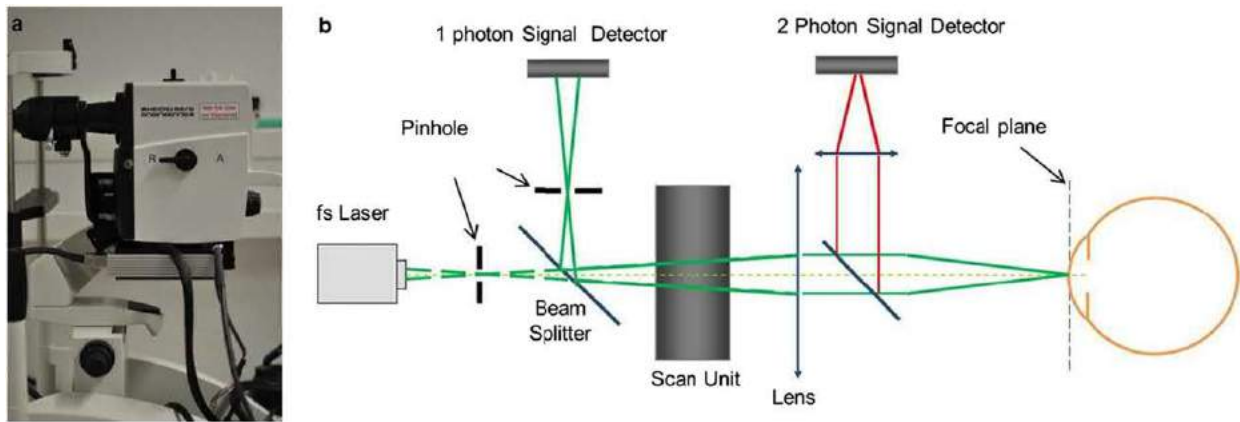


Fig. 8 (a) Heidelberg Engineering two-photon microscopy and (b) schematic diagram of the optical set up of two-photon microscopy based on HRT. Reproduced from Qu, Y., Thomas, K.E., Vola, M.E., *et al.*, 2012. Rapid identification of microorganisms using the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 53 (14), 6175; Qu, Y., Thomas, K.E., Schuster, A.K., *et al.*, 2011. Identification of microorganisms utilizing the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 52 (14), 5823.

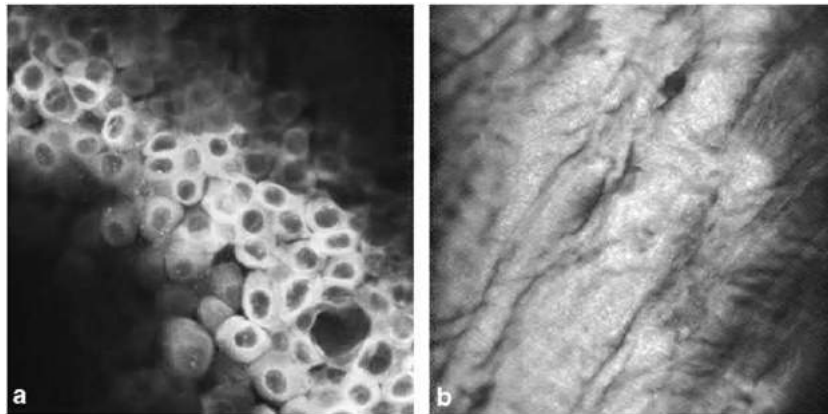


Fig. 9 Two-photon imaging of corneal epithelial cells (a) and stromal collagen fibers (b) of human cornea. Reproduced from Qu, Y., Thomas, K. E., Vola, M.E., *et al.*, 2012. Rapid identification of microorganisms using the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 53 (14), 6175; Qu, Y., Thomas, K.E., Schuster, A.K., *et al.*, 2011. Identification of microorganisms utilizing the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 52 (14), 5823.

both the tissue structures and the metabolic state can be well addressed. Two-photon microscopy provides excellent resolution and high contrast images of the cornea similar to confocal microscopy. High resolution two-photon images of the corneal epithelium and the collagen fibers from the human corneal buttons are shown in [Fig. 9](#). The collagen fibers in the stroma can be well established with TPEF and SHG imaging.

Two-Photon Imaging of Limbal Stem Cells

The limbal stem cells are located in the basal layer of the corneal limbus, and the proliferation of the limbal cells maintains the cornea by replacing the lost cells by tears. Limbal stem cells also prevent the conjunctival epithelial cells from migrating onto the surface of the cornea. The limbal stem cells have never been visualized with confocal microscopy and with the use of two-photon microscopy the limbal stem cells have been identified in mice. The limbal stem cells are smaller in size and it is shown by two-photon microscopy ([Fig. 10](#)). The two-photon imaging of limbal stem cells would help in the diagnosis of limbal stem cell deficiency and also differentiate and explore the nature of the limbal stem cells in a precise way.

Two-Photon Imaging of Infectious Keratitis

Confocal microscopy plays a significant role in detecting the microorganisms causing corneal infections. It is difficult to identify the microorganisms like *Acanthamoeba* on the ocular surfaces since the appearance can impersonate as herpetic or bacterial keratitis. In most cases, the *Acanthamoeba* keratitis is diagnosed either by tissue culture or corneal biopsy, and histological analysis.

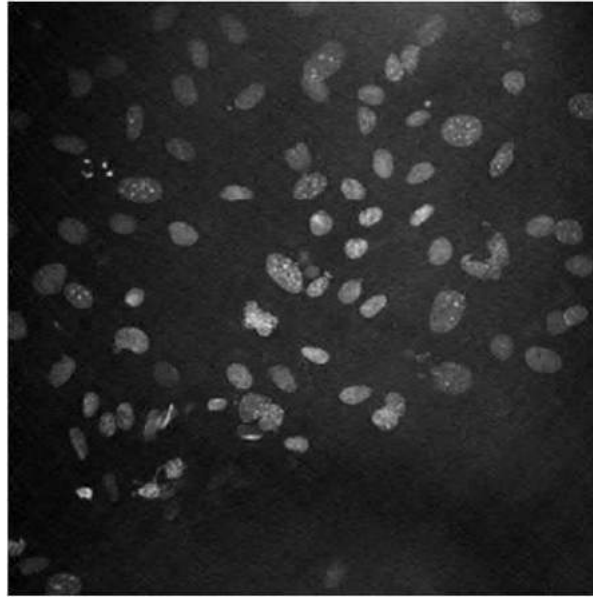


Fig. 10 Two-photon imaging of mice limbal stem cells. Reproduced from Qu, Y., Thomas, K.E., Vola, M.E., *et al.*, 2012. Rapid identification of microorganisms using the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 53 (14), 6175; Qu, Y., Thomas, K.E., Schuster, A.K., *et al.*, 2011. Identification of microorganisms utilizing the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 52 (14), 5823.

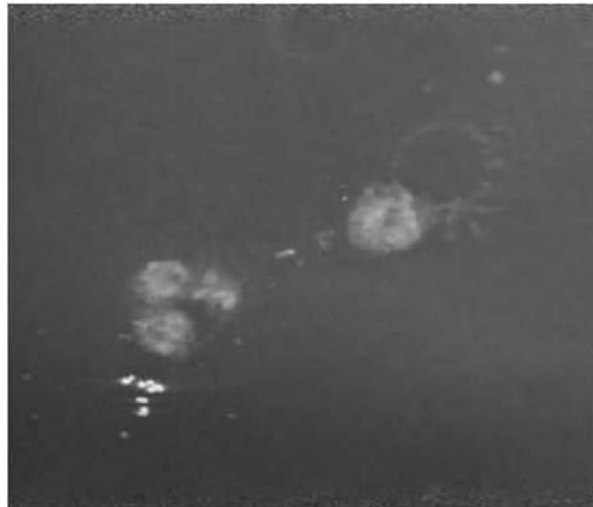


Fig. 11 Two-photon imaging of *Acanthamoeba* keratitis. Reproduced from Qu, Y., Thomas, K.E., Vola, M.E., *et al.*, 2012. Rapid identification of microorganisms using the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 53 (14), 6175; Qu, Y., Thomas, K.E., Schuster, A.K., *et al.*, 2011. Identification of microorganisms utilizing the two-photon ophthalmoscope. *Investigative Ophthalmology and Visual Science* 52 (14), 5823.

Although confocal microscopy lacks sufficient resolution in detecting the *Acanthamoeba* keratitis, it is used for diagnostic purposes since confocal microscopy is the only available diagnostic tool for patients. In confocal microscopy, the *Acanthamoeba* is visualized as round or ovoid highly reflective structures and the same is depicted by the two-photon microscopy (Fig. 11).

Second Harmonic Imaging of Cornea

Collagen is the major component of connective tissues including cornea, and due to its transparency to visual and near infrared light the human eye is ideal for laser-based diagnostic and imaging applications. Among different multiphoton implementations, SHG imaging is particularly suitable to investigate non-centrosymmetric structures like collagen fibrils (Figs. 12 and 13). The basic structure of the collagen fibril is a triple helix composed of three proteins chains, which gives collagen the intrinsic ability of SHG

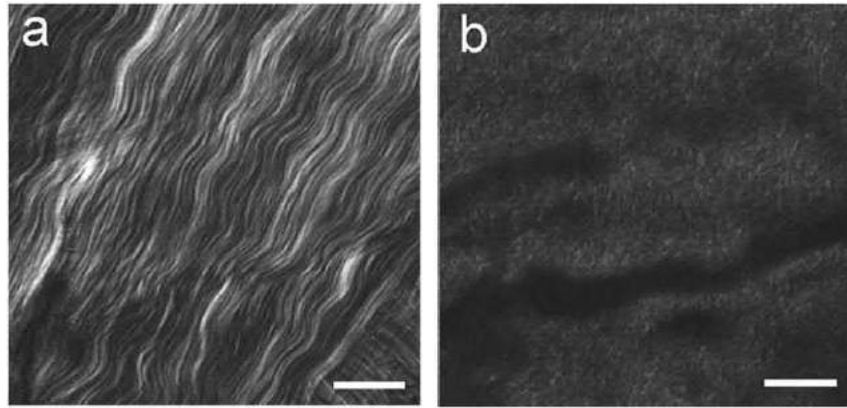


Fig. 12 SHG imaging of corneal collagen fibrils in (a) forward and (b) backward directions from porcine cornea. The fibrillar structures resolved in (a) correspond to collagen bundles which are composed of regularly packed collagen fibrils. Scale=10 μ m. Reproduced from Han, M., Giese, G., Bille, J., 2005. Second harmonic generation imaging of collagen fibrils in cornea and sclera. *Optics Express* 13, 5791–5797.

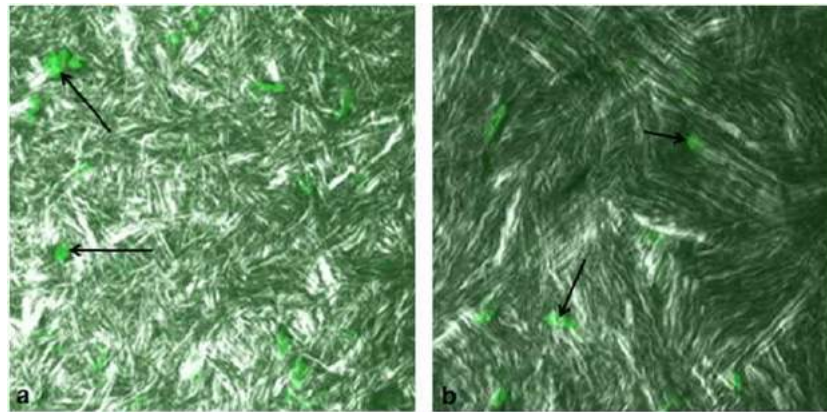


Fig. 13 SHG imaging and two-photon imaging of human cornea from (a) Bowman's membrane and (b) corneal stroma. The collagen fibers were visualized using the SHG imaging and the keratocytes (arrows) were visualized with the two-photon imaging. Reproduced from Han, M., Giese, G., Bille, J., 2005. Second harmonic generation imaging of collagen fibrils in cornea and sclera. *Optics Express* 13, 5791–5797.

imaging. Similar to TPEF, SHG is produced in only a small focal volume, permitting high resolution 3D optical sectioning of thick tissues. In contrast to TPEF, SHG from collagen is an intrinsic and a coherent process. Intrinsic imaging avoids the complications of slicing and labeling, and samples can be investigated under physiological conditions. Coherent constructive or destructive interference of SHG provides extra hints to the ultrastructure of collagen fibrils and their organizations.

The collagen fibrils of the porcine cornea using SHG imaging are shown in [Fig. 12](#). Most collagen fibrils run parallel to the cornea surface and share similar orientations with their neighbors. One of the unique features of SHG is that the geometry of the SHG emission field reflects the size and shape of the collagen fibrils. The SHG radiation pattern is mainly determined by the phase matching condition and the objects with the axial size on the order of the second harmonic wavelength exhibit forward directed SHG, while objects with an axial size less than $\lambda/10$ (approx. 40 nm) are estimated to produce nearly equal backward and forward SHG signals.

In [Fig. 12\(b\)](#), the backward SHG from cornea was extremely weak and there was no correlation between the structures revealed by forward and backward SHG imaging. The predominant forward SHG implies that the corneal collagen fibrils are not randomly distributed. It's worth mentioning that the fibrillar structures revealed in [Fig. 12\(b\)](#) actually correspond to collagen bundles composed of many corneal collagen fibrils. In cornea, the collagen fibrils ($n=1.47$) and the extrafibrillar matrix ($n=1.35$) have different refractive indices, thus the collagen fibrils have to be treated as individual scatters. Since the dimension of the collagen bundle (approximately 0.5 μ m) is close to the second harmonic wavelength (400 nm), the phase correlation of SHG from neighboring collagen fibrils should be taken into account. The forward predominant SHG radiation from cornea indicates the regular arrangement of the collagen fibrils, at least in the domain of the collagen bundles.

In [Fig. 13](#), SHG and two-photon imaging of human cornea from Bowman's membrane ([Fig. 13\(a\)](#)) and corneal stroma ([Fig. 13\(b\)](#)) is depicted. The collagen fibers were visualized using the SHG imaging and the keratocytes (arrows) were imaged with TPEF imaging.

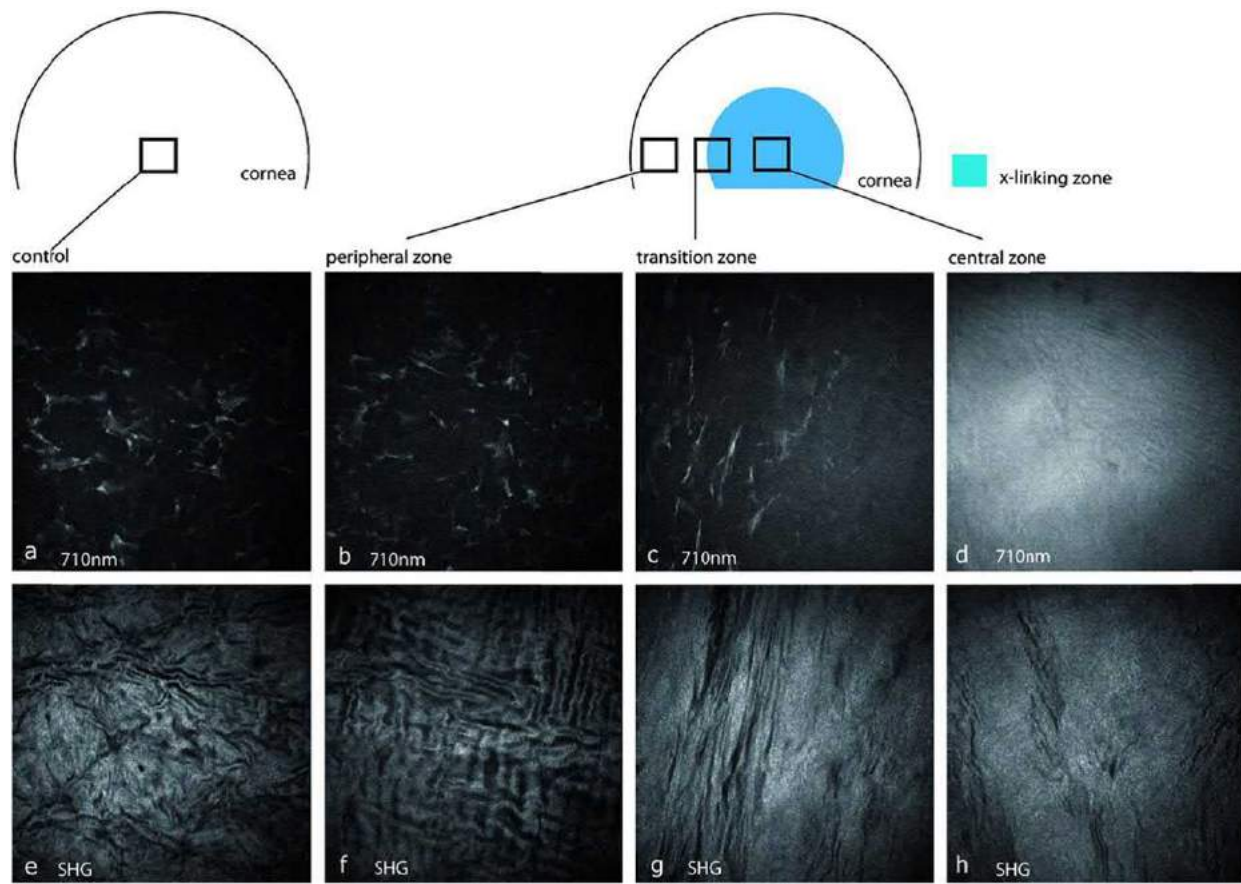


Fig. 14 Two-photon microscopy of enucleated rabbit eyes 2 weeks after standard corneal CXL, upper row: two-photon imaging at 710 nm excitation light and lower row: second harmonic generation. (a) Two-photon imaging of the central zone of control cornea; (b) the peripheral zone of the treated cornea shows keratocytes in a dense network; (c) at the tangential zone, there is a sharp transition between the treated area and the untreated area, as shown by diminished keratocytes and a strong autofluorescence signal toward the central zone; (d) in the central zone, there are no keratocytes and the autofluorescence signal is uniform and strong; (e) and (f) second harmonic generation of stromal collagen in the control zone and the peripheral zone of the treated cornea shows a wavelike pattern; (g) at the transition zone toward the treated area, the wavelike pattern changes to a uniform signal; (h) in the central zone, a uniform SHG signal is visible. Reproduced from Steven, P., Hovakimyan, M., Guthoff, R.F., Hüttmann, G., Stachs, O., 2010. Imaging corneal crosslinking by autofluorescence 2-photon microscopy, second harmonic generation, and fluorescence lifetime measurements. *Journal of Cataract Refractive Surgery* 36 (12), 2150–2159.

Two-Photon Imaging of Corneal Cross-linking

Keratoconus is a progressive, noninflammatory and bilateral ectatic corneal disease characterized by corneal thinning and weakening. The thinning is most apparent at the apex of the cornea, and the steep conical protrusion leads to high myopia with irregular astigmatism. The visual loss occurs primarily from irregular astigmatism and myopia, and secondarily from the corneal scarring. Computerized corneal topography has become a gold standard for the early identification of keratoconus and recently confocal microscopy has also been used to assess the keratoconus cornea. Corneal cross-linking (CXL) is an invasive treatment procedure for keratoconus in order to stop or delay the process. The CXL procedure increases the mechanical and biomechanical stability of the cornea by application of ultraviolet-A light of 370 nm and riboflavin (vitamin B2). The assessment of cornea after cross-linking is limited because the treatment effect cannot be visualized by slit-lamp biomicroscopy. Confocal microscopy can detect the side effects like stromal edema and rarefaction, the reappearance of stromal keratocytes over time. However, the effect of crosslinks between the collagen fibers after CXL cannot be visualized by confocal microscopy. Two-photon imaging with SHG is effective in measuring the crosslinks between collagen fibers as shown in [Fig. 14](#).

Laser Safety Considerations of Two-Photon Microscopy

The safe use of lasers has been a primary consideration since lasers are widely used for treatments and in research laboratories. The American National Standards Institute (ANSI) has specific standards for eye safety that take into account many of the important experimental variables such as laser power delivered at the cornea, viewing time, beam diameter or full-field viewing, the exposure area on the retina, and the wavelength. Near infrared radiation (NIR) relative to visible wavelength is considered to be less

phototoxic and has deeper penetration in thick biological tissues, which is one of the key reasons for its widespread use in imaging devices. Compared to retinal imaging, the safety standards for illuminating the cornea and anterior structures of the eye are tolerant. The laser safety limits for corneal imaging can be calculated by the maximum permissible exposures (MPE) according to the ANSI standards as follows:

$$\text{Recommended MP corneal irradiance} = 25 \times t - 0.75 \text{ (W/cm}^2\text{)} \text{ for } t < 10\text{s}$$

$$\text{Recommended MP corneal irradiance} = 4.0 \text{ (W/cm}^2\text{)} \text{ for } t > 10\text{s}$$

For example, if the irradiated area is described by $1536 \mu\text{m} \times 1536 \mu\text{m} = 0.0236 \text{ cm}^2$, the MPE should be 94.4 mW for longer exposure time ($t > 10\text{s}$) and should be 515 mW for shorter exposure time.

Future Directions

Before implementing multiphoton microscopy for in vivo imaging of the human cornea, the shortcomings of these nonlinear microscopy techniques have to be addressed. For TPEF imaging, the drawback of depositing the energy into the tissue has to be evaluated to prevent photodamage. Although SHG and THG are absorption-free and can be used without photobleaching effects, harmonic imaging has another disadvantage that currently obstructs its use in ophthalmic practice. The harmonic signals are emitted predominantly in forward direction, whereas in backward direction the harmonic signals are weak. This limits SHG imaging for in vivo applications, and the backscattered signal might be enhanced using an absorbing dye. Also, new optical techniques such as wavefront aberration optimization and adaptive optics may be a key for future in vivo applications. Until these restrictions of multiphoton microscopy are eliminated, confocal microscopy remains exclusive in the assessment of cellular morphology in living human cornea.

Acknowledgment

The authors like to thank Heidelberg Engineering GmbH (Heidelberg, Germany) for providing equipment support.

See also: Optical Coherence Tomography and Its Application to Imaging of Skin and Retina. Optics of the Human Eye

Further Reading

- Denk, W., Strickler, J.H., Webb, W.W., 1990. Two-photon laser scanning fluorescence microscopy. *Science* 248, 73–76.
- Guthoff, R.F., Baudouin, C., Stave, J., 2006. *Atlas of Confocal Laser Scanning In-Vivo Microscopy in Ophthalmology*. Germany: Springer.
- Guthoff, R.F., Zhivov, A., Stachs, O., 2009. In vivo confocal microscopy, an inner vision of the cornea – A major review. *Clinical and Experimental Ophthalmology* 37, 100–117.
- Han, M., Giese, G., Bille, J., 2005. Second harmonic generation imaging of collagen fibrils in cornea and sclera. *Optics Express* 13, 5791–5797.
- Hao, M., Flynn, K., Nien-Shy, C., *et al.*, 2010. In vivo non linear optical (NLO) imaging in live rabbit eyes using the heidelberg two-photon laser ophthalmoscope. *Experimental Eye Research* 91 (2), 308–314.
- Mantopoulos, D., Cruzat, A., Hamrah, P., 2010. In vivo imaging of corneal inflammation: New tools for clinical practice and research. *Seminars in Ophthalmology* 25 (5-6), 178–185.
- Steven, P., Hovakimyan, M., Guthoff, R.F., Hüttmann, G., Stachs, O., 2010. Imaging corneal crosslinking by autofluorescence 2-photon microscopy, second harmonic generation, and fluorescence lifetime measurements. *Journal of Cataract & Refractive Surgery* 36 (12), 2150–2159.

Refractive Error and Wavefront Sensing

Larry N Thibos, Indiana University, Bloomington, IN, United States

© 2018 Elsevier Ltd. All rights reserved.

Definitions and Conventions

Refractive Error, Refractive Correction, and Their Physical Units

The purpose of the eye's optical system is to cast an image of the external world onto the photoreceptor layer of rods and cones in the retina. If the system were perfect it would focus all rays of light from a point source into a single image point on the retina. Real eyes, however, suffer from aberrations, which are optical imperfections that degrade the quality of the retinal image and hamper vision. Since aberrations are flaws in the refraction of light, one might presume that the terms aberrations and refractive errors are synonymous. However, in the field of ophthalmic and visual optics, the term refractive error is restricted to two specific types of aberrations: defocus and astigmatism. Historically, these are the only two aberrations that could be measured routinely in a clinical environment and then corrected easily with spectacles or contact lenses. Although the existence of other ocular aberrations (e.g., coma, spherical aberration, etc.) has been known since the time of Newton, those aberrations could not easily be measured or corrected clinically. Since these "higher-order" aberrations fell outside the domain of conventional clinical practice, they were excluded from the traditional definition of refractive error. Depending on context, astigmatism may also be excluded leaving just defocus as the only aberration implied by the term refractive error.

The sign convention for specifying refractive errors of eyes can appear confusing. Although scientists may be inclined to describe the refractive errors of the eye, per se, clinicians prefer to specify the lens needed to correct the eye's refractive errors. As illustrated in Fig. 1, the ideal condition of emmetropia exists when distant objects produce well-focused retinal images.

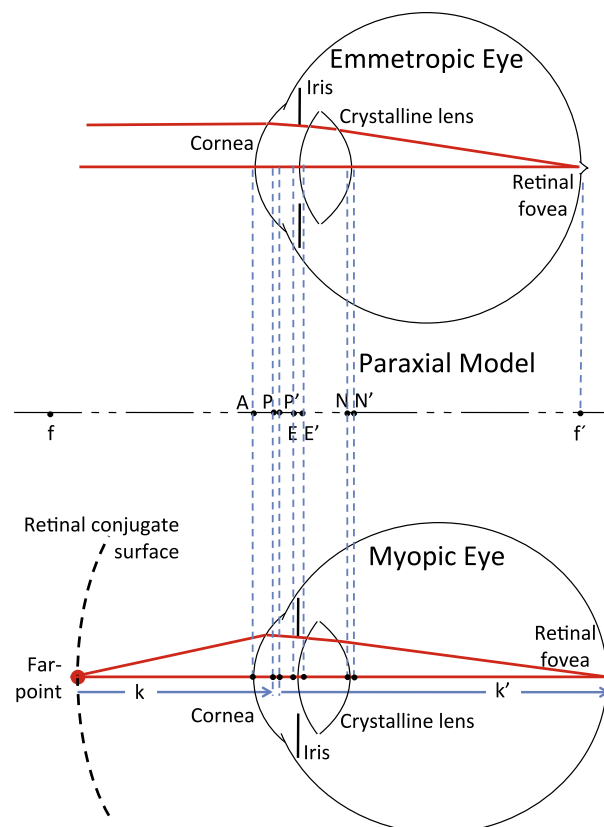


Fig. 1 Schematic optical diagrams of the eye. Upper panel shows an emmetropic eye focusing rays from a distant point source onto the retinal photoreceptors in the fovea. Lower panel shows a myopic eye focusing rays from a point source at finite distance onto the retinal photoreceptors. Middle panel is a Gaussian paraxial model for both eyes showing cardinal points (f, f' – focal points; P, P' – principal points; N, N' – nodal points), the entrance and exit pupils (E, E') and the corneal apex (A).

However, a short-sighted (i.e., myopic) eye has too much refractive power to clearly focus the retinal image of a distant object, and therefore requires a correcting spectacle or contact lens of negative power. Thus, the sign of the defocus component of refractive error for myopic eyes is negative by standard clinical conventions. Conversely, a hyperopic eye has too little power to focus on distant objects and therefore requires a positive correcting lens, so hyperopic refractive error is positive.

Refractive errors and viewing distance of objects are both specified in units of inverse meters, more commonly referred to as diopters (D). Diopters is the preferred unit in ophthalmic and visual optics because practical problems are more easily solved when the Gaussian conjugate equation of paraxial optics

$$n'/k' = n/k + (n' - n)/r \quad (1)$$

where k is the object distance, k' is the image distance, n is the refractive index of object space, n' is the refractive index of image space, and r is the radius of curvature of the refracting surface, is rewritten as

$$K' = K + F \quad (2)$$

where K is the object vergence, K' is the image vergence, and F is the refracting power of the surface attributed to the Gaussian principal planes. All three quantities in this linear equation have physical units of diopters. The situation here is analogous to electrical circuits involving parallel resistors, for which solutions are most easily obtained by introducing the term conductance when referring to the inverse of resistance. That maneuver requires a corresponding change of physical units of resistance (ohms) to inverse units of conductance (mhos). In the present optical analogy, inverse distance from object or image points to their respective principal points is called vergence, which has units of diopters. Thus, the Gaussian conjugate Eq. (2) reads "the vergence K' of light leaving a refracting surface (or thin lens) is equal to the incident vergence of light K plus the refracting power F of the lens". If the lens is astigmatic, this same equation applies to refraction in any meridian but the refracting power F varies sinusoidally with meridian angle. A simple mnemonic for distinguishing the closely related terms power and vergence is: "lenses have the power to change the vergence of light".

Given the terminology introduced above, the defocus component of refractive error may be defined three equivalent ways that are mutually consistent but have different interpretations. They may be understood by considering Fig. 1, which illustrates a simple case of myopia caused by axial elongation of the eye. Thus, the myopic optical system in this example has the same paraxial cardinal points as the illustrated emmetropic eye, but axial elongation causes the retinal conjugate surface to move from infinity to a finite distance from the eye. The myopic eye's far point, which is the point in object space that is optically conjugate to the retinal fovea, is located at some finite distance k from the first principal point P . So the first interpretation of refractive error occurs in object space: it is the vergence of the eye's far point,

$$\text{Refractive error} = K = n/k \quad (3)$$

The second interpretation is in image space: it is the focusing error for a distant object point. Since $K=0$ for objects at infinity, the focusing error is computed from Eq. (2) as the difference between the required power K' (sometimes called the dioptric length of the eye) and the eye's available power F ,

$$\text{Refractive error} = K = K' - F \quad (4)$$

The third interpretation conceives of refractive error as a lens prescription: it is the power of the correcting lens needed to focus a distant target on the retina. Again $K=0$, so the required correcting lens has to make up the difference between the eye's dioptric length and its power, as indicated in Eq. (4). In principle, the correcting lens should be placed in the eye's first principal plane but in practice it is placed on the cornea (i.e., a contact lens, which requires only a small adjustment in power), or in the spectacle plane (which requires a larger adjustment in lens power to compensate for the significant separation of the lens from the eye's first principal point).

Specification of Sphero-Cylindrical Lenses and Refractive Errors

Several conventions are in common use for specifying the sphero-cylindrical lens needed to fully correct an eye's refractive errors. Differences between these conventions arise from alternative ways of envisioning how a prescribed sphero-cylindrical lens is constructed from component lenses. Two of these conventions presuppose that the correcting lens is the combination of a spherical lens of power S juxtaposed with a cylindrical lens of power C oriented with the axis of the cylinder at some angle α . The difference between these two conventions is the sign of C . Optometrists typically prefer to think of C as a negative number, whereas ophthalmologists typically prefer to think of C as a positive number. Simple transposition formulae allow conversion from one convention to another. A third convention preferred by many researchers replaces S with M , the so-called mean spherical-equivalent power. If one imagines correcting an eye with sphero-cylindrical refractive error using only a spherical lens of power M , the eye would remain astigmatic by some amount J . The value of J refers to the power of a so-called Jackson crossed-cylinder, which is constructed from two conventional cylinder lenses of equal but opposite powers oriented with their axes mutually perpendicular.

All three of the conventions described above may be said to be in polar form since the astigmatic component of the correction is given in terms of a magnitude (lens power) and an angle (the orientation of the lens axis). For computational purposes, conversion from polar form to rectangular form is usually advantageous. This is easily accomplished for the cross-cylinder

convention using the following transposition formulae:

$$\begin{aligned} M &= S + C/2 \\ J_0 &= -\frac{C}{2} \cos 2\alpha \\ J_{45} &= -\frac{C}{2} \sin 2\alpha \end{aligned} \quad (5)$$

The three-dimensional vector $[M, J_0, J_{45}]$ is known in the ophthalmic optics literature as a power vector and the length of a power vector is called blur strength. Blur strength quantifies the overall blurring effect of sphero-cylindrical refractive error on image quality according to geometrical optics. The components of a power vector may be converted to second-order Zernike aberration coefficients using the following formulae:

$$\begin{aligned} M &= \frac{-c_2^0 4\sqrt{3}}{r^2} \\ J_0 &= \frac{-c_2^2 2\sqrt{6}}{r^2} \\ J_{45} &= \frac{-c_2^{-2} 2\sqrt{6}}{r^2} \end{aligned} \quad (6)$$

where c_n^m is the n th order Zernike coefficient of meridional frequency m and r is the pupil radius. If Zernike coefficients are specified in microns (10^{-6} m) and pupil radius is specified in millimeters (10^{-3} m) then the computed values of M , J_0 and J_{45} will have units of diopters (m^{-1}).

Specification of Higher-Order Aberrations of Eyes

The accepted convention for reporting Zernike aberration coefficients in the ophthalmic optics field is given in ANSI standard Z80.28. (ANSI, 2010) The normalization scheme used in this standard allows an interpretation of individual Zernike coefficients as the RMS wavefront error of the corresponding aberration mode. Accordingly, we may infer from Eq. (6) that defocus (M) and astigmatism (J_0, J_{45}) in diopters are proportional to the ratio of RMS wavefront error to pupil area.

Zernike polynomials Z_n^m are indexed according to their radial order n and their meridional frequency m . Historically, the attention of clinical practitioners has been directed at the lower Zernike orders 1–2 but increasing attention to the higher orders ($n > 2$) has occurred in recent years due to the technological development of wavefront aberrometers (see Section Aberrometry and Wavefront Sensing). The mathematical formula describing individual Zernike basis functions, or modes, is given by the equation

$$Z_n^m = N_n^m R_n^{|m|}(\rho) M(m\theta) \quad (7)$$

which is separable into a function of pupil-plane coordinates ρ (the normalized radial coordinate) and θ (meridian angle, measured counterclockwise from the horizontal when viewing the patient). The radial term for the Zernike polynomial function with indices n and m is given by the equation

$$R_n^{|m|}(\rho) = \sum_{s=0}^{0.5(n-|m|)} \frac{(-1)^s (n-s)!}{s! (0.5(n+|m|)-s)! (0.5(n-|m|)-s)!} \rho^{n-2s} \quad (8)$$

where s is an integer summation index incremented by one unit. The meridional (or azimuthal) term for a Zernike polynomial function with index m is given by the equations

$$\begin{aligned} M(m\theta) &= \cos(m\theta) \quad \text{if } m \geq 0 \\ M(m\theta) &= \sin(|m|\theta) \quad \text{if } m < 0 \end{aligned} \quad (9)$$

The normalization term N in Eq. (7) makes every Zernike mode have unit variance over the unit circle and is defined by the equation

$$N_n^m = \sqrt{(2 - \delta_{0,m})(n+1)} \quad \text{where } \delta_{0,m} = 1 \text{ if } m = 0, \quad \delta_{0,m} = 0 \text{ if } m \neq 0 \quad (10)$$

Zernike aberration coefficients provide the weighting factors needed to reconstruct a wavefront aberration for an individual eye with a weighted sum of orthonormal basis functions. The circle polynomials of Zernike were selected over the traditional Seidel series for the ANSI standard because they are a complete set that can be used to represent any continuous, smooth function over a circular domain. Moreover, individual basis functions in the set correspond to common aberrations found in optical systems (e.g., defocus, coma, spherical aberration). One disadvantage of Zernike basis functions is that their spatial derivatives are not orthogonal, which hampers the accurate calculation of coefficients from wavefront slope measurements. This disadvantage is overcome by fitting slope data with Zernike radial slope polynomials (Nam *et al.*, 2009).

Another disadvantage stems from the fact that Zernike basis functions of a given order are balanced with basis functions of lower orders to achieve orthogonality and minimize RMS variation. For example, the basis function for Zernike spherical aberration is constructed by subtracting an amount of defocus (which varies as ρ^2) from Seidel spherical aberration (which varies as ρ^4). Thus the total amount of defocus (in the ρ^2 sense) is distributed across multiple Zernike modes, a fact that can hamper interpretation and complicates the calculation of power vector description of refractive errors when higher-order aberrations are present. These problems are reduced by an alternative set of basis functions (Stephenson, 2009) that combine the benefits of Zernike polynomials (completeness) and Seidel polynomials (un-balanced).

Methods for Determining Refractive Error and Refractive State

Refraction is a clinical process for determining the refractive error of an eye. Refractions can be performed objectively with a clinical autorefractor, an instrument that determines spherical and astigmatic focusing errors of the eye, or with an aberrometer, an instrument that also measures higher-order aberrations. Traditionally, refractive errors are measured subjectively by clinicians using psychophysical methods to determine the best combination of spherical and cylindrical correcting lens for optimizing vision and therefore, by implication, optimizing focusing of the retinal image. Clinicians generally regard the subjective refraction as definitive since it includes the patient's judgment of optical focus based on their own visual perception using their own retina and brain. However, other considerations (e.g., matching image size in the two eyes) may lead the clinician to prescribe correcting lenses that are somewhat suboptimal optically in order to achieve overall objectives of visual comfort and performance. Consequently, the prescribed lens is often different from that indicated by the refraction procedure.

Locating the Eye's Far Point With Optical Measurements

The far point of an eye is that point in object space which is optically conjugate to the entrance apertures of the foveal photoreceptors when accommodation is relaxed. Since refractive error is equal to the vergence of the far point according to Eq. (3), locating an eye's far point is a straightforward means of determining refractive error. Unfortunately, an unambiguous far point exists only for an ideal eye which is free of all optical aberrations except defocus. Real eyes invariably suffer from other aberrations besides defocus, and therefore an operational definition of the far point is required. For example, an astigmatic eye may be described as having two far points, one for line objects oriented parallel to the axis of astigmatism and another for the orthogonal orientation. The point midway between these two extremes is a widely accepted approximation for the far point of an astigmatic eye since this is where the far point would be located if the astigmatism were corrected by a lens. This approximation assumes higher-order aberrations are negligible, in which case refractive errors may be computed directly from second-order Zernike coefficients using Eqs. (5) and (6).

One reasonable approximation for the far point of an eye with higher-order aberrations is that location in object space where an object must be located in order to maximize the quality of its retinal image. Unfortunately, this approximation does not locate the far point uniquely for several reasons. Firstly, image quality depends on the characteristics of the visual stimulus (e.g., luminance, size, spatial frequency spectrum, wavelength spectrum) and therefore may change as the stimulus changes. Secondly, there are many ways to define and measure image quality (Thibos *et al.*, 2004), none of which are universally endorsed by the clinical or vision science communities. Thirdly, the psychophysical judgment of optimum focus by an observer depends on factors other than optical image quality, such as neural sampling of the retinal image, cortical neural processing, and perceptual weighting given to the various features of optical blur (e.g., contrast attenuation, loss of sharpness, and doubling or ghosting of the image). Moreover, the selection of an objective method for locating the nominal far point of an eye is likely to depend on the intended application.

Definition of the Eye's Refractive State

The human eye is an active optical system that responds to changes in viewing distance by changing its refractive power, a process called accommodation. At any given state of accommodation, each point on the retinal surface is optically conjugate (at least approximately) to some location in visual space. The locus of all such points is called the retinal conjugate surface. Refractive state is defined as the vergence of the retinal conjugate surface, which in general will vary across the visual field. Thus, refractive state is essentially the same concept as refractive error defined above, but extended to apply to the accommodating eye. Although it would seem reasonable to use the general term "refractive state" for an eye in any state of accommodation, the historical precedent is to use the term "refractive error" for the special case of relaxed accommodation (i.e., minimum power). Clearly, the same difficulty in establishing a universally accepted method for specifying refractive error described above applies equally to specifying refractive state.

Methods for Estimating Refractive State From Zernike Aberration Coefficients

Wavefront aberrometers make comprehensive measurements of an eye's aberrations reported in the form of Zernike aberration coefficients, C_n^m . From these measurements one may compute the eye's wavefront aberration function $W(\rho, \theta)$ as a weighted sum of

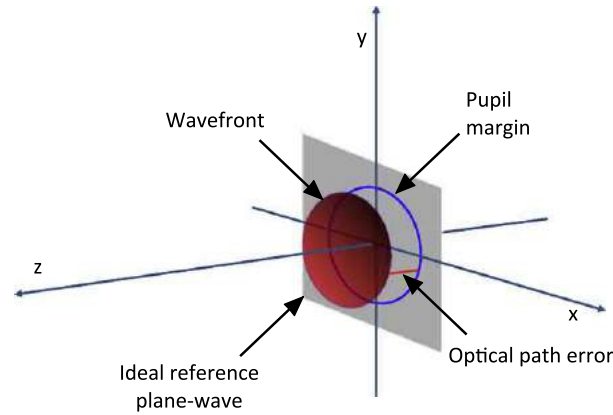


Fig. 2 A reference coordinate system for specifying wavefront errors. A right-hand (x, y, z) coordinate reference frame is constructed such that the z -axis is the chief ray and the (x, y) plane is perpendicular to the z -axis at the center of the exit pupil.

Zernike polynomials,

$$W(\rho, \theta) = \sum_{n,m} c_n^m Z_n^m \quad (11)$$

A given aberration function refers to a specific visual direction, which is typically the foveal line of sight defined by the path of the chief ray in object space joining the point of regard (i.e., the fixation point) with the center of the eye's entrance pupil. More generally, the chief ray for any point in visual space defines a line-of-sight along which an aberration function may be specified.

Wavefront aberration functions, or maps, have a variety of physical, mathematical, and practical interpretations and uses. Physically, the map describes the optical distance from an object point to the corresponding retinal image point through different parts of the pupil, relative to the optical distance along a path through the center of the pupil (i.e., along the chief ray). An equivalent interpretation, illustrated in Fig. 2, is the sag (relative to an ideal plane-wave) in the z -direction of a wavefront reflected from the eye produced by a point source on the retina. These differential errors in optical path length cause phase errors in light arriving at the image point, which in turn degrades image quality. Mathematically, the aberration map $W(\rho, \theta)$ is the phase portion of the generalized pupil function $P(\rho, \theta)$ defined in optical theory as

$$P(\rho, \theta) = A(\rho, \theta) \exp(-ikW(\rho, \theta)) \quad (12)$$

where $A(\rho, \theta)$ is the corresponding amplitude function and $k=2\pi/\lambda$. The generalized pupil function is a fundamental description of an optical system that enables calculation of the optical point spread function (PSF) and optical transfer function (OTF) (Goodman, 2005), both of which are functional measures of optical quality. Scalar metrics of image quality computed from the PSF or OTF are useful in practical optimization problems, such as determining refractive state or prescribing lenses to correct the eye's aberrations.

Aberrometry and Wavefront Sensing

Wavefront sensors, the core technology used in ocular aberrometers, are both new and old. In the clinical literature, the first applications of aberrometry to refractive surgery (Applegate and Howland, 1997) and to ocular disease (Thibos and Hong, 1999) appeared near the end of the 20th century years. However, the fundamental principle by which modern aberrometers measure ocular wave-front aberrations of eye was discovered 400 years ago by the celebrated Jesuit philosopher-astronomer, Christopher Scheiner (Scheiner, 1619), a contemporary of Galileo and Kepler. All three astronomers conceived of the eye as an optical instrument that forms retinal images in the same way that a telescope forms images. However, it was Scheiner who first demonstrated how to measure the refractive error of the human eye using a simple device that is known today as Scheiner's disk. Although Scheiner used his invention to measure and study only spherical refractive error (i.e., defocus, which is the simplest and most prevalent of all ocular aberrations), the basic idea was generalized by Thomas Young at the beginning of the 19th century to measure astigmatism in his own eye (Young, 1801), and eventually by Liang *et al.* at the end of the 20th century to measure higher-order aberrations (Liang *et al.*, 1994). Thus, Scheiner's disk is the prototype of the modern Shack-Hartman wavefront sensor for measuring ocular aberrations.

Determining Optical Path Lengths From Ray Deflections

The challenge of using a Shack-Hartmann sensor with a living eye is that the eye is a closed system, so it can be measured only by reflecting light off the patient's retina as illustrated in Fig. 3. Thus, for measurement purposes, object space is inside the eye and image space is the external world. To make measurements, a narrow laser beam is introduced into the eye using a beam-splitter to produce a source of reflected light, a so-called retinal beacon. Light reflected out of the eye passes through an array of small,

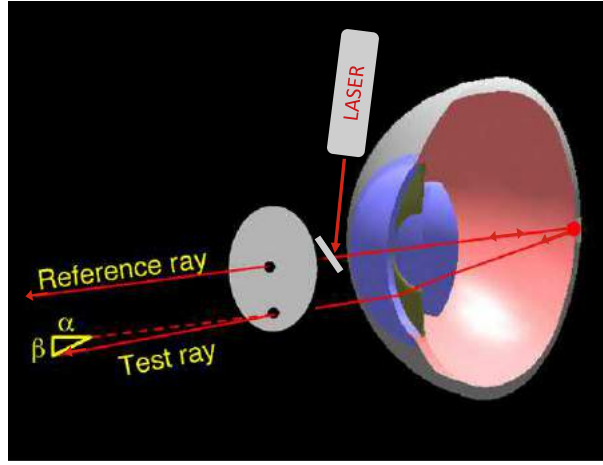


Fig. 3 Principle of operation of the Scheiner–Hartmann–Shack wavefront aberrometer. A laser beam injected into the eye produces a retinal beacon that reflects light out of the eye for measurement. Angular deviation of an isolated test ray relative to the direction of the reference ray is a measurement of local wavefront slope. In a practical instrument, multiple apertures defined by a micro-lenslet array produce simultaneous measurements at many points in the eye's pupil.

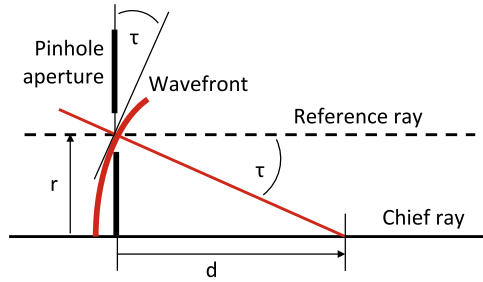


Fig. 4 Geometry for defining wavefront vergence. A pinhole aperture located distance r from the z -axis isolates a ray perpendicular to the wavefront and intersecting the z -axis at some distance Z from the aperture plane. Skew rays that do not intersect the chief ray are projected into the plane of the diagram to determine the radial component of vergence.

pinhole apertures (i.e., Scheiner's disk or a Hartmann screen) to isolated rays exiting through specific locations in the eye's pupil. The ray's direction of propagation relative to the chief ray is specified by (α, β) , the angles of deviation from a secondary reference ray (dashed line) parallel to the chief ray and passing through the aperture. In a modern device, the array of pinholes is replaced by an array of micro-lenslets that subdivides the beam of light reflected out of the eye from the retinal beacon. Each lenslet forms a spot image that deviates horizontally and vertically from the optical axis of the lenslet by an amount that is directly proportional to the angular deviations (α, β) . Since rays are, by definition, normal to the exiting wavefront, the sensor is essentially a mechanism for measuring wavefront slope, from which wavefront shape is recovered by integration.

Converting Ray Aberrations to Wavefront Aberrations

As described in Section Determining Optical Path Lengths From Ray Deflections, the principle of operation of the Shack–Hartman wavefront sensor used in ocular aberrometers is to measure the deviation of individual rays of light passing through various locations in the eye's pupil. From the geometry of Fig. 4 we see that, for small angles of deviation of the ray from the reference direction, angle τ and wavefront slope are approximately equal,

$$\text{wavefront slope} = \tan(\tau) \cong \tau = \text{deviation angle} \quad (13)$$

In general, the aberrated ray does not lie in the meridional plane defined by the optical axis of the system and a point on the wavefront located at the foot of the ray. However, if we project the ray onto the meridional plane of Fig. 4, then another useful relationship can be deduced from the geometry:

$$\frac{\text{wavefront slope}}{\text{location in pupil}} = \frac{\partial W(\rho, \theta) / \partial \rho}{\rho} = \frac{\tan(\tau)}{r} = \frac{1}{d} = \text{wavefront vergence} \quad (14)$$

This equation says that, for any meridian θ , the radial component of wavefront vergence (in diopters) is equal to the ratio of radial wavefront slope (in radians) to pupil location (in meters). So, although wavefront vergence is a vector quantity with

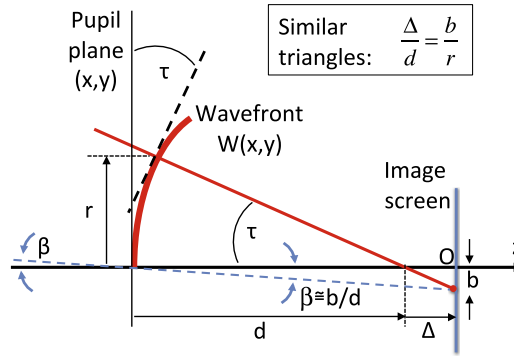


Fig. 5 Geometry for expressing the relationship between wavefront aberrations and the size of blurred images. The diagram depicts a cross-section of a reflected wavefront of light emerging from the eye produced by a point source on the retinal fovea. The plane of the diagram is perpendicular to the eye's entrance pupil and contains the foveal line of sight (i.e., the z-axis). An isolated ray passing through the pupil plane at distance r from the pupil center intersects the image screen at distance b from the z-axis and thus contributes one point to the blurred image. (For simplicity, we assume the ray is in the plane of the diagram but in general this is not necessarily true.) The angular size (in radians, subtended at the pupil center) of this blurred point is approximately equal to the focusing error E (in diopters) of the aberrated ray (which depends on the axial location of the image screen) and radial distance r (in meters).

orthogonal components proportional to (α, β) , the radial component is a scalar quantity that may be used to recover wavefront phase by integration (Nam *et al.*, 2009).

A similar geometrical analysis reveals the relationship between wavefront aberrations and the size of blurred images. Although we are ultimately interested in blurring of the retinal image, aberration measurements are performed on the reflected wavefront emerging from the eye as illustrated in Fig. 5. Consider an image screen that is placed some small distance Δ from the focused plane where an isolated ray intersects the chief ray (i.e., the z-axis). This focusing error causes the image point contributed by the illustrated ray to be displaced by amount b from the image point O contributed by the chief ray. In other branches of optics, this linear displacement in the image plane is of primary concern. However, in vision science we are typically interested more in the angle β subtended by b at the pupil center. If we assume the pupil is near the eye's nodal point then the angles of blur subtended in image and object space are the same. The extent of this angular blur is easily computed if we express ray aberrations as a vergence error in diopters. That is, let the vergence of the image screen be $1/(d + \Delta)$ and the vergence of the ray in this example be $1/d$, so the vergence error E is

$$E = \frac{1}{d + \Delta} - \frac{1}{d} = -\frac{\Delta}{d^2 + d\Delta} \cong -\frac{\Delta}{d^2} \quad \text{for } \Delta \ll d \quad (15)$$

The negative sign in this result means a negative correcting lens would be required to focus this aberrant ray. A negative sign also means the ray crosses the optical axis before striking the image surface.

From the geometry of similar triangles in Fig. 5, $\Delta/d = b/r$ and for small values of defocus, $\beta \cong b/d$. Substituting these two relationships into Eq. (15) reveals that the angular subtense β of the blurred image point (in rad) produced by the aberrated ray is the product of focusing error E (in diopters) and radial position r of the ray in the pupil plane (in meters),

$$E = -\frac{b}{rd}, \quad -Er = \frac{b}{d} \cong \beta \quad (16)$$

The minus sign in Eq. (16) indicates that the ray intersected the chief ray before arriving on the opposite side of the image screen in this example. Although the derivation described the deviated ray as a focusing error for simplicity, Eq. (16) applies even when higher-order aberrations are present if E is computed from wavefront vergence as defined by Eq. (14).

The three fundamental concepts of wavefront phase, wavefront slope, and wavefront curvature can each be used to assess optical quality of the eye and determine the optimum sphero-cylindrical correcting lens as described in Section Optimizing Retinal Image Quality by Correcting Refractive Errors. These properties of a wavefront are linked together by geometry and calculus in a manner directly analogous to distance, velocity, and acceleration. Slope is the spatial derivative of wavefront phase and curvature is the spatial derivative of slope. Conversely, slope is the integral of curvature and phase is the integral of slope. Thus, measurements of wavefront slope provided by a wavefront sensor may be mathematically integrated to retrieve the shape of the aberrated wavefront. Just as distance, velocity, and acceleration each have different physical units, wavefront phase, slope, and curvature have different units. Wavefront phase is measured in microns, which is perhaps more easily interpreted if normalized by the wavelength of light to yield waves of phase shift. Wavefront slope is specified in micrometer per millimeter, which is the same as milliradians (1 mrad = 3.438 arcmin) and wavefront curvature is specified in micrometer per square millimeter which is the same as inverse meters or diopters.

An important practical application of Eq. (14) is to derive a formula for paraxial refractive error that takes into account the fact that 2nd order Zernike modes are included in higher-order modes for the purpose of building an orthogonal set of basis functions with minimum variance. Thus to determine paraxial refractive errors, it is necessary to apply Eq. (14) to the Zernike expansion given by Eq. (11) and evaluate the result at the pupil center. Since all terms in ρ^n with $n > 2$ have zero slope when $\rho = 0$, they make no contribution to paraxial wavefront vergence according to Eq. (14). For a 6th order Zernike expansion, the resulting formula for paraxial refractive error reduces to

$$\begin{aligned} M &= \frac{-c_2^0 4\sqrt{3} + c_4^0 12\sqrt{5} - c_6^0 24\sqrt{7}}{r^2} \\ J_0 &= \frac{-c_2^2 2\sqrt{6} + c_4^2 6\sqrt{10} - c_6^2 12\sqrt{14}}{r^2} \\ J_{45} &= \frac{-c_2^{-2} 2\sqrt{6} + c_4^{-2} 6\sqrt{10} - c_6^{-2} 12\sqrt{14}}{r^2} \end{aligned} \quad (17)$$

Optimizing Retinal Image Quality by Correcting Refractive Errors

The ultimate decision of whether a retinal image is optimally focused by a sphero-cylindrical correcting lens is complicated by the presence of higher-order aberrations and by how the question is framed. If the decision is an esthetic judgment, the observer's subjective criterion for making the decision may be unknown or difficult to articulate. If the decision is based on performance of an objective task, then optimum focus may depend on the nature of the task (e.g., reading printed text versus detecting a target) or on stimulus parameters (e.g., size, velocity, intensity, etc.). Experiments on human observers have shown that, for the task of reading small letters, refractive corrections based on Eq. (6) tend to leave the eye slightly myopic whereas corrections based on Eq. (17) tend to leave the eye slightly hyperopic (Martin *et al.*, 2011). The optimum correction lies between these limits, suggesting that optimization of some metric of retinal image is a better way to compute refractive error from wavefront aberration measurements. Numerous metrics have been evaluated for this purpose, several of which provide unbiased measures of optimum focus. One widely used unbiased metric called the visual Strehl ratio is analogous to the common Strehl ratio but is determined for a neutrally weighted optical PSF obtained by convolving the optical PSF with a neural weighting function that represents the spatial filtering properties of the visual system.

See also: Interferometric Imaging

References

- ANSI, 2010. American national standard Z80.28 for ophthalmics – Methods for reporting optical aberrations of eyes, ANSI Z80.28, American National Standards Institute.
- Applegate, R.A., Howland, H.C., 1997. Refractive surgery, optical aberrations, and visual performance. *Journal of Refractive Surgery* 13 (3), 295–299.
- Goodman, J.W., 2005. *Introduction to Fourier Optics*. Greenwood Village, CO: Roberts & Co.
- Liang, J., Brimm, B., Goelz, S., Bille, J.F., 1994. Objective measurement of the wave aberrations of the human eye with the use of a Hartmann–Shack wave-front sensor. *Journal of the Optical Society of America, A* 11, 1949–1957.
- Martin, J., Vasudevan, B., Himebaugh, N., Bradley, A., Thibos, L., 2011. Unbiased estimation of refractive state of aberrated eyes. *Vision Research* 51 (17), 1932–1940.
- Nam, J., Thibos, L.N., Iskander, D.R., 2009. Zernike radial slope polynomials for wavefront reconstruction and refraction. *Journal of the Optical Society of America A Optics, Image Science, and Vision* 26 (4), 1035–1048.
- Scheiner, C., 1619. *Oculus, Sive Fundamentum Opticum*. Innspruk: Agricola.
- Stephenson, P.C., 2009. Optical aberrations described by an alternative series expansion. *Journal of the Optical Society of America A Optics, Image Science, and Vision* 26 (2), 265–273.
- Thibos, L.N., Hong, X., 1999. Clinical applications of the Shack–Hartmann aberrometer. *Optometry and Vision Science* 76, 817–825.
- Thibos, L.N., Hong, X., Bradley, A., Applegate, R.A., 2004. Accuracy and precision of objective refraction from wavefront aberrations. *Journal of Vision* 4 (4), 329–351.
- Young, T., 1801. The Bakerian lecture: On the mechanism of the eye. *Philosophical Transactions of the Royal Society of London* 91, 23–88.

Methods of Vision Correction

Len Zheleznyak, University of Rochester, Rochester, NY, United States

Ramkumar Sabesan, University of Washington School of Medicine, Seattle, WA, United States

Geunyoung Yoon, University of Rochester, Rochester, NY, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

For centuries humans have been inventing technologies to correct the eye's optical aberrations. As our understanding of the eye's optical defects has evolved, so too has the complexity of correction methods. But despite technological advances, from the medieval looking-stone to the modern-day femtosecond laser, aberration-free vision remains out of reach for most of the population.

Two primary approaches to correcting the eye's refractive error are wide-spread:

1. Introducing corrective lenses in front of the eye (e.g., spectacles or contact lenses).
2. Altering the eye's optical properties by modifying the cornea (e.g., refractive surgery) or replacing the crystalline lens with an intraocular lens.

In this article, we will review both approaches to correcting the eye's optical aberrations to improve visual performance. Finally, we consider promising future approaches to vision correction.

Background

Spectacles

The inventor of spectacles is unknown, however, it is thought that spectacles were invented in the late 1200s (Dreyfus, 1988). The earliest known depiction of spectacles in a painted work of art is the portrait of Cardinal Hugo of Providence, painted by Tommaso de Modena in 1352 (Fig. 1).

Spectacle technology has advanced in terms of lightweight optical materials and anti-reflection coatings, but their main function of correcting the eye's lower order aberrations (sphere and cylinder) has remained relatively unchanged.

Fig. 2 illustrates a myopic (i.e., near-sighted) eye before and after correction with a spectacle lens. In the case of myopia, the eye has excess refractive power (sphere), resulting in a focus in front of the retina (Fig. 2(a)) for light coming from infinity (i.e., a collimated beam). Therefore, the required corrective lens has negative defocus, which pushes the focus posteriorly towards the retina (Fig. 2(b)).

Despite the efficacy and convenience of spectacles, their limitations for certain optical corrections are significant. For example, as the eye's directional gaze changes, a different area of the spectacle is used. This is akin to a decentration, or misalignment of the correction-lens' wavefront from the eye's wavefront. This is not an issue for conventional lenses that only correct lower order aberrations, however it does prevent efficacious higher-order aberration correction in spectacles.

Contact lenses

While the initial ideas behind contact lenses can be traced back to the 16th century to Leonardo da Vinci, it was not until 1888 that a German ophthalmologist, Adolf Fick, successfully fit the first contact lens. Fick used glass blowing to create lenses which rested on the sclera of the eye and neutralized the refractive power of the cornea by filling the space between with dextrose. The solution, given its similarity in refractive index to the cornea, would minimize the power arising from the corneal front surface, otherwise exposed to air. Indeed, the same refractive index matching approach to alter the corneal power was proposed earlier by Leonardo da Vinci, and later used by Thomas Young to confirm the presence of astigmatism in the cornea (Fig. 3).

Since their inception, contact lenses have taken different forms varying in their size, material properties, oxygen permeability and optical characteristics. Their growth is marked by two key inventions. First, the development of polymethyl methacrylate (PMMA) made lenses lighter and smaller than those made from glass, leading to contacts which could be worn on the cornea as opposed to earlier glass sclerals. These plastic polymers were later refined to allow for better oxygen permeability and led to a class of lenses which are now more commonly called rigid gas permeable (RGP) contact lenses. The second revolutionary advance came with the development of soft hydrogel materials (Wichterle and Lim, 1960) which substantially increased the wear time of contacts by improving oxygen permeability and led to the conventional "soft" contact lenses. While the permeability characteristics of such hydrogels were advanced following their invention, a major breakthrough came about from the development of silicone hydrogels in the 1990s that allowed for extremely high oxygen permeability leading to extended wear times and wearing comfort. Note however that all the aforementioned types and materials (except for glass) of contacts are still in use today for varying degree of abnormalities in the cornea and the ocular surface. In addition to optical correction discussed below, some applications of contact lenses pertain to non-refractive cases also. Scleral lenses ranging in diameter from 13 to 24

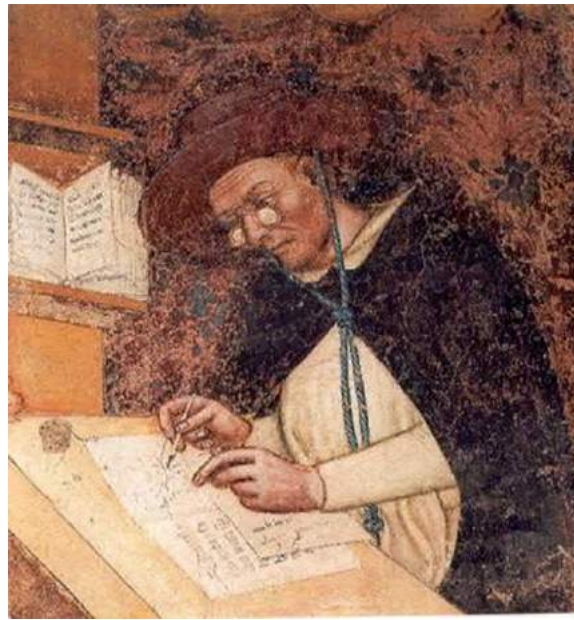


Fig. 1 Earliest depiction of spectacles in art: Cardinal Hugo of Providence. Painted by Tommaso de Modena, 1352.

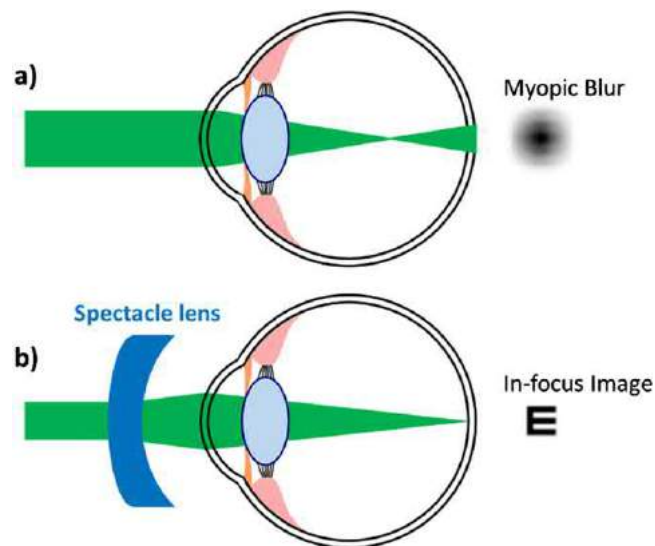


Fig. 2 Schematic of an (a) un-corrected and (b) corrected myopic eye with a negative diopter spectacle lens.

mm rest on the sclera of the eye and create a reservoir tear layer between the cornea and the lens. By virtue of being placed on the sclera and not the sensitive corneal tissue, these lenses are often used to provide relief in patients with corneal disease or following injury- for instance in extreme cases of dry eye, corneal ectasia, post-LASIK or corneal transplant complications, Stevens-Johnson syndrome and Sjogren's syndrome. A similar protective role for the cornea is met by bandage soft contact lenses as well.

Contact lenses provide a unique vehicle for delivering optical corrections to the eye – both for conventional refraction (sphere and cylinder) and for those corneas affected by higher order aberrations beyond sphero-cylindrical errors. The advantage of contacts in providing optical correction arises mainly because the lenses sit in close axial proximity to the eye's optical and physical pupil leading to two key desirable features. First, when the eyeball moves, the contacts move in unison, thereby continuously providing the requisite optical correction despite large changes in gaze. Second, the accessible field of view remains large and consistent across a variety of viewing conditions. Conversely, spectacles, by virtue of being stationary, do not align the eye's optical axis when the gaze changes. Furthermore, by being displaced axially from the eye's pupil, the isoplanatic patch (the field of view over which the aberrations remain relatively constant) shrinks and leads to a smaller effective field of view when using spectacles.

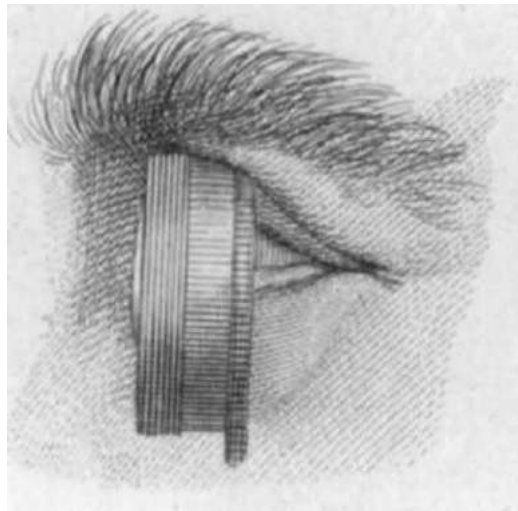


Fig. 3 Thomas Young's illustration of neutralizing the cornea by refractive index matching by applying a lens with water reservoir. Reproduced figure from 1801 publication.

These two factors are less critical for conventional sphero-cylindrical corrections however pose a significant challenge for compensating large magnitudes of asymmetric optical aberrations.

The overall principle of a contact lens in providing refractive error correction is similar to that of spectacles, albeit with a few key differences. Similar to a spectacle prescription, a positive/negative powered lens is used to correct for hyperopia/myopia, while unequal spherical powers across the horizontal and vertical meridians are used to correct for cylinder. However, the magnitude of refraction needs to accommodate for the different axial locations of the spectacle and contact lens corrections. In a clinical setting, prescriptions are sought using trial lenses placed at the spectacle plane. To convert the spectacle prescription to contact lens power at the cornea, the following equation is used, where K : refraction at the corneal/contact lens plane, F : refraction at the spectacle plane, d : vertex distance or the distance between the spectacle and the corneal plane (typically 12–13 mm).

$$K = \frac{F}{1 - dF}$$

In addition to the optical considerations, contacts are designed to accommodate for the physiological demand of supporting metabolism in the cornea via replenishing the supply of the tear film. The physical characteristics (curvature, diameter, edge profile, thickness) are thus optimized for the lens to move with every blink and realign with the visual axis afterwards. The alignment is less critical for pure spherical error correction compared to cylindrical and higher order aberration. This is because spherically asymmetric errors are not corrected optimally when the optical axes of the eye and the corrective optics are not co-aligned. Common strategies to align the lens include prism-ballast or double slab-off wherein different thickness profiles across the surface of the lens help in its optimum interaction with the eyelids and stabilizing it post-blink (Holden, 1975).

The efforts to compensate for higher order aberrations using ophthalmic lenses received a major boost mainly due to the advent of wavefront sensing techniques (Liang *et al.*, 1994). Using these advanced techniques, the corneal abnormality of keratoconus (KC) and its optical consequences have been evaluated (Pantanelli *et al.*, 2007; Maeda *et al.*, 2002). Given the large magnitude of higher order aberrations found in KC, it was immediately appreciated that these patients stand to benefit greatly from an optical correction aimed at minimizing these aberrations.

Soft contact lenses with wavefront-guided surface profiles have been shown to have potential in correcting HOA and improving vision in KC (López-Gil *et al.*, 2003; Sabesan *et al.*, 2007; Marsack *et al.*, 2008). In such a scheme of correction in two eyes with KC shown by López-Gil *et al.*, the lens was designed according to the aberration profile of the eye, but the reduction in higher order aberration and improvement in vision were relatively small (López-Gil *et al.*, 2003). One possible explanation for the small average reduction in higher order aberration could be the failure to account for the decentration and rotation of the contact lens on the eye. Static and dynamic contact lens movements critically affect correction performance of HOA, as the lens is not aligned to the center of the visual axis, especially in eyes with abnormal corneal surface profiles (Navarro *et al.*, 2000; Guirao *et al.*, 2001). Sabesan *et al.* demonstrated a scheme of correction in KC using soft contact lenses where the lens was designed by accounting for both the eye's aberration profile and the static lens decentration and rotation on eye (Sablesan *et al.*, 2007). Using these wavefront-guided customized soft contact lenses, an improvement in optical quality by a factor of 3 in HOA with respect to the conventional lens on average in 3 KC patients was demonstrated. The improved optics resulted in an average improvement of 2.1 lines in visual acuity over the conventional correction of defocus and astigmatism alone. However, the residual higher order wavefront error was still nearly double of what is observed in normal eyes. This residual error was explained to a reasonable extent by the manufacturing error and lens movement. To reduce the variability of lens position by conferring mechanical stability between blinks, Chen *et al.* (2007) employed back surface customized soft contact lenses whose posterior surface profiles were sculpted to match

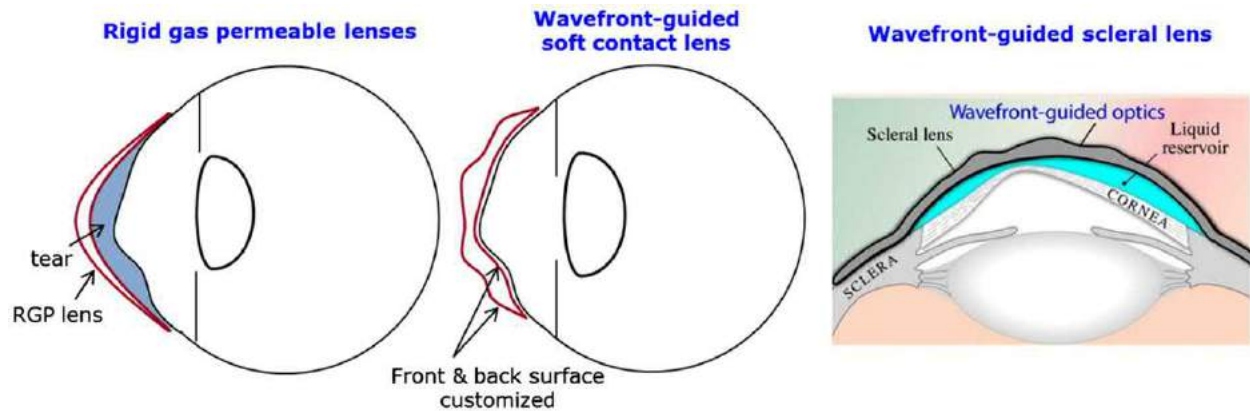


Fig. 4 Different contact lens strategies to overcome corneal higher order aberrations in highly aberrated eyes, such as keratoconus.

the anterior corneal surface in KC. On-eye performance of the back surface customized soft contact lens demonstrated that the lens stability was improved by a factor of 2 for horizontal and vertical decentration, and a factor of 5 in rotational orientation over conventional lens. However, significant residual HOA induced by internal optics, especially posterior corneal surface still degraded retinal image quality. Additional customization of the front surface of the lens thus has the potential to further correct these residual aberrations.

Clinically, RGPs and scleral lenses are considered the standard of correction in KC. These lenses achieve correction by masking corneal irregularities with the tear lens between the posterior lens surface and the anterior corneal surface, in the same manner as the first attempts of fitting contacts. RGP corneal contact lenses are 8.5–10.5 mm in diameter and cover only 75–80% of the cornea. Mini-scleral, corneo-scleral and scleral lenses which range in diameter from 13–24 mm, depending on type or fit, may rest partly on the cornea. A scleral lens prosthetic device (SLPD) with diameters ranging from 17.5–24 mm is designed and fit to vault the cornea entirely. A noteworthy difference between RGP corneal lens and SLPD is the dynamic movement of these corrective devices on the eye. A well-fitted corneal RGP lens slides with each blink to allow for tear exchange necessary for physiological tolerance at the corneal surfaces where contact is made. Optical correction in an RGP corneal lens is thus inherently unstable. A carefully fitted SLPD is expected to exhibit minimal movement because the bearing haptic aligns with a large area of conjunctiva overlying the sclera. Suction is avoided by precise alignment with the sclera or by creation of channels on the posterior surface (Rosenthal, 2009, 2010). The clinical benefits of an SLPD in terms of improvement in visual acuity and visual function across a wide range of diagnoses including KC have been established (Stason *et al.*, 2010). By virtue of their larger diameter and broader bearing zone, these devices have been employed in the treatment of various other forms of corneal ectasia, corneal irregularity following transplant, ulcers, dry eye syndrome, ocular surface disease and others (Pullum *et al.*, 2005; Segal *et al.*, 2003). Reduction of HOA has been reported in normal and KC eyes with corneal RGP lens (Hong and Himebaugh, 2001; Dorronsoro *et al.*, 2003; Marsack *et al.*, 2007; Negishi *et al.*, 2007) and SLPD (Gumus *et al.*, 2011). However, since these lenses minimize only anterior corneal aberrations, significant posterior corneal aberrations remain uncompensated (Chen and Yoon, 2008). In addition, aberrations such as coma and astigmatism are induced due to RGP corneal lens rotation and decentration, further degrading retinal image quality. The positional stability between blinks due to the large surface coverage in SLPD makes it an ideal platform for wavefront-guided correction of HOA in KC. Sabesan *et al.* demonstrated significant reduction in root-mean-square of HOA and improvement in vision in 11 advanced KC eyes wearing wavefront-guided SLPD (Sabesan *et al.*, 2007). HOA were effectively corrected by the customized SLPD with wavefront-guided optics and RMS was reduced 3.1 times on average to $0.37 \pm 0.19 \mu\text{m}$ for a 6 mm pupil, comparable to normal levels. This correction resulted in significant improvement of 1.9 lines in mean visual acuity ($P < 0.05$). Contrast sensitivity was also significantly improved by a factor of 2.4, 1.8 and 1.4 on average for 4, 8 and 12 cycles/degree, respectively ($P < 0.05$ for all frequencies). Interestingly, although the residual aberration was comparable to that of normal eyes, the average visual acuity in logMAR with the customized SLPD was 0.21, substantially worse than normal acuity. This implicates the role of chronic exposure to poor retinal image quality that may restrict the visual benefit achievable immediately after wavefront-guided customized treatment in these eyes (Fig. 4).

Visual Benefit of Correcting Higher Order Aberrations

The amount of benefit to be gained from correcting the eye's higher order aberrations depends on subjects' native wavefront profiles and pupil size. However, the visual benefit has been shown to be significant for various visual tasks (e.g., visual acuity, contrast sensitivity), especially in low light levels where pupils are large or in ocular pathologies such as KC. Higher order aberration correction in the human eye has been demonstrated with several methods, such as soft contact lenses (Chen *et al.*, 2007; Sabesan *et al.*, 2007), scleral lenses (Sabesan *et al.*, 2013; Marsack *et al.*, 2014), phase plates (Yoon *et al.*, 2004; Allen and Raasch, 2005) and adaptive optics (Yoon and Williams, 2002; Legras and Rouger, 2008; Rossi *et al.*, 2007; Sabesan *et al.*, 2007, 2012). However, the degree of visual benefit can vary widely depending on the individual subjects.

In a seminal population study ($n=109$ eyes) by Williams *et al.* (2000), it was found that contrast sensitivity was expected to improve by a factor between 1 (no improvement) and 8 times (16 cyc/deg, 5.7 mm pupil) after higher order aberration correction. The variability in visual benefit indicates that some subjects are more suited for higher order aberration correction than others. Patients with large amounts of higher order aberrations or abnormal eyes (e.g., KC, high spherical aberration post-laser refractive surgery) would clearly stand to benefit from higher order aberration correction. However, aside from the magnitude of aberrations, other factors such as subjects' native pupil size should also be considered. As pointed out by Williams *et al.*, visual benefit is minimal in bright light conditions (i.e., when the pupil is small) or in the presence of focusing error, which is commonplace due to accommodative lag (Williams *et al.*, 2000).

Furthermore, the vast majority of studies on the impact of higher order aberration correction have been monocular. Using a binocular adaptive optics vision simulator, Sabesan *et al.* (2012) showed that binocular summation decreased when subjects' native higher order aberrations were corrected. Therefore, monocular estimates of visual benefit tend to overestimate in the case of binocular correction.

While individual higher order aberrations vary randomly from patient to patient, the exception is primary spherical aberration (Porter *et al.*, 2001; Thibos *et al.*, 2002), the correction of which has been shown to lead to significant improvements in both visual acuity and contrast sensitivity (Sabesan *et al.*, 2012; Piers *et al.*, 2004, 2007; Schwarz *et al.*, 2014). The primary source of the eye's spherical aberration is the cornea (Artal *et al.*, 2002), and thus intraocular lenses have been a natural platform for its correction (Holladay *et al.*, 2002; Muñoz *et al.*, 2006). However, it is important to note that depth of focus suffers with the correction of spherical aberration (Piers *et al.*, 2004). This is particularly important in regard to intraocular lenses, as cataract surgery leaves patients unable to accommodate.

Chromatic Aberration

Longitudinal chromatic aberration of the human eye is approximately 2 diopters across the visible spectrum (Thibos *et al.*, 1992; Atchison and Smith, 2005). Due to the dispersion of ocular media (i.e., positive Abbe number), the longer wavelengths are relatively hyperopic compared to shorter wavelengths. Because this is caused by the inherent material properties of the eye's anatomy, chromatic aberration is more consistent across the population relative to the monochromatic aberrations (e.g., sphere, cylinder and higher order aberrations). With advances in diffractive intraocular lens technology, there has been a renewed interest in correcting ocular chromatic aberration (Schwarz *et al.*, 2014; López-Gil and Montés-Micó, 2007; Artal *et al.*, 2010). However, as discussed by López-Gil and Bradley, the clinical benefit of correcting spherical and chromatic aberrations may be modest (López-Gil and Bradley, 2012), especially under photopic conditions. In addition, misaligned intraocular lenses providing spherical and chromatic aberration correction lead to unwanted aberrations, such as transverse chromatic aberration, coma and astigmatism (López-Gil and Montés-Micó, 2007; Cheng *et al.*, 2010).

Presbyopia

Past the age of approximately 40 years, even emmetropic eyes require vision aids. As the eye ages, it loses its ability to accommodate (i.e., change focus from far to near objects) due to lifelong crystalline lens growth and loss of elasticity. The age-related loss of accommodation is known as *presbyopia*. As first measured by Donders in 1864 (Donders and Moore, 1864) and later by Duane in 1912 (Duane, 1912), the eye's amplitude of accommodation suffers a steady decline, starting in childhood until a plateau is reached at approximately 50 years (Fig. 5).

Evidence of visual aids can be traced to magnifying lenses known as "reading stones" developed in the 9th century, CE. The English friar Roger Bacon wrote in 1250 CE of such devices to correct for his presbyopia:

"It may be observed that old people hold objects they wish to examine further from the eye. ...If anyone examines letters or other minute objects through the medium of crystal or glass or other transparent substance, if it be shaped like the lesser segment of a sphere, with the convex side being towards the eye, and the eye being in the air, he will see the letters far better, and they will seem larger to him. ...For this reason, such an instrument is useful to old persons..." (James, 1928).

Until the 18th century invention of Benjamin Franklin's bifocal spectacles, little progress was made for correcting presbyopia. However, the last several decades have experienced a tremendous increase in research and development devoted to presbyopia correction. As a result, there are currently a wide variety of strategies available, which are outlined in Table 1.

The field of presbyopia correction can be segmented into two categories: (1) the restoration of the eye's ability to accommodate, i.e., change focal length and (2) increase the eye's depth of focus.

Restoring Accommodation

The majority of accommodation restoration techniques have focused on using eye's lens as the engine for dynamic optical power change. For example, pharmacological agents (Renna *et al.*, 2016; Crawford *et al.*, 2014) and femtosecond laser induced micro-incisions (Holzer *et al.*, 2009; Ripken *et al.*, 2008) are currently in development for restoring elasticity to the presbyopic crystalline lens.

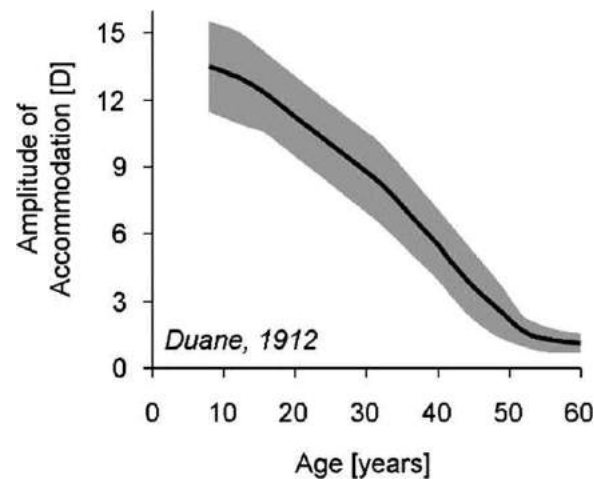


Fig. 5 Amplitude of accommodation as a function of age. Average and upper and lower limits are indicated by black line and grey area, respectively. Figure adapted from Duane, A., 1912. Normal values of the accommodation at all ages. *Journal of the American Medical Association* 59 (12), 1010–1013.

Table 1 Methods for correcting presbyopia

	<i>Method</i> <i>Spectacles</i>	<i>Description</i> <i>Reading glasses</i> <i>Bifocals, trifocals, progressives</i>	<i>Mechanism</i> <i>Single vision</i> <i>Change of gaze</i>
Restoring Dynamic Accommodation (<i>pseudophakic accommodation</i>)	Intraocular Lenses	Accommodating IOL	Positional and shape changes induced by ciliary body
Extending Depth of Focus (<i>pseudoaccommodation</i>)	Contact Lenses	Bifocal, multifocal	Wavefront aberrations
	Corneal Surgery	Tissue removal: LASIK, PRK	Wavefront aberrations
		Implantable refractive corneal inlay	Wavefront aberrations
		Implantable small-aperture inlay	Pinhole effect
	Intraocular Lenses	Multifocal IOL	Wavefront aberrations and/or diffraction
	Monovision	Dominant eye corrected for distance, non-dominant eye corrected for near	Binocular combination of two eyes' retinal images

Otherwise, novel accommodating intraocular lens (IOL) designs have been proposed to restore accommodation (Pepose *et al.*, 2017). Such lenses are designed to change surface curvature or shift axially within the eye during accommodative effort. One such curvature-change design is the Powervision (Belmont, CA) FluidVision IOL, which is comprised of a hollow lens and liquid-filled reservoir haptics. When the eye is in its relaxed state, the liquid resides within the haptics and the lens rests in a relatively flat shape, defining the geometry for viewing distant objects. As the eye exerts accommodative effort, the ciliary muscle contracts, decreasing the equatorial diameter of the capsular bag. This forces the fluid from the haptic reservoir into the central lens cavity, increasing its curvature and optical power, thereby providing near vision (Sheppard *et al.*, 2010).

Alternatively, accommodating IOLs may function on axial displacement, such as the Crystalens (Bausch and Lomb, Rochester, NY). The Crystalens is a single-optic IOL mounted on hinged haptics, allowing the lens to vault anteriorly during accommodative effort, as illustrated in Fig. 6. However, in a theoretical analysis by Hunter *et al.*, it was shown that a 1 mm axial shift of a 20D IOL induces a 1.2D optical change. Therefore, single-optic accommodating IOLs must travel an excessive distance through the anterior chamber to provide a meaningful optical change for presbyopia correction.

Several investigators (Koepl *et al.*, 2005; Marcos *et al.*, 2014; Zheleznyak, 2014) have assessed accommodative IOL movement by measuring changes in anterior chamber depth during accommodative effort. They have found that on average, the single-optic accommodating IOL (Crystalens, Bausch + Lomb) undergoes a posterior shift during accommodative effort by approximately 0.15 mm. Similarly, studies measuring changes in wavefront aberrations during accommodative effort show inconsistent changes in defocus, and in some cases, an incorrect change in the eye's power (Zheleznyak, 2014; Pérez-Merino *et al.*, 2014).

To overcome this limitation, dual-optic IOL designs have been proposed (e.g., Synchrony IOL, Abbott Medical Optics, Santa Clara, CA). In comparison to single-optic designs, the dual-optic accommodating IOL achieves approximately 2.2D increase in power with a 1 mm shift of the anterior lens (McLeod *et al.*, 2007). However, the expected benefit of the dual-optic design over the

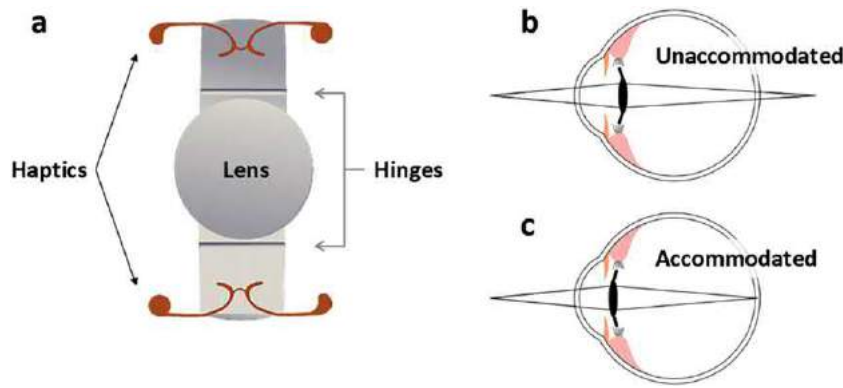


Fig. 6 (a) Single-optic accommodating intraocular lens (Crystalens, Bausch + Lomb, Rochester, NY). Illustration of the intended axial movement of a single-optic accommodating IOL in the (b) unaccommodated and (c) accommodated states of the pseudophakic eye.

single-optic accommodating IOL has not been reflected clinically (Alió *et al.*, 2012) and may be due to capsular bag fibrosis limiting IOL movement (Koepl *et al.*, 2005).

Extending the Eye's Depth of Focus

In the absence of true accommodation restoration, extending the eye's depth of focus has been a common path taken by clinicians to alleviate the symptoms of presbyopia. This can be achieved by reducing pupil size or introducing multifocality to the wavefront aberrations via aspheric refractive surfaces, diffractive optical elements, or both.

The Pinhole Effect

Reducing the pupil size of the eye is the easiest way to lessen the impact of presbyopic blur. For example, eyelid squinting is a common means of truncating the pupil, albeit only vertically, to improve acuity in the presence of refractive error (Sheedy *et al.*, 2003). Pupil size also effects aberration-free image quality at OD of defocus. The cutoff frequency of the modulation transfer function (MTF) is linearly dependent upon the pupil size, as shown below.

$$f_{\text{cutoff}} = \frac{d}{\lambda}$$

Where f_{cutoff} , r and λ represent cutoff frequency (in cycles/radian), pupil diameter and wavelength, respectively. Therefore, as pupil size decreases, so does the cutoff frequency of the MTF and the spatial bandwidth of the image, thereby degrading distance image quality. Therefore, decreasing pupil size for extended depth of focus is a trade-off with distance optical quality. This concept is illustrated below in Fig. 7.

Extended depth of focus via small pupils has been implemented with pinhole corneal inlay (KAMRA, AcuFocus, Irvine, CA). The inlay is implanted within the corneal stroma and has a 1.6 mm inner diameter. Although significant improvements in near vision have been demonstrated (Yilmaz *et al.*, 2011), it may only be implanted monocularly (typically the non-dominant eye), due to the reduction in retinal illuminance.

Multifocal Wavefront Aberrations

In contrast with the pinhole effect, manipulating the wavefront aberrations is a means of extending depth of focus without reducing retinal image brightness. However, wavefront multifocality is also accompanied by the trade-off with distance optical quality. As shown in Fig. 8(a), monofocal wavefronts result in all rays across the pupil converging to a single point, or a single focus. An example of a refractive bifocal with a central add power for near vision is illustrated in Fig. 8(b). In this case, when the blue ray bundle is in focus on the retina, the red ray bundle is out of focus. Thus, the out of focus (red) ray bundle degrades retinal image quality and is commonly referred to as halo or glare. An example of a continuous distribution of refractive powers across the pupil is spherical aberration (Fig. 8(c)). Although spherical aberration is known to increase depth of focus, efforts of mitigating negative side-effects (e.g., degraded distance vision) have included incorporating higher order spherical aberrations (Yi *et al.*, 2011; Benard *et al.*, 2011) and pupil amplitude apodization (Zheleznyak *et al.*, 2014).

A novel concept in multifocal wavefront design segmenting the wavefront angularly into zones of refractive power, also known as a light-sword wavefront (Petelczyc *et al.*, 2009, 2011; de Gracia *et al.*, 2013; Mira-Agudelo *et al.*, 2016). A benefit of this technique is that the relative energy allocated to far and near vision is unaffected by pupil size. Recent theoretical (Petelczyc *et al.*, 2011; de Gracia *et al.*, 2013) and psychophysical (Mira-Agudelo *et al.*, 2016) studies have shown that angular multifocal designs have promise in offering superior through-focus image quality over traditional radial designs (e.g., Fig. 8(b)).

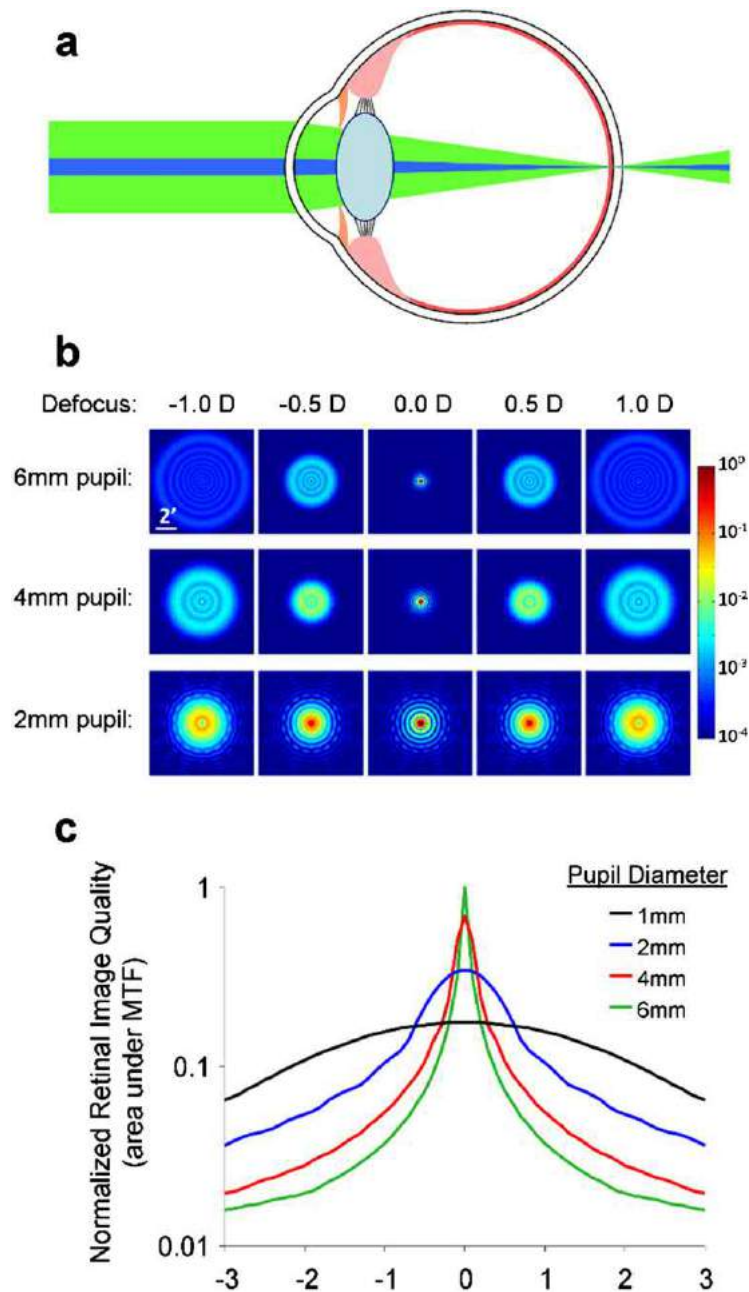


Fig. 7 (a) Schematic eye with ray bundle from a distant object for large and small pupils. Through-focus blur circles are shown for comparison. (b) Through-focus point spread functions with 6, 4 and 2 mm pupil diameters. (c) Through-focus normalized retinal image quality in monochromatic light (550 nm) of an aberration-free eye with various pupil sizes.

In addition to refractive multifocal wavefronts, diffractive multifocals also offer a means of increasing the eye's depth of focus. Typically implemented in intraocular lenses, diffractive multifocals consist of a radially symmetric diffractive structure which separates the light into several orders of diffraction. The phase height (in multiples of the design wavelength, λ) determines the diffraction efficiency, or the amount of energy allocated to far or near image quality. Three diffractive multifocals with various phase heights exhibiting (a) an emphasis on far vision, (b) equal energy to both far and near vision or (c) an emphasis on near vision are shown in Fig. 9. Diffraction efficiency may also be modified by means of apodizing the phase height as a function of pupil radius, as in the case of the ReSTOR (Alcon) apodized multifocal intraocular lens (Davison and Simpson, 2006). Recent advances in diffractive multifocal design have also resulted in trifocal intraocular lenses aimed at improving vision for intermediate object distances (Gatinel *et al.*, 2011).

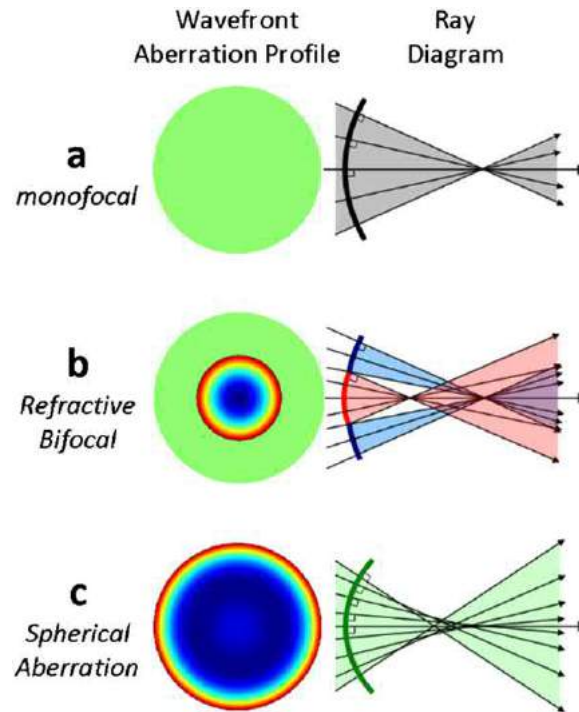


Fig. 8 Wavefront aberration maps and ray diagrams of (a) monofocal, (b) refractive bifocal and (c) spherical aberration wavefronts.

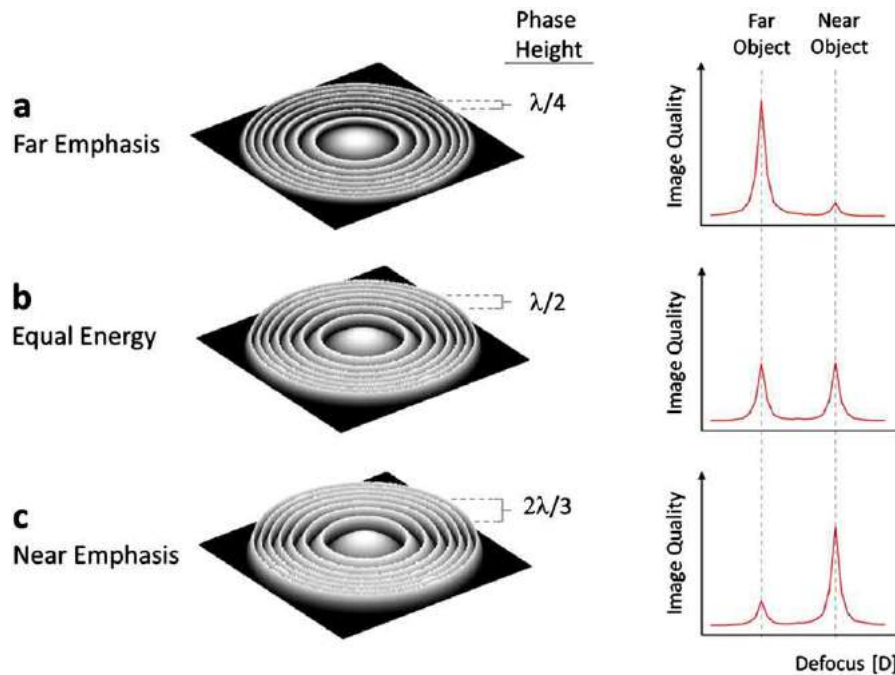


Fig. 9 Diffractive multifocal wavefronts and their associated through-focus image quality are shown for designs with (a) an emphasis on far vision, (b) equal energy to both far and near vision or (c) and emphasis on near vision.

Through-focus image quality (in-vitro) of a monofocal and various multifocal IOLs is shown in [Fig. 10](#). In this case, retinal image quality is defined as the fidelity of a resolution target, which has been shown to correlate well with clinically measured visual acuity in pseudophakic patients ([Plaza-Puche et al., 2015](#)). Optical bench metrology of IOLs is critical to understanding their optical performance. Alternatively, while Shack–Hartmann wavefront sensing measure the wavefront, it does not typically quantify

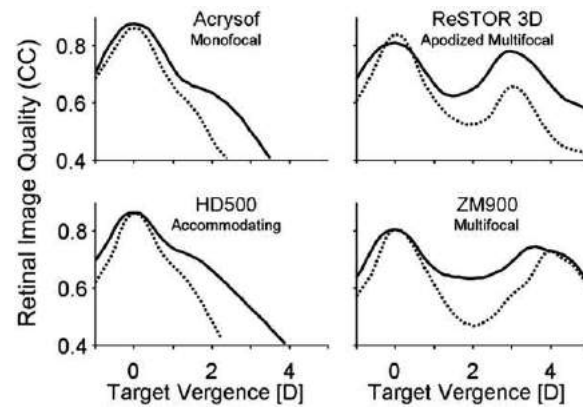


Fig. 10 Through-focus retinal image quality of a monofocal and various diffractive multifocal IOLs. Adapted from Zheleznyak, L., Kim, M.J., MacRae, S., Yoon, G., 2012. Impact of corneal aberrations on through-focus image quality of presbyopia-correcting intraocular lenses using an adaptive optics bench system. *Journal of Cataract & Refractive Surgery* 38 (10), 1724–1733.

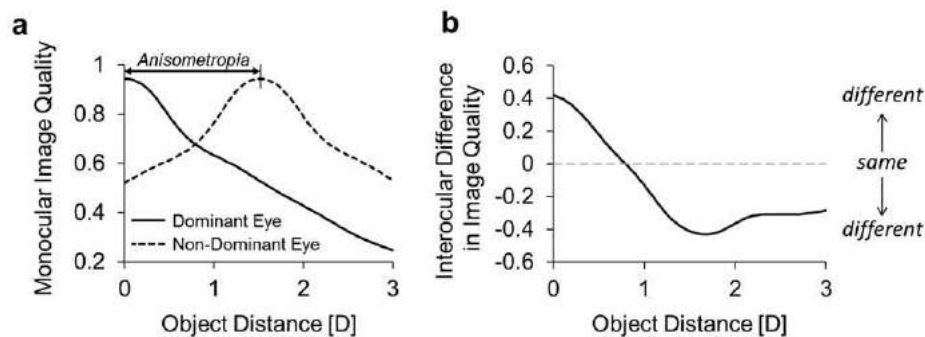


Fig. 11 (a) Through-focus monocular image quality for the dominant and non-dominant eyes in traditional monovision. The magnitude of anisometropia is defined as the dioptric difference in peak monocular image quality. (b) Through-focus interocular difference in monocular image quality.

scatter and its measurement may be confounded by diffractive multifocal designs. By observing a point-spread function or resolution target through a pseudophakic model eye, the eye's optics may be more fully understood. An important side note: optical bench tests are not conclusive in their representation of a patient's perception. A combination of optical and neural factors, such as the Stiles–Crawford effect (Applegate and Lakshminarayanan, 1993), retinal sampling (Hirsch and Curcio, 1989) and cortical processing inform a patient's perception of the visual world.

Binocular Approaches to Presbyopia Correction

Monovision is a paradigm of presbyopia correction which capitalizes on the binocularity of the visual system. In traditional monovision, the dominant eye is corrected for distance vision, whereas the fellow non-dominant eye is corrected for near vision by introducing positive defocus. To illustrate the impact of anisometropia on optical quality of the two eyes, through-focus monocular image quality in the presence of 1.5 diopters (D) of anisometropia is shown in Fig. 11(a). Typical amounts of anisometropia induced clinically in traditional monovision vary from 1D to 2D (Evans, 2007; Jain *et al.*, 1996; Johannsdottir and Stelmach, 2001).

Traditional monovision requires patients to suppress the image from the blurred eye and to rely solely on the eye with superior image quality for high acuity tasks. While this technique does improve presbyopic visual performance at near object distances, binocular visual functions such as binocular summation (Pardhan and Gilchrist, 1990; Legras *et al.*, 2001) and stereopsis (Lovasik and Szymkiw, 1985) are sacrificed. As shown in Fig. 11(b), anisometropia leads to a large difference in interocular image quality. In this example, the largest interocular difference occurred at distance (0D) and intermediate (1.5D) object distances.

To overcome the limitations of traditional monovision, it has been proposed to ease the large difference in interocular image quality by increasing each eye's depth of focus. This technique, also known as modified monovision, may be implemented with spherical aberration (Vandermeer *et al.*, 2015; Zheleznyak *et al.*, 2013, 2015), multifocal intraocular lenses (Ravikumar *et al.*, 2016) or by constricting the eye's entrance pupil diameter, as with a small-aperture corneal inlay (Fernández *et al.*, 2013).

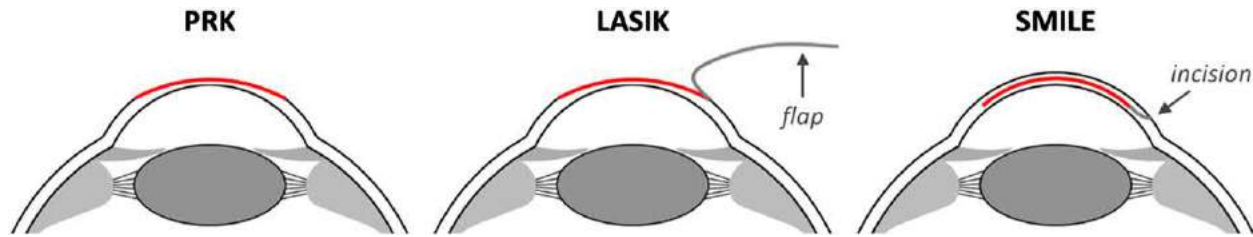


Fig. 12 Illustrations of PRK, LASIK and SMILE laser refractive surgeries. The red segment in each figure indicates the ablation zone.

Refractive Surgery

In 1948, Jose I. Barraquer Moner introduced a surgical procedure to correct the eye's refractive error by sculpting the cornea's topography, a procedure he named *keratomileusis*. Almost 70 years later, laser vision correction is among the most widely performed elective surgical procedures worldwide (Reinstein *et al.*, 2012).

The premise behind refractive surgery is the reshaping of the cornea's topography to control the eye's wavefront aberrations. This can be achieved in a variety of ways, although currently there are three primary techniques: PRK, LASIK and SMILE (illustrated in Fig. 12).

In *photorefractive keratectomy*, or PRK, the cornea's epithelial layer is removed, exposing the underlying stroma. An excimer laser (193 nm wavelength) is subsequently used to ablate undesired tissue, resulting in a new thickness profile of stroma (Seiler *et al.*, 2000). The epithelial layer regrows on the newly sculpted stromal bed within a week of the procedure (Caldwell and Reilly, 2008; Lee *et al.*, 2005).

Alternatively, in *laser-assisted in situ keratomileusis*, or LASIK, the epithelial layer is not removed. Instead, to expose the stromal bed for laser sculpting, a flap is created in the cornea with a microkeratome blade, or more recently, with a femtosecond laser (~ 1040 nm wavelength). The flap is cut such that a section of tissue remains, serving as a hinge for the folding of the flap to expose the underlying stroma. After excimer laser ablation is complete on the stroma (similar to PRK), the flap is folded back into place.

The desire for less-invasive refractive surgery has ushered in a relatively new procedure known as *small incision lenticule extraction*, or SMILE, which avoids both epithelial removal or creating a flap. This is achieved via femtosecond laser used to remove a section of tissue (i.e., the *lenticule*) from within the cornea's stroma through a small incision outside the optical zone.

Due to the invasive nature of these procedures, which inherently sever collagen fibrils in the stromal bed, there may be unintended consequences, such as dry eye (due to nerve damage) and ectasia (due to biomechanical instability) (Randleman *et al.*, 2008).

A common patient complaint after refractive surgery is poor visual quality, especially under low light conditions (e.g., starburst, halos and glare during night driving) (Villa *et al.*, 2007; Hammond *et al.*, 2004). The root of most of these visual symptoms is the unintended induction of higher order aberrations that accompanies the correction of myopia and hyperopia with refractive surgery due to the cornea's biological response and ablation efficiency (Yoon *et al.*, 2005). For example, several studies have shown a significant correlation between the diopters of correction (i.e., myopia or hyperopia) and induction of spherical aberration (Yoon *et al.*, 2005; Marcos *et al.*, 2001), subsequently degrading contrast and visual quality. Decentration of the ablation zone has also been shown to induce unwanted optical aberrations (Mrochen *et al.*, 2001).

The Future of Vision Correction

To address the limitations of ablative refractive surgery, several less invasive procedures are under development, such as corneal collagen crosslinking and femtosecond laser induced refractive index correction.

Corneal collagen crosslinking is a technique in which ultra-violet radiation, when paired with doping agents such as riboflavin, cross-links the corneal collagen, thereby increasing the mechanical strength of the corneal tissue (Snibson, 2010). A challenge of this technique however is accurate doping of the cornea. Because the epithelial cell layer acts as a barrier, the most common application (e.g., the *Dresden protocol*) requires epithelial cell debridement (similar to PRK). While this procedure is invasive, the subsequent radiation is non-ablative and tissue-sparing. Corneal crosslinking is typically used for mechanically stabilizing the progression of keratoconus or ectasia, and research and development are underway to transition this method to refractive error (Kanellopoulos and Asimellis, 2014) and presbyopia (Kanellopoulos and Asimellis, 2015) correction by way of targeted ultra-violet radiation.

Another recent advance in non-invasive vision correction includes the use of femtosecond lasers to alter the refractive index of corneal stroma (Savage *et al.*, 2014; Ding *et al.*, 2008; Xu *et al.*, 2011; Wozniak *et al.*, 2017). This procedure does not require any epithelial cell layer debridement and is non-ablative and tissue-sparing. In this technique, a pulsed laser is focused into the cornea, triggering a multiphoton absorption process which results in a highly localized refractive index change, limited to a voxel of several micrometers. Due to the multiphoton nature of this process, the laser focus is scanned throughout the optical zone of the

treatment. The short pulse duration (~ 100 femtoseconds) and high repetition rate of laser pulses (MHz regime) allows for relatively little energy required to alter the cornea's refractive index, as compared to femtosecond lasers specifically designed for cutting tissue (e.g., flap cutting). By varying laser parameters such as pulse energy and scan speed, the magnitude of refractive index change is highly controlled during scanning. This approach has also been demonstrated in intraocular lens material (Sahler *et al.*, 2016), and contact lens material (Gandara-Montano *et al.*, 2015). In addition, visual performance has been demonstrated with phase plates modified by the laser induced refractive index correction technique (Zheleznyak *et al.*, 2017).

Summary

Methods of vision correction have come a long way, from medieval looking-stones to advances in modern technology, such as surgical implants and laser refractive surgery. Whether it is sphero/cylindrical error, or the more subtle higher order and chromatic aberrations, all eyes suffer from some degree of optical aberrations. Thus, the need for a robust, non-invasive means of providing clear vision persists. While many methods for correction exist, they are each limited in their efficacy for reasons detailed above. Therefore, advances in technology and our fundamental understanding of vision are critical to ushering in the future of vision correction.

See also: Aberrations

References

- Alió, J.L., Plaza-Puche, A.B., Montalbán, R., Ortega, P., 2012. Near visual outcomes with single-optic and dual-optic accommodating intraocular lenses. *Journal of Cataract & Refractive Surgery* 38 (9), 1568–1575.
- Allen, Y.Y., Raasch, T.W., 2005. Design and fabrication of a freeform phase plate for high-order ocular aberration correction. *Applied Optics* 44 (32), 6869–6876.
- Applegate, R.A., Lakshminarayanan, V., 1993. Parametric representation of Stiles–Crawford functions: Normal variation of peak location and directionality. *JOSA A* 10 (7), 1611–1623.
- Artal, P., Berrio, E., Guirao, A., Piers, P., 2002. Contribution of the cornea and internal surfaces to the change of ocular aberrations with age. *JOSA A* 19 (1), 137–143.
- Artal, P., Manzanera, S., Piers, P., Weeber, H., 2010. Visual effect of the combined correction of spherical and longitudinal chromatic aberrations. *Optics Express* 18 (2), 1637–1648.
- Atchison, D.A., Smith, G., 2005. Chromatic dispersions of the ocular media of human eyes. *JOSA A* 22 (1), 29–37.
- Benard, Y., Lopez-Gil, N., Legras, R., 2011. Optimizing the subjective depth-of-focus with combinations of fourth- and sixth-order spherical aberration. *Vision Research* 51 (23), 2471–2477.
- Caldwell, M., Reilly, C., 2008. Effects of topical nepafenac on corneal epithelial healing time and postoperative pain after PRK: A bilateral, prospective, randomized, masked trial. *Journal of Refractive Surgery* 24 (4), 377–382.
- Cheng, X., Bradley, A., Ravikumar, S., Thibos, L.N., 2010. The visual impact of Zernike and Seidel forms of monochromatic aberrations. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 87 (5), 300.
- Chen, M., Sabesan, R., Ahmad, K., Yoon, G., 2007. Correcting anterior corneal aberration and variability of lens movements in keratoconic eyes with back-surface customized soft contact lenses. *Optics Letters* 32 (21), 3203–3205.
- Chen, M., Yoon, G., 2008. Posterior corneal aberrations and their compensation effects on anterior corneal aberrations in keratoconic eyes. *Investigative Ophthalmology & Visual Science* 49 (12), 5645.
- Christina, K., Oliver, F., Rupert, M., *et al.*, 2005. Pilocarpine-induced shift of an accommodating intraocular lens: At-45 Crystalens. *Journal of Cataract & Refractive Surgery* 31 (7), 1290–1297.
- Crawford, K.S., Garner, W.H., Burns, W., 2014. Dioptin™: A novel pharmaceutical formulation for restoration of accommodation in presbyopes. *Investigative Ophthalmology & Visual Science* 55 (13), 3765.
- Davison, J.A., Simpson, M.J., 2006. History and development of the apodized diffractive intraocular lens. *Journal of Cataract & Refractive Surgery* 32 (5), 849–858.
- de Gracia, P., Dorronsoro, C., Marcos, S., 2013. Multiple zone multifocal phase designs. *Optics Letters* 38 (18), 3526–3529.
- Ding, L., Knox, W.H., Bühren, J., Nagy, L.J., Huxlin, K.R., 2008. Intrastissue refractive index shaping (IRIS) of the cornea and lens using a low-pulse-energy femtosecond laser oscillator. *Investigative Ophthalmology & Visual Science* 49 (12), 5332–5339.
- Donders, F.C., Moore, W.D., 1864. *On the Anomalies of Accommodation and Refraction of the Eye: With a Preliminary Essay on Physiological Dioptrics*. vol. 22. London: New Sydenham Society.
- Dorronsoro, C., Barbero, S., Llorente, L., Marcos, S., 2003. On-eye measurement of optical performance of rigid gas permeable contact lenses based on ocular and corneal aberrometry. *Optometry and Vision Science* 80 (2), 115–125.
- Dreyfus, J., 1988. The invention of spectacles and the advent of printing a revised version of paper given to the bibliographical society 20 January 1987. *The Library* 6 (2), 93–106.
- Duane, A., 1912. Normal values of the accommodation at all ages. *Journal of the American Medical Association* 59 (12), 1010–1013.
- Evans, B.J., 2007. Monovision: A review. *Ophthalmic and Physiological Optics* 27 (5), 417–439.
- Fernández, E.J., Schwarz, C., Prieto, P.M., Manzanera, S., Artal, P., 2013. Impact on stereo-acuity of two presbyopia correction approaches: monovision and small aperture inlay. *Biomedical Optics Express* 4 (6), 822–830.
- Gandara-Montano, G.A., Ivansky, A., Savage, D.E., Ellis, J.D., Knox, W.H., 2015. Femtosecond laser writing of freeform gradient index microlenses in hydrogel-based contact lenses. *Optical Materials Express* 5 (10), 2257–2271.
- Gatinel, D., Pagnoulle, C., Houbrechts, Y., Gobin, L., 2011. Design and qualification of a diffractive trifocal optical profile for intraocular lenses. *Journal of Cataract & Refractive Surgery* 37 (11), 2060–2067.
- Guirao, A., Williams, D.R., Cox, I.G., 2001. Effect of rotation and translation on the expected benefit of an ideal method to correct the eye's higher-order aberrations. *Journal of the Optical Society of America A-Optics Image Science and Vision* 18 (5), 1003–1015.
- Gumus, K., Gire, A., Pflugfelder, S.C., 2011. The Impact of the Boston ocular surface prosthesis on wavefront higher-order aberrations. *American Journal of Ophthalmology* 151 (4), 682–690.

- Hammond, S.D., Puri, A.K., Ambati, B.K., 2004. Quality of vision and patient satisfaction after LASIK. *Current Opinion in Ophthalmology* 15 (4), 328–332.
- Hirsch, J., Curcio, C.A., 1989. The spatial resolution capacity of human foveal retina. *Vision Research* 29 (9), 1095–1101.
- Holden, B.A., 1975. The principles and practice of correcting astigmatism with soft contact lenses. *Clinical and Experimental Optometry* 58 (8), 279–299.
- Holladay, J.T., Piers, P.A., Koranyi, G., van der Mooren, M., Norrby, N.E.S., 2002. A new intraocular lens design to reduce spherical aberration of pseudophakic eyes. *Journal of Refractive Surgery* 18 (6), 683–691.
- Holzer, M.P., Mannsfeld, A., Ehmer, A., Auffarth, G.U., 2009. Early outcomes of INTRACOR femtosecond laser treatment for presbyopia. *Journal of Refractive Surgery* 25 (10), 855–861.
- Hong, X., Himebaugh, N., 2001. On-eye evaluation of optical performance of rigid and soft contact lenses. *Optometry & Vision Science* 78 (12), 872.
- Jain, S., Arora, I., Azar, D.T., 1996. Success of monovision in presbyopes: Review of the literature and potential applications to refractive surgery. *Survey of Ophthalmology* 40 (6), 491–499.
- James, R., 1928. The father of british optics: Roger bacon, c. 1214–1294. *The British Journal of Ophthalmology* 12 (1), 1.
- Johannsdottir, K.R., Stelmach, L.B., 2001. Monovision: A review of the scientific literature. *Optometry & Vision Science* 78 (9), 646–651.
- Kanellopoulos, A.J., Asimellis, G., 2014. Hyperopic correction: Clinical validation with epithelium-on and epithelium-off protocols, using variable fluence and topographically customized collagen corneal crosslinking. *Clinical ophthalmology (Auckland, NZ)* 8, 2425.
- Kanellopoulos, A.J., Asimellis, G., 2015. Presbyopic PiXL cross-linking. *Current Ophthalmology Reports* 3 (1), 1–8.
- Hyung, K.L., Kyung, S.L., Jin, K.K., *et al.*, 2005. Epithelial healing and clinical outcomes in excimer laser photorefractive surgery following three epithelial removal techniques: Mechanical, alcohol, and excimer laser. *American Journal of Ophthalmology* 139 (1), 56–63.
- Legras, R., Hornain, V., Monot, A., Chateau, N., 2001. Effect of induced anisometropia on binocular through-focus contrast sensitivity. *Optometry & Vision Science* 78 (7), 503–509.
- Legras, R., Rouger, H., 2008. Calculations and measurements of the visual benefit of correcting the higher-order aberrations using adaptive optics technology. *Journal of Optometry* 1 (1), 22–29.
- Liang, J., Williams, D.R., 1994. Objective measurement of wave aberrations of the human eye with the use of a Hartmann–Shack wave-front sensor. *Journal of the Optical Society of America A, Optics, Image Science, and Vision* 11 (7), 1949–1957.
- López-Gil, N., Bradley, A., 2012. The Potential for and Challenges of Spherical and Chromatic Aberration Correction With New IOL Designs. *BMJ Publishing Group Ltd.*
- López-Gil, N., Montés-Micó, R., 2003. Correcting ocular aberrations by soft contact lenses. *South African Optometric Association* 62, 173–177.
- López-Gil, N., Montés-Micó, R., 2007. New intraocular lens for achromatizing the human eye. *Journal of Cataract & Refractive Surgery* 33 (7), 1296–1302.
- Lovaski, J.V., Szymkiw, M., 1985. Effects of aniseikonia, anisometropia, accommodation, retinal illuminance, and pupil size on stereopsis. *Investigative Ophthalmology & Visual Science* 26 (5), 741–750.
- Maeda, N., Fujikado, T., Kuroda, T., *et al.*, 2002. Wavefront aberrations measured with Hartmann–Shack sensor in patients with keratoconus. *Ophthalmology* 109 (11), 1996–2003.
- Marcos, S., Barbero, S., Llorente, L., Merayo-Llodes, J., 2001. Optical response to LASIK surgery for myopia from total and corneal aberration measurements. *Investigative Ophthalmology & Visual Science* 42 (13), 3349–3356.
- Marcos, S., Ortiz, S., Pérez-Merino, P., *et al.*, 2014. Three-dimensional evaluation of accommodating intraocular lens shift and alignment in vivo. *Ophthalmology* 121 (1), 45–55.
- Marsack, J.D., Parker, K.E., Applegate, R.A., *et al.*, 2008. Performance of wavefront-guided soft lenses in three keratoconus subjects. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 85 (12), E1172.
- Marsack, J.D., Parker, K.E., Pesudovs, K., Donnelly III, W.J., Applegate, R.A., 2007. Uncorrected wavefront error and visual performance during RGP wear in keratoconus. *Optometry & Vision Science* 84 (6), 463.
- Marsack, J.D., Ravikumar, A., Nguyen, C., *et al.*, 2014. Wavefront-guided scleral lens correction in keratoconus. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 91 (10), 1221–1230.
- McLeod, S.D., Vargas, L.G., Portney, V., Ting, A., 2007. Synchrony dual-optic accommodating intraocular lens: part 1: Optical and biomechanical principles and design considerations. *Journal of Cataract & Refractive Surgery* 33 (1), 37–46.
- Mira-Agudelo, A., *et al.*, 2016. Compensation of presbyopia with the light sword lenscompensation of presbyopia with the light sword lens. *Investigative Ophthalmology & Visual Science* 57 (15), 6870–6877.
- Mrochen, M., Kaemmerer, M., Mierdel, P., Seiler, T., 2001. Increased higher-order optical aberrations after laser refractive surgery: A problem of subclinical decentration. *Journal of Cataract & Refractive Surgery* 27 (3), 362–369.
- Muñoz, G., Albarrán-Diego, C., Montés-Micó, R., Rodríguez-Galitero, A., Alió, J.L., 2006. Spherical aberration and contrast sensitivity after cataract surgery with the Tecnis Z9000 intraocular lens. *Journal of Cataract & Refractive Surgery* 32 (8), 1320–1327.
- Navarro, R., Moreno, E., Dorronsoro, C., 2000. Phase plates for wave-aberration compensation in the human eye. *Optics Letters* 25 (4), 236–238.
- Negishi, K., Kumanomido, T., Utsumi, Y., Tsubota, K., 2007. Effect of higher-order aberrations on visual function in keratoconic eyes with a rigid gas permeable contact lens. *American Journal of Ophthalmology* 144 (6), 924–929.
- Pantanelli, S., MacRae, S., Jeong, T.M., Yoon, G., 2007. Characterizing the wave aberration in eyes with keratoconus or penetrating keratoplasty using a high-dynamic range wavefront sensor. *Ophthalmology* 114 (11), 2013–2021.
- Pardhan, S., Gilchrist, J., 1990. The effect of monocular defocus on binocular contrast sensitivity. *Ophthalmic and Physiological Optics* 10 (1), 33–36.
- Pepose, J.S., Burke, J., Qazi, M.A., 2017. Benefits and barriers of accommodating intraocular lenses. *Current Opinion in Ophthalmology* 28 (1), 3–8.
- Pérez-Merino, P., Birkenfeld, J., Dorronsoro, C., *et al.*, 2014. Aberrometry in patients implanted with accommodative intraocular lenses. *American Journal of Ophthalmology* 157 (5), 1077–1089.
- Petelczyc, K., Garcia, J.A., Bará, S., *et al.*, 2009. Presbyopia compensation with a light sword optical element of a variable diameter. *Photonics Letters of Poland* 1 (2), 55–57.
- Petelczyc, K., Ares García, J., Bará, S., *et al.*, 2011. Strehl ratios characterizing optical elements designed for presbyopia compensation. *Optics Express* 19 (9), 8693–8699.
- Piers, P.A., Fernandez, E.J., Manzanera, S., Norrby, S., Artal, P., 2004. Adaptive optics simulation of intraocular lenses with modified spherical aberration. *Investigative Ophthalmology & Visual Science* 45 (12), 4601–4610.
- Piers, P.A., Manzanera, S., Prieto, P.M., Gorceix, N., Artal, P., 2007. Use of adaptive optics to determine the optimal ocular spherical aberration. *Journal of Cataract & Refractive Surgery* 33 (10), 1721–1726.
- Plaza-Puche, A.B., Alió, J.L., MacRae, S., *et al.*, 2015. Correlating optical bench performance with clinical defocus curves in varifocal and trifocal intraocular lenses. *Journal of Refractive Surgery* 31 (5), 300–307.
- Porter, J., Guirao, A., Cox, I.G., Williams, D.R., 2001. Monochromatic aberrations of the human eye in a large population. *JOSA A* 18 (8), 1793–1803.
- Pullum, K.W., Whiting, M.A., Buckley, R.J., 2005. Scleral contact lenses: The expanding role. *Cornea* 24 (3), 269.
- Randleman, J.B., Woodward, M., Lynn, M.J., Stulting, R.D., 2008. Risk assessment for ectasia after corneal refractive surgery. *Ophthalmology* 115 (1), 37–50.
- Ravikumar, S., Bradley, A., Bharadwaj, S., Thibos, L.N., 2016. Expanding binocular depth of focus by combining monovision with diffractive bifocal intraocular lenses. *Journal of Cataract & Refractive Surgery* 42 (9), 1288–1296.
- Reinstein, D.Z., Archer, T.J., Gobbe, M., 2012. The history of LASIK. *Journal of Refractive Surgery* 28 (4), 291–298.
- Renna, A., Vejarano, L.F., Cruz, E., Alió, J.L., 2016. Pharmacological treatment of presbyopia by novel binocularly instilled eye drops: A pilot study. *Ophthalmology and Therapy* 1–11.

- Ripken, T., Oberheide, U., Fromm, M., 2008. fs-Laser induced elasticity changes to improve presbyopic lens accommodation. *Graefes Archive for Clinical and Experimental Ophthalmology* 246 (6), 897–906.
- Rosenthal, P., 2010. Scleral lens with scalloped channels or circumferential fenestrated channels. US Patent 7,695,135.
- Rosenthal, P., 2009. Scleral contact lens with grooves and method of making lens. US Patent 7,591,556.
- Rossi, E.A., Weiser, P., Tarrant, J., Roorda, A., 2007. Visual performance in emmetropia and low myopia after correction of high-order aberrations. *Journal of Vision* 7 (8), 14.
- Sabesan, R., Ahmad, K., Yoon, G., 2007. Correcting highly aberrated eyes using large-stroke adaptive optics. *Journal of Refractive Surgery* 23 (9), 947–952.
- Sabesan, R., Johns, L., Tomashevskaya, O., *et al.*, 2013. Wavefront-guided scleral lens prosthetic device for keratoconus. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 90 (4), 314.
- Sabesan, R., Moon Jeong, T., Carvalho, L., 2007. Vision improvement by correcting higher-order aberrations with customized soft contact lenses in keratoconic eyes. *Optics Letters* 32 (8), 1000–1002.
- Sabesan, R., Zheleznyak, L., Yoon, G., 2012. Binocular visual performance and summation after correcting higher order aberrations. *Biomedical Optics Express* 3 (12), 3176–3189.
- Sahler, R., Bille, J.F., Enright, S., Chhoeung, S., Chan, K., 2016. Creation of a refractive lens within an existing intraocular lens using a femtosecond laser. *Journal of Cataract & Refractive Surgery* 42 (8), 1207–1215.
- Savage, D.E., Brooks, D.R., DeMagistris, M., *et al.*, 2014. First demonstration of ocular refractive change using blue-IRIS in live catsrefractive changes using blue-IRIS. *Investigative Ophthalmology & Visual Science* 55 (7), 4603–4612.
- Schwarz, C., Cánovas, C., Manzanera, S., *et al.*, 2014. Binocular visual acuity for the correction of spherical aberration in polychromatic and monochromatic light. *Journal of Vision* 14 (2), 8.
- Segal, O., Barkana, Y., Hourovitz, D., *et al.*, 2003. Scleral contact lenses may help where other modalities fail. *Cornea* 22 (4), 308.
- Seiler, T., Kaemmerer, M., Mierdel, P., Krinke, H.-E., 2000. Ocular optical aberrations after photorefractive keratectomy for myopia and myopic astigmatism. *Archives of Ophthalmology* 118 (1), 17–21.
- Sheedy, J.E., Truong, S.D., Hayes, J.R., 2003. What are the visual benefits of eyelid squinting? *Optometry & Vision Science* 80 (11), 740–744.
- Sheppard, A.L., Bashir, A., Wolffsohn, J.S., Davies, L.N., 2010. Accommodating intraocular lenses: A review of design concepts, usage and assessment methods. *Clinical and Experimental Optometry* 93 (6), 441–452.
- Snibson, G.R., 2010. Collagen cross-linking: A new treatment paradigm in corneal disease—a review. *Clinical & Experimental Ophthalmology* 38 (2), 141–153.
- Stason W.B., Razavi M., Jacobs D.S., *et al.*, 2010. Clinical benefits of the boston ocular surface prosthesis. *American Journal of Ophthalmology* 149 (1), 54–61.
- Thibos, L.N., Hong, X., Bradley, A., Cheng, X., 2002. Statistical variation of aberration structure and image quality in a normal population of healthy eyes. *JOSA A* 19 (12), 2329–2348.
- Thibos, L.N., Ye, M., Zhang, X., Bradley, A., 1992. The chromatic eye: A new reduced-eye model of ocular chromatic aberration in humans. *Applied Optics* 31 (19), 3594–3600.
- Vandermeer, G., Rio, D., Gicquel, J.-J., Pisella, P.-J., Legras, R., 2015. Subjective through-focus quality of vision with various versions of modified monovision. *British Journal of Ophthalmology* 99 (7), 997–1003.
- Villa, C., Gutiérrez, R., Jiménez, J.R., Gonzalez-Méijome, J.M., 2007. Night vision disturbances after successful LASIK surgery. *British Journal of Ophthalmology* 91 (8), 1031–1037.
- Wichterle, O., Lim, D., 1960. Hydrophilic gels for biological use. *Nature* 185 (4706), 117–118.
- Williams, D., Yoon, G.Y., Porter, J., *et al.*, 2000. Visual benefit of correcting higher order aberrations of the eye. *Journal of Refractive Surgery* 16 (5), S554–S559.
- Wozniak, K.T., Gearhart, S.M., Savage, D.E., *et al.*, 2017. Comparable change in stromal refractive index of cat and human corneas following blue-IRIS. *Journal of Biomedical Optics* 22 (5). doi:10.1117/1.JBO.22.5.055007.
- Xu, L., Knox, W.H., DeMagistris, M., Wang, N., Huxlin, K.R., 2011. Noninvasive intratissue refractive index shaping (IRIS) of the cornea with blue femtosecond laser light. *Investigative Ophthalmology & Visual Science* 52 (11), 8148–8155.
- Yi, F., Iskander, D.R., Collins, M., 2011. Depth of focus and visual acuity with primary and secondary spherical aberration. *Vision Research* 51 (14), 1648–1658.
- Yilmaz, O.F., Alagöz, N., Pekel, G., *et al.*, 2011. Intracorneal inlay to correct presbyopia: Long-term results. *Journal of Cataract & Refractive Surgery* 37 (7), 1275–1281.
- Yoon, G., Jeong, T.M., Cox, I.G., Williams, D.R., 2004. Vision improvement by correcting higher-order aberrations with phase plates in normal eyes. *Journal of Refractive Surgery* 20 (5), S523–S527.
- Yoon, G., MacRae, S., Williams, D.R., Cox, I.G., 2005. Causes of spherical aberration induced by laser refractive surgery. *Journal of Cataract & Refractive Surgery* 31 (1), 127–135.
- Yoon, G.-Y., Williams, D.R., 2002. Visual performance after correcting the monochromatic and chromatic aberrations of the eye. *JOSA A* 19 (2), 266–275.
- Zheleznyak, L., Alarcon, A., Dieter, K.C., Tadin, D., Yoon, G., 2015. The role of sensory ocular dominance on through-focus visual performance in monovision presbyopia corrections. *Journal of Vision* 15 (17), 1–12.
- Zheleznyak, L., Gandara-Montano, G., MacRae, S., *et al.*, 2017. First demonstration of human visual performance through refractive-index-modified ophthalmic devices written in hydrogels [ARVO E-ABSTRACT 1274]. *Investigative Ophthalmology & Visual Science* 58, 1274.
- Zheleznyak, L., Sabesan, R., Oh, J.-S., MacRae, S., Yoon, G., 2013. Modified monovision with spherical aberration to improve presbyopic through-focus visual performance. *Investigative Ophthalmology & Visual Science* 54 (5), 3157–3165.
- Zheleznyak, L.A., 2014. Overcoming Presbyopia by Manipulating the Eyes' Optics. University of Rochester.
- Zheleznyak, L., Jung, H., Yoon, G., 2014. Impact of pupil transmission apodization on presbyopic through-focus visual performance with spherical aberration. *Investigative Ophthalmology & Visual Science* 55 (1), 70–77.

Laser and Light in Ophthalmology

Michael Mrochen, IROC Science AG, Zurich, Switzerland and Swiss Eye Research Foundation, Reinach, Switzerland

Nicole Lemanski, Mabel Cheng MD PLLC, Albany, NY, United States

Bojan Pajic, Orasis Ag, Reinach, Switzerland and Swiss Eye Research Foundation, Reinach, Switzerland

© 2018 Elsevier Ltd. All rights reserved.

The human eye is a complex transparent optical system that receives light in the electromagnetic spectrum. The key structures of the eye include the pupil, iris, lens, cornea, conjunctiva, sclera, conjunctiva, fovea, choroid, optic nerve, optic disc, ciliary body, macula and retina, the latter of which contain photoreceptors responsible for light perception and vision including color differentiation and the perception of depth. The eye can be divided into three layers: the outermost fibrous layer comprised of the cornea and sclera; the middle layer, which consists of the choroid, ciliary body, pigmented epithelium and the iris; and, the retina which represents the innermost layer. A transparent, watery fluid, called aqueous humour occupies the anterior and posterior chambers, i.e., between the cornea and the iris and between the iris and the lens, respectively, while vitreous humour, with a more jelly like consistency, fills the posterior cavity behind the lens. The ciliary body connects the iris to the choroid and is responsible for focusing the crystalline lens and secreting aqueous humour. The lens is attached to ciliary body by the Zonule of Zinn.

In order for the eye to operate correctly, each structure of the eye must function correctly. Given the complexity of the eye it is perhaps not surprising that it may be affected by numerous and systemic disorders, ranging from conjunctivitis and dry eye syndrome to age-related disease including macular degeneration and cataracts. Additionally, refractive disorders including hyperopia, myopia, astigmatism and presbyopia are thought to affect more than two billion of the world population; the World Health Authority estimates that 153 million people worldwide have uncorrected refractive errors, excluding presbyopia ([World Health Organization, 2009](#)).

Light and Laser Therapy

Lasers and light technologies have become important diagnostic and therapeutic tools in ophthalmology. They can be used for imaging the eye and can also be used to treat eye conditions, including glaucoma and retinal disease as well as refractive conditions involving the cornea and lens.

Laser and light treatments used in ophthalmology have photochemical, photothermal and photomechanical mechanisms. The Interactions of light with biological tissues depend on its wavelength, pulse duration and irradiance ([Fig. 1](#)). However, all are caused primarily by the interaction of photons with the molecules and the molecular compounds of the tissue.

Reducing the illumination time (micro- to millisecond range) and increasing the intensities the tissue temperature can be increased and thermal processes might occur. As temperature rises, the cellular and tissue structural proteins undergo denaturation and conformational changes a process defined as thermal coagulation. Increasing the temperature up to 100°C or more, may cause

Laser-tissue-interaction processes

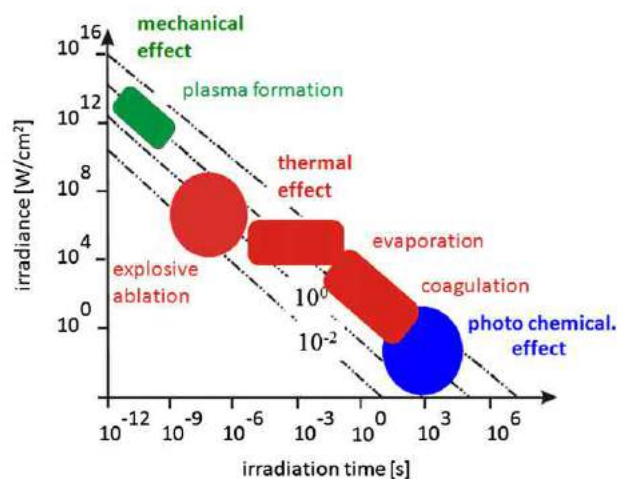


Fig. 1 The interaction between laser and biological tissue.

vaporization of the liquid within the tissue. Further reduction of the illumination time might finally lead to an explosive tissue removal (photoablation).

Photochemical Interaction

The primary interaction mechanism between tissue and laser light are mainly depending on the wavelength, irradiation time (pulse duration) and the intensity or energy dose. For long irradiation times (seconds to minutes) and low intensities one can observe mainly photochemical processes. Photochemical interactions are the basis for cross linking of corneal collagen (CXL).

Therapeutic photochemical interactions are typically performed at low irradiances ($<1 \text{ W/cm}^2$) for long durations, i.e., tens of seconds, to avoid tissue heating. During photochemical interactions, a photon of a specific wavelength is absorbed by a photosensitizer which transforms the molecule (e.g., porphyrin, riboflavin) into the excited singlet state. The molecule decays back to the ground state, or it can be converted into the excited triplet state which in turn is able to transfer energy to a triplet ground state of oxygen. This energy transfer results in highly toxic singlet oxygen. This toxic singlet oxygen in turn produces irreversible oxidation of the cell structures in the areas to which the photosensitizer has been concentrated (Palanker, 2014).

As noted earlier, CXL is an example of a photochemical reaction. It involves four important factors: energy dose of the applied ultraviolet (UV) light, concentration or formulation of riboflavin, oxygen concentration within the tissue, timing of riboflavin diffusion, oxygen diffusion, and light exposure. Riboflavin acts as a photosensitizer that is activated by the absorbed UV light to create reactive oxygen species (ROS) which help to produce bonds between molecules in the stroma, thus increasing the biomechanical strength of the cornea.

The Dresden Protocol uses a 0.1%–0.15% aqueous solution of riboflavin-5-phosphate (vitamin B2) that provides a high absorption and, therefore, good shielding-effect at a wavelength of 360–370 nm. The typically used radiant exposure (energy dose) of 5.4 J/cm^2 allows for biomechanical stiffening of the cornea following treatment.

CXL is an established treatment for keratoconus and post-LASIK ectasia with numerous clinical studies showing that it helps to slow disease progression, regularize the cornea and improve visual acuity in patients with these disorders. Other indications include pellucid marginal degeneration, iatrogenic keratectasia after refractive laser surgery and corneal melting that does not respond to conventional therapy. More recently, CXL has also been investigated as a treatment for refractive errors including myopia, hyperopia and astigmatism (Kanellopoulos, 2014) and for infectious keratitis.

It is known that upon photochemical activation, oxygen species are released from each of the carboxylic groups present in riboflavin, leading to the generation of a light activated riboflavin and single reactive oxygens in solution ($\text{O}_2^{\cdot-}$). As there are two by-products of the photochemical activation, there are two mechanisms available to interact with the collagen structure. In the direct mechanism (Type I), the light activated riboflavin interacts directly with the collagen molecule by hydrogen abstraction of amines to stabilize its own carbon double bonds therefore enabling intermolecular covalent crosslinks to form in the collagen. In the indirect mechanism (Type II), single reactive oxygens form hydrogen peroxide (H_2O_2) and free radicals (e.g., superoxide anion $\text{O}_2^{\cdot-}$) which in turn oxidize the collagen molecule and therefore enable the formation of indirect covalent crosslinks. Most cross-linking occurs through Type I photochemical reactions (Lichtman and Conchello, 2005; Rich *et al.*, 2014; Balasubramanian *et al.*, 1990).

It seems to be of relevance that when an anaerobic solution of riboflavin is exposed to blue or ultraviolet light in the presence of an electron donor, the yellow color of the solution is reduced in intensity. The spectrum of the resulting solution is identical to that of chemically reduced riboflavin and aeration of the solution results in a complete return of the color such that the spectrum of the original and aerated solutions. The photoreaction of riboflavin in the presence of an electron donor would appear to be a photoreduction. In aqueous media, photoreduction in the presence of an electron donor is considered to be more efficient than photobleaching (Carr, 1960; Smith, 1963; Cairns, 1970).

As a result of Type I and Type II reactions, are two reaction mechanisms competing for oxygen in the cornea: excited riboflavin and reduced-state riboflavin. Consequently oxygen is rapidly consumed in the cornea. Unfortunately, reduced-state riboflavin is not reactive and does not act as a photosensitizer. Only new oxygen molecules that diffuse into cornea from atmosphere can lead to further crosslinking activities. The depletion of oxygen during CXL suggests a combined Type I and Type II mechanism after several minutes as oxygen begins to replenish itself. Of note, Herekar (2007) reported on the depletion of oxygen during corneal cross linking and found a significant reduction of oxygen available for Type II reactions within the first 3–5 min at an intensity of 3 mW/cm^2 . At an intensity of 16 mW/cm^2 , the depletion occurred in less than 1 min. Thus, during CXL, both anaerobic and aerobic conditions are created in the cornea. Anaerobic conditions occur where reduced state riboflavin is formed and all oxygen is consumed. Aerobic conditions occur when oxygen from the air diffuses into the cornea. It is reasonable to conclude then that it is the oxygen diffusion time might be primary driver for the depth and effectiveness of CXL. Specifically, reducing the diffusion time will create a thinner aerobic layer, producing a CXL effect only at anterior part of cornea. Higher diffusion of oxygen will cause deeper CXL effects and higher efficacy (Fig. 2).

Photothermal Interactions

Photothermal interactions are the basis for treatments such as panretinal photocoagulation (PRP). Simply speaking, when tissue absorbs laser energy, a photothermal interaction occurs. Tissue temperature changes both during and after the laser irradiation are crucial. Depending on the temperature change and duration of exposure, different effects may occur, i.e., necrosis, coagulation and

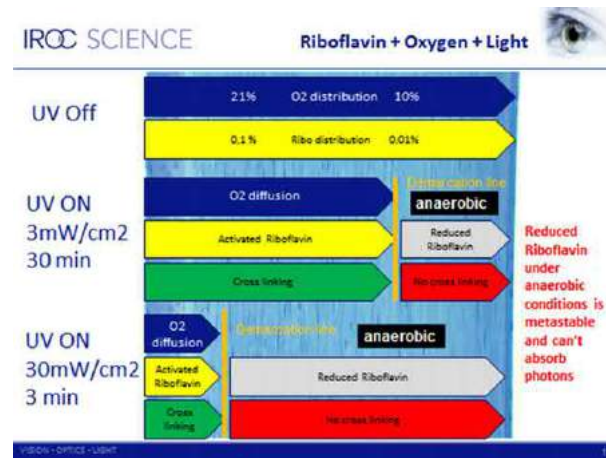


Fig. 2 The photochemical relationship between oxygen, riboflavin and UVA light.

vaporization. Daniel Palanker Department of Ophthalmology and Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA, notes that heat diffusion during the laser pulse becomes significant when the pulse duration exceeds the time it takes for heat to spread over the distance equal to the laser penetration depth in tissue. For pulses shorter than the characteristic diffusion time across the light absorption zone, the heat cannot escape from the energy deposition zone during the pulse; this results in the most efficient heating (Palanker, 2014).

Panretinal photocoagulation was first evaluated in the 1970s using both xenon arc laser and argon laser treatments on patients with proliferative diabetic retinopathy (PDR). The study found that while both laser types were similarly effective, the argon laser was associated with fewer adverse effects.

Panretinal photocoagulation for the treatment of PDR typically uses visible light to create a thermal burn in the retinal tissue. When the retinal pigment epithelium (RPE) which overlays the choroid, just below the retina) absorbs laser light, localized coagulation necrosis occurs, i.e., the laser coagulates the RPE and adjacent photoreceptors, but the retina remains intact. As a result, PRP reduces the area of ischemic tissue, which in turn reduces total vascular endothelial growth factor (VEGF) production in the eye. This in turn helps to reduce neovascularization (American Academy of Ophthalmology).

Ischemic diseases, namely diabetic retinopathy, retinal vein occlusions and sickle cell retinopathy, are the main therapeutic targets for PRP. Indications for focal photocoagulation treatments include retinal breaks, ocular tumors, extrafoveal choroidal neovascular membrane, and microaneurysms.

The overall goal of panretinal photocoagulation is to reduce the oxygen consumption of the outer retina and re-establish a balance between the oxygen supply and demand. PRP typically employs or PRP, typically yellow, green, or red laser light to create micro burns. Specifically, absorbed laser energy is converted into thermal energy, raising the tissue temperature to 20°C–30°C. Tissue burns denature tissue proteins leading to local cell death and coagulative necrosis. An average PRP treatment administers 1200–1500 burns across an area 0.5 mm diameter. This number of burns/area may help to reduce the number of photoreceptors and oxygen consumption of the outer retina by around 20%.

There are several PRP protocols including long duration and short duration protocols. A short duration protocol uses pattern scanning that delivers a pattern of multiple burns in roughly the same time it takes for a conventional laser to deliver one burn. Consequently, the pulse duration is approximately 10–30 ms per spot and the number of spots administered is approximately 3000–5000. As the name suggests, the long duration protocol is more time consuming because the burns are delivered individually. The pulse duration is usually 100 ms with a large spot size (200–500 μ). A power of 200–250 mW is typically applied (American Academy of Ophthalmology).

Photomechanical Interactions

Photomechanical interactions (photodisruption) form the basis for precise tissue cutting, useful in refractive and cataract surgery. Photodisruption occurs when absorption of the laser intensity results in an optical break down due to the high electric field strength beyond the binding energy of molecules. The created electron plasma requires more space leading to fast expansion of the plasma. Expanding and collapsing cavitation bubbles can rupture nearby tissue or eject tissue fragments from the exposed surface. To avoid heat diffusing away from the laser absorption site, energy should be delivered within the thermal confinement time accounting for the use of relatively short pulse durations, i.e., in femto- to picosecond range (Palanker, 2014).

The Small Incision Lenticule Extraction (SMILE) procedure, which is a minimally invasive procedure, designed to correct higher degrees of myopia is an example of photomechanical laser application. Specifically, a femtosecond laser is used to create a 3 mm corneal micro-tunnel through which the lenticule is extracted. SMILE received FDA approval in 2016.

SMILE is indicated for myopia (-1.0 D to -8 D) – and may be particularly useful for myopic patients with dry eye because there is less disruption of corneal surface, and in turn, the tear film, as compared with LASIK. Patients should also have ≤ -0.50 D cylinder and MRSE -8.25 D in the affected eye. SMILE is unsuitable for patients with astigmatism.

The SMILE procedure employs a femtosecond laser. During the procedure, the individually calibrated curved glass of the femtosecond laser is placed against the cornea. Formation of a meniscus tear film allows the patient to see the fixation target because the vergence of the fixation beam is focused according to the patient's refraction. The patient looks at the fixation target and corneal suction ports are activated to fixate the eye. The lower interface of the intrastromal lenticule is created first, followed by the upper interface of the lenticule (which is known as a cap). The laser is then used to create a 2–3 mm tunnel incision that links the cap interface to the corneal surface (Reinstein *et al.*, 2014).

An applanation-free femtosecond laser has also been proposed by Pajic and colleagues and is designed to circumvent the problems associated with weak and strong applanation, respectively. In the case of weak applanation on one hand, an adaption piece of almost the same curvature as the cornea is brought onto the eye. In the case of strong applanation on the other hand, a painful applanation is generated by pressing the eye down with a glass plate. In an applanation-free procedure, physiological saline with a refractive index ($n=1.338$) close to that of the cornea (regularly $n=1.376$) is brought on top of the eye into the suction ring. This liquid ensures a very good refractive index matching and thus minimizes the wave front distortion of the fs-laser pulses upon entrance into cornea. The thickness of the liquid can be varied from a few hundred μm to 1 mm. In this way, an applanation-free processing of the cornea becomes possible (Pajic and Hafezi, 2011).

Laser and Light Interaction and Ocular Target Tissue

Cornea

Laser and light technology play an important role in the treatment of corneal conditions, such as keratoconus. Keratoconus is a bilateral disorder characterized by steepening and thinning of the central cornea, which leads to loss of visual acuity. In early stages keratoconus is treated with contact lenses; however, traditionally, when patients become contact lens-intolerant, keratoplasty, or corneal transplant surgery is performed. Advances in lasers and light, decreased the need for keratoplasty. For example, CXL helps to increase the biomechanical strength of the cornea, with the goal of avoiding corneal transplant.

The procedure utilizes a 0.1%–0.15% aqueous solution of riboflavin-5-phosphate (vitamin B2) and UVA light at a wavelength of 360–370 nm (intensity 3 mW/cm²), usually for 30 min. This is known as the Dresden Protocol. The typically used radiant exposure (energy dose) of 5.4 J/cm² allows for biomechanical stiffening of the cornea following treatment. Newer accelerated protocols with shorter treatment times are currently under development.

Findings from a meta-analysis of randomized controlled trials that assessed the effect of CXL in slowing progression of keratoconus show that CXL significantly decreased mean keratometry, maximum keratometry and minimum keratometry. There were also significant improvements in best spectacle-corrected visual acuity (BCVA). Manifest cylinder error decreased significantly in patients undergoing CXL procedure compared with control patients in sensitivity analysis. Thus, CXL appears to be an effective option in stabilizing the cornea in keratoconus (Li *et al.*, 2015; Cummings *et al.*, 2016; Choi *et al.*, 2017). Some authors found that the accelerated CXL protocol with higher UV intensity and reduced irradiation time showed a smaller topographic flattening effect when compared to the standard Dresden protocol. Sadoughi *et al.* (2016) also compared the Dresden protocol with accelerated CXL (9 mW/cm² for 10 min) in patients with progressive keratoconus. Like Cummings *et al.*'s study, they found that accelerated CXL had similar refractive, visual, keratometric, and aberrometric results and significant less adverse effects on the corneal thickness and endothelial cells as compared with the conventional method after 12 months follow-up.

There is a large body of published studies describing outcomes in the treatment of keratoconus. Collagen corneal cross-linking was first pre-clinically evaluated as a conservative treatment for keratectasia in 1998 by Spoerl *et al.* (1998). Enucleated porcine eyes were treated with UV-light ($\lambda=254$ nm), 0.5% riboflavin, 0.5% riboflavin and UV-light (365 nm) blue light (436 nm) and sunlight, and the chemical agents, glutaraldehyde (1% and 0.1%, 10 min) and Karnovsky's solution (0.1%, 10 min), and were then compared with untreated corneas. They found that compared to untreated corneas treatment with riboflavin and UV-irradiation as well as weak glutaraldehyde or Karnovsky's solutions significantly increased corneal stiffness (Spoerl *et al.*, 1998). This early in vitro study paved the way for Wollensak *et al.*'s prospective, nonrandomized clinical pilot study in 23 eyes with moderate or advanced progressive keratoconus. Data from this study showed that treatment with riboflavin and ultraviolet-A (UVA) irradiation (370 nm, 3 mW/cm² for 30 min) halted the progression of keratoconus in all eyes and caused disease regression in 16 eyes (70%) (Wollensak *et al.*, 2003).

The initial clinical results have been confirmed and reproduced during the past 15 years in multiple clinical studies and have lead to regulatory approvals around the world including FDA approval of the Dresden protocol in 2016. Overall, it appears that CXL's main benefit is in halting, or at least slowing, disease progression. However, it may also help to improve visual acuity. A systemic literature review and meta-analysis of ocular functional and structural parameters of patients with keratoconus undergoing cross-linking procedures found an improvement in visual acuity of 1–2 Snellen lines 3 months or more after undergoing CXL; changes were more pronounced in uncorrected visual acuity. Some topography parameters were found to be improved (0.6–1 diopters) 12–24 months after CXL (Meiri *et al.*, 2016). According to a study by Greenstein and Hersh, outcomes following CXL may be predicted by preoperative patient characteristics. Results of the study, which comprised 104 eyes (66 keratoconus;

38 corneal ectasia), showed that patients with worse preoperative CDVA and higher K values, (particularly with a CDVA of 20/40 or worse or a maximum K of 55.0 D or more), were most likely to have improvement after CXL. However, the authors noted that no preoperative characteristics were predictive of CXL failure (Greenstein and Hersh, 2013).

Lens

Lasers play a key role in modern-day cataract surgery. For many cataract surgeons, femtosecond lasers have replaced hand-held tools for creating the corneal incision and anterior capsulotomy as well as lens and cataract fragmentation. It is also possible to correct astigmatism through corresponding arcuate incisions. It has been suggested that use of femtosecond laser may help to improve safety and precision and reduce postoperative complications versus traditional cataract surgery. A study by Abell and colleagues of outcomes and safety in more than 4000 cases at a single center, however, showed that the two cataract surgery techniques appear to be equally safe. The authors reported that the safety of femtosecond-assisted cataract surgery (FLACS) in terms of posterior capsule complications was equal to that of phacoemulsification, and that anterior capsule tears was a concern in both groups (Abell *et al.*, 2015). A meta analysis of 14,567 eyes that also compared FLACS to manual cataract surgery similarly found that there were no statistically-significant differences between groups in terms of patient-important visual and refractive outcomes and overall complications (Popovic *et al.*, 2016).

Lasers are often also use post-cataract surgery. The neodymium: yttrium-aluminum-garnet (Nd: YAG) laser is used to perform postoperative opacification, and has largely replaced surgical cutting or polishing of the posterior capsule in the management of opacification. The Nd: YAG laser is a solid-state laser with a wavelength of 1064 nm and has a photomechanical effect (Steinert, 2013).

Vitreoretinal

Light is also employed in the treatment of vitreoretinal disorders. Retinal laser photocoagulation is used to treat retinal tears (laser retinopexy) in order to inhibit progression to a retinal detachment. For this indication, green light is the wavelength typically because it is well absorbed by melanin and hemoglobin but is poorly absorbed by xanthophylls thus limiting the potential of macular damage. Red wavelengths are also used and are particularly helpful in patients with media opacity such as cataracts. However, since red light has a longer wavelength than green light, there may be more unpredictable and more heterogeneous absorption by the choroid; this in turn may increase pain and may also lead to focal damage to Bruch's membrane (Silva and Blumenkranz, 2013).

Panretinal photocoagulation is used to treat diabetic retinopathy; it may also be used to retinal ischemia and retinal neo-vascularization of avascular areas. A number of lasers have been used in panretinal photocoagulation including ruby lasers (694-nm wavelength), argon lasers which can use the blue (488-nm wavelength) and green (514-nm wavelength) light emission, and dye lasers. In the 1990s diode lasers (810-nm wavelength) and micropulsed diode lasers, and later, transpupillary thermotherapy (TTT) using a near-infrared (810-nm wavelength) laser were introduced. Newer developments include pattern scan laser photocoagulator with a 532-nm laser and 532-nm pattern-type eye-tracking lasers (Kozak and Luttrull, 2015).

There are a number of landmark studies describing PRP outcomes in diabetic retinopathy. These include the Diabetes Control and Complications Trial, the Diabetic Retinopathy Study, the Early Treatment Diabetic Retinopathy Study (ETDRS) and the Diabetic Retinopathy Vitrectomy Study Research Group trial (Royle *et al.*, 2015).

These studies highlighted a number of points. The randomized, controlled DRS trial, undertaken by the National Eye Institute in the 1970s, compared the efficacy of PRP (versus no treatment) in delaying the progression of PDR and preventing vision loss. The study included 1700 patients with non-proliferative PDR in both eyes or PDR in at least one eye. Patients received PRP treatment in one eye, and no treatment in the fellow eye. Patients were randomly assigned to receive either xenon arc treatment or argon treatment. Findings showed that PRP reduced the risk of severe visual loss by more than 50%. The vision loss rate was 16.3% in untreated eyes versus 6.4% in treated eyes over two years. The study also showed that argon laser was more effective than the xenon arc laser. Importantly, the DRS study also showed that at 2 years, 11% of treated eyes versus 26% of untreated eyes with high-risk retinopathy developed severe visual loss. At 4 years, 20% of treated eyes and 44% of untreated eyes developed severe visual loss (> 5/200). In patients with early proliferative retinopathy – 3% of treated eyes and 7% of untreated eyes developed severe visual loss at 2 years. At 4 years 7% of treated eyes and 21% of untreated eyes developed severe visual loss. The ETDRS group subsequently examined whether early PRP was more effective than delayed PRP treatment. The ETDRS study group concluded that scatter photocoagulation should not be offered in mild cases of PDR as the potential adverse effects and relative risks of treatment may outweigh minimal benefits gained by treatment at early stages. They advised that PRP treatment should commence when non-proliferative PDR becomes severe or progresses into PDR (American Academy of Ophthalmology; Royle *et al.*, 2015).

Lasers are also being increasingly used to reduce symptomatic vitreous floaters. Vitreous floaters are deposits of various size, shape, consistency, refractive index, and motility within the eye's vitreous humour, which is normally transparent. Previously, neodymium–yttrium-aluminum-garnet (Nd: YAG) were employed which was associated with mixed success. More recently, modified YAG lasers have been used. The Ultra Q Reflex™ system (Ellex Medical Lasers, Australia,) is a Q-switched Nd: YAG laser that vaporizes floaters at lower energy levels and with fewer shots than older laser systems. Unlike traditional Nd: YAG lasers, this modified laser is focused onto the front surface of the floater. The laser emits a short and small burst of energy that provides a

potent power density (10^9 J/cm²). This energy converts collagen and hyaluronic molecules found in a floater into a gas, which is then reabsorbed into the eye (*The Ultra Q Reflex*). Findings from a retrospective, observational study by Inder Paul Singh, MD (*Cummings et al., 2016*) of 168 eyes of 124 patients (mean age, 66 years (range, 42–89 years)) who underwent YAG vitreolysis (laser power range 2.0–5.5 mJ) demonstrated that 92% of patients were satisfied with the procedure. There was one case of intraocular pressure (IOP) spike, and two phakic lenses were damaged, one of which required cataract surgery. No retinal detachment or other retinal complications were seen and there was no anterior chamber or vitreous reactions (*Singh, 2014*).

Refractive

Lasers and light therapy has perhaps received the widest recognition in the treatment of refractive errors, in particular the use of excimer lasers.

Excimer laser (PRK and LASIK)

Excimer lasers are used in the treatment of myopia, hyperopia and astigmatism; operates in the ultraviolet spectral region. Laser-assisted in situ keratomileusis (LASIK) and PRK both employ this type of laser technology. While PRK uses an excimer to reshape the surface of the cornea, LASIK uses an excimer laser reshapes the cornea under a corneal flap at the stromal level.

Excimer lasers typically operate in the UV spectrum and use a gas mixture (known as the excimer gain medium) containing a noble gas such as argon, and a halogen, e.g., chlorine or fluorine. When a high voltage electric discharge is delivered into the laser cavity these gases combine to form a high energy excited gas state (excited dimer). When this high-energy compound subsequently dissociates, a photon of energy is released corresponding to the energy of the noble gas-halogen molecule. This light energy is amplified in the laser system resulting in the production of a high-energy discrete pulse of laser energy which disrupts chemical bonds of proteins, glycosaminoglycans, and nucleic acids in the cornea. Importantly, laser energy is absorbed by these compounds but with minimal damage to surrounding tissue because of a high energy density and short pulse duration (*Manche et al., 1998*).

The efficacy of LASIK and PRK is well established. However, a literature analysis that included randomized controlled trials comparing LASIK and PRK for the correction of any degree of myopia (1135 participants, 1923 eyes) concluded that LASIK gives a faster visual recovery and is a less painful technique than PRK. Nevertheless, the authors added that the two techniques appear to give similar outcomes 1 year after surgery (*Shortt et al., 2013*).

Photorefractive keratectomy (PRK) is a laser that utilizes excimer laser technology, and is typically used in the treatment of keratoconus. Current regimes for PRK utilize topography-guided ablation profiles designed to help reduce corneal surface irregularities with the goal of improving vision. In a retrospective follow-up study by *Tambe et al. (2015)*, 28 eyes of 23 patients (age 17–60) with grade 1–3 keratoconus received topography-guided PRK between 1998 and 2013. Of the 28 eyes, 18% underwent corneal transplantation at a median of 7 years (range 3–10) after PRK, four eyes were not available for follow-up, and in the remaining 19 eyes, corrected distance visual acuity (CDVA) was improved in 16 eyes (84.3%), reduced in 2 eyes (10.5%), and unchanged in 1 eye (5.2%). The mean spherical equivalent was reduced from -6.2 to -3.7 diopters (D) after 3 months and to -2.1 D at the late follow-up visit. Similarly, the mean cylinder was reduced from -4.2 to -3.0 D after 3 months and at the late follow-up. The authors suggested, therefore, that topography-guided PRK in keratoconus may help to reduce myopia and astigmatism and may offer a temporary or permanent alternative to keratoplasty in contact lens-intolerant keratoconus.

Femtosecond laser

Femtosecond laser technology is now also being employed in LASIK. Known as femtosecond laser-assisted LASIK (FS-LASIK), the femtosecond laser is used for flap creation while the excimer laser is used for stromal bed ablation. Thus the femtosecond laser allows the surgeon to create a thin-hinged corneal flap and a significant less deviation in flap thickness without using a micro-keratome. An analysis published in 2017 concluded that FS-LASIK yielded more predictable corneal flaps, lesser ocular aberrations, better uncorrected visual acuity, less variations in IOP and a reduced risk of developing dry eye. However, FS-LASIK may cause transient light sensitivity, diffuse lamellar keratitis, opaque bubble layer, corneal haze and rainbow glare (*Bashir et al., 2017*).

The SMILE procedure also uses femtosecond laser technology. SMILE evolved from Femtosecond Lenticule Extraction (FLEX) and has been shown to be safe and effective treatment for myopia (*Vestergaard et al., 2012; Wang et al., 2013; Kamiya et al., 2014*). The procedure involves cutting a small lenticule in the corneal stroma with a femtosecond laser. The lenticule is subsequently removed through a small incision.

There are number of published studies describing SMILE visual and refractive outcomes (*Vestergaard et al., 2012; Wang et al., 2013; Kamiya et al., 2014; Sekundo et al., 2014; Agca et al., 2014; Lin et al., 2014*). A systematic review and meta-analysis that compared (SMILE) with (FS-LASIK) for treating myopia concluded that both FS-LASIK and SMILE are safe, effective and predictable surgical options for treating myopia. However, they also found that dry eye symptoms and loss of corneal sensitivity may occur less frequently after SMILE than after FS-LASIK (*Shen et al., 2016*). Nevertheless, limitations of SMILE have been identified. A literature analysis published by Seiler *et al.* noted that the visual rehabilitation after SMILE takes significantly longer (weeks) compared to LASIK (days). They added that the refractive success rate of SMILE is still not as good as that of LASIK (88% vs. 95% within ± 0.5 D) (*Seiler et al., 2017*). The disadvantage of SMILE is that a rest refractive error it is said an enhancement is not possible with the FS laser. Thus for this procedure an excimer laser is needed.

Femtosecond lasers are also used in the treatment of keratoconus. The femtosecond laser is a near-infrared laser that works at a wavelength of 1053 nm. This laser is utilized to cut small precise channels in the cornea, for implantation of intrastromal corneal ring segments (ICRS). ICRS are thin curved pieces of plastic that are inserted into the cornea in order to provide biomechanical stability in keratoconus. Historically these channels were made with a handheld instrument, but this often resulted in the instrument inadvertently puncturing through the already thinned corneal stroma. The femtosecond laser is now used to cut precise channels prior to implantation of ICRS. The laser has an ultra-short-pulse duration (in the range of 10^{-15} s), allowing it to produce smaller shock waves which helps to minimize collateral damage (Kullman and Pineda, 2010; Mamalis, 2011). There are only a few studies comparing the effectiveness of femtosecond versus manual channel creation; however, published data suggest that femtosecond laser channel creation is safe with few complications. Coskunseven *et al.* (2011) conducted a retrospective chart review of 531 patients (850 eyes) who underwent ICRS using a femtosecond laser for channel creation. They noted that the overall complication rate was 5.7% (49 cases out of 850 eyes). The main complication was incomplete channel formation (2.7% cases).

Other

Lasers also play an important role in the treatment of glaucoma (Table 1) and include selective laser trabeculoplasty (SLT), argon laser trabeculoplasty (ALT) (both of which target the trabecular meshwork and are used in the treatment of open-angle glaucoma) and laser iridotomy, which targets the outer edge of the iris and is used to treat closed-angle glaucoma. Lasers are also used in cyclophotocoagulation (targeting the ciliary body) and for suture lysis.

Selective laser trabeculoplasty has largely replaced ALT, mainly due to its improved safety profile. The improved safety of SLT versus ALT is due to the former's very short pulse duration of three nanoseconds. This permits selective photothermolysis, i.e., the laser selectively targets cells of the trabecular meshwork. Unlike ALT, it does not cause coagulative damage to the trabecular meshwork and reduces collateral damage to surrounding tissues (Song, 2016).

Future Applications

There is a large variety of application developments and new technologies under development in ophthalmology. In this article we are highlighting some of the current developments that might affect the current standard of care for treating patients with vision disorders.

Photochemical– Corneal cross linking for refractive procedures

Encouragingly, CXL is also emerging as a treatment for pathologies other than corneal ectasia. As noted, CXL is being evaluated as a treatment for refractive errors. Known as photorefractive intrastromal corneal collagen cross-linking (PiXL™), the procedure uses riboflavin and a ultraviolet-A (UVA) 365 nm wavelength light source with programmable illumination pattern delivered by a proprietary device to cause differential corneal tissue “stiffening” to reshape the cornea. A key advantage of this approach is that, unlike tissue ablation/removal in current corneal refractive procedures, there is no removal of stromal tissue. Early study findings are encouraging.

Thus far, studies have focused on the use of PiXL as a treatment for low myopia. Feasibility study findings published in 2014 (Kanellopoulos, 2014) showed that application of 12 J/cm² applied transepithelially in a predetermined pattern led to an average improvement of 2.3 diopters (D) 1 week postoperative. Findings from a 6-month prospective study of 14 patients with myopia or myopic astigmatism undertaken by Matthias Elling, FEBO, showed statistically-significant improvement in UCVA at 1, 3 and 6 months postoperatively. A significant improvement in BCVA was also observed. Data also improvements in mean manifest sphere and manifest refractive spherical equivalent at all visits versus baseline (Elling *et al.*, 2017). Data presented at XXXIV Congress of the European Society of Cataract and Refractive Surgeons (ESCRS) from a 12-month study of 39 eyes, also undertaken by Matthias Elling showed significant improvements in UCVA and no loss of BCVA. In a 9-month study of eight patients, a mean myopia reduction of 0.75 D with no loss of BCVA was noted (Primary Care Optometry News, 2017).

Photochemical – Corneal cross linking for corneal infections

Collagen corneal cross-linking is also being investigated as a treatment of infectious keratitis. Known in this context as PACK-CXL (photo activated chromophore for infectious keratitis), the procedure may provide an alternative to standard antibiotic therapy in treating infectious corneal disorders (Tabibian *et al.*, 2015). The antimicrobial effect occurs due to the effect of UV light interacting with riboflavin. The resulting reactive oxygen species damage the pathogen cell walls; riboflavin may also interact with the nucleic acids of the pathogen, preventing replication.

Larger randomized prospective clinical trials are few for PACK-CXL thus far, but smaller case reports findings from small series prove promising. Makdoui and colleagues reported a series of 16 patients with culture proven bacterial ulcers who all improved after being treated with CXL alone. All eyes achieved epithelial healing and less inflammation, although two patients required antibiotic therapy to completely resolve their infection (Makdoui *et al.*, 2012). Said *et al.* (2014) compared PACK-CXL with standard anti-microbial therapy in 40 eyes of 40 patients with advanced infectious keratitis and coexisting corneal melting. Twenty-one patients underwent PACK-CXL and antimicrobial therapy while 19 patients received antimicrobial therapy only.

Table 1 Application of laser and light technologies according to tissue type/indication

<i>Intended use</i>	<i>Application</i>	<i>Tissue structure</i>	<i>Laser/light source</i>	<i>Wavelength</i>	<i>Interaction mechanism</i>
Treatment of corneal ectasia	Corneal cross linking	Cornea	UV-LED (continuous wave or pulsed)	365 nm	Photochemical
Corneal scars or other corneal disorders	Phototherapeutic keratectomy (PTK)	Cornea	ArF Excimer laser	193 nm	Photoablation
Treatment of refractive errors	Photorefractive keratectomy (PRK) and Laser assisted in-situ keratomileusis (LASIK)	Cornea	ArF Excimer laser	193 nm	Photoablation
	LASIK- Flaps and Small Incision Lenticule Extraction (SMILE)	Cornea	Femtosecond lasers (pulse duration 220–580 fs)	1043 nm	Photomechanical/ photodisruption
Treatment of glaucoma	Laser trabeculoplasty	Trabecular mesh work	2w-Nd: YAG Argon laser	532 nm 515 nm	Photothermal
	Laser iridotomy/laser peripheral iridoplasty	Iris	Nd:YAG (2w)	532 nm 1064 nm	Photothermal/ photomechanical
	Cyclophotocoagulation	Ciliary body	Argon Nd:YAG	515 nm 1064 nm	Photothermal
	Glaucoma laser suture lysis	Cutting a suture of trabecular ectomy scleral flap	Diode laser 2w-Nd:YAG Argon laser	810 nm 532 nm 515 nm	Photothermal
Treatment of cataract	Femtosecond laser-assisted cataract surgery (FLACS)	Cataractous lens fragmentation and incisions for cataract surgery	Neodymium: glass (pulse duration 300 – 500 fs)	1053 nm	Photomechanical/ photodisruption
	Posterior Capsulotomy	Turbid posterior lens capsule after intra-ocular lens implantation	Nd:YAG (Q-switched)	1064 nm	Photomechanical/ photodisruption
Treatment retinal or choroidal pathology	Retinal ischemic processes (e.g., diabetes, retinopathy of prematurity)	Retina	2w-Nd: YAG Argon laser	532 nm 515 nm	Photothermal
	Retinal tears	Retina	2w-Nd: YAG Argon laser	532 nm 515 nm	Photothermal
	Focal macular laser therapy for focal macular oedema	Retina	2w-Nd: YAG Argon laser	532 nm 515 nm	Photothermal
	Photodynamic therapy for choroidal neovascularization	Choroid	Diode laser	690 nm	Photochemical
	Transpupillary thermotherapy (TTT)	Malignant choroidal melanoma	Diode laser	810 nm	Photothermal
Treatment of vitreous disorders	Vitreous strand in the anterior chamber, posterior vitreous detachment and vitreous floaters	Vitreous	Nd: YAG (Q-switched)	1064 nm	Photomechanical
Treatment of lid and orbit	Blepharoplasty	Upper and lower eye lid	CO ₂ laser	10,600 nm	Photoablation

Although PACK-CXL did not shorten the time to corneal healing, the complication rate was 21% in the antimicrobial group, whereas there was no incidence of corneal perforation or recurrence of the infection in the PACK-CXL group.

Photothermal – Computer guided treatments

Studies of PRP in diabetic macular edema (DME) using conventional slit-lamp based laser application have shown low success rates in terms of either visual acuity gains or reductions in anti-VEGF use burden, likely due to inadequate accuracy and completeness associated with conventional laser therapy (ETDRS Study Group, 1985; Michael *et al.*, 2011; Mitchell *et al.*, 2011; Brown *et al.*, 2013). Consequently, technology has been developed to improve laser spot accuracy. Computer-guided technology for navigated laser photocoagulation offers significantly greater accuracy in laser spot application to areas of edema than traditional laser treatment. Study findings also show that navigated laser photocoagulation reduces the retreatment rate (as compared with conventional laser therapy) by almost half in patients with DME (Neubauer *et al.*, 2013). An additional development is sub-threshold diode micropulse laser. This was developed with the treatment of DME in mind, but it may also be effective in the treatment of PDR.

Photomechanical – The SMILE procedure

As noted earlier, SMILE is indicated for myopia with clinically-insignificant astigmatism. However, trials are underway to evaluate SMILE in myopic patients with astigmatism and hyperopic patients. Use of 345 nm UV femtosecond laser pulses, which may have applications in SMILE, but also flap creation and ultrathin Descemet-stripping endothelial keratoplasty has also been tested in vivo by Hammer and colleagues (Hammer *et al.*, 2015). The Laser energy levels required by a ultraviolet femtosecond laser to disrupt corneal tissue are tenfold lower than those required by an infrared laser offering significant safety advantages. Additionally, light of shorter wavelength can be focused more precisely and may therefore allow more precise photodisruption of the corneal stroma. Clinical studies are needed to determine the safety and efficacy of 345 nm UV femtosecond laser.

Conclusion

Laser and light therapy continues to play a key role in the management of many ophthalmic disorders. Ongoing developments suggest that improvements may be made to existing treatments, and may even provide new ways of treating problems, for example, guided crosslinking in the treatment of myopia. Although, a greater body of evidence is needed to ascertain whether these new technologies will have a clinically-significant impact, it appears likely that lasers and light therapies will have a growing role in ophthalmology.

References

- Abell, R.G., Darian-Smith, E., Kan, J.B., *et al.*, 2015. Femtosecond laser-assisted cataract surgery versus standard phacoemulsification cataract surgery: Outcomes and safety in more than 4000 cases at a single center. *J Cataract Refract Surg.* 41 (1), 47–52.
- Agca, A., Demirok, A., Cankaya, K.I., *et al.*, 2014. Comparison of visual acuity and higher-order aberrations after femtosecond lenticule extraction and small-incision lenticule extraction. *Cont Lens Anterior Eye* 37 (4), 292–296.
- American Academy of Ophthalmology, Panretinal photocoagulation. Available at: http://eyewiki.aao.org/Panretinal_Photocoagulation.
- Balasubramanian, D., Du, X., Zigler, J.S., 1990. The reaction of singlet oxygen with proteins, with special reference to crystallins. *Photochem Photobiol.* 52, 761–768.
- Bashir, Z.S., Ali, M.H., Anwar, A., Ayub, M.H., Butt, N.H., 2017. Femto-LASIK: The recent innovation in laser assisted refractive surgery. *J Pak Med Assoc.* 67 (4), 609–615.
- Brown, D.M., Nguyen, Q.D., Marcus, D.M., *et al.* RIDE and RISE Research Group, 2013. Long-term outcomes of ranibizumab therapy for diabetic macular edema: The 36-month results from two phase III trials: Rise and RIDE. *Ophthalmology* 120 (10), 2013–2022.
- Cairns, W.L., 1970. Products of riboflavin photodegradation, Iowa State University. Available at: <http://lib.dr.iastate.edu/rtd/4214/>.
- Carr, D.O., 1960. Mechanism of riboflavin-catalyzed oxidations, Iowa State University. Available at: <http://lib.dr.iastate.edu/rtd/2813/>.
- Choi, M., Kim, J., Kim, E.K., Seo, K.Y., Kim, T.I., 2017. Comparison of the conventional dresden protocol and accelerated protocol with higher ultraviolet intensity in corneal collagen cross-linking for keratoconus. *Cornea* 36 (5), 523–529.
- Coskunseven, E., Kymionis, G.D., Tsiklis, N.S., *et al.*, 2011. Complications of intrastromal corneal ring segment implantation using a femtosecond laser for channel creation: A survey of 850 eyes with keratoconus. *Acta Ophthalmol* 89 (1), 54–57.
- Cummings, A.B., McQuaid, R., Naughton, S., Brennan, E., Mrochen, M., 2016. Optimizing corneal cross-linking in the treatment of keratoconus: A comparison of outcomes after standard- and high-intensity protocols. *Cornea* 35 (6), 814–822.
- Elling, M., Kersten-Gomez, I., Dick, H.B., 2017. Photorefractive intrastromal corneal cross-linking for the treatment of myopic refractive error: Six month interim findings of an ongoing 12-month prospective study. *J Cataract Refract Surg* 43 (6), 789–795.
- Elman, M.J., Bressler, N.M., Qin, H., *et al.*, 2011. Expanded 2-year follow-up of Ranibizumab plus prompt or deferred laser or triamcinolone plus prompt laser for diabetic macular edema. *Ophthalmology* 118 (4), 609–614.
- ETDRS Study Group, 1985. Photocoagulation for diabetic macular edema. Early Treatment Diabetic Retinopathy Study report number 1. Early Treatment Diabetic Retinopathy Study research group. *Arch Ophthalmol.* 103 (12), 1796–1806.
- Greenstein, S.A., Hersh, P.A., 2013. Characteristics influencing outcomes of corneal collagen crosslinking for keratoconus and ectasia: Implications for patient selection. *J Cataract Refract Surg.* 39 (8), 1133–1140.
- Hammer, C.M., Petsch, C., Klenke, J., *et al.*, 2015. Corneal tissue interactions of a new 345 nm ultraviolet femtosecond laser. *J Cataract Refract Surg.* 41 (6), 1279–1288.
- Herekar, S., 2007. Fractionated cross-linking: Preliminary lab investigation. In: 3rd International Congress on Corneal Cross Linking, Zurich, Switzerland (accessed 08.12.07).
- Kamiya, K., Shimizu, K., Igarashi, A., Kobashi, H., 2014. Visual and refractive outcomes of femtosecond lenticule extraction and small-incision lenticule extraction for myopia. *Am J Ophthalmol* 157 (128–134), e122.

- Kanellopoulos, A.J., 2014. Novel myopic refractive correction with transepithelial very high-fluence collagen cross-linking applied in a customized pattern: Early clinical results of a feasibility study. *Clin Ophthalmol* (Auckland, NZ) 8, 697–702.
- Kozak, I., Luttrull, J.K., 2015. Modern retinal laser therapy. *Saudi Journal of Ophthalmology* 29 (2), 137–146.
- Kullman, G., Pineda 2nd, R., 2010. Alternative applications of the femtosecond laser in ophthalmology. *Semin Ophthalmol* 25, 256–264.
- Li, J., Ji, P., Lin, X.L., 2015. Efficacy of corneal collagen cross-linking for treatment of keratoconus: A meta-analysis of randomized controlled trials. *PLOS ONE* 10 (5), e0127079. doi:10.1371/journal.pone.0127079.
- Lichtman, J.W., Conchello, J.A., 2005. Fluorescence microscopy: Review. *Nature Methods* 2, 910–919.
- Lin, F., Xu, Y., Yang, Y., 2014. Comparison of the visual results after SMILE and femtosecond laser-assisted LASIK for myopia. *J Refract Surg* 30, 248–254.
- Makdoui, K., Mortensen, J., Sorkhabi, O., Malmvall, B.E., Crafoord, S., 2012. UVA-riboflavin photochemical therapy of bacterial keratitis: A pilot study. *Graefes Arch Clin Exp Ophthalmol* 250, 95–102.
- Mamalis, M., 2011. Femtosecond laser: The future of cataract surgery? *J Cataract Refract Surg* 37, 1177–1178.
- Manche, E.E., Carr, J.D., Haw, W.W., Hersh, P.S., 1998. Excimer laser refractive surgery. *Western Journal of Medicine* 169 (1), 30–38.
- Meiri, Z., Keren, S., Rosenblatt, A., *et al.*, 2016. Efficacy of corneal collagen cross-linking for the treatment of keratoconus: A systematic review and meta-analysis. *Cornea* 35 (3), 417–428.
- Mitchell, P., Bandello, F., Schmidt-Erfurth, U., *et al.* RESTORE Study Group, 2011. The RESTORE study: Ranibizumab monotherapy or combined with laser versus laser monotherapy for diabetic macular edema. *Ophthalmology* 118 (4), 615–625.
- Neubauer, A.S., Langer, J., Liegl, R., *et al.*, 2013. Navigated macular laser decreases retreatment rate for diabetic macular edema: A comparison with conventional macular laser. *Clin Ophthalmol* 7, 121–128.
- Pajic, B., Hafezi, F., 2011. Applanation-free femtosecond laser processing of the cornea. In: Pajic-Eggspuehler, B. (Ed.), *Femtosecond Laser Techniques and Technology*, OSA Publishing.
- Palanker, D., 2014. Ophthalmic laser therapy: Mechanisms and applications. Available at: https://web.stanford.edu/~palanker/publications/Ophthalmic_Laser_Therapy.pdf.
- Popovic, M., Campos-Möller, X., Schlenker, M.B., Ahmed, I.I., 2016. Efficacy and safety of femtosecond laser-assisted cataract surgery compared with manual cataract surgery: A meta-analysis of 14,567 eyes. *Ophthalmology* 123 (10), 2113–2126.
- Primary Care Optometry News, 2017. Studies show promising results for PiXL procedure in low myopia.
- Reinstein, D.Z., Archer, T.J., Gobbe, M., 2014. Small incision lenticule extraction (SMILE) history, fundamentals of a new refractive surgery technique and clinical outcomes. *Eye and Vision* 1, 3. doi:10.1186/s40662-014-0003-1.
- Rich, H., Odlyha, M., Cheema, U., Mudera, V., Bozec, L., 2014. Effects of photochemical riboflavin-mediated crosslinks on the physical properties of collagen constructs and fibrils. *Journal of Materials Science Materials in Medicine* 25 (1), 11–21.
- Royle, P., Mistry, H., Auguste, P., *et al.*, 2015. Pan-retinal photocoagulation and other forms of laser treatment and drug therapies for non-proliferative diabetic retinopathy: Systematic review and economic evaluation. Southampton (UK): NIHR Journals Library. doi:10.3310/hta19510.
- Sadoughi, M.M., Einollahi, B., Baradaran-Rafii, A., *et al.*, 2016. Accelerated versus conventional corneal collagen cross-linking in patients with keratoconus: An inpatient comparative study. *Int Ophthalmol*. doi:10.1007/s10792-016-0423-0.
- Said, D.G., Elalfy, M.S., Gatzoufas, Z., *et al.*, 2014. Collagen cross-linking with photoactivated riboflavin (PACK-CXL) for the treatment of advanced infectious keratitis with corneal melting. *Ophthalmology* 121 (7), 1377–1382.
- Seiler, T., Koller, T., Wittwer, V.V., 2017. Limitations of SMILE (Small Incision Lenticule Extraction). *Klin Monbl Augenheilkd* 234 (1), 125–129.
- Sekundo, W., Gertner, J., Bertelmann, T., Solomatin, I., 2014. One-year refractive results, contrast sensitivity, high-order aberrations and complications after myopic small-incision lenticule extraction (ReLEx SMILE). *Graefes Arch Clin Exp Ophthalmol* 252 (5), 837–843.
- Shen, Z., Shi, K., Yu, Y., *et al.*, 2016. Small incision lenticule extraction (SMILE) versus femtosecond laser-assisted in situ keratomileusis (FS-LASIK) for myopia: A systematic review and meta-analysis. (Virgili G, ed.) *PLOS ONE* 11 (7), e0158176. doi:10.1371/journal.pone.0158176.
- Shortt, A.J., Allan, B.D., Evans, J.R., 2013. Laser-assisted in-situ keratomileusis (LASIK) versus photorefractive keratectomy (PRK) for myopia. *Cochrane Database Syst Rev* 1), CD005135.
- Silva, R.A., Blumenkranz, M.S., 2013. Prophylaxis for retinal detachments. *American Academy of Ophthalmology*. Available at: <https://www.aao.org/munnerlyn-laser-surgery-center/prophylaxis-retinal-detachments>.
- Singh, I.P., 2014. How YAG Laser Vitreolysis can be Used in Practice for Treatment of Floaters. *Ophthalmology Times*.
- Smith, E.C., 1963. The photochemical degradation of riboflavin, Iowa State University. Available at: <http://lib.dr.iastate.edu/rtd/2947/>.
- Song, J., 2016. Complications of selective laser trabeculoplasty: A review. *Clinical Ophthalmology* (Auckland, NZ) 10, 137–143.
- Spoerl, E., Huhle, M., Seiler, T., 1998. Induction of cross-links in corneal tissue. *Exp Eye Res* 66 (1), 97–103.
- Steinert, R.F., 2013. Nd:YAG laser posterior capsulotomy. Available at: <https://www.aao.org/munnerlyn-laser-surgery-center/ndyag-laser-posterior-capsulotomy-3>.
- Tabibian, D., Richoz, O., Hafezi, F., 2015. PACK-CXL: Corneal cross-linking for treatment of infectious keratitis. *Journal of Ophthalmic & Vision Research* 10 (1), 77–80.
- Tambe, D.S., Ivarsen, A., Hjortdal, J., 2015. Photorefractive keratectomy in keratoconus. *Case Reports in Ophthalmology* 6 (2), 260–268.
- The Ultra Q Reflex. Ellex medical lasers. Available at: <http://www.ellex.com/physicians/product-portfolio/laser-floater-removal/ultra-q-reflex/>.
- Vestergaard, A., Ivarsen, A.R., Asp, S., Hjortdal, J.O., 2012. Small-incision lenticule extraction for moderate to high myopia: predictability, safety, and patient satisfaction. *J Cataract Refract Surg* 38, 2003–2010.
- Wang, Y., Bao, X.L., Tang, X., *et al.*, 2013. Clinical study of femtosecond laser corneal small incision lenticule extraction for correction of myopia and myopic astigmatism. *Zhonghua Yan Ke Za Zhi* 49 (292–298), 19.
- Wollensak, G., Spoerl, E., Seiler, T., 2003. Riboflavin/ultraviolet-a-induced collagen crosslinking for the treatment of keratoconus. *Am J Ophthalmol* 135 (5), 620–627.
- World Health Organization, 2013. What is refractive error? Available at: <http://www.who.int/features/qa/45/en/>.

Probing Ocular Mechanics With Light

Susana Marcos, Instituto de Optica, CSIC, Madrid, Spain

Giuliano Scarcelli, University of Maryland, College Park, MD, United States

Antoine Ramier and Seok H Yun, Massachusetts General Hospital, Boston, MA, United States

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

FEM Finite Element Model

OCT Optical Coherence Tomography

sOCT Spectral OCT

ssOCT Swept Source OCT

CXL Cross-Linking

UVX Riboflavin Ultraviolet Light CXL

RGX Rose Bengal Green Light CXL

IOP Intraocular Pressure

FP Fabry–Perot

VIPA Virtually Imaged Phased Array

DSEK Descemet's stripping endothelial keratoplasty

DALK Deep anterior lamellar keratoplasty

LASIK Laser-assisted in situ keratomileusis

OCE Optical coherence elastography

SNR Signal to noise ratio

ARF Acoustic radiation force

λ Wavelength of light in vacuum

ρ Density

ν Poisson's ratio

E Young's modulus

c_g Group velocity

$\Delta\phi$ Phase difference

ΔL Optical path length difference

Δu_z Geometric displacement (component along the optical axis z)

n Refractive index

$\sigma_{\Delta L}$ Sensitivity (noise floor) for the optical path length difference

Glossary

Acoustic radiation force Physical force exerted by acoustic waves.

Brillouin Scattering Inelastic light scattering phenomenon where incident light interacts with acoustic phonons within a material to produce scattered light with different color (wavelength or frequency).

Corneal Refractive Surgery Surgical reshaping of the cornea (generally using laser ablation or corneal lenticule extraction) to produce a refractive change to correct refractive errors.

Corneal Videokeratography Technique to measure anterior corneal surface topography, generally base on the projection of rings onto the cornea and analysis of the acquired image.

Cross-Linking Surgical treatment of keratoconus using a photosensitizer and a light irradiation to produce chemical bonds between collagen fiber and increase corneal stiffness.

Ectasia Posterior corneal bulging resulting from reduced corneal thinning in corneal refractive surgery or disease.

Elastography Imaging modality that maps the elastic properties of tissues.

Etalon Transparent slab with two reflecting surfaces used to separate light components of different colors based on multi-pass interference.

Glaucoma Ocular disease in which damage is produced in the optic nerve head (generally due to increased Intraocular Pressure), resulting in nerve fiber loss and peripheral vision loss.

Gonioscopy lens Specialized lens used in ophthalmology placed on the surface of the cornea for visualizing the anterior segment angle.

Interference (of light) Physical phenomenon where two light waves are superposed and form a resultant wave of larger or smaller amplitude depending on the relative delay of the two.

Intrastromal Corneal Ring Segments Surgical treatment of keratoconus consisting on sections of ring segments (with triangular or hexagonal section) implanted in the corneal periphery through a manual or femto-second laser channel in order to flatten the cornea and increase surface regularity.

Keratoconus Ocular disease produced in which corneal weakening leads to corneal bulging and degraded optical quality and vision.

Modulus (of elasticity) Parameter that quantifies the resistance of a sample to be deformed.

Myopia Ocular condition in which distant objects are focused in front of the retina resulting in a blurred image.

Optical Coherence Tomography Low coherence interferometric imaging technique allowing high resolution tomographic images of the eye. Common modalities (Fourier domain) include spectral and swept source operation principles.

Presbyopia Age-related loss of accommodation capacity with age, associated to increased stiffness in the crystalline lens.

Rayleigh surface waves Elastic waves propagating at the surface of a solid, comprising both longitudinal and transverse components, solutions of the elastic wave equation for a semi-infinite medium.

Scheimpflug Imaging Eye anterior segment imaging modality in which a slit is projected onto the eye and cross-section images of the eye from the anterior cornea to the posterior cornea (or to the posterior lens) are collected. The

camera axis is tilted 45° with respect to the optical axis in order to increase the depth of focus. Geometrical and optical distortion needs to be corrected for proper quantification.

Spectrometer Instrument capable of splitting incident light in its sub-components of different color (wavelength or frequency).

Introduction

The cornea and crystalline lens of the eye are optical elements that project images of the outside world onto the retina. Although primarily regarded as lenses and investigated in terms of their optical performance, understanding the mechanical properties of the ocular elements is also crucial. The mechanical structure of the cornea determines its shape, which in turn determines its optical quality. A compromise of the corneal mechanical structure will therefore highly impact vision of the patient (Dupps and Wilson, 2006). On the other hand, the crystalline lens in the young eye has the capability of changing its shape to focus near and far objects, so called accommodation. Increased in the crystalline lens stiffness results in the loss of accommodation capacity (Weeber *et al.*, 2007), a condition called presbyopia, affecting 100% of the population beyond the age of 45. The quest for a treatment restoring crystalline lens elasticity, or alternatively, accommodating intraocular lens with the mechanical capacity to respond to an accommodative stimulus would be the ultimate solutions for presbyopia. The mechanical properties of another ocular tissue, the sclera, have also been linked with the development of myopia, which results from the excessive elongation of the eyeball. It has been acknowledged that understanding the matrix and cellular factors contributing to a weakened sclera in myopes may aid in the development of a clinically appropriate treatment for myopia (McBrien and Jobling, 2009). Finally, because pressure is a mechanical phenomenon, and because of elevated intraocular pressure (IOP) is a risk factor for glaucoma, the role of ocular biomechanics in understanding the development and progression of glaucomatous optics neuropathy has long been recognized. On the one hand, standard measurements of IOP are largely affected by corneal biomechanics (Liu and Roberts, 2005). On the other hand, the mechanical effects of IOP on the optical nerve head, and the contribution of the tissue mechanical properties to the susceptibility of IOP insult are key to understand individual differences in the mechanical, ischemic and cellular events occurring in glaucoma (Sigal and Grimm, 2012).

While the implication of ocular tissue mechanics in presbyopia, myopia and glaucoma is clear, the cornea has been a primary focus of attention to the investigation and clinical measurement of biomechanics. Knowledge of the corneal biomechanical properties is key in the understanding of the progression of certain corneal diseases (such as keratoconus) or the corneal response to treatments. Certain corneal treatments such as surgeries involving incisions (cataract, incisional refractive surgery) (Denoyer *et al.*, 2013) or intrastromal corneal ring segment implantation to treat keratoconus rely specifically on the mechanical response of the cornea (Ortiz *et al.*, 2012; Pérez-Merino *et al.*, 2013, 2014). Emerging treatments of keratoconus, such as corneal collagen cross-linking, aim at increasing corneal stiffness, therefore changing the corneal mechanical response.

Despite the agreement on the clinical need to monitor the mechanical properties of the cornea, most knowledge of the inherent constituent properties of the cornea comes from tests in vitro (uniaxial extensimetry, two-dimension flap extensimetry, or corneal or whole eye inflation) (Andreassen *et al.*, 1980; Wollensak *et al.*, 2003; Kling *et al.*, 2010, 2012; Boyce *et al.*, 2007; Elsheikh *et al.*, 2007; Kling and Marcos, 2013a,b,c). The reports of the corneal Young modulus in the literature span almost two orders of magnitude. The differences across studies likely arise from the differences across methods, the relative complexity of the corneal structure, and the differences in the experimental conditions across studies (Andreassen *et al.*, 1980; Wollensak *et al.*, 2003; Kling *et al.*, 2010, 2012; Boyce *et al.*, 2007; Elsheikh *et al.*, 2007; Kling and Marcos, 2013a,b,c). Some methods attempt to reproduce the geometry and the integrity of the cornea of an intact eye (i.e., whole eye inflation), however it is likely that the effect of hydration and post-mortem time in experiments *ex vivo* are an important source of variability in the measurements (Kling and Marcos, 2013a,b,c). Besides, there is a clinical need for in vivo evaluation of corneal biomechanical parameters, at the patient level, for diagnostics and monitoring of disease and treatment. The availability of a method to quantify the corneal biomechanical properties of a patient in the clinic will allow to identify the risk of post-LASIK ectasia in a refractive surgery candidate (Ambrosio *et al.*, 2011). A clinical tool will also aid in the diagnosis of keratoconus, which at present is solely based on topographic and geometrical evaluation of the patient's cornea. State-of-the-art treatment techniques for keratoconus will also largely benefit for quantification of corneal biomechanics. The selection of the appropriate Intrastromal Ring Segment (ICRS) design in refractive or keratoconus treatment (Kling and Marcos, 2013a,b,c) can be guided by individual assessment of the corneal biomechanical response. On the other hand, quantitative biomechanical parameters will help to monitor the effect of collagen cross-linking (Bak-Nielsen *et al.*, 2013), a new treatment for keratoconus aiming at stiffening corneal tissue.

Requirements of in vivo measurements of corneal mechanics include non-invasiveness, rapid acquisition times, and the availability of data allowing direct reconstruction of the inherent mechanical properties of the cornea, in isolation from other factors. The use of optical imaging methods fulfills most of those requirements, and is ideal for in vivo assessment. Several approaches have been suggested to measure corneal biomechanical properties in vivo, either using optical, acoustic or magnetic imaging. These include: corneal videokeratography upon stepwise central indentation with a cantilever that measures corneal bending-resistance (Grabner *et al.*, 2005); ultrasonic and magnetic resonance techniques that use the shear wave propagation velocity to map corneal elasticity in vivo (Litwiller *et al.*, 2010); corneal optical coherence elastography producing a strain mapping based on speckle tracking (Muthupillai *et al.*, 1995; Ford *et al.*, 2006); air-puff corneal deformation imaging; phase-sensitive

Optical Coherence Tomography based on phase changes of successive corneal cross-sectional images (Chang *et al.*, 2002); and Brillouin microscopy providing a spatially resolved map of corneal Brillouin modulus, which is correlated to the local relative values of corneal elastic modulus, upon induction of quasi-static oscillations (GHz) in the tissue (Scarcelli *et al.*, 2007a,b). A drawback of several of these techniques is that they are slow (magnetic imaging), show a relatively low spatial resolution (ultrasonic/magnetic resonance imaging) or require contact of a probe with the patient's corneal surface and hence the use of anesthetics (physical applanation and ultrasound based techniques). The current article will address the three last based techniques, given the advantages of using light to capture mechanical information.

Besides, imaging techniques track corneal deformation, but quantification is needed to retrieve the inherent corneal bio-mechanical properties from the acquired images. Corneal tissue is a complex mechanical biomaterial, frequently described as a non-isotropic viscoelastic material. As in most soft-biological tissues, viscoelastic properties arise from the time-dependent nature of the biomechanical properties, are expected to be of major importance. For example, progressive deformation of the cornea (ectasia) occurring in keratoconus, and after some LASIK procedures, is a result of a viscoelastic phenomenon. Nevertheless, *ex vivo* techniques have primarily quantified the elastic properties (Young's modulus), which is in fact non-linear. Most *in vivo* techniques describe corneal deformation parameters, which are only partially dependent on the mechanical properties (other factors such as corneal thickness and IOP play a role in deformation). Besides, the corneal properties vary across corneal depth, which is directly addressed by the depth-resolved nature of Brillouin Microscopy.

The full power of light-based imaging techniques to quantify inherent mechanical properties is only achieved when those are combined with Finite Element Modeling (FEM). Assuming a certain material model and an accurate knowledge of the applied forces it is possible to retrieve the material variables through inverse optimization. Applied forces can be an air puff (which dynamically changes through a deformation event) or an acoustic stimulus. In Brillouin microscopy it is necessary to assume relationships between the Brillouin modulus and the modulus of elasticity and values for tissue density and index of refraction. Relationships between the shear modulus and the Brillouin modulus can also be validated with the aid of FEM analysis.

While these techniques have been primarily used to analyze the cornea, some of them are also suited to study mechanical properties of other tissues such as the crystalline lens (Brillouin microscopy) or the sclera (air puff deformation or OCT vibrography).

This article describes three optical imaging techniques to measure ocular biomechanics: (1) Air Puff Corneal Dynamic Imaging, using Scheimpflug imaging or Ocular Coherence Tomography (OCT); (2) Phase-sensitive OCT Elastography (3) Brillouin Microscopy; also known as OCT Vibrography. Special emphasis will be made on applications in corneal tissue. Inherent corneal properties in human and animal models obtained from FEM analysis and Inverse Optimization using input data from imaging techniques will be presented.

Air Puff Corneal Dynamic Imaging

High-speed corneal imaging in combination with an air puff combines the advantages of non-invasiveness and high-speed acquisition suitable for measurements *in vivo*. This concept, using Scheimpflug imaging has been recently commercialized by Oculus (Corvis ST system by Oculus). Air puff OCT has been demonstrated by groups at Copernicus University (Torun, Poland) and at the Institute of Optics CSIC (Madrid, Spain). This system is FDA approved as a tonometer, but provides also quantitative analysis of corneal deformation parameters (but not inherent material properties). Previous attempts (Ocular Response Analyzer by Reichert) to use air-puff applanation measurements to derive corneal mechanical properties were limited by the collection of relatively indirect data.

Direct imaging of the temporal and spatial corneal deformation profile during the air-puff event allows quantitative analysis and, through the use of Finite-Element modeling, retrieval of corneal elasticity moduli. In addition, the selected time for deformation (in the order of milliseconds) allows quantification of viscoelastic properties (relative viscoelasticity modulus and relaxation time constant) in a regime of interest for characterization of the corneal mechanical response to disease and treatments.

Air Puff Scheimpflug Imaging

The Corvis instrument by Oculus operates under the principle of Scheimpflug imaging. In Scheimpflug imaging a slit is projected onto the eye. A tilt of the imaging plane with respect to the optical axis of the instrument allows for a large depth-of-focus. While this generates geometrical distortions, these can be corrected by software. In the Corvis, the Ultra High Speed Scheimpflug camera takes over 4300 frames per second to monitor corneal response to a metered collimated air puff with fixed profile that has symmetrical configuration and fixed maximal internal pump pressure of 25 kPa. The Scheimpflug camera has blue light LED (455 nm) and covers 8.5 mm horizontally of a single slit. Recording measurement time is 30 ms, which allows for acquiring 140 digital frames. Each image has 576 measuring points.

Air Puff OCT Imaging

Direct OCT measurements of the human corneal dynamics during an air-puff induced deformation were first presented in 2011 (Alonso-Caneiro *et al.*, 2011). The authors reported a novel apparatus that combined swept source OCT (ssOCT) system

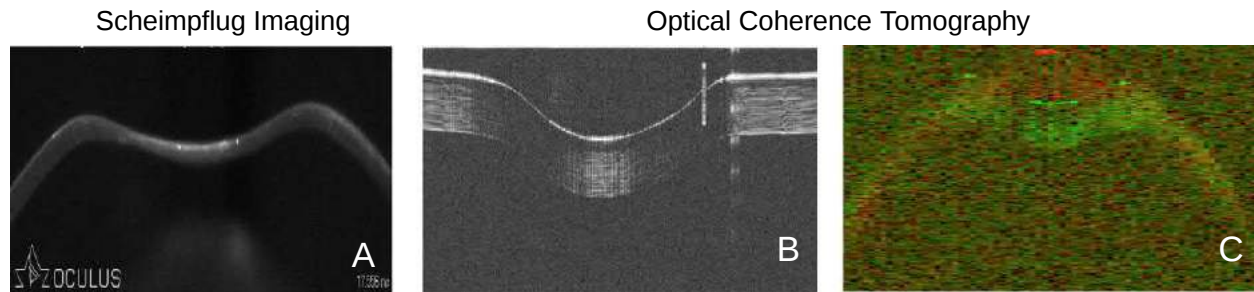


Fig. 1 Air-puff corneal deformation imaging (porcine eyes ex vivo). A. Spatial deformation from Scheimpflug-based imaging, Corvis ST by Oculus; B. Temporal deformation (A-scan of apex as a function of time) from air-puff ssOCT imaging (Nicolaus Copernicus University); C. Spatial deformation imaging (B-scan) from air-puff sOCT imaging (Instituto de Optica, CSIC). Merged undeformed cornea and maximally deformed cornea. Air-puff deformation event ≈ 20 ms in all cases. Images are from Alonso-Caneiro, D., Karnowski, K., Kaluzny, B.J., *et al.*, 2011. Assessment of corneal dynamics with high-speed swept source optical coherence tomography combined with an air puff system. *Optics Express* 19 (15), 14188–14199 and Dorronsoro, C., Pascual, D., Pérez-Merino, P., Kling, S., Marcos, S., 2012. Dynamic OCT measurement of corneal deformation by an air puff in normal and cross-linked corneas. *Biomedical Optics Express* 3 (3), 473–487.

(Karnowski *et al.*, 2011) with an air-puff chamber adapted from a commercial non-contact tonometer (**Fig. 1(A)**). A commercial prototype of swept source laser (AXSUN Technologies Inc., Billerica, MA) was used as a light source (50 kHz; 1300 nm; 110 nm optical bandwidth centered at 1300 nm, axial resolution of 9 μm in air; 0.9 mm imaging depth range; 102 dB; for 2.5 mW average optical power at the cornea). The module of an air-puff pressure chamber was extracted from a clinically approved tonometer (XPert NCT; Reichert Inc., Buffalo, NY) and was built into the imaging head of ssOCT system (**Fig. 1(B)**). A custom designed electrical circuit was used to power the air-puff chamber motor with high-voltage pulse. Additionally during the air-puff the induced corneal deformation reference pressure signal is obtained via pressure sensor originally built in the chamber.

Single A-scans are collected in time during the corneal deformation from the single/central point on the cornea where air stream is applied (M-scan). As a result the direct measurement of air induced deformation of the corneal surfaces and the anterior surface of the lens is presented. The corneal applanation/recovery time is approximately 20 ms.

The combination of an air-puff with a sOCT system was presented by Dorronsoro *et al.* (2012). In this configuration, a tilted mirror was used to couple a commercial non-contact tonometer (NT 2000, Nidek, Hiroishi, Japan) with a custom spectral domain OCT system (Alonso-Caneiro *et al.*, 2011). The tube of the tonometer providing the air puff passes through the center of the tilted mirror through a drilled hole. This configuration allows both instruments, the OCT system and the tonometer, point at the corneal apex in an almost normal orientation. This configuration allows dynamic imaging of corneal cross-sections during the air puff deformation event. The OCT imaging axis and the axis of air puff tube are perpendicularly oriented, and lying in a horizontal plane, with the tilted mirror located at the point of intersection between these two axes, at 45° (See **Fig. 1(C)**). In the vertical direction the mirror an additional tilt of 8° was imposed to avoid the horizontal scanning beam of the sOCT to overlay the mirror hole (tube tip) and therefore vignette the image. As in conventional non-contact corneal tonometry, the air puff impacts the cornea directly and produces a fast deformation event (< 30 ms). The distance from the non-contact tonometer tube tip to the corneal apex was set to 11-mm. This distance is the identical configuration to that used in the standard use for corneal applanation tonometry of the instrument. The air puff configuration was manually set, in order to obtain the same standard air puff spatio-temporal profile in all measurements. The full air pressure provided by the instrument was used in the experiments, and delivery timing was electronically triggered, in synchronization with the image acquisition by the OCT.

Corneal Deformation Parameters

This Corvis processing software allows dynamic inspection of the actual deformation process during deformation. Advanced algorithms for edge detection of the corneal contours are applied for every frame. The recording starts with the cornea at the natural convex shape. The air puff forces the cornea inwards (ingoing phase) through applanation (first or ingoing applanation) into a concavity phase until it achieves the highest concavity. There is an oscillation period before the outgoing or returning phase. The cornea undergoes a second applanation before achieving its natural shape when there is a possible oscillation. The Corvis ST commercial system identifies the timing and corresponding pressure of the air puff at the first and second applanations and at the moments of highest concavity in the spatial deformation images. Intraocular pressure (IOP) is calculated based on the timing of the first applanation event. This initial IOP measurement is dependent of corneal resistance. The deformation amplitude is determined as the highest displacement of the apex in the highest concavity moment image. Other measured parameters are: the radius of curvature at highest concavity; applanation lengths and corneal velocities during ingoing and outgoing phases (**Fig. 2**).

Quantitative parameters can also be extracted from the captured set of OCT images, using custom routines (Dorronsoro *et al.*, 2012). These routines involve corneal edge detection algorithms and numerical computations of temporal and spatial deformation parameters (deformation amplitude, speeds and durations of deformation) to extract the parameters, the stable non-deformed cornea immediately before the deformation event is compared with the deformed cornea during the air puff event. A subset of used descriptive parameters are graphically illustrated in **Fig. 3**: corneal thickness (1, red) at the corneal apex before and

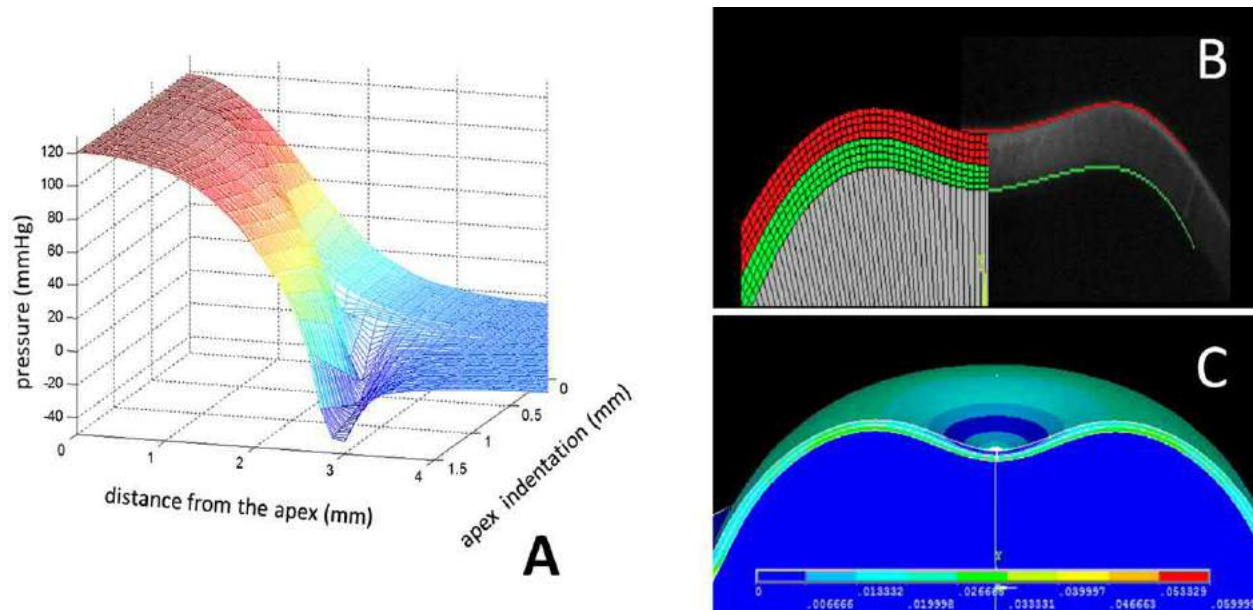


Fig. 2 A. Air-puff spatio-temporal profile (from pressure measurements and hydrodynamics calculations); B. Cornea at maximum air-puff spatial deformation: Experimental and Fine-Element model in inverse optimization. C. Simulated corneal deformation with retrieved biomechanical parameters. Data are from Kling, S., Bekesi, N., Dorronsoro, C., *et al.*, 2014. Corneal viscoelastic properties from finite-element analysis of in vivo air-puff deformation. PLOS ONE 9 (8), e104904.

after the deformation, and during maximum deformation; *deformation amplitude* (4, red) at the corneal apex (representing the maximum deformation depth); *overall deformation period* (9); *increasing deformation period* (7); *decreasing deformation period* (8); *mean increasing and decreasing deformation speeds* (2 and 6, black); *peak increasing and decreasing deformation speeds* (3 and 5, orange) (maximum positive and negative slopes of the deformation curve, represented as orange lines).

The parameters above were obtained from high-speed A-scan measurements at the corneal apex, a measurement mode available in both described approaches. Additional information can be extracted from the dynamic B-scan measurements (along the horizontal meridian; Fig. 3), providing a direct view of the dynamic spatial deformation of the cornea. Analysis of the B-scans during the deformation event could theoretically provide similar results on the apex dynamics than the A-scan measurements. However, the temporal sampling in those measurements is lower resulting in a lower precision of the extracted temporal parameters. The B-scans were also used to obtain additional deformation parameters at maximum deformation: *diameter of the deformed zone* (10, green), *displaced corneal volume* at the moment of peak deformation (11, gray), *non-deformed corneal radius* (12) and *peak deformed corneal radius* (13).

Corneal Mechanical Properties From Air Puff Corneal Dynamic Imaging

Although the air puff corneal deformation can be characterized using several temporal and spatial deformation parameters deformation is affected not only by mechanical properties, but also by geometrical parameters (corneal thickness and, to a lesser extent, corneal curvature) as well as Intraocular Pressure (IOP) (Kling and Marcos, 2013a,b,c). Finite-Element-Models (FEM) have been recently been applied to reconstruct the inherent mechanical parameters of corneal tissue from Scheimpflug-based temporal and spatial deformation imaging, and on measurements of the temporal and spatial properties of the air-puff (Fig. 1) (Marcos *et al.*, 2014). To simulate the experimental conditions accordingly, the pressure characteristics provided by the air-puff system were determined. The temporal pressure profile was measured experimentally using a pressure sensor and then a computational fluid dynamics simulation was performed to determine the spatial pressure profile, and its change during the deformation event. The corneal tissue was modeled as a linear viscoelastic material, in 2-D. Inverse modeling, with a multi-step optimization procedure, was performed in order to find the biomechanical parameter set that best represented the different experimental conditions. Retrieved parameters include the modulus of elasticity, viscoelastic relative modulus and relaxation constant, and represent the suitable descriptive parameters to monitor corneal biomechanical properties and to predict the mechanical response to modeling.

The reconstruction of the inherent material properties from air-puff corneal deformation has been validated using phantom corneas of different hydrogel polymer materials and thicknesses and on porcine corneas, at a constant IOP of 15 mmHg (Bekesi *et al.*, 2016a,b). Uniaxial tensile tests were performed on the model cornea materials as well as on corneal strips, and the results were compared to stress-strain simulations assuming the reconstructed material parameters. The simulated stress-strain curves of the studied hydrogel corneal materials fitted well the experimental stress-strain curves from uniaxial extensimetry. Equivalent Young's moduli of the reconstructed material properties from air-puff were 0.31, 0.58 and 0.48 MPa for the three polymer materials respectively which differed <1% from those obtained from extensimetry. The simulations of the same material but

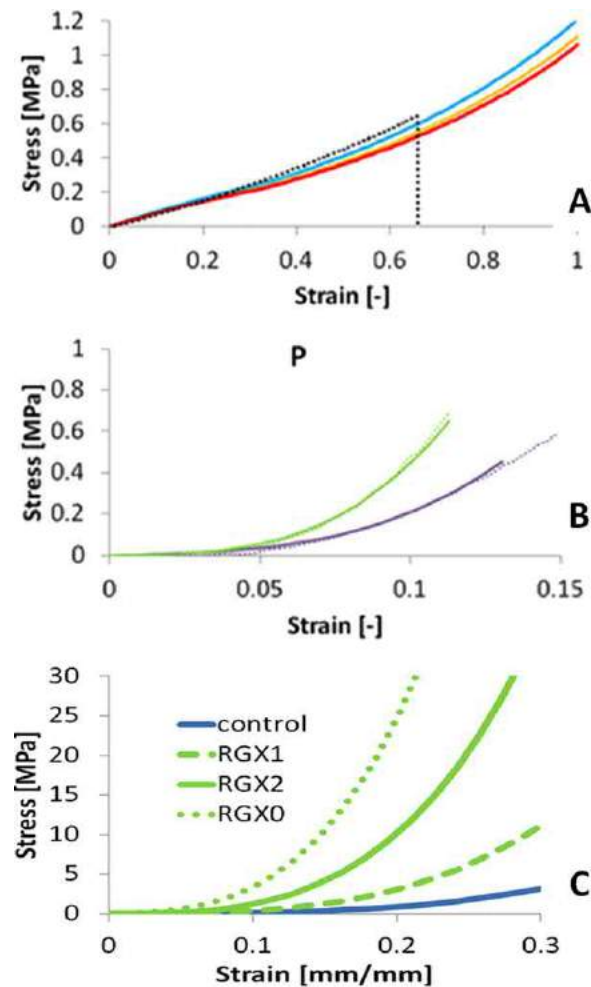


Fig. 3 Stress-strain curves in phantom, porcine corneas and rabbit corneas. A. Obtained from extensimetry on polymer strips (dotted line) and from air-puff corneal deformation imaging and inverse optimization. Red, Yellow and Red lines correspond to phantom corneas (mounted in an artificial eye) of the same material but different thicknesses B. Obtained from extensimetry on corneal strips, and from air-puff corneal deformation imaging and inverse optimization. Green and Purple correspond to two different corneas. C. Obtained from air-puff corneal deformation imaging and inverse optimization in untreated rabbit corneas, rabbit corneas 1-month and 2-month following cross-linking treatment (RGX, Rose Bengal Cross Linking) in vivo, rabbit corneas following cross-linking ex vivo (measurements immediately after enucleation and treatment). Data are from Bekesi, N., Kochevar, I.E., Marcos, S., *et al.*, 2016a. Corneal biomechanical response following collagen cross-linking with rose bengal–green light and riboflavin–UVA. *Investigative Ophthalmology & Visual Science* 57, 992–1001; Bekesi, N., Dorronsoro, C., de la Hoz, A., Marcos, S., 2016b. Material properties from air puff corneal deformation by numerical simulations on model corneas. *PLOS ONE* 11 (10), e0165669 and Bekesi, N., *et al.*, 2017. Biomechanical changes after in vivo collagen cross-linking with rose bengal–green light and riboflavin–UVA. *Investigative Ophthalmology & Visual Science* 58, 1612–1620.

different thickness resulted in similar reconstructed material properties. The air-puff reconstructed average equivalent Young's modulus of the porcine corneas was 1.3 MPa, within 18% of that obtained from extensimetry.

Clinical and Experimental Applications of Air Puff Deformation Imaging: Cross-Linking

Corneal cross-linking (CXL) is an emerging treatment to halt the progression of ectatic disorders by crosslinking collagen fibrils and glycoproteins in the corneal stroma. Reconstruction of corneal mechanical parameters from air puff corneal deformation has been applied to the characterization of different modalities of CXL (Riboflavin UVA UVX and Rose Bengal Green light RGX) on rabbit eyes ex vivo and in vivo (under constant IOP) (Bekesi *et al.*, 2016b, 2017) as well as in a human eye in vivo (Kling *et al.*, 2014).

Corneal deformation amplitude decreased significantly after both CXL methods. The material parameters obtained from inverse modeling were consistent with corneal stiffening after both RGX and UVX. Within the treated corneal volume, the elasticity was found to decrease by a factor of 11 after RGX and by a factor of 6.25 after UVX following treatments ex vivo. Conversely, elasticity increased by factors of 3.4 (RGX) and 1.7 (UVX) 1 month and by factors of 10.7 (RGX) and 7.3 (UVX)

2 months after treatment in vivo. The equivalent modulus of elasticity of rabbits eye ranged from 1.43 MPa (untreated eyes) and 15 MPa (2-month post RGX, treated area). The estimated equivalent modulus of elasticity in a human eye was 0.45 MPa. Other estimates include viscoelastic properties, such as relative modulus and time relaxation constants. The viscoelastic effect was evidenced by the remaining deformation (between 18.7% and 24.1% of the overall deformation at different IOPs), which gradually decreased with time after the air-puff event. The viscoelasticity of the cross-linked porcine corneas ex vivo was described by a relative modulus of 0.7 and a relaxation time constant of 1.5 ms, and contributed with 46% to the maximal apex indentation.

Phase-Sensitive Optical Coherence Elastography

Principles and Implementations of Optical Coherence Elastography

The interferometric nature of OCT allows phase measurement for the detection of tissue deformation at micro- and nano-scales in amplitude and from static to ultrasound frequencies in the time scale. This capability has been harnessed for elastography, also known as optical coherence elastography (OCE), by employing appropriate mechanical stimuli. In the previous section, we have described an air-puff stimulus to induce macroscopic corneal deformation. Stimulus with reduced amplitudes can induce microscopic corneal deformation, which can be readily measured by phase-sensitive OCT. Compared to macroscopic air-puff, microscopic perturbation involves proportionally smaller strain changes in the tissue and less IOP changes. The smaller perturbations avoid the need of nonlinear mechanical analysis and minimize secondary effects, such as corneal tension and scleral expansion, induced by the IOP change. In this section, we describe a variety of elastography techniques based on phase measurement in OCT.

Early OCE studies used intensity-based measurement to track displacement of structures under an applied load. Intensity-based OCE has been applied for ophthalmic applications (Ford *et al.*, 2006, 2011, 2014; Torricelli *et al.*, 2014; Armstrong *et al.*, 2013), in which displacement vector is calculated from cross-correlation between initial and mechanically perturbed images. However, nowadays phase-sensitive detection has become a far more popular choice for OCE because it provides better displacement sensitivity compared to the intensity-based method.

In Fourier domain OCT, the phase of each image pixel is readily measured from the Fourier transform of the measured interferometric signal. The phase change $\Delta\phi$ can be directly related to the change in optical path length delay ΔL by a simple relationship: $\Delta\phi(z, t) = \frac{4\pi}{\lambda} \Delta L(z, t)$. Converting the optical delay to a true geometrical displacement requires knowledge of or assumptions about the refractive index distribution $n(z)$. It is also important to take into account the discontinuity of the refractive index at the corneal surface (Song *et al.*, 2013). The cornea is generally simplified as a two-layer structure consisting of a coupling medium of index n_0 (often the air) and a tissue layer of index n_1 . Then, the geometrical displacement is calculated from the phase profile within the tissue with respect to the phase at the corneal surface $\Delta\phi_0$ (Wang and Larin, 2014):

$$\Delta u_z(z, t) = \frac{\lambda}{4\pi n_1} \left(\Delta\phi(z, t) + \Delta\phi_0(t) \frac{n_1 - n_0}{n_0} \right)$$

The amplitude sensitivity, $\sigma_{\Delta L}$, of shot-noise-limited, phase-sensitive measurement is given by the signal-to-noise ratio (SNR) via $\sigma_{\Delta L} = \frac{\lambda}{4\pi} \frac{1}{\sqrt{\text{SNR}}}$. For example, for regions with SNR of 40 dB in OCT images, a displacement sensitivity better than a thousandth of the optical wavelength (i.e., sub-nanometer) is achieved.

Other noise sources, such as mechanical vibrations of optical components in an OCT system can increase the phase noise above the fundamental SNR limit. Some of the noise may be reduced by common-path interferometer designs (Choma *et al.*, 2005). Wavelength swept lasers in swept-source OCT tend to introduce additional noise to due phase jitter between wavelength sweep cycles and a data acquisition clock. A calibration mirror (Vakoc *et al.*, 2005) and real-time interferometric clock generation (Braaf *et al.*, 2011) have been developed to mitigate the swept laser-induced noise. Phase-sensitive detection can only measure the axial component of the displacement parallel to the axis of an OCT beam, and phase measurement may present phase-wrapping errors when the displacement amplitude is greater than the optical wavelength. Techniques to get around these problems have also been put forward.

Depending on the type of mechanical perturbations the applied stimuli induce on the cornea, we may sort elastography techniques in three categories: static, standing waves (i.e., vibrational resonance), and traveling (acoustic) waves, which are illustrated in Fig. 4. All approaches share common steps in operation. First, a mechanical force is applied to the tissue to induce deformation. Second, OCT is used to measure the deformation. And, finally a biomechanical model is applied to relate the observed deformation to tissue stiffness.

Static OCE was the first proposed approach for general OCE (Schmitt, 1998) and for ophthalmic applications (Ford *et al.*, 2006). It is based on the simple concept that softer areas deform more than stiffer areas. The well-known static mechanics is applied to derive tissue elasticity from measured strain for given external stress. Typically, the external stress is applied by compressing the tissue with a contact device, such as an objective lens or a transparent plate. Although this simple mechanical method is widely used for general measurements of ex vivo tissue samples, it is not as commonly used for applications to the eye because a noncontact approach is preferred for clinical diagnosis in patients. Nonetheless, several studies using ex vivo rabbit and human eyes showed the proof of concept by using a gonioscopy lens to compress the cornea (Ford *et al.*, 2011, 2014). Another

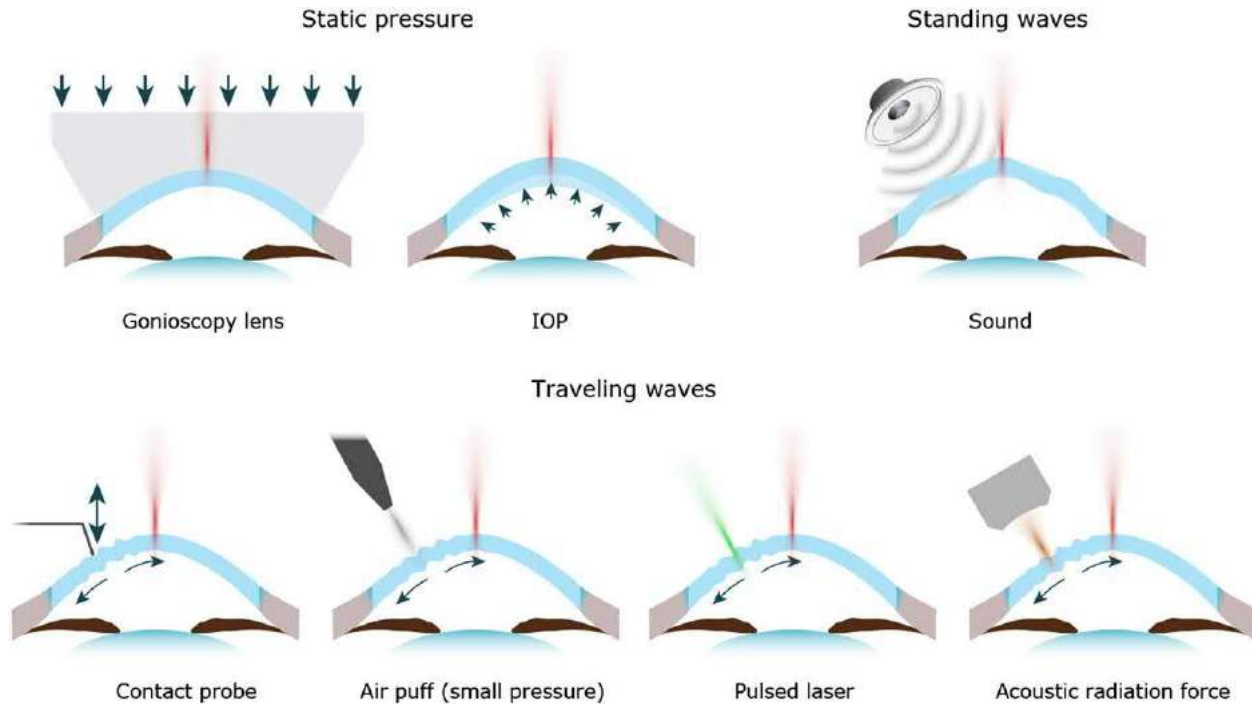


Fig. 4 OCT elastography techniques. Various mechanical loading methods have been employed to induce different types of perturbations ranging from static deformation and vibrational resonance to propagating acoustic waves.

variant of static OCE is to externally modulate IOP as a loading mechanism for cornea elastography. Proof-of-concept experiments were performed on ex vivo corneal explants (Ford *et al.*, 2006, 2016). Although this invasive technique is unlikely to be applicable in clinical settings, it could be useful for ex vivo measurements and animal studies.

Traveling wave OCE has been extensively used for ocular elastography. The basic principle of operation was adapted from shear wave elastography in ultrasound imaging. It uses a localized energy source to excite mechanical acoustic waves, which then propagate in the tissue. The wave parameters, such as the amplitude, velocity, and decay rate, are extracted from OCT measurements and related to the mechanical properties of the tissue. Various non-contact approaches are available to excite the traveling waves in the eye. As a direct analogue of ultrasound shear wave elastography, amplitude-modulated ultrasound was used to excite shear waves via acoustic radiation force (ARF), which are then detected by phase-sensitive OCT (Nguyen *et al.*, 2015). Intense pulsed laser light can excite acoustic waves via the photo-acoustic effect (Li *et al.*, 2012). However, both of these techniques require high ultrasound and optical energy to generate measurable acoustic amplitudes, raising safety concern to the eye. Some techniques to mitigate this issue have been proposed, such as numerical ultrasound pulse compression (Nguyen *et al.*, 2015) and surface reflection-induced ARF (Ambroziński *et al.*, 2016).

As described in the previous section, large-amplitude air puffs can induce millimeter-scale deformations in the cornea but such large deformations involve the non-linear response of collagen fibrils in the corneal stroma and thus add complexity to the mechanical analysis. Air puffs with a significantly lowered amplitude and shorter duration have been introduced to OCE as a way to excite traveling acoustic waves in the cornea (Wang *et al.*, 2013) and have extensively been investigated.

An advantage of traveling-wave OCE is that analytical relationships between Young's modulus and the acoustic wave velocity can be established. A few expressions have been derived, which are valid under different circumstances depending on the frequency and location of the excitation force. Vibrations generated at the surface of the cornea were often modeled as Rayleigh surface waves (Li *et al.*, 2012; Wang *et al.*, 2013), and Young's modulus E is obtained from the wave group velocity c_g using $E = \frac{2\rho(1+\nu)^3}{(0.87+1.12\nu)^2} c_g^2$. This formula is accurate only in a high-frequency limit in which the acoustic wavelength is much smaller than the corneal thickness. In more general cases, the impact of the thickness and curvature of the cornea, as well as the aqueous humor, have been investigated in detail (Han *et al.*, 2015b,c; Aglyamov *et al.*, 2015). A modified Rayleigh-Lamb wave model that accounts for the layered structure and viscosity of the cornea has been developed (Han *et al.*, 2015a, 2017).

Standing wave OCE is based on modal analysis, with the rationale that for a given structure geometry, a stiffer material will cause a higher resonance frequency. A frequency-domain method using sound wave stimulus has been proposed, in which the vibrational modes of the cornea or the whole eye globe are analyzed by sweeping the acoustic frequency. The eye has a fundamental resonance frequency of about 100 Hz (Akca *et al.*, 2015). As in the case of the air-puff methods, to extract elastic modulus numerical simulations would be necessary, which takes into account the entire eye globe as well as the cornea.

Applications of Optical Coherence Elastography in Ocular Tissues

Corneal biomechanics has been the major application area of OCE. This is in part because cornea is readily accessible for the induction of a mechanical stimulus but more due to the fact that the mechanical properties of the cornea are involved in number of pathological processes and are considered an important factor to consider in refractive surgeries and some vision treatments. The first demonstrations of cornea OCE were performed in the mid-2000s (Ford *et al.*, 2006; De La Torre-Ibarra *et al.*, 2006). But the interest for ophthalmic OCE increased considerably in the early 2010s with the progress of phase-sensitive OCE (Li *et al.*, 2012; Manapuram *et al.*, 2012; Ford *et al.*, 2011).

OCE can measure the biomechanical changes induced by CXL (Ford *et al.*, 2014; Singh *et al.*, 2016). In particular, OCE has been used to quantify the effects of standard epithelium-off CXL procedures for various transepithelial crosslinking reagents and femtosecond laser-assisted riboflavin-UVA CXL (Armstrong *et al.*, 2013). A follow-up study (Torricelli *et al.*, 2014) showed that benzalkonium chloride-ethylenediaminetetraacetic acid (BKC-EDTA) induced a larger stiffening effect than the standard epithelium-off riboflavin-UVA procedure. More recently, stiffness mapping of customized CXL treated corneas has been demonstrated (Singh *et al.*, 2017).

The age dependence of corneal stiffness was examined by using an oscillating tip as the source of mechanical waves on live mice (Manapuram *et al.*, 2012). The observed age-related increase in corneal stiffness with age was in agreement with previous measurements done by conventional mechanical measurements (Cartwright *et al.*, 2011; Elsheikh *et al.*, 2007). The relaxation rate after a mechanical stimulus was also shown to increase with age (Li *et al.*, 2014).

Besides the cornea, phase-sensitive OCE has been used to characterize the mechanical properties of the posterior sclera in ex vivo porcine eyes (Singh *et al.*, 2017) and measure the mechanical properties of crystalline lens in rabbit eyes (Wu *et al.*, 2015). Similarly, retinal OCE was performed on ex vivo porcine eyeballs (Song *et al.*, 2015).

Brillouin Microscopy

Brillouin Microscopy Technique

Brillouin microscopy takes advantage of a light scattering phenomenon first described by Brillouin (1922). When incident light hits a sample, it interacts with microscopic sound waves (acoustic phonons) within the material so that the scattered light has a slightly different frequency than the incidence light. Such acoustic phonons are inherently present in any material and are specific to the mechanical properties of the material. As such, the Brillouin frequency shift between incident and scattered light, can provide local mechanical information of the sample (Fig. 5). The Brillouin frequency shift is described by the equation:

$$(v)_B = \pm \frac{2n}{\lambda_p} \sqrt{\frac{M'}{\rho}} \cos\left(\frac{\theta}{2}\right) \quad (1)$$

In Eq. (1), n is the refractive index of the sampled material, M' is the real part of the longitudinal elastic modulus of the material, ρ is the density of the material and θ is the angle between incident and scattered optical radiation; importantly.

Because the light scattering is driven by acoustic waves, the changes in wavelength between incident and scattered light are in the 1–10 picometer wavelength range, or GHz frequency range. This is more than 1000 times smaller than changes observed with Raman scattering or fluorescence. Thus, to detect Brillouin signals, ordinary spectrometers cannot be used as they lack the spectral resolution to resolve Brillouin signal from input laser light and the spectral extinction to detect Brillouin scattered light in the presence of strong light with similar wavelength. A solution to this problem was first developed by JR Sandercock who developed a spectrometer using multiple-pass scanning Fabry–Perot (FP) interferometers (Sandercock, 1975). Sandercock's spectrometer has been used since the 1970s enabling noncontact mechanical analysis of materials in a variety of fields (Dil, 1982) including the first

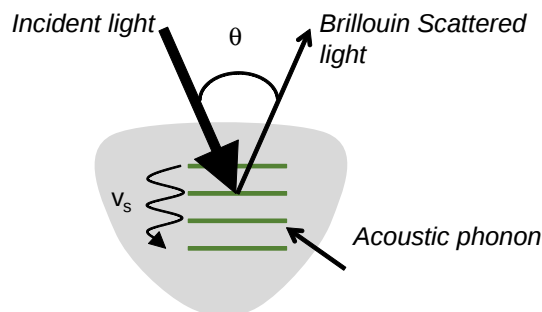


Fig. 5 Brillouin Principle: When incident light hits a sample, it interacts with microscopic sound waves (acoustic phonons) within the material so that the scattered light has a slightly different frequency than the incidence light. Such acoustic phonons are inherently present in any material and are specific to the mechanical properties of the material.

characterization of ocular tissue (Vaughan and Randall, 1980). However, FP spectrometers are characterized by slow acquisition times (~ 10 min to hours per spectrum), therefore their applications to biomedical sciences is difficult.

In 2008, the introduction of virtually-imaged phased array (VIPA) etalons (Shirasaki, 1996) in lieu of FP interferometers greatly improved acquisition time (Scarcelli *et al.*, 2008). The VIPA etalon is similar to a FP etalon with few key differences: (1) the etalon is tilted and input light is diverging so that different incidence angle find different resonances; this enables to collect all the spectrum in one shot (Koski *et al.*, 2002); (2) the first surface of a VIPA is totally reflective and only allows light to enter the interferometer through a narrow anti-reflection-coated window; this prevents the formation of a reflected interference pattern which translates in about two-order of magnitude higher throughput. VIPA-based Brillouin spectrometers have been further improved in recent years thus enabling the transition from single-point Brillouin spectroscopy to an imaging modality through the combination with confocal microscopes (Scarcelli *et al.*, 2007a,b, 2008, 2011). This configuration has now been used for several years to measure tissue, cells and sub-cellular structures (Scarcelli *et al.*, 2015a; Scarcelli and Yun, 2012; Zhang *et al.* 2017a,b).

Brillouin modulus and relationship with standard mechanical properties

From a mechanical standpoint, Brillouin spectroscopy probes the response of a sample to a one-dimensional oscillatory strain of hypersonic ($\sim$ GHz) frequency. The measured frequency shift is related to the longitudinal modulus of material, as shown in Eq. (1). However, in Eq. (1) both index of refraction and density of material are not constant within a sample and from sample to sample therefore the extraction of Brillouin longitudinal modulus needs to be carefully examined. In biological material such as cells and tissues, density and refractive index are strongly correlated; for ocular tissues approximating density and index to constant values yields only a small percentage error. Therefore, local, three-dimensional, direct measurement of the longitudinal elastic modulus of material can be obtained by measuring the Brillouin spectrum. In crystalline materials, the longitudinal modulus (M') is related to traditional Young's modulus (E') in a straightforward manner through the Poisson's ratio (σ) via $M' = E'(1 - \sigma)/(1 + \sigma)(1 - 2\sigma)$. However, in biological material the fundamental relationship to Young's or shear moduli is not yet established because of the large difference in frequency of the test (quasi-static vs GHz) which is relevant due to the frequency-dependent moduli of most biomaterials (Mofrad and Kamm, 2006) and because of the near incompressibility of biological tissue due to the high water content.

For both cornea and lens tissue, an empirical correlation between Brillouin modulus and traditional Young's or shear modulus has been established experimentally. Such correlation has been found to be linear in log-log scale (Scarcelli *et al.*, 2011), which suggests a power-law relationship, consistent with what was previously found in tissue, polymers and cells (Deng *et al.*, 2006; Fabry *et al.*, 2001; Gittes and MacKintosh, 1998; Jonas *et al.*, 2008; Duck, 1990; Holm and Sinkus, 2010).

This offers the possibility of quantitatively measure cornea and lens mechanical properties after proper calibration. Specifically, for a given material we can write: $\log(M') = a \log(E') + b$, where the coefficients a and b are intrinsic properties of material that have a fundamental basis on atomic and molecular interactions. We can also write $\frac{\Delta M'}{M'} = a \frac{\Delta E'}{E'}$ where $\Delta M'$, $\Delta E'$ are respective derivatives or variations. This allows to estimate the sensitivity of the instrument to actual variations in Young's modulus. Current Brillouin instruments reach frequency sensitivity of ± 1 MHz/ $\sqrt{\text{Hz}}$, which corresponds to a relative error in longitudinal moduli, $\Delta M'/M' \approx \pm 0.3\%$ for an integration time of 0.1 s. Using for example $a=0.03$ measured with porcine corneal tissues, we determine that the Brillouin measurement can detect a difference in Young's modulus, $\Delta E'/E'$ of about 9% at an integration time of 0.1 s.

Applications of Brillouin Microscopy in the Cornea

Recent Brillouin measurements for the analysis of keratoconus, corneal collagen crosslinking and refractive surgery are promising (Fig. 6) (Scarcelli *et al.*, 2014, 2015b). For keratoconus, both ex vivo and in vivo human experiments were performed. For ex vivo

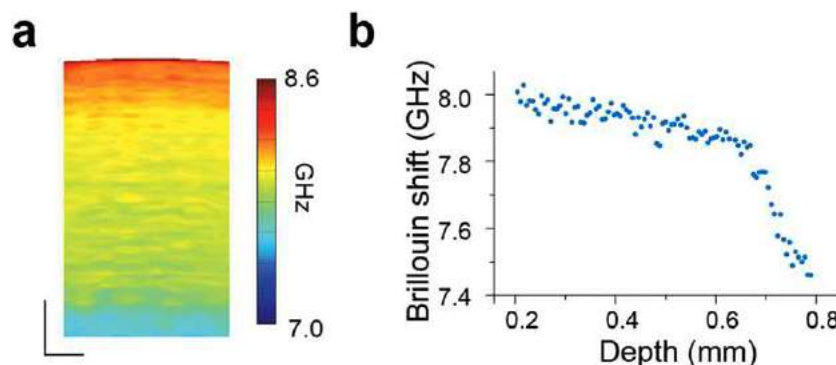


Fig. 6 Cornea Applications: (a) Representative Brillouin cross-section of a bovine cornea ex vivo reveals depth-dependence of Brillouin shift. Scale Bar 200 μm . (b) Representative depth profile of a bovine cornea ex vivo. A high correlation in depth is observed between regions of higher Brillouin shift and regions with higher content and more branching of corneal collagen.

analysis, normal tissue samples were obtained from donor corneas used in Descemet's stripping endothelial keratoplasty (DSEK); keratoconic tissue samples were obtained from deep anterior lamellar keratoplasty (DALK) procedure in advanced keratoconus patients. For the in vivo study, patients with advanced keratoconus were measured just before undergoing DALK surgery and compared to healthy volunteers. Ex vivo and in vivo analysis are consistent with each other: keratoconic corneas had overall lower Brillouin shift, thus lower elastic modulus. Thanks to Brillouin spatial resolution, we performed a spatially resolved analysis: the Brillouin shift of the normal corneas was observed to be relatively uniform across the central region of the cornea; while in keratoconus patients a strong spatial variation was measured with the cone showing the lowest Brillouin shift and indicating a large (up to ~70%) focal reduction in Young's modulus. Interestingly, away from the cone, the elastic modulus was very similar to normal cornea in both ex vivo and in vivo investigations.

For the analysis of corneal crosslinking, Brillouin has the potential to quantify the mechanical efficacy of various procedures. Because a universally recognized standard CXL protocol (Dresden) exists, we can use the log-log correlation to relate the changes of elastic modulus induced by CXL without calibration by comparing the unknown procedure to the standard Dresden protocol via the following equation:

$$\frac{\Delta M'_X}{\Delta M'_{Dresden}} = \frac{\Delta E'_X}{\Delta E'_{Dresden}}$$

Thus far, we have been able to use demonstrate Brillouin's ability to assess the efficacy of several cross-linking protocols while varying parameters of the CXL procedure (Reiss *et al.*, 2011). For example, corneal cross-linking currently requires an epithelium debridement prior to exposure to Riboflavin and UV-A. However, recent efforts have been given towards assessing the efficacy of the procedure without removing the epithelium (epi-On CXL) and, thereby, decreasing patient discomfort. The epi-on CXL procedure was estimated to induce about 33% of the stiffening done by the standard Dresden procedure. For comparison, the mechanical stiffening of the epi-on procedure was tested with gold-standard quasi-static rheology and was found to induce approximately 39% of the stiffening of epi-off CXL. Prior studies (Wollensak and Iomdina, 2009) found the mechanical efficacy of transepithelial CXL to be approximately 20% of standard CXL.

For the analysis of refractive surgery, we recently tested Brillouin's ability to assess corneal mechanical changes associated with LASIK surgery. LASIK surgery is expected to decrease the stiffness of the cornea due to the creation of a flap and the subsequent corneal ablation. To revert this potentially harmful effect, an accelerated CXL procedure has begun to be tested in combination with LASIK surgery. We used Brillouin microscopy to observe the properties of the cornea during the various stages of the LASIK/CXL combined procedure. After the LASIK flap was created, the cornea showed lower stiffness, especially in the anterior and central portions, where the flap was created. However, we observed no statistically significant difference after the flap-induced cornea was exposed to accelerated CXL. Therefore, using Brillouin microscopy, we were able to conclude the accelerated CXL procedure commonly paired with LASIK did not yield significant benefits towards increasing the compromised stiffness.

Applications of Brillouin Microscopy in the Crystalline Lens

The condition commonly known as presbyopia is related to the age-related loss of accommodation. Accommodation is the crucial biomechanical process of human vision that enables young persons to see near and distant objects by dynamically adjusting the focusing power of the eye. As age increases, the accommodation power drops. From a biomechanical point of view, this process is

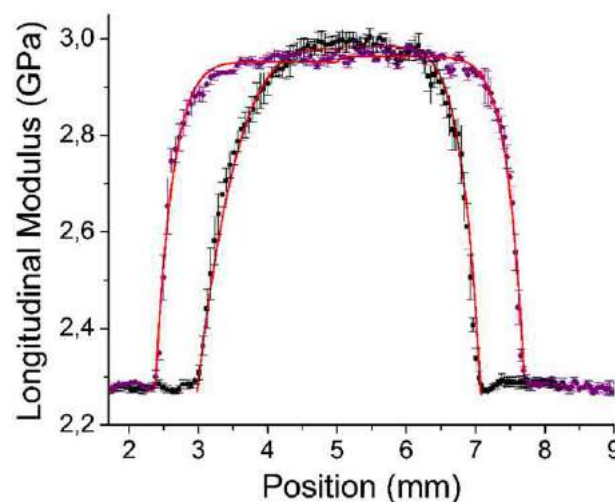


Fig. 7 Lens Applications: Brillouin microscopy can detect the crystalline lens longitudinal modulus with high-spatial resolution in vivo. Here shown the mean sagittal longitudinal elastic modulus in 28-(black square) and 60-(purple dot) year-old human lenses ($n=3$). The red curves are power fits of the lens profiles.

due to the inability of the eye muscles to significantly change the shape of the crystalline lens. It has long been suspected that the loss of accommodation power is due to the diminished capacity of ciliary muscles to contract the lens with age. However, recent data indicate that ciliary muscles maintain ability to contract after the loss of accommodation (Ostrin and Glasser, 2007). Other potential candidate factors for presbyopia onset are age-dependent changes of the capsule and zonular fibers; but experiments have found too little changes to account for the decrease in accommodation power (Brian and Melinda, 2009; Fisher, 1969). Thus, the crystalline lens seems to be the main factor in the loss of accommodation; specifically age-related increased stiffness of the crystalline lens is suspected as the primary cause of presbyopia: as age progresses, the lens tissue stiffen which diminishes its ability to change shape (Glasser and Campbell, 1998, 1999).

Brillouin microscopy is a promising technique to characterize lens mechanical properties. In particular, the high 3D spatial resolution of Brillouin microscopy enables to characterize the spatial distribution of the longitudinal modulus within the crystalline lens (Fig. 7). In animal lenses, Brillouin microscopy revealed a dramatic age-dependent increase in the longitudinal modulus of the murine lens in vivo (Scarcelli *et al.*, 2011) and in porcine/bovine lenses ex vivo. In human lens, both the spatial distribution and the age dependence have been characterized (Besner *et al.*, 2016). In terms of spatial distribution, at all ages, the central portion of the lens was stiffer than the periphery, a result that was confirmed by ex vivo experiments using traditional standard shear rheometry. This result is consistent with some previous studies (Scarcelli *et al.*, 2011; Reiss *et al.*, 2011, 2012; Randall and Vaughan, 1982; Bailey *et al.*, 2010; Yoon *et al.*, 2013; Hollman *et al.*, 2007; Pau and Kranz, 1991) but not others (Fisher, 1971; Heys *et al.*, 2004; Weeber *et al.*, 2007; Wilde *et al.*, 2012). In terms of age dependence, Brillouin analysis did not find a statistically significant change of the peak longitudinal modulus with age, which confirmed previous Brillouin analysis performed ex vivo with FP interferometers (Bailey *et al.*, 2010). The most prominent age-related feature revealed by Brillouin was the increase of the stiff portion of the lens. Thus, Brillouin analysis does not support the previously reported “massive” increase of lens modulus as main factor to loss of accommodation (Fisher, 1971; Heys *et al.*, 2004; Weeber *et al.*, 2007; Wilde *et al.*, 2012). However, although more calibration mechanical studies are needed, with the given profiles and log-log correlation, the increase of the stiff portion of the lens would yield a 2-fold increase in the overall stiffness of the lens with age, which is consistent with previous ex vivo data (Glasser and Campbell, 1999; Augusteyn *et al.*, 2011).

Towards a Clinical Brillouin Microscopy System

A widely-available clinical Brillouin device that could measure biomechanical properties of the cornea and lens could find widespread application in both ophthalmic research and in the clinic. Currently available Brillouin microscopes have reached the performance levels to enable clinical testing and clinical studies to test the potential of Brillouin microscopy for keratoconus, crosslinking and presbyopia are currently on-going. For both cornea and lens applications, Brillouin microscopy has already shown biologically- and clinically-relevant information.

However, the success of Brillouin microscopy is highly dependent on its technology development. At the current mechanical sensitivities of in vivo instruments, Brillouin technology offers only ~10 discernable values for the stratification of patients of varying keratoconus stages. Also, current imaging sessions to collect a full Brillouin map of cornea or lens tissue may take 10–30 min. Finally, both the cost of the Brillouin microscope and the complexity of the setup represent an obstacle for widespread adoption in the clinic. Technological improvements are needed to address all of these issues. On this front, significant improvements have been shown already, although still at proof-of-principle stages: improved spectral selection at low loss has the potential to increase the throughput of Brillouin spectrometers (Fiore *et al.*, 2016; Shao *et al.*, 2016; Edrei *et al.*, 2017; Berghaus *et al.*, 2015a,b); novel multiplexed geometries could potentially speed-up the acquisition time by one to two order of magnitudes (Zhang *et al.*, 2016); standardized procedures for spectrometer alignment have been developed (Berghaus *et al.*, 2015a,b); the sensitivity of clinical instruments is currently limited by drifts due to temperature changes or laser instabilities, which could be removed with frequency locking or absolute frequency standards. Finally, improved speed and sensitivity could come from using higher laser power as all the experiments reported so far have used smaller illumination levels than the ones used in other clinical ophthalmic instruments. As a result, there is reason for optimism that Brillouin technology could be adopted in the clinic, focused engineering efforts aimed at making Brillouin microscopy faster, more sensitive, portable and easy to operate by non-experts will be crucial for the widespread use of the technology.

See also: Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

References

- Aglyamov, S.R., *et al.*, 2015. The dynamic deformation of a layered viscoelastic medium under surface excitation. *Physics in Medicine and Biology* 60 (11), 4295–4312.
- Akca, B.I., *et al.*, 2015. Observation of sound-induced corneal vibrational modes by optical coherence tomography. *Biomedical Optics Express* 6 (9), 3313.
- Alonso-Caneiro, D., Karnowski, K., Kaluzny, B.J., *et al.*, 2011. Assessment of corneal dynamics with high-speed swept source optical coherence tomography combined with an air puff system. *Optics Express* 19 (15), 14188–14199.
- Ambrosio, J.R., *et al.*, 2011. Evaluation of corneal shape and biomechanics before lasik. *International Ophthalmology Clinics* 51 (2), 11–38.
- Ambroziński, Ł., *et al.*, 2016. Air-coupled acoustic radiation force for non-contact generation of broadband mechanical waves in soft media. *Applied Physics Letters* 109 (4), 43701.
- Andreassen, T.T., Simonsen, A.H., Oxlund, H., 1980. Biomechanical properties of keratoconus and normal corneas. *Experimental Eye Research* 31 (4), 435–441.

- Armstrong, B.K., *et al.*, 2013. Biological and biomechanical responses to traditional epithelium-off and transepithelial riboflavin-UVA CXL techniques in rabbits. *Journal of Refractive Surgery* 29 (5), 332–341.
- Augusteyn, R., *et al.*, 2011. Age-dependence of the optomechanical responses of ex vivo human lenses from India and the USA, and the force required to produce these in a lens stretcher: The similarity to in vivo disaccommodation. *Vision Research* 51 (14), 1667–1745.
- Bailey, S.T., *et al.*, 2010. Light-scattering study of the normal human eye lens: Elastic properties and age dependence. *Transactions on Biomedical Engineering* 57 (12), 2910–2917.
- Bak-Nielsen, S., *et al.*, 2013. The rigidity of corneas before and after corneal cross-linking – As measured by Corvis® ST. *Investigative Ophthalmology & Visual Science* 54, 1613. (E-Abstract).
- Bekesi, N., *et al.*, 2017. Biomechanical changes after in vivo collagen cross-linking with rose bengal–green light and riboflavin-UVA. *Investigative Ophthalmology & Visual Science* 58, 1612–1620.
- Bekesi, N., Dorronsoro, C., de la Hoz, A., Marcos, S., 2016b. Material properties from air puff corneal deformation by numerical simulations on model corneas. *PLOS ONE* 11 (10), e0165669.
- Bekesi, N., Kochevar, I.E., Marcos, S., 2016a. Corneal biomechanical response following collagen cross-linking with rose bengal–green light and riboflavin-UVA. *Investigative Ophthalmology & Visual Science* 57, 992–1001.
- Berghaus, K.V., Yun, S.H., Scarcelli, G., 2015a. High speed sub-GHz spectrometer for Brillouin scattering analysis. *Jove-Journal of Visualized Experiments* 106, e53468.
- Berghaus, K., Zhang, J., Yun, S.H., Scarcelli, G., 2015b. High-finesse sub-GHz-resolution spectrometer employing VIPA etalons of different dispersion. *Optics Letters* 40 (19), 4436–4439.
- Besner, S., Scarcelli, G., Pineda, R., Yun, S.H., 2016. In vivo Brillouin analysis of the aging crystalline lens. *Investigative Ophthalmology & Visual Science* 57 (13), 5093–5100.
- Boyce, B.L., Jones, R.E., Nguyen, T.D., Grazier, J.M., 2007. Stress-controlled viscoelastic tensile response of bovine cornea. *Journal of Biomechanics* 40 (11), 2367–2376.
- Braaf, B., *et al.*, 2011. Phase-stabilized optical frequency domain imaging at 1- μ m for the measurement of blood flow in the human choroid. *Optics Express* 19 (22), 20886.
- Brian, P.D., Melinda, K.D., 2009. The lens capsule. *Experimental Eye Research* 88.
- Brillouin, L., 1922. Diffusion de la lumière et des rayons X par un corps transparent homogène influence de l'agitation thermique. *Annals of Physics* 17, 88–122.
- Cartwright, N.E.K., Tyrer, J.R., Marshall, J., 2011. Age-related differences in the elasticity of the human cornea. *Investigative Ophthalmology and Visual Science* 52 (7), 4324–4329.
- Chang, E.W., Kobler, J.B., Yun, S.H., 2002. Subnanometer optical coherence tomographic vibrography. *Optics Letters* 27 (17), 3678–3680.
- Choma, M.A., *et al.*, 2005. Spectral-domain phase microscopy. *Optics Letters* 30 (10), 1162–1164.
- De La Torre-Ibarra, M.H., Ruiz, P.D., Huntley, J.M., 2006. Double-shot depth-resolved displacement field measurement using phase-contrast spectral optical coherence tomography. *Optics Express* 14 (21), 9643–9656.
- Deng, L.H., *et al.*, 2006. Fast and slow dynamics of the cytoskeleton. *Nature Materials* 5 (8), 636–640.
- Denoyer, A., Ricaud, X., Van Went, C., Labbé, A., Baudouin, C., 2013. Influence of corneal biomechanical properties on surgically induced astigmatism in cataract surgery. *Journal of Cataract & Refractive Surgery* 39 (8), 1204–1210.
- Dil, J.G., 1982. Brillouin-scattering in condensed matter. *Reports on Progress in Physics* 45 (3), 285–334.
- Dorronsoro, C., Pascual, D., Pérez-Merino, P., Kling, S., Marcos, S., 2012. Dynamic OCT measurement of corneal deformation by an air puff in normal and cross-linked corneas. *Biomedical Optics Express* 3 (3), 473–487.
- Duck, F.A., 1990. *Physical Properties of Tissue*. London: Academic.
- Dupps, W.J., Wilson, S.E., 2006. Biomechanics and wound healing in the cornea. *Experimental Eye Research* 83 (4), 709–720.
- Edrei, E., Gather, M., Scarcelli, G., 2017. Integration of spectral coronagraphy within VIPA-based spectrometers for high extinction Brillouin imaging. *Optics Express* 25, 6895.
- Elsheikh, A., *et al.*, 2007. Assessment of corneal biomechanical properties and their variation with age. *Current Eye Research* 32 (1), 11–19.
- Fabry, B., Maksym, G.N., Butler, J.P., *et al.*, 2001. Scaling the microrheology of living cells. *Physical Review Letters* 87 (14), 1481021–1481024.
- Fiore, A., Zhang, J., Shao, P., Yun, S., Scarcelli, G., 2016. High-extinction virtually imaged phased array-based Brillouin spectroscopy of turbid biological media. *Applied Physics Letters* 108 (20), (ARTN 203701).
- Fisher, R., 1969. Elastic constants of the human lens capsule. *The Journal of Physiology* 201 (1), 1–20.
- Fisher, R.F., 1971. Elastic constants of human lens. *Journal of Physiology-London* 212 (1), 147–180.
- Ford, M., *et al.*, 2006. OCT corneal elastography by pressure-induced optical feature flow. In: *SPIE Proceedings* 6138, *Ophthalmic Technologies XVI*, 61380P.
- Ford, M.R., *et al.*, 2011. Method for optical coherence elastography of the cornea. *Journal of Biomedical Optics*. 16005.
- Ford, M.R., *et al.*, 2014. Serial biomechanical comparison of edematous, normal, and collagen crosslinked human donor corneas using optical coherence elastography. *Journal of Cataract & Refractive Surgery* 40 (6), 1041–1047.
- Fu, J., *et al.*, 2016. Depth-resolved full-field measurement of corneal deformation by optical coherence tomography and digital volume correlation. *Experimental Mechanics*.
- Gittes, F., MacKintosh, F.C., 1998. Dynamic shear modulus of a semiflexible polymer network. *Physical Review E* 58 (2), R1241–R1244.
- Glasser, A., Campbell, M., 1998. Presbyopia and the optical changes in the human crystalline lens with age. *Vision Research* 38 (2), 209–238.
- Glasser, A., Campbell, M., 1999. Biometric, optical and physical changes in the isolated human crystalline lens with age in relation to presbyopia. *Vision Research* 39 (11), 1991–2006.
- Grabner, G., *et al.*, 2005. Dynamic corneal imaging. *Journal of Cataract & Refractive Surgery* 31, 163–174.
- Han, Z., *et al.*, 2017. Optical coherence elastography assessment of corneal viscoelasticity with a modified Rayleigh–Lamb wave model. *Journal of the Mechanical Behavior of Biomedical Materials* 66, 87–94.
- Han, Z., Aglyamov, S.R., *et al.*, 2015a. Quantitative assessment of corneal viscoelasticity using optical coherence elastography and a modified Rayleigh–Lamb equation. *Journal of Biomedical Optics* 20 (2), 20501.
- Han, Z., Li, J., Singh, M., Aglyamov, S.R., *et al.*, 2015b. Analysis of the effects of curvature and thickness on elastic wave velocity in cornea-like structures by finite element modeling and optical coherence elastography. *Applied Physics Letters* 106 (23), 1–5.
- Han, Z., Li, J., Singh, M., Wu, C., *et al.*, 2015c. Quantitative methods for reconstructing tissue biomechanical properties in optical coherence elastography: A comparison study. *Physics in Medicine and Biology* 60 (9), 3531–3547.
- Heys, K.R., Cram, S.L., Truscott, R.J.W., 2004. Massive increase in the stiffness of the human lens nucleus with age: The basis for presbyopia. *Molecular Vision* 10, 956–963.
- Hollman, K.W., O'Donnell, M., Erpelding, T.N., 2007. Mapping elasticity in human lenses using bubble-based acoustic radiation force. *Experimental Eye Research* 85 (6), 890–893.
- Holm, S., Sinkus, R., 2010. A unifying fractional wave equation for compressional and shear waves. *Journal of the Acoustical Society of America* 127 (1), 542–548.
- Jonas, M., Huang, H.D., Kamm, R.D., So, P.T.C., 2008. Fast fluorescence laser tracking microrheology, II: Quantitative studies of cytoskeletal mechanotransduction. *Biophysical Journal* 95 (2), 895–909.
- Karnowski, K., Kaluzny, B.J., Szkulmowski, M., *et al.*, 2011. Corneal topography with high-speed swept source OCT in clinical examination. *Biomedical Optics Express* 2 (9), 2709–2720.
- Kling, S., Bekesi, N., Dorronsoro, C., Pascual, D., Marcos, S., 2014. Corneal viscoelastic properties from finite-element analysis of in vivo air-puff deformation. *PLOS ONE* 9 (8), e104904.
- Kling, S., Ginis, H., Marcos, S., 2012. Corneal biomechanical properties from two-dimensional corneal flap extensimetry: Application to UV-Riboflavin cross-linking. *Investigative Ophthalmology and Vision Sciences* 53 (8), 5010–5015.

- Kling, S., Marcos, S., 2013a. Effect of hydration state and storage media on corneal biomechanical response from in vitro inflation tests. *Journal of Refractive Surgery* 29 (7), 490–497.
- Kling, S., Marcos, S., 2013b. Finite element modeling of intrastromal ring segment implantation into a hyperelastic cornea. *Investigative Ophthalmology and Vision Sciences* 54 (1), 881–889.
- Kling, S., Marcos, S., 2013c. Contributing factors to corneal deformation in air puff measurements. *Investigative Ophthalmology and Vision Sciences* 54 (7), 3961–3968.
- Kling, S., Remón, L., Pérez-Escudero, A., *et al.*, 2010. Corneal biomechanical changes after collagen cross-linking from porcine eye inflation. *Investigative Ophthalmology and Vision Sciences* 51 (8), 3961–3968.
- Koski, K.J., Muller, J., Hochheimer, H.D., Yarger, J.L., 2002. High pressure angle-dispersive Brillouin spectroscopy: A technique for determining acoustic velocities and attenuations in liquids and solids. *Review of Scientific Instruments* 73 (3), 1235–1241.
- Li, C., *et al.*, 2012. Noncontact all-optical measurement of corneal elasticity. *Optics Letters* 37 (10), 1625–1627.
- Li, J., *et al.*, 2014. Air-pulse OCE for assessment of age-related changes in mouse cornea in vivo. *Laser Physics Letters* 11 (6), 65601.
- Litwiler, D., Lee, S., Kolipaka, A., Mariappan, Y., Glaser, K., *et al.*, 2010. MR elastography of the ex vivo bovine globe. *Journal of Magnetic Resonance Imaging* 32 (1), 44–51.
- Liu, J., Roberts, C.J., 2005. Influence of corneal biomechanical properties on intraocular pressure measurement: Quantitative analysis. *Journal of Cataract & Refractive Surgery* 31 (1), 146–155.
- Manapuram, R.K., *et al.*, 2012. In vivo estimation of elastic wave parameters using phase-stabilized swept source optical coherence elastography. *JBO Letters* 17 (10), 15–18. Available at: <https://biomedicaloptics.spiedigitallibrary.org/article.aspx?articleid=1367083>.
- Marcos, S., Kling, S., Bekesi, N., Dorronsoro, C., 2014. Corneal biomechanical properties from air-puff corneal deformation imaging. In: *SPIE Proceedings 8946, Optical Elastography and Tissue Biomechanics*, 894609.
- McBrien, N.A., Jobling, A.I., 2009. Biomechanics of the sclera in myopia: Extracellular and cellular factors. *Optometry and Vision Science* 86 (1), E23–E30.
- Mofrad, M.R.K., Kamm, R.D., 2006. *Cytoskeletal Mechanics*. New York: Cambridge University Press.
- Muthupillai, R., *et al.*, 1995. Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. *Science* 269 (5232), 1854–1857.
- Nguyen, T.-M., *et al.*, 2015. Shear wave elastography using amplitude-modulated acoustic radiation force and phase-sensitive optical coherence tomography. *Journal of Biomedical Optics* 20 (1), 16001.
- Ortiz, S., Perez-Merino, P., Alejandre, N., *et al.*, 2012. Quantitative OCT-based corneal topography in keratoconus with intracorneal ring segments. *Biomedical Optics Express* 3, 814–882.
- Ostrin, L.A., Glasser, A., 2007. Edinger-Westphal and pharmacologically stimulated accommodative refractive changes and lens and ciliary process movements in rhesus monkeys. *Experimental Eye Research* 84 (2), 302–313.
- Pau, H., Kranz, J., 1991. The increasing sclerosis of the human lens with age and its relevance to accommodation and presbyopia. *Graefes Archive for Clinical and Experimental Ophthalmology* 229 (3), 294–296.
- Perez-Merino, P., Ortiz, S., Alejandre, N., *et al.*, 2014. Ocular and optical coherence tomography-based corneal aberrometry in keratoconic eyes treated by intracorneal ring segments. *American Journal of Ophthalmology* 157, 116–127.
- Pérez-Merino, P., Ortiz, S., Alejandre, N., Jiménez-Alfaro, I., Marcos, S., 2013. Quantitative OCT-based longitudinal evaluation of intracorneal ring segment implantation in keratoconus. *Investigative Ophthalmology & Visual Science* 54, 6040–6051.
- Randall, J., Vaughan, J.M., 1982. The measurement and interpretation of Brillouin scattering in the lens of the eye. *Proceedings of the Royal Society of London Series B – Biological Sciences* 214 (1197), 449–470.
- Reiss, S., *et al.*, 2012. Ex vivo measurement of postmortem tissue changes in the crystalline lens by Brillouin spectroscopy and confocal reflectance microscopy. *Transactions on Biomedical Engineering* 59 (8), 2348–2354.
- Reiss, S., Burau, G., Stachs, O., Guthoff, R., Stolz, H., 2011. Spatially resolved Brillouin spectroscopy to determine the rheological properties of the eye lens. *Biomedical Optics Express* 2 (8), 2144–2159.
- Sandercock, J.R., 1975. Some recent developments in Brillouin scattering. *RCA Review* 36 (1), 89–107.
- Scarcelli, G., *et al.*, 2015a. Noncontact three-dimensional mapping of intracellular hydromechanical properties by Brillouin microscopy. *Nature Methods* 12 (12), 1132.
- Scarcelli, G., Besner, S., Pineda, R., Kalout, P., Yun, S.H., 2015b. In vivo biomechanical mapping of normal and keratoconus corneas. *JAMA Ophthalmology* 133 (4), 480–482.
- Scarcelli, G., Besner, S., Pineda, R., Yun, S.H., 2014. Biomechanical characterization of keratoconus corneas ex vivo with Brillouin microscopy. *Investigative Ophthalmology & Visual Science* 55 (7), 4490–4495.
- Scarcelli, G., Kim, P., Yun, S., 2011. In vivo measurement of age-related stiffening in the crystalline lens by Brillouin optical microscopy. *Biophysical Journal* 101 (6), 1539–1584.
- Scarcelli, G., Kim, P., Yun, S.H., 2008. Cross-axis cascading of spectral dispersion. *Optics Letters* 33 (24), 2979–2981.
- Scarcelli, G., Yun, S.H., 2007a. Confocal Brillouin microscopy for three-dimensional mechanical imaging. *Nature Photonics* 9 (2), 39–43.
- Scarcelli, G., Yun, S.H., 2007b. Three-dimensional Brillouin confocal microscopy. In: *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference and Photonic Applications Systems Technologies*, OSA Technical Digest Series (CD), paper CTuV5.
- Scarcelli, G., Yun, S.H., 2012. In vivo Brillouin optical microscopy of the human eye. *Optics Express* 20, 9197.
- Schmitt, J.M., 1998. OCT elastography: Imaging microscopic deformation and strain of tissue. *Optics Express* 3 (6), 199.
- Shao, P., *et al.*, 2016. Etalon filters for Brillouin microscopy of highly scattering tissues. *Optics Express* 24 (19), 22232–22238.
- Shirasaki, M., 1996. Large angular dispersion by a virtually imaged phased array and its application to a wavelength demultiplexer. *Optics Letters* 21 (5), 366–368.
- Sigal, I.A., Grimm, J.L., 2012. A few good responses: Which mechanical effects of IOP on the ONH to study? *Investigative Ophthalmology & Visual Science* 53 (7), 4270–4278.
- Singh, M., *et al.*, 2016. Noncontact elastic wave imaging optical coherence elastography for evaluating changes in corneal elasticity due to crosslinking. *IEEE Journal of Selected Topics in Quantum Electronics* 22 (3), 266–276.
- Singh, M., Li, J., *et al.*, 2017. Optical coherence elastography for evaluating customized riboflavin/UV-A corneal collagen crosslinking. *Journal of Biomedical Optics* 22 (9), 91504.
- Singh, M., Nair, A., Aglyamov, S.R., *et al.*, 2017. Noncontact optical coherence elastography of the posterior porcine sclera in situ as a function of IOP. In: Manns, F., Söderberg, P.G., Ho, A. (Eds.), *Proceedings SPIE 10045, Ophthalmic Technologies XXVII*, p. 1004524.
- Song, S., *et al.*, 2015. Quantitative shear-wave optical coherence elastography with a programmable phased array ultrasound as the wave source. *Optics Letters* 40 (21), 5007.
- Song, S., Huang, Z., Wang, R.K., 2013. Tracking mechanical wave propagation within tissue using phase-sensitive optical coherence tomography: Motion artifact and its compensation. *Journal of Biomedical Optics* 18 (12), 121505.
- Torricelli, A.A.M., *et al.*, 2014. BAC-EDTA transepithelial riboflavin-UVA crosslinking has greater biomechanical stiffening effect than standard epithelium-off in rabbit corneas. *Experimental Eye Research* 125, 114–117.
- Vakoc, B., *et al.*, 2005. Phase-resolved optical frequency domain imaging. *Optics Express* 13 (14), 5483–5493.
- Vaughan, J.M., Randall, J.T., 1980. Brillouin scattering, density and elastic properties of the lens and cornea of the eye. *Nature* 284 (5755), 489–491.
- Wang, S., *et al.*, 2013. A focused air-pulse system for optical-coherence-tomography-based measurements of tissue elasticity. *Laser Physics Letters* 10 (7), 75605.
- Wang, S., Larin, K.V., 2014. Noncontact depth-resolved micro-scale optical coherence elastography of the cornea. *Biomedical Optics Express* 5 (11), 3807–3821.
- Weeber, H.A., Eckert, G., Pechhold, W., *et al.*, 2007. Stiffness gradient in the crystalline lens. *Graefes Archive for Clinical and Experimental Ophthalmology* 245 (9), 1357–1366.
- Wilde, G.S., Burd, H.J., Judge, S.J., 2012. Shear modulus data for the human lens determined from a spinning lens test. *Experimental Eye Research* 97 (1), 36–48.

- Wollensak, G., Iomdina, E., 2009. Biomechanical and histological changes after corneal crosslinking with and without epithelial debridement. *Journal of Cataract and Refractive Surgery* 35 (3), 540–546.
- Wollensak, G., Spoerl, E., Seiler, T., 2003. Stress-strain measurements of human and porcine corneas after riboflavin-ultraviolet-a-induced cross-linking. *Journal of Cataract & Refractive Surgery* 29, 1780–1785.
- Wu, C., *et al.*, 2015. Assessing age-related changes in the biomechanical properties of rabbit lens using a coaligned ultrasound and optical coherence elastography system. *Investigative Ophthalmology & Visual Science* 56 (2), 1292–1300.
- Yoon, S., Aglyamov, S., Karpouk, A., Emelianov, S., 2013. The mechanical properties of ex vivo bovine and porcine crystalline lenses: Age-related changes and location-dependent variations. *Ultrasound in Medicine and Biology* 39 (6), 1120–1127.
- Zhang, J., Fiore, A., Yun, S.H., *et al.*, 2016. Line-scanning Brillouin microscopy for rapid non-invasive mechanical imaging. *Scientific Reports*, 6, ARTN 35398.
- Zhang, J., Nou, X., Kim, H., Scarcelli, G., 2017a. Brillouin flow cytometry for label-free mechanical phenotyping of the nucleus. *Lab on a Chip* 17, 663–670.
- Zhang, J., Wu, C., Raghunathan, R., Larin, K.V., Scarcelli, G., 2017b. Non-invasive structural and biomechanical imaging of the developing embryos using OCT and Brillouin microscopy. In: *SPIE Proceedings 10067, Optical Elastography and Tissue Biomechanics IV*, 100670W.

Further Reading

- Ortiz, S., Siedlecki, D., Perez-Merino, P., *et al.*, 2011. Corneal topography from spectral optical coherence tomography. *Optics Express* 2 (12), 3232–3247.
- Scarcelli, G., Yun, S.H., 2008. Confocal Brillouin microscopy for three-dimensional mechanical imaging. *Nature Photonics* 2 (1), 39–43.
- Scarcelli, G., Yun, S.H., 2011. Multistage VIPA etalons for high-extinction parallel Brillouin spectroscopy. *Optics Express* 19 (11), 10913–10922.

Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

Michael Pircher, Medical University of Vienna, Vienna, Austria

© 2018 Elsevier Ltd. All rights reserved.

Abbreviations

AMD Age related macular degeneration
BCC Basal cell carcinoma
CNV Choroidal neovascularization
DOF Depth of focus

FWHM Full width at half maximum
OCT Optical coherence tomography
SLO Scanning laser ophthalmoscope
UHR Ultra high resolution

Nomenclature

A-scan Depth profile recorded with OCT
B-scan Cross-sectional image consisting of several laterally displaced A-scans

C-scan En-face image extracted from an OCT volume scan

Introduction

Optical coherence tomography (OCT) provides cross-sectional images of tissue with high axial resolution. Thereby the sample is probed using light in the infrared range. Image acquisition is performed noninvasively and nondestructively. The axial resolution of OCT lies in the micrometer range and is decoupled from the transverse resolution. This overcomes limitations associated with the numerical aperture of the objective and allows imaging with high axial resolution despite a moderate transverse resolution. Thus layered structures, such as human skin or the human retina, are well suited to be investigated with OCT. The penetration depth in tissue depends on the corresponding scattering properties and typically lies in the order of 1–2 mm. Structures that lie deeper within tissue can be visualized either because of overlying transparent tissue (as is the case in the human eye) or because of endoscopic probes that can be moved inside the body. OCT can be regarded as bridge between microscopic imaging technologies such as confocal microscopy (with limited depth penetration) to macroscopic imaging technologies such as ultrasound or computer tomography.

OCT is a highly sensitive technique and can achieve sensitivities ranging up to 110 dB or more. This allows detection of light even from weak backscattering samples. Image acquisition in OCT is done through recording of depth profiles (A-scans). Several laterally displaced A-scans form a cross-sectional image or B-scan. In a volume scan of the sample several laterally displaced B-scans are recorded. Another advantage of OCT is the high imaging speed. A-scan rates of several megahertz have been reported. Thus highly sampled volumes or several volumes per second can be recorded.

The highest impact of OCT certainly lies in the field of ophthalmology. No other imaging modality is capable to provide three-dimensional (3D) information of retinal tissue. Diagnosis and treatment control of retinal diseases are strongly depending on OCT images and the impact to the field can be compared with the invention of the direct ophthalmoscope more than 200 years ago. Many instruments are commercially available and OCT imaging has found a wide range of applications. Apart from ophthalmology OCT imaging can be found in dermatology, cardiovascular research, pulmonology, and many more. In addition to the biomedical field OCT has been applied to material sciences and art preservation (Stifter, 2007).

Basic Principle of OCT

The technique evolved from low coherence interferometry and was initially intended to measure axial distances in the eye. The basic principle is similar to ultrasound technology. However, instead of sound waves light is used to generate an image. The speed of light is much higher than that of sound waves. Thus, an interferometric technique is needed in order to distinguish between light that has traveled different distances in the micrometer range. Fig. 1(A) shows a scheme of an OCT interferometer. Light is emitted from an OCT light source and is split by the beam splitter into two arms, the reference arm and the sample arm, respectively. In the reference arm, the light is back reflected by the reference arm mirror. In the sample arm the beam is deflected on a scanning unit (for x- and y-scanning) before the light is sent to the sample. The light is backscattered from the sample, de-scanned and brought to interference with light from the reference arm at the beam splitter. The interfering light is then detected by the detector. It is very common to use fiber optics for the interferometer part. This eliminates the need for interferometer alignment and enables the fabrication of very compact systems. There are several techniques available for OCT.

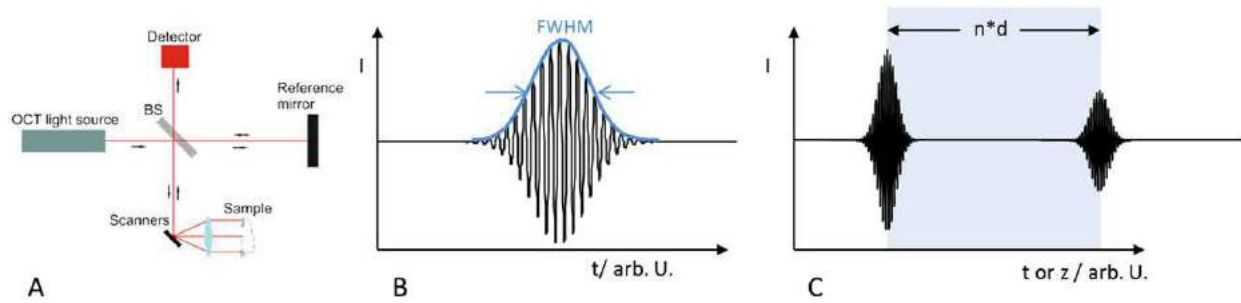


Fig. 1 (A) Scheme of an optical coherence tomography (OCT) interferometer. BS denotes the beam splitter of the interferometer. For time domain optical coherence tomography (TD-OCT) the reference mirror is mounted on a translation stage. (B) Measured intensity I at the detector (interference pattern) when the mirror is moved in axial direction (reflectivity in sample and reference arm are equal) full width at half maximum (FWHM) denotes the full width at half maximum of the coherence function which is defined as axial resolution in OCT. (C) Measured interference pattern when a glass plate with thickness d is measured (n denotes the refractive index of the glass plate).

Timed Domain OCT

Initially the so-called time domain OCT (TD-OCT) technique has been used. In this technique the interferometer is illuminated with a broadband light source. To generate an A-scan (depth profile) the reference mirror is moved axially. In the case of a mirror in the sample arm, the detector will record an intensity pattern that is shown in Fig. 1(B). Due to the broad bandwidth of the source, interference fringes will only be visible if sample and reference arm lengths are matched (within the coherence gate). Thus light originating from different depths can be separated. In the case of a glass plate the light will be backscattered (or back reflected) at the front and back surface of the plate. During axial scanning of the reference arm mirror two coherence functions (arising at the surface and back surface of the glass plate) as illustrated in Fig. 1(C) will be detected. The distance between the two signals corresponds to the optical thickness of the plate (i.e., the geometrical thickness multiplied with the refractive index of the plate). The full width at half maximum (FWHM) of the coherence function is defined as the axial resolution of OCT and equals the round trip coherence length ($= 1/2$ of the coherence length) of the light source.

The A-scan rate of this technique is limited by the inertia of the mirror to a few 100 Hz. However, the method can be highly sensitive. Motion of the mirror will introduce a Doppler frequency shift for each wavelength. Thus, a carrier frequency for the OCT signal is generated which allows for band pass filtering of the signal. This procedure filters $1/f$ noise out and increases the sensitivity. Depending on the bandwidth of the light source and the speed of motion a certain electronic signal bandwidth needs to be detected. Faster speeds will require larger electronic bandwidths (of the detector and amplifying electronics) and thus lower system sensitivity.

To overcome the limitation given by the inertia of the mirror, a galvanometer scanning mirror and a grating (rapid scanning optical delay line) can be placed in the reference arm. Instead of a translational motion in spatial domain, this configuration allows for a rotational motion in wavelength domain. The method has the advantage that the group delay (axial scanning of the sample) and the phase delay (modulation frequency of the OCT signal) can be adjusted independently. Therefore the imaging speed can be improved up to a few kilohertz. TD-OCT detects only light backscattered from within the coherence volume of the sample. The coherence volume is defined through the round trip coherence length and the transverse resolution of the system. Light from other depths (outside the coherence volume) is rejected which limits the sensitivity of the method. However, the depth range of the system depends on the travel range of the reference mirror. Thus, in principle very long axial distances without loss in sensitivity can be scanned with TD-OCT.

Fourier Domain OCT

With the development of Fourier domain (FD) OCT the sensitivity could be greatly improved. The gain in sensitivity can be directly translated into imaging speed that allows A-scan rates an order of magnitude higher than in TD-OCT. Two different FD-OCT schemes are possible. The spectral domain (SD) OCT technique uses a similar broadband light source as in TD-OCT. However, instead of a single detector at the interferometer exit the light is spectrally dispersed through a grating and detected with a line scan camera (cf. Fig. 2(A)). Thereby, the reference arm mirror is kept stationary. In this way light from the entire sample depth is recorded simultaneously which results in a sensitivity advantage compared to TD-OCT.

In case of a mirror in the sample arm (same light power returning from sample and reference arm) the spectrometer camera will detect an interference pattern that is displayed in Fig. 2(B). The modulation of the spectrum arises from constructive and destructive interference for each wavelength and depends on the path length difference between sample and reference arm beam. Thus the frequency of the modulation depends on the amount of path length difference. Smaller differences result in low frequencies, while larger differences cause higher frequencies. In addition, the spectrum (and the interference pattern) is detected in the wavelength domain. Thus, each modulation frequency that corresponds to a certain depth will have a chirp along the spectrum. To remove this chirp, the spectrum needs to be transformed into k -space ($= 1/\lambda$ -space). Associated with this chirp will

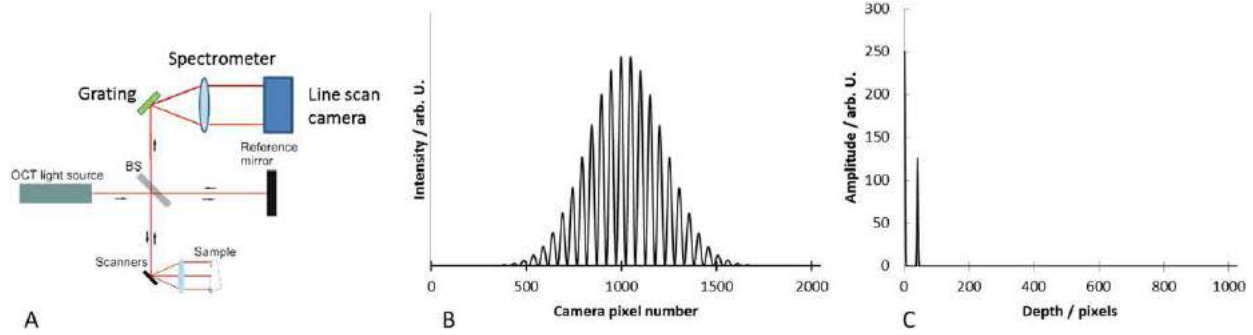


Fig. 2 (A) Scheme of a spectral domain optical coherence tomography (SD-OCT) instrument using a spectrometer in the detection arm (BS, beam splitter). (B) Measured interference pattern when a mirror is placed in the sample arm (light power returning from sample and reference arm are equal). (C) Fast Fourier transformation of the spectrum measured in (B) corresponding to a depth profile.

be a drop of sensitivity (sensitivity fall off) with depth because of the finite width of the detector elements (camera pixels). Higher frequency components (constructive and destructive interference of neighboring wavelengths) will therefore be detected by a single detector element and thus not be visible. It should be noted that FD-OCT requires a minimum detection speed. Jitter within the interferometer or axial movement of the sample (in case of in vivo imaging) will cause a change (phase shift) in the interference pattern which degrades the visibility of the interference pattern when this change occurs during the exposure time of the camera. In order to avoid this “fringe washout” the exposure time should be shorter than 1 ms.

In the case displayed in [Fig. 2\(B\)](#) one arm is only slightly longer than the other. At this point it should be noted that it is not possible to distinguish (without any other means) between positive (reference arm is longer) or negative (sample arm is longer) path length differences because both will result in the same frequency modulation of the spectrum. In order to generate an A-scan the interference pattern (modulated spectrum) needs to be Fourier transformed. The corresponding depth profile is plotted in [Fig. 2\(C\)](#). Because of the inability to distinguish between positive and negative frequencies, the transformation of the spectrum yields in fact two depth profiles that are mirrored around zero delay. Zero delay refers to the position where sample and reference arm lengths are exactly matched. The total number of sample points after Fourier transformation stays constant. However, this means that the number of sample points per A-scan is only half the number of camera pixels. Because the mirrored A-scan does not contain additional information it is very common in OCT to simply omit this part as is done in [Fig. 2\(C\)](#).

The depth range of a SD-OCT system depends on the spectral resolution of the spectrometer as increasing depth is associated with higher modulation frequencies that need to be sampled according to the Nyquist theorem. This resolution is determined by the dispersion properties of the grating (number of lines that are illuminated on the grating), the spectrometer optics and the number of pixels that are available on the line scan camera. According to [Leitgeb et al. \(2003\)](#) the maximum depth range can be calculated via

$$z_{\max} = \frac{\lambda^2}{4\delta\lambda} \quad (1)$$

where λ and $\delta\lambda$ denote the wavelength and the spectral resolution, respectively. A typical OCT configuration uses a light source with a central wavelength of 840 nm, a bandwidth of 50 nm, and a 2048 pixel camera. Assuming that 150 nm bandwidth is imaged onto the camera and that the separation between spectral lines (corresponding to the resolution provided by the grating) is smaller than the size of a pixel an imaging depth of 2.4 mm is achieved. Back reflections from larger imaging depths will appear attenuated as aliasing frequencies and may cause unwanted artifacts inside an OCT image.

Another technique for realizing FD-OCT is swept source (SS) OCT. SS-OCT uses fast (wavelength) tunable lasers as light source in combination with a single (fast) photo detector. During a laser sweep the interference pattern is recorded depending on wavelength and separated in time. The pattern will be similar to the one displayed in [Fig. 2\(B\)](#). It should be noted that the instantaneous power at each wavelength can be higher compared to SD-OCT which compensates for the nonparallel data acquisition scheme of SS-OCT compared to SD-OCT. Similar as in SD-OCT, light backscattered from the entire sample depth contributes to the interference pattern and the depth profile is reconstructed in a similar way via Fourier transformation of the recorded spectral information. The instantaneous linewidth of the source determines the instantaneous coherence length and defines the depth range of the system. The instantaneous linewidth thus plays an equivalent role as the spectral resolution of SD-OCT. Within this range light from the reference arm and sample arm are coherent and can interfere. However, the instantaneous line width is a completely different measure as the round trip coherence length. The latter is defined by the FWHM of the laser sweep, while the former is the width of a laser line emitted at a certain wavelength within the sweep. [Fig. 3\(A\)](#) illustrates the emission of a SS laser.

The output power of many swept laser sources fluctuates. In order to compensate for this fluctuations balanced photo detection is employed. [Fig. 3\(B\)](#) shows a schematic of a fiber-based SS-OCT instrument. The light emitted from the SS is already confined in a single mode fiber. Before the entrance of the interferometer the light traverses a fiberized circulator and is split thereafter by a 50:50 beam splitter into reference and sample beam, respectively. The light back reflected from both arms is recombined by the 50:50

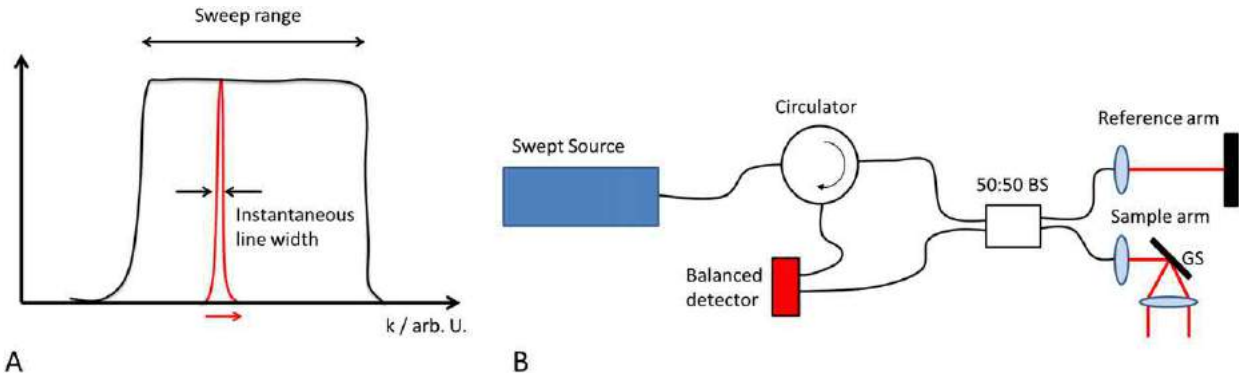


Fig. 3 (A) Illustration of a swept source (SS) laser sweep. The sweep range determines the axial resolution of the system. The instantaneous line width (drawn not to scale) determines the imaging depth provided by the light source. (B) Representative scheme of a SS optical coherence tomography (OCT) instrument based on fiber optics and balanced photo detection. BS 50:50, fiber beam splitter; GS, galvanometer scanner.

beam splitter and is detected with a balanced photo detector. The circulator ideally transmits nearly 100% of the light from the light source to the interferometer, while nearly 100% of the light coming from the interferometer is transmitted to the detector. There are several advantages of SS-OCT compared to SD-OCT. In SD-OCT the light is detected with a spectrometer. The efficiency of the spectrometer cannot be 100% which results in a lower sensitivity compared to SD-OCT. In addition the wavelength sweep can be performed in $1/\lambda$ space which results in an elimination of the frequency chirp that is present in spectrometer-based systems and correspondingly into an elimination of the sensitivity drop with depth. Finally, the spectrometer needs to be perfectly aligned in order to achieve a high sensitivity which sets high demands, from a technical point of view, on the spectrometer design in commercial instruments. No such alignment is required for SS-OCT instruments.

Resolution

The FWHM of the interference pattern (cf. Fig. 1(B)) defines the axial resolution of OCT and is inversely proportional to the bandwidth of the light source. In case of a Gaussian-shaped spectrum the resolution can be calculated via following equation (Drexler and Fujimoto, 2015):

$$\Delta z = \frac{2 \ln(2) \bar{\lambda}^2}{n \pi \Delta \lambda} \quad (2)$$

where $\bar{\lambda}$ denotes the central wavelength, $\Delta \lambda$ the bandwidth, and n the refractive index of the medium. As can be seen from this equation, the resolution in OCT depends on the central wavelength of the light source and the corresponding bandwidth. Fig. 4 shows the axial resolutions for three wavelength regions that are commonly used in OCT. Standard OCT resolution refers to bandwidths around 50 nm corresponding to an axial resolution between 5 and 15 μm . The use of larger bandwidths (100 nm or more) is known as ultrahigh resolution OCT and requires ultra-broadband light sources, such as titanium sapphire laser or supercontinuum light sources. Important for achieving the resolution in OCT that is provided by the light source is a sufficient optical transparency of the optical components and matching of dispersion (see below) introduced in reference and sample arm, respectively. Especially fiberized interferometers may cause unwanted effects because single mode propagation (multimode propagation results in coherent image artifacts) depends on the size of the core diameter in respect to the wavelength. Thus larger wavelengths will be stronger attenuated if the core size is chosen to ensure single mode propagation for the shortest wavelengths.

The transverse resolution mainly depends on the imaging optics in the sample arm. In order to achieve maximum fringe visibility it is common in OCT to use light with high spatial coherence. Light emitted from a single mode fiber has this property. Because of its directionality, divergence, and Gaussian intensity profile, the beam propagation can be described using Gaussian optics. In many OCT systems the objective lens is not the limiting aperture as the beam is scanned over this lens in order to achieve transverse scanning of the sample. Hence, the beam remains Gaussian and the resolution can be calculated via the FWHM of the Gaussian beam waist ω_0 and the angular spread θ_s of the Gaussian beam (Fercher, 2010):

$$\Delta x = 2 \sqrt{\ln 2} \omega_0 = 2 \sqrt{\ln 2} \frac{\bar{\lambda}}{\pi \theta_s} \quad (3)$$

$\bar{\lambda}$ denotes the central wavelength of the beam. Strictly speaking, this resolution can be obtained only at the focal (or certain depth) plane. Outside this focal plane the resolution is degraded. A measure for the imaging depth where the resolution is degraded only by a factor of $\sqrt{2}$ is the Rayleigh range z_R . Twice this range defines the depth of focus (DOF), which can be calculated with

$$2z_R = \text{DOF} = 2 \frac{\bar{\lambda}}{\pi \theta_s^2} \quad (4)$$

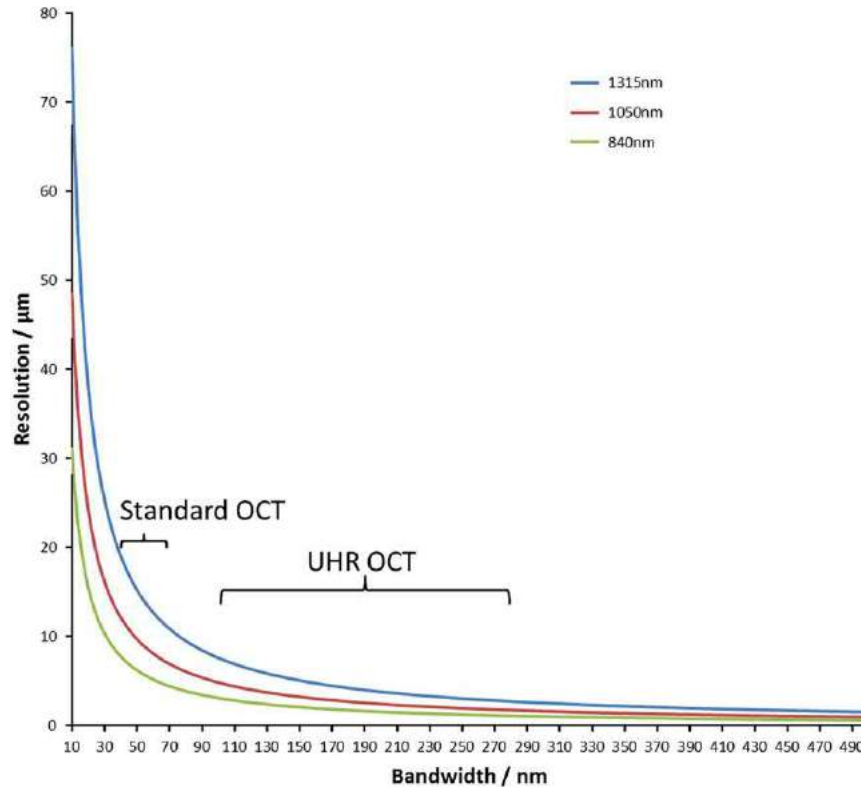


Fig. 4 Axial resolution of an optical coherence tomography (OCT) system depending on the bandwidth and the central wavelength for three commonly used wavelength regions. Standard OCT uses a bandwidth between 30 and 70 nm, which results in a resolution below 20 μm . Ultrahigh resolution (UHR) OCT refers to the use of larger bandwidths (> 100 nm) and corresponding resolutions that are below 5 μm .

Thus for high-resolution systems only part of the imaged volume will be in focus and imaged sharply. For an OCT system at 1300 nm and 20 μm transverse resolution the DOF will be 2 mm. This DOF will be reduced to 20 μm in the case of 2 μm transverse resolution. Hence, the imaging optics needs to be carefully designed in order to optimize the OCT performance. Several methods have been proposed to maintain high transverse resolution throughout the imaging depth. These are based on a dynamic shift of focus with the coherence gate (only possible for TD-OCT) or on the use of special illumination techniques (such as Bessel beams).

Sensitivity

Several noise sources influence the sensitivity in OCT. Receiver noise from the photo detectors, excess noise arising from fluctuations in light power, and shot noise that originates from the quantum nature (and random arrival) of photons. Ideally, an OCT system is limited by shot noise only. This limit can be reached through the coherent amplification process inherent to OCT as the light power returning from the reference arm greatly exceeds the power returning from the sample. In the case of sufficient light power from the light source, the amount of light returning from the reference arm can be optimized in order to reach the shot noise limit. In TD-OCT the shot noise limited sensitivity depends on the detection bandwidth and the power efficiency of the interferometer (Izatt *et al.*, 2015):

$$\text{Sensitivity} = \frac{\rho P_s}{2e\Delta B_{\text{TD}}} \quad (5)$$

where ρ denotes the quantum efficiency of the detector, P_s is the light power returning from the sample arm (with a 100% reflective mirror as sample), e is the electron charge ($e = 1.6 \cdot 10^{-19}$ C), and ΔB_{TD} is the electronic detection bandwidth of the TD-OCT systems. The light power returning from the sample depends on the power incident on the sample, the interferometer configuration, and the optical losses within the system. Let us consider a fiber-based Michelson interferometer with a 50/50 splitting ratio and 4 mW of optical power (central wavelength at 1300 nm with a bandwidth of 50 nm) on the sample. The light is emitted from a single mode fiber. Thus the coupling efficiency (of the light returning from the sample) will be $\sim 50\%$. Another 50% of the light is lost due to the fiber beam splitter which results in an optical power returning from the sample arm (with a 100% reflective mirror as sample) of 1 mW. When the above mentioned TD-OCT instrument is operated at 1 kHz A-scan rate and 3 mm depth range a minimum electronic bandwidth of 800 kHz will be required. This results in a sensitivity of the OCT system (assuming 0.8 quantum efficiency of the detector) of 103 dB.

In FD-OCT the sensitivity can be expressed in a similar way (Leitgeb *et al.*, 2003):

$$\text{Sensitivity} = \frac{\rho P_s}{4e\Delta B_{\text{FD}}} M \quad (6)$$

with ΔB_{FD} the electronic bandwidth of FD-OCT systems (inversely proportional to the exposure time of the camera in SD-OCT) and M the number of spectral channels of the FD-OCT systems. By comparing this equation with Eq. (5) we can see that the sensitivity of FD-OCT is a factor of $M/2$ larger than in TD-OCT. Thus, for a FD-OCT instrument operating with the same light source and interferometer as before will have an improved sensitivity. Nevertheless, following issues need to be considered. The light returning from the sample will be slightly attenuated by the efficiency of the spectrometer (this typically lies in the range of 60%) and the exposure time of the camera will be shorter than the time corresponding to the A-scan rate because the camera requires a certain readout time. For an A-scan rate of 70 kHz we assume that 1000 pixels on the camera are (equally) illuminated and that the camera requires 4 μs for readout. This results in an exposure time of $\sim 10 \mu\text{s}$. The detection bandwidth B of FD-OCT is given by (Leitgeb *et al.*, 2003)

$$B_{\text{FD-OCT}} = \frac{2z_R}{l_c \tau} \quad (7)$$

with z_R denoting the depth range of the system and l_c the double pass coherence length. τ is the exposure time of the camera. In our case ($z_R = 6 \text{ mm}$, $l_c = 15 \mu\text{m}$, and $\tau = 10 \mu\text{s}$) the detection bandwidth is 80 Mhz which results in a shot noise limited sensitivity of 108 dB. Thus, a higher sensitivity can be achieved with this SD-OCT system although the imaging speed is increased by a factor of 70.

Dispersion

Dispersion describes the effect of a wavelength dependent refractive index of a medium. For example, the refractive index of water (cf. Fig. 5) is not constant and decreases with wavelength. For OCT dispersion mismatch between the interferometer arms results in a degradation of the axial point spread function. This is associated with a degradation of axial resolution and sensitivity. Thus the interferometer arms need to be balanced which is normally achieved by inserting corresponding materials that are present in the sample arm (such as lenses for scanning optics) into the reference arm of the interferometer. For compensating dispersion of the eye cuvettes that are filled with water are commonly used. FD-OCT records spectral data. Therefore it is possible to compensate for dispersion in post-processing.

Wavelength Range

The choice of imaging wavelength in OCT depends on the investigated sample as well as on the availability of light sources. For a high penetration depth the total attenuation coefficient of tissue at the imaging wavelength should be minimal. Both, absorption and scattering contribute to the total attenuation coefficient. Water is the main constituent of tissue (apart from teeth

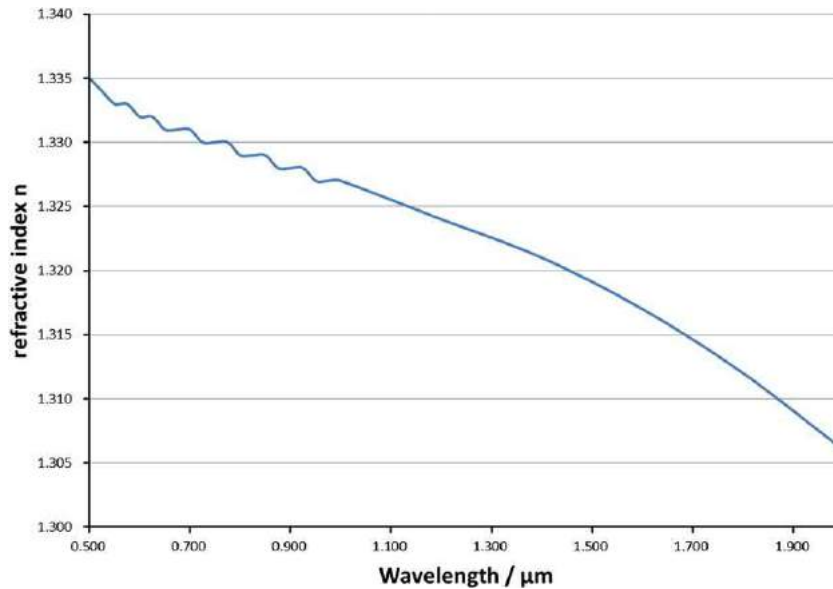


Fig. 5 The wavelength dependence of the refractive index of water in the visible and near-infrared (NIR) region. Data from Hale, G.M., Querry, M. R., 1973. Optical-constants of water in 200-nm to 200- μm wavelength region. *Applied Optics* 12(3), 555–563.

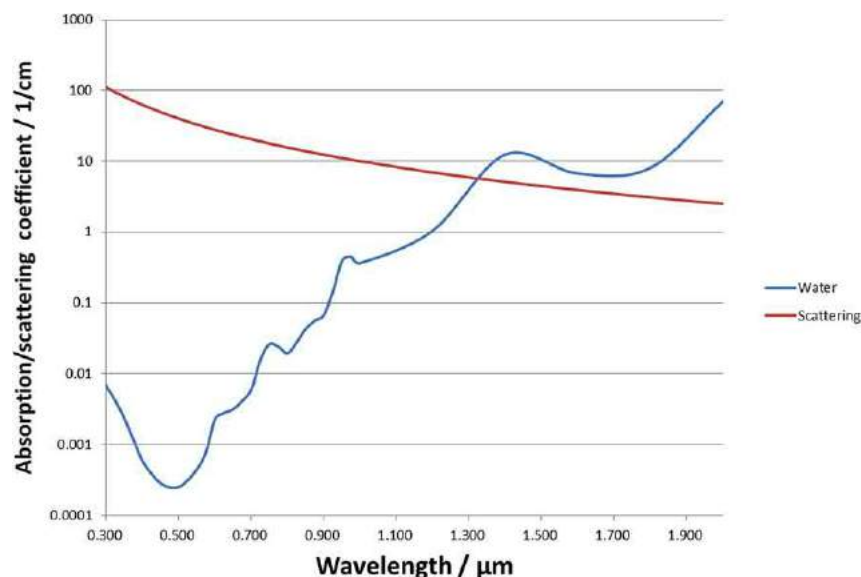


Fig. 6 Absorption coefficient of water and scattering coefficient of dense tissue. The three main optical coherence tomography (OCT) imaging wavelengths (840, 1060, 1300 nm) are outlined in the graph. Data from Hale, G.M., Querry, M.R., 1973. Optical-constants of water in 200-nm to 200- μ m wavelength region. *Applied Optics* 12(3), 555–563.

or bones) and its absorption properties will be dominating. As can be seen in [Fig. 6](#) the absorption coefficient of water strongly increases from visible light to infrared light. On the other hand, scattering gradually decreases in this wavelength range. At the 1300 nm wavelength region the contribution of scattering and absorption are largely within the same range. Thus, this wavelength region is the preferred choice for imaging dense tissue such as human skin as for shorter wavelengths the penetration will be limited by scattering, while for longer wavelengths absorption will be dominating. Imaging at 1300 nm requires detector materials that are sensitive in this region. These are much more expensive than comparable materials that are sensitive in the visible range. As such the preferred OCT imaging technology will be based on SS-OCT because this technology requires a single-balanced detector (instead of a detector array as in SD-OCT) and provides some benefits (as outlined above) compared to the SD-OCT counterpart.

For imaging the retina, the light has to traverse the ocular media (eye length of ~ 2 cm) twice which significantly changes the situation. The scattering properties of the retina are of less importance compared to the transmittance of the light through the eye. Ocular media are very weakly scattering (apart from pathologies such as cataract). Thus the imaging wavelength should be as close to the visible range as possible to minimize absorption by water. However, for highly sensitive measurements the quantum efficiency of the detector should be high. The used detector material, silicon, has a peak around 900 nm and levels off toward shorter wavelength. Thus imaging wavelength close to the maximum of the quantum efficiency of the detector is desirable. An additional aspect is that in the visible range the photoreceptors of the retina will be very sensitive to light (tails of the absorption spectra reach into the infrared region). Hence, the subjective perception of the light power will increase rapidly at wavelength regions that are close or within the visible range. This is associated with increasing subject discomfort during imaging, even though the light levels that are sent to the eye have to be below the limits for safe exposure. In addition dispersion is more pronounced for shorter wavelength regions. Because of these issues the wavelength region around 840 nm has been found as compromise and is widely used in OCT for retinal imaging. Since line scan cameras in this wavelength region are easily available and relatively cheap, the preferred imaging technique in this wavelength region is SD-OCT.

Nevertheless, another wavelength region at 1060 nm which is around a local minimum of the water absorption spectrum (cf. [Fig. 6](#)) has gained interest in recent years. Due to the longer wavelength region the penetration depth within the retina can be increased and the influence of media opacities (such as cataract) on the image quality will be lower. Another advantage of this wavelength region is the availability of fast SSs. Thus high sensitivity can be maintained throughout the imaging depth which further improves the retinal image quality. Due to technical issues (dispersion introduced by fiber optics), SSs at 840 nm are relatively slow compared to their counterparts at 1060 or 1300 nm.

OCT Angiography

In recent years, OCT angiography (OCTA), a method to visualize vessel structure within tissue, has gained increasing interest. There are many different approaches to realize OCTA. The underlying basic principle, however, is the same for all methods. The idea is that two (or more) measurements (A-scans or B-scans) are taken at the same location but separated in time. By comparing these measurements changes (intensity or phase or both) caused by blood flow can be detected. The time separation between these

measurements needs to be short enough in order to minimize the influence of sample motion and long enough in order to ensure sufficient changes introduced by blood flow. With A-scan rates of 100 kHz or more the most common OCTA technique is based on recording several B-scans at the same location. This, however, increases the overall measurement time.

The major benefit of OCTA is that it provides information on the microvascular perfusion of tissue noninvasively. In addition, due to the spatial separation of capillaries, these can be visualized even though the lateral extension lies below the transverse resolution of the system. Other techniques that are capable to provide information on the vascular structure require the application of contrast agents. Some subjects may show allergic reactions or other complications after administration of such contrast agents. Thus state of the art angiographic imaging is always associated with residual risk factors. Another benefit of OCTA lies in the fact that depth resolved information on the vasculature is provided. This allows for a separation between individual vascular beds and thus an improved image contrast. It should be noted, however, that OCTA is not able to detect vessel leakages because the technique relies on motion contrast. Within leakages motion occurring between individual B-scan recordings will be too slow compared to subject motion to be detected in vivo. However, in vitro experiments have shown that without bulk sample motion slow fluid motion (such as Brownian motion) can be visualize with OCTA techniques.

Imaging of Human Skin

The human skin is easily accessible with optical methods. In clinical routine, lesions will be first inspected visually with high magnification using a microscope (dermatoscope). In the case of suspicious lesions a biopsy will be taken which will be sliced and prepared for histologic examination. The cellular details provided by this histologic examination can then be used to determine the malign or benign status of a lesion and finally determines the further clinical procedure. OCT may represent an alternative to the histologic examination as it provides in vivo and noninvasively cross-sectional images of skin. Current studies focus on improving the sensitivity and specificity of OCT methods for detecting malign tissue (Ulrich *et al.*, 2016). Thereby, additional contrast mechanism such as polarization sensitive OCT or OCTA are used to gather additional information. Especially OCTA provides essential additional information as tumors, such as basal cell carcinoma (BCC) are known to present with increased vascularization.

Using fiber-based or handheld systems practically every location can also be reached with OCT. In order to minimize scattering and absorption, the preferred imaging wavelength for human skin will be in the 1300 nm range. The penetration depth of OCT in skin is in the order of 1–2 mm. Fig. 7 shows a representative B-scan image of a finger recorded with a 1310 nm SD-OCT instrument. The A-scan rate of the instrument was 91 kHz. The image consists of 500 A-scans and was recorded in ~5 ms. The bandwidth of the light source was 50 nm which results in an axial resolution of 10 μ m in tissue (assuming a refractive index of 1.4). The power onto the skin was 8 mW. Individual layers of the skin can be clearly seen in Fig. 7 and are labeled in the image.

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/B978-0-12-803581-8.09787-3>.

An interesting skin region is the nail fold of a finger. In this region capillaries are close to the surface (close to the nail some capillaries extend parallel to the surface) and can be investigated with optical methods. In order to assess the status of the capillary network, capillaroscopy is used in clinical routine. This method examines the skin with a microscope after applying a contrast agent (such as oil) to the skin region. Using OCTA this network can be visualized completely noninvasively and without the

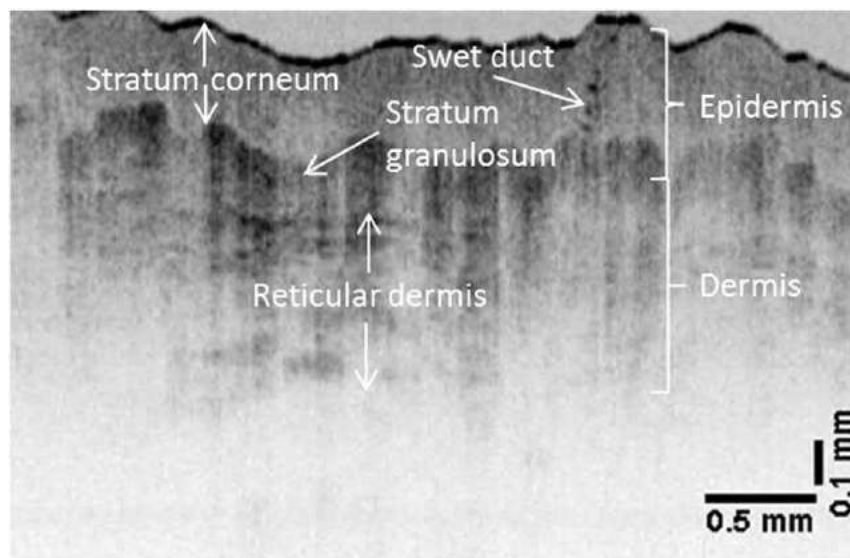


Fig. 7 Representative B-scan image recorded at 1300 nm of the inner side of a finger. Superficial layers down to deeper layers of the dermis can be visualized. The multimedia file (media 1) shows a fly through of B-scans of the entire three-dimensional (3D) data set.

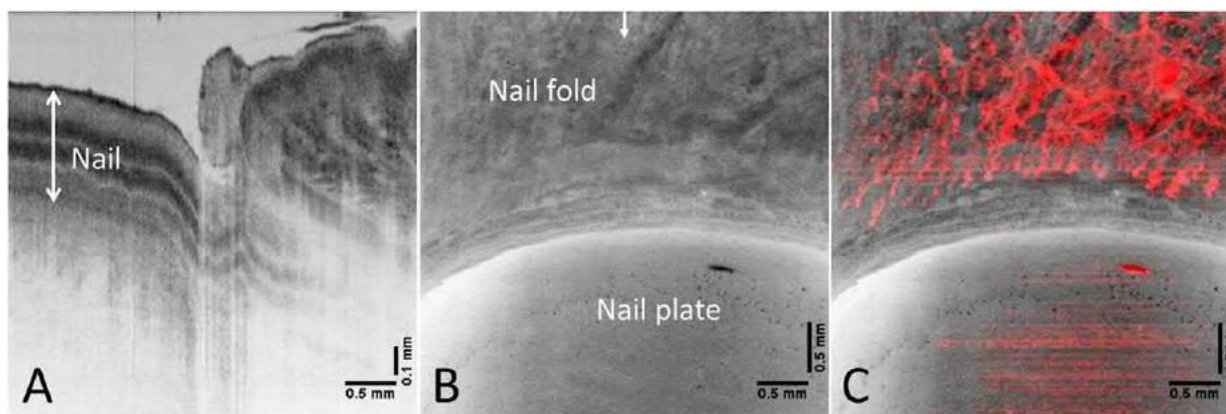


Fig. 8 Representative optical coherence tomography (OCT) images recorded in the nail fold region of a healthy volunteer. (A) OCT-intensity B-scan (averaged over 5 B-scans). (B) Depth integrated en-face intensity image. The arrow indicates the location of the B-scan displayed in (A). (C) Depth integrated en-face intensity image and overlaid (red) vessel map.

application of a contrast agent. **Fig. 8** shows representative image data recorded with the SD-OCT instrument described above. The intensity B-scan (**Fig. 8(A)**) clearly shows different structures such as the nail plate and different skin layers of the nail fold. The banding structure within the nail plate is an artifact that originates from birefringence of the plate. The sensitivity of the system is sufficient to penetrate through the nail plate as structures underneath the plate can be seen. The axial displacement of the nail plate beneath the nail fold (cf. right hand side in **Fig. 8(A)**) is caused by the increased optical path length that is introduced by light propagation through the overlying nail fold tissue. The en-face intensity projection clearly shows the different regions of the nail fold and nail plate. **Fig. 8(C)** shows a composite image where the depth integrated vessel structure is overlaid to the intensity image of **Fig. 8(B)**. In this image the capillary network can be clearly seen. Close to the nail plate individual capillary loops can be seen which are typical for this skin region.

Imaging of Human Retina

The human retina consists of several layers and represents an essential part for the vision process. All layers that can be seen in histology can be visualized with OCT. **Fig. 9** shows a representative OCT B-scan image of the fovea region with a labeling of the different retinal layers. In healthy subjects the fovea is the location of best visual acuity. In this region the highest density of cone photoreceptors can be found. These cells are responsible for photopic vision and can be distinguished into three types (sensitive in the short, medium, and long wavelength region). In the fovea no rod photoreceptors (responsible for scotopic vision) are present. The light is incident from the top of this image and has to penetrate different retinal layers until it is detected by the photoreceptors. The pigments of the photoreceptors are located within the outer segments of cones and rods. The image was recorded with a research grade prototype SD-OCT instrument operating at 863 nm with a bandwidth of 60 nm which resulted in an axial resolution of $\sim 4 \mu\text{m}$ in tissue (assuming a refractive index of 1.4). The A-scan rate was 70 kHz and one B-scan consisting of 1024 A-scans was recorded in ~ 15 ms. Although the axial resolution is sufficient to resolve individual retinal layers, individual cells (such as photoreceptors) cannot be visualized. In order to increase the transverse resolution, aberrations that are introduced by imperfections of the eye optics need to be compensated for. This can be done with the implementation of adaptive optics, a technology known from astronomy (Pircher and Zawadzki, 2007).

Many commercial OCT devices for imaging the retina are available. **Fig. 10** shows representative data recorded with a commercial SD-OCT instrument. The instrument provides a scanning laser ophthalmoscope (SLO) overview mode that is used for retinal tracking. The location of the recorded B-scan is marked with a green line in the SLO image. The image quality is very similar to the previous image. The entire 3D data set can be viewed in media 2.

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/B978-0-12-803581-8.09787-3>.

In order to outline the difference of retinal images recorded at different wavelengths, **Fig. 11** shows a retinal image recorded at 1040 nm. Compared to the previous images enhanced penetration into the choroid and even into the sclera can be seen. Otherwise all retinal layers can be seen with similar quality. The image was recorded with a research grade SS-OCT instrument operating at 100 kHz A-scan rate and 20×20 field of view. The axial resolution of the system was $5.6 \mu\text{m}$ in tissue.

Through a change in the scanning pattern (recording of several B-scans at the same location) angiographic data can be retrieved. **Fig. 12** shows representative data of a healthy volunteer recorded with a commercial instrument. The en-face projection images were generated through depth integration over the depth extension indicated with white arrows in **Fig. 12(A)**. **Fig. 12(B)** shows cross-sectional angiographic data (red) overlaid to the intensity image. The contrast of retinal vessels is greatly increased using the angiographic data evaluation. The en-face projection of the intensity images (cf. **Fig. 12(C)**) shows larger vessels as shadows.

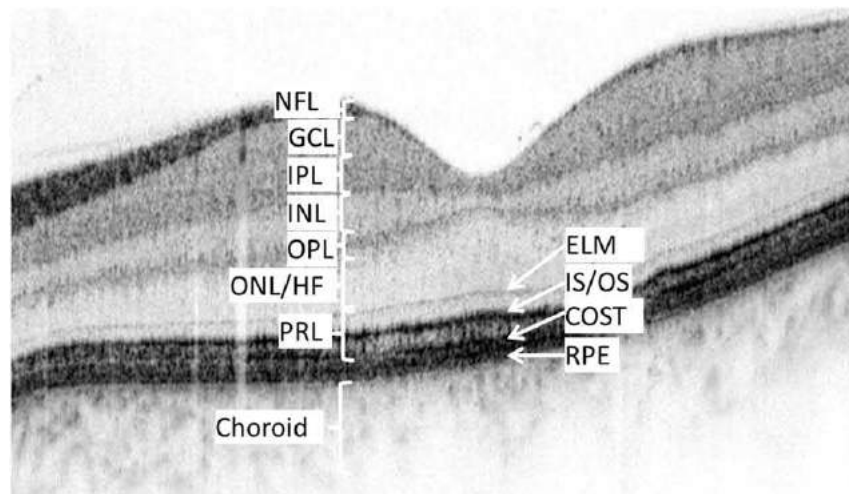


Fig. 9 Optical coherence tomography (OCT) image of the retina with a labeling of the different layers (the image spans an angle of 28 degree which corresponds to ~ 8 mm on the retina). COST, cone outer segment tips; ELM, external limiting membrane; GCL, ganglion cell layer; INL, inner nuclear layer; IPL, inner plexiform layer; IS/OS, junction between inner and outer segments of cone photoreceptors; ONL/HF, outer nuclear layer/Henle's fiber layer; OPL, outer plexiform layer; PRL photoreceptor layer; RNFL, retinal nerve fiber layer; RPE, retinal pigment epithelium.

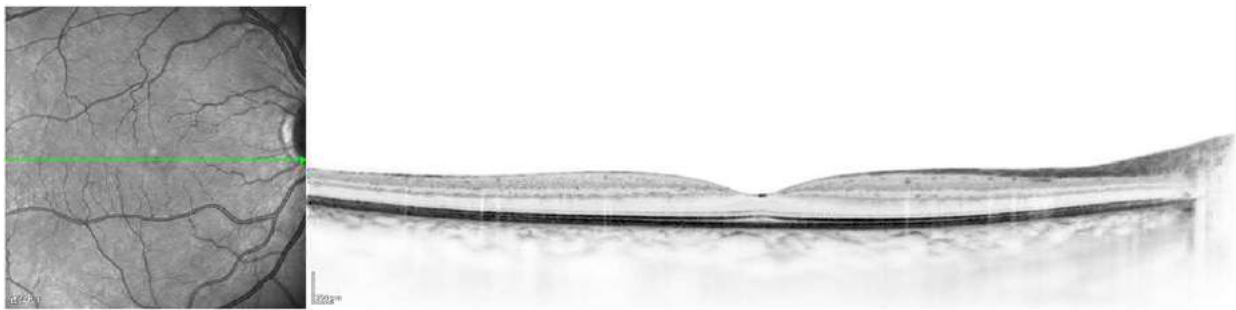


Fig. 10 Representative retinal image data of the fovea of a healthy volunteer recorded with a commercial instrument (imaging wavelength = 840 nm). Left: overview scanning laser ophthalmoscope (SLO) image. Right: optical coherence tomography (OCT) B-scan recorded at the location indicated in the left image with the green arrow. The multimedia file (media 2) shows a fly through of B-scans of the entire 3D data set.

However, the entire vascular structure can be seen in the corresponding OCTA projection map. This map clearly shows the avascular zone of the fovea region. From these maps deteriorations (such as non-perfused areas) from the normal vasculature structure can be easily determined.

Fig. 13 shows representative image data of patients with age related macular degeneration (AMD). AMD is regarded as one of the major causes of blindness in the industrialized world. The disease mainly affects the macula region (containing the fovea) of subjects that are older than 60 years. AMD can be differentiated into several forms: dry AMD and wet AMD. The latter is the most severe form which results (untreated) into a rapid degradation of visual acuity. An early indicator of AMD is the presence of Drusen. Drusen are elevations of the retinal pigment epithelium (RPE)/photoreceptor bands because of accumulation of retinal debris underneath these layers. **Fig. 13(A)** shows a typical example of OCT images recorded in a patient with Drusen. The elevations of the photoreceptors cause image distortions and degradation in visual acuity of the patient. At locations of the Drusen an additional layer (Bruch's membrane) can be seen. In the healthy case, this layer is obscured by the overlying RPE. Due to the highly scattering of the pigments within the RPE, this layer appears as broadened diffuse layer although it is a mono cellular layer. The 3D visualization of Drusen allows for a quantitative assessment of the size and volume of the Drusen. As such the effect of therapeutic interventions on the Drusen size can be precisely monitored.

The most severe form of AMD is choroidal neovascularization (CNV). In this case new vasculature is generated within the retina and leakages of these lead to damage of the sensitive neural tissue of the retina. **Fig. 13(B)** shows a representative B-scan of the fovea region of a patient with CNV. Scarring of retinal tissue and associated disintegration of photoreceptors is very often accompanied by this disease. Using OCT the size of edema or cysts can be quantified. In addition vascular growth can be determined using OCTA. Therapeutic intervention (such as anti-vascular endothelial growth factor (anti-VEGF) therapy) can be easily monitored and the technology can be meanwhile regarded as gold standard for this kind of diagnosis and for treatment control.

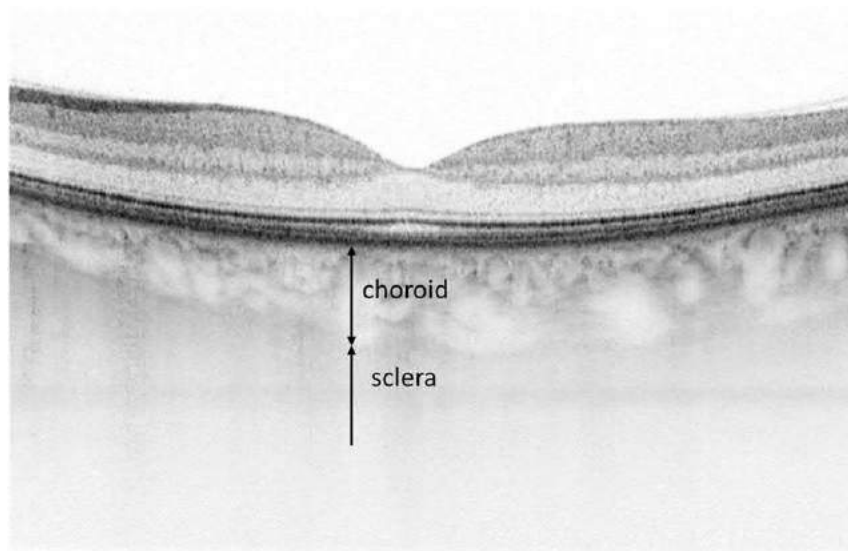


Fig. 11 Retinal image of the fovea of a healthy volunteer recorded with a swept source optical coherence tomography (SS-OCT) instrument at 1040 nm. Enhanced penetration into the choroid and sclera can be observed.

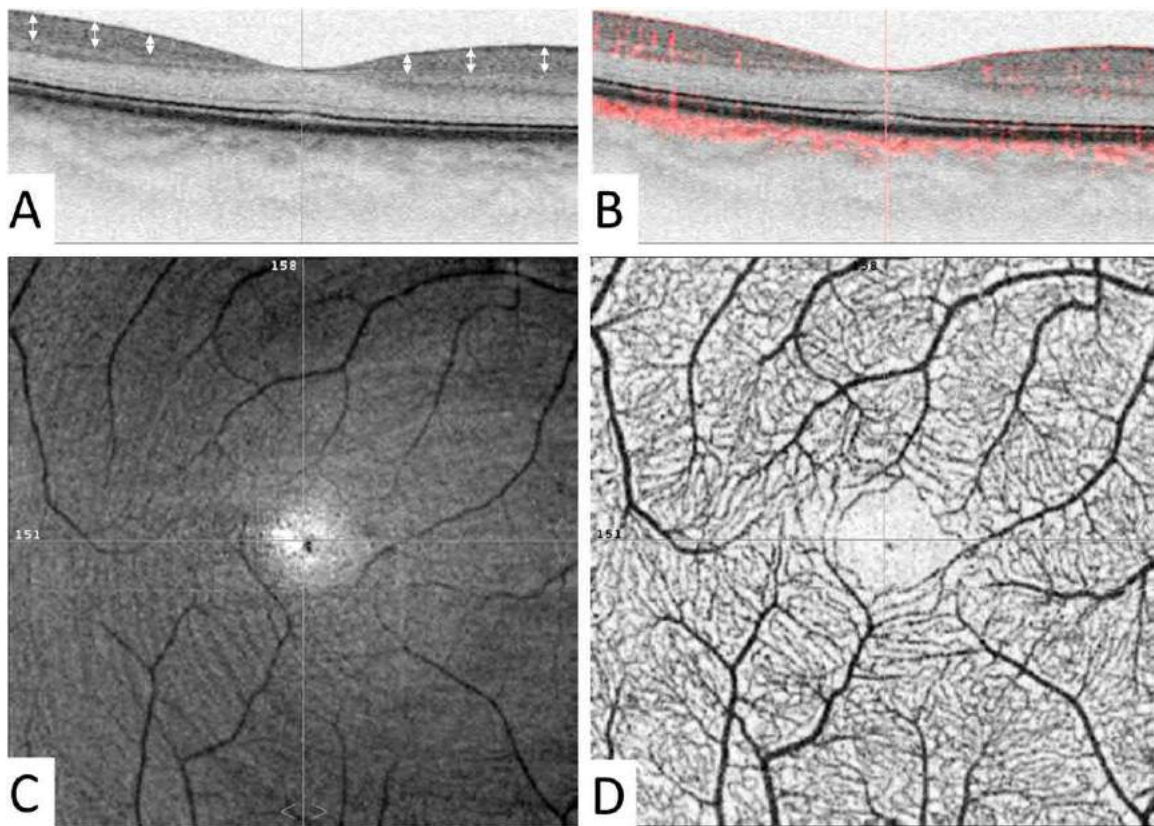


Fig. 12 Retinal images of the fovea of a healthy volunteer recorded with a commercial instrument. (A) Central B-scan of the volume scan. The arrows indicate the image depth that was used to generate the en-face images in (C) and (D). (B) Central B-scan of the volume scan and overlaid vessel structure (in red) retrieved with the angiography algorithm. (C) En-face intensity image (white=high intensity, black=low intensity) generated through depth integration over the area indicated in (A). En-face vessel map generated through maximum intensity projection of the angiographic data within the area indicated in (A).

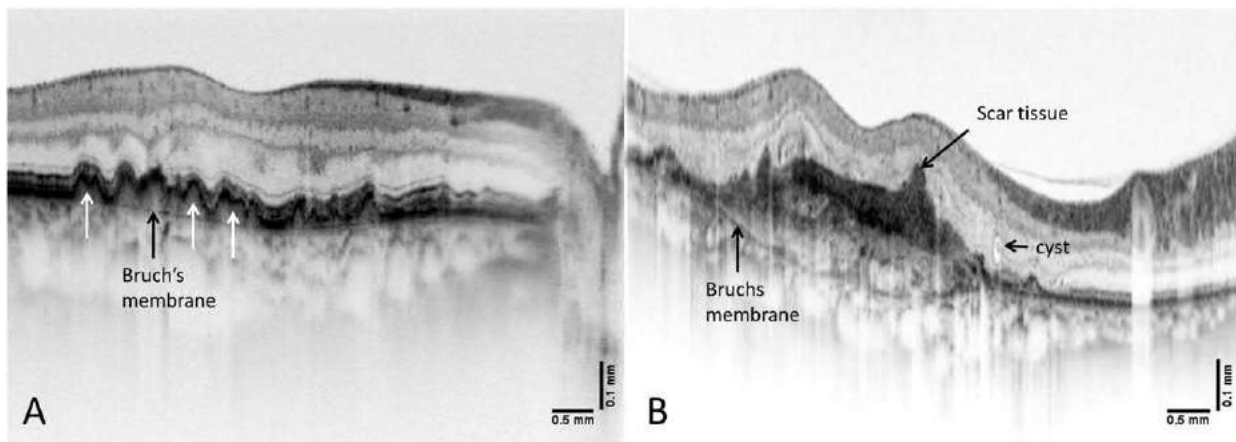


Fig. 13 Representative optical coherence tomography (OCT) B-scans (averaged over 50 frames) at 840 nm of patients with age related macular degeneration (AMD). (A) OCT B-scan image of Drusen (some are indicated by white arrows). The black arrow indicates Bruch's membrane. (B) OCT B-scan image of a choroidal neovascularization (CNV). Images are displayed on a logarithmic intensity scale.

There are many other retinal diseases where OCT plays a key role in diagnosis. A summary of the impact of OCT to all these diseases will be beyond the scope of this article. Nevertheless, one specific example should be mentioned: Glaucoma, a disease that is associated with visual field defects and one of the major causes for blindness worldwide. At an early stage of the disease, these defects occur in the periphery of the retina which is not easily noticeable by patients themselves because for reading etc. only the fovea region is used. Before changes in the visual field can be detected, a loss of retinal nerve fiber tissue (loss in layer thickness) will occur. The 3D imaging capabilities of OCT enable a precise measurement of the retinal nerve fiber layer thickness. Thus OCT provides an excellent indicator for early signs of Glaucoma and for monitoring therapeutic success.

Conclusion and Outlook

OCT is a very powerful imaging technology. With the development of FD-OCT imaging speed could be greatly improved and A-scan rates in the megahertz range have become possible. This paved the way for many new applications of OCT. This article focused on skin and retinal imaging but many more applications of OCT are available. Associated with the improved imaging speed is the possibility to investigate dynamic processes or to increase the contrast through the introduction of angiographic evaluation of OCT data. The noninvasive visualization of capillaries in tissue, such as skin or retina, opens completely new perspectives for clinical research. The high clinical demand of this technology manifests itself through the rapid commercialization of OCTA for both, skin, and retinal imaging. Thereby, OCT certainly plays a key role for accurate diagnosis especially in ophthalmic routine. Nevertheless, there are new developments in the field. Currently new SSs are developed that provide higher speeds (A-scan rates) and very long coherence lengths. Meanwhile, imaging depth ranges of some meters have been demonstrated with OCT. This opens new possibilities such as imaging of the whole eye (from the cornea to the retina) with a single SS-OCT system. In addition further parallelization of OCT is investigated. Line field OCT, for example, records one B-scan in parallel. Thereby SD-OCT (this requires an area detector in the spectrometer) or SS-OCT can be used. The latter uses a SS in combination with a line scan camera. Very high speeds (several volumes per second) can be achieved using a full-field illumination in combination with a SS and a very fast area camera. The demands on the camera are very high because the phase stability (to avoid "fringe washout") requires the recording of a volume within a short time frame (ideally below 1 ms). In order to have sufficient sampling points in depth (determined by the number of spectral lines) the frame rate of the camera needs to be in the order of 50 kHz. Apart from the high imaging speeds that are possible and that have been demonstrated for in vivo retinal imaging, the full-field technique provides phase stability over the entire field of view (in lateral direction). OCT provides direct access to the phase of the electric light field and computational techniques can be applied in order to compensate for aberrations introduced, for example, by the eye. Thus high resolution can be achieved even without the use of hardware-based adaptive optics. In addition, the focal plane of the image can be adjusted in post-processing. As such, an entirely sharp volume can be obtained with high transverse resolution despite of a limited DOF.

Apart from these developments, OCT has become a valuable tool in cardiovascular research. Using fiberized probes, OCT imaging can be done from inside the body, for example, within a coronary artery. As such plaques, deposits, or stents within the artery can be visualized in vivo. Key element for this development is the high imaging speed provided by OCT. Currently clinical trials are performed to demonstrate the benefit of cardiovascular OCT compared to state of the art examination tools.

Another development in the field is the combination of OCT with other imaging technologies. For example, the combination of OCT with photo acoustics has gained increasing interest. The photo acoustic technique relies on absorption of the probing light beam and provides a higher penetration depth but lower resolution than OCT. Therefore vasculature that is located deeper within

tissue can be visualized. The combination with OCTA enables the visualization of the entire vascular network as superficial capillaries are detected with OCT, while the deeper vascular structure is detected with photo acoustics.

Acknowledgments

The author likes to thank following people from the Medical University of Vienna for their assistance: Christoph K. Hitzenberger, Matthias Salas, Martin Fürst, Mitsuro Sugita, Teresa Torzicky, and Ursula Schmidt-Erfurth.

See also: Interferometric Imaging. Overview: Coherence

References

- Drexler, W., Fujimoto, J., 2015. Optical Coherence Tomography, Heidelberg. New York, NY: Springer.
- Fercher, A.F., 2010. Optical coherence tomography – Development, principles, applications. *Zeitschrift Fur Medizinische Physik* 20, 251–276.
- Izatt, J.A., Choma, M.A., Dhalla, A., 2015. Theory of optical coherence tomography. In: Drexler, W., Fujimoto, J.G. (Eds.), *Optical Coherence Tomography*. Heidelberg; New York, NY; Dordrecht; London: Springer.
- Leitgeb, R., Hitzenberger, C.K., Fercher, A.F., 2003. Performance of Fourier domain vs. time domain optical coherence tomography. *Optics Express* 11, 889–894.
- Pircher, M., Zawadzki, R., 2007. Combining adaptive optics with optical coherence tomography: Unveiling the cellular structure of the human retina in vivo. *Expert Review of Ophthalmology* 2, 16.
- Stifter, D., 2007. Beyond biomedicine: A review of alternative applications and developments for optical coherence tomography. *Applied Physics B-Lasers and Optics* 88, 337–357.
- Ulrich, M., Themstrup, L., De Carvalho, N., *et al.*, 2016. Dynamic optical coherence tomography in dermatology. *Dermatology* 232, 298–311.

TIRF Microscopy and Variants

Daniel Axelrod, University of Michigan, Ann Arbor, MI, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Total internal reflection fluorescence microscopy (TIRFM) is a method of illuminating a fluorescent sample so that only the zone proximal to the solid support (the “substrate”) will be excited. This high excitation selectivity greatly reduces background fluorescence from regions farther from the substrate. TIRFM has found many applications in studies of cell/substrate contact, intercellular organelles near those contact regions, and exo-and endocytosis, as well as single molecules studies where out-of-focus background would otherwise make detection difficult.

Numerous reviews of TIRFM have been written (Schneckenburger, 2005; Axelrod, 2014), including a chapter in the Encyclopedia of Cell Biology, also published by Elsevier/Academic Press (Axelrod, 2016). Those works (and their references) and the references in this article should be consulted for mathematical details of the basic theory and optical configurations. The emphasis here is on qualitative effects and intuitive explanations, with no equations presented. This chapter will focus on TIR-related issues of lateral and axial resolution, polarization, multiphoton variations, refractive index measurement, scattering, and combinations of TIRFM with other optical techniques.

Qualitative Basic Principles

TIR occurs at a planar interface between two dielectrics (defined as $z=0$) of different refractive index when the angle of incidence as measured from the direction normal to the interface in the higher index material exceeds some critical angle. From the viewpoint of simple ray optics and Snell's Law, at that critical angle, the refracted beam propagating into the lower index material refracts away from the normal and skims along the surface. At higher angles, the beam reflects 100% back into the higher index material. From the viewpoint of wave optics, the incident plane waves approaching the interface at subcritical angles change direction at the interface due to their increased speed in the lower index medium. However, the spacing of the wavefronts on both sides of the interface, as they intersect the surface, must be equal. That requirement, along with the fixed wavelength for propagation in the low index medium (set by the frequency and speed of light there) determines the orientation of wavefronts in the low index medium. At the critical angle, the wavefronts in the lower index material are oriented normal to the interface. At higher incidence angles, the geometry would force the spacing of the wavefronts to be closer than can be supported as propagating light in the lower index material, so instead, the wavefronts becoming exponentially decaying in amplitude with distance along the normal direction (the “ z -direction”) to the substrate (Axelrod, 2013). This spatially-decaying field in z is called the “evanescent field” of TIR. It still has wavefronts moving along the surface with an amplitude at $z=0$ at least as large as the incident field's amplitude (and usually higher), and those sinusoidal electric field fluctuations are fully capable of exciting light-absorbing (and possibly fluorescing) molecules. The evanescent field can also be scattered by refractive index variations in the low index material and thereby partially converted back into propagating light. Those two mechanisms – absorption and scattering – of the evanescent field, decrease the amplitude of the reflected beam to less than 100%, but that decrement typically is very small.

Lateral Resolution

In its standard, high aperture microscope objective-based form, TIRF is made possible by focusing the incident beam to a small spot at the objective's back focal plane, sufficiently far off-axis so the beam emerges collimated from the objective, and heads toward the sample at greater than the critical angle. For this to work, the objective aperture must be greater than the refractive index of the medium on the low density side of the interface (1.33 for water) and preferably at least 1.4. In the space between the BFP and final emergence, the beam expands from its focus, so TIRF is necessarily a wide-field method, possibly illuminating the entire field of view. Reducing the area of illumination is certainly possible (by decreasing the angle of convergence/divergence) while retaining the incident beam's focus at the objective back focal plane (BFP). But as that convergence angle decreases, the size of the focal spot at the BFP necessarily increases by diffraction. The result is that the angle of incidence starts to spread, and some light may even spread into sub-critical angles, thereby polluting the purity of the evanescence.

But for some purposes, a very small lateral area of illumination at the TIR/sample plane is desirable. One such purpose is improve lateral resolution by combining TIRF with structured illumination microscopy (SIM). In standard SIM (Gustafsson *et al.*, 2000; Hirvonen *et al.*, 2009), excitation field is given a periodic illumination (in one or two dimensions). Then at least three successive images are recorded, each with the periodicity shifted by 120 deg in phase; a computer algorithm then recombines these images into a rendition of the original image with substantially increased resolution. (Qualitatively, this works because two fluorescent points spaced apart by a sub-resolution distance will “see” different excitation intensities on the steep slopes of the

periodic illumination, and that difference will change or even reverse as the phase is shifted.) One commercially-available way of producing a periodicity is simply to project onto the sample the shadow of a grating positioned at an equivalent sample plane in the excitation path. But the finer the periodic spacing, the higher will be the ultimate reconstructed resolution. By interfering pairs of oppositely directed and mutually coherent TIR beams (Chung *et al.*, 2007; Kner *et al.*, 2009; Fiolka *et al.*, 2012; Beach *et al.*, 2014), a striped interference pattern in the evanescent field can be generated with a spacing that is significantly smaller than any that could be produced by either a projected grating or by interfering subcritical propagating beams.

This TIR-SIM procedure produces resolution improvement in one dimension only, the dimension transverse to the interference stripes. But two additional oppositely-directed coherent TIR beams (all split from the same laser source) can be introduced from the orthogonal direction, producing an interference checkerboard. This leads to improved resolution in both dimensions in the sample plane.

Conceptually extending this approach, evanescent fields could be configured to approach from many azimuthal directions simultaneously. These would all constructively interfere completely in only one location: at the very center of the field. In the most extreme case, an azimuthal circular ring of supercritical illumination will form a very tiny evanescent spot in the center, with a radius only about 2/3 of that obtainable by focusing subcritical laser beam (Axelrod, 2013). That spot would have not only the thin $\lambda/5$ to $\lambda/10$ axial decay length of standard wide-area TIRF but also an extremely small lateral size, thereby defining an illumination volume of only $\sim 0.005 \mu\text{m}^3$. Such a small volume could be used for scanning (Chon and Gu, 2004; Terakado *et al.*, 2009), for highly confined FRAP, and for FCS.

To create this highly concentrated TIR illumination pattern, the incident light must be confined to a supercritical radius ring at the back focal plane. If the light is focused to a sharp annulus in that BFP region, the polar angle of incidence at the sample (and thereby the evanescent field depth at the sample) will be a well defined value. If the light is merely limited to that region but not focused there, the polar incidence angle (and evanescent field depth) will span a range of values. This is the approach used by Chon and Gu (2004); Terakado *et al.* (2009). The polarization of the ring at the BFP should be such that any azimuthal direction will produce an evanescent field polarization that can interfere with that of every other azimuthal direction, as done by Terakado *et al.* (2009); otherwise, the smallest possible spot at the sample plane will not be produced. The only possible BFP polarization that can do this is radial polarization, which can be implemented by commercially available circular grating plates (Beresna *et al.*, 2011).

In standard intersecting beam TIR-SIM pairs with a finite number of azimuthal angles, the best resolution depends on how small can be the spacing of the interference fringes. Use of the highest refractive index substrate helps, because that decreases the wavelength of the incident light in the substrate. Another approach is to use a substrate that is imprinted with a periodic pattern spacing even smaller than the spacing of intersecting TIR beams. This can improve the resolution even farther. A grating also gives the optical system a resonance that increases the amplitude of the evanescent field for particular incidence angles and wavelengths. This occurs, for example, if the periodic grating spacing is an integral multiple of the incident light wavelength as measured along the TIR surface (Sentenac *et al.*, 2009).

TIR Polarization

The evanescent field is polarized according to the polarization of the incident beam. S-pol incident light (say, in the y direction, for incident light in the x-z plane) gives rise to s-pol evanescence (in this case, the y direction). P-pol incident light (say, in the x-z plane of incidence) gives rise to a p-pol evanescent field. The p-pol takes the form of an elliptical polarization, mainly along the z-direction but with some amplitude in the x-direction at a 90 deg phase difference. The characteristic decay distance of the evanescent electric field in the axial (z) direction is the same regardless of polarization.

The p-pol ability to produce a polarization in the z-direction is unique: standard EPI-illumination with excitation light directed along the axial direction produces only polarizations oriented in the sample plane, the x-y directions. That z-direction polarization cannot excite fluorophores whose dipoles lie parallel to the sample plane. But any deviation from parallelism can be excited. For example, a planar membrane labeled with a fluorophore (such as a carbocyanine dye with a double hydrocarbon tails such as diI or diD coupled by a conjugated bridge) whose excitation dipole lies in the local plane of the membrane will not be excited if the membrane itself lies parallel to the TIR substrate. But if the membrane is wavy or folded or has protrusions or indentations, those region will selectively “light up” under p-pol TIRF, against a dark background. This effect has been used to detect sites of endo- and exocytosis as they occur on living secretory cells (Steyer and Almers, 1999), and also to study the interaction and kinetics of fusion between lipid vesicles and planar supported membranes (Oreopoulos and Yip, 2009; Kiessling *et al.*, 2010). The technique requires that two images be recorded, one with p-pol, and one with s-pol for normalization. Optical designs to speed the switching and image acquisition have been devised (Johnson *et al.*, 2014).

Variable Polar Angle Illumination, Including Subcritical

The depth of the evanescent field is a function of incidence polar angle, decreasing rapidly from (theoretically) infinite right at the critical angle, down to about a single wavelength of light at about one degree higher than the critical angle, and farther down to about a tenth of the wavelength at skimming (90 deg) polar incidence angle. (However, the intensity of the evanescent field drops

monotonically with increasing incidence angle.) Another chapter in this Encyclopedia describes how this variation may be used to determine z-distance separation of a single fluorophore from a TIR surface, and previous reviews contain mathematical details. In principle, the z-dependent concentration $C(z)$ of a fluorophore near a substrate could be inferred by performing an inverse Laplace transform to the TIRF intensity vs. polar incidence angle I vs. θ profile. In practice, various reasonable guesses for $C(z)$ are tested by comparing their predicted I vs. θ profiles against actual measured profiles (Olveczky *et al.*, 1997; Oheim *et al.*, 1998). Variable-angle TIRF has also been used to measure the distribution of z-distances of whole secretory granules from the cell/substrate contact region (Rohrbach, 2000).

If the polar incidence angle is made to be slightly less than the critical angle, the incident beam refracts into the sample at a very high (but less than 90 deg) polar angle. If the subcritical incident beam were truly an infinitely-broad plane wave, as generally assumed in basic optical theory for TIR, then the z-depth in the lower index medium would be infinite. But in practice, a refracted beam skimming along the surface above the substrate has a finite depth. That depth is not easily controlled in commercial TIRF setups. But it can be estimated by viewing a fluorescent sample under epi-illumination with the angle of incidence adjusted to zero degrees (i.e., straight up in an inverted microscope, called “EPI-fluorescence”). The illuminated region will be roughly circular with fuzzy edges but with some characteristic radius on the sample. That radius (usually many micrometers) will be the approximate z-depth of the field of a barely subcritical beam when viewed at its position of emergence from the substrate. The emerging beam will form a “pencil” of light slowly diverging up from the substrate. By viewing the illuminated region “downbeam”, several micrometers from where it emerged from the substrate, the full diameter of the beam propagates in the lower density medium. By laterally shifting the emergence location from the region under observation, the height of that pencil above the substrate can be adjusted. This allows the ability to explore fluorescence from regions in that shaft of light at user-controllable varying depths into the sample, depths that are inaccessible to standard TIRF and yet do not excite as much out-of-focus background as does standard “straight-up” EPI. With a custom setup based around a free laser beam, rather than commercial setup based around an optical fiber, the radius of the illumination region can be controlled by changing the angle of convergence/divergence of the incident beam at the objective BFP: the smaller that angle, the smaller the radius and consequently the smaller the depth of the subcritical refracted beam. It should be possible to increase the lateral extent of the illumination without increasing the z-depth (i.e., to produce a sheet rather than a pencil) by use of a cylindrical lens.

Barely subcritical illumination has been given a variety of names in the literature: leaky TIRF, quasi-TIRF, pseudo-TIRF, and Highly Inclined and Laminated Optical sheet (HILO), and Variable Angle Epifluorescence microscopy, the last two of which are the most scientifically accurate since subcritical illumination is not TIR at all (Konopka and Bednarek, 2008; Lu *et al.*, 2013; Sun *et al.*, 2011; Tokunaga *et al.*, 2008).

Somewhat related to high angle subcritical illumination is “Light Sheet Microscopy”, in that a relatively thin sheet of illumination (but not as thin as in TIR) is imposed on the sample at an adjustable height from the substrate. The optical setup can be elaborate, involving two objectives trained upon the sample from orthogonal directions, one for sheet illumination, the other for detection, with a specially configured sample holder at the intersection of the two objectives’ focal planes (Weber *et al.*, 2014). Light-sheet microscopy has also been adapted for two-photon excitation, which produces a thinner effective sheet with less background (Zong *et al.*, 2015).

Even without varying the incidence angle, but just by setting it a slightly supercritical angle (to achieve a fairly deep evanescent field), the z-position of discrete sources can be deduced (at least over a couple of wavelength distance range), by analyzing the height and breadth of point spread function of the source (Mondal and Hess, 2017).

Two Photon TIRF

All of the above variants of TIRF are in the realm of linear optics (except for the very last one), in which the observed fluorescence is proportional to the first power of the illumination intensity. Multiphoton microscopy, by producing emission proportional to at least the second power of the illumination, has been used to greatly improve axial resolution in scanning focal-spot microscopies. It is reasonable to think that two-photon TIRF excitation should have the ability to halve the effective evanescent field depth and thereby produce even more z-selectivity than standard TIRF. In practice, the standard multiphoton illumination source is a pulsed laser with a near IR emission. In linear mode, that would produce a deeper evanescent field than would visible light, because the evanescent field depth is proportional to wavelength. In two-photon mode, the evanescent field depth is indeed halved, but that only brings it back down to what would have been obtained from single photon visible light.

The real advantage of multiphoton TIRF is further suppression of background scattering. Scattering from a variety of sources (see sections below) can be significant, but generally it is less intense than the evanescent field itself (at least near $z=0$). With its non-linear favoring of the brightest places, multiphoton techniques give an advantage to the pure evanescent field over background scattering (Oheim and Schapper, 2005; Lane *et al.*, 2012).

The Reflected Beam in TIR

The depth of the evanescent field is a function of the local refractive index on the low-index side of the interface. For any fixed supercritical incidence angle and particular substrate, the depth increases as the refractive index increases, and the dependence is

rather steep for barely supercritical angles. Unfortunately, the sample's refractive index is usually not known accurately, and may even vary laterally with heterogeneous samples. TIR itself, without the fluorescence, can provide the unknown information.

In TIRF, attention is on the evanescent field, and upon the detection of fluorescence that it excites. But the (nonfluorescent) reflected beam itself contains information about the sample. By replacing the standard dichroic mirror (used to reflect the incident beam up through the objective to the sample and block the transmission of the reflected beam) with a standard 100% mirror positioned only partway across the aperture, the reflected beam can be captured. The microscope/detector optical system that normally would image the fluorescence will now image the reflected beam itself. The image is not sharp (because only a fraction of the objective aperture is used), but it has an interesting feature. At image locations corresponding to locations in the sample where the refractive index is low (say, off-cell on a cell culture), TIR is achieved and the reflection is bright whereas locations with low refractive index are subcritical and the reflection is dim. As the incidence angle is increased stepwise over a stack of images, the transition from subcritical to supercritical is quite distinct, and the transition occurs over a specific narrow range of angles that is very sensitive to local refractive index. By analyzing the image stack, the spatially-resolved refractive indices of a cell culture's contact regions with the substrate can be calculated. A bonus result is a measure of the separation between the bottom of the cell and the substrate because increased separation affects the steepness of the local reflected intensity increase with angle ([Bohannon et al., 2017](#)).

In principle, the reflected beam's polarization also contains information about sample refractive index. If the incident beam is linear polarized at 45 deg (i.e., half s-pol and half p-pol), the two polarizations suffer different phase delays upon TIR. The reflected beam then will be elliptically polarized, to an extent that depends on local refractive index of the sample.

A coherent collimated supercritical angle incident beam before the TIR surface is a plane wave, but interaction with the laterally variable sample index sample distorts the wavefront flatness. By interfering the reflected beam with a plane wave reference beam from the same source, a hologram of the sample can be recorded digitally ([Ash et al., 2010](#)).

Supercritical Angle Emission

As discussed in another article in this Encyclopedia, an excited fluorophore embedded in a low refractive index near a higher index planar substrate will produce an emission pattern that is very different from what it would produce in isolation: see [Axelrod \(2013\)](#) for mathematical details. Part of this effect is simple refraction of the emitted light into the substrate, thereby changing its propagation direction. All of the propagating emission light is bent into angles that are "subcritical", leaving higher polar emission "supercritical" angles dark. But a major part of the effect is due to the existence of the fluorophore's "near field", evanescent light that does not propagate away from the fluorophore vicinity. If the substrate is close enough (within about a wavelength), it will capture some of the near field, convert it to propagating light, and send it into those high supercritical polar angles in the substrate material. In fact, the brightest emission propagates exactly at the border between sub and supercritical, at the "critical" angle itself. The exact critical angle is quantitatively related to the refractive index of the material in which the fluorophore is embedded.

To measure that critical angle, the sample's emission is collected with a high aperture objective, often the same as one that would be used for TIR excitation (i.e., NA generally greater than 1.4). The objective's back focal plane reveals a one-to-one mapping of emission angles into radial positions. By recording images of the brightest ring in the back focal plane, and measuring its radius, the sample's refractive index can be computed ([Brunstein et al., 2017](#)). Unfortunately, spatial information is lost, because emission light from any location in the image plane passes through all areas of the back focal plane. However, if the excitation region on the sample could be restricted and scanned, then spatially-localized refractive index information could be recovered.

TIR Scattering

An important limitation of TIRF is scattering of the incident light. Part of this scattering originates in the excitation path of the optical system itself, particular in objective-based TIRF. [Mattheyses and Axelrod \(2005\)](#) found this to constitute about 10% of the intensity at $z=0$. This is probably much less in prism-type TIRF because the excitation path there is largely out of the field of view. [Brunstein et al. \(2014a,b\)](#) found that the scattering originating in the optical system is generally larger than the scattering of evanescent light by a cell culture sample, at least for flat and confluent cultures. However, for thicker or denser cells, sample-induced scattering could be greater. This is a problem because the scattered light can excite fluorescence at z -distances well beyond the reach of the evanescent field itself.

One kind of scattering is generated when an "object", a portion of the sample generally larger than one wavelength, has a refractive index too high to support TIR at the incidence angle used. That higher refractive index can be spaced away from the TIR surface within about a wavelength, and still capture some of the evanescent field energy and convert it to propagating light: this is called "frustrated TIR". That propagating light can then scatter or even become trapped in structures that cause it to total reflect multiple times (such as in spherical beads). Another complication occurs because frustrated TIR disrupts the evanescent field, leaving downbeam objects in the sample in a shadow. It also can lead to multiple reflections between the sample object and the TIR surface, so that the object no longer "sees" a pure exponentially decaying field.

Of particular interest is scattering resulting from refractive index gradients in the sample, even if all regions have a low enough refractive index to support TIR at the given incidence angle. The simplest case is a sphere residing in an evanescent field. (The

sphere may be considered a simplified model of a biological cell in a TIRF experiment.) This is a special case of Mie scattering, which results from a propagating plane wave hitting a dielectric object larger than a small fraction of the wavelength. The results of such a calculation ([Ganic et al., 2002](#)) show that the scattering pattern is very different for propagating plane wave scattering vs. evanescent scattering; much of the symmetry is lost in the evanescent case. The angular pattern is also very dependent on the incident polarization. An alternative calculation ([Bekshaev et al., 2013](#)) recasts the plane wave to have an imaginary incidence angle; this is equivalent to an evanescent wave. In both cases, the results explicitly shown are for the case where the refractive index of the spheres is high enough to cause “frustrated” TIR, but the mathematical techniques should be valid also for spheres with much lower index. Neither approach accounts for optical interactions with the TIR surface (such as multiple reflections, capture of evanescent fields on the surface of the sphere by the TIR substrate, etc.), nor for multiple scattering, refocusing, and shadowing among nearby spheres. There is room for useful theoretical work in this area.

Interactions between the TIR surface and the sphere are taken into account by [Liu et al. \(2000\)](#). The radius of the sphere, in combination with the wavelength, leads to sharp resonances in the intensity of scattering, which also depends on separation distance of the sphere to the substrate compared to the radius. The calculations, however, assumed that the sphere was made of the same material as the planar substrate, and therefore is not directly relevant to practical TIRF, where the biological cell under view has a refractive index much lower than the TIR substrate so that an evanescent field can exist inside the cell.

An alternative approach to the TIR scattering problem dispenses with the single sphere and instead postulates a rough surface ([Nieto-Vesperinas and Sanchez-Gil, 1992](#)). In the theory, the roughness could be periodic, or a known function, or a randomly generated function. However, predicting scattering in practical TIRF is a somewhat different case, because the “roughness” is not in the (high index) substrate, but in a lower index layer (the cells) adhered to a smooth substrate.

TIR fluorescence microscopy was originally an outgrowth of prism-based TIR illumination in which the (non-fluorescent) scattering was collected and imaged through a microscope ([Ambrose, 1956](#)). TIR scattering, both static and dynamic (in which coherent scattered light from multiple scattering centers interferes and causes random intensity fluctuations), can be used to observe the behavior – positions and motions – of small objects or colloidal particles near the TIR surface ([Sigel, 2009](#)).

TIR Combined With FRAP or FCS or FLIM

A very early application of TIRF microscopy was actually not for imaging at all, but for measuring the dynamics of biomolecules at surfaces: on/off chemical kinetic rates and surface diffusion. When looking at a fluorescent solution with subcritical angle illumination (“EPI”) through a microscope, the fluorescence looks unchanging and uniform (apart from interference fringes if a coherent laser source is used). But when the incidence angle is increased to the critical angle and beyond, the view quickly changes to a more speckly appearance that also “twinkles” as molecules diffuse in and out of the thin evanescent region. By measuring the amplitude and rapidity of these fluctuations in a small lateral region (by autocorrelation of the fluorescence intensity), the absolute concentration, the reversible binding rates, and any surface diffusion can all be inferred ([Thompson et al., 1981, 2009](#); [Thompson and Steele, 2007](#)), a technique called TIR-Fluorescence Correlation Spectroscopy (TIR-FCS).

A different but related procedure, called TIR-Fluorescence Recovery After Photobleaching (TIR-FRAP) also measures kinetic rates and diffusion but calls for photobleaching the region near the substrate with a bright flash of TIR “bleach” illumination and then monitoring the post-bleach fluorescence recovery with a much dimmer TIR “probe” illumination. To see diffusion in a short experimental time, it is useful to reduce the area of illumination. One way is to shape it into a line ([Burghardt and Axelrod, 1981](#)). Another is to intersect two TIR beams, forming interference stripes, much in the manner of structured illumination but for a different purpose ([Fulbright and Axelrod, 1993](#)).

TIR-FRAP has been applied to measuring the diffusion of fluorescent-marked proteins inside secretory granules of living cells ([Ngatchou-Weiss et al., 2014](#)). These spherical granules are small (~ 300 nm in diameter), but the evanescent field depth is even smaller (usually <100 nm). Therefore, the TIR bleaching flash will photobleach mainly those proteins in the part of the granule most proximal to the TIR substrate. Recovery will occur during the probe phase of the experiment due to diffusive redistribution of protein within the granule, and its rate can be interpreted in terms of a diffusion coefficient. Care must be taken, as in all FRAP experiments, to account for spontaneous “reversible” recovery after photobleaching that may occur in the millisecond to tens of millisecond time scale, even in the absence of diffusion.

The nearby presence of a dielectric surface can affect the fluorescence lifetime of a fluorophore. This occurs by several possible mechanisms. One of them is the interaction of the surface with the fluorophore’s near field. This provides an additional path for energy to be drained from the excited state, shortening the lifetimes. Another mechanism is the change in the chemical environment near a surface: viscosity, ionic strength, accessibility of quenchers, etc. Since much biological fluorescence microscopy observes cells very near a supporting substrate (the coverslip), fluorescence lifetime imaging microscopy (FLIM) techniques have been combined with TIRF to specifically study the surface effects. The setup and applications are reviewed in [Gadella \(2009\)](#).

TIRF on Single Molecules

Detection of single molecules with single attached fluorophores is challenging in part because the number of photons available from a single fluorescent molecule before it photobleaches away is quite limited, typically less than one million, and many of these

photons may be absorbed or may miss the detector system. With few available photons, shot noise in the signal (due to the randomness of photon emission and detection) makes it imperative that other extraneous noise sources be reduced. One of the most common features in fluorescence microscopy is background, and even if it is constant in intensity, it can still produce its own shot noise that will bury the signal from the single fluorophore. TIRF has been effective in suppressing background, so it is used commonly in single molecule fluorescence detection. For reviews of the single molecule field in general, see [Lord et al. \(2010\)](#) and for a theoretical focus on the optical aspects, see [Axelrod \(2012\)](#).

A particularly unique aspect of single molecule studies is the ability to “see” the orientation of an immobilized molecule, not just to detect its existence. Some information about orientation can be gleaned from viewing the emission focused at the sample plane with polarizing filters, but the information can be ambiguous. As an alternative, analysis of the azimuthal distribution of emission at the back focal plane (BFP) of the objective ([Burghardt and Ajtai, 2009](#)) removes the ambiguity.

See also: Overview: Microscopy

References

- Ambrose, E.J., 1956. A surface contact microscope for the study of cell movements. *Nature* 178, 1194.
- Ash III, W.M., Clark, D., Lo, C.-M., Kim, M.K., 2010. Total internal reflection holographic microscopy for quantitative phase characterization of cellular adhesion. *Proc. of SPIE* 7568. doi:10.1117/12.842470.
- Axelrod, D., 2012. Fluorescence excitation and imaging of single molecules near dielectric-coated and bare surfaces: A theoretical study. *J. Microscopy* 247, 147–160.
- Axelrod, D., 2013. Evanescent excitation and emission in fluorescence microscopy. (Invited review) *Biophys. J.* 104, 1401–1409. (Correction, 104:2321).
- Axelrod, D., 2014. Evanescent excitation and emission. In: Cornea, A., Conn, P.M. (Eds.), *Fluorescence Microscopy: Super-Resolution and Other Novel Techniques*. Elsevier, pp. 1–14.
- Axelrod, D., 2016. Total internal reflection fluorescence microscopy. In: Bradshaw, R.A., Stahl, P.D. (Eds.), *Encyclopedia of Cell Biology 2*. Academic Press, pp. 62–69.
- Beach, J.R., Shao, L., Remmert, K., et al., 2014. Nonmuscle myosin II isoforms coassemble in living cells. *Curr. Biol.* 24, 1160–1166.
- Bekshaev, A.Y., Bliokh, K.Y., Nori, F., 2013. Mie scattering and optical forces from evanescent fields: A complex-angle approach. *Opt. Express* 21, 7082–7095.
- Beresna, M., Gecevičius, M., Kazansky, P.G., Gertus, T., 2011. Radially polarized optical vortex converter created by femtosecond laser nanostructuring of glass. *Appl. Phys. Lett.* 98. doi:10.1063/1.3590716.
- Bohannon, K.P., Holz, R.W., Axelrod, D., 2017. Refractive index imaging of cells with variable-angle near-TIR microscopy. *Microsc. Microanal.* 1–11. doi:10.1017/S1431927617012570.
- Brunstein, M., He’rault, K., Oheim, M., 2014b. Eliminating unwanted far field excitation in objective-type tifr. Part II. Combined evanescent-wave excitation and supercritical-angle fluorescence detection improves optical sectioning. *Biophys. J.* 106, 1044–1056.
- Brunstein, M., Roy, L., Oheim, M., 2017. Near-membrane refractometry using supercritical angle fluorescence. *Biophys. J.* 112, 1940–1948.
- Brunstein, M., Teremetz, M., He’rault, K., Tourain, C., Oheim, M., 2014a. Eliminating unwanted far-field excitation in objective-type TIRF. Part I. Identifying sources of nonevanescent excitation light. *Biophys. J.* 106, 1020–1032.
- Burghardt, T.P., Ajtai, K., 2009. Mapping microscope object polarized emission to the back focal plane pattern. *J. Biomed. Opt.* 14, 034036.
- Burghardt, T.P., Axelrod, D., 1981. Total internal reflection/fluorescence photobleaching recovery study of serum albumin adsorption dynamics. *Biophys. J.* 33, 455–468.
- Chon, J.W.M., Gu, M., 2004. Scanning total internal reflection fluorescence microscopy under one-photon and two-photon excitation: Image formation. *Appl. Opt.* 43, 1063–1071.
- Chung, E., Kim, D., Cui, Y., Kim, Y.-H., So, P.T.C., 2007. Two-dimensional standing wave total internal reflection fluorescence microscopy: Superresolution imaging of single molecular and biological specimens. *Biophys. J.* 93, 1747–1757.
- Fiolka, R., Shao, L., Rego, E.H., Davidson, M.W., Gustafsson, M.G., 2012. Time-lapse two-color 3D imaging of live cells with doubled resolution using structured illumination. *Proc. Natl. Acad. Sci. USA* 109, 5311–5315.
- Fulbright, R.M., Axelrod, D., 1993. Dynamics of nonspecific adsorption of insulin to erythrocyte membrane. *J. Fluor.* 3, 1–16.
- Gadella, T.W.J., Jr., (Ed.), 2009. Total internal reflection fluorescence lifetime imaging microscopy. In: *Laboratory Techniques in Biochemistry and Molecular Biology*, 33.
- Ganic, D., Gan, X., Gu, M., 2002. Three-dimensional evanescent wave scattering by dielectric particles. *Optik* 113, 135–141.
- Gustafsson, M.G.L., Agard, D.A., Sedat, J.W., 2000. Doubling the lateral resolution of wide-field fluorescence microscopy using structured illumination. *Proc. SPIE* 3919, 141–150.
- Hirvonen, L.M., Wicker, K., Mandula, O., Heintzmann, R., 2009. Structured illumination microscopy of a living cell. *Eur. Biophys. J.* 38, 807–812.
- Johnson, D.S., Toledo-Crow, R., Mattheyses, A.L., Simon, S.M., 2014. Polarization-controlled TIRFM with focal drift and spatial field intensity correction. *Biophys. J.* 106, 1008–1019.
- KieSSLing, V., Domanska, M.K., Tamm, L.K., 2010. Single SNARE-mediated vesicle fusion observed in vitro by polarized TIRFM. *Biophys. J.* 99, 4047–4055.
- Kner, P., Chhun, B.B., Griffis, E.R., Winoto, L., Gustafsson, M.G.L., 2009. Super-resolution video microscopy of live cells by structured illumination. *Nat. Methods* 6, 339–342.
- Konopka, C.A., Bednarek, S.Y., 2008. Variable-angle epifluorescence microscopy: A new way to look at protein dynamics in the plant cell cortex. *Plant J.* 53, 186–196.
- Lane, R.S.K., Macpherson, A.N., Magennis, S.W., 2012. Signal enhancement in multiphoton TIRF microscopy by shaping of broadband femtosecond pulses. *Opt. Express* 25948–25959.
- Liu, C., Weigel, T., Schweiger, G., 2000. Structural resonances in a dielectric sphere on a dielectric surface illuminated by an evanescent wave. *Opt. Commun.* 185, 249–261.
- Lord, S.J., Lee, H.-L.D., Moerner, W.E., 2010. Single-molecule spectroscopy and imaging of biomolecules in living cells. *Anal. Chem.* 82, 2192–2203.
- Lu, W., Xia, T., Xu, L., Chen, Y.-G., Fang, X., 2013. Visualization of the post-Golgi vesicle mediated transportation of TGF- β receptor II by quasi-TIRFM. *J. Biophotonics* 7, 788–798.
- Mattheyses, A., Axelrod, D., 2005. Direct measurement of evanescent field profile and depth in TIRF microscopy. *J. Biomed. Opt.* 11, 014006.
- Mondal, P.P., Hess, S.T., 2017. Total internal reflection fluorescence based multiplane localization microscopy enables super-resolved volume imaging. *Appl. Phys. Lett.* doi:10.1063/1.4983786.
- Ngatchou-Weiss, A., Bittner, M.A., Holz, R.W., Axelrod, D., 2014. Protein mobility within secretory granules. *Biophys. J.* 107, 16–25.
- Nieto-Vesperinas, M., Sanchez-Gil, J.A., 1992. Light scattering from a random rough interface with total internal reflection. *J. Opt. Soc. Am. A* 9, 424–436.
- Oheim, M., Loerke, D., Preitz, B., Stuhmer, W., 1998. A simple optical configuration for depth-resolved imaging using variable angle evanescent-wave microscopy. “Optical Microscopy”. *SPIE* 3568, 131–140.
- Oheim, M., Schapper, F., 2005. Non-linear evanescent-field imaging. *J. Phys. D Appl. Phys.* 38, R185–R197.

- Oliveczky, B.P., Periasamy, N., Verkman, A.S., 1997. Mapping fluorophore distributions in three dimensions by quantitative multiple angle-total internal reflection fluorescence microscopy. *Biophys. J.* 73, 2836–2847.
- Oreopoulos, J., Yip, C.M., 2009. Combinatorial microscopy for the study of protein–membrane interactions in supported lipid bilayers: Order parameter measurements by combined polarized TIRFM/AFM. *J. Struct. Biol.* 168, 21–36.
- Rohrbach, A., 2000. Observing secretory granules with a multiangle evanescent wave microscope. *Biophys. J.* 78, 2641–2654.
- Schneckenburger, H., 2005. Total internal reflection fluorescence microscopy: Technical innovations and novel applications. *Curr. Opin. Biotechnol.* 16, 13–18.
- Sentenac, A., Belkebir, K., Giovannini, H., Chaumet, P.C., 2009. High-resolution total internal-reflection fluorescence microscopy using periodically nanostructured glass slides. *J. Opt. Soc. Am. A* 26, 2550–2557.
- Sigel, R., 2009. Light scattering near and from interfaces using evanescent wave and ellipsometric light scattering. *Curr. Opin. Colloid Interface Sci.* 14, 426–437.
- Steyer, J.A., Almers, W., 1999. Tracking single secretory granules in live chromaffin cells by evanescent-field fluorescence microscopy. *Biophys. J.* 76, 2262–2271.
- Sun, W., Xu, A., Marchuk, K., Wang, G., Fang, N., 2011. Whole-cell scan using automatic variable-angle and variable-illumination-depth pseudo-total internal reflection fluorescence microscopy. *J. Lab. Autom.* 255–262.
- Terakado, G., Watanabe, K., Kano, H., 2009. Scanning confocal total internal reflection fluorescence microscopy by using radial polarization in the illumination system. *Appl. Opt.* 48, 1114–1118.
- Thompson, N.L., Burghardt, T.P., Axelrod, D., 1981. Measuring surface dynamics of biomolecules by total internal reflection with photobleaching recovery or correlation spectroscopy. *Biophys. J.* 33, 435–454.
- Thompson, N.L., Steele, B.L., 2007. Total internal reflection with fluorescence correlation spectroscopy. *Nat. Protocols* 2, 878–890.
- Thompson, N.L., Wang, X., Navaratnarajah, P., 2009. Total internal reflection with fluorescence correlation spectroscopy: Applications to substrate-supported planar membranes. *J. Struct. Biol.* 168, 95–106.
- Tokunaga, M., Imamoto, N., Sakata-Sogawa, K., 2008. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nat. Methods* 5, 159–161. (Addendum, 455).
- Weber, M., Mickoleit, M., Huisken, J., 2014. Light sheet microscopy. *Methods Cell Biol.* 123, 193–215.
- Zong, W., Zhao, J., Chen, X., *et al.*, 2015. Large-field high-resolution two-photon digital scanned light-sheet microscopy. *Cell Res.* 25, 254–257.

Further Reading

- Axelrod, D., 2007. Total internal reflection fluorescence microscopy. In: Török, P., Kao, F.-J. (Eds.), *Optical Imaging and Microscopy: Techniques and Advanced Systems*, Springer Series in Optical Sciences., second ed., Berlin: Springer, pp. 195–236.

Super-Resolution Depth Measurements: Variable Angle TIRF, Super-Critical Angle Fluorescence, MIET

Alexey Chizhik, University of Göttingen, Göttingen, Germany

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Fluorescence microscopy is one of the most commonly used techniques for studying biological samples. It allows one to observe live samples in real time; the technical simplicity of fluorescence imaging techniques makes it accessible for a broad community of life-science researchers; and finally, specific labelling allows one to directly visualize sub-cellular structures. Moreover, in contrast to X-ray, electron, or atomic force microscopy, optical microscopy allows one to perform 3D imaging of thick samples up to hundreds of micrometers. However, the wave nature of light limits the spatial resolution of a conventional fluorescence microscope, i.e., it cannot resolve structures smaller than the diffraction limit. In the visible spectral range, this corresponds to a spatial resolution of roughly half a micrometer along the optical (z -) axis, and of about a quarter of a micrometer in the x - y plane. Abbe's equation,

$$d = \frac{\lambda}{2n \sin \alpha} \quad (1)$$

which was formulated in 1873, relates the resolution d of a lens with its numerical aperture $NA = n \cdot \sin(\alpha)$, and with the wavelength λ of light. **Fig. 1** shows a simulated diffraction limited focal spot formed by focusing a linearly polarized light of 488 nm with a perfectly designed lens of 1.2 NA into water. These are the parameters of a typical microscope that is used for the biological imaging, and the size of the shown focal spot defines the resolution of the system.

For more than a century, the diffraction barrier imposed by the wave nature of light seemed to be insurmountable. The fundamental step towards a change of perception of optical resolution was made only by the end of the 20th century, when 4Pi microscopy enhanced the axial (along optical axis) resolution by almost a factor of 7 (Hell and Stelzer, 1992). This invention was followed by the invention of numerous fluorescence imaging techniques, which meanwhile allow one to achieve nearly nanometer resolution (Hell, 2003, 2009; Huang *et al.*, 2009). All existing super-resolution microscopy methods can be mostly divided into the two general groups according to their fundamental principle of resolution enhancement: deterministic, and stochastic techniques (Hell, 2003, 2009; Huang *et al.*, 2009). Increase of optical resolution in deterministic methods is based on the non-linearity of fluorescence excitation. This group of methods includes STED, I⁵M, RESOLFT, SIM and their derivatives. The second group of super-resolution imaging techniques employs reversible stochastic photo-switching of fluorophores between photo-active and inactive states, which can be partially tuned using various physico-chemical parameters. Among these methods are STORM, dSTORM, PALM, GSD, SPDM, SOFI and variants thereof.

Whereas some of the above techniques squeeze the lateral resolution down to some tens of nanometers, the much less axial resolution remains a key limiting factor for applications where z -sectioning of a sample is required. This drawback of almost all super-resolution microscopy methods stimulated the appearance of another class of high-resolution axial imaging methods, either independent or complementary to the methods that enhance lateral resolution.

To avoid confusion, it should be noted that not all of the high-resolution axial imaging methods, which will be discussed below, can be technically attributed to the family of super-resolution microscopy. Whereas the term "super-resolution microscopy" implies a resolution of two or more closely located emitters located *on the same axis*, these techniques axially localize *single*

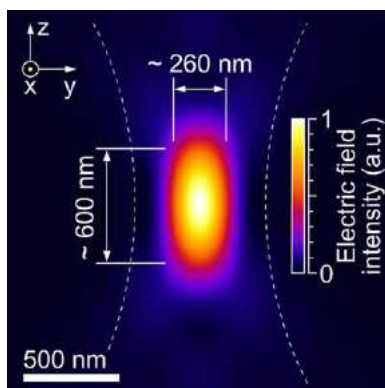


Fig. 1 The side view of the simulated diffraction limited focal spot created by 488 nm linearly polarized light when focused in water with a 1.2 NA objective lens. The dashed curves schematically show the direction of light focusing. Calculations are done by the methods of Jörg Enderlein, see Ref. Enderlein *et al.* (1999).

fluorophores (or in more general case fine structures) with nanometer accuracy. Later in the text, we will not emphasize this difference; however, it is important to keep it in mind to understand the limitations of the techniques discussed below.

Existing high-resolution axial imaging methods are based on several general approaches. First, the axial localization of a fluorophore can be determined by detecting it with different positions of the focal plane (biplane imaging (Juette *et al.*, 2008)), or by tailoring the shape of the point spread function (PSF) (astigmatic imaging (Huang *et al.*, 2008), helical wavefront shaping (Pavani *et al.*, 2009)). With such methods, one achieves ca. 70 nm axial resolution and ca. 30 nm lateral resolution. In all cases, the axial resolution is typically worse by a factor of 3–5 than the lateral one.

The second general approach is deducing the axial position of an emitter from relative intensities of single-fluorophore images when employing an interferometric detection system. Interferometric PALM (iPALM) (Shtengel *et al.*, 2009) and 4pi-STORM (Aquino *et al.*, 2011) approach single nanometer z-localization accuracy. However, the high technical complexity of the required optical system, and the need for tedious sample preparation and adjustment limit their applicability for routine microscopy measurements. Technically simpler coherent optical microscopy methods such as RCM (Limozin and Sengupta, 2009), or iSCAT (Kukura *et al.*, 2009) routinely achieve nanometer or even sub-nanometer localization accuracy of scattering/reflecting objects along the optical axis. However, the drawbacks of these methods are the lack of specificity of the imaged structures, and the reduced sensitivity when compared with most fluorescence-based methods.

In this article, we will focus on three distinct techniques for high resolution axial imaging, which combine unprecedented nanometer resolution with technical simplicity: Supercritical Angle Fluorescence (SAF) (Ruckstuhl and Verdes, 2004; Deschamps *et al.*, 2014), variable-angle Total Internal Reflection Fluorescence (TIRF) (Cardoso Dos Santos *et al.*, 2016; Stock *et al.*, 2003), and Metal Induced Energy Transfer (MIET) (Chizhik *et al.*, 2014; Karedla *et al.*, 2014; Baronsky *et al.*, 2017) microscopy. Despite the differences in their technical principle, all three methods are based on the same common effect of evanescence of light close to an interface. This general principle is in contrast to all other super-resolution microscopy techniques, which achieve resolution enhancement by reversible switching between different states of fluorescence emitters (either stochastically or deterministically). The axial localization methods discussed here are based on evanescent light, either in emission (SAF and MIET), or in excitation (variable angle TIRF).

In this article, we will discuss the physical principle of these methods, their technical implementation, applications, advantages, limitations, and potential future perspectives. In view of the numerous excellent articles and comprehensive reviews about each of these techniques, we will here not focus much on the mathematical formalism of the methods, but rather on recent technical implementations and an overview of recent applications that have not been discussed in other reviews.

Variable-Angle Total Internal Reflection Fluorescence Microscopy

In contrast to standard epifluorescence microscopy, in total internal reflection fluorescence (TIRF) microscopy the incident angle of the plane-wave excitation light is chosen above the critical angle of total internal reflection between the coverslide and the sample medium (see Fig. 2). Due to the fact that the refractive index n_2 of an aqueous solution is lower than that of glass, i.e., $n_1 > n_2$, this results in total internal reflection of the incident light at all angles that are greater than the critical angle

$$\theta_c = \arcsin(n_2/n_1) \quad (2)$$

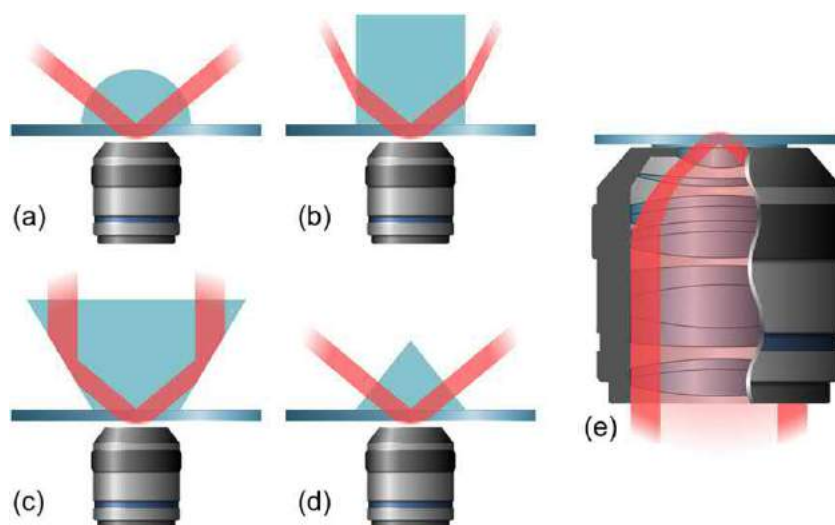


Fig. 2 (a)–(d) Schematic of various prism-based TIRF excitation/detection systems. (e) Schematic of prismless TIRF excitation/detection scheme, where a high NA objective lens is used for transmitting both excitation and emission light. The optical components inside the objective lens are not shown; the light path is shown schematically.

Despite being totally reflected, the incident light generates an evanescent electromagnetic field that extends into the sample and decays exponentially with the distance from the interface according to

$$I(z) = I_0 \exp[-z/d(\theta)] \quad (3)$$

The penetration depth $d(\theta)$ of the electric field $I(z)$ at wavelength λ is given by [Axelrod et al. \(1984\)](#)

$$d(\theta) = \frac{\lambda/4\pi}{\sqrt{n_1^2 \sin^2 \theta - n_2^2}} \quad (4)$$

Exponential decay of the field intensity results in a confinement of the excitation volume within ca. the first 100 nm above the sample surface. Confining illumination of a sample within a narrow region allows one to enhance the signal-to-noise ratio by suppressing the background signal and out of focus fluorescence and to limit the photo-damage of the specimen, thus making TIRF an ideal tool for imaging cellular structures near the sample surface. The idea of using TIRF to illuminate cells in close contact with the sample surface was first proposed by [Ambrose \(1956\)](#) and later developed by [Axelrod \(1981\)](#). We will omit here further details on TIRF microscopy and its variants, which are extensively described in the article by Daniel Axelrod. In the current article, we will focus on one of its advanced varieties that allow for extracting the fluorophore-substrate separation distance with axial resolution of 10 nm for certain types of samples.

The exponential decrease of the evanescent field upon TIRF illumination of the sample allows one to estimate the axial localization of a fluorophore by measuring its relative brightness. However, quantitative interpretation of a TIRF image for extracting the distance between an emitter and the substrate surface is far from being trivial ([Cardoso Dos Santos et al., 2016](#)). The contrast of TIRF images depends on several parameters, such as the local concentration of fluorophores, their orientation, absorption cross section, and angular emission distribution, as well as on local variations of the fluorophore's quantum yield and the sample's refractive index ([Chizhik et al., 2011](#); [Mertz, 2000](#)).

The situation becomes even more complicated by the fact that some of these parameters can vary as a function of height ([Chizhik et al., 2014](#); [Mertz, 2000](#); [Enderlein et al., 1999](#); [Enderlein, 1999](#)). This complication can be addressed by recording a series of TIRF images at different incident angles of the incident light. Since all the above parameters do not depend on θ , their values will not change by varying the angle of incidence. By including all the unknown parameters in a new factor $\chi(x, y, z)$, one can rewrite [Eq. \(3\)](#) as

$$I(x, y, z, \theta_i) = \chi(x, y, z) \gamma(\theta_i) \exp(-z/d(\theta_i)) \quad (5)$$

where $\gamma(\theta_i)$ is the correction factor that takes into account the variation of the field intensity at the substrate-sample interface as a function of the angle of incidence θ_i ([Axelrod et al., 2002](#)). Taking the logarithm of [Eq. \(5\)](#), one obtains

$$\ln \frac{I(x, y, z, \theta_i)}{\gamma(\theta_i) - \gamma(\theta_1)} = \ln[\chi(x, y, z) \gamma(\theta_1)] - \frac{z}{d(\theta_i)} \quad (6)$$

The logarithm on the left side of [Eq. \(6\)](#) depends linearly on $d^{-1}(\theta_i)$. Therefore, by using a linear fit in each pixel of the fluorescence image measured at different angles of incidence, one can retrieve the fluorophore-substrate distance, without requiring any information on parameters related to the molecular parameters, such as quantum yield, absorption cross-section, or concentration. This technique, known as variable-angle TIRF (vaTIRF), was introduced in the mid-eighties ([Lanni et al., 1985](#); [Gingell et al., 1985](#)) for studies of sub-cellular structures.

One of the advantages of vaTIRF is that it allows one to simultaneously determine the value of the effective refractive index (n_2 in [Eq. \(4\)](#)) of the sample, which includes the combined influence of any water gap between a cell membrane and the substrate surface, and the cytoplasm inside the cell. This index determines the penetration depth $d(\theta)$ of the evanescent field ([Cardoso Dos Santos et al., 2016](#)).

VaTIRF was used for quantification of the membrane height on focal adhesion zones ([Burmeister et al., 1994](#)), for mapping the topography of cells above a substrate ([Olveczky et al., 1997](#)), for studying axial motion of secretory granules in the ventral side of living cells ([Loerke et al., 2002](#)), and to investigate the influence of 5-aminolevulinic acid on the adhesion of tumor cells ([Lassalle et al., 2007](#)).

Experimentally, the first realizations of vaTIRF were based on an optical coupling of the substrate to a glass or quartz prism of cubic, trapezoidal, hemi-cylindrical or hemi-spherical shape ([Fig. 2\(a–b\)](#)). One of the key challenges is to keep the position of the illuminating light spot on the sample constant while changing the angle of incidence. This is possible only when using a hemi-spherical or hemi-cylindrical prism or a multiple focus laser scanning system ([Stock et al., 2003](#)). A drawback of all these vaTIRFM systems is that when a prism is used to generate the evanescent wave, the microscope objective is then mounted on the opposite side of the sample for fluorescence collection, which often diminishes the image quality, for instance in case of low sample transparency or strong light scattering. To address this issue, a prism-less approach in TIRF has been proposed ([Axelrod, 2001](#)). Indeed, a prism-less setup based on an inverted microscope equipped with a high numerical aperture (NA) objective ($NA > 1.4$) allows for obtaining images with the full resolution defined by the NA. In this configuration, the objective is used both to create the evanescent excitation and to collect the signal ([Fig. 2\(e\)](#)). The prism-less approach simplifies installation and adjustment of the optical system which is required for tuning incidence angle of the excitation light. Recently, [Boulanger et al. \(2014\)](#) demonstrated the superior capability of a prism-less setup in vaTIRFM by imaging the cortical actin network. Jaffiol and co-workers used a similar system for studying adhesion of MDA-MB-231 cancer cells to a glass substrate ([Cardoso Dos Santos et al., 2016](#)). By recording a

series of 10 TIRF images with different incidence angles of excitation light, the authors achieved an axial resolution close to 10–20 nm that makes vaTIRF suitable for quantifying cell adhesion structures or cell topography. Sequential vaTIRF imaging with selective photobleaching of molecules within the lower layers close to the interface can be used for straightforward observation of deeper structures within a sample without masking signals from farther distances by the brighter signals closer to the substrate (Fu *et al.*, 2016).

Super-Critical Angle Fluorescence Microscopy

Almost all fluorophores can be considered as ideal electric dipole emitters. Along with the far-field component, the dipole also has a near-field part which is not connected to far-field radiation, but which can strongly interact with nearby structure and interfaces. Let us consider a fluorophore that is located in close proximity to an interface separating two media with different refractive indices n_1 and n_2 . The electric field of the emitting fluorophore can be represented as a superposition of plane waves, where each of these plane waves is refracted according to Snell's law when travelling from medium with refractive index n_2 into the medium with refractive index n_1 ($n_2 < n_1$). As a result, the emission into medium 1 is restricted within a cone limited by the critical angle

$$\theta_c = \arcsin(n_2/n_1) \quad (7)$$

This component is referred to as under-critical angle fluorescence (UAF). However, if the fluorophore-interface distance is smaller than the fluorescence wavelength, then additional super-critical angle fluorescence (SAF) emission is observed (Enderlein *et al.*, 1999) due to the coupling of evanescent near-field components of the dipole field into medium 1. As a result, one can now also observe, in medium 1, emission along angles larger than the critical angle θ_c .

Fig. 3(a) and (b) show the angular distribution of emission for an isotropically oriented molecule that is located directly at the glass-water interface (a), or separated by a third of its emission wavelength (b). Thus, displacing the fluorophore by only a few tens of nanometers away from the interface leads to a strong angular redistribution of emission, a large part of which can be found

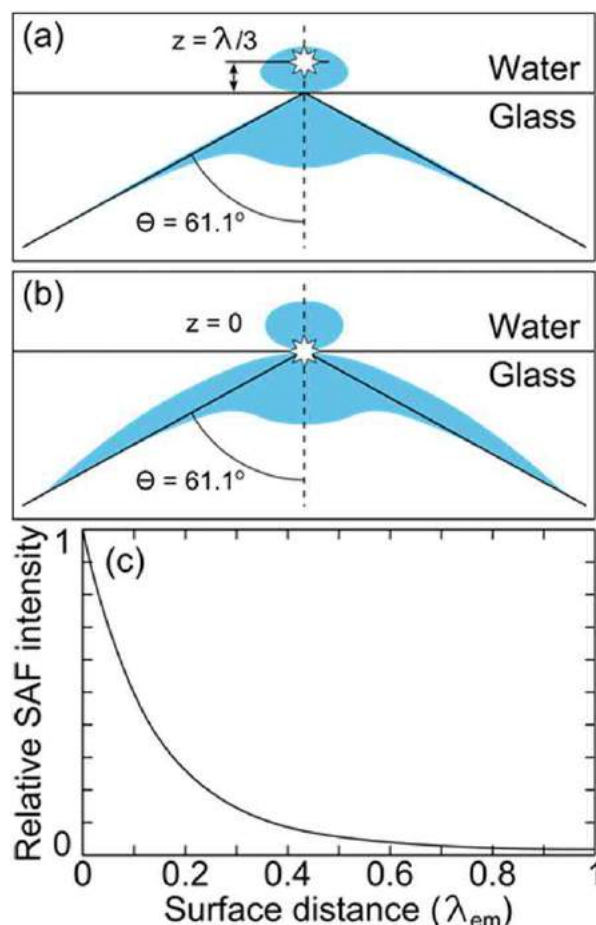


Fig. 3 Polar plots of the emission direction of molecules with isotropically oriented transition dipole (a) at the water/glass interface and (b) a surface distance equal to a third of the emission wavelength. (c) Dependence of SAF emission on the molecule's axial position. Calculations are done by the methods of Jörg Enderlein, see Ref. Enderlein *et al.* (1999).

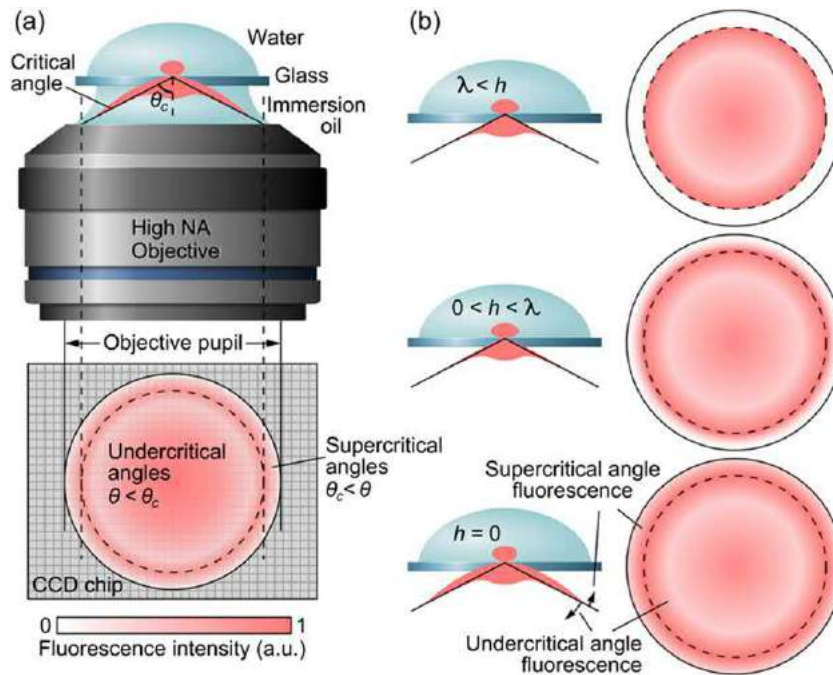


Fig. 4 (a) Schematic of the SAF and UAF fluorescence detection system. (b) Angular distribution of emission of a fluorophore at different axial positions and the corresponding distributions of the SAF and UAF at the back focal plane of the objective lens. Calculations are done by the methods of Jörg Enderlein, see Ref. Enderlein *et al.* (1999).

above the critical angle of 61.1° for the given ratio of refractive indices. This modulation of emission can be directly visualized by collecting the fluorescence with a high NA objective lens and measuring the profile of the collimated beam with a sensitive EMCCD chip. Fig. 4(a) shows a schematic of such a measurement. The dashed ring corresponds to the photons emitted at the critical angle. All the light which is detected inside this ring is the under-critical fluorescence. The light that is detected beyond the dashed ring corresponds to the super-critical fluorescence. Depending on the axial position of the emitter, the ratio of the super- and under-critical fluorescence changes (Fig. 4(b)).

It should be emphasized that the dependence between the axial position of the emitter and the angular distribution of the emission is analogous to and the reverse of the exponential decay of a TIRF excitation intensity. In that case, there is a close relationship between the angle of incidence and the penetration depth of the evanescent light above the interface, as we have seen in the previous section (vaTIRF microscopy).

The number of photons emitted beyond the critical angle is maximized when the fluorophore is directly bound to the surface, and it equals $\sim 34\%$ in case of a random dipole orientation at a glass-water interface (Ruckstuhl *et al.*, 2003). Moving the molecule away from the interface leads to a steep decrease of the intensity of supercritical fluorescence (Fig. 3(c)), which drops by a factor of 2 when the emitter is separated by nearly one sixth of the emission wavelength.

The quantitative measurement of the fluorophore's axial position requires either the determination of the supercritical-to-undercritical angle fluorescence intensity ratio, or the measurement of the absolute intensity of the supercritical angle fluorescence. The core idea of SAF microscopy and the first technical realization of the corresponding optical system which allows for collecting the fluorescence emitted above the critical angle was proposed by the group of Stefan Seeger (Enderlein *et al.*, 1999). The fluorescence detection system consisted of a parabolic glass element (Fig. 5(a)), which acts as a TIR mirror that collects the fluorescence light around the angles of maximum emission. Ruckstuhl *et al.* have used this light collection scheme for simultaneous measurement of both under- and supercritical angle fluorescence (see ref. Ruckstuhl and Verdes (2004)). One of the drawbacks of the system is that the paraboloid glass segment also serves as the substrate, which however can be simplified by using a coverslip that is placed on the paraboloid's front face with index-matching immersion oil. Recently, Hill *et al.* have proposed the use of low-cost disposable biochips for supercritical angle fluorescence detection for multiplexed clinical diagnostic applications (Hill *et al.*, 2011), which simplified the application of parabolic light collectors for routine measurements.

The more critical issues of using the parabolic light collector are low compatibility with standard microscopes and the impossibility to take wide-field images (Enderlein *et al.*, 2011). In recent years, the parabolic light collector has been replaced by high NA objectives which allow also for wide field imaging. For collecting all the under-critical fluorescence, one needs an objective the numerical aperture of which is equal or larger than the critical TIR angle between the sample and the objective's immersion medium. A simple calculation shows that for emitters bound to a glass-water interface, one has to use at least a NA of 1.33 for collecting all the under-critical angle fluorescence, which is distributed within a cone of opening angle 61.1° . Collection of the supercritical angle emission requires a large value of the numerical aperture. For this reason, SAF microscopy has become

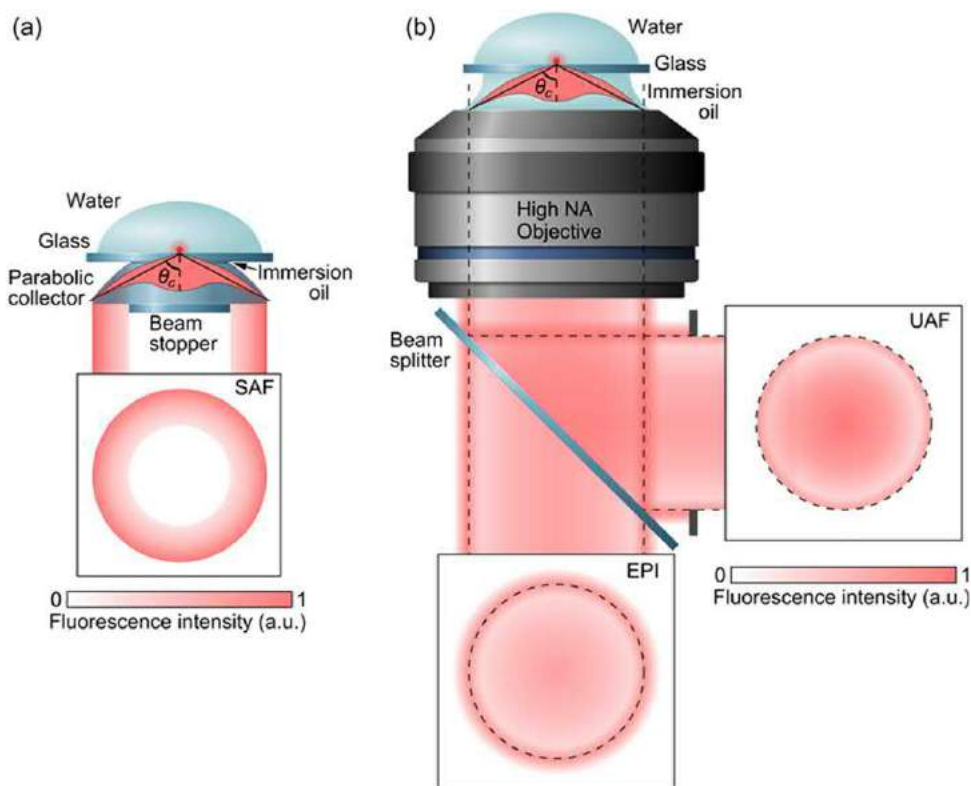


Fig. 5 (a) Schematic the parabolic collector for detection of the SAF. (b) Schematic of the high NA objective lens-based vSAF light collection scheme. Fluorescence is split in two parts: in the first detection channel, the EPI part is measured; in the second channel, only the UAF part is recorded (the SAF component is blocked in the back focal plane). As the result of the subtraction one obtains the virtual mask where only the SAF component is conserved.

widespread with the accessibility of TIRF objective lenses, which can possess NA values up to 1.49. This corresponds to a collection angle of about 79° and allows one to collect nearly all the supercritical angle fluorescence.

The first realization of SAF microscopy based on collection of both super- and under-critical fluorescence with a commercial 1.49 NA objective lens has been demonstrated in the group of Emmanuel Fort (Barroca *et al.*, 2012). A modified method that is referred to as virtual SAF (vSAF) is based on measuring fluorescence wide-field images both over all emission angles, and only over UAF emission angles (Fig. 5(b)). The latter was done by mechanically blocking the SAF part. As a result, the images that are obtained with the full angular distribution of emission allow for maximizing the lateral resolution, while calculating the difference of emission in the two channels results in the SAF image.

The use of a standard objective lens for SAF microscopy allows one to combine SAF microscopy which enhances axial resolution with super-resolution imaging techniques that increase the lateral resolution. By combining dSTORM and SAF microscopy, Bourg *et al.* have achieved close to 20 nm isotropic three-dimensional localization precision of fluorescent molecules in the basal cell membrane and on F-actin fibers within the first 150 nm above the substrate surface (Bourg *et al.*, 2015). Simultaneously, Ries and co-workers have used a similar experimental approach for imaging immunostained clathrin-coated pits and microtubules (Deschamps *et al.*, 2014).

Metal-Induced Energy Transfer Microscopy

As was predicted by Purcell (1946), placing a fluorescent molecule in the vicinity of a metal surface quenches its fluorescence and decreases its excited state lifetime. The mechanism of this phenomenon is similar to that of Förster resonance energy transfer (FRET): energy from the excited molecule is transferred, via electromagnetic coupling, into plasmons of the metal, where energy is either dissipated or re-radiated as light. The fluorophore-metal interaction was extensively studied in the 1970s and 1980s (Drexhage, 1974), and a quantitative theory of this interaction was developed on the basis of semi-classical quantum optics (Enderlein, 1999; Barnes, 1998). One of the key parameters that determine the character of the interaction between a quantum emitter and a metal structure is the size of the structure. If the metallic nanostructure is smaller than the wavelength of light that is emitted by the fluorophore, the fluorophore's radiation leads to a uniform oscillatory displacement of electrons inside the nanostructure, leading to strong restoring forces that act on these charges. These restoring forces become biggest at a certain

oscillation frequency, which is called plasmon resonance. Furthermore, these coherent oscillations of free electrons in metallic nanoparticles generate a highly localized electric field in close proximity to the particle's surface. Depending on the type of metal or the nanoparticle's shape and size, the plasmon resonance can be tuned to different spectral ranges for maximizing fluorophore-particle interaction. Interaction of fluorophores with localized surface plasmons has found numerous applications, with fluorescence enhancement being the most important one. Here, we will not further focus on the details of this kind of fluorophore-metal interaction and redirect the reader to the numerous existing extensive reviews.

A fundamentally different type of the fluorophore-metal interaction occurs when the fluorophore is located in close proximity to a smooth metal surface. The fluorophore's radiation creates charge density oscillations in the metal surface that are called propagating plasmons. Here, we will show how the interaction of fluorophore with a smooth metal surface can be used for axial localization of a fluorophore with nanometer accuracy. The core idea is that this interaction accelerates the return of an excited fluorescent molecule to its ground state, which manifests itself as a shortening of its excited state lifetime (also called fluorescence lifetime) (Chizhik *et al.*, 2014; Karedla *et al.*, 2014). Owing to the fact that the energy transfer rate is dependent on the distance of a molecule from the metal layer, the fluorescence lifetime can be directly converted into a distance value (Fig. 6). The theoretical basis for this conversion is the perfect quantitative understanding of the interaction between a quantum emitter and a metallic surface (Enderlein, 1999). It is important to emphasize that the energy transfer from the molecule to the metal is dominated by the interaction of the molecule's near-field with the metal, and is thus similar to FRET. However, due to the planar geometry of the metal film, which acts as the acceptor, the distance dependency of the energy transfer efficiency is much weaker than the sixth power of the distance, and shows a monotonous relation between lifetime and distance within roughly the first 150–200 nm above the surface for fluorescence in the visible spectral range. The distance range where a monotonous relation is observed is proportional to the wavelength of emission. Therefore, one can extend it by using a fluorophore with emission in the red or near-IR spectral region.

One of the key advantages of this method, which is called metal-induced energy transfer (MIET), is that it does not require any hardware modification to a conventional fluorescence-lifetime imaging microscope (FLIM), thus preserving its full lateral resolution. Fig. 7 schematically shows the key elements of the microscope that are required for MIET imaging. The only additional

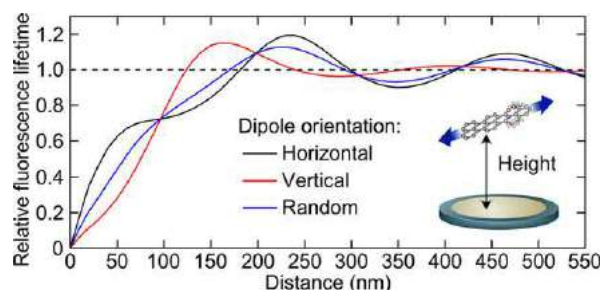


Fig. 6 Calculated dependence of molecule's excited state lifetime on its axial position over the metal film. Curves are calculated for an emission wavelength of 650 nm and a gold film thickness of 20 nm deposited on the glass cover slide. Calculations are done by the methods of Jörg Enderlein and Daja Ruhlandt, see Ref. Karedla *et al.* (2015).

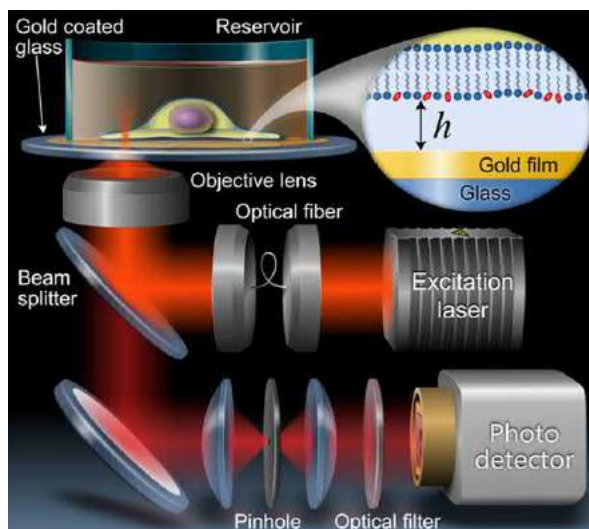


Fig. 7 Scheme of experimental setup for live cell MIET imaging.

requirement for MIET, when compared to conventional FLIM, is the presence of a thin semitransparent gold film (typically 10–20 nm thick) deposited on the glass cover slide supporting the sample.

Using a high NA objective lens (preferably at least $NA=1.4$) one achieves two essential goals that are needed for MIET imaging. Firstly, since the measurements are done within a layer of ca. 150 nm height above the surface, the supercritical fluorescence constitutes a significant part of the total signal and should be collected by a high-NA objective to maximize fluorescence collection efficiency (see Section Super-Critical Angle Fluorescence Microscopy). Secondly, high NA objectives allow for minimizing the axial extension of the excitation focus, allowing for preferentially exciting molecules only within the MIET range, thus enhancing signal-to-noise ratio. As a result, when focused on the sample surface, the focused light excites only molecules within the first ~ 200 nm above the sample. As a result, because the apical cell membrane is typically at least 500 nm away from the substrate, only dye molecules within the basal membrane are efficiently excited and detected.

For evaluating MIET measurements, one has to model the emission properties of a fluorescing molecule over a metal surface. The molecule is assumed to be an electric dipole emitter. Let us consider the emission of a single dipole emitter with orientation angles (α, β) , where β denotes the inclination towards the vertical axis and α the orientation angle around that axis. The electric field amplitude of the dipole's emission into direction (θ, ϕ) is given by the general formula

$$E_{em} = e_p \left[A_{\perp} \cos \beta + A_{\parallel}^c \sin \beta \cos((\phi) - \alpha) \right] + e_s A_{\parallel}^s \sin \beta \sin((\phi) - \alpha) \quad (8)$$

where the $A_{\perp}$, $A_{\parallel}^c$, and $A_{\parallel}^s$ are functions of emission angle θ but not on α , β or ϕ . Explicit expressions for $A_{\perp}$, $A_{\parallel}^c$, and $A_{\parallel}^s$ can be found by expanding the electric field of the dipole emission into a superposition of plane waves, and the tracing each plane wave component through the planar structures using Fresnel's relations (for details see Enderlein *et al.* (1999), Enderlein (1999), Enderlein and Ruckstuhl (2005), Enderlein (2000)). It is important to note that the functions $A_{\perp}$, $A_{\parallel}^c$, and $A_{\parallel}^s$ depend also on the emission wavelength.

Knowing the electric field amplitude of the emission into a given direction $(\theta, (\phi))$, one can then derive the total power of emission as

$$S_{total}(\beta, \alpha) \propto B_{\perp} \cos^2 \beta + B_{\parallel} \sin^2 \beta \quad (9)$$

with the weight factors $B_{\perp}$ and $B_{\parallel}$ that take into account the absorption of emitted energy into the metal layer (Enderlein *et al.*, 1999). Knowing the total emission power S_{total} , one can then calculate the lifetime of the emitter by

$$\tau/\tau_0 = S_0 / [(\phi)S_{total} + 1 - (\phi)] \quad (10)$$

where (ϕ) is the quantum yield, τ_0 the free lifetime, and S_0 the total emission power of the emitter in free space (sample space).

For finally calculating the actual lifetime-distance curve, one has to average the result over all possible molecular orientations and also the emission spectrum of the emitter.

The applicability of MIET for live-cell imaging has been recently shown by mapping the basal membrane of living cells with nanometer accuracy (Chizhik *et al.*, 2014). Knowledge of the precise cell-substrate distance as a function of time and location with unprecedented resolution provides a new means to quantify cellular adhesion and dynamics, as is required for a deeper understanding of fundamental biological processes such as cell differentiation, tumor metastasis and cell migration.

As a biological model system three adherent cell lines were chosen: MDA-MB-231 human mammary gland adenocarcinoma cells and A549 human lung carcinoma cells, which are able to form metastasis in vivo models, as well as MDCK II from canine kidney tissue as a benign epithelial cell line. Interestingly, significant differences in the cell-interface distance between a normal epithelial cell and cancerous cell lines were observed (Chizhik *et al.*, 2014). Recently, Baronsky *et al.* have used MIET imaging for studying the epithelial-to-mesenchymal transition of NMuMG cells (Baronsky *et al.*, 2017).

Fig. 8 shows the measured intensity and lifetime images that were used to obtain the 3D reconstruction of the basal cell membrane. Because the variation of the fluorescence intensity is not only dependent on the metal-induced quenching, but also on the homogeneity of labelling, exclusively the lifetime information was used for reconstructing a three-dimensional map of the

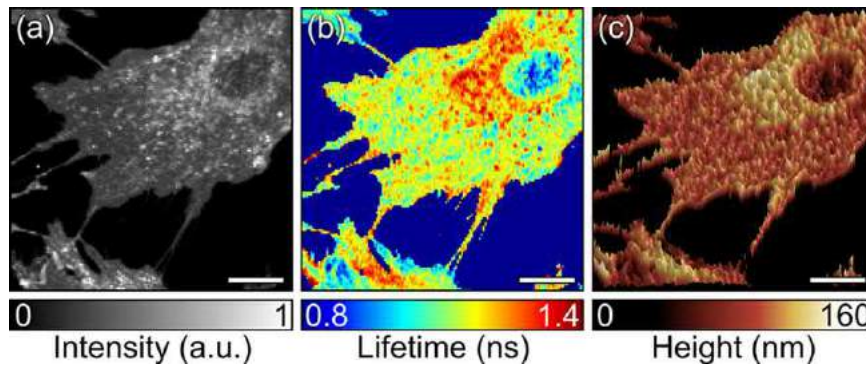


Fig. 8 Simultaneously acquired fluorescence intensity (a) and lifetime (b) images of the basal membrane of living NMuMG cells grown on a gold-coated glass cover slide. (c) Three-dimensional profile of the basal cell membrane computed from fluorescence lifetime image. Scale bars: 20 μm .

basal membrane. On the other hand, the intensity distribution was used to discriminate the membrane fluorescence against the background. Regions with no cells are difficult to identify from the lifetime images alone, as the lifetime values can become exceedingly scattered at low signal-to-noise ratios.

The axial resolution of the recorded images can be determined by calculating the standard deviation of cell-substrate distance (Chizhik *et al.*, 2014). The resolution depends on the photon rate and varies between 2 nm and 4 nm for typical fluorescence intensities measured and can be further enhanced to 1 nm by increasing the number of detected photons.

Karedla *et al.* (2014) have used MIET microscopy for axial localization of single molecules immobilized on the surface. The standard deviation for the observed single-molecule lifetime values is less than 0.1 ns, which corresponds to an axial localization accuracy of better than 2.5 nm for horizontal dipoles.

Although, we have focused on the application of MIET to map the membrane of a living cell and the single molecule axial localization, demonstrating its versatility, relative technical simplicity, and its applicability for life-science, the potential scope of applications of MIET is much greater. The distance range covered by MIET nicely bridges (and complements) the realm of conventional FRET and all the recently developed super-resolution imaging techniques. Although MIET has to rely on a perfect theoretical understanding of the fluorophore-metal interaction, which is certainly more involved than that of FRET, this is no serious obstacle in the current age of powerful desktop computers. In contrast to FRET, the infamous orientation-factor problem is greatly relaxed because one has to know only the relative orientation of one species of fluorescent molecules with respect to the planar metal film. Moreover, in contrast to FRET, one needs to label only one site of a sample with one sort of fluorophore, instead of double-labelling with a donor and acceptor. MIET is exceedingly simple to set up and neither requires modification of the FLIM system nor preparation of complex sample substrates. Coating glass cover slides with a thin metal film is the only prerequisite of the technique. Thus, the technical simplicity of MIET will allow its application in a wide range of studies where nanometer resolution is required.

Acknowledgement

The author is grateful to Jörg Enderlein and Ingo Gregor for helpful comments on this article.

See also: Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

References

- Ambrose, E.J., 1956. *Nature* 178, 1194.
- Aquino, D., Schonle, A., Geisler, C., *et al.*, 2011. *Nature Methods* 8, 353.
- Axelrod, D., 1981. *The Journal of Cell Biology* 89, 141.
- Axelrod, D., 2001. *BIOMEDO* 6, 6.
- Axelrod, D., Burghardt, T.P., Thompson, N.L., 1984. *Annual Review of Biophysics and Bioengineering* 13, 247.
- Axelrod, D., Hellen, E.H., Fulbright, R.M., 2002. In: Lakowicz, J.R. (Ed.), *Topics in Fluorescence Spectroscopy*. Boston, MA: Springer US, p. 289.
- Barnes, W.L., 1998. *Journal of Modern Optics* 45, 661.
- Baronsky, T., Ruhlandt, D., Brückner, B.R., *et al.*, 2017. *Nano Letters* 17, 3320.
- Barroca, T., Balaa, K., Lévêque-Fort, S., Fort, E., 2012. *Physical Review Letters* 108, 218101.
- Boulanger, J., Gueudry, C., Münch, D., *et al.*, 2014. *Proceedings of the National Academy of Sciences* 111, 17164.
- Bourg, N., Mayet, C., Dupuis, G., *et al.*, 2015. *Nature Photonics* 9, 587.
- Burmeister, J.S., Truskey, G.A., Reichert, W.M., 1994. *Journal of Microscopy* 173, 39.
- Cardoso Dos Santos, M., Déturche, R., Vézy, C., Jaffiol, R., 2016. *Biophysical Journal* 111, 136.
- Chizhik, A.I., Chizhik, A.M., Khoptyar, D., *et al.*, 2011. *Nano Letters* 11, 1700.
- Chizhik, A.I., Rother, J., Gregor, I., Janshoff, A., Enderlein, J., 2014. *Nature Photonics* 8, 124.
- Deschamps, J., Mund, M., Ries, J., 2014. *Optics Express* 22, 29081.
- Drexhage, K.H., 1974. In: Wolf, E. (Ed.), *Progress in Optics* 12. Elsevier, p. 163.
- Enderlein, J., 1999. *Chemical Physics* 247, 1.
- Enderlein, J., 2000. *Biophysical Journal* 78, 2151.
- Enderlein, J., Gregor, I., Ruckstuhl, T., 2011. *Optics Express* 19, 8011.
- Enderlein, J., Ruckstuhl, T., 2005. *Optics Express* 13, 8855.
- Enderlein, J., Ruckstuhl, T., Seeger, S., 1999. *Applied Optics* 38, 724.
- Fu, Y., Winter, P.W., Rojas, R., *et al.*, 2016. *Proceedings of the National Academy of Sciences* 113, 4368.
- Gingell, D., Todd, I., Bailey, J., 1985. *The Journal of Cell Biology* 100, 1334.
- Hell, S., Stelzer, E.H.K., 1992. *Optics Communications* 93, 277.
- Hell, S.W., 2003. *Nature Biotechnology* 21, 1347.
- Hell, S.W., 2009. *Nature Methods* 6, 24.
- Hill, D., McDonnell, B., Hearty, S., *et al.*, 2011. *Biomedical Microdevices* 13, 759.
- Huang, B., Bates, M., Zhuang, X., 2009. *Annual Review of Biochemistry* 78, 993.
- Huang, B., Wang, W., Bates, M., Zhuang, X., 2008. *Science* 319, 810.
- Juette, M.F., Gould, T.J., Lessard, M.D., *et al.*, 2008. *Nature Methods* 5, 527.
- Karedla, N., Chizhik, A.I., Gregor, I., *et al.*, 2014. *ChemPhysChem* 15, 705.

- Karedla, N., Ruhlandt, D., Chizhik, A.M., Enderlein, J., Chizhik, A.I., 2015. In: Kapusta, P., Wahl, M., Erdmann, R. (Eds.), *Advanced Photon Counting: Applications, Methods, Instrumentation*. Cham: Springer International Publishing, p. 265.
- Kukura, P., Ewers, H., Müller, C., *et al.*, 2009. *Nature Methods* 6, 923.
- Lanni, F., Waggoner, A.S., Taylor, D.L., 1985. *The Journal of Cell Biology* 100, 1091.
- Lassalle, H.-P., Baumann, H., Strauss, W.S.L., Schneckenburger, H., 2007. *Journal of Environmental Pathology, Toxicology and Oncology* 26, 83.
- Limozin, L., Sengupta, K., 2009. *ChemPhysChem* 10, 2752.
- Loerke, D., Stühmer, W., Oheim, M., 2002. *Journal of Neuroscience Methods* 119, 65.
- Mertz, J., 2000. *Journal of the Optical Society of America B* 17, 1906.
- Olveczky, B.P., Periasamy, N., Verkman, A.S., 1997. *Biophysical Journal* 73, 2836.
- Pavani, S.R.P., Thompson, M.A., Biteen, J.S., *et al.*, 2009. *Proceedings of the National Academy of Sciences* 106, 2995.
- Purcell, E.M., 1946. *Physical Review* 69, 681.
- Ruckstuhl, T., Rankl, M., Seeger, S., 2003. *Biosensors and Bioelectronics* 18, 1193.
- Ruckstuhl, T., Verdes, D., 2004. *Optics Express* 12, 4246.
- Shtengel, G., Galbraith, J.A., Galbraith, C.G., *et al.*, 2009. *Proceedings of the National Academy of Sciences* 106, 3125.
- Stock, K., Sailer, R., Strauss, W.S.L., *et al.*, 2003. *Journal of Microscopy* 211, 19.

Overview: Microscopy

CJR Sheppard, National University of Singapore, Singapore

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The basic construction and imaging properties of the optical microscope, one of the most important optical instruments, are described. But the restrictive meaning of a microscope as a device that the user looks down to see a magnified image is far from the present terminology that includes scanning systems (where the image is stored in a computer), nonvisible radiation, such as electrons or X-rays, and a variety of different contrast mechanisms.

According to the *Oxford English Dictionary* a microscope is 'an instrument magnifying objects by means of lenses so as to reveal details invisible to the naked eye'. Thus, basically a microscope forms a magnified image of an object. The first microscopes used visible light and formed an image showing variations in the intensity of light scattered by the object. This intensity is in general related to the optical properties of the object in a complicated way.

Nowadays there are many different types of microscope. Our understanding of a microscope must therefore be generalized in the following ways.

- The microscope may not be a conventional imaging system using lenses or mirrors. For example, it may be a scanning imaging system. A confocal system is a combination of a conventional and a scanning system. Another important category of microscopes is that of probe microscopes, including near-field optical microscopes. Some probe microscopes, for example the atomic force microscope (AFM) do not even seem at first glance to rely on the direct use of radiation.
- The microscope may not use visible light, but other electromagnetic radiation, or even other forms of radiation.
- The image can be formed using a variety of different contrast mechanisms. Some of these, such as phase contrast, are designed to image particular optical properties of the sample. Others image the generation of a form of radiation when the object is stimulated by another form of radiation. In particular, in principle, virtually any form of spectroscopy can be the basis of building up an image by measuring the spatial variations in the signal.

Different Forms of Microscopes

The Conventional Microscope

In a simple microscope, an objective lens forms a real, magnified, and inverted image of an object. In a compound microscope, an eyepiece is added, forming a virtual image that is viewed by eye to give real image on the retina. The eye can be replaced by CCD camera to record moving or still images. In early microscopes, the objective lens and eyepiece are mounted in the two ends of a brass tube. The objective screws into the tube using an RMS (Royal Microscopical Society) thread (0.8" Whitworth 36 threads per inch), until its mounting face is flush with the end of the tube. The length of the tube is the mechanical tube length. The distance from the mounting face of the objective to the plane of the real image is the optical tube length, which varied according to different manufacturers in the range 160–210 mm. Now most objectives have infinity tube length, so that they collimate the light from the object. The collimated light is brought to a focus by an additional tube lens. An objective of correct tube length should always be used as otherwise spherical aberration is introduced. As the objective, tube lens, and eyepiece are designed together as a system, great care should be exercised when mixing components from different manufacturers.

Abbe showed that in order to form a perfect lateral image the objective must satisfy a different condition called the Herschel condition. Otherwise the image of an off-axis point will suffer from coma. An aberration-free system satisfying the sine condition is called an aplanatic system. A microscope produces a three-dimensional (3D) image of a 3D object. A point of the object closer to the lens appears further from the lens in the image. If the lateral magnification is M , the longitudinal magnification is approximately equal to M^2 . However, it can be shown that it is impossible to devise an instrument that can produce a perfect 3D image of a 3D object, because to have a perfect transverse image the system must satisfy Abbe's sine condition, but to have a perfect axial image the system must satisfy the Herschel condition. These two conditions are not mutually compatible. Actually, we find that the longitudinal magnification is not exactly equal to M^2 , and is not even constant in space. Different depths of the object are imaged differently: the microscope objective only produces a good image for one plane of focus, and other planes will exhibit spherical aberration. However, all of these problems can be overcome by bringing the part of the object under observation into focus, without refocusing the eye or other detector. A 3D image can thus be generated by scanning the object stage in the axial direction, and recording a stack of images from different depths in a computer. For modern objectives of infinite tube length, focusing can be alternatively achieved by piezoelectric scanning of the objective lens.

Resolution

For an object consisting of a single bright point object in a dark background, the image produced by a perfect microscope, according to paraxial diffraction theory, is the so-called Airy disk, consisting of a bright central spot, surrounded by a series of rings. The intensity in a meridional cross-section through the Airy disk is the same for any aperture of objective, or wavelength of light, but its width varies. The intensity of the first bright ring is 1.75% of that at the peak. The radius of the first dark ring of the Airy disk is

$$r_0 = \frac{0.61\lambda}{n \sin \alpha} \quad (1)$$

where λ is the wavelength, n is the refractive index of the immersion medium, and α is the angle subtended at the edge of the objective aperture. The radius depends on $n \sin \alpha$, which is called the numerical aperture and should be as large as possible for good resolution. Thus, high-magnification lenses often use an immersion fluid, usually oil with a refractive index of 1.518 (ISO 8036/1).

The optical coordinate v is defined as

$$v = \frac{2\pi}{\lambda} r n \sin \alpha \quad (2)$$

so that the dark ring of the Airy disk occurs at a value of v of $2\pi \times 0.61 = 3.83$. Resolution is sometimes expressed in terms of the full width at half maximum (FWHM) of the image of a point object, which is 3.232 in optical coordinates. A full nonparaxial theory does not give a very different figure.

Resolution of a microscope is often specified by two-point resolution, which describes whether the image of two points can be distinguished from that of a single point. According to the Rayleigh criterion of two-point resolution, two points are just resolved if the second point is placed on the first dark ring of the first. The separation is then $0.61\lambda/(n \sin \alpha)$. A cross-section through the image in an incoherent optical system of two bright point objects of equal strength and different separations is shown in Fig. 1. The intensity in the image between the points decreases as the separation is increased. The ratio of the intensity midway between the points to that at the points for the Rayleigh criterion to be satisfied is then found to be 0.735.

The concept of resolution should be distinguished from sensitivity or precision. An object much smaller than the resolution limit can still be detected, perhaps weakly, in a microscope: this detection of weak contrast depends on the sensitivity of the microscope rather than its resolution. The size of an object can also be measured with a precision much greater than the resolution of the microscope.

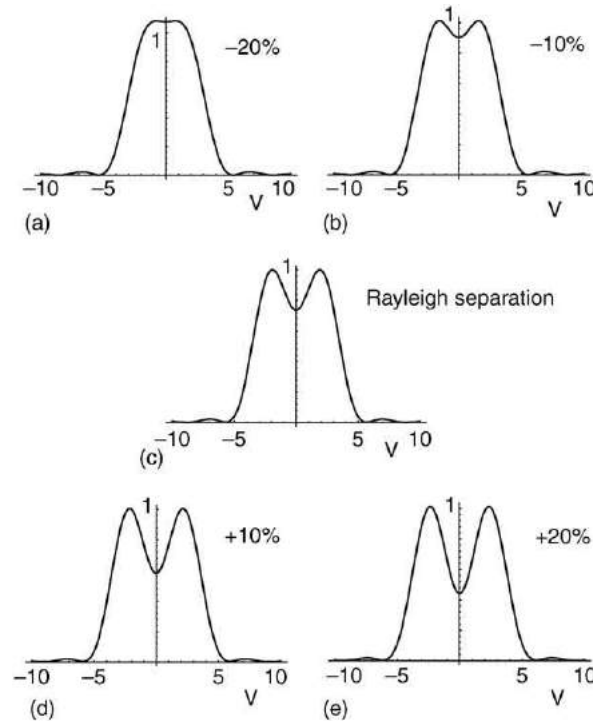


Fig. 1 A cross-section through the image of two equal point objects in an incoherent microscope. The separation of the points is that for the Rayleigh criterion to be satisfied, and for changes of ± 10 , 20% of the Rayleigh separation.

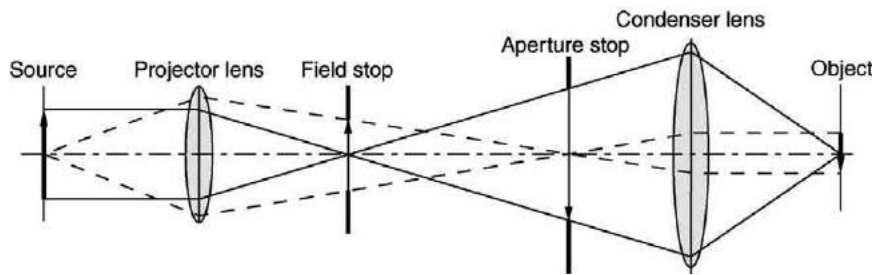


Fig. 2 Geometry of the Köhler illumination system. In practice an arrangement with a smaller number of lenses is usually used.

Illumination

If the object is not self-luminous, in order to be imaged in a microscope it must be illuminated. A semitransparent object, such as a biological slice, is illuminated in transmission. For observation of bulk objects, or surfaces, we use illumination in the reflection geometry, called epi-illumination. Either a tungsten halogen lamp or an arc source is usually used. In critical illumination, the source is focused on to the object by a condenser lens. The disadvantage of this approach is that variations in emission of the source are imaged directly into the image. Cheaper microscopes avoid this problem by using a diffuser. Better microscopes use instead the Köhler illumination system, where the source is placed in the front focal plane of the condenser lens. **Fig. 2** shows a Köhler illumination system in which the source is imaged into the front focal plane of the condenser lens by another, projector, lens. The system incorporates a separate field stop and aperture stop. In practice, the size of the field stop should be reduced to illuminate as small a region of the object as necessary in order to minimize stray light. An important part of setting up a microscope to operate properly consists in centering the aperture stop of the condenser. This is often done using a so-called Bertrand lens that allows imaging of the aperture stop.

Image formation

The Rayleigh criterion was originally specified for an incoherent optical system. It is thus applicable in fluorescence microscopy, as fluorescence is an incoherent process as two fluorescent point objects fluoresce independently. But for the image of a trans-illuminated object, the resolution depends on the coherence of the illumination. This is controlled by the numerical aperture of the condenser lens as compared with that of the objective. The ratio of these numerical apertures is called the coherence parameter, S , so that

$$S = \frac{n_c \sin \alpha_c}{n \sin \alpha} \quad (3)$$

For $S=0$ the illumination is purely coherent, while for $S \rightarrow \infty$ imaging becomes purely incoherent, thus corresponding to the Rayleigh criterion, so S should really be called an incoherence parameter. An important in-between case is when $S=1$, when the apertures of objective and condenser are equal. This is termed matched, full, or complete illumination. According to the generalized Rayleigh criterion, the points are just resolved when the ratio of the intensity midway between the points to that at the points is 0.735, which is called the generalized Rayleigh criterion. We require that the distance between the points is as small as possible when this occurs. The generalized Rayleigh separation is the same for incoherent illumination and for $S=1$, and corresponds to a distance of $2\nu_0=3.83$ in optical coordinates. Note, however, that for other separations of the points, the ratio is different for these two cases: it is not true, but often erroneously stated, that $S=1$ corresponds to incoherent imaging. Resolution improves as the aperture of the condenser is increased ($2\nu_0=5.15$ in optical coordinates for coherent illumination) reaching a maximum when $2\nu_0=1.46$, when resolution is 9% better than for incoherent imaging ($2\nu_0=3.58$). This maximum in resolution is achieved when the numerical aperture of the condenser is larger than that of the objective, which is not physically achievable if the objective has the highest possible numerical aperture. Similarly, it is impossible to achieve incoherent illumination with very large objective numerical apertures.

The effect of coherence on the resolution of an image depends on the form of the object, and on the particular resolution criterion employed. In practice, contrast also depends on the aperture of the condenser. Usually, opening the condenser improves resolution, but reduces contrast, so there will be an optimum condenser aperture size.

Abbe theory of microscope imaging

Abbe argued that for coherent illumination of a grating, the grating spectra can be observed in the back focal plane of the objective. The tube lens then forms an image from the grating components. Thus, this is an early description of Fourier optics. The strength of the grating components can be altered in the back focal plane to give various optical effects, such as phase contrast. Abbe's theory was not properly appreciated at the time because it was known that in practice, resolution could be improved by using a larger condenser aperture, which does not give coherent illumination. There was much controversy about the merits of Köhler versus critical illumination, and different sizes of condenser aperture.

We now know that Köhler and critical illumination are in principle equivalent. The aberrations of the condenser lens are not important, so that the source and condenser together behave simply as a partially coherent effective source. Imaging can still be described by Fourier optics within the framework of partially coherent imaging theory.

Although Köhler and critical illumination are equivalent in principle, Köhler illumination is in practice better because of its improved illumination uniformity.

Depth of focus

The longitudinal image of a point object is found to be approximately invariant when expressed in terms of an optical coordinate u , defined as

$$u = \frac{8\pi}{\lambda} z n \sin^2 \frac{\alpha}{2} \quad (4)$$

The axial resolution, and thus also the depth of focus, varies strongly with aperture. The FWHM for longitudinal imaging is 11.13 in optical coordinates.

Microscope objectives

Because microscope objectives often have large apertures, aberrations will be very strong, unless the objective is designed correctly. The basic design principle for a high aperture lens is the aplanatic front element (Fig. 3). It is found that all rays from a point A in a sphere of refractive index n , where A is a distance r/n from its center, appear to come from a point B, distance m , without any aberration. This principle is used by employing a front element that has one surface of the same radius of curvature as the sphere. The front surface is a sphere centered on A (Fig. 3(b)), so that the rays are not deviated on crossing it.

A low-power achromat uses just two components with different optical dispersion to cancel chromatic aberration and correct for other aberrations (an achromatic doublet). For higher powers, two separated doublets are necessary. For the highest apertures, an aplanatic front is used. However, this can never converge the light, but only makes it less divergent. It is thus used together with an achromatic doublet. For oil immersion, the interface between the oil and the front element is not very important, so it is made planar to simplify manufacture. In practice, it is usually made planar even for a dry lens, and the resulting aberrations cancelled out elsewhere in the objective. Apochromats, corrected for three colors, use more elements, which may include two or more stages of aplanatic front. The final objective thus consists of a number of elements that must be accurately aligned relative to one another. The objective is adjusted by the manufacturer to give a good star image (image of a point object). First a spacer is selected to

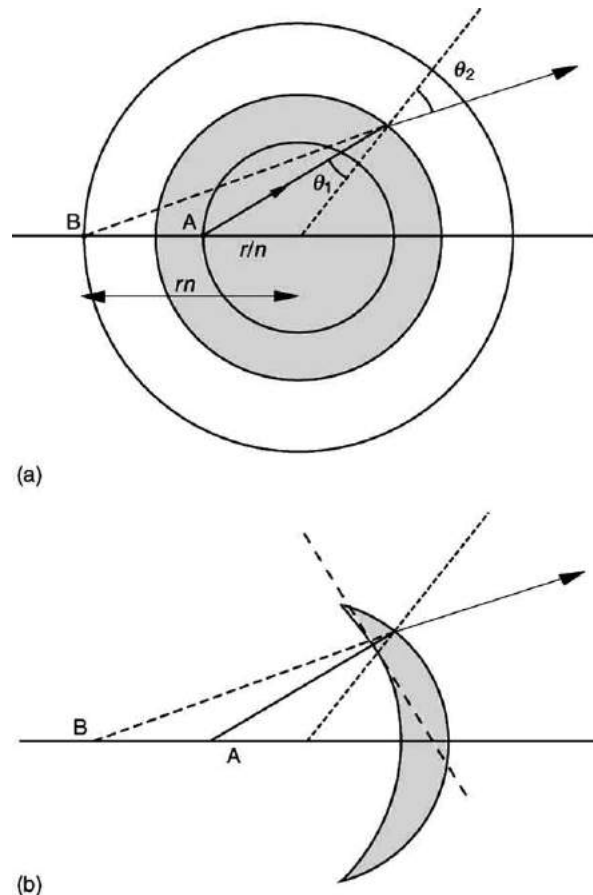


Fig. 3 The principle of the aplanatic front: (a) an aplanatic surface; (b) a meniscus lens that acts as an aplanatic system.

optimize spherical aberration. Then a sleeve is adjusted to make the objective parfocal with others. Finally a screw is used to center the assembly to remove coma.

Conventional and Scanning Microscopes

Conventional microscope

Various different types of transmission microscope are illustrated in Fig. 4. We consider image formation in a conventional microscope as illustrated schematically in Fig. 4(a), showing a microscope with critical illumination. A large area incoherent source is focused by the condenser lens on to the specimen, illuminating a comparatively large area of the specimen, corresponding to the whole field of view of the objective. Information from each illuminated point in the specimen is simultaneously transmitted by the objective lens to form the primary image. The objective is responsible for forming the image, with the condenser playing only a secondary role in determining the resolution of the system, through control of the coherence of the illumination. Fig. 4(b) shows a conventional microscope in which the image is measured point by point by a detector. In practice this could be achieved by using a CCD detector.

Scanning microscope

An image can be generated by a scanning system, as illustrated in Fig. 4(c). A probe of light is formed by demagnification of a source, and is scanned over the object in a raster. The transmitted (or reflected) light is detected by a photodetector and thus builds up an image. The size of the probe of light limits the resolution of the system, that is, the smallest detail in the object that can be seen in the image. It should be noted that the magnification of the image is given simply by the ratio of the distance scanned in the image to the distance scanned by the probe. It is thus unrelated to the demagnification of the source. The image magnification can be altered without changing the lens and is not in any way related to the resolution.

It has been shown that the imaging properties of scanning and conventional microscopes are identical under analogous conditions, this being known as the principle of equivalence. This property is based on the principle of reciprocity, which is a very general physical law that holds even with diffraction, absorption, multiple scattering, and stray light. The principle of equivalence is generally valid except if there are nonreciprocal magnetic or polarization effects, or energy losses involved (such as in fluorescence microscopy, and with inelastic scattering in electron microscopy).

Scanning systems exhibit a number of important advantages over conventional systems. Broadly, these are based on two classes of property. First, in a scanning system the image is in the form of an electronic signal. As a result, this is advantageous for quantitative measurements, as well as for image processing, including image enhancement, image restoration, and image analysis. Second, in a scanning system the object is illuminated by a focused spot, which extends the range of imaging modes available.

Confocal microscope

Finally in Fig. 4(d), we combine the arrangements of Figs. 4(b) and (c) to give a confocal scanning optical microscope, in which a point source illuminates just a small region of the object, and a confocal point detector detects light from this illuminated region.

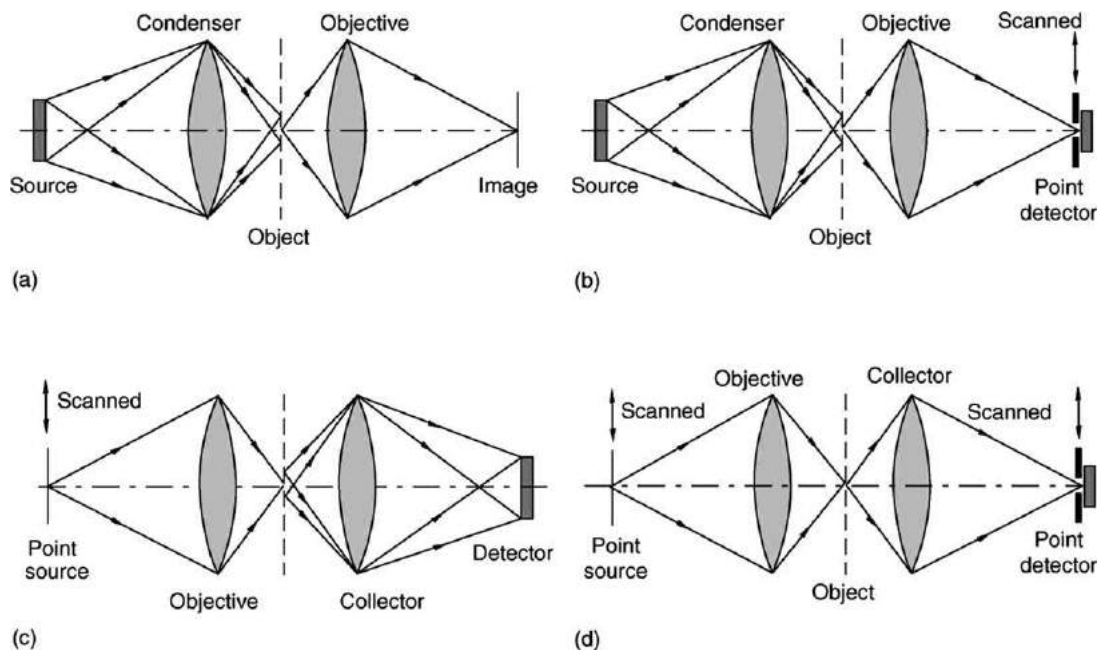


Fig. 4 Different geometries of transmission microscope: (a) a conventional microscope; (b) a conventional microscope with a point detector; (c) a scanning microscope; (d) a confocal microscope.

If the point source and detector are scanned in unison, a two-dimensional image is generated. However, this system now behaves very differently from the previous ones. The confocal microscope is thus not a special case of the general partially coherent conventional imaging system. From the symmetry of Fig. 4(d), it is clear that the two lenses play an equal part in the imaging process. This results in an improvement in resolution. In fact, the confocal system behaves as a coherent imaging system, but with a sharper effective point spread function than in a conventional coherent microscope.

Although Fig. 4 is drawn for the transmission geometry, in practice most confocal systems operate in the reflection or epi-illumination mode, in which the same objective lens is used both for illumination and detection. For a specimen placed in the focal plane, the properties of confocal transmission and reflection systems are identical. However, once the object is moved from the focal plane, certain differences arise: in particular in confocal reflection a strong optical sectioning effect occurs that allows a single section through a thick object to be imaged. This is the major advantage of the confocal microscope arrangement. Fig. 4 also applies equally well to fluorescence imaging. However, in this case, because of the incoherent nature of fluorescence emissions after excitation by coherent light, imaging is then incoherent.

Confocal microscopes can be achieved by using either a point detector, in practice performed by placing a pinhole in front of a photodetector, or by using a coherent detector. Such a coherent detector, sensitive to the amplitude of the incident radiation, is not directly available for light, but is available for acoustic radiation, as in a scanning acoustic microscope. For light, a confocal effect can also be achieved by using an interferometric method to synthesize a coherent detector. Thus, in interference microscopes using the high aperture condenser and objective lenses, an optical sectioning effect arises similar to that in the confocal microscope. Another way of producing a coherent detector is to use a single-mode optical fiber.

Probe Microscopes

Scanning tunneling microscope

The scanning tunneling microscope was the first of the family of probe microscopes, relying on a fundamentally different principle from the usual forms of microscope. A physical tip in the nanometer range is brought close, also in the nanometer range, to a conducting sample. If an electric potential is applied, electrons can tunnel across between the sample and the tip. By scanning the tip mechanically across the sample, an image can be generated with a resolution in the subnanometer range. The surface topography can also be measured with subnanometer sensitivity. While interference microscopy can measure surface topography with subnanometer sensitivity, in this case the profile is averaged over the lateral resolution of the microscope, of the order of the wavelength in dimensions. The scanning tunneling microscope can image individual atoms. The signal is related to the work function of the material. Altering the bias voltage allows the band structure of the material to be investigated.

Atomic force microscope

In the atomic force microscope, the force between the surface and the tip is measured using a cantilever beam and used to build up an image. Atomic force microscopy can be performed in either a contact or noncontact mode. It can be used with insulating specimens. Optical methods, often using a position-sensitive detector, are used to measure displacements of the cantilever.

Near-field scanning optical microscope (NSOM)

In a near-field scanning optical microscope, a very small tip or aperture is scanned relative to the specimen to attain resolution greater than the classical limit set by the wavelength of the radiation. As Fig. 5 shows, various different designs of near-field

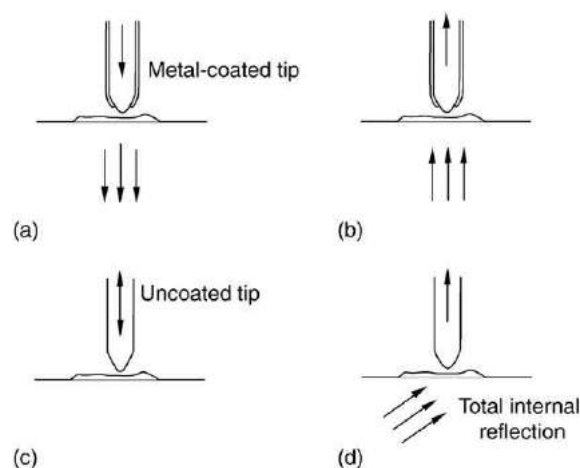


Fig. 5 Different forms of near-field microscope: (a) illumination-mode NSOM; (b) collection-mode NSOM; (c) a near-field probe used for both illumination and collection in a confocal arrangement; (d) photon tunneling microscope (PTM). Reproduced with permission from Sheppard CJR (1978) The scanning optical microscope. *The Physics Teacher* 16: 648–651. © 1978 American Association of Physics Teachers.

microscope have been proposed. It is evident that these again form into the conventional/scanning categories described earlier. The sample can be illuminated using the near-field probe, as in Fig. 5(a). This is called the illumination-mode NSOM. Or the sample can be uniformly illuminated, and a signal detected using a near-field probe as in Fig. 5(b), giving a collection-mode NSOM. Or a confocal arrangement can be used, with a near-field probe used for both illumination and collection (c). In this case, an uncoated tip can be used; as otherwise the detected signal will be very small. Finally, Fig. 5(d) shows the photon tunneling microscope (PTM), in which the sample is illuminated with evanescent waves produced by total internal reflection and an uncoated tip used to probe the evanescent field in the presence of the sample.

Different Types of Radiation

Electromagnetic Radiation

UV radiation and X-rays

The first microscopes used visible light, but, in fact, we do not need to use visible light, but can use any electromagnetic radiation, over a broad range of different available wavelengths. In Fig. 6 we show the wavelength of the electromagnetic spectrum, illustrating how the wavelength in air varies with frequency. Instead of using visible light, shorter wavelengths allow greater resolution to be achieved. Ultra-violet light is commonly used to excite fluorescence. X-rays have also successfully been used to image biological samples.

IR and microwave radiation

In the longer wavelength region, infra-red radiation is used for observation of semiconductors or for molecular spectroscopic imaging. Imaging of emitted midinfra-red radiation can show up variations in temperature, such as hot spots in semiconductors. Microwaves have also been used, but resolution is limited by the long wavelength. For this reason microwaves (and mid-IR) have also been used for near-field microscopy.

Other Radiation

Electron microscope

The other important class of radiation that can be used for illumination is matter waves. The most common of these, of course, is electrons, the wavelength of which, as a function of acceleration voltage, is shown in Fig. 7. The bend in the curve at very high

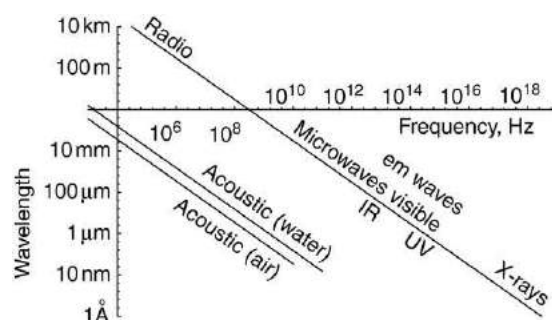


Fig. 6 The wavelength of different forms of electromagnetic radiation and acoustic radiation. Reproduced with permission from Sheppard CJR (2002) *The generalized microscope*. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

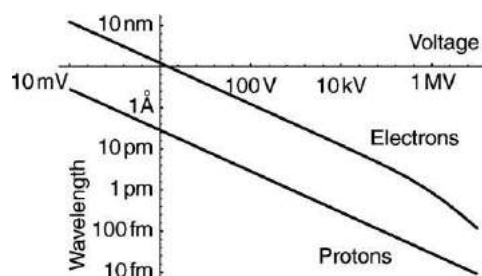


Fig. 7 The wavelength of matter waves (electrons and protons) as a function of accelerating voltage. Reproduced with permission from Sheppard CJR (2002) *The generalized microscope*. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

voltages is caused by relativistic effects. The wavelength of the electrons is very small: even for 100 V electrons, the wavelength is only 0.1 nm. But aberrations caused by the electron lenses limit resolution so that at present atomic resolution is only just achievable, even with much higher voltages.

In a transmission electron microscope (TEM), a beam of electrons is produced by an electron gun. The specimen is illuminated via a condenser lens (or lenses). Usually these are magnetic lenses in which a current in a coil produces a magnetic field that focuses the electrons. The electrons that are transmitted through the specimen are then focused by a series of lenses to form a final image that can be viewed directly on a phosphor screen. An X-ray detector can be used to characterize the elements in the specimen. An electron-energy-loss spectrometer, together with a detector, can measure the energy of the electrons in a particular region of the image.

Other matter radiation

Instead of electrons, other particle beams can be used for illumination. [Fig. 7](#) shows the wavelength of a beam of protons, as function of the accelerating voltage. Because of the larger mass of protons, relativistic effects are not apparent for protons in the curve in [Fig. 7](#). Microscope images have also been formed using ions, and even neutral particles such as neutrons. Neutral atom microscopes are also under development.

Acoustic radiation

Instead of electromagnetic waves, acoustic radiation can be used to form a scanning acoustic microscope (SAM). Ultrasonic waves are focused into a water immersion medium. Mechanical scanning is again used to build up an image. [Fig. 6](#) shows the wavelength of the acoustic wave as a function of frequency. By using microwave frequencies, wavelengths, and therefore resolution, in the submicron range, can be generated. The SAM acts as a confocal microscope because the detector is sensitive to the amplitude of the acoustic wave. The acoustic microscope provides information on variations in the elastic properties (viscosity and elasticity) of the sample.

Different Contrast Mechanisms

The bright field microscope detects the intensity of the light transmitted through the object. A microscope with a condenser aperture of appreciable size behaves as a partially coherent system, and the image will, in general, depend on the phase of the light coming from the object as well as its intensity. Many objects are weakly scattering, so the contrast of this phase information is weak. In this case, the microscope effectively images variations in the transmittance or reflectance of the sample. Different designs of microscope can be used to image various properties of the object, thus providing new and complementary information. Phase contrast microscopy provides but one example of a different contrast mechanism. In fact, in general, any physical or chemical interaction of the illuminating radiation with the sample can be used as the basis for a contrast mechanism. These different contrast mechanisms can be performed in either a conventional or scanning arrangement, or in a confocal system.

Phase Contrast Microscopy

Phase contrast is a widely used technique in both biological microscopy (in transmission) and materials microscopy (in reflection). Its importance in biological microscopy stems, firstly, from the fact that contrast may otherwise be too weak to be visible, but secondly because the phase variations are caused by changes in optical thickness, related to physical density. In reflection microscopy, phase is related to surface height variations, but also depends on the optical material properties of the sample.

Polarization Microscopy

The polarization state of light transmitted through or reflected from a sample can also be changed. This change is detected in polarization microscopy, for example by placing the sample between crossed polarizers. The birefringence detected can be either material birefringence, or form birefringence (caused by the shape of the microstructure). Polarization microscopy is widely used in biological microscopy, and in mineralogy.

Fluorescence Microscopy

In fluorescence microscopy, electrons excited by illumination decay to the ground state and emit photons of light. Auto-fluorescence is natural fluorescence of the sample. Fluorescent dyes can be used as labels for particular biological or chemical constituents. Fluorescence microscopy is usually performed in the epi-illumination geometry to assist rejection of light from the source. Often it is performed in a confocal microscope, allowing 3D localization of the fluorescent labels.

Raman Microscopy

Raman spectroscopy can be performed on a microscopic scale to produce images of presence of particular molecular bonds. It is usually performed in a scanning system, often of the confocal type.

This is one example of the wider class of spectroscopic microscopies: in principle, any form of spectroscopy can be used as the basis for a microscope contrast mode.

Nonlinear Microscopy

Another class of contrast mechanisms is based on nonlinear optical interactions. Again, any nonlinear mechanism can be used as the basis for a contrast mechanism. These include two- (or multiphoton) fluorescence, second- (or higher-order) harmonic generation, sum or difference frequency generation, and coherent anti-Stokes Raman scattering (CARS). Probably other nonlinear optical mechanisms will be exploited in microscopy in the future.

Other Contrast Mechanisms

The contrast mechanisms described above, all rely on illumination and detection of light. But in many other modes, radiation of one form can be converted into another. These include cathodoluminescence microscopy (electrons to light), X-ray microanalysis (electrons to X-rays), photo-emission microscopy (light, usually UV, to electrons), photo-acoustic microscopy (light to ultrasonics), photothermal microscopy (light to thermal waves), optical or electron beam induced current (OBIC/EBIC). Again, in many of these, the illuminating or emitted radiation can be focused, or a combination can be used in a confocal arrangement.

Further Reading

- Born, M., Wolf, E., 1999. *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.
- Bradbury, S., Bracegirdle, B., 1997. *Introduction to the Optical Microscope*. Oxford: Bios Scientific Publishers.
- Inoue, S., Spring, K.R., 1997. *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.
- Martin, L.C., 1966. *Theory of the Microscope*. London: Blackie.
- Paesler, M.A., Moyer, P., 1996. *Near-Field Optics: Theory, Instrumentation, and Applications*. New York: Wiley InterScience.
- Pluta, M., 1988. *Advanced Light Microscopy, Volume 1: Principles and Basic Properties*. Amsterdam: Elsevier.
- Sheppard, C.J.R., 1987. Scanning optical microscopy. In: Barer, R., Cosslett, V.E. (Eds.), *Advances in Optical and Electron Microscopy*, vol. 10. London: Academic Press, pp. 1–98.
- Sheppard, C.J.R., 2002. The generalized microscope. In: Diaspro, A. (Ed.), *Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances*. New York: Wiley-Liss, pp. 1–18.
- Slayter, E.M., Slayter, H.S., 1992. *Light and Electron Microscopy*. Cambridge: Cambridge University Press.
- Wiesendanger, R., 1994. *Scanning Probe Microscopy and Spectroscopy*. Cambridge: Cambridge University Press.
- Wilson, T., Sheppard, C.J.R., 1984. *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Confocal Microscopy

T Wilson, University of Oxford, Oxford, UK

© 2005 Elsevier Ltd. All rights reserved.

Introduction

It is probably fair to say that the development and wide commercial availability of the confocal microscope has been one of the most significant advances in light microscopy in the recent past. The main reason for the popularity of these instruments derives from their ability to permit the structure of thick specimens of biological tissue to be investigated in three dimensions. It achieves this important goal by resorting to a scanning approach together with a novel (confocal) optical system.

The traditional wide-field conventional microscope is a parallel processing system which images the entire object field simultaneously, which is a severe requirement for the optical components. We can relax this requirement if we no longer try to image the whole object at once. The limit of this relaxation is to require a good image of only one object point at a time. The price that we have to pay is that we must scan in order to build up an image of the entire field.

A typical arrangement of a scanning confocal optical microscope is shown in Fig. 1 where the system is built around a host conventional microscope. The essential components are some form of mechanism for scanning the light beam (usually from a laser) relative to the specimen and appropriate photodetectors to collect the reflected or transmitted light. Most of the early systems were analog in nature, but it is now universal, due to the serial nature of the image formation, to use a computer both to drive the microscope and to collect, process, and display the image.

In the beam scanning confocal configuration of Fig. 1, the scanning is typically achieved by using vibrating galvanometer-type mirrors or acousto-optic beam deflectors. The use of the latter gives the possibility of TV-rate scanning, whereas vibrating mirrors are often relatively slow when imaging an extended region of the specimen, although significantly higher scanning speeds are achievable over smaller scan regions. We should note that other approaches to scanning may be implemented, such as specimen-scanning and lens-scanning. These methods, although not generally available commercially, do have advantages in certain specialized applications. Since this article is necessarily limited in length much additional material may be found in other sources listed in the Further Reading at the end of this article.

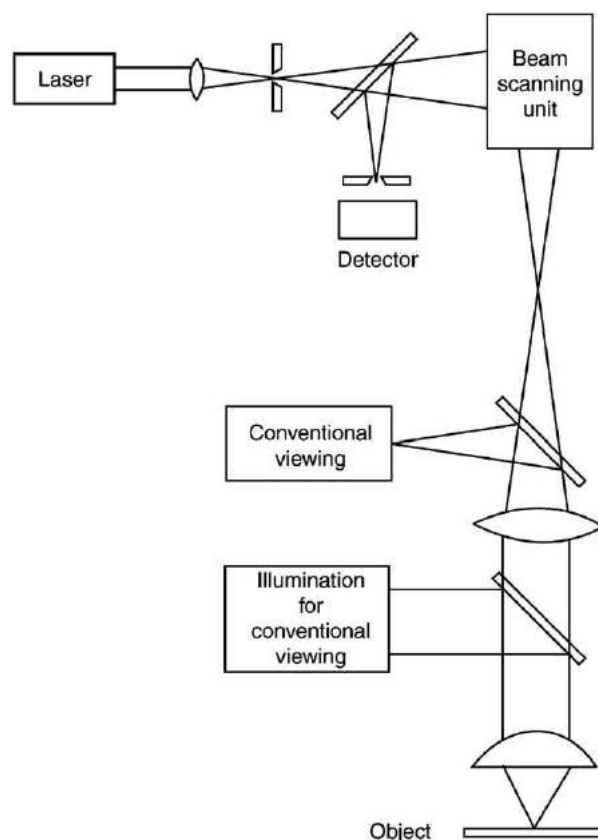


Fig. 1 Schematic diagram of a confocal microscope.

Image Formation in Scanning Microscopes

We will not discuss the fine detail of the optical properties of confocal systems. However, the essence is shown in Fig. 1 where we see that the confocal optical system consists simply of a point source of light which is then used to probe a single point on the specimen. The strength of the reflected or fluorescence radiation from the single object point is then measured via a point, pinhole detector. The confocal – point source and point detector – optical system therefore merely produces an ‘image’ of a single object point and hence some form of scanning is necessary to produce an image of an extended region of the specimen. However, the use of single-point illumination and single-point detection results in novel imaging capabilities which offer significant advantages over those possessed by conventional wide-field optical microscopes. In essence, these are enhanced lateral resolution and, perhaps more importantly, result in a unique depth discrimination or optical sectioning property. It is this latter property which leads to the ability to obtain three-dimensional images of volume specimens.

The improvement in lateral resolution may at first seem implausible. However, it can be explained simply by the principle which states that resolution can be increased at the expense of field of view, which can then be increased by scanning. One way of taking advantage of this principle is to place a very small aperture extremely close to the object. The resolution is now determined by the size of the aperture rather than the radiation. In the confocal microscope we do not use a physical aperture in the focal plane but rather use the back-projected image of a point detector in conjunction with the focused point source. Fig. 2 indicates the improvement in lateral resolution that may be achieved.

The confocal principle was introduced in an attempt to obtain an image of a slice within a thick specimen which was free from the distracting presence of out-of-focus information from surrounding planes. The confocal optical systems fulfills this requirement and its inherent optical sectioning or depth discrimination property has become the major motivation for using confocal microscopes, and is the basis of many of the novel imaging modes of these instruments. The origin of the depth discrimination property may be understood easily from Fig. 3, where we show a reflection-mode confocal microscope and consider the imaging of a specimen with an undulating surface. The full lines show the optical path when an object feature lies in the focal plane of the lens. At a later scan position, the object surface is supposed to be located in the plane of the vertical dashed line. In this case, simple ray tracing shows that the light reflected back to the detector pinhole arrives as a defocused blur, only the central portion of which is detected and contributes to the image. In this way the system discriminates against features which do not lie within the focal region of the lens. A very simple method of both demonstrating the effect and giving a measure of its strength is to scan a perfect reflector axially through focus and measure the detected signal strength. Fig. 4 shows a typical response. These responses are frequently termed the $V(z)$, by analogy with a similar technique in scanning acoustic microscopy. A simple paraxial theory

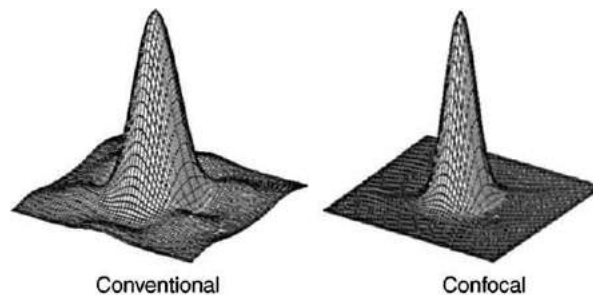


Fig. 2 The point spread functions of the conventional and confocal microscopes showing the improvement in lateral resolution which may be obtained in the confocal case.

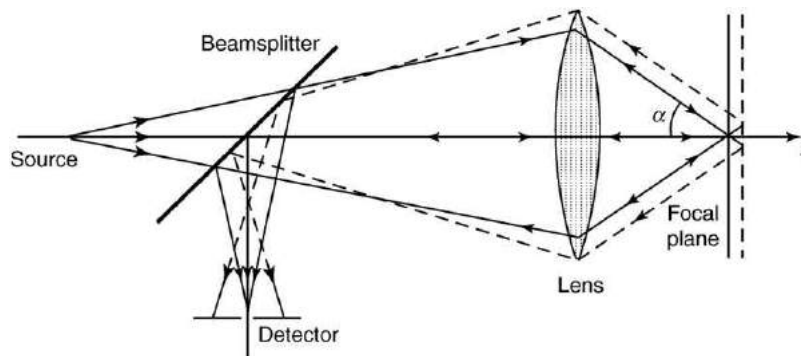


Fig. 3 The origin of the optical sectioning or depth discrimination property of the confocal optical system.

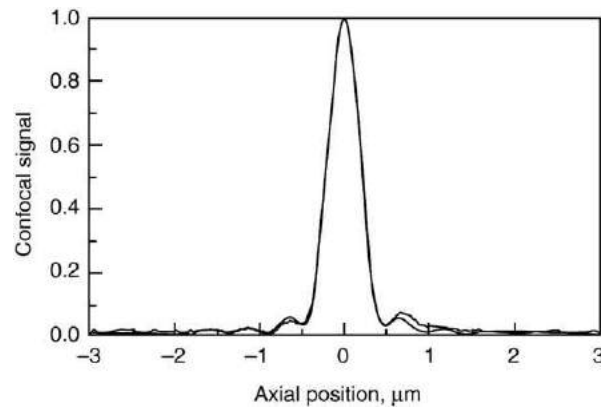


Fig. 4 The variation in detected signal as a plane reflector is scanned axially through focus. The measurement was taken with a 1.3 numerical aperture objective and 633 nm radiation.

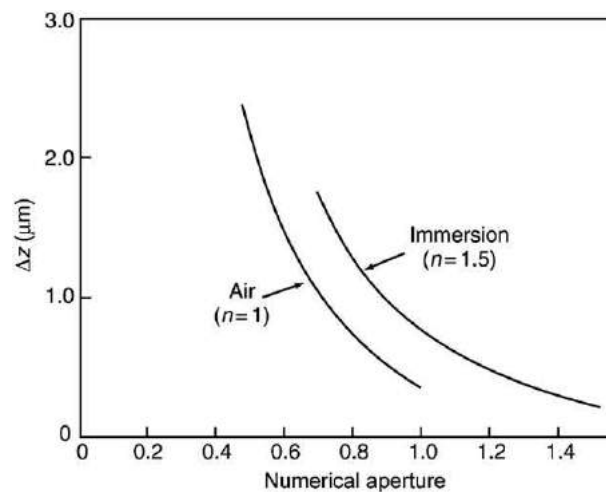


Fig. 5 The optical sectioning width as a function of numerical aperture. The curves are for red light (0.6328 μm wavelength). Δz is the full width at the half-intensity points of the curves of $I(u)$ against u .

models this response as

$$I(u) = \left[\frac{\sin(u/2)}{u/2} \right]^2 \quad (1)$$

where u is a normalized axial coordinate which is related to real axial distance, z , via

$$u = \frac{8\pi}{\lambda} n z \sin^2(\alpha/2) \quad (2)$$

where λ is the wavelength and $n \sin \alpha$ the numerical aperture. As a measure of the strength of the sectioning, we can choose the full width at half intensity of the $I(u)$ curves. **Fig. 5** shows this value as a function of numerical aperture for the specific case of imaging with red light from a helium neon laser. These curves were obtained using a high aperture theory which is more reliable than **Eq. (1)** at the highest values of numerical aperture. We note, of course, that these numerical values refer to nonfluorescence imaging. The qualitative explanation of optical sectioning, of course, carries over to the fluorescence case but the actual value of the optical sectioning strength is different in the fluorescence case.

Applications of Depth Discrimination

Since this property is one of the major reasons for the popularity of confocal microscopes, it is worthwhile, at this point, to review briefly some of the novel imaging techniques which have become available with confocal microscopy.

Fig. 6 illustrates the essential effect: **Fig. 6(a)** shows a conventional image of a planar microcircuit which has deliberately been mounted with its normal at an angle to the optic axis. We see that only one portion of the circuit, running diagonally, is in focus.

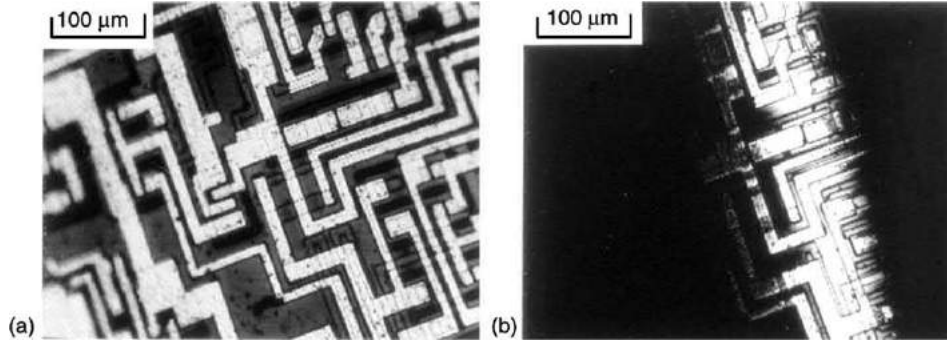


Fig. 6 (a) Conventional scanning microscope image of a tilted microcircuit: the parts of the object outside the focal plane appear blurred. (b) Confocal image of the same microcircuit: only the part of the specimen within the focal region is imaged strongly.

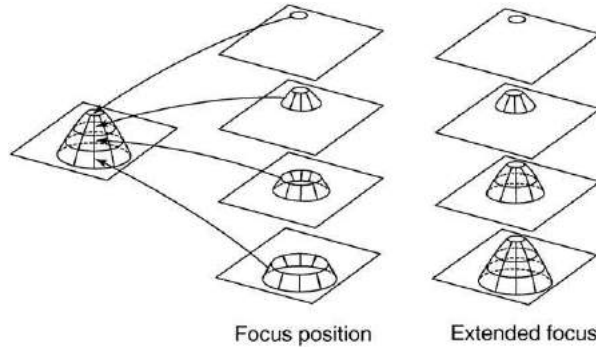


Fig. 7 An idealization of the optical sectioning property showing the ability to obtain a through-focus series of images, which may then be used to reconstruct the original volume object at high resolution.

Fig. 6(b) shows the corresponding confocal image: here the discrimination against detail outside the focal plane is clear. The areas which were out of focus in **Fig. 6(a)** have been rejected. Furthermore, the confocal image appears to be in focus throughout the visible band, which illustrates that the sectioning property is stronger than the depth of focus.

This suggests that if we try to image a thick translucent specimen, we can arrange, by the choice of our focal position, to image detail exclusively from one specific region. In essence, we can section the specimen optically without having to resort to mechanical means. **Fig. 7** shows an idealized schematic of the process. The portion of the beehive-shaped object that we see is determined by the focus position. In this way it is possible to take a through-focus series and obtain data about the three-dimensional structure of the specimen. If we represent the volume image by $I(x, y, z)$, then by focusing at a position $z=z_1$, we obtain, ideally, the image $I(x, y, z_1)$. This, of course, is not strictly true in practice because the optical section is not infinitely thin.

It is clear that the confocal microscope allows us to form high-resolution images with a depth of focus sufficiently small that all the detail which is imaged appears in focus. This suggests immediately that we can extend the depth of focus of the microscope by adding together (integrating) the images taken at different focal settings without sacrificing the lateral resolution. Mathematically, this extended-focus image is given by

$$I_{\text{EF}}(x, y) = \int I(x, y, z) dz \quad (3)$$

As an alternative to the extended-focus method, we can form an auto-focus image by scanning the object axially and, instead of integrating, selecting the focus at each picture point by recording the maximum in the detected signal. Mathematically, this might be written as

$$I_{\text{AF}}(x, y) = I(x, y, z_{\text{max}}) \quad (4)$$

where z_{max} corresponds to the focus setting giving the maximum signal. The images obtained are somewhat similar to the extended focus and, again, substantial increases in depth of focus may be obtained. We can go one step further and turn the microscope into a noncontacting surface profilometer. Here we simply display z_{max} .

It is clear by now that the confocal method gives us a convenient tool for studying three-dimensional structures in general. We essentially record the image as a series of slices and play it back in any desired fashion. Naturally, in practice it is not as simple as this, but we can, for example, display the data as an $x-z$ image rather than an $x-y$ image. This is somewhat similar to viewing the specimen from the side. As another example, we might choose to recombine the data as stereo pairs by introducing a slight lateral offset to each image slice as we add them up. If we do this twice, with an offset to the left in one case and the right in another, we

obtain, very simply, stereo pairs. Mathematically, we form images of the form:

$$\int I(x \pm \gamma z, y, z) dz \quad (5)$$

where γ is a constant. In practice it may not be necessary to introduce offsets in both directions to obtain an adequate stereo view.

Our discussion so far on these techniques has been by way of a simplified introduction. In particular, we have not presented any fluorescence images. The key point is that, in both brightfield and fluorescence modes, the confocal principle permits the imaging of specimens in three dimensions. The situation is, of course, more involved than we have implied. A thorough knowledge of the image formation process, together with the effects of lens aberrations and absorption, is necessary before accurate data manipulation can take place.

In conclusion to this section it is important to emphasize that the confocal microscope does not produce three-dimensional images. It essentially produces very high-quality two-dimensional images of a (thin) slice within a thick specimen. A three-dimensional rendering of the entire volume specimen may then be generated by suitably combining a number of these two-dimensional image slices from a through focus series of images.

Fluorescence Microscopy

We now turn our attention to confocal fluorescence microscopy, because this is the imaging mode which is usually employed in biological applications. Although we have introduced the confocal microscope in terms of bright-field imaging the comments we have made concerning the origin of the optical sectioning, etc., carry over directly to the fluorescence case although the numerical values describing the strength of the optical sectioning are, of course, different and we will return to this point later.

If we assume that the fluorescence in the object destroys the coherence of the illuminating radiation and produces an incoherent fluorescent field proportional to the intensity of the incident radiation, $I(v, u)$, then we may write the effective intensity point spread function, which describes image formation in the incoherent confocal fluorescence microscope, as

$$I(v, u) I\left(\frac{v}{\beta}, \frac{u}{\beta}\right) \quad (6)$$

where the optical coordinates u and v are defined relative to the primary radiation and $\beta = \lambda_2/\lambda_1$ is the ratio of the fluorescence to the primary wavelength. We note that $v = (2\pi/\lambda)r \sin \alpha$, where r denotes the actual radial distance.

This suggests that the imaging performance depends on the value of β . In order to illustrate this we show in Fig. 8 the variation in detected signal strength as perfect fluorescent sheet scans through focus. This serves to characterize the strength of the optical sectioning in fluorescence microscopy in the same way that the mirror was used in the brightfield case. We note that the half-width of these curves is essentially proportional to β and so for optimum sectioning the wavelength ratio should be as close to unity as possible.

We have just discussed what we might call one-photon fluorescence microscopy in the sense that a fluorophore is excited by a single photon of a particular wavelength. It then returns to the ground state and emits a photon at the (slightly longer)

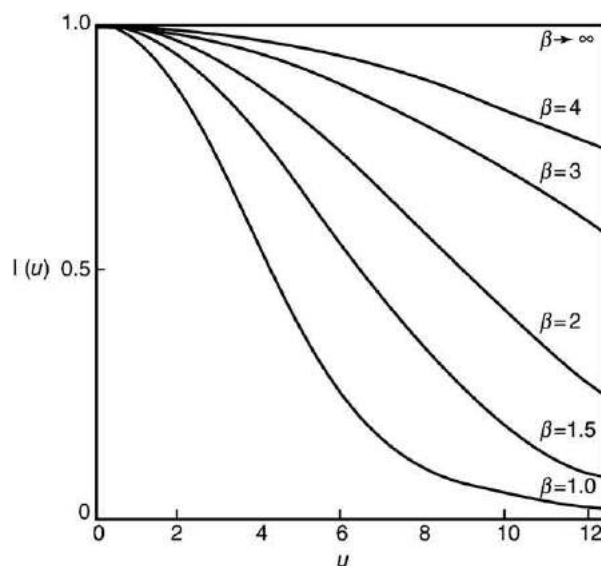


Fig. 8 The detected signal as thin fluorescent sheet is scanned axially through focus for a variety of fluorescent wavelengths. We note that if we measure the sectioning by the half width of these curves, the strength of the sectioning is essentially proportional to β .

fluorescence wavelength. It is this radiation which is detected via the confocal pinhole. Recently, however, much interest has centered on two-photon excitation fluorescence microscopy. This process relies on the simultaneous absorption of two, longer wavelength photons, following which a single fluorescence photon is emitted. The excitation wavelength is typically twice that used in the one-photon case. The beauty of the two-photon approach lies in the quadratic dependence of the fluorescence intensity on the intensity of the illumination. This leads to fluorescence emission which is always confined to the region of focus. In other words the system possesses an inherent optical sectioning property. Other benefits of two-photon fluorescence over the single photon case include the use of red or infra-red lasers to excite ultra-violet dyes, confinement of photo-bleaching to the focal region (the region of excitation), and the reduced effects of scattering and greater penetration. However, it should also be remembered that, compared with single photon excitation, the fluorescence yield of many fluorescent dyes under two-photon excitation is relatively low.

In order to make a theoretical comparison between the one- and two-photon modalities, we shall assume that the emission wavelength λ_{em} is the same, irrespective of the mode of excitation. Since the wavelength required for single-photon excitation is generally shorter than the emission wavelength, we may write it as $\gamma\lambda_{em}$ where $\gamma < 1$. Since two-photon excitation requires the simultaneous absorption of two photons of half the energy we will assume that the excitation wavelength may be written as $2\gamma\lambda_{em}$ which has been shown to be a reasonable approximation for many dyes.

If we now introduce optical coordinates u and v , normalized in terms of λ_{em} , we may now write the effective point spread functions in the one-photon confocal and two-photon case as

$$I_{1p-conf} = I\left(\frac{v}{\gamma}, \frac{u}{\gamma}\right) I(v, u) \quad (7)$$

and

$$I_{2p} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) \quad (8)$$

respectively. We note that although a pinhole is not usually employed in two-photon microscopy it is possible to include one if necessary. In this confocal two-photon geometry the effective point spread function becomes

$$I_{2p-conf} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) I(v, u) \quad (9)$$

If we now look at Eqs. (7) and (8) in the $\gamma = 1$ limit, we find that

$$I_{1p-conf} = I^2(v, u) \quad (10)$$

and

$$I_{2p} = I^2\left(\frac{v}{2}, \frac{u}{2}\right) \quad (11)$$

We now see that because of the longer excitation wavelength used in two-photon microscopy the effective point spread function is twice as large as that of the one-photon confocal in both the lateral and axial directions. The situation is somewhat improved in the confocal two-photon case but it is worth remembering that the advantages of two-photon excitation microscopy are accompanied by a reduction in optical performance as compared to the single-photon case.

From the practical point of view it is worth noting that the two-photon approach has certain very important advantages over one-photon excitation, in terms of image contrast, when imaging through scattering media apart from the greater depth of penetration afforded by the longer wavelength excitation. In a single-photon confocal case it is quite possible that the desired fluorescence radiation from the focal plane may be scattered after generation in such a way that it is not detected through the confocal pinhole. Since the fluorescence is generated throughout the entire focal volume it is also possible that undesired fluorescence radiation, which was not generated within the focal region, may be scattered so as to be detected through the confocal pinhole. In either case this leads to a reduction in image contrast. The situation is, however, completely different in the two-photon

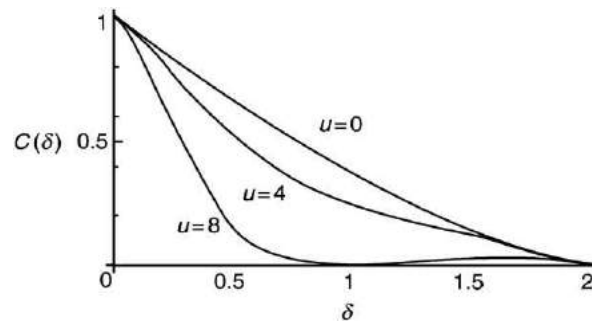


Fig. 9 The transfer function for a conventional fluorescence microscope for a number of values of defocus (measured in optical units). We notice that the response for all nonzero spatial frequencies decays with defocus.

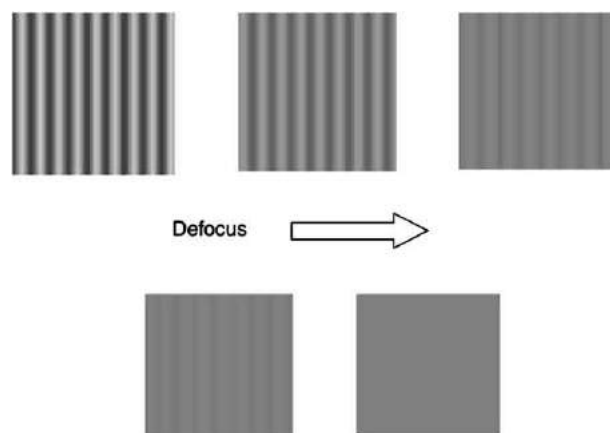


Fig. 10 The image of a one-dimensional single spatial frequency bar pattern for varying degrees of defocus. We note that for sufficiently large values of defocus the bar pattern is not imaged at all.

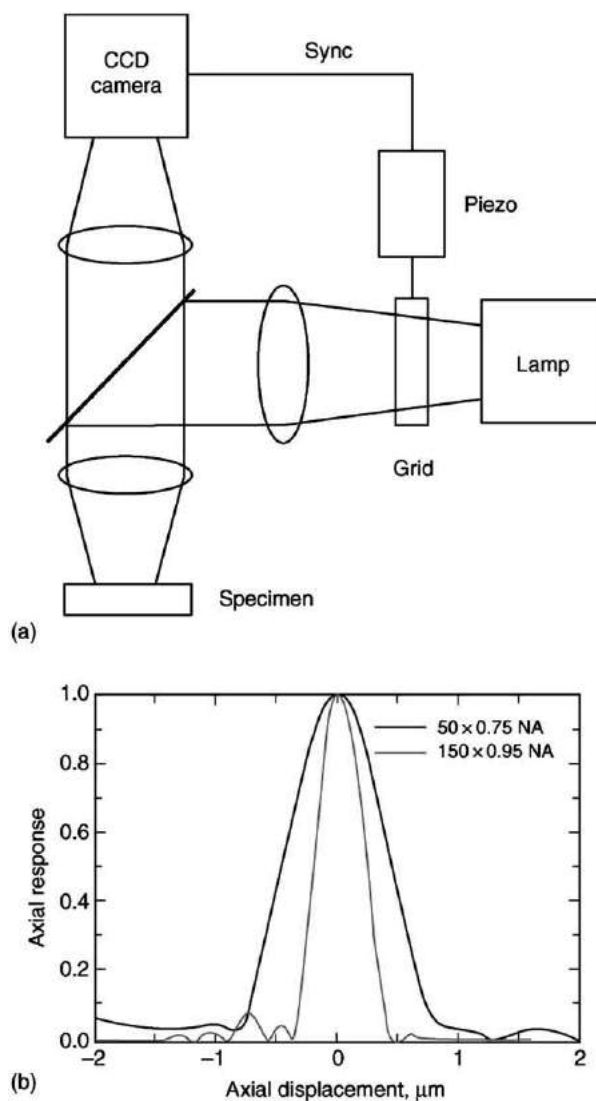


Fig. 11 (a) Shows the optical system of the structured illumination microscope together with experimentally obtained axial responses in (b) which confirm the optical sectioning ability of the instrument.

case. Here the fluorescence is generated only in the focal region and not throughout the focal volume. Furthermore since all the fluorescence is detected via a large area detector – no pinhole is involved – it is not so important if further scattering events take place. This leads to high-contrast images which are less sensitive to scattering. This is particularly important for specimens which are much more scattering at the fluorescence ($\lambda/2$) wavelength than the excitation (λ) wavelength.

The Use of Structured Illumination to Achieve Optical Sectioning

An alternative approach to obtain optical sectioning is to use structured illumination in a conventional microscope. This approach is attractive since the conventional light microscope already possesses many desirable properties: real-time image capture, standard illumination, ease of alignment, etc. However, it does not produce optically sectioned images in the sense usually understood in confocal microscopy. In order to see how this deficiency may be corrected via a simple modification of the illumination system let us begin by looking at the theory of image formation in a conventional fluorescence microscope and start by asking in what way the image changes as the microscope is defocused. We know that in a confocal microscope the image signal from all object features attenuates with defocus and that this does not happen in a conventional microscope. However, when we look closely at the image formation process we find that it is only the zero spatial frequency (constant) component which does not change with defocus (Fig. 9); all other spatial frequencies actually do attenuate with defocus to a greater or lesser extent. Fig. 10 illustrates this by showing the image of a single spatial frequency one-dimensional bar pattern object for increasing degrees of defocus. When the specimen is imaged in focus, a good image of the bar pattern is obtained. However with increasing defocus the image becomes progressively poorer and weaker until it eventually disappears leaving a uniform gray level. This observation is the basis of a simple way to perform optical sectioning in a conventional microscope.

If we simply modify the illumination path of the microscope so as to project onto the object the image of a one-dimensional, single spatial frequency fringe pattern, then the image we see through the microscope will consist of a sharp image of those parts of the object where the fringe pattern is in focus together with an out-of-focus blurred image of the rest of the object. In order to obtain an optically sectioned image it is necessary to remove the blurred out-of-focus portion as well as the fringe pattern from the in-focus optical section. There are many ways to do this – one of the simplest involves simple processing of three images taken at three different spatial positions of the fringe pattern. The out-of-focus regions remain fairly constant between these images and the

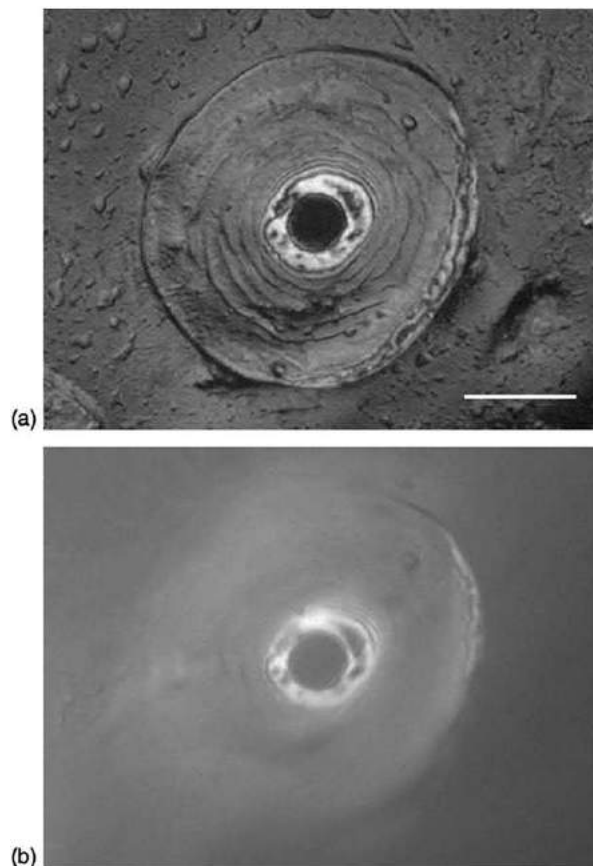


Fig. 12 Two images of the region around the spiracle of a head louse. As in the previous figure, (a) is an autofocus image and (b) shows a mid-focus conventional image. The scale bar depicts a length of 10 μm .

relative spatial shift of the fringe pattern allows the three images to be combined in such a way as to remove the fringes. This permits us to retrieve both an optically sectioned image as well as a conventional image in real time.

Since the approach involves processing three conventional microscope images, the image formation is fundamentally different from that of the confocal microscope. However, the depth discrimination or optical sectioning strength is very similar and this approach, which requires very minimal modifications to the instrument, has been used to produce high-quality three-dimensional images of volume objects which are directly comparable to those obtained with confocal microscopes. A schematic of the optical arrangement is shown in [Fig. 11\(a\)](#), together with experimentally obtained axial responses in [Fig. 11\(b\)](#). It will be noted that these responses are substantially similar to those obtained in the true confocal case. As an example of the kind of images which can be obtained with this kind of microscope, we show in [Fig. 12](#) two images of a spiracle of a head louse. The first is an auto-focus image of greatly extended depth of field constructed from a through-focus series of images. The second is a conventional image taken at mid-focus. The dramatic increase in depth of field is clear when compared with a mid-focus conventional image. We note that these images were taken using a standard microscope illuminator as light source. Indeed, the system is so light efficient that good-quality optical sections have been obtained on transistor specimens using simply a candle as light source!

Imaging using fluorescence light is also possible using this technique. However, an alternative approach that does not require a physical grid is possible if a laser is used as the light source. In this system the laser illumination is split into two beams which are allowed to interfere at an angle in the fluorescent specimen. This has the effect of directly 'writing' a one-dimensional fringe pattern in the specimen. Spatial shift of the fringe pattern is achieved by varying the phase of one of the interfering beams. As before, three images are taken from which both the optically sectioned image and conventional image may be obtained. The beauty of this approach is that no imaging optics are required at the illuminating wavelength and system.

Summary

We have discussed the origin of the optical sectioning property in the confocal microscope in order to introduce the range of imaging modes which this unique form of microscope leads to. A range of optical architectures have also been described. By far the most universal is that shown in [Fig. 1](#) where a confocal module is integrated around a conventional optical microscope. Other, more recent, real-time implementations have also been described. A number of practical aspects of confocal microscopy have not been discussed. This is because they are readily available elsewhere, such as advice on the correct choice of detector pinhole size or because they are still the focus of active research, such as the development of new contrast mechanisms for achieving enhanced three-dimensional resolution such as the stimulated emission depletion method, STED or 4-Pi which are still under development but offer great promise.

See also: Overview

Further Reading

- Corle, T.R., Kino, G.K., 1996. *Confocal Scanning Optical Microscopy and Related Imaging Systems*. New York: Academic Press.
- Denk, W., Strikler, J.H., Webb, W.W., 1990. Two photon laser scanning fluorescence microscopy. *Science* 248, 73.
- Diaspro, A., 2002. *Confocal and Two-Photon Microscopy*. New York: Wiley-Liss.
- Gu, M., 1996. *Principles of Three-Dimensional Imaging in Confocal Microscopes*. Singapore: World Scientific.
- Hell, S., Stelzer, E.H.K., 1992. Fundamental improvements of resolution with a 4Pi-confocal fluorescence microscope using two-photon excitation. *Optics Communication* 93, 277.
- Hell, S., Wichmann, J., 1994. Breaking the diffraction limit by stimulated emission. *Optics Letters* 19, 870.
- Lukosz, W., 1966. Optical systems with resolving power exceeding the classical limit. *Journal of the Optical Society of America* 56, 1463.
- Masters, B.R., 1996. *Confocal Microscopy*. Bellingham: SPIE.
- Minsky, M., 1961. *Microscopy Apparatus*. United States Patent 3,013,467, Dec 19, 1961 (Filed Nov 7, 1957).
- Neil, M.A.A., Juskaitis, R., Wilson, T., 1997. Method of obtaining optical sectioning by using structured illumination in a conventional microscope. *Optics Letters* 22, 1905.
- Pawley, J.B. (Ed.), 1995. *Handbook of Biological Confocal Microscopy*. New York: Plenum Press.
- Wilson, T. (Ed.), 1990. *Confocal Microscopy*. London: Academic Press.
- Wilson, T., Sheppard, C.J.R., 1984. *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Atom Optics

AD Cronin and DE Pritchard, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Elsevier Ltd. All rights reserved.

In 1923 Louis de Broglie suggested that particles of matter propagate as waves with a wavelength

$$\lambda_{dB} = h/p \quad (1)$$

where h is Planck's constant and p is the particle's momentum. This breakthrough idea motivated Schrödinger's equation, which is the equation of motion of quantum mechanics.

Schrödinger's equation is a wave propagation equation, and therefore implies the existence of a whole new type of optics – matter wave optics – in which electron waves, neutron waves, atom waves, and entire molecule waves can be manipulated coherently. Such optics enable new measurement technologies and devices that depend on matter wave interference.

Guided by a knowledge of light waves, one can understand that the de Broglie wavelength of a particle sets the scale for focusing, diffraction, and interference. However, there are differences between matter and light. For example, the vacuum is tremendously dispersive for matter waves. In other words, the group velocity for a particle depends on its de Broglie wavelength, while for light waves the group velocity is independent of wavelength.

To illuminate the mathematical connection with light optics, we note the wave equation for light is

$$\nabla^2 \vec{E} = \mu \varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} \quad (2)$$

where $\vec{E} = \vec{E}(\vec{x}, t)$ is the optical electric field, and the parameters μ and ε may vary throughout space, describing an index of refraction. The wave equation for matter waves is Schrödinger's equation:

$$\left(-\frac{\hbar^2 \nabla^2}{2m} + V \right) \psi = i\hbar \frac{\partial}{\partial t} \psi \quad (3)$$

where $\psi = \psi(\vec{x}, t)$ is the probability amplitude, or wavefunction, for a particle of mass m . Here, the potential energy V can vary throughout space, and thus control the phase and amplitude of propagating atoms. Just as density for photons is proportional to $\vec{E}(\vec{x})^2$, the probability density for particles is given by $|\psi(\vec{x})|^2$.

In a vacuum $V=0$ and $(\mu_0 \varepsilon_0)^{-1/2}=c$, and if we assume that both ψ and E have a time dependence described by $e^{-ik\omega}$, the wave equations become

$$\left[\nabla^2 + \frac{\omega^2}{c^2} \right] \vec{E} = 0 \quad (4)$$

and

$$\left[\nabla^2 + \frac{2m\omega}{\hbar} \right] \psi = 0 \quad (5)$$

which look similar, but have a critical difference stemming from the single time derivative in Schrödinger's equation. To see how this makes the vacuum dispersive for matter waves, but not for light, consider a propagating plane wave of either ψ or $\vec{E}$ described by $e^{i(kx - \omega t)}$ (for either kind of wave, $k \equiv 2\pi(\lambda)^{-1}$). Using the wave equation for light waves, the dispersion relation is

$$\omega = ck \quad (6)$$

where c is the phase (and group) velocity of light, which is independent of k . The same plane wave substituted into Schrödinger's equation for matter waves yields the dispersion relation

$$\omega = \frac{\hbar k^2}{2m} \quad (7)$$

This is quadratic for matter waves, which means the group velocity $v_g = d\omega/dk$ is twice the phase velocity $v_p = \omega/k$, and both still depend on k .

From an engineering perspective, matter wave optics consists of lenses, mirrors, beam-splitters, and other components of an optics toolkit. The fact that electrons and neutrons can be transmitted through solid material, and that electrons can be reflected or refracted by static electromagnetic fields, has enabled optical elements to be created for these kinds of matter waves. For atoms, however, finding suitable materials for lenses or mirrors is a daunting task, since with rare exceptions (e.g., cold H on liquid He) atoms thermalize on surfaces rather than elastically bouncing off or passing through them. This was the major obstacle which delayed the blossoming of atom optics until the mid-1980s.

This obstacle has now been circumvented, but only with difficulty. Electromagnetic fields (whose spatial configurations are severely constrained by Maxwell's equations) can change the potential energy of neutral atoms, but since an atom's net charge is zero, provide only a weak and highly dispersive index of refraction based on the atom's polarizability or magnetic moment. However, these effects are tremendously enhanced if the electromagnetic fields oscillate near a transition frequency of the atom.

Hence, propagating waves and standing waves of single-frequency laser light have become standard tools in atom optics. Standing light waves are often referred to as a light crystal, because the periodic potential can diffract atom beams.

An alternative source of atom optical components is based on nanofabrication. Thin sheets of material with patterned holes serve as absorptive atom optics by transmitting atom waves only through the nanometer scale openings. A variety of techniques have been used to make physical diffraction gratings, zone plates, and holograms for atom waves. Methods such as electron beam lithography and ultraviolet photolithography both rely on sophisticated etching procedures to make the final free-standing structures. Thus, in either approach – near-resonant light, or miniature physical structures – current technologies such as the tunable laser or nanofabrication techniques have been required for the development of useful atom optics.

While coherent manipulation of atom waves is often the goal, incoherent manipulations are possible also, e.g., if photons are spontaneously scattered. Dissipative processes offer an additional class of possibilities such as increasing the brightness of an atom source, and are therefore an important part of atom optics. This article excludes atom slowing and trapping, and focuses instead on atom diffraction – a coherent process which can be used to make beam splitters for atom interferometers.

Nanofabricated Atom Optics

Because atoms stick to surfaces, mechanical structures are absorptive atom optics. Yet the de Broglie waves transmitted through slots of a mask are nearly unaffected, provided the mask is thin enough. This is because the van der Waals potential energy of interaction with nearby surfaces falls off rapidly for neutral atoms.

Consider, for example, an atom wave packet being diffracted by a mechanical grating made with physical bars and slots as shown in Fig. 1. Initially, let the atom wave packet have a slowly varying spatial envelope, $f=f(x, y, z)$, and momentum only in the $\hat{x}$ -direction, $\vec{p} = \hbar \vec{k}_0 = \hbar(\lambda_{dB})^{-1} \hat{x}$, so the wavefunction can be written as

$$\psi = f e^{i(k_0 x - \omega t)} \quad (8)$$

Immediately after the grating, only the portions of the wavefunction passing through slots remain nonzero, as depicted in Fig. 2. Here, the wavefunction is now periodic both due to the grating lattice vector $k_g \hat{y}$ and the initial wavevector $k_0 \hat{x}$. According to the Schrödinger equation, a periodicity in any direction implies a momentum in that direction. Thus, variation of ψ in the $\hat{y}$ variation gives rise to a new momentum distribution.

The square wave modulating the envelope of the wave packet can be described by a Fourier sum times the incident wavefunction

$$\psi = f e^{i(k_0 x - \omega t)} \sum_{n=-\infty}^{\infty} a_n \cos(nk_g y) \quad (9)$$

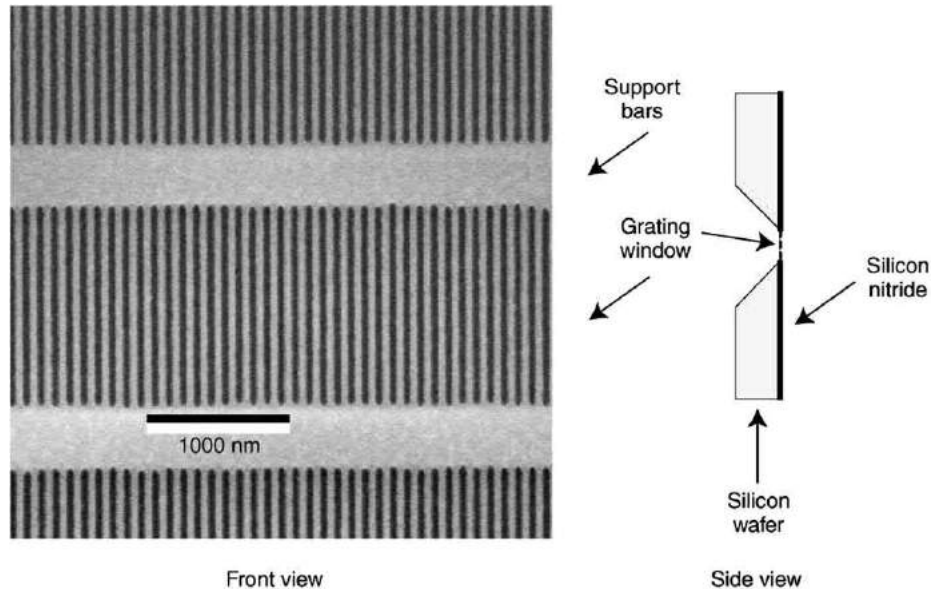


Fig. 1 Scanning electron microscope image of a silicon nitride diffraction grating for atom waves. The bars (light) absorb, and the slots (dark) transmit atom waves. The grating period is 100 nm, the open fraction is 0.67, and the free-standing bars are supported every 5 μm with a thicker strip of silicon nitride material. The grating was fabricated by Tim Savas and Henry I Smith at MIT NanoStructures laboratory.

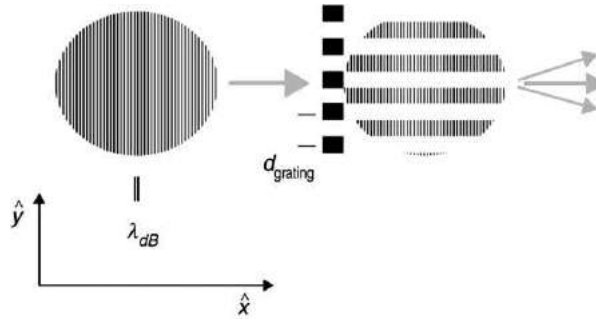


Fig. 2 Schematic diagram of an atom wavepacket passing through an absorbing grating, as seen from above. The atom wavepacket is depicted before and after passing through the grating from left to right.

where $k_g = 2\pi(d_g)^{-1}$ is the grating lattice vector. The only other physical parameter of the grating is the open fraction, γ , and this determines the Fourier amplitudes a_n

$$a_n = \frac{\sin(n\pi\gamma)}{n\pi} \quad (10)$$

From the principle of superposition (which is applicable here because the Schrödinger equation is linear) the wavefunction can be written as a sum of traveling waves with wavevectors $\vec{k}_n$:

$$\psi = f \sum_{n=-\infty}^{\infty} a_n e^{(i\vec{k}_n \cdot \vec{r} - \omega t)} \quad (11)$$

$$\vec{k}_n = \frac{k_0 \hat{x} + nk_g \hat{y}}{N} \quad (12)$$

where, just as in standard optics, the factor N serves to keep $|\vec{k}_n|$ on a spherical shell, i.e., the grating does not add energy to the atom.

Inserting this expression for the wavefunction into Schrödinger's equation we get the dispersion relation for matter waves and a different group velocity, v_n , for each k_n

$$\hbar\omega = \frac{\hbar^2 k_n^2}{2m} \quad (13)$$

$$v_n = \frac{d\omega}{dk_n} = \frac{\hbar k_n}{m} = p_n/m \quad (14)$$

Thus, diffraction through a grating puts an atom into a coherent superposition of different momenta, each separated by $\sim \hbar k_g \hat{y}$. Some time later, the atom will be in a coherent superposition of multiple locations, and the far-field probability density will be peaked at integer multiples of the diffraction angle

$$\theta_D = \frac{k_g}{k_0} = \frac{\lambda_{dB}}{d_{\text{grating}}} \quad (15)$$

where d_{grating} is the lattice spacing of the grating and the small-angle approximation for $\sin(\theta)$ has been used. Exactly as in light optics, these diffraction gratings are good momentum spectrometers, i.e., the diffraction angle is a measure of longitudinal momentum. (Contrary to light optics, group velocity can also be used to measure matter wave momentum.) Coherence between the diffraction orders remains until a perturbation occurs which would (even in principle) be sufficient to determine which path the atom took. This assertion has been tested with atoms in an interferometer (discussed below).

Atom flux in the n th diffraction order is given by

$$I_n = \left(\frac{\sin(n\pi\gamma)}{n\pi} \right)^2 \quad (16)$$

To resolve the diffracted orders the initial transverse momentum distribution of the atoms must be smaller than k_g , and the time until the far-field observation must be sufficiently large given the beam width. At intermediate times, the atom waves are in the Fresnel diffraction region where the Talbot effect can be observed. This effect causes the atom waves to form a replica of the grating in free space at integer multiples of the Talbot length, $Z_T = 2d_{\text{grating}}^2 \lambda_{dB}^{-1}$.

Fig. 3 shows the far-field diffraction pattern from a beam of sodium (^{23}Na) atoms using a material grating with a period of 100 nm. The atom beam velocity peaked at 1500 ms^{-1} corresponds to a de Broglie wavelength of $0.11 \text{ \AA}$ thus, $\theta_D \approx 10^{-4} \text{ rad}$. (The beam of 3000 ms^{-1} atoms diffracts at half this angle.) The spread in longitudinal velocity of $\sigma_v/v \approx 6\%$ can be deduced from the data because it slightly increases the width for higher order diffraction peaks. The open fraction of the grating, $\gamma \approx 60\%$, can also be determined from the diffraction data, because it determines the relative intensity of the orders. Effects of the atoms' finite size and interaction with the grating bars make small corrections to this simple theory.

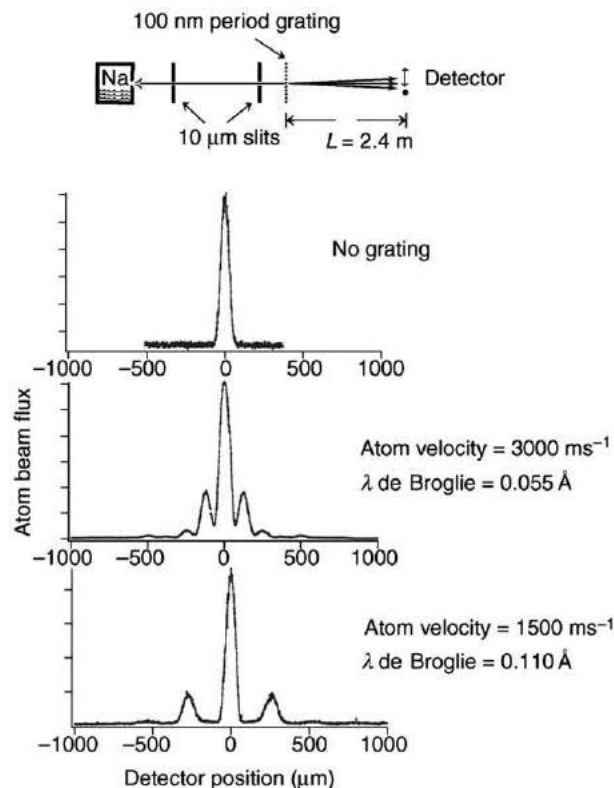


Fig. 3 Diffraction data from a beam of sodium formed by passing atoms through two collimating slits followed by a nanofabricated diffraction grating. The momentum of the atoms determines their de Broglie wavelength, and thus their diffraction angle $\sin(\Theta_D) = \lambda_{dB}/d_{\text{grating}}$. These data were obtained at MIT by the authors.

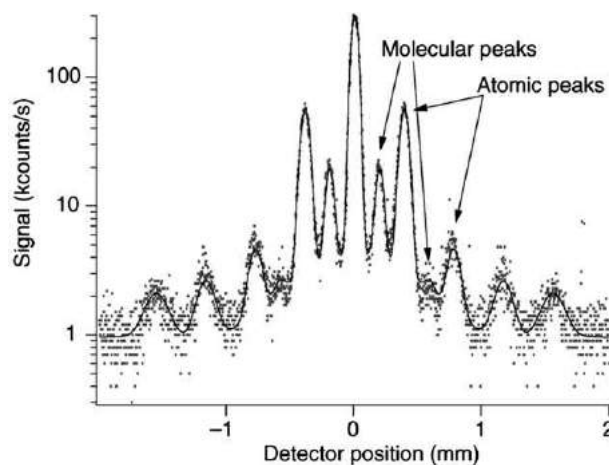


Fig. 4 Diffraction data using a mixed beam of sodium atoms (Na) and sodium dimer molecules (Na₂). From this source molecules have the same speed as atoms, thus twice the momentum. Thus, molecule waves are diffracted at half the diffraction angle compared to atom waves. These data were obtained at MIT by the authors.

Nanofabricated atom optics differ from those based on light forces in a number of ways: they are amplitude structures (with corresponding loss of transmission intensity), they are species-independent, their scale size can be several times smaller than attainable with light, and they can be arbitrarily patterned since they are fabricated by electron beam lithography. Insensitivity to species is demonstrated in Fig. 4, where a beam of Na₂ molecules and Na atoms is diffracted and separated by a nanofabricated diffraction grating. The flexibility of electron beam lithography has allowed fabrication not only of diffraction gratings, but also of spherical and cylindrical zone plates, as well as a combination of lens and a hologram that generates a focused atom image. One stunning application laid to rest a long-standing argument concerning whether a stable bound state of the ⁴He₂ dimer exists. For

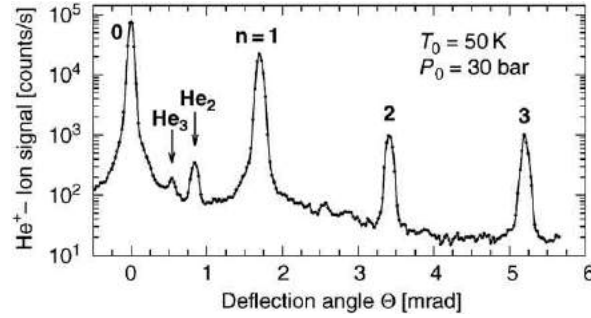


Fig. 5 Diffraction of helium atoms and helium molecules through a nanofabricated grating. These data, reproduced with permission from Wieland Schoellkopf, were obtained at the Max-Planck-Institute in Göttingen.

this a diffraction grating was used to separate and resolve $^4\text{He}_2$ and more massive clusters (Fig. 5). Subsequently a nano-sieve was used to estimate the size of the $^4\text{He}_2$ dimers to be $62 \pm 10 \text{ \AA}$.

Standing Waves of Light

In contrast to nanofabricated structures, standing waves of near-resonant light are phase gratings for atoms. The optical electric field dynamically polarizes the atoms and causes a shift to the ground state energy (the ac-Stark shift). For optical electric fields oscillating above (below) the atomic resonance frequency, the ac-Stark shift increases (decreases) potential energy for atoms and in either case perturbs the de Broglie wave phase.

In many respects, the interaction of an atom with a standing light wave is richer than the more familiar topic of light-atom interactions in a traveling light wave. Part of this richness reflects two ways in which a standing wave can be considered, either as two counter-propagating traveling waves or as a single, stationary standing wave. The standing wave-atom interaction is capable of transferring momentum in well-determined quantities, coherently splitting atomic wave packets, and generating forces much larger than possible with spontaneously scattered light. Not surprisingly, this interaction has many distinct facets that only appear in different regimes of the interaction parameters (intensity, detuning, and pulse duration) and atomic parameters (mass, initial momentum with respect to the standing wave, and excited state natural lifetime).

We begin with the Bragg scattering of atoms from a standing wave light grating. Although it can be difficult to realize the physical conditions that assure its occurrence, pure Bragg scattering is simpler than intermediate cases involving spontaneous decay (from light too close to resonance) or Kapitza-Dirac diffraction (from shorter interaction times).

Consider a standing wave light grating formed by two counter-propagating plane waves (traveling parallel to the z -axis) of equal amplitude, E_0 , wavevector, k , frequency, ω , polarization vector, $\hat{e}$, and temporal envelope function, $f(t)$:

$$\vec{E}(z, t) = E_0 f(t) \sin(kz - \omega t) \hat{e} + E_0 f(t) \sin(kz + \omega t) \hat{e} \quad (17)$$

$$= 2E_0 f(t) \sin(kz) \cos(\omega t) \hat{e} \quad (18)$$

We would like to work with momentum states as our atomic basis, thus it is easiest to consider the description of the electric field driving the transitions in terms of two counter-propagating traveling waves of definite momentum (Eq. (17)), as opposed to the single standing wave they jointly form (Eq. (18)).

Momentum is transferred by paired stimulated absorption and emission processes, resulting in a transfer of photons between the traveling waves. An N_B^{th} order diffraction process transfers N_B photons from one traveling wave to the counter-propagating traveling wave and changes the atomic momentum by $2N_B \hbar k \hat{z}$. Furthermore, atomic population is transferred only between $|g, -N_B \hbar k\rangle$ and $|g, +N_B \hbar k\rangle$, where $|g(e), \pm N \hbar k\rangle$ denotes a two-level atom in its ground (excited) state with momentum $\pm N \hbar k$ parallel to the standing wave axis. The excited state remains nominally unpopulated so long as the temporal envelope function, $f(t)$, does not have strong frequency components near the laser detuning

$$\delta = \omega - \omega_0 \quad (19)$$

where ω_0 is the unperturbed frequency of the atomic transition. Furthermore, for a given initial state $(|g, -N_B \hbar k\rangle)$ the uniqueness of the final state $(|g, +N_B \hbar k\rangle)$ occurs because of the fundamental assumption that makes Bragg scattering so simple to describe; the uncertainty in the photon energy driving the transitions is small compared to the energy separation between neighboring momentum states. A quantitative discussion of the validity of this assumption will be given later.

We now calculate the probability, $P_1^B(\tau)$, of the first order ($N_B = 1$) Bragg process. Scattering transfers an atom from $|g, -\hbar k\rangle$ to $|g, +\hbar k\rangle$ when the atoms interact with a constant light intensity for a time τ (i.e., in Eqs. (17) and (18) $f(t)$ is a square wave of unit amplitude and duration τ). The geometry for Bragg diffraction from a standing wave of light is shown in Fig. 6, and the energy/momentum states involved in the transition are depicted in Fig. 7.

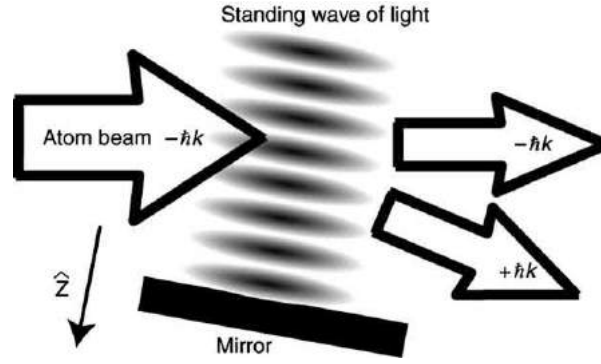


Fig. 6 Two counter-propagating running waves of light superimpose to form a standing wave of light. This 'light crystal' acts as a diffraction grating for atoms, shown here in the configuration for Bragg diffraction.

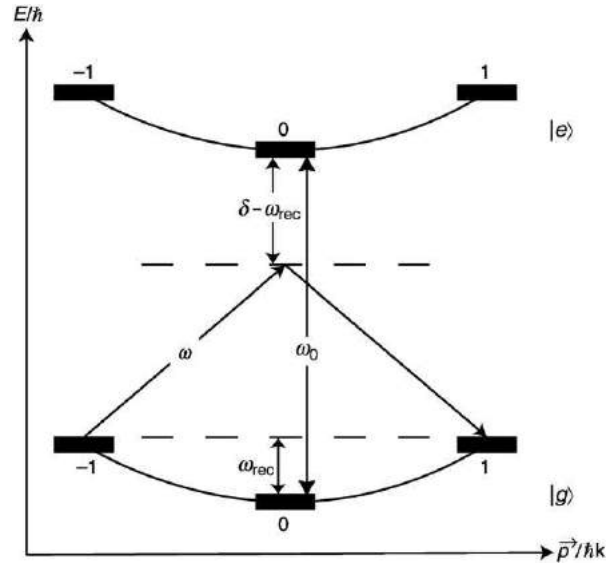


Fig. 7 Energy vs. momentum diagram describing atomic states involved in Bragg diffraction. The total energy is due to the electronic state ($|g\rangle$, $|e\rangle$) plus the kinetic energy associated with the recoil from photon emission (or absorption). In Bragg diffraction, atoms make a coherent transition from $|g, -\hbar k\rangle \leftrightarrow |g, +\hbar k\rangle$. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

Considering the electronic degree of freedom, we rewrite the Schrödinger equation as

$$(\mathcal{H}_0 + \mathcal{H}_{\text{int}})|\psi\rangle = i\hbar \frac{\partial}{\partial t} |\psi\rangle \quad (20)$$

where the total Hamiltonian $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\text{int}}$ is a matrix. The wavefunction is a vector of coefficients c_i in the basis $\{|g, -\hbar k\rangle, |e, 0\rangle, |g, +\hbar k\rangle\}$, with normalized population amplitudes for each component:

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t) \\ c_0(t) \\ c_{+1}(t) \end{pmatrix} \quad (21)$$

In the electric dipole approximation, the interaction Hamiltonian is $\mathcal{H}_{\text{int}}(t) = -\vec{\mu} \cdot \vec{E}(t)$, where $\mu = \langle e | q\vec{r} | g \rangle \cdot \hat{e}$ is the electric dipole matrix element connecting the ground ($|g\rangle$) and excited ($|e\rangle$) states of the atom. $\hat{e}$ is the polarization vector of the light and q is the charge of the electron. By momentum conservation, only the plane wave traveling in the $\pm \hat{z}$ direction couples the $|e, 0\rangle \leftrightarrow |g, \pm \hbar k\rangle$ transitions. By using this argument, we are effectively viewing the electric field as a quantum mechanical operator. Expanding the sinusoidal variation of the electric field in terms of complex exponentials and treating the spatially dependent

complex exponential terms as quantum mechanical momentum translation operators yields the interaction Hamiltonian:

$$\mathcal{H}_{\text{int}} = \frac{\hbar\omega_R}{2} \begin{pmatrix} 0 & ie^{i\omega t} & 0 \\ -ie^{-i\omega t} & 0 & ie^{-i\omega t} \\ 0 & -ie^{i\omega t} & 0 \end{pmatrix} \quad (22)$$

For example, $e^{\pm ikz}|g, n\hbar k\rangle = |e, (n\pm 1)\hbar k\rangle$ describes absorption, and for stimulated emission, $e^{\mp ikz}|e, n\hbar k\rangle = |g, (n\pm 1)\hbar k\rangle$. This procedure includes the rotating wave approximation. In addition, we have neglected any frequency components associated with the sudden switch on of the fields and the finite duration of the light-atom interaction. In formulating $\mathcal{H}_{\text{int}}$, the strength of the light-atom interaction is parameterized by the single-photon Rabi frequency:

$$\omega_R = \frac{\mu E_0}{\hbar} \quad (23)$$

Without loss of generality, we will take μ and hence ω_R to be positive, real-valued quantities.

The total Hamiltonian follows simply by including the electronic and kinetic energy terms:

$$\mathcal{H}_0 = \hbar \begin{pmatrix} \omega_{\text{rec}} & 0 & 0 \\ 0 & \omega_0 & 0 \\ 0 & 0 & \omega_{\text{rec}} \end{pmatrix} \quad (24)$$

where the single-photon recoil energy, E_{rec} , of an atom of mass m is given by

$$E_{\text{rec}} = \hbar\omega_{\text{rec}} = \frac{\hbar^2 k^2}{2m} \quad (25)$$

Making the *ansatz* for the solution wavefunction as

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t)e^{-i\omega_{\text{rec}}t} \\ c_0(t)e^{-i\omega_0t} \\ c_{+1}(t)e^{-i\omega_{\text{rec}}t} \end{pmatrix} \quad (26)$$

and substituting into the Schrödinger equation yields three coupled first-order differential equations:

$$\dot{c}_{\pm 1}(t) = \mp \frac{\omega_R}{2} e^{i\Delta t} c_0(t), \quad (27)$$

$$\dot{c}_0(t) = \frac{\omega_R}{2} e^{-i\Delta t} (c_{+1}(t) - c_{-1}(t)), \quad (28)$$

where $\Delta = \delta + \omega_{\text{rec}} = (\omega - \omega_0) + \omega_{\text{rec}}$. Differentiating Eq. (27) and substituting Eq. (28) into the result yields two coupled second-order differential equations:

$$\ddot{c}_{\pm 1}(t) - i\Delta \dot{c}_{\pm 1}(t) + \frac{\omega_R^2}{4} (c_{\pm 1}(t) - c_{\mp 1}(t)) = 0 \quad (29)$$

With the initial conditions:

$$c_{-1}(0) = 1 \quad (30)$$

$$c_0(0) = 0 \Rightarrow \dot{c}_{\pm 1}(0) = 0 \quad (31)$$

$$c_{+1}(0) = 0 \quad (32)$$

and the assumption $\Delta^2 \gg \omega_R^2$, the solutions to Eq. (29) are

$$c_{-1}(t) = e^{-i\frac{\omega_R^{(2)}}{2}t} \cos\left(\frac{\omega_R^{(2)}}{2}t\right) \quad (33)$$

$$c_{+1}(t) = ie^{-i\frac{\omega_R^{(2)}}{2}t} \sin\left(\frac{\omega_R^{(2)}}{2}t\right) \quad (34)$$

where the two-photon Rabi frequency is

$$\omega_R^{(2)} = \frac{\omega_R^2}{2\Delta} \rightarrow \frac{\omega_R^2}{2\delta}, \quad |\delta| \gg \omega_{\text{rec}} \quad (35)$$

with both transitions driven at equal single-photon Rabi frequencies, ω_R .

Substituting the solution of Eqs. (33) or (34) into Eq. (27) yields an expression for the excited state amplitude:

$$c_0(t) = i\frac{\omega_R}{2\Delta} e^{-i\Delta t} e^{-i\omega_R^{(2)}t} \quad (36)$$

This will be important in calculating the rate of spontaneous emission events later.

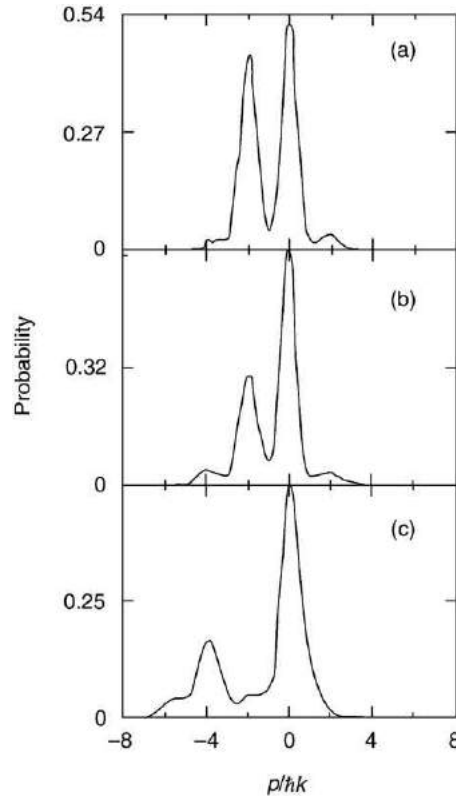


Fig. 8 Bragg scattering of an atom beam from an optical standing wave. First-order Bragg scattering at (a) lower power, (b) higher power of the laser beams (Pendellösung has set in, increasing the amplitude of the undeflected peak), (c) second-order Bragg scattering. Reprinted with permission from *Physical Review Letters* 60(6): 515–518. Copyright (1998) by the American Physical Society.

The solutions for $c_{-1}(t)$ and $c_{+1}(t)$ oscillate with the interaction duration, τ , yielding the result for the $|g, -\hbar k\rangle \rightarrow |g, +\hbar k\rangle$ transition probability:

$$P_1^B(\tau) = |c_{+1}(\tau)|^2 = \sin^2\left(\frac{\omega_R^{(2)}}{2}\tau\right) \quad (37)$$

Thus, the system oscillates between the two momentum states $|g, -\hbar k\rangle$ and $|g, +\hbar k\rangle$ in a manner analogous to the Rabi oscillation of atomic population between two resonantly coupled states. This solution with oscillatory probabilities for the two Bragg coupled states is known as the Pendellösung and has been observed for atoms, neutrons, and X-rays (Fig. 8). The nice feature here is that the strength of the grating can be actually controlled by the intensity of the light.

Viewing Bragg scattering as a two-photon transition from the initial ground state to the final ground state with opposite momentum, illuminates the close connection with a Raman transition between two internal sub-states of the ground state manifold, each with its own external momentum state. The formalism describing the Raman transition is basically the same as that presented here, except the two transitions can be driven at different single-photon Rabi frequencies, ω_{R1} and ω_{R2} , so that the generic two-photon Rabi frequency is given by $\omega_R^{(2)} = \omega_{R1}\omega_{R2}/2\Delta$, where Δ is the detuning from the intermediate state.

An N_B^{th} order Bragg process (similar to a $2N_B$ -photon Raman process) is a coherent succession of N_B two-photon transitions from $|g, -N_B\hbar k\rangle$ to $|g, +N_B\hbar k\rangle$ with $2N_B - 1$ intermediate states of the form:

$$|e, (-N_B + 1)\hbar k\rangle, |g, (-N_B + 2)\hbar k\rangle, \dots \\ \dots |g, (N_B - 2)\hbar k\rangle, |e, (N_B - 1)\hbar k\rangle$$

Such a process is characterized by a $2N_B$ -photon Rabi frequency given by

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{2N_B-1}\Delta_1\Delta_2\cdots\Delta_{2N_B-1}} \quad (38)$$

where Δ_n is the detuning from the n^{th} intermediate state. Fig. 9 shows what this process would look like for an N_B^{th} order Bragg transition where the intermediate state detunings are given by

$$\Delta_n = \begin{cases} \delta + (2N_B n - n^2)\omega_{\text{rec}} & n \text{ odd} \\ (2N_B n - n^2)\omega_{\text{rec}} & n \text{ even} \end{cases} \quad (39)$$

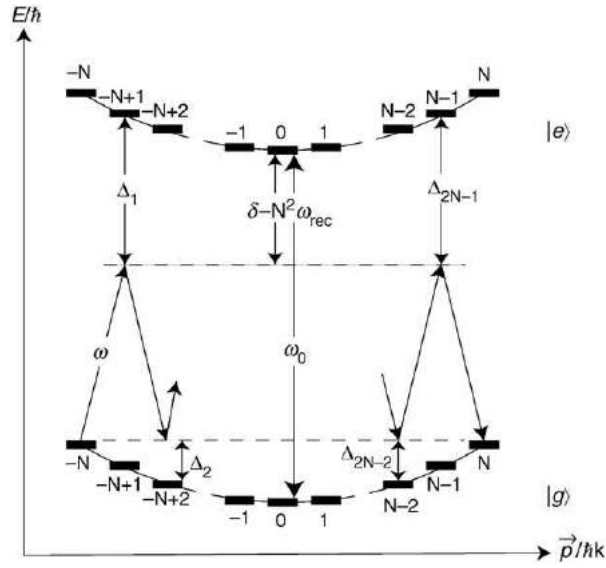


Fig. 9 Energy vs. momentum diagram describing atomic states involved in higher-order Bragg diffraction. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

Substituting these detunings into Eq. (38) yields the N_B^{th} order Bragg transition $2N_B$ -photon Rabi frequency, $\omega_R^{(2N_B)}$:

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{4N_B-3} [(N_B - 1)!]^2 \delta^{N_B} \omega_{\text{rec}}^{N_B-1}} \quad (40)$$

where we have assumed $|\delta| \gg N_B^2 \omega_{\text{rec}}$.

To ensure that the system truly undergoes Bragg scattering, and validate the assumption that only states of equal kinetic energy and opposite momentum are coupled, the overall exposure time, τ , of the atoms to the fields must be limited both from below and above.

The lower bound on τ is necessary to resolve the final momentum state ($|g, N_B \hbar k\rangle$) from neighboring momentum states ($|g, (N_B \pm 2) \hbar k\rangle$) two photon recoil momenta away. Avoiding population transfer into these states requires that the energy separation be resolvable, or

$$\tau(\delta E) \gg h \quad (41)$$

For first-order Bragg scattering processes ($N_B = 1$) the nearest lying momentum states that may be mistakenly populated are $|g, \pm 3 \hbar k\rangle$, for which

$$\delta E = \frac{p_{n=3}^2}{2m} - \frac{p_{n=1}^2}{2m} = 8\hbar\omega_{\text{rec}} \quad (42)$$

Therefore

$$\tau \gg \frac{h}{8\hbar\omega_{\text{rec}}} = \frac{\pi}{4\omega_{\text{rec}}} \quad (43)$$

For sodium atoms (^{23}Na) in 590 nm light, $\tau \gg 6 \mu\text{s}$ is the lower bound required for first-order Bragg scattering.

For all higher-order ($N_B > 1$) Bragg scattering processes, the nearest momentum states are $|g, \pm (N_B - 2) \hbar k\rangle$, which limits the interaction time to

$$\tau \gg \frac{\pi}{2(N_B - 1)\omega_{\text{rec}}} \quad (44)$$

which for $N_B \gg 1$ reduces to

$$\tau \gg \frac{\pi}{2N_B\omega_{\text{rec}}} \quad (45)$$

The upper bound on the interaction duration is necessary to avoid spontaneous emission. The interaction duration must be short enough so that the expected number, N_s , of spontaneous emission events per atom during the time τ is negligible. N_s is simply given by the product of the excited state fraction (Eq. (36)) and the probability of spontaneous decay given that the atom is in the excited state:

$$N_s = |c_0(t)|^2 \Gamma \tau = \frac{\omega_R^2}{4\Delta^2} \Gamma \tau \quad (46)$$

where Γ is the natural decay rate of the excited state. Avoiding spontaneous emission ($N_s \ll 1$) while still having a significant probability for transitions ($\omega_R^{(2N_B)} \tau \simeq \pi$) forces one to work in the regime where $\Delta \gg \Gamma$, which is a practical requirement to avoid spontaneous emission and maintain coherence. For sodium atoms on the 590 nm transition, $\Gamma = 6.3 \times 10^7 \text{ s}^{-1}$, and a typical choice in Bragg diffraction experiments is $\Delta = 50\Gamma$ or $\Delta/2\pi = 500 \text{ MHz}$.

Briefly, to relate Rabi frequency to optical intensity, the transition rate W on resonance with a Lorentzian spectral line is given by Fermi's golden rule:

$$W = \frac{\omega_R^2}{\Gamma} = \frac{\lambda_{\text{photon}}^2}{2\pi} \frac{I}{\hbar\omega} \quad (47)$$

where the last term can be regarded as the cross section for absorption multiplied by intensity in units of photon flux. Hence

$$\omega_R^2 = \Gamma \frac{\lambda_{\text{photon}}^2}{2\pi} \frac{I}{\hbar\omega} \quad (48)$$

For Na atoms in $12 \text{ mW}(\text{cm})^{-2}$ resonant light, $\omega_R = \Gamma = 6.3 \times 10^7 \text{ s}^{-1}$. Both the probability of Bragg scattering and the spontaneous emission increase linearly with $\omega_R^2 \tau \Gamma^{-1}$. However, spontaneous emission decreases as Δ^2 while the Bragg scattering probability decreases as Δ . Thus, as shown in Fig. 10, at sufficient detuning and laser power, Bragg scattering occurs with no spontaneous emission. For reference, $12 \text{ mW}(\text{cm})^{-2} \times 10 \mu\text{s}$ corresponds to $\omega_R^2 \tau \Gamma^{-1} = 630$ for sodium atoms.

Bragg scattering of atoms from a standing light wave was first observed at MIT in 1988. A supersonic atomic beam was diffracted from a standing wave of near-resonant laser light as depicted in Fig. 6. The angle between the atomic beam (of thermal wavelength λ_{dB}) and the light grating (of periodicity $\lambda_L/2$) was tuned to the appropriate Bragg angle, θ_B , where:

$$\lambda_{dB} = \lambda_L \sin(\theta_B) \quad (49)$$

Population transfer corresponding to both first- and second-order Bragg scattering was observed (Fig. 7). The experiment required a sub-recoil transverse momentum spread of the atomic beam in order to resolve the different momentum states in the far field and limit the final state to only one diffracted order. The Pendellösung was observed as an oscillation in population transfer as a function of standing wave intensity, $I \propto \omega_R^{(2)}$, for a fixed interaction time, τ .

Atomic beam diffraction from an optical standing wave is a continuous-wave (CW) experiment in which the selectivity needed for the Bragg process is imposed by good angular resolution of the particle beam and a high degree of parallelism between the light crystal planes. This ensures that of the various final Bragg orders allowed by momentum conservation, only one conserves energy (energy conservation is exact in a CW experiment). For atoms scattering from a light crystal, parallelism of the crystal planes requires highly parallel photon momentum that implies a minimum width of the standing wave (the diffraction limit for the collimated photons). The transit time, τ , of the atoms across this width then exceeds the lower bound given in Eq. (44).

The excellent collimation required of the atomic beam to ensure resolution of the Bragg scattered atoms reduces the intensity of the source by many orders of magnitude. A Bose-Einstein Condensate (BEC) is an attractive alternative source of atoms because its momentum spread is typically an order of magnitude below a single-photon recoil momentum. To Bragg diffract atoms initially in a stationary BEC, it is easier to move the light crystal than to accelerate the condensate. This is done by simply frequency shifting one of the traveling waves so that the resultant standing wave formed by its interference with the unshifted traveling wave moves with the proper velocity ($N_B \hbar k/m$) relative to the stationary atoms to impart the necessary momentum. The Bragg scattered atoms will then have momentum $2N_B \hbar k$ in the laboratory frame. The resonance condition thus becomes a condition on relative detuning, δ_{N_B} , between the two laser beams forming the diffraction grating. For N_B^{th} -order Bragg diffraction, the relative detuning is given by

$$\delta_{N_B} = \frac{2N_B \hbar k^2}{m} = 4N_B \omega_{\text{rec}} \quad (50)$$

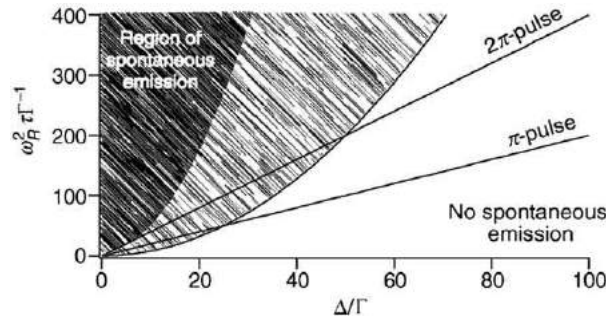


Fig. 10 The conditions for which $N_s > 0.02(0.10)$ are hashed in light(dark) gray. ω_R^2/Γ is proportional to optical intensity. The line labeled π -pulse indicates the parameters which cause complete first-order Bragg diffraction. The line labeled 2π -pulse indicates conditions where population is completely transferred back to the initial state (Pendellösung oscillation).

The first demonstration of Bragg scattering in a BEC was at NIST in 1999. They used Bragg scattering mainly as a tool to manipulate the momentum of the BEC, observing up to sixth-order processes. At MIT the interaction time was lengthened (to ≈ 100 times the lower bound of Eq. (44) for first-order Bragg scattering), creating a new type of spectroscopy called Bragg Spectroscopy. It is a spectroscopic measurement of the shift of the Bragg scattering frequency due to the Doppler shift ($k \cdot v$) from the atoms' motion together with any mean field interaction. (These effects can be separated by going to higher-order Bragg scattering to enhance the Doppler shift.)

Bragg Spectroscopy was used to observe the momentum distribution of a BEC in a magnetic trap. The width of the Bragg resonance curve was primarily due to a 2 kHz Doppler-broadening that yielded the momentum distribution of the condensate. The spread in the corresponding velocity distribution was very small ($\approx 0.5 \text{ mm s}^{-1}$), even smaller than allowed by the Heisenberg uncertainty limit for a particle confined in the ground state of a harmonic trap. This reflects the increase in size of the condensate due to the mean field repulsion (^{23}Na , used in the MIT experiment, has a positive scattering length). The distribution was Heisenberg uncertainty limited at the observed size of the BEC, establishing for the first time that the coherence length of the condensate was equal to its size. In addition, the narrow Bragg resonance was shifted by the repulsive interactions within the condensate, resulting in a spectroscopic measurement of the mean-field energy.

Kapitza–Dirac Scattering

Diffraction of neutral atoms from a standing wave of near-resonant light with a short interaction time ($\tau_{KD} \ll 1/\omega_{\text{rec}}$) has come to be called Kapitza–Dirac scattering, in honor of their pioneering suggestion. In 1933, Kapitza and Dirac predicted that an electron beam propagating in a standing light wave would undergo stimulated Compton scattering and be reflected. This process has a tiny cross-section, equal to the classical electron radius squared: $\sigma_{\text{Compton}} = 8\pi/3(e^2/mc^2)^2 \approx 6 \times 10^{25} \text{ cm}^2$, and has only recently been observed using extremely high laser intensities. If the electrons are replaced by atoms, however, the scattering cross-section for resonant light $\sigma_{\text{atom}} = (1/2\pi)\lambda_{\text{photon}}^2 \approx 4 \times 10^{-10} \text{ cm}^2$ is 15 orders of magnitude larger. This large cross-section, together with the ready availability of tunable lasers, has allowed stimulated scattering, both in the Kapitza–Dirac and Bragg regimes, to become the primary tool for the coherent manipulation of atoms.

In Kapitza–Dirac scattering, atomic motion during the interaction time is small compared to the characteristic dimensions of the interaction potential. This is equivalent to the eikonal approximation for scattering or the thin-lens approximation in optics. The idea is that the phase of the incident particle changes along each classical trajectory, but not the amplitude. Thus, there may be momentum transfer perpendicular to the trajectory, but the trajectory is not significantly displaced (until after the interaction is over). Mathematically, this regime can be treated by neglecting the atomic kinetic energy term in the Hamiltonian (the Raman–Nath approximation).

The standing wave interaction may be treated by considering the standing wave (ac-Stark shift) potential resulting from the applied fields given in Eq. (18):

$$V(z, t) = \frac{\hbar\omega_R^2}{\delta} f^2(t) \sin^2(kz) \quad (51)$$

where we have assumed $\delta \gg \Gamma$, because the constraint $N_s \ll 1$ (Eq. (46)) holds in this regime as well, where N_s is the expected number of spontaneous emission events per atom during the time τ . Although we neglect the kinetic energy term in the Hamiltonian, we continue to will treat z in Eq. (51) as an operator.

Given the initial atomic wavefunction in momentum space, $\psi(p_0)$, the atomic wavefunction immediately after the interaction is given by

$$\psi(p) = e^{-\frac{i}{\hbar} \int dt' U(z, t')} \psi(p_0) \quad (52)$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} e^{i\frac{\omega_R^2}{2\delta}\tau \cos(2kz)} \psi(p_0) \quad (53)$$

where $\tau = \int dt' f^2(t')$ and the integral is over the interaction duration. With the use of the identity for Bessel functions of the first kind, $e^{iz \cos(\beta)} = \sum_{n=-\infty}^{\infty} i^n J_n(\alpha) e^{in\beta}$, the atomic wavefunction can be written as

$$\psi(p) = e^{-i\frac{\omega_R^2}{2\delta}\tau} \sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right) e^{i2nkz} \psi(p_0) \quad (54)$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} \sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right) \psi(p_0 + 2n\hbar k) \quad (55)$$

States with $2N\hbar k$ of momentum are populated with the probability:

$$P_N = J_N^2(\theta), \quad N = 0, \pm 1, \pm 2, \dots \quad (56)$$

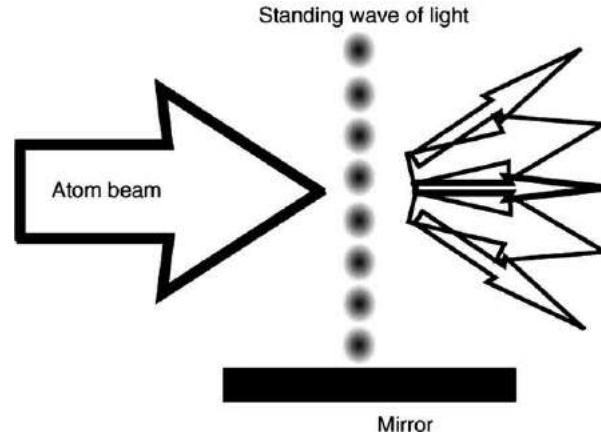


Fig. 11 Two counter-propagating running waves of light superimpose to form a standing wave of light. The more narrow beam shown here represents the configuration for Kapitza-Dirac diffraction.

where

$$\theta = \frac{\omega_R^2}{2\delta} \tau = \omega_R^{(2)} \tau \quad (57)$$

is the pulse area. This leads to a transverse rms momentum of the diffracted atoms that is linearly proportional to the pulse area:

$$p_{\text{rms}} = \sum_{n=-\infty}^{\infty} (n\hbar k)^2 P_n = 2^{1/2} \theta \hbar k \quad (58)$$

The maximum and minimum bounds on the interaction time, τ , and the amount of momentum transfer to the atoms are discussed in the next section.

Kapitza-Dirac diffraction of atoms was first observed at MIT in 1986. Diffraction of a well-collimated (subrecoil) supersonic atomic beam was observed after passage through the tightly focused waist of a near-resonant standing wave (Fig. 11). Significant diffraction into momentum states $|g, \pm 10\hbar k\rangle$ was observed (Fig. 12). Even higher diffracted orders should be observable in the future using laser beams directed at small BECs for somewhat longer times.

Comparison of Atom-Standing Wave Interactions

The atomic motion induced by interaction with a standing light wave varies qualitatively with the interaction parameters. Fig. 13 summarizes the various interaction regimes of an atom with a standing wave in terms of the two most basic parameters: the two-photon Rabi frequency, $\omega_R^{(2)} = \omega_R^2/2\delta$, and the duration of the interaction, τ . To render the plot independent of atomic species, $\omega_R^{(2)}$ is given in units of the single-photon recoil frequency, ω_{rec} , and τ is given in units of the inverse single-photon recoil frequency, ω_{rec}^{-1} . Thus, for ^{23}Na the point (1,1) on Fig. 13 corresponds to (6.4 μs , $2\pi \times 25$ kHz). The product of the two-photon Rabi frequency and the duration of the interaction is the pulse area, $\theta = \omega_R^{(2)} \tau$ (Eq. (57)), which is likewise the product of the pair of coordinates forming a point on the plot.

The lines KD1(KD10) on Fig. 13 show where the first maxima of $P_{1,10} = J_{1,10}^2(\theta)$ (Eq. (56)) occur, corresponding to the maximum possible population transfer into the first and tenth Kapitza-Dirac diffracted order respectively. Since the momentum distribution of diffracted atoms depends only on the pulse area (θ), all Kapitza-Dirac orders are parallel to each other and further offset from the origin in increasing order number.

The Kapitza-Dirac regime ends at large interaction times where the Raman-Nath approximation fails due to motion of the atoms down the slope of the standing wave potential. We show these lines dashed as the interaction time approaches the beginning of the classical oscillation regime, and terminate these lines where the oscillation produces its first focus, at the first focus line. This line corresponds to the atoms completing a quarter period of oscillation in the standing wave potential, $\tau = \tau_{\text{osc}}/4$, where the oscillation time is derived from the approximation that the potential $\langle V(z,t) \rangle_t$ (Eq. (51)) is harmonic about the minimum:

$$\tau_{\text{osc}} = \frac{\pi}{\omega_R} \left(\frac{|\delta|}{\omega_{\text{rec}}} \right)^{1/2} \quad (59)$$

In the shaded region of Fig. 13, the atoms will oscillate classically about the potential minimum, causing a periodic focusing of the atoms alternately in position and momentum space. However, anharmonicities of the potential away from its minimum will degrade the quality of the focusing effects.

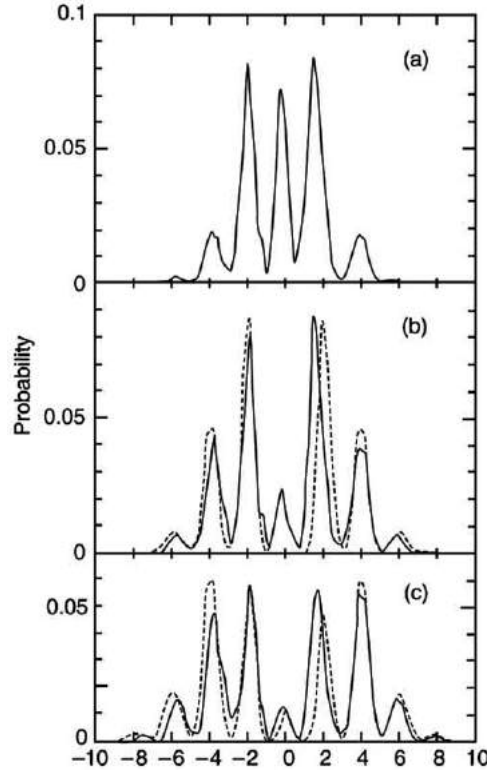


Fig. 12 Kapitza–Dirac diffraction of an atomic beam from a standing wavelength grating. The solid lines are experimental data, the dashed lines are theoretical curves. (a) $\theta = 1.69$, (b) $\theta = 2.33$, (c) $\theta = 2.84$. Reprinted with permission from *Physical Review Letters* 56(8): 827–830. Copyright (1986) by the American Physical Society.

The maximum diffracted order that can be significantly populated by the light–atom interaction is limited by energy conservation. Classically, the maximum momentum transfer to the atoms due to the sudden switch on of the standing wave is delivered to atoms that convert the full height of the standing wave potential ($\hbar\omega_R^2/\delta$), into kinetic energy. This gives

$$\frac{p_{\max}^2}{2m} \simeq \frac{\hbar\omega_R^2}{\delta} \Rightarrow p_{\max} \simeq \omega_R \left(\frac{2\hbar m}{\delta} \right)^{1/2} \quad (60)$$

Equating the maximum absorbed momentum, $p_{\max}$, to an integer number of two-photon recoil momenta yields the maximum expected diffracted order, $N_{\max}$:

$$p_{\max} = 2N_{\max}\hbar k \Rightarrow N_{\max} \simeq \frac{\omega_R}{2} (\omega_{\text{rec}}\delta)^{-1/2} \quad (61)$$

This is just half of the square-root of the standing wave potential height measured in units of the single-photon recoil energy:

$$N_{\max} = \frac{1}{2} \left(\frac{\omega_R^2}{\omega_{\text{rec}}} \right)^{1/2} = \frac{1}{2} \left(\frac{\omega_R^{(2)}}{\omega_{\text{rec}}} \right)^{1/2} \quad (62)$$

As the two-photon Rabi frequency is reduced, $N_{\max}$ falls below unity. As a result, there is no longer time for the higher-order multi-photon processes to generate significant amplitudes in diffracted orders with $N > 1$ before dephasing due to the kinetic energy term (Eq. (24)) becomes significant and the higher-order processes become negligible. Only a small population is ever transferred to the $N = \pm 1$ orders and it does not oscillate at ω_{rec} . Therefore the classical oscillation regime does not extend below $N_{\max} \simeq 1$ where it is terminated on Fig. 13.

Classical oscillation would result in atoms with momentum from zero up to the maximum allowed by energy conservation. Therefore, the classical focusing of atoms must be manifest as an atomic population distribution over many of the quantized momentum states (separated by $2\hbar k$) allowed by energy conservation (Eq. (61)). However, as the interaction time lengthens and extends into the Bragg regime, the populated momentum states become restricted by energy conservation until only the oscillatory Pendellösung into and back from only one final state remains. Therefore, we have ended the classical oscillation regime where the Bragg condition (Eq. (44)) is satisfied. The Bragg regime presupposes a smooth light pulse shape. For a pulse with sharp edges, classically oscillatory behavior can still be observed at longer times than included in the shaded region of Fig. 13, which is why we show the large τ boundary of the classical oscillation regime as dashed.

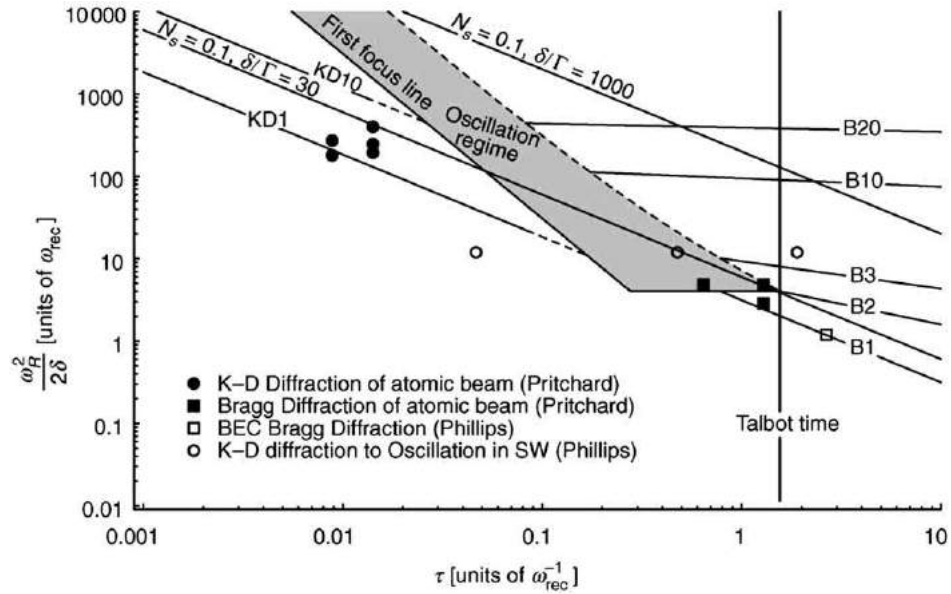


Fig. 13 Atomic diffraction from a standing light wave. The vertical axis is the two-photon Rabi frequency, $\omega_R^2/2\delta$, in units of the single-photon recoil frequency, ω_{rec} , and the horizontal axis is the pulse duration, τ , in units of the inverse single-photon recoil frequency, ω_{rec}^{-1} . The scaling is chosen to eliminate the atomic species dependence of the plot. All coherent momentum transfer processes are destroyed by spontaneous decay, which occurs with probability $N_s=0.1$ along the lines labeled accordingly and parameterized by the given ratio δ/Γ . Lines KD1 and KD10 show conditions for the maximum transfer into the first and tenth Kapitza-Dirac diffracted order respectively. As the interaction time is increased the Raman-Nath approximation is violated (termination of Kapitza-Dirac curves) and the atoms enter the oscillatory regime, executing at least a quarter period of oscillation above the first focus line (shaded area). Cures B1, B2, B3, B10, and B20 correspond to conditions that generate complete Bragg reflection in the first-, second-, third-, tenth-, and twentieth-order respectively. Experimental conditions are shown as points. Filled circles: Kapitza-Dirac diffraction of an atomic beam (Pritchard 1986). Filled squares: Bragg diffraction of an atomic beam (Pritchard 1988). Open squares: Bragg diffraction of a BEC (Phillips 1999). Open circles: transition from Kapitza-Dirac diffraction to oscillation of a BEC in a standing wave light pulse (Phillips 1999). Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

Table 1 Natural parameters and typical experimental parameters involved in standing wave diffraction. The relevant frequencies together with the corresponding times ($1/\omega$) are tabulated. The parameters for ^{23}Na have been used for system dependent quantities

ω	$\omega/2\pi$	$1/\omega$
ω_0	500 THz	0.3 fs
δ (Bragg)	500 MHz	0.3 ns
ω_R (Bragg)	100 MHz	1.6 ns
ω_R (K-D)	100 MHz	1.6 ns
Γ	10 MHz	16 ns
ω_{rec}	25 kHz	6.4 μs
ω_{osc}	1 MHz	160 ns

In the Bragg regime, transfer of population is restricted to (and back from) only one final momentum state. The allowable final states are restricted by limiting the frequency bandwidth (i.e., energy uncertainty) of the light fields in the atomic rest frame. This is accomplished by lengthening the interaction time and smoothing the rise and fall of the electromagnetic fields. The parameters for a first-order Bragg transition (Table 1) are typically $\delta/\Gamma \simeq 50$, $\Gamma\tau \simeq 1$, and $\omega_R/\Gamma \simeq 10$ giving $N_s \simeq 10^{-2}$ and $\omega_R^{(2)}\tau \simeq 1$. Obtaining significant population transfer with higher-order Bragg processes requires larger intensities. Various orders (1, 2, 3, 10, and 20) of Bragg diffraction are shown as lines on Fig. 13 corresponding to $\omega_R^{(2N_B)}\tau = \pi$, where $\omega_R^{(2N_B)}$ is given in Eq. (40). The lines are terminated at the appropriate interaction time determined from the final momentum state resolution condition (Eq. (44)). The Bragg regime extends indefinitely to larger interaction times, which might be termed the region of Bragg spectroscopy. In experiments in this regime with ^{23}Na , atomic velocity resolution below 1 mm s^{-1} was obtained at $\tau = 80\omega_{rec}^{-1}$. The study of adiabatically expanded BECs would require larger interaction times to resolve their smaller velocity spread and weaker mean field shifts.

The Kapitza–Dirac and Bragg regimes assume a different initial atomic momentum (in the rest frame of the standing wave) parallel to the standing wave axis. Kapitza–Dirac scattering assumes no component of the initial atomic momentum along this axis. The efficiency of Kapitza–Dirac diffraction falls rapidly as the initial momentum in units of the photon momentum approaches $1/\omega_{\text{rec}}\tau$. To observe N_B^{th} -order Bragg scattering, the initial atomic momentum along the standing wave axis must have (nonzero) magnitude $N_B\hbar k$. Without adhering to this constraint, no final momentum state will be energetically degenerate with the initial state and the atomic sample will not respond to the presence of the standing wave, even if the interaction parameters are appropriate ($\omega_R^{(2N_B)}\tau \simeq 1$) for N_B^{th} -order Bragg scattering.

The condition $N_s = 0.1$ (Eq. (46)), where N_s is the expected number of spontaneous emission events per atom during the light–atom interaction, is drawn on Fig. 13 for two ratios (30 and 1000) of standing wave detuning, δ , to the atomic excited state natural lifetime, Γ . For fixed $\delta\Gamma$, we see that $N_s \propto \theta$, thus conditions with an increased pulse area (above and to the right of the $N_s = 0.1$ lines) result in a proportionally increased N_s . These $N_s = 0.1$ lines are extended across all regimes of Fig. 13 because restricting spontaneous emission is required for both Kapitza–Dirac and Bragg scattering. As N_s approaches unity, the correct optical potential describing the light–atom interaction no longer is Eq. (51), and scattering of atoms becomes an incoherent process which can lead to new possibilities such as a complex valued optical potential. Thus, while the standing wave fields in pure Bragg or Kapitza–Dirac scattering behaved as phase gratings for atoms, light gratings on resonance can effectively become amplitude diffraction gratings.

Atom Interferometry

In the 19th century, Fizeau (1853), Michelson (1881), Rayleigh (1881), and Fabry and Perot (Fabry 1899) exploited the interference properties of light waves to create the light interferometer which has since resulted in many beautiful experiments and precise measurements. Using technologies invented since the Second World War, the initial ideas of de Broglie and Schrödinger have evolved into construction of interferometers for neutrons, electrons, and atoms. These new interferometers are proving to be valuable tools for probing fundamental physics, for studying quantum mechanical phenomena, and for making inertial measurements.

The scientific value of interferometry with atoms and molecules has long been recognized. In fact, the concept of an atom interferometer was patented in 1973 and it has been extensively discussed since. Atom interferometry offers great richness stemming from the varied internal structure of atoms, the wide range of properties possessed by different atoms (e.g., mass, magnetic moment, absorption frequencies, and polarizability), and the great variety of interactions between atoms and their environment (e.g., static E–M fields, radiation, and other atoms).

Light interferometers are generally based either on achromatic beamsplitters such as half-silvered mirrors or on other semi-transparent membranes whose structure is small compared to the wavelength of the wave they are splitting. Lacking material structures that are either transparent to atoms or smaller than their de Broglie wavelength, diffraction gratings have been pressed into service both as beamsplitters and mirrors for atom waves. This means that atom interferometers are constrained to designs which somehow compensate for the dependence of diffraction angle on the wavelength of the individual atoms. In spite of this challenge, a surprising variety of atom and molecule interferometers have been built since 1991. A majority have used the three-grating configuration in which the first grating splits the incident beam, the second reverses the differential momenta given by the first, and the third recombines the two beams at the location where they overlap. Both material gratings and standing waves of light have been used in the Raman–Nath, Bragg, and adiabatic regimes to obtain interference fringes. Some designs render the interference fringes in position space, others in internal state space.

Fig. 14 shows the MIT setup for a three-grating atom interferometer which features sufficient separation between interfering paths to accommodate an interposed metal foil. This allows different fields or media to act on atom waves only in one of the two arms of the interferometer. Such an interaction region has been used to measure phase shifts due to an electric field applied to only one of the interfering paths. The phase difference between the two paths is then given by the Feynman path integral:

$$\phi = \frac{-i}{\hbar} \int V(x) dt \quad (63)$$

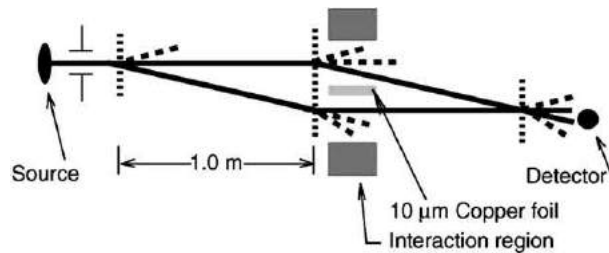


Fig. 14 An interferometer for atoms built with three nanofabricated gratings. Each atom propagates in a superposition of two paths, shown in bold. The beam of atoms is well enough collimated, and the diffraction angle is large enough, that a metal barrier can be inserted between the two paths inside the interferometer. The interference fringes are observed by translating one grating transverse to the atom beam. See for example: Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

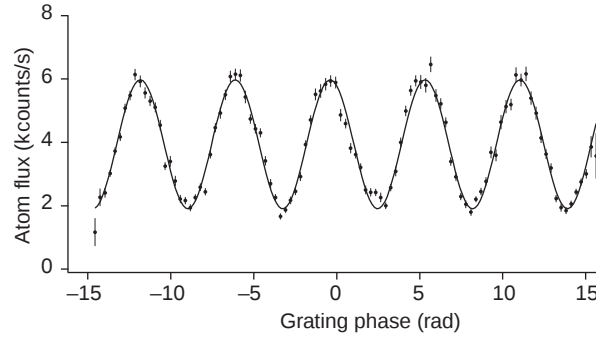


Fig. 15 Interference fringes from atoms. 2π of grating phase corresponds to translating the grating one grating period, or 100 nm. These data were obtained at MIT.

where

$$V = -\frac{1}{2}\alpha E^2 \quad (64)$$

is the potential energy due to the electric field E , and α is the electric polarizability of the atoms. The resulting phase shift of the interference fringes (Fig. 15) has been used to measure the polarizability of sodium atoms to unprecedented accuracy.

Phase shifts have also been measured due a dilute target gas in one arm of the interferometer. In this case there is a complex index of refraction for the forward scattered atoms (some attenuation, and some phase shift) which depends on matter wave wavelength. Recent measurements reveal glory oscillations in the matter wave index of refraction due to passing through a dilute gas, and are a sensitive probe of interatomic potentials.

In the domain of inertial sensing, atom interferometers have been used to the local gravitational acceleration, $\vec{g}$, Newton's constant, G , gravity gradients, $\nabla\vec{g}$, and rotations $\vec{\Omega}$, each with sensitivity rivaling if not exceeding any other method. The Sagnac phase shift, due to rotations, for a two-path interferometer is

$$\phi_S = \frac{4\pi}{\lambda_{dB}v} \vec{\Omega} \cdot \vec{A} \quad (65)$$

where v is the velocity of atoms, $\vec{A}$ is the enclosed area of the interferometer, and $\vec{\Omega}$ is the rotation rate. The two factors in the denominator, λ_{dB} and v , are both $\sim 10^5$ times larger for light than for atoms, which is why atom interferometers have the potential to be 10^{11} times more accurate than optical interferometers. In practice, atom interferometers only outperform laser gyros by a small amount because they have dramatically smaller enclosed areas and particle flux than optical interferometers.

Since the first atom interferometers for Na and He* began working in 1991, others have been made for Li, Ar*, Ca, Cs, K, Mg, Ne*, Rb, and Li₂, Na₂, I₂, C₆₀ and C₇₀ molecules, interferometers starting with trapped atoms have been made for Cs, Ca, He*, Mg, and Rb. Interferometers using Bose-Einstein condensates have been demonstrated with Na and Rb. These lists are still growing. Diffraction from a single grating has been observed for molecules as large as C₇₀ buckyballs, which may someday be used in a three-grating interferometer.

Conclusion

The manipulation of atoms using light forces from standing light waves is a rich subject. The seminal suggestion of Kapitza and Dirac laid dormant for 50 years due to lack of experimental technology. However, in the 1980s this suggestion was realized with atomic sources and considerably extended both experimentally and theoretically. In the early 1990s, coherent standing wave manipulation and nonfabricated atom optics became the two major routes to making atom interferometers. As the new century begins, these techniques are being refined further with improved sources, new detection schemes, and better atom optics. Replacing thermal atomic beams by Bose-Einstein condensates (BECs), as sources of atoms, will revolutionize the field of atom optics just as lasers did in light optics.

Conversely, stimulated light scattering and nanostructures provide new ways to study atomic and molecular properties, such as Bragg spectroscopy for BEC characterization. As progress continues, exciting developments should be forthcoming, such as definitive measurements of the fine structure constant, α , atom gyroscopes that are far superior to the best laser gyroscope technology, and improved studies of BECs. Another area that is on the brink of spectacular development is the confinement of coherent matter waves in atom waveguides and the development of the scientific and technological opportunities that these represent, in analogy to fiber optic waveguides for light.

While we have described the application of nanofabrication techniques to atom optics, it is possible to imagine technology transfer in the other direction. The fundamental problem of fabricating ever smaller structures might be tremendously advanced by

atom optics because the de Broglie wavelength of atoms is so much smaller than that of light, permitting them to be focused to directly deposit much smaller features than possible with photolithography.

See also: Coherent Lightwave Systems

Further Reading

- Cohen-Tannoudji, C., Diu, B., Laloe, F., 1977. Quantum Mechanics. New York: Wiley and Sons.
Hecht, E., 1987. Optics. Reading: Addison-Wesley.
Berman, P.R. (Ed.), 1997. Atom Interferometry. San Diego, CA: Academic Press.
Mandel, L., Wolf, E.M., 1995. Optical Coherence and Quantum Optics. London: Cambridge University Press.
Balykin, V.I., Letokhov, V.S., 1995. Atom Optics with Laser Light. Chur, Switzerland: Harwood Academic Publishers.
Meystre, P., Sargent III, M., 1999. Elements of Quantum Optics. Heidelberg, Germany: Springer Verlag.
Rauch, H., 2000. Neutron Interferometry. Oxford, UK: Oxford University Press.
Meystre, P., 2001. Atom Optics. New York: Springer-Verlag.
Campargue, R. (Ed.), 2001. Atomic and Molecular Beams, the State of the Art 2000. Berlin: Springer.

Organic Semiconductors

Fang-Chung Chen, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic semiconductors (OSCs) are receiving increasing attention these days because they have many attractive properties – including light weight, low-cost production, low-temperature processing, mechanical flexibility, and abundant availability – that distinguish them from their conventional inorganic counterparts. The recent successful commercialization of active-matrix organic light-emitting diodes (OLEDs) for high-end smartphone displays suggests that organic compounds have promising prospects for use in various other electrical and optoelectronic devices. For example, the fabrication of solar cells, transistors, photodetectors, and lasers with OSCs can be highly feasible and highly efficient. Another attractive property of OSCs is that they can be processed using simple solution processing techniques (e.g., ink-jet printing, reel-to-reel fabrication), making the fabrication of electronic devices much easier and cheaper. As Professor Heeger said in his Nobel lecture, “I’m convinced that we are on the verge of a revolution in ‘Plastic Electronics’.” We foresee rapid advances in the field of OSCs bringing increasingly more inexpensive, workable, and smart organic devices into our daily lives.

Organic materials are substances comprising mostly carbon and hydrogen atoms; they had been considered for many years to be exclusively electrically insulating. The discovery of electrical conductivity in organic compounds can be traced back to the 1950s; it has opened up a new fascinating field of research. Electroluminescence (EL) from anthracene single crystals was reported by Pope *et al.* in 1963. In the 1970s, Shirakawa, Heeger, and MacDiarmid found that the conductivity of polyacetylene could be increased tremendously – up to a level close to that of a typical metal – after exposure to vapors of chlorine, bromine, or iodine. Such high conductivity from a traditional “insulator” of electricity drew much attention and encouraged the rapid growth of the field of organic electronics. The Nobel Prize in Chemistry in 2000 was awarded to Heeger, MacDiarmid, and Shirakawa for “the discovery and development of conducting polymers.”

Among the various organic electronic devices known today, the technology of OLEDs has definitely been under the spotlight. Initially, the EL observed from organic single crystals was not energy efficient and required a large voltage to drive the devices. In 1987, Tang and Van Slyke developed light-emitting devices based on amorphous organic thin films and achieved a quantum efficiency of 1% under an electrical bias of less than 10 V. In 1990, OLEDs incorporating conjugated polymers were also reported by Professor Friend at Cavendish Laboratory. Since then, relevant technologies – including new materials, device structures, and device engineering approaches – evolved rapidly, gradually establishing a new industry.

As organic-based devices have matured from scientific concepts in the laboratory to applications in the real market, there is a requirement to understand the unique properties of OSCs and determine the opportunities that this new generation of electronic materials might offer. In this article, we discuss the electronic structures of OSCs and several fundamental charge storage scenarios in OSCs. We also review the concept of doping in OSCs. We conclude with a brief overview of interesting device applications, including electrophotographic devices, thin film transistors, and light-emitting transistors.

Electronic Structures

OSCs are usually categorized into two main groups, depending on the molecular weights of their organic components: conjugated polymers (Fig. 1) and small molecules (Fig. 2). Polymers are constituted by repeated structural units (monomers) connected by covalent bonds. Because the number of repeating units usually cannot be controlled precisely, they are prepared with a distribution of molecular weights. On the other hand, small molecules have precise and much smaller molecular weights. The other distinction between the two OSC categories is in their methods of fabrication. Most conjugated polymers are soluble in organic solvents and can be processed using various solution-processing approaches. In contrast, small molecules are usually deposited through thermal evaporation under high vacuum – a deposition method that allows good control over the thickness, morphology, and other properties of the organic films. Accordingly, very thin (even down to a few nanometers), high-quality films of small molecules can be obtained quite readily. The chemical structures of the small molecules can be also tailored so that they become soluble in organic solvents, facilitating the use of solution fabrication processes. For example, pentacene (Fig. 2(d)) can be deposited only through thermal evaporation, whereas 6,13-bis[(triisopropylsilyl)ethynyl]pentacene (TIPS-pentacene) (Fig. 2(f)) features side chains that improve its solubility, simplify its processing while maintaining high crystallinity.

To understand the electronic structures of OSCs, we begin with the atomic orbitals (AOs) of carbon, the most important element of any organic materials. A carbon atom has six electrons; its electronic configuration is $(1s)^2(2s)^2(2p)^2$. It can be bonded to other atoms using various hybridization orbitals, including sp , sp^2 , and sp^3 hybrid orbitals (Fig. 3). For example, one s and two p orbitals are superposed to form three sp^2 hybrid orbitals; the orbitals are distributed in the same plane with an angle of 120 degree between them. In a simple model molecule, ethylene, each carbon atom offers three sp^2 hybrid orbitals to bond with other atoms (Fig. 4). The $1s$ orbital of a hydrogen atom overlaps with one of the sp^2 hybrid orbitals to form a molecular orbital (MO),

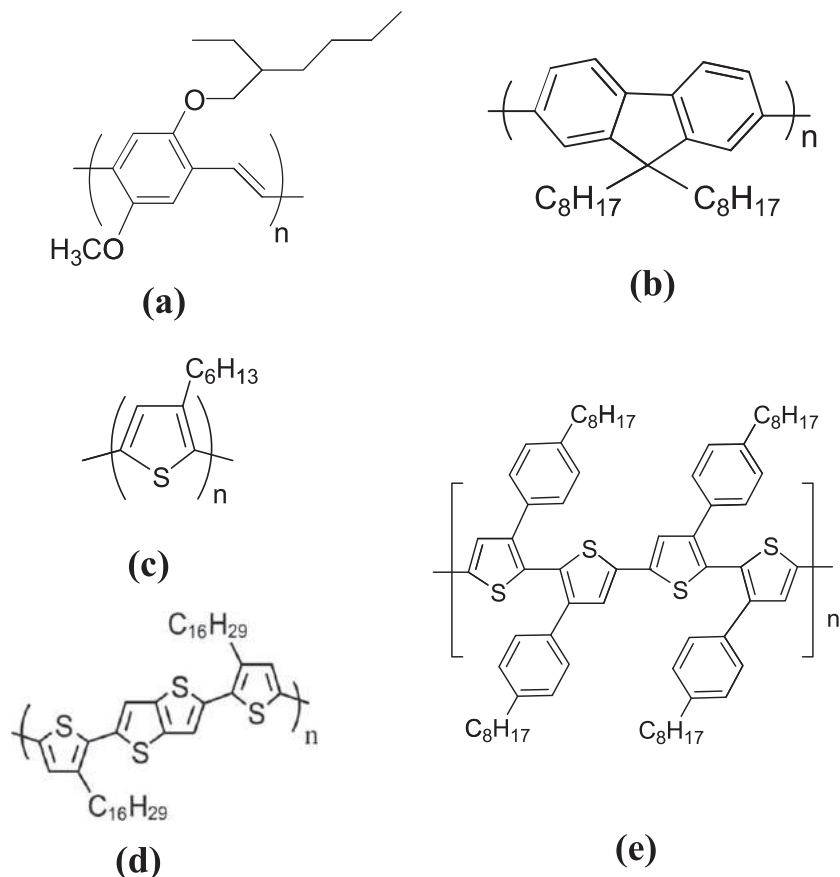


Fig. 1 Chemical structures of polymer-based OSCs: (a) poly[1-methoxy-4-(2-ethylhexyloxy)-*p*-phenylene vinylene] (MEH-PPV), (b) poly(9,9-dioctylfluorene) (PFO), (c) poly(3-hexylthiophene-2,5-diyl) (P3HT), (d) poly[2,5-bis(3-hexadecylthiophen-2-yl)thieno[3,2-*b*]thiophene] (PBTBT), and (e) poly[3-(4-octylphenyl)thiophene] (POPT).

a so-called σ -orbital; the resulting bonding is a σ -bond. The two carbon atoms are also bonded through sp^2 hybrid orbitals. Each carbon atom, however, retains one p atomic orbital, aligned orthogonal to the plane of the sp^2 hybrid orbitals. The two p_z atomic orbitals overlap side-by-side with the electron density distributed above and below the internuclear axis, forming a different type of MO called a π -orbital. Note that the π -electron cloud is delocalized between the two carbon atoms. Consequently, the carbon atoms are bonded by two bonds: a σ -bond and a π -bond; this arrangement is termed a “double bond.” In view of the electron wave functions, the overlapping of two AOs produces two MOs: a bonding state and an anti-bonding state, after constructive and destructive interference, respectively. Because destructive interference results in lower electron density between the atoms, the energy of the anti-bonding state is higher than that of the bonding state. For π -orbitals, the bonding and anti-bonding states are denoted as π - and π^* -MOs, respectively.

Benzene, a larger molecule (**Fig. 5**), has six carbon atoms that are also bonded using sp^2 hybrid orbitals. Similarly, the six p_z AOs that remain form six π -bonding MOs: three bonding and three anti-bonding orbitals. As a consequence of interference among the six p_z orbitals, the lower bonding and upper anti-bonding states are degenerate (**Fig. 5(b)**). Six p_z electrons occupy the three bonding states in a neutral benzene molecule. Similar to ethylene, the π -electron cloud is delocalized over the six carbon atoms. In other words, the electrons can move from one carbon atom to another; such a delocalized structure is said to feature “conjugation.”

The MO of highest energy that electrons occupy is known as the highest occupied molecular orbital (HOMO). Meanwhile, the lowest unoccupied molecular orbital (LUMO) is the MO of lowest energy that electrons do not occupy. The difference in energy between the HOMO and LUMO is the energy required to excite an electron in the molecule. In benzene, for example, **Fig. 5(b)**, the upper bonding MOs (π_2, π_3) are the HOMOs and the lower anti-bonding MOs (π_4^*, π_5^*) are the LUMOs. For OSCs, because the energy splitting of π -bonds is usually smaller than that of σ -bonds, electronic processes (e.g., photon absorption and emission) occur energetically favorably on π -orbitals. Similarly, the charges injected from the metal contacts of organic electronic devices would tend to occupy π -MOs. In other words, the electronic functions of OSCs originate from the smaller and favorable energy levels of the frontier orbitals of the π -electron system; the σ -electrons are hardly excited because much higher energy is required for electronic transitions between σ -orbitals. Furthermore, the HOMO–LUMO energy gap decreases upon increasing the size of the π -conjugation system, because of the greater extent of π -electron delocalization.

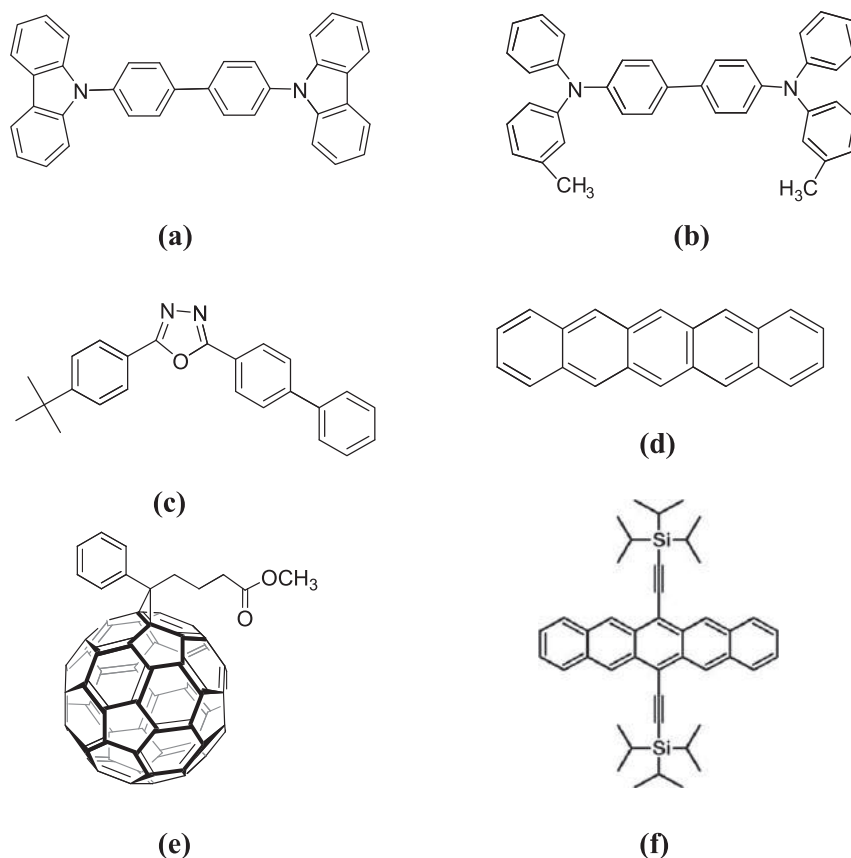


Fig. 2 Chemical structures of small-molecule semiconductors: (a) 4,4'-*N,N'*-dicarbazolyl-biphenyl (CBP), (b) *N,N'*-diphenyl-*N,N'*-bis(3-methylphenyl)-1,1'-biphenyl-4,4'-diamine (TPD), (c) 5-(4-biphenyl)-2-(4-*tert*-butylphenyl)-1,3,4-oxadiazole (t-PBD), (d) pentacene, (e) [6,6]-phenyl- C_{61} -butyric acid methyl ester (PCBM), and (f) 6,13-*bis*[(triisopropylsilyl)ethynyl]pentacene (TIPS-pentacene) soluble pentacene).

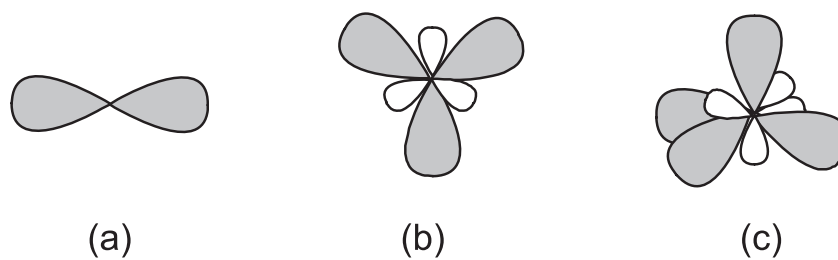


Fig. 3 Schematic representation of (a) sp , (b) sp^2 , and (c) sp^3 hybrid orbitals.

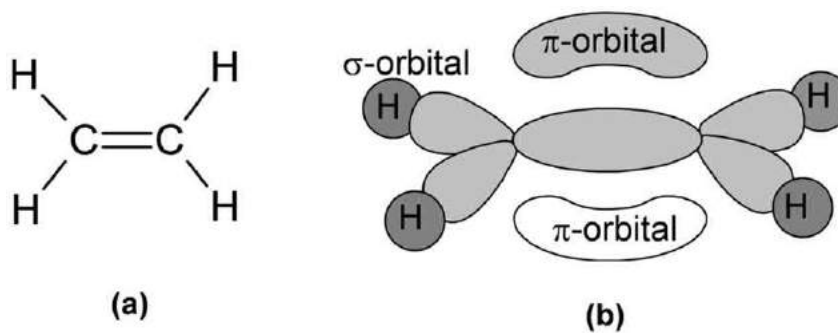


Fig. 4 (a) Chemical structure of ethylene and (b) schematic orbital diagram.

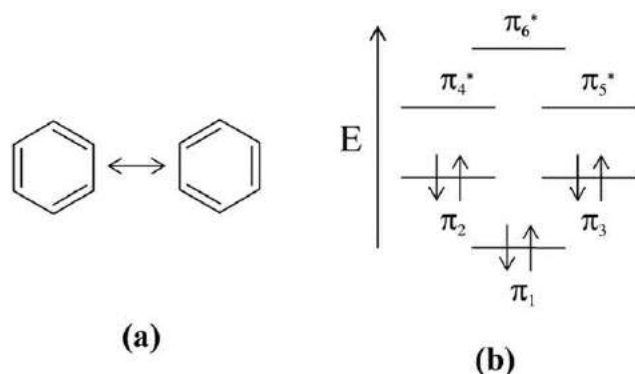


Fig. 5 (a) Chemical structure of benzene, it consists of two resonance forms and (b) energy diagram of the π -MOs in benzene.

Trans-polyacetylene was the first conducting polymer to be discovered (Fig. 6). For a single chain of neutral polyacetylene, the π -band, which can be considered to be composed of an almost infinite number of p_z orbitals, would be half-filled, resulting theoretically in metallic behavior (Fig. 6(a)). Experimental results have strongly indicated, however, that polyacetylene is a semiconductor having an energy gap of approximately 1.5 eV. In reality, the one-dimensional (1D) “metal” is unstable and the structure tends to distort, as displayed in Fig. 6(b). Accordingly, the main chain of polyacetylene actually has alternately long and short bonds. The so-called Peierls distortion converts the repeating unit of *trans*-polyacetylene to *trans*- $(-\text{HC}=\text{CH}-)_n$, instead of *trans*- $(\text{CH})_n$. As displayed in Fig. 6(g), dimerization of the molecule opens an energy band gap. The π -band is fully occupied and the upper π^* -band is empty; this band structure is similar to that of a typical inorganic semiconductor, which features a full-filled valence band and an empty conduction band. While a 1D π -electron metal is not stable, higher-dimensional π -electron systems are not subject to such Peierls instability. For example, carbon nanotubes and graphene can be metallic or semi-metallic.

Solitons, Polarons, and Bipolarons

The electronic structure of a polymer can be described using the Su-Schrieffer-Heeger (SSH) model, based on a quasi-one-dimensional tight-binding model. It assumes that the π -electrons are strongly coupled to the polymer backbone through the electron-phonon interactions. Charge storage occurs through formation of solitons, polarons, and bipolarons. Solitons are structural defects on the polymer chains. As illustrated in Fig. 6(b) and (c), polyacetylene has two resonance forms. The soliton is a domain boundary of these two degenerate resonance configurations (Fig. 6(d)); the domain wall can actually extend over several carbon atoms (Fig. 6(e)). A soliton may be neutral or carry a charge. The energy of the soliton lies between the π - π^* gap, because of its nonbonding nature. This defect state is mobile because the system energy does not depend on the position of the soliton. For non-degenerate ground-state polymers – for example, poly(phenylene vinylene) (PPV, Fig. 7(a)) and polyparaphenylene (PPP, Fig. 7(b)) – in which the interchange of single and double bonds leads to changes in energy, the solitons cannot be stable because the structure is not symmetric. Instead, double defects are required for such non-degenerate polymers. Such double defects are called polarons; they can be considered as a bound state of a charged soliton and a neutral soliton (Fig. 7(c)). Consequently, its energy levels hybridize the two states, which are split into a bonding, less energetic level and an anti-bonding, more energetic one (Fig. 7(d)). The two energy levels can be occupied by zero, one, or two electrons, resulting in the formation of positive or negative polarons. A bipolaron is a bound state of two charged solitons of equal charge; it also consists of a split pair of energy levels between the energy gap. Furthermore, solitons, polarons, and bipolarons are coupled with structural relaxation, implying strong electron-phonon interactions of the charge storage in conjugation polymers. In other words, the charges are surrounded by lattice distortions that can stabilize the system.

Excitons

Excitons are states of electronic excitations in OSCs. An exciton is a particle composed of an electron/hole pair bounded by coulombic force. The total charge of an exciton is zero. An exciton is, however, capable to move through the organic materials carrying the excitation energy. There are three types of excitons (Fig. 8): Frenkel, Wannier-Mott, and charge-transfer (CT) excitons. The main difference among them is the binding energy and/or the distance between the electron/hole pair. When the radius of the electron/hole pair is comparable with the size of a single molecule or the crystalline constant (a_L), the exciton is of Frenkel type. The small radius also implies a strong coulombic interaction. In other words, it is very difficult to overcome the binding energy and dissociate the exciton into free charges. If the radius of the exciton is much larger than a_L , it is a Wannier-Mott exciton. Therefore, the binding energy of this type of exciton is small. The intermediate case, in which the exciton extends over several molecular units, it is termed a CT exciton. The Wannier-Mott exciton is the most common one found in inorganic semiconductors. Usually it can

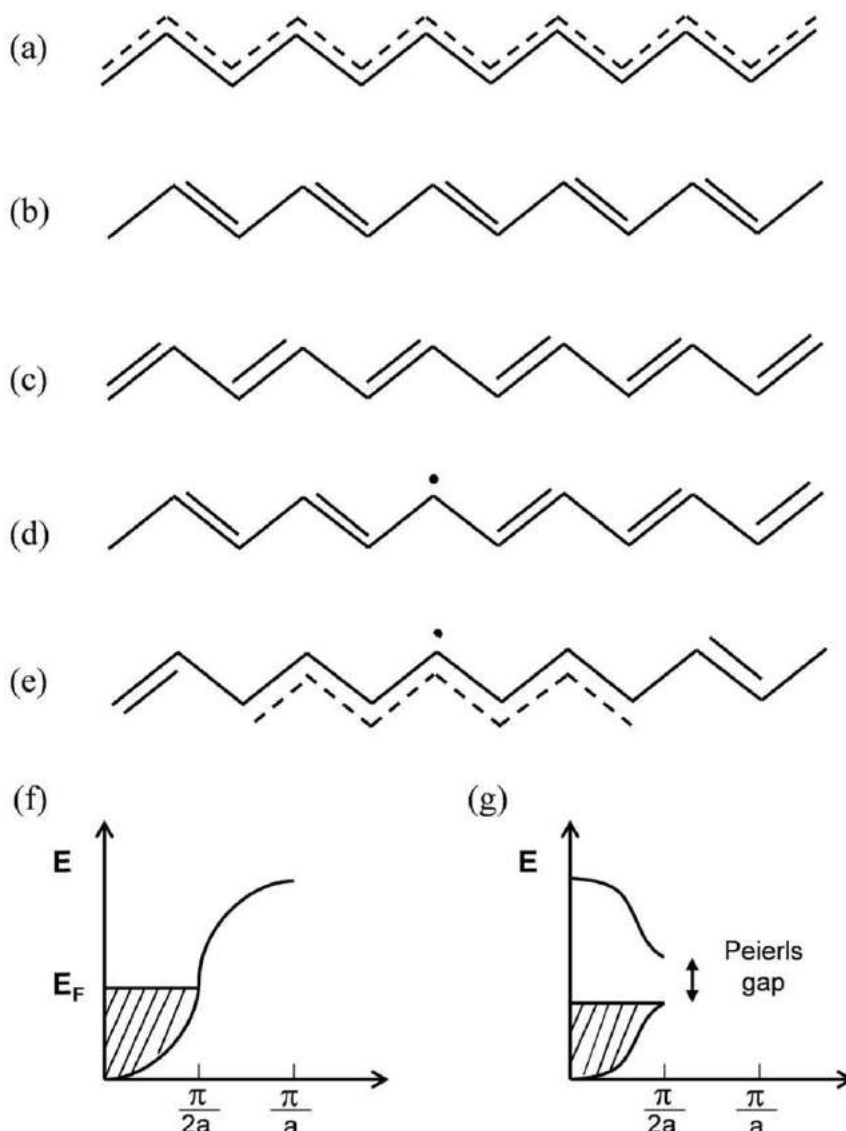


Fig. 6 (a–e) Polyacetylene: (a) undimerized structure; (b, c) two resonance forms resulting from Peierls effects; (d) soliton in *trans*-polyacetylene; (e) domain wall extension of the soliton. (f, g) Electronic band dispersion $E(k)$ plotted with respect to k for (f) non-dimerized and (g) dimerized systems. E_F is the Fermi energy; a gap appears in the dimerized system, the so-called “Peierls gap.”

be detected only at low temperature, because the thermal energy at room temperature is sufficiently high to overcome its binding energy. In OSCs, most excitons are Frenkel-type because of their strong binding energy, which originates from the low dielectric constants of organic materials and their relatively localized electronic clouds. A typical binding energy for excitons in OSCs is approximately 0.5 eV.

The generation of excitons is a very important process in organic optoelectronic devices. In OLEDs, the recombination of electrons and holes results in excitons that subsequently decay to produce EL. In organic photovoltaic devices (e.g., organic solar cells), photon absorption generates the excitons, which move freely within the semiconductors, due to the strong binding energy, but are subject to effective charge dissociation at the interface between strong organic electron-donors and organic electron-acceptors. Such exciton dissociation generates free charges, eventually leading to a photocurrent after the charges have been collected by the electrodes.

Because of the strong binding energy, the directions of spin of the excitons are not likely to change readily when generated in common OSCs. Therefore, the excitons can be categorized into two different types depending on their combinations of spin directions. As displayed in Fig. 9(a), the total spin angular momentum of a molecule in the ground state is a sum of the two spin vectors, equal to zero because of the opposite spin directions. In other words, the overall spin quantum number (S) is zero

$$S = |s_1 + s_2| = |(+1/2) + (-1/2)| = 0 \quad (1)$$

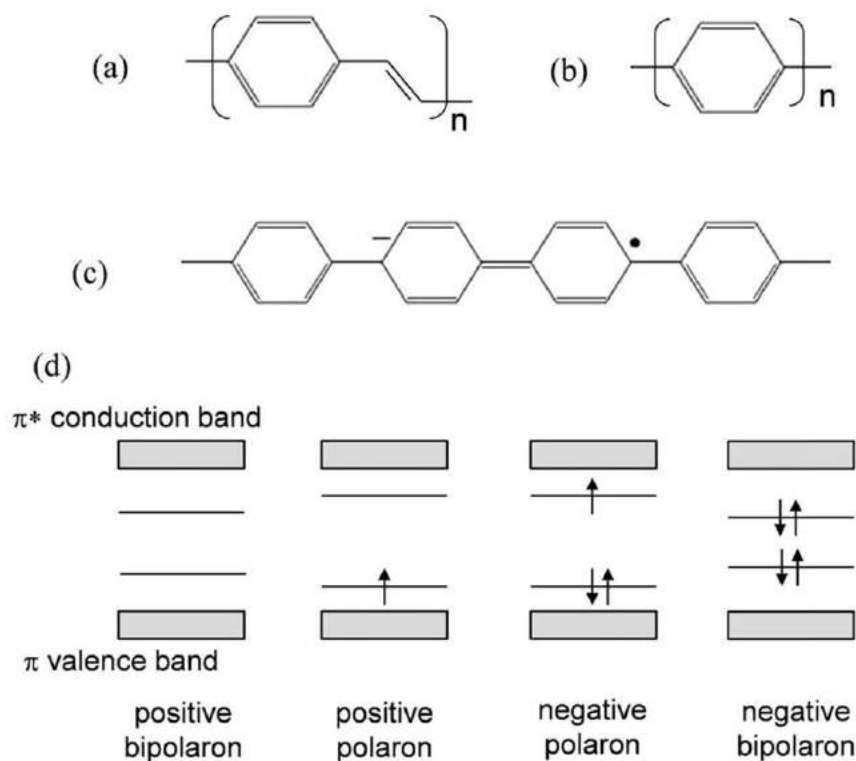


Fig. 7 Chemical structure of (a) poly(phenylene vinylene) (PPV) and (b) polyparaphenylene (PPP). (c) Schematic representation of a negative polaron in PPP. (d) Band diagram of various polarons.

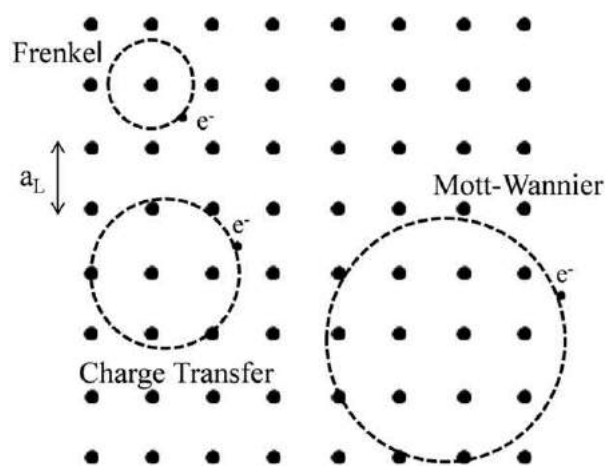


Fig. 8 Schematic representation of three different types of excitons in solid state materials; a_L is the lattice constant.

where s_1 and s_2 are the spin quantum numbers of the individual electron. If one electron is excited to the π -orbital, the resulting exciton is called a singlet because no net spin angular momentum exists in the molecule (Fig. 9(b)). On the other hand, if the two spin vectors have the same direction, the total spin angular momentum is now equal to 1, as revealed in the following equation

$$S = |s_1 + s_2| = |(+1/2) + (+1/2)| = 1 \quad (2)$$

Note that if s_1 and s_2 were both equal to $-1/2$, the sum would remain the same. Accordingly, the excited state (exciton) now exhibits a net angular momentum. Assuming the presence of a magnetic field, the energy of the exciton could be split into three different energy levels. Therefore, it is called a triplet (Fig. 9(c)).

Fig. 9(d) displays possible decay processes of a typical organic molecule, where S_0 , S_1 , and T_1 represent the ground state, first singlet state, and first triplet state, respectively; k_{nr} and k'_{nr} are the rate constants of non-radiative decay processes and k_r and k'_r are the rate constants of radiative decay processes. The singlet exciton can flip its spin direction and transform into a triplet exciton

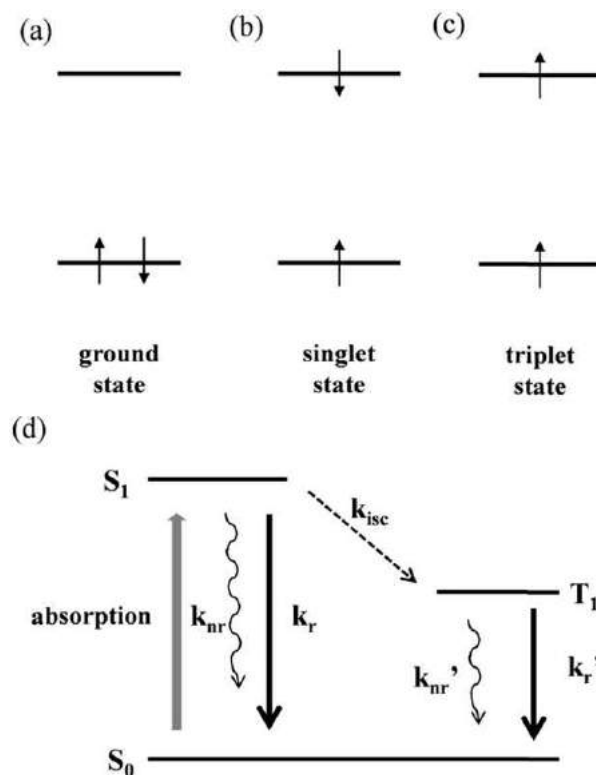


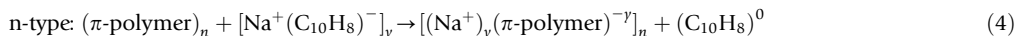
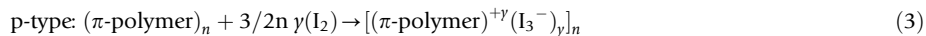
Fig. 9 Schematic representation of electron spins in the (a) ground state, (b) singlet exciton, and (c) triplet exciton of an organic molecule. (d) Excitation and possible decay processes in an OSC.

through a process called inter-system crossing (ISC); its rate constant is k_{isc} . As indicated in Fig. 9(d), the relaxation of a singlet exciton (S_1) to the ground state (S_0) is possible because the spin angular momentum is conserved. In quantum mechanics, such a transition is allowed; its radiative decay may lead to a highly efficient emission, called fluorescence (k_r). On the other hand, the transition of a triplet exciton to its ground state is forbidden. Consequently, the efficiency of the radiative decay is very low and, for most organic molecules, can be observed only at a very low temperature. The emission of triplet excitons is called phosphorescence (k_r').

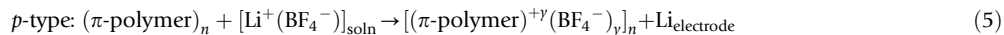
Concept of Doping and p- and n-Type OSCs

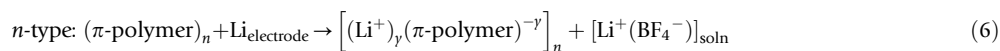
Like inorganic semiconductors, OSCs can also be classified as p- and n-type semiconductors. The doping concept and process are, however, substantially different from those of inorganic materials, even though the same terminology of “doping” is used. For example, the inorganic semiconductor silicon is doped through ion implantation, in which impurities (phosphorus (P) or boron (B) atoms) are introduced. These foreign atoms substitute some of the silicon atoms and lead to new energetic levels in the band gap. Silicon becomes an n- or p-type semiconductor after the introduction of P or B atoms, respectively.

On the other hand, many different methods can be used to “dope” OSCs, including chemical doping, electrochemical doping, photo-doping, and charge injection from metal contacts. Chemical doping involves CT redox chemistry, as illustrated in the following examples for polymers

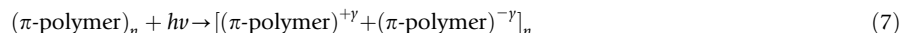


From a chemical point of view, p- and n-type doping is analogous to electrochemical oxidation and reduction, respectively. The doping process generates charge carriers in the OSCs. The electrical conductivity increases upon increasing the doping level and can reach a level comparable with that of a metal at a very high doping concentration. The drawback of this chemical method is that the doping level is hard to control, especially for intermediate doping levels. Fortunately, electrochemical doping can solve this problem. An electrode supplies redox charges to the semiconductor, and ions simultaneously diffuse in or out of the polymer chains from the electrolyte to compensate the charge mismatch. Through control of the voltage of the electrode, the doping level can be determined precisely. Electrochemical doping can be illustrated using the following equations:



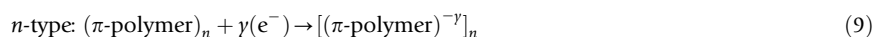
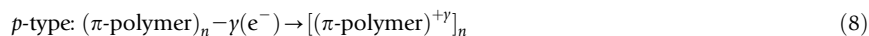


The third method, photo-doping, is a process in which photon absorption generates an electron/hole pair, which could lead to locally oxidized (p-type doping) and nearby reduced (n-type doping) states of the material. The unique mechanism is illustrated by the following equation:



After photon absorption, recombination to the ground state can be either radiative or nonradiative. The value of γ in the equation depends on the excitation rate in competition with the recombination rate.

The last scenario involves charge injection from metallic contacts into the OSCs. Holes and electrons are introduced into the π - and π^* -bands, respectively, as illustrated by the following equations:



Eq. (8) describes the injection of a hole into the filled π -band, corresponding to an oxidation reaction (removal of an electron). Eq. (9) describes the injection of an electron into an empty π^* -band, analogous to a reduction reaction. The major distinction of doping through charge injection from other doping processes (e.g., chemical doping) is the absence of counter ions to compensate the charges. Therefore, the doping process is only temporary; the states are relatively unstable. On the other hand, for chemical or electrochemical doping, the doping is permanent, unless the carriers are removed by an “undoping” process. As for photo-doping, the resulting photoconductivity is transient and limited by recombination and decay processes.

Even though OSCs can indeed be “doped” as described above and become p- or n-type semiconductors, many reports define OSCs as p- or n-type by judging whether they are good conductors for electrons or holes. For instance, pentacene is usually referred to as an excellent p-type semiconductor (Fig. 2(d)) because it exhibits a very high hole mobility. TPD (Fig. 2(d)) is often employed as a hole-transporting material in OLEDs. As a result, TPD is also considered a p-type OSC because it transports holes effectively and because holes are injected readily from the metal electrodes. Note that these OSCs are not intentionally doped and they are usually purified to as high a degree as possible to remove impurities before they are processed for device fabrication. Therefore, we cannot assume extra energy levels appear near the HOMOs, as they do for inorganic semiconductors. Another way of defining OSCs is to determine whether they function as electron donors (p-type) or electron acceptors (n-type) during device operation. For example, fullerene derivatives (Fig. 2(e)) are defined as n-type OSCs because they accept electrons upon photoexcitation and transport the photogenerated electrons in organic solar cells. Meanwhile, many conjugated polymers blended with fullerene derivatives are considered as p-type OSCs because they donate electrons upon photoexcitation in solar cells.

Device Applications

In this section, we review some of the most important organic electronic devices, including electrophotographic devices, OLEDs, organic thin-film transistors (OTFTs), and organic light-emitting transistors (OLETs). For practical applications, organic photoconductors were the first technology to use OSCs as a key component in commercially available products; they are often employed in offices in copy machines and laser printers (Fig. 10). In such an electrophotographic device, the image of a document is transferred to a photoconductive plate (photoreceptor), previously charged with static electricity, through a set of optical components. Because only parts of the photoconductive plate are discharged, as a result of its high photoconductivity, upon exposure to light, an image duplicates the original document on the surface of the photoconductive plate. The next development step involves the placement of the electrostatically charged toner particles over the partially charged surface; some of the particles are adhered to the image by the electric field. An electrostatic field then transfers the image composed of toner particles to a sheet of paper. Finally, the image is fixed through pressure and heating. After swiping out the toner particles (clean) and exposing to light to remove the remaining static electricity (erase), the photoconductive plate is ready for the next copy process.

The simplest device structure of an OLED is displayed in Fig. 11(a). It typically consists of two organic layers: an electron-transport layer (ETL) and a hole-transport layer (HTL). This design obviates the need to find an OSC possessing balanced injection and transport between electrons and holes. Unlike an individual single crystal (e.g., anthracene) that allows only effective hole injection, the bilayer structure facilitates hole injection from an indium tin oxide (ITO) electrode and electron injection from a metal alloy (MgAg) electrode. The charge balance can be further finely tuned through control over the thicknesses of the HTLs and ETLs. The excitons formed after recombination of the injected electrons and holes relax to the ground states, resulting in photon emission. Because the ITO anode is transparent, a strong green emission is observed through the ITO side of this device.

In solid state organic materials, only very weak van der Waals forces exist between the molecules. Many experimental results have indicated, however, that π -electrons can be transported effectively within OSCs. Organic field effect transistors (OFETs) provide a useful platform for examining the charge transport properties of OSCs. Transistors are fundamental building blocks in modern electrical circuits; they function as either signal amplifiers or on/off switches. The phenomenon of the conductivity of a semiconductor being modulated by applying an electric field normal to its surface is called the “field effect.” Most OFETs are OTFTs, implying that the active (semiconducting) layer is nearly two-dimensional. Fig. 11(b–d) presents the three common device structures of OTFTs; Fig. 11(e) illustrates the operation mechanism. The charge carriers travel in a channel at the dielectric–OSC

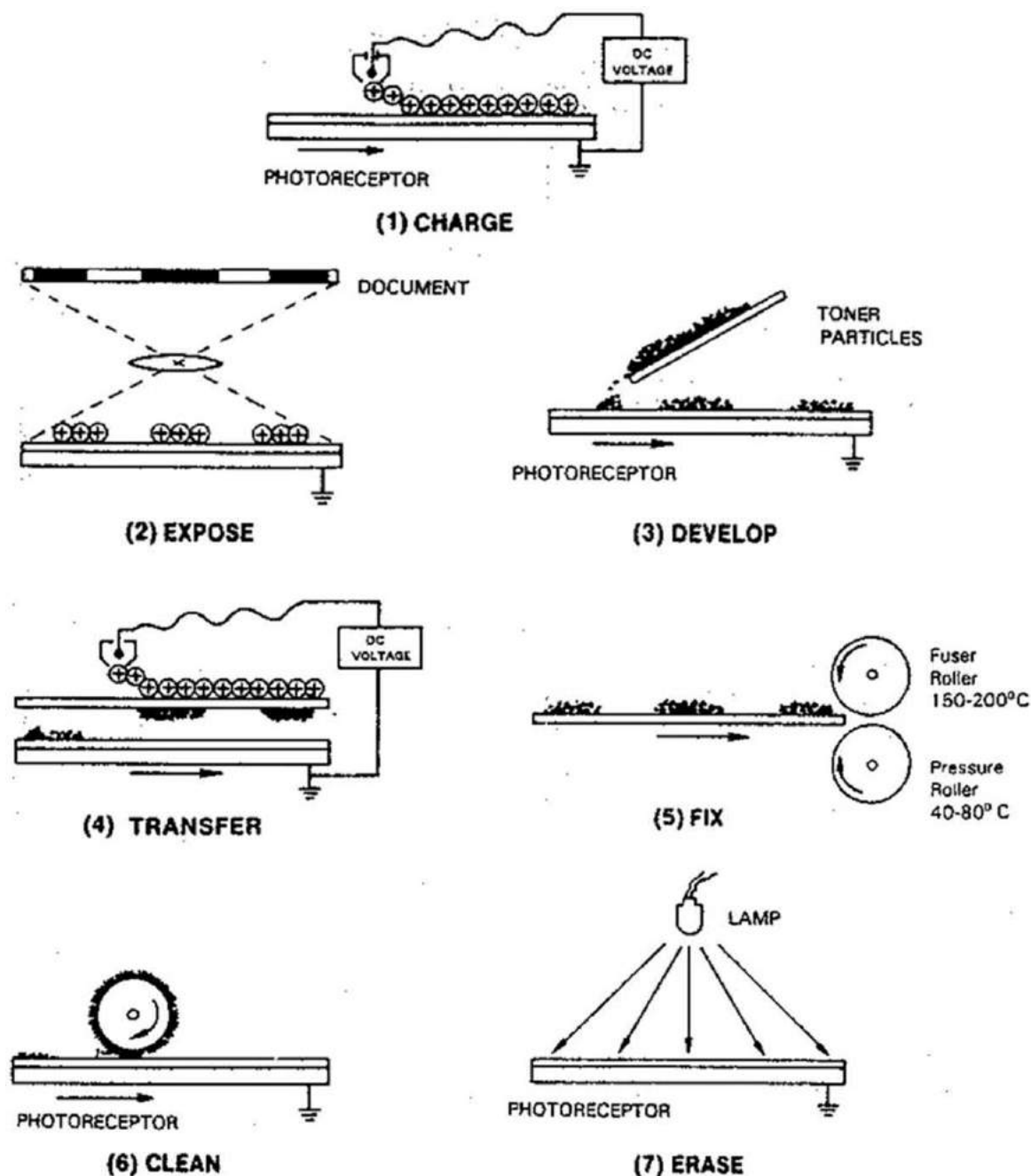


Fig. 10 Working principle of an electrophotographic device: (1) charge, (2) expose, (3) develop, (4) transfer, (5) fix, (6) clean, and (7) erase. Reproduced from Pope, M., Swenberg, C., 1999. *Electronic Processes in Organic Crystals and Polymers*, with permission from Oxford University Press, Copyright (1999).

interface. The source and drain electrodes provide contacts to the OSCs, and can inject and collect, respectively, the charges in the channel. The conductance of the channel can be modulated by a third gate electrode, which is separated by a layer of gate insulator. In a p-channel OTFT, for example, a negative voltage is applied to the gate (V_G), inducing holes in the channel through field effects. A negative bias is also applied between the source and drain electrodes (V_D), resulting in a driving force to flow the field-induced holes from the source to drain electrodes. The device behaves as a resistor, with the drain current (I_D) increasing linearly with respect to the increasing value of V_D . From the relationship between I_D and V_D (Fig. 11(f)), this operation region is called the linear region. While a substantial value of V_D is supplied, the magnitude of the field inducing the charges is partially canceled near the drain. When the effective voltage is lower than the threshold voltage (V_t), the concentration of induced charges drops to almost zero, forming a high resistance regime in the channel close to the drain end. The channel now pinches at one side and the value of I_D starts to saturate. This phenomenon is known as "pinch-off" and this operation region is the saturation region

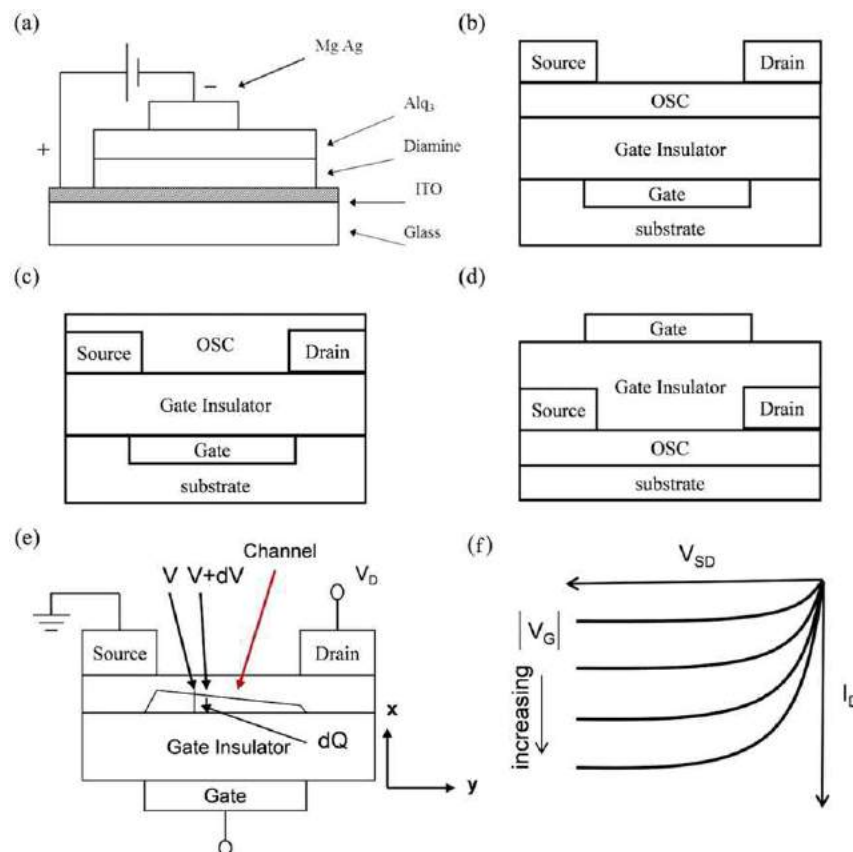


Fig. 11 (a) Device structure of a typical organic light-emitting diode (OLED); the layers of Alq_3 and diamine are the electron-transport layer (ETL) and hole-transport layer (HTL), respectively. Common structures of organic thin-film transistors (OTFTs): (b) top-contact, bottom-gate, (c) bottom-contact, bottom-gate, and (d) bottom-contact, top-gate. (e) Illustration of the operating mechanism of a typical OTFT. (f) Output characteristics (I_D – V_D) of a typical OTFT. ITO, indium tin oxide.

(Fig. 11(f)). Notably, the magnitude of the saturation current depends on the applied gate voltage. For n-channel OTFTs, the electrical behavior is similar, but the current–voltage curves are located in the first quadrant, instead of the third. The charge carriers are now electrons.

In the linear region, the drain current is described by:

$$I_D = C_{\text{ox}} \mu (W/L) [(V_G - V_{\text{th}}) V_D - (1/2) V_D^2] \quad (9)$$

where C_{ox} is the capacitance per unit area of the dielectric layer, μ is the mobility of the OSC, and W and L are the channel length and width, respectively. In the saturation region, the value of I_D can be expressed as

$$I_D = C_{\text{ox}} \mu (W/2L) (V_G - V_{\text{th}})^2 \quad (10)$$

Therefore, fitting the curves with the above equations can allow determination of the mobilities of the OSCs in either the linear or saturation regions. The other important parameter for evaluating the performance of an OTFT is the on/off ratio, defined as $I_{\text{on}}/I_{\text{off}}$, where I_{on} and I_{off} are the drain currents when the channel is turn on and off, respectively. A noticeable leakage current usually exists because charges can leak through the gate and/or because some charges are stored within the semiconductors. For state-of-the-art OTFTs, the mobility can readily reach up to $1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. Mobilities greater than $10 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ have also been demonstrated for solution-processed OTFTs. For example, a highly aligned metastable structure of 2,7-diocetyl[1]benzothieno [3,2-*b*][1]benzothiophene (C8-BTBT), grown from a blended solution of C8-BTBT and polystyrene, has displayed a mobility of $43 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ($25 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ on average), suggesting great potential for future electronic applications.

Because OLEDs are typically two-terminal devices, field-effect transistors are essential to control their luminescence levels and to switch individual OLED pixels when they serve as the elements of active-matrix display panels. In contrast, no transistors are needed when OLETs are employed for similar applications; a single OLET can perform both light-generation and electrical switching functionalities. For an unipolar OTFT, only a single type of charge carrier, electron or hole, is injected and transported along the conducting channel. Two carriers, however, are required to form excitons for light emission in OLETs. Therefore, an ideal OLET is an ambipolar device, although unipolar OLETs have also been reported. Light emission in unipolar OLETs occurs close to one of the electrodes, and the excitons are readily quenched by the metals, potentially decreasing the luminescence efficiencies.

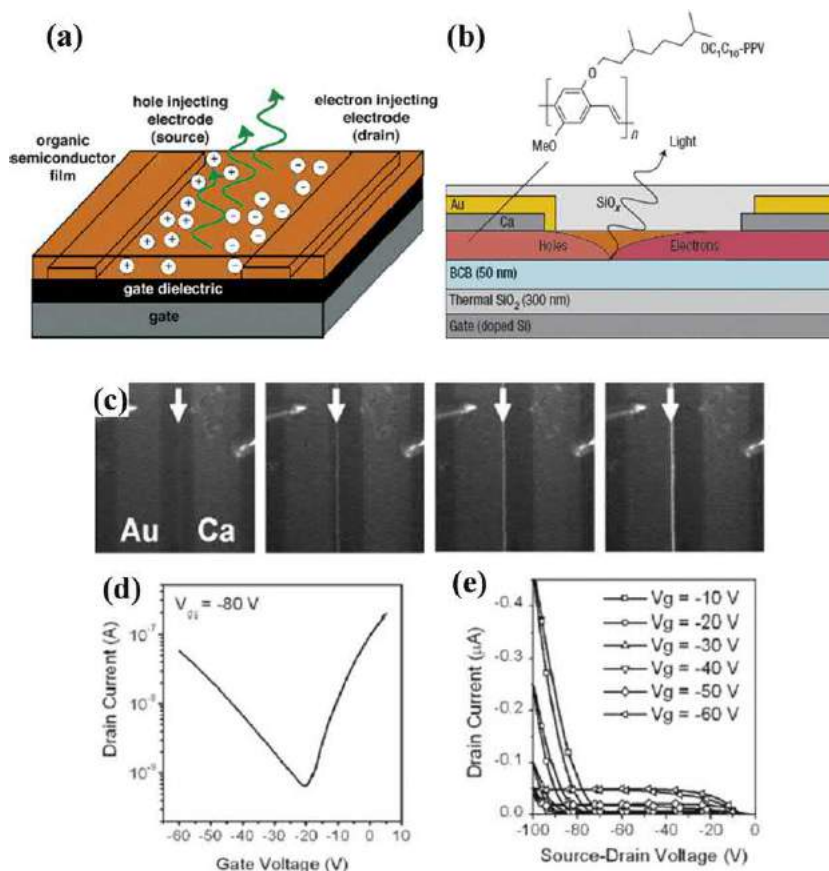


Fig. 12 (a) Schematic representation of a typical organic light-emitting transistor (OLET) device structure and its working principle. (b) Device structure of the first ambipolar OLET; the Au and Ca metal sources are positioned off-center, opposite each other, leading to two electrodes for hole and electron injection, respectively; the chemical structure of MDMO-PPV is also depicted. (c) Movement of the recombination zone as a function of drain bias: $V_G = -90$ V; $V_D = -79, -86, -93$, and -100 V (from left to right). (d) Transfer characteristics at $V_D = -80$ V. (e) Output characteristics. Reproduced from Cicoira, F., Santato, C., 2007. Organic light emitting field effect transistors: Advances and perspectives. *Advanced Functional Materials* 17, 3421–3434 with permission from WILEY-VCH Verlag GmbH & Co, Copyright (2007); and Zaumseil, J., Friend, R.H., Sirringhaus, H., 2006. Spatial control of the recombination zone in an ambipolar light-emitting organic transistor. *Nature Materials* 5, 69–74 with permission from Nature Publishing Group, Copyright (2007).

Fig. 12(a) displays the structure of a typical OLET having a bottom contact geometry; it is similar to that of a conventional OTFT. In other words, other structures, like the two other geometries in **Fig. 11**, are also possible choices. The source and drain electrodes are hole-injecting and electron-injecting electrodes, and the OSC should be emissive. Upon application of a reasonable gate bias (V_G), holes and electrons are injected from the source and drain contacts, respectively. The two opposite charges form excitons in the channel and, subsequently, they decay radiatively to produce light.

Ideal performance of an ambipolar device is, in practice, very difficult to achieve. First, the source and drain should be capable for injecting both holes and electrons efficiently. The OSC should also exhibit balanced hole and electron mobilities. In 2006, Zaumseil *et al.* reported the first ambipolar OLETs incorporating poly(2-methoxy-5-(3,7-dimethyloctyloxy)-*p*-phenylene-vinylene) (MDMO-PPV), a conjugated polymer displaying high photoluminescence efficiency (**Fig. 12(b)**). On the dielectric surface, they deposited a layer of crosslinked benzocyclobutene derivative (BCB) as a buffer dielectric to prevent the trapping of electrons at the dielectric–semiconductor interface. This approach is one of the keys to significantly improve electron (n-channel) transport on oxide dielectric surfaces. Au and Ca electrodes were then deposited for effective hole and electron injection, respectively. The use of Au electrodes only for both contacts would have led to high electron injection barrier at one side. Similarly, the use of Ca electrodes only would have led to a high hole injection barrier at one of the contacts. **Fig. 12(c)** illustrates the light emission from the OLET, and the movement of the light emitting zone between the two electrodes. In fact, no light emission was observed when the value of $|V_D|$ was below a certain threshold, because of limited electron injection. A narrow line of light appeared once the value of $|V_D|$ exceeded the threshold voltage. The light emission line moved toward the Au electrode upon increasing the value of $|V_D|$. The transfer characteristics of the device (**Fig. 12(d)**) revealed ambipolar behavior; the minimum point, at a value of V_G of approximately -20 V, indicates the transition from the electron- to hole-dominated current in the OLET. The output characteristics in **Fig. 12(e)** also reveal ambipolar currents. High-performance OLETs can also be fabricated by using mixed p- and n-type OSCs as the semiconducting layer. This approach can be accomplished through a bilayer structure, or co-deposition of hole and electron

transporting materials. More recent results have indicated that OLETs can outperform equivalent light-emitting diodes when using a trilayer semiconducting heterostructure.

Summary

The properties of OSCs are much different from those of inorganic semiconductors in many aspects. For example, π -electrons are responsible for the charge transport in OSCs, and the concept of doping is fundamentally different. Nevertheless, OSCs are very attractive for use in electronic devices because they are formed from inexpensive starting materials and have many useful properties, including solution-processability and flexibility. These unique properties offer many possibilities for various applications. For example, high-performance inorganic light-emitting diodes are usually fabricated through delicate crystal growth, and large-area devices are very difficult to implement. In contrast, OLEDs are readily pixelated and, therefore, can be used for flat-panel displays. We must emphasize, however, that organic electronic devices are not designed intentionally to compete with those containing inorganic counterparts. Rather, we believe that there will be many new applications in which inorganic semiconductors will not be useful; therefore, the exploration of organic devices should continue in the future.

See also: Organic Lasers. Organic Photodetectors

Further Reading

- Capelli, R., Toffanin, S., Generali, G., *et al.*, 2010. Organic light-emitting transistors with an efficiency that outperforms the equivalent light-emitting diodes. *Nature Materials* 9, 496–503.
- Cicoira, F., Santato, C., 2007. Organic light emitting field effect transistors: Advances and perspectives. *Advanced Functional Materials* 17, 3421–3434.
- Heeger, A.J., 2001. Semiconducting and metallic polymers: The fourth generation of polymeric materials (Nobel lecture). *Angewandte Chemie International Edition* 40, 2591–2611.
- Helfrich, W., Schneider, W.G., 1965. Recombination radiation in anthracene crystals. *Physical Review Letters* 14, 229–231.
- Pope, M., Swenberg, C.E., 1999. *Electronic processes in organic crystals and polymers*. New York, NY: Oxford University Press.
- Tang, C.W., VanSlyke, S.A., 1987. Organic electroluminescent diodes. *Applied Physics Letters* 51, 913–915.
- Turro, N.J., 1991. *Modern molecular photochemistry*. Sausalito, CA: University Science Book.
- Yuan, Y., Giri, G., Alexander, L., *et al.*, 2014. Ultra-high mobility transparent organic thin film transistors grown by an off-centre spin-coating method. *Nature Communications* 5, 3005.
- Zaumseil, J., Friend, R.H., Sirringhaus, H., 2006. Spatial control of the recombination zone in an ambipolar light-emitting organic transistor. *Nature Materials* 5, 69–74.

OLED/Introduction

Dae-Gyu Moon, Soonchunhyang University, Asan-si, Chungcheongnam-do, Republic of Korea

© 2018 Elsevier Ltd. All rights reserved.

OLED Architectures

Organic light-emitting diode (OLED) is a thin film electroluminescent device that emits electromagnetic radiation from organic materials sandwiched between two electrodes to an external electric field. Since the first observation on the organic electroluminescence in 1963, Tang *et al.* made a great progress in efficiency and lifetime using a double layer structure based on small molecules in 1987. Friend *et al.* first reported polymer light-emitting devices in 1990. Baldo *et al.* made a quantum leap in efficiency of OLED using a phosphorescent molecule in 1998.

A simple OLED structure is illustrated in Fig. 1. Organic thin films are situated between the anode and cathode. At least, one of both electrodes should be optically transparent. When a sufficient forward bias is applied to the anode, electrons, and holes are injected into the organic electron and hole transport layers (HTLs) from the cathode and the anode, respectively. The injected carriers transport through the respective carrier transport layers by the internal electric field toward the electrode of opposite polarity, recombining to form excitons in the emission layer. The radiative decay of these excitons generates photons in the visible range that are extracted out of the device.

Carrier injection is improved by lowering the potential barrier height at the electrode/organic interface, which can be approximately estimated by the difference between the work function of electrode and the highest occupied (or lowest unoccupied) molecular orbital (HOMO or LUMO) energy level. ITO is widely used as a transparent anode material because it is electrically conductive and its work function is relatively high (4.5–5.0 eV). ITO surface is treated with oxygen plasma or ozone in order to increase the work function and to remove surface contaminants. Low work function metals such as Mg, Ca, Mg:Ag, or LiF/Al are used as cathode materials. The light emission processes in OLEDs provide several features such as wide viewing angle, fast response time, and current driving of luminance. OLEDs can be fabricated on many kinds of substrates including large area glasses, silicon wafers, plastic films, metal foils, and textiles. This high degree of freedom in selecting substrates makes OLEDs very promising for future optoelectronic applications.

Various functional organic layers are used to improve the efficiency and lifetime in modern OLED architectures. Typical layers are hole injection layer (HIL), HTL, electron block layer (EBL), hole block layer (HBL), electron transport layer (ETL), and electron injection layer (EIL). EBL and HBL confine the recombination and emission processes to the emission layer (EML) composed of host and guest molecules. The host molecules can be mixed with two or three kinds of molecules to control the recombination and emission processes.

Small Molecule and Polymer OLEDs

OLEDs can be prepared using small molecules or polymers. Thin film of small molecules is usually deposited by thermal evaporation in high or ultrahigh vacuum (typically 10^{-6} – 10^{-7} Torr). The thermal evaporation of small molecules allows for easy deposition of multilayer devices with sophisticated architectures that precisely control all the emission processes from carrier injection to photon emission. Small molecules often have a thermal decomposition temperature much higher than the sublimation temperature so that the materials can be purified by repeated sublimation in vacuum or inert gas. High purity materials are essential in achieving high efficiency and long lifetime devices. Polymers cannot be deposited by thermal evaporation because their decomposition temperature is generally lower than their sublimation temperature. Polymer films are prepared by solution processing such as inkjet printing and spin coating. Hence, the polymer devices have advantages in simple and large area fabrication.

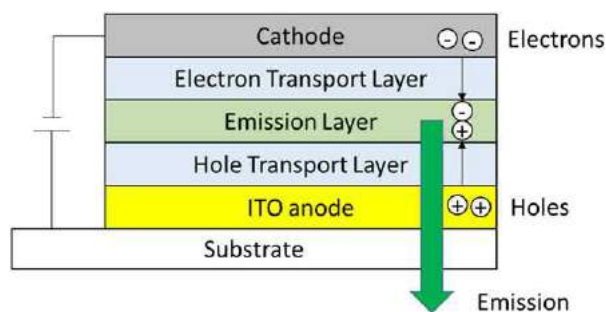


Fig. 1 Device structure of organic light-emitting diode (OLED) with triple organic layers.

Carrier Injection and Materials

Organic materials are regarded as insulator because they generally have very high resistance at low electric field. It means that a strong electric field is required to inject carriers from the electrodes to the LUMO or HOMO levels of organic layers. Charge carriers can be injected into the organic layers via Schottky or tunnel emission as shown in Fig. 2. Electrons or holes acquire sufficient energy at high temperature. When the energy of charge carriers is higher than the potential barrier at the interface, the carriers are thermally excited over the barrier into the organic layer. This phenomenon is called Schottky or thermionic emission. If the barrier is thin enough, electrons tunnel into the organic layer through the barrier. This mechanism is called field emission and it is expressed by Fowler–Nordheim equation assuming triangular shape of energy barrier. In both processes, small barrier height is important in improving carrier injection so that the organic materials with shallow HOMO and deep LUMO levels are preferred as the hole and electron injection layers, respectively.

Fig. 3 shows the organic materials for the HIL. Triphenylamine derivatives are widely used as a HIL material. Typical examples are 4,4',4''-tris[phenyl(*m*-tolyl)amino]triphenylamine (*m*-MTDATA) (*m*-MTDATA) and 4,4',4''-tris[2-naphthyl(phenyl)amino]triphenylamine (2-TNATA) (2-TNATA). copper phthalocyanine (CuPc) is also frequently used as a HIL material. HOMO levels of these materials are less than 5.2 eV. Hole injection can also be improved by using a material with a very deep LUMO level, where the holes are directly injected into the LUMO level rather than the HOMO level. For example, dipyrzino[2,3-f:2',3'-h]quinoxaline-2,3,6,7,10,11-hexacarbonitrile (HAT-CN) has a LUMO level of 5.5–6.0 eV. poly(3,4-ethylenedioxythiophene)-poly(styrenesulfonate) (PEDOT:PSS) is widely used as a hole injecting polymer layer. This conducting polymer has a work function of about 5.2 eV. Metal oxides such as MoO₃, WO₃ are often used as inorganic buffer layer for hole injection.

The selection of electron injecting material is somewhat complicated. A low work function metal results in a small barrier for electron injection. However, they are susceptible to atmospheric oxidation and corrosion. They easily react with air, water, and organic materials. More stable metals such as Al are widely used in electronics. However, they have too high work functions to efficiently inject electrons. Therefore, very thin electron injection buffer layer is inserted between ETL and Al cathode. Alkali metal halides and metal oxides are widely used as an electron injection buffer. For example, LiF/Al is most commonly used for the cathode. Lithium quinolate (Liq) is also widely used as a small molecule EIL for Al cathode. Low binding energy between lithium and hydroxyquinoline ligand liberates lithium in the Liq complex during deposition of Liq or cathode. The interface dipole layer can be formed by charge transfer, surface rearrangement, and chemical reaction at the organic/electrode interface. The dipole layer results in the vacuum level shift at the interface so that the injection barrier is different from the simple expectation based on the work function and the HOMO or LUMO level. The sign of vacuum level shift is generally negative, attributing to the downward shift of organic energy levels. Large amount of shift is often observed at the metal/organic interface.

Carrier Transport and Materials

When the number of injected charges is higher than the equilibrium charge density in organic material, the excess charges form a space charge inducing an electric field inside the organic layer. That is, local polarization layer is formed near the electrode, which limits the charge injection from the electrode. This polarized layer is called space-charge layer. Current flow is limited by the rate at which carriers drift out of the space-charge layer, so it is called space-charge-limited current (SCLC). As the organic materials generally contain defects and impurities acting as charge carrier traps, the charge carrier transport is also limited by the number of traps. In both cases of trap-limited and space-charge-limited conduction, the current is controlled by the mobility of charge carriers. Organic molecular crystals such as anthracene exhibit band-like charge carrier transport because the charge carriers are highly delocalized. Thus, the charge carrier mobility reaches to about 1 cm²/Vs. However, in the amorphous organic materials that are generally used in OLEDs, the charge carriers are highly localized due to weak intermolecular interaction and strong local polarization. Here, hopping process from one molecule to a neighboring molecule dominates the charge carrier transport. The resulting charge carrier mobility is low (10⁻⁷–10⁻² cm²/Vs) in amorphous organic materials.

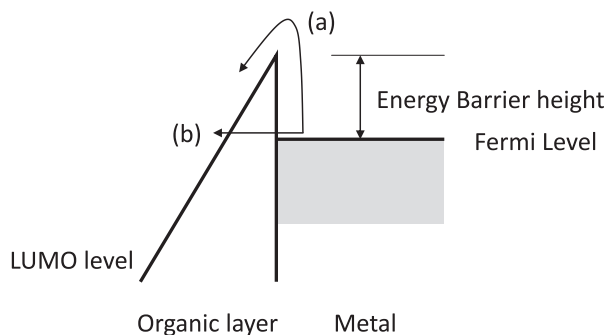


Fig. 2 Injection mechanisms of electrons from the metal into the organic layer under an applied voltage: (a) Schottky emission and (b) tunnel emission.

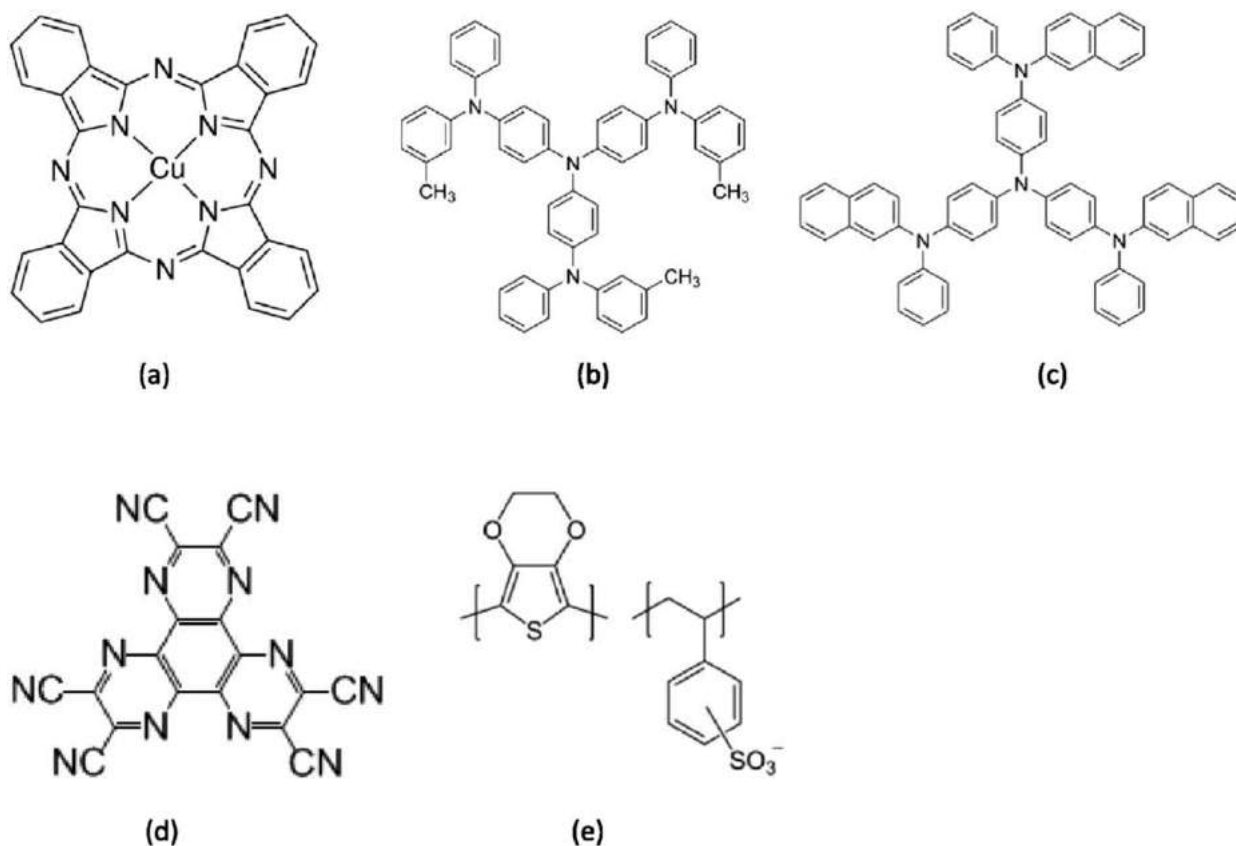


Fig. 3 Organic materials for the hole injection layer (HIL): (a) copper phthalocyanine (CuPc), (b) 4,4',4''-tris[phenyl(*m*-tolyl)amino]triphenylamine (*m*-MTDATA), (c) 4,4',4''-tris[2-naphthyl(phenyl)amino]triphenylamine (2-TNATA), (d) dipyrzino[2,3-*f*:2',3'-*h*]quinoxaline-2,3,6,7,10,11-hexacarbonitrile (HAT-CN), and (e) poly(3,4-ethylenedioxythiophene)-poly(styrenesulfonate) (PEDOT:PSS).

HTL provides conductive pathway for the injected holes to migrate into the EML. The HOMO level of HTL material is designed to be slightly shallower than that of EML material for improving the charge flow into the EML with minimal hole accumulation at the HTL/EML interface. Another function of HTL is to act as electron or exciton block layer that prevents the transfer of electrons or excitons in EML to the adjacent organic or anode layer. For this purpose, the HTL material should have a shallow LUMO level and wide energy gap. Fig. 4 shows various HTL materials. Typical examples are aromatic amine-based materials such as 4,4'-cyclohexylidenebis[*N,N*-bis(4-methylphenyl) benzenamine] (TAPC), *N,N'*-bis(3-methylphenyl)-*N,N'*-diphenylbenzidine (TPD), and *N,N'*-Di(1-naphthyl)-*N,N'*-diphenyl-(1,1'-biphenyl)-4,4'-diamine (NPD). They have shallow HOMO levels of 5.4–5.8 eV and shallow LUMO levels of 2.4–2.7 eV. They have a small energy for hole creation and the resulting radical cation is stable. They have relatively high hole mobility of 10^{-4} – 10^{-3} cm²/Vs. However, they are easily crystallized at high temperature due to their low glass transition temperatures (T_g) of 65–96°C. Carbazole-based *tris*(4-carbazoyl-9-ylphenyl)amine (TCTA) is also widely used as HTL material. TCTA has a high T_g of 151°C and its hole mobility is about 10^{-5} cm²/Vs. Polymers such as poly(9-vinylcarbazole) (PVK) and poly[(9,9-dioctylfluorenyl-2,7-diyl)-*co*-(4,4'-(*N*-(4-*sec*-butylphenyl)diphenylamine))] (TFB) are also widely used as HTL materials for solution-processed OLEDs.

ETL material should have good electron injection property, high electron mobility, high hole and exciton blocking ability. There are much smaller number of ETL materials compared with the HTL materials. The electron mobility of ETL materials is typically lower than the hole mobility of HTL materials. In operating OLEDs, it is difficult to achieve the chemical stability in ETL materials. Imidazole, siloles, boranes, and phosphine oxide, and metal complexes are used as an ETL material. Fig. 5 shows various ETL materials. *tris*(8-hydroxyquinoline) aluminum (Alq₃) is well-known electron-transporting material. The LUMO level (about 3.1 eV) of Alq₃ is sufficient to inject electrons whereas its HOMO level (<6.0 eV) is too shallow to block holes. 2,2',2''-(1,3,5-benzinetriyl)-*tris*(1-phenyl-1-*H*-benzimidazole) (TPBi) is an important ETL material for high efficiency OLEDs. The electron mobility of TPBi is 10^{-6} – 10^{-5} cm²/Vs. TPBi has large energy gap and its LUMO level is 2.7 eV while its HOMO level is 6.2–6.7 eV. 3-(biphenyl-4-yl)-5-(4-*tert*-butylphenyl)-4-phenyl-4 *H*-1,2,4-triazole (TAZ) and BCP are wide energy gap ETL material. Pyridine-containing 2,9-dimethyl-4,7-diphenyl-1,10-phenanthroline (BCP), (e) 1,3,5-*tris*(3-pyridyl-3-phenyl)benzene (TmPyPB) has a good electron-transporting ability, exhibiting high electron mobility of about 10^{-3} cm²/Vs. HOMO and LUMO levels of TmPyPB are 2.7 and 6.7 eV, respectively. Boron-containing 3TPYMB is also typically used as wide gap ETL material. ETL materials with wide energy gap and deep HOMO level such as TAZ, BCP, TPBi, and TmPyPB act as hole or exciton block layer. Various benzothiadiazole, pyridine, and oxadiazole polymers are generally used as an ETL material for solution-processed OLEDs.

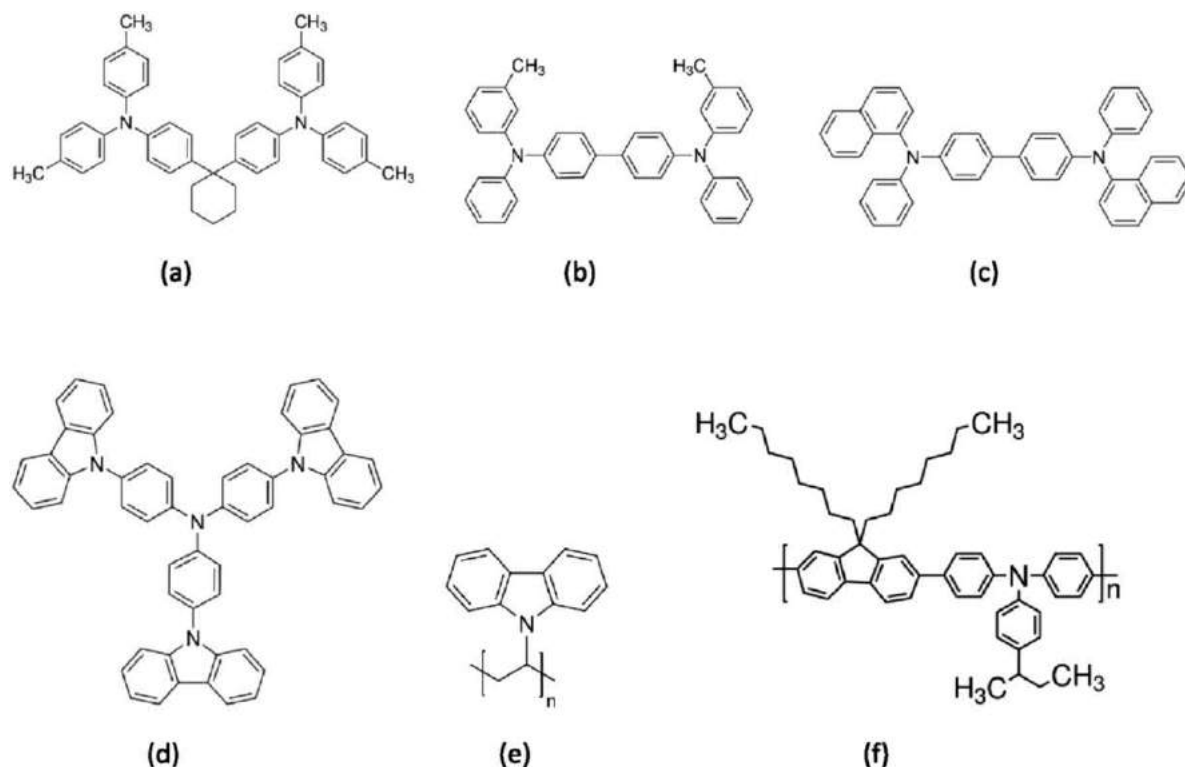


Fig. 4 Organic materials for the hole transport layer (HTL): (a) 4,4'-cyclohexylidenebis[*N,N*-bis(4-methylphenyl) benzenamine] (TAPC), (b) *N,N*-bis(3-methylphenyl)-*N,N*-diphenylbenzidine (TPD), (c) *N,N*-Di(1-naphthyl)-*N,N*-diphenyl-(1,1'-biphenyl)-4,4'-diamine (NPD), (d) *tris*(4-carbazoyl-9-ylphenyl)amine (TCTA), (e) poly(9-vinylcarbazole) (PVK), and (f) poly[(9,9-dioctylfluorenyl-2,7-diyl)-*co*-(4,4'-(*N*-(4-sec-butylphenyl) diphenylamine))] (TFB).

Doped Transport Layer

OLEDs generally require high electric field to inject charge carriers and to drive the injected carriers into the emission region because of contact barriers at the interface and low carrier mobility of organic materials. High operating voltage exceeding thermodynamic limit is typically required to drive an OLED at high luminance level. Consequently, doped transport layer is frequently used for minimizing the voltage drop in the transport layer and improving the charge carrier injection from the contacts into the neighboring organic transport materials. Hole transport or injection layer can be doped with the electron accepting atoms or molecules that releases the mobile holes in the layer (p-type doping). **Fig. 6** shows the various dopant molecules for the doped transport layer. For example, if *m*-MTDATA HIL is doped with 2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane (F4-TCNQ) molecules, electron transfer occurs from *m*-MDATA to F4-TCNQ as the electron affinity of F4-TCNQ is comparable to the ionization energy of *m*-MTDATA. It leads to negative charged F4-TCNQ molecules and positive charged *m*-MTDADA molecules, i.e., holes in the *m*-MTDATA layer that can move relatively freely throughout the *m*-MTDATA layer. F4-TCNQ molecule thus improves the electrical conductivity of the organic layer, eventually resulting in low voltage drop. HAT-CN is also widely used as a p-type dopant. The LUMO level of HAT-CN (5.7 eV) is very similar with those of typical HTL materials, thus the extra holes can be generated in the HTL by transferring electrons from HTL to HAT-CN molecules. MoO₃ is another effective p-type dopant in improving the injection efficiency and reducing ohmic losses in the transport layers. Other materials such as metals (Au, Pt), gases (Cl, Br, I), and small molecules (1,3,4,5,7,8-hexafluorotetracyanonaphthoquinodimethane – F6-TNAP) can be used as p-type dopants. Similarly, electron injection and transport can be improved by n-type dopants that donate electrons to the neighboring ETL molecules. Organic molecules (Liq, pyronine B – PyB chloride), inorganic materials (Cs₂CO₃), low work metals (Li, Cs) can modify the interface electronic structure, enhancing the conductivity of ETLs. The doped transport layer is typically separated by thin undoped organic layer from the emission region because the p-type and n-type dopants tend to provide nonradiative exciton recombination centers.

Exciton Formation and Decay

When electrons and holes recombine on organic molecules, two kinds of excitons, singlet and triplet excitons, are generated. Singlet excitons have spin antisymmetry states and their transitions to ground state are quantum mechanically allowed, giving rise to fluorescence as shown in **Fig. 7**. Triplet states with spin even symmetry lead to phosphorescence. Radiative transition rate of

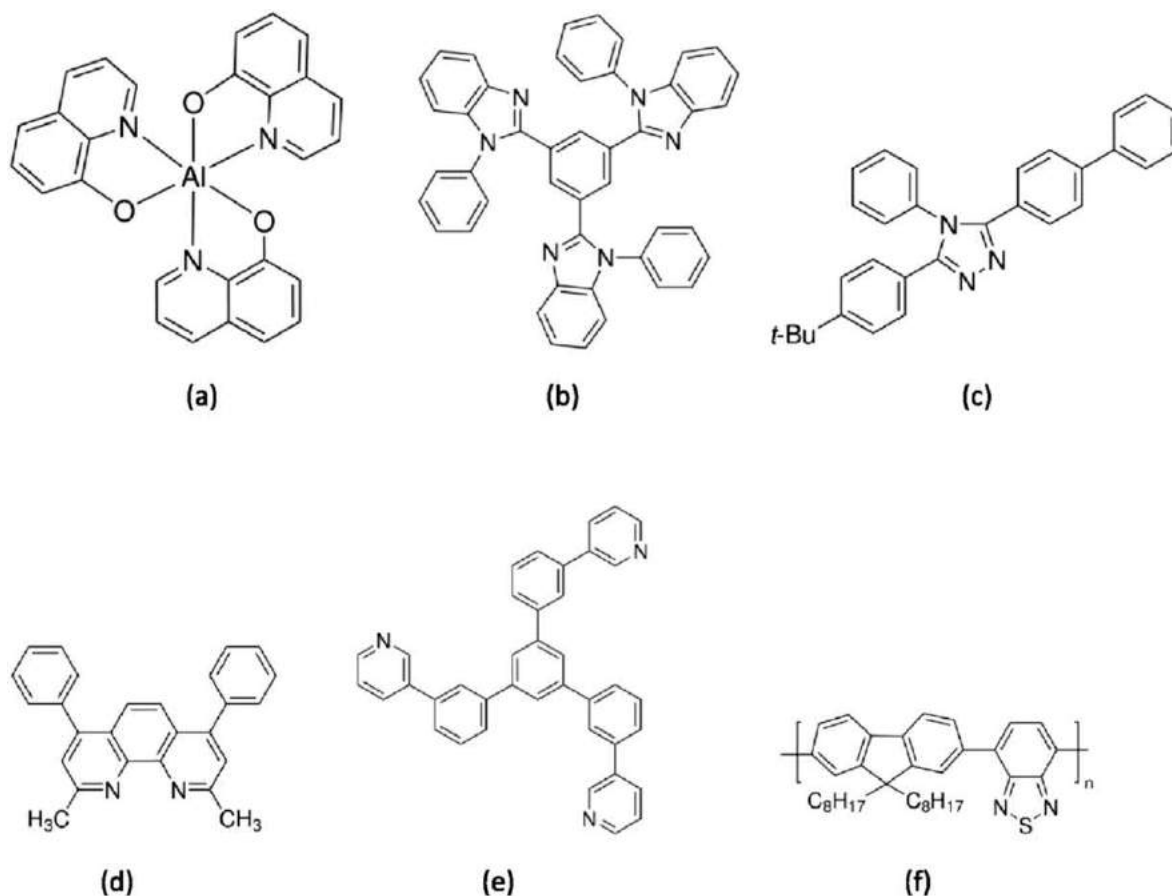


Fig. 5 Organic materials for the electron transport layer (ETL): (a) *tris*(8-hydroxyquinoline) aluminum (Alq_3), (b) 2,2',2''-(1,3,5-benzinetriyl)-*tris*(1-phenyl-1-*H*-benzimidazole) (TPBi), (c) 3-(biphenyl-4-yl)-5-(4-*tert*-butylphenyl)-4-phenyl-4-*H*-1,2,4-triazole (TAZ), (d) 2,9-dimethyl-4,7-diphenyl-1,10-phenanthroline (BCP), (e) 1,3,5-*tris*(3-pyridyl-3-phenyl)benzene (TmPyPB), (f) poly(9,9-dioctylfluorene-*alt*-benzothiadiazole) (F8BT).

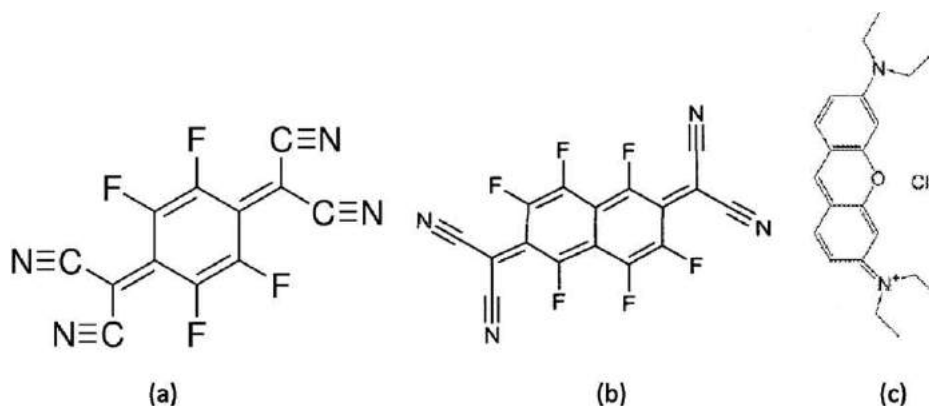


Fig. 6 Organic materials for the doped transport layer: (a) 2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane (F4-TCNQ), (b) 1,3,4,5,7,8-hexafluorotetracyanonaphthoquinodimethane (F6-TNAP), and (c) pyronine B (PyB) chloride.

singlet exciton is fast (< 10 ns), whereas the decay rate of triplet exciton is slow ($100 \mu\text{s}$ – 10 s) because it is a disallowed transition. Statistically 25% singlet and 75% triplet excitons are generated in the OLED due to the random nature of spin orientation when the carriers are injected into the organic materials from the electrodes. Therefore, the internal quantum efficiency (IQE) (photons/electrons) of fluorescent OLED is limited to 25%. In some fluorescent conjugated polymer or small molecules, the formation fraction of singlet excitons is higher than 25%. Triplet excitons can be converted to singlet excitons via reverse intersystem crossing (ISC) or triplet-triplet fusion in some fluorescent materials. In these cases, the IQE higher than 25% can be achieved in fluorescent OLEDs. In general, it is difficult to observe phosphorescence from fluorescent materials because the

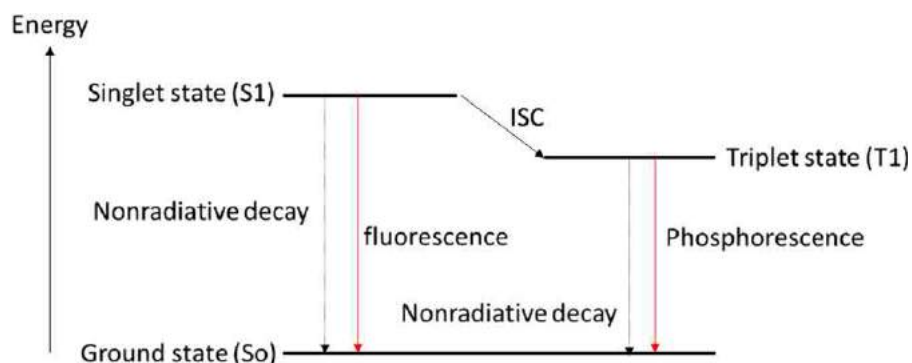


Fig. 7 Energy diagram of singlet and triplet excited states relative to the ground state, fluorescence, phosphorescence, and intersystem crossing (ISC).

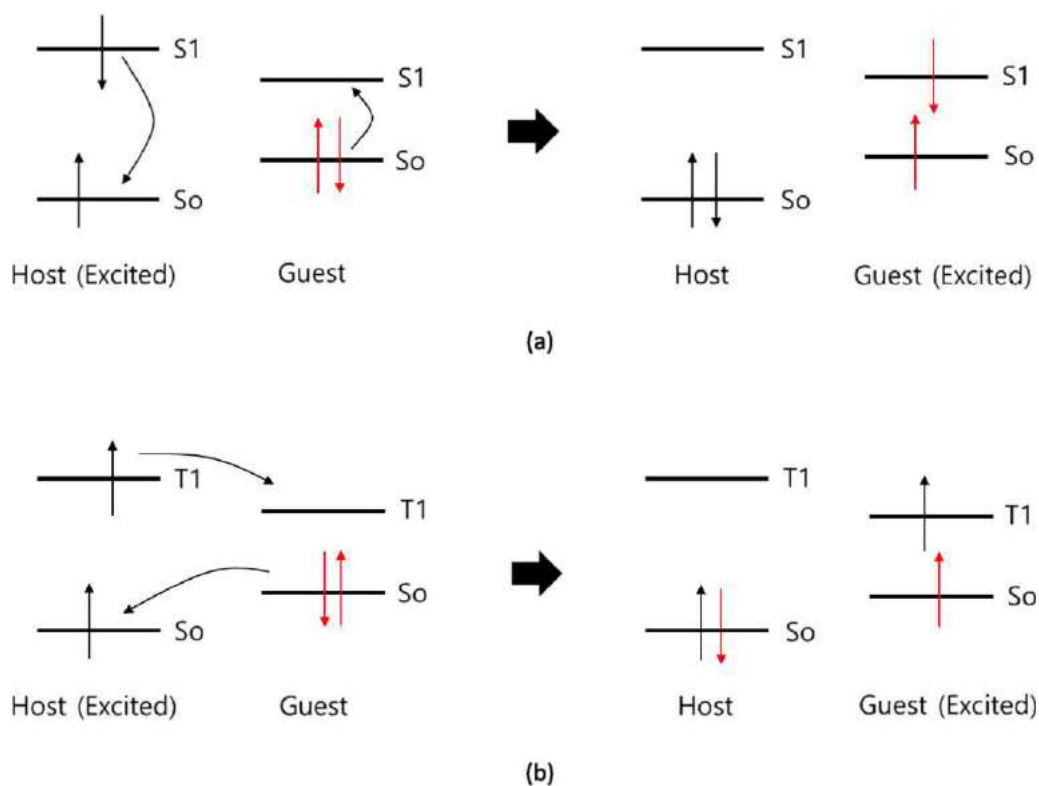


Fig. 8 Energy transfer processes from the host to the guest molecules: (a) Förster energy transfer and (b) dexter energy transfer.

radiative decay rate of triplets is much slower than that of the competing nonradiative decay process. However, phosphorescent molecules containing heavy metals such as iridium or platinum exhibit strong radiative transition from triplet excited states to ground states and fast ISC from singlets to triplet states, leading to 100% IQE device. Significant fraction of light generated by exciton decay is absorbed by surface plasmons and waveguide modes. Only about 20% of the light, i.e., roughly estimated by $0.5n^2$ where n is the refractive index of organic emitter, is coupled out of the device. Thus, the maximum achievable external quantum efficiency (EQE) is limited to 20% in OLEDs. However, high efficiency device with 20–40% EQE can be also obtained by outcoupling efficiency improvement techniques such as microlens array and horizontal dipole orientation.

Host–Guest System

Emission layer is composed of host and guest molecules in OLEDs (Fig. 8), where at least one electroluminescent emissive material is doped into the charge transporting host material. The host–guest system is an effective way to achieve high efficiency in small molecule OLEDs because most of emitters exhibit significant concentration quenching. That is, low photoluminescence

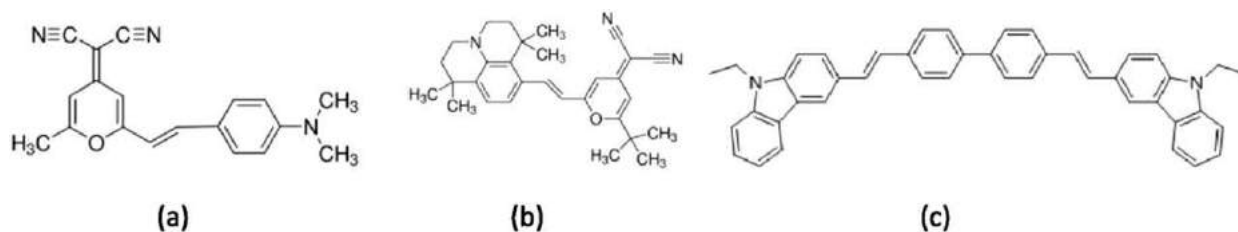


Fig. 9 Organic molecules for the fluorescent guest emitter: (a) 4-(dicyanomethylene)-2-methyl-6-(4-dimethylaminostyryl)-4 *H*-pyran (DCM), (b) 4-(dicyanomethylene)-2-*tert*-butyl-6-(1,1,7,7-tetramethyljulolidin-4-yl-vinyl)-4 *H*-pyran (DCJTB), and (c) 4,4'-*bis*(9-ethyl-3-carbazovinylenyl)-1,1'-biphenyl (BCzVBi).

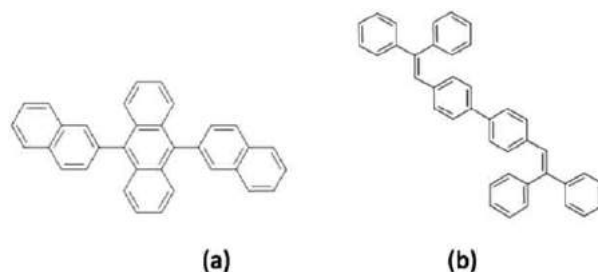


Fig. 10 Host materials for the blue fluorescent guest molecule. (a) 9,10-di(2-naphthyl)anthracene (ADN) and (b) 4,4'-*bis*(2,2-diphenylvinyl)-1,1'-biphenyl (DPVBi).

quantum yield (PLQY) is obtained in solid state even though the PLQY is about unity in gas phase or solution state. In polymer OLEDs, the EML material is often composed of a single polymer comprising multiple monomer units that provide emissive and charge transporting properties. However, the host–guest system is popular in both small molecule and polymer devices. The system also takes advantageous features of easy color tuning and operational stability. The dopant concentration in the host–guest system should be sufficiently low for avoiding nonradiative aggregate traps.

Excitons in the emission layer can be generated on the conductive host molecules. The energy of excited host molecules can be transferred to the guest molecules by Förster or Dexter mechanism. Förster energy transfer is a long range (up to about 10 nm) dipole–dipole coupling of host donor and guest acceptor molecules. This process contributes to the emission process in fluorescent OLEDs because of its singlet–singlet transfer nature. Dexter energy transfer is a short range (typically about 1 nm), intermolecular electron exchange process between donor and neighboring acceptor sites. Dexter transfer involves the triplet–triplet transfer process in phosphorescent OLEDs. Another exciton formation process is a direct recombination of electrons and holes on the guest molecules. It involves trapping of charge carriers and consequent recombination on the guest molecules.

4-(dicyanomethylene)-2-methyl-6-(4-dimethylaminostyryl)-4 *H*-pyran (DCM) and 4-(dicyanomethylene)-2-*tert*-butyl-6-(1,1,7,7-tetramethyljulolidin-4-yl-vinyl)-4 *H*-pyran (DCJTB) are widely used as fluorescent guest molecules as shown in Fig. 9. Coumarin and quinacridones are generally used as green fluorescent emitters. Distyrylarylene derivatives such as 4,4'-*bis*(9-ethyl-3-carbazovinylenyl)-1,1'-biphenyl (BCzVBi) are widely used as efficient fluorescent blue dopants. Alq₃ is the most popular electron-transporting host material for the fluorescent dopants. Fig. 10 shows the typical blue fluorescent guest molecules. 9,10-di(2-naphthyl)anthracene (ADN) and 4,4'-*bis*(2,2-diphenylvinyl)-1,1'-biphenyl (DPVBi) are thermally stable and good phase-compatible large bandgap host for blue fluorescent guest molecules. Iridium complexes such as Ir(piq)₃, Ir(ppy)₃, Irpic are typically doped into the wide energy gap host materials having high triplet energy levels in phosphorescent OLEDs.

Degradation of OLEDs

The luminance decreases with operating time as a result of physical, morphological, chemical degradation in OLEDs. The half-lifetime (T₅₀ or LT₅₀) is defined as the time the luminance drops to 50% of its initial value at the constant current density. Similarly, T₇₀ and T₉₇ refer to the time the luminance drops to 70% and 97% of its initial value, respectively. Typically, the lifetime becomes shorter at higher initial luminance value. The OLED luminance is approximately proportional to current density over a substantial range. Hence, the lifetime decreases in proportion to 1/(current density)^{*n*}, where *n* is acceleration factor of 1.2–1.9. The continuous drop of luminance under the constant current means the continuous decrease of device efficiency. This luminance degradation can be attributed to many causes such as electrochemical/photochemical reaction, crystallization, and impurities. Another common degradation phenomenon is the formation and growth of small and non-emissive regions which is called dark spots. High purity organic materials, high vacuum processes, and thorough device encapsulations are essential for long lifetime OLEDs.

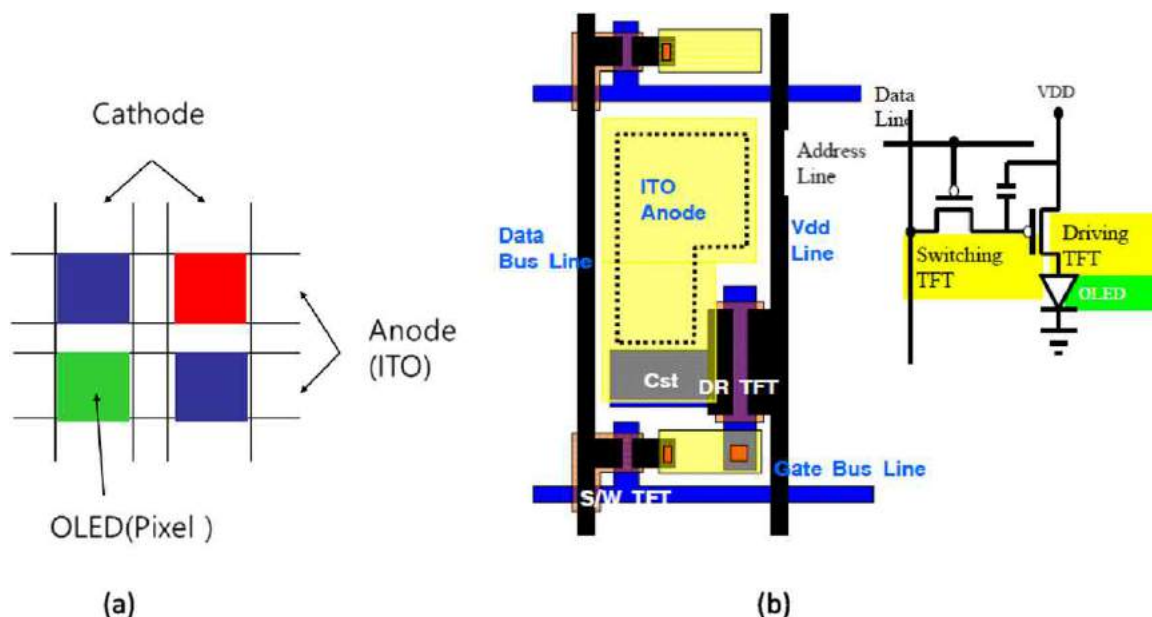


Fig. 11 Schematic pixel structures for organic light-emitting diode (OLED) displays: (a) passive matrix OLED (PMOLED) pixel and (b) active matrix OLED (AMOLED) pixel and equivalent circuit.

OLED Applications

The major applications of OLEDs are displays and lightings. OLEDs can be pixel elements in the passive matrix (PM) and active matrix (AM) displays. Vertically oriented data lines are crossed by the horizontally oriented scan bus lines in passive matrix OLED (PMOLED) displays (Fig. 11). OLED pixels are located at the crossing points. OLED pixels are fabricated by solution printing or evaporation process. PMOLED displays are limited to the low resolution and small size applications. On the other hand, active matrix OLED (AMOLED) displays are widely used for the high resolution, large size applications. AMOLED display contains thin film transistors (TFTs) for providing a constant current to an OLED pixel. Polycrystalline silicon and oxide TFTs are preferred to drive OLED pixels due to their high field effect mobility and current stability. Color pixels are fabricated by the several patterning techniques such as shadow mask process, transfer from donor substrates, and solution printing. Another method of using color filter and white OLED is widely used to fabricate high resolution and large area AMOLED displays. White OLED structure for displays is very similar with that for lightings. White OLED can be realized by various structures such as multi-emission layers, stacked device, and down-conversion method. Although the display techniques are mostly applicable in lighting applications, the light sources having a high color rendering index value are preferred for the reasonable color reproduction of the illuminated objects. Therefore, the white OLED for lightings has an emission spectrum of multiple wavelength peaks.

OLEDs also offer new possibilities not accessible in other device technologies. OLEDs can be prepared on various kinds of flexible and stretchable substrates. OLEDs can provide bendable, foldable, stretchable, and rollable displays or lightings. Transparent or mirror-type displays and lightings are also possible with OLEDs. They provide widow information displays, head-mounted displays, digital speed meters, and so on. The OLED technology will open up the numerous new fields of novel applications.

See also: Organic Semiconductors

Further Reading

- Baldo, M.A., O'Brien, D.F., You, Y., *et al.*, 1998. Highly efficient phosphorescent emission from organic electroluminescent devices. *Nature* 395, 151–154.
- Burroughes, J.H., Bradley, D.D.C., Brown, A.R., *et al.*, 1990. Light-emitting diodes based on conjugated polymers. *Nature* 347, 539–541.
- Gaspar, D.J., Polikarpov, E., 2015. *OLED Fundamentals: Materials, Devices, and Processing of Organic Light-Emitting Diodes*. New York, NY: CRC Press.
- Li, Z.R., 2015. *Organic Light-Emitting Materials and Devices*. New York, NY: CRC Press.
- Nalwa, H.S., Rohwer, L.S., 2003. *Handbook of Luminescence, Display Materials, and Devices: Organic Light-Emitting Diodes*. Valencia, CA: ASP.
- Scholz, S., Kondakov, D., Lüssem, B., Leo, K., 2015. Degradation mechanisms and reactions in organic light-emitting devices. *Chem. Rev.* 115, 8449–8503.
- Tang, C.W., VanSlyke, S.A., 1987. Organic electroluminescent diodes. *Appl. Phys. Lett.* 51, 913–915.
- Tsujimura, T., 2012. *OLED Displays: Fundamentals and Applications*. Hoboken, NJ: Wiley.

Harvesting Triplet Excitons in OLED

Gufeng He, Shanghai Jiao Tong University, Shanghai, China

© 2018 Elsevier Ltd. All rights reserved.

Introduction

As early as 1960s, Pope observed an electroluminescence (EL) phenomenon from organic semiconductor, where he used a single crystal of anthracene as EL material. The driving voltage was around 400 V and the luminance was very low. Vincett brought a major step-forward in organic EL in 1982, who applied a vacuum-deposited anthracene film as the emissive layer (EML) and the driving voltages dropped to below 30 V. However, these discoveries had not drawn many attentions until Tang and VanSlyke reported a double layer heterojunction organic light-emitting diode (OLED) in 1987. The device was composed of a vacuum thermal evaporated aromatic diamine layer for hole transporting, and a tris(8-hydroxyquinoline) aluminum (Alq3) layer for electron transporting and emission. This was the first time that an organic electroluminescent device could reach a high brightness ($> 1000 \text{ cd/m}^2$) with a relatively low operating voltage ($< 10 \text{ V}$) and an external quantum efficiency (EQE) about 1%, demonstrating its exciting prospect for applications in display and lighting. Since then, OLED has attracted tremendous interests and been developed rapidly. Up to now, small size OLED displays have been widely adopted in high-end smart phones, digital cameras, mp3 players, and so on; large size OLED TVs have also entered the market. With further development of OLED, more and more OLED displays will be around us and energy saving OLED lighting will definitely shine our daily life as well.

As of a novel flat panel display technology, OLED has its remarkable advantages of wide viewing angle, high contrast ratio, fast response, large color gamut, and low power consumption, and considered as the next generation of display competing with liquid crystal display (LCD). Furthermore, because of high efficiency, large area, light and thin factors, flexibility, OLED is also an ideal diffusive light source, showing a great potential in environmental friendly lighting applications.

A simple OLED structure consists of a thin organic electroluminescent film, sandwiched between two electrodes. When an external bias is applied on the device, holes and electrons are injected from the anode and the cathode into the organic semiconductor layer, respectively, then migrate in the organic layer to the opposite side. Subsequently, they meet in the EML and recombine to form excitons. Finally, the excitons radiatively decay to the ground states and emit light. However, to improve the charge carrier injection and transport and to confine excitons within EML to achieve high performance, multilayer structure is commonly adopted in advanced OLED. A typical high efficiency OLED consists of five organic layers (Fig. 1): hole injection layer (HIL), hole transport layer (HTL), EML, electron transport layer (ETL) and electron injection layer (EIL), where the HTL is often used as electron blocking layer (EBL), while ETL is used as hole blocking layer (HBL).

The OLED development focuses mainly on the materials and structures. Soon after Tang's first OLED report on small molecules, in 1990, Richard Friend *et al.* successfully discovered an OLED using solvable spin-coated polymer as EML, which is usually called as polymer light-emitting diode (PLED). In their publication, a single layer of poly(p-phenylenevinylene) (PPV), sandwiched between indium tin oxide (ITO) and Al electrodes, emitted green light under applied voltages. The relatively high efficiency and low turn-on voltage promised for a large-area, low-cost solution processable commercial application. However, all these small molecule and polymer materials are fluorescent, whose emission is from singlet excited state. The electrons in organic materials are

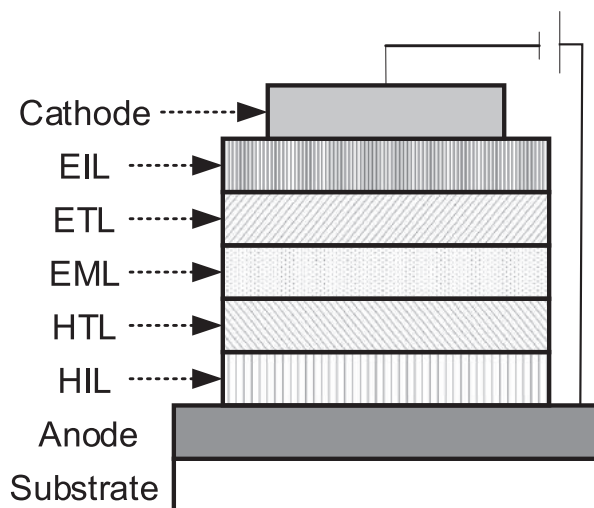


Fig. 1 Typical multilayer organic light-emitting diode (OLED) structure. EIL, electron injection layer; EML, emissive layer; ETL, electron transport layer; HIL, hole injection layer; HTL, hole transport layer.

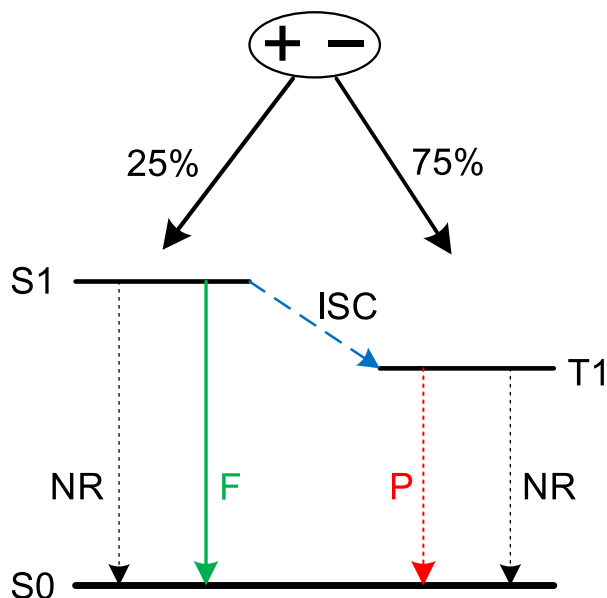


Fig. 2 Fluorescence (F) process. ISC, intersystem crossing.

mostly paired up with opposite electron spin orientations because of the Pauli Exclusion Principle. When receiving enough energy such as from light, one of the paired electrons can be excited from ground state to excited states. These two electrons have either the antisymmetric or symmetric spin orientation. Each electron spin is denoted as “ s ”, which equals to either $+1/2$ or $-1/2$. The spin multiplicity (total spin number, S), calculated by $S=2\sum s+1$, is either 1 for singlet if the pair of electrons are spin-antisymmetric or 3 for triplet if the pair of electrons are spin-symmetric. If an organic material is excited from the ground state, the excited state usually remains as singlet due to the Wigner–Witmer selection rule in quantum mechanics for the electronic transition process. Similarly, the transition from the triplet excited state ($T1$) to ground state ($S0$) is prohibited, i.e., the phosphorescence is not observable at room temperature for normal organic electroluminescent materials. In the OLED, the excitons are generated by injected holes and electrons from the anode and the cathode, respectively, which have their own spin orientations. Therefore, the electrically generated excitons can be either singlets or triplets. According to the statistic calculation of quantum mechanics, there will be 3 triplet excitons for every singlet exciton formed. Since radiative decay from triplet excited state to singlet ground state is spin-forbidden, only a quarter of electrically generated excitons, singlets, contribute to the light emission, which limits the internal quantum efficiency of OLED at 25%, as shown in Fig. 2.

To achieve a real high efficiency, it is necessary to harvest these non-radiative (NR) triplet excitons. One strategy is to make the radiative decay possible from triplet state $T1$ to ground state $S0$ by incorporating heavy metals in the molecules to increase spin-orbit coupling, which are called phosphorescent materials. Another strategy is to make the reverse intersystem crossing (RISC) process possible by narrowing the gap between the singlet and triplet excited states. By using these two methods, all electrically generated singlets and triplets can contribute to the light emission, and the internal quantum efficiency can be up to 100% theoretically, which is four times as high as conventional fluorescence.

In 1998, Thompson and Forrest groups boosted OLED internal quantum efficiency limit from 25 to 100% by using phosphorescent electroluminescent materials. Heavy metal atoms were widely used in phosphorescent materials to mix the triplet and the singlet excited states to make the radiative decay possible from triplet state to ground state. Recently, Addachi group from Kyushu University realized 100% internal quantum efficiency via thermally activated delayed fluorescence (TADF), without using heavy metal chelate. It is another breakthrough in the development of organic electroluminescent materials. If we call fluorescent materials as the first generation, high efficiency phosphorescent materials as the second generation, then TADF materials are considered as the third generation.

Phosphorescent EL

Although the OLED had been extensively studied since the first practical OLED was invented by Kodak, a breakthrough of its development is the discovery of electro-phosphorescence by Forrest *et al.* Compared with fluorescent OLED using only singlet excitons for light emission, phosphorescent OLED can utilize both singlet and triplet excitons, which boosts the internal quantum efficiency (IQE) of OLED by a factor of 4 theoretically. Hence, phosphorescence has attracted tremendous research interests in both academia and industry for its significant performance potential. In fact, highly efficient and long living red and green phosphorescent electroluminescent materials have already been employed in commercial OLED displays. Compared with the great achievements in red and green phosphorescent materials, although the EQE of blue phosphorescent OLED has already exceeded 20%, lacking of deep color and poor lifetime still restrain its commercialization.

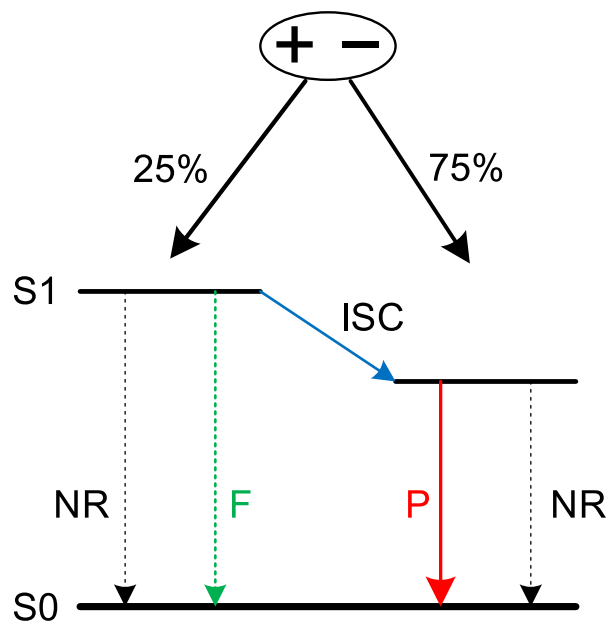


Fig. 3 Phosphorescence (P) process. ISC, intersystem crossing.

The typical lifetime of triplet excited state of organic material is usually longer than millisecond (ms), which is long enough for releasing its energy via NR decay channels such as the thermal vibration motion. There is one way to enable phosphorescence from organic materials is to suppress the thermal vibration motion by lowering temperatures down to liquid nitrogen temperature (77K). However, it is surely not practical for OLEDs, which usually operate at room temperature. A practical and effective way is to reduce the lifetime of triplet excited state so that the transition process from T1 to S0 can compete with the NR thermal vibration motion.

To enhance the phosphorescence, the most effective way is to use heavy metal atom in organic molecule, which enhances spin-orbit coupling, the interaction between spin magnetic moment and orbital magnetic moment. The strength of such coupling is proportional to the proton number in atomic nucleus, which is in turn related to the electromagnetic field generated by atomic electrons circling around. This is called as “heavy atom effect”. The distinguish between triplet and singlet state becomes diminished when spin-orbit coupling constant is large. The selection rule is losing its restriction and radiative decay from the triplet excited state to singlet ground state becomes feasible, and the triplet state lifetime is shortening, faster than molecular thermal vibration motion (**Fig. 3**). Up to now, the heavy atoms used in phosphorescent materials are all transition metals, such as Ru, Re, Os, Ir, Pt, Au, and Hg. Among them, most phosphorescent materials focus on those group VIII B elements, Os, Ir, and Pt. Especially, Ir-containing phosphorescent materials attract the most attentions.

The presence of a heavy metal atom in complexes increases singlet–triplet mixing and adds excited states to the electronic spectrum. In addition to states localized on the organic groups (ligands), the metal atom can participate in an excited state formed by the transfer of an electron from the metal to a ligand, known as metal–ligand charge transfer (MLCT) state. The MLCT state has a larger overlap with the heavy metal atom than the ligand excited states. Therefore, the spin-orbit coupling is higher, and the mixing between singlets and triplets is greater. Consequently, the radiative decay from the MLCT triplet state to ground state is more feasible. The phosphorescence often occurs from both MLCT and ligand excited states, and the relative importance of the MLCT and ligand components depends on their relative energies. If the ligand has a significantly lower triplet energy, the MLCT component will be very small, and the phosphorescence is mainly from the ligand triplet state. In this case, the triplet experiences less spin-orbit coupling induced by the central heavy metal atom, and consequently, its lifetime may not be reduced greatly. Triplet–triplet annihilation (TTA) is more pronounced with long triplet lifetime, and the OLEDs based on such materials may experience poor efficiency at large current densities due to strong efficiency roll-off. Thus, a high efficiency phosphorescent material should preferably have a triplet lifetime of less than $\sim 10 \mu\text{s}$, by enhancing the singlet–triplet mixing, which can be realized by ensuring that the energy of $^3\text{MLCT}$ triplet state is lower than that of ligand triplet state.

The first phosphorescent material employed in OLED was a red phosphorescent dye based on heavy metal Pt, called 2,3,7,8,12,13,17,18-octaethyl-21 H,23 H-porphine platinum(II) (PtOEP) (**Fig. 4**). The external quantum efficiency (EQE) reaches a maximum value of 4% with saturated red emission of (0.7, 0.3), which was not extremely high, relative to high efficiency fluorescent red emitter. However, PtOEP still attracted enough attention due to its phosphorescence nature. Firstly, PtOEP has an absorption peak at around 540 nm, and its fluorescence emission peak is at longer wavelength around 580 nm. In contrast, the observed EL peak is at 650 nm with a very nice narrow red emission band, which is much longer than the fluorescence emission of PtOEP. Furthermore, the lifetime of the emission at $\sim 650 \text{ nm}$ in solution was measured about $\sim 50 \mu\text{s}$, which is significantly longer than typical fluorescence lifetime of 0.1 $\sim$ 10 ns. All these spectroscopic data infer that the EL emission at 650 nm is

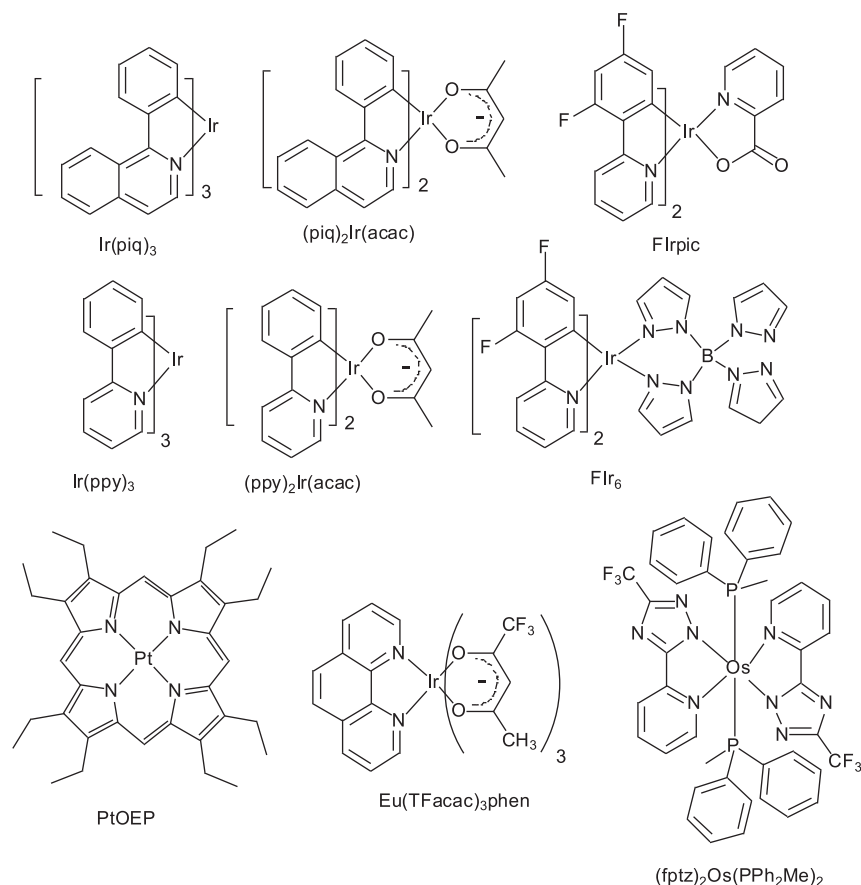


Fig. 4 Chemical structures of several typical phosphorescent materials.

phosphorescence rather than fluorescence. However, the PtOEP phosphorescent OLED shows serious efficiency roll-off, a typical EL behavior for phosphorescent OLEDs at elevated current density, which is mainly ascribed to their relatively long triplet lifetime. Furthermore, the flat square planar structure of Pt complexes often causes molecular π - π stacking that quenches the emission readily, leading to poor efficiencies in the devices.

Iridium is another mostly investigated heavy metal atom in phosphorescent materials. The $5d^6$ orbitals of Ir^{3+} provide a flexible platform for $^3\text{MLCT}$ phosphorescence. These complexes possess a chemically stable octahedral symmetry, with three bidentate ligands complexed to the central Ir atom. $\text{Ir}(\text{piq})_3$ (tris(1-phenylisoquinoline)iridium(III)) and $(\text{piq})_2\text{Ir}(\text{acac})$ (bis(1-phenylisoquinoline) (acetylacetonate)iridium(III)) are two typical red phosphorescent emitters based on Iridium (Fig. 4). The common feature of these two Ir complexes is the cyclometalate ligand 1-phenylisoquinoline (piq). In fact, cyclometalate ligand piq is very powerful for red phosphorescence. Most of the red phosphorescent materials developed later are based on derivatives of piq ligand.

If the ligands are changed to ones with higher triplet energies, the phosphorescence emission can be shifted to the shorter wavelength. For example, the ligand 2-phenylpyridine (ppy) is the parent structure for many other cyclometalation ligand of green phosphorescent materials, such as $\text{Ir}(\text{ppy})_3$ (tris(2-phenylpyridine)iridium(III)) and $(\text{ppy})_2\text{Ir}(\text{acac})$ (Bis[2-(2-pyridinyl-N)phenyl-C](acetylacetonato)iridium(III)) (Fig. 4). The ancillary ligand acac has a weaker ligand-field strength than ppy, leading to smaller d-orbitals splitting, which results in smaller energy of MLCT. Therefore, the emission peak wavelength of $(\text{ppy})_2\text{Ir}(\text{acac})$ is longer than that of $\text{Ir}(\text{ppy})_3$ (525 vs. 514 nm). The photophysical properties of $\text{Ir}(\text{ppy})_3$ and $(\text{ppy})_2\text{Ir}(\text{acac})$ are quite similar, although $\text{Ir}(\text{ppy})_3$ is homoleptic and $(\text{ppy})_2\text{Ir}(\text{acac})$ is heteroleptic. The phosphorescence quantum yield of $\text{Ir}(\text{ppy})_3$ is 0.40 and that of $(\text{ppy})_2\text{Ir}(\text{acac})$ is 0.34. The triplet lifetime of $\text{Ir}(\text{ppy})_3$ is 1.9 μs and that of $(\text{ppy})_2\text{Ir}(\text{acac})$ is 1.6 μs . The color coordinates of $\text{Ir}(\text{ppy})_3$ are (0.27, 0.63), whereas those of $(\text{ppy})_2\text{Ir}(\text{acac})$ are (0.31, 0.64). Such differences are not large, but very similar trend have been observed for other homoleptic and heteroleptic coordination complexes as well. In terms of molecular dipole moment, facial homoleptic coordination complexes are smaller than heteroleptic coordination complexes, which in turn are smaller than meridional homoleptic coordination complexes. And it is indicated by the experimental data that the molecular dipole will quench phosphorescence of the material.

Alternatively, if the ligand is made strongly electron accepting, then $^3\text{MLCT}$ emission may be blue shifted. The most famous blue phosphorescent materials are heteroleptic bis[2-(4,6-difluorophenyl)pyridinato-C2,N](picolinato)iridium(III) (FIrpic) and bis (2,4-difluorophenylpyridinato)-tetrakis(1-pyrazolyl)borate iridium(III) (FIr6) using fluoro-substituted phenylpyridine ligands

and an anionic 2-picolinic acid or poly(pyrazolyl)borate as an auxiliary ligand, respectively (Fig. 4). However, these emissions are either greenish-blue (0.17, 0.34) or sky-blue (0.16, 0.26), far from a saturated blue color as the National Television Standards Committee (NTSC) standard of (0.14, 0.08).

As the first blue phosphorescent material, FIrpic has two emission peaks at 470, 494 nm with phosphorescence quantum yield of 50–60%. The strong vibration peak at longer wavelength of main emission peak makes the color actually cyan containing a lot of green emission. With improved blue color, FIr6 has a shorter emission wavelength (λ_{max} at 457, 485 nm) and a higher phosphorescence quantum yield (96%) than those of FIrpic, but its device stability was poorer. Since the emission is mainly from the $\pi\pi^*$ state of cyclometalated ligands mixing with d-orbital of heavy transition metal via MLCT, such emission state due to the mixing of d-orbital is often inadequate, particularly for the short wavelength blue phosphorescence. Consequently, such phosphorescence usually has a longer triplet lifetime, closer to pure organic species (organic compound without containing heavy atom). Triplet state lifetime of microsecond range is required for fast emissive decay process and high phosphorescence quantum yield is desirable to achieve theoretical maximum EQE over 20%. It has been known that Ir-based organometallic compounds show short excited state lifetime due to singlet–triplet mixing and easy control of emission spectra by managing the chemical structure of ligands.

FIrpic has two fluorines on the phenyl unit to shift the highest occupied molecular orbital (HOMO) level downward for high triplet energy. The triplet energy is increased by the electron withdrawing fluorine and a bulky picolinic acid ancillary ligand. Although the emission spectrum of FIrpic is blue-shifted, it exhibits only sky blue color with y-coordinate over 0.30. The spectrum is further blue-shifted by changing the ancillary ligand from picolate (pic) to bulky tetrakis(1-pyrazolyl)borate ligand in FIr6. The FIr6 has bulky tetrakis(1-pyrazolyl)borate ligand to reduce the conjugation of the main ligand through steric hindrance. The peak wavelength of the FIr6 is shifted to 458 nm compared with 470 nm of FIrpic.

Besides platinum and iridium, europium (Eu) is also employed as the central metal to develop phosphorescent emitters. For example, Zheng *et al.* obtains Eu(TFacac)₃phen using 1,1,1-trifluoroacetylacetone as ligand, and Male *et al.* demonstrates Eu (TTFA)₃phen using tris-(thiophenyltrifluoromethylacetylacetone)(phenanthroline) as the ligand. The EL emission from Eu complexes is extremely sharp since the emission is from the f–f transition of Eu atom.

Os is another heavy metal atom often used in phosphorescent materials, and the emission color can be tuned by changing the ligand. Jiang *et al.* have designed a series of Os complexes. The phosphorescence lifetime is 0.6–1.9 μs , and the EL emission peak is from 620 to 650 nm, with color coordinates around (0.65, 0.33). For example, when (fptz)₂Os(PPh₂Me)₂ was doped into tris[4,9-phenylfluoren-9-ylphenyl]amine (TFPA), the device exhibits red EL having EQE as high as 18% (power efficiency ~ 25 lm/W). It is quite remarkable that the device has minor efficiency roll-off, which is mainly attributed to the short phosphorescence lifetime of ~ 0.7 μs . On the other hand, symmetrical structure induces small net dipole moment and reduce dipole–dipole interaction and quench emission.

As of relatively long phosphorescence lifetime, the phosphorescent emitter materials usually need to be doped into suitable host materials to avoid severe TTA. Besides some common rules to be followed, such as excellent film-forming property, good thermal and electrochemical stability, high glass transition temperature (T_g), there are still some special criteria to meet for host materials of phosphorescent emitters. Firstly, the host material should possess appropriate the HOMO and the lowest unoccupied molecular orbital (LUMO) levels to ensure effective hole and electron injection; secondly, the host material should have higher triplet energy than the phosphorescent emitter to prevent reverse energy transfer; Finally, the host material should possess high hole or electron transport property to reduce the operating voltages. Recently, bipolar host materials, capable of transporting both holes and electrons, have drawn more attentions. It is confirmed that employing bipolar host materials can not only reduce the operating voltages but also improve the device stabilities.

Advanced Delayed Fluorescence

After nearly 20 years development, all red, green and blue phosphorescent OLEDs have achieved extremely high EQE over 30%. However, these phosphorescent emissive materials contain heavy metals such as iridium and platinum, which are rather expensive and resource limited. To overcome this issue, a lot of studies focus on delayed fluorescent materials, which do not use heavy metals and can convert triplet excitons to singlet excitons through RISC. There are mainly two methods to achieve efficient RISC: TTA and TADF. More and more ultra-high efficiencies of fluorescent OLEDs exceeding 10%, far beyond the theoretical limitation of fluorescent emitters, have been reported recently, attracting wide research interests in both academic institutions and industry (Figs. 5 and 6).

In 2009, Kodak demonstrated some high EQE OLED devices using a fluorescent material 4,4'-bis[4-(di-p-tolylamino)styryl]biphenyl doped in 9,10-bis(2-naphthyl)-2-phenylanthracene (PADN) as EML, which are attributed to TTA contribution. Based on the assumption that spin statistics dictate 20% probability of generating a singlet exciton through TTA, they proposed a formula to estimate the maximum EQE of such system: $0.2 \times (25\% + 0.2 \times 75\%) = 8\%$. Soon later they obtained an EQE of $> 11\%$ at 3 V using rubrene as emitter. This value is far beyond the theoretical limit of the fluorescent emitter imposed by spin statistics, and cannot be satisfactorily explained by previous equation. They surmise that the fluorescent OLED devices are capable of using a considerably larger fraction of triplet states than it was previously believed. With the appropriate emissive material and host material, every two triplet excitons can form one singlet exciton to contribute to the light emission, i.e., the upper limit for the singlet excited state yield in the TTA process is 0.5. Therefore the maximum internal quantum efficiency of fluorescent OLEDs is to be $25\% + 0.5 \times 75\% = 62.5\%$. The estimated maximum EQE of the fluorescent OLEDs should be revised to at least $0.2 \times 62.5\% = 12.5\%$ after considering a light out-coupling efficiency of $\sim 20\%$ in the device. However, in order to promote the nonlinear up-conversion TTA process, high driving voltages or concentrated sensitizers are generally essential, accompanied by proportionally increased device efficiency roll-off.

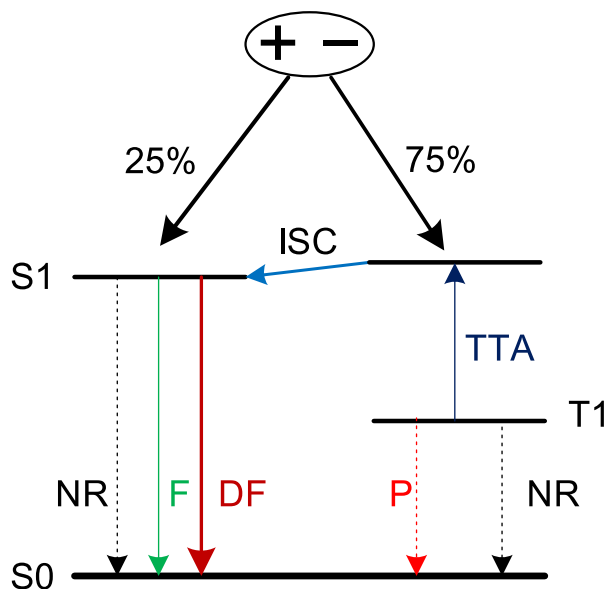


Fig. 5 Triplet-triplet annihilation (TTA) process. ISC, intersystem crossing.

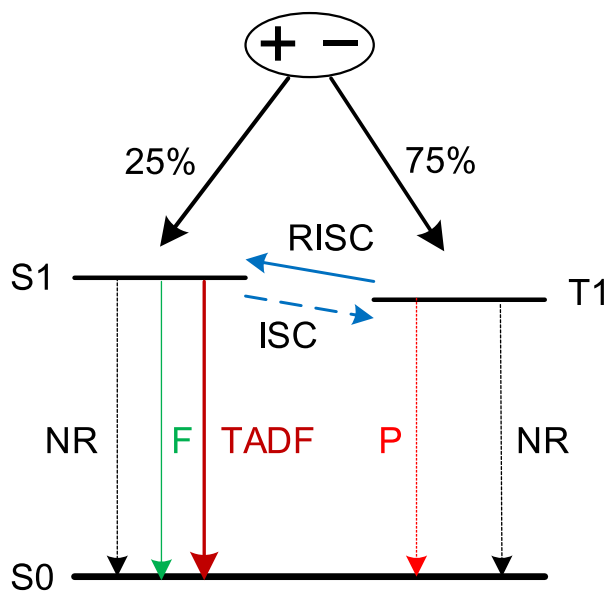


Fig. 6 Thermally activated delayed fluorescence (TADF) process. ISC, intersystem crossing; RISC, reverse intersystem crossing.

A more efficient way to harvest triplet excitons is via the direct RISC process from T1 to S1. According to Hund's rule, the energy of T1 is always lower than S1, therefore such RISC process must be stimulated or activated by external energy. If the energy difference between T1 and S1 (ΔE_{ST}) is small enough, the thermal activation of the molecules can overcome the barrier and commit the endothermic RISC process. The triplet excitons at T1 are back-transferred to S1, followed by radiative decay to the ground state S0, called TADF. There are two distinct unimolecular mechanisms involved in TADF: the prompt fluorescence (PF) from the direct electrical excited singlet S1, and the delayed fluorescence (DF) from the back-transferred singlet S1. Apparently, PF and DF have different luminescent lifetimes, which can be distinguished from the transient photoluminescence spectroscopy. And their individual fluorescence efficiency can also be calculated by integrating the intensity of each component.

In these processes, the efficient RISC from T1 to S1 is the key to harvest triplet excitons to enhance the DF. However, when the RISC process is enhanced by reducing the singlet-triplet energy splitting ΔE_{ST} , the intersystem crossing (ISC) process from S1 to T1 is also enhanced. Moreover, the RISC rate ($\sim 10^6 \text{ s}^{-1}$) is lower than the ISC rate in TADF molecules generally. In the end, there is a balance of exciton distribution between S1 and T1 states in TADF molecules. Consequently, ΔE_{ST} , rate constant of ISC and RISC are the key parameters to design efficient TADF materials.

In traditional fluorescent and phosphorescent materials, S1 energy is considerably higher than T1, and RISC hardly occurs. To facilitate RISC, a small ΔE_{ST} is particularly important. At S1 or T1 state, the unpaired two electrons are mainly distributed on the frontier orbitals of the HOMO and the LUMO, respectively. To obtain small ΔE_{ST} in the molecules, a spatial wave function separation of HOMO and LUMO need to be adopted. The general strategy is to introduce a large steric hindrance structure or a donor–acceptor system with twist/spiro/bulky connection, which can reduce the overlap between the HOMO and LUMO.

RISC process is an endothermic transition, hence it is temperature sensitive, and high temperature facilitates it. At low temperature (<100K), the fluorescence efficiency of TADF molecules are significantly suppressed and is strongly dependent on photoexcitation intensity, which are considered as characteristic features of TADF emitters.

At the early stage, such TADF materials always exhibit low fluorescence efficiencies, making it really difficult to verify the effectiveness for EML in OLED. It was not until 2009 that Adachi group at Kyushu University first introduced a series of Sn(IV) fluoride-porphyrin complexes with small ΔE_{ST} around 0.4 eV. It is demonstrated that the photoluminescence efficiency of such TADF material can be increased from ~1 to 3% with an increase in temperature due to RISC. However, the activation energy of the RISC (ΔE_{ST}) was still rather high at 0.24 eV, leading to inefficient RISC, and the overall EL efficiency is still very low.

Donor–acceptor systems with pronounced intramolecular charge transfer (CT) characteristics are very suitable to realize small ΔE_{ST} through separated HOMO in donor groups and LUMO in acceptor groups for TADF emission. A steric hindrance, such as twist, bulky, or spirojunction, incorporated between the donor and acceptor groups can further effectively separates the spatial overlaps between HOMO and LUMO. For example, a typical molecule 2-biphenyl-4,6-bis(1,2-phenylindolo[2,3-a]carbazol-11-yl)-1,3,5-triazine (PIC-TRZ) has ΔE_{ST} of only 0.11 eV, which realizes a significant contribution of 30% RISC efficiency under both photoluminescence and EL processes.

Another strategy proposed again by Adachi group to realize high RISC efficiency is via exciplex of electron-donating and electron-accepting molecules, also called as intermolecular donor–acceptor system. The electron transition from the LUMO of an acceptor to the HOMO of a donor is in a large electron–hole separation distance for exciplex emission. It is possible that the hole and electron wavefunctions are spatially separated and electronic exchange energy is very small compared to that of intramolecular excited states, resulting in the triplet levels being very close to the singlet levels, i.e., a small ΔE_{ST} for efficient TADF emission via RISC. However, the limited overlap of hole and electron wavefunctions generally results in very low fluorescence efficiency of exciplex according to the Franck–Condon principle. For strong TADF emission in the exciplex, the donor and acceptor molecules should have high triplet energy levels to confine the triplet exciplex state to prevent the quenching of the triplet state via the triplet energy back transfer to the donor or acceptor. Furthermore, such intramolecular donor–acceptor system should have not only high RISC efficiency but also high fluorescence efficiency. For example, an exciplex system of 4,4',4''-tris[3-methylphenyl(phenyl)amino]triphenylamine (m-MTDATA) as a donor and tris-[3-(3-pyridyl)mesityl]borane (3TPYMB) as acceptor has a high RISC efficiency of 86.5%. Using this emission system, the OLED achieves a maximum EQE of 5.4%, even with the rather low photoluminescence efficiency of 26%, showing a significant contribution of efficient RISC.

Within just several years' development, the EQE of TADF OLED is already over 21%, and further increased to 30% in a suitable cohost system. The device stability is significantly improved as well, with lifetimes of 2800 h at an initial luminance of 1000 cd/m², and of over 10,000 h at 500 cd/m². To take advantage of the triplet harvesting property, TADF molecules can act as host materials or singlet exciton sensitizers to harvest triplet excitons in OLEDs for fluorescent emitters. The devices show a significantly increased EQE and an enhanced device operational stability.

Although TADF materials have the potential to harvest ~100% excitons, the mechanism of TADF still requires further clarification. The efficiency roll-off is quite serious for most TADF emitter based OLED, which could be attributed to pronounced triplet–triplet or singlet–triplet quenching because of slow RISC process. It is necessary to further improve the performance of TADF materials to meet the requirements of real application. Novel device configurations should be developed according to the molecular features of TADF materials, which will promote the overall device performances for advanced optoelectronic applications.

See also: Organic Semiconductors

Further Reading

- Baldo, M.A., O'Brien, D.F., You, Y., *et al.*, 1998. Highly efficient phosphorescent emission from organic electroluminescent devices. *Nature* 395 (6698), 151–154.
- Chiang, C.-J., Kimyonok, A., Etherington, M.K., 2012. Ultrahigh efficiency fluorescent single and bi-layer organic light emitting diodes: The key role of triplet fusion. *Advanced Functional Materials* 23, 739–746.
- Huang, S.H., Zhang, Q.S., Shiota, Y., 2013. Computational prediction for singlet- and triplet-transition energies of charge-transfer compounds. *Journal of Chemical Theory and Computation* 9, 3872–3877.
- Li, Y.F., 2015. *Organic Optoelectronic Materials*. Switzerland: Springer.
- Li, Z.G., Meng, H., 2007. *Organic Light-Emitting Materials and Devices*. London: Taylor & Francis.
- Tao, Y.T., Yang, C.L., Qin, J.G., 2011. Organic host materials for phosphorescent organic light-emitting diodes. *Chemical Society Reviews* 40, 2943–2970.
- Tao, Y., Yuan, K., Chen, T., *et al.*, 2014. Thermally activated delayed fluorescence materials towards the breakthrough of organoelectronics. *Advanced Materials* 26, 7931–7958.
- Uoyama, H., Goushi, K., Shizu, K., *et al.*, 2012. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature* 492 (7428), 234–238.
- Xiao, L.X., Chen, Z.J., Qu, B., *et al.*, 2011. Recent progresses on materials for electrophosphorescent organic light-emitting devices. *Advanced Materials* 23, 926–952.
- Yook, K.S., Lee, J.Y., 2012. Organic materials for deep blue phosphorescent organic light-emitting diodes. *Advanced Materials* 24, 3169–3190.

Organic Light-Emitting Diodes/Light Extraction

Jiun-Haw Lee, National Taiwan University, Taipei City, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic light-emitting diode (OLED) consists of organic thin-film stacks with different functions, sandwiched by two electrodes. When the electrical voltage applied to the OLED, electrons and holes are injected from cathode and anode, respectively, into the organic layers. Usually, there are several layers such as hole-injection layer (HIL) and hole-transporting layer (HTL) to carry holes, as well as electron-injection layer (EIL) and electron-transporting layer (ETL) to carry electrons. Electrons and holes recombine in the emitting layer (EML) and form excitons. Excitons relax the energy radiatively to generate photons. There are several optical losses when the light propagates from the OLED to the air.

Due to the difference of refractive index (n) between the organic material (typically $n \sim 1.6$ – 2.0) and the air ($n = 1$), part of photons emitted from the EML are trapped due to total internal reflection (TIR). Typically, the OLED was fabricated on the glass substrate ($n = 1.5$). And a transparent anode (indium tin oxide (ITO), $n = 1.8$ – 2.0) and a reflective cathode (Al) are used. Optical wave tends to propagate at the medium with higher refractive index. Hence, part of light is trapped inside the ITO/organic layers, which is called “waveguiding mode.” And the light inside the glass substrate is called “substrate mode,” to separate the TIR effects at different interfaces.

Note that the carrier mobility of the organic materials for the OLED application are pretty low (typically $< 10^{-2} \text{ cm}^2/\text{Vs}$), and hence total film thickness of the thin-film stacks in an OLED is typically less than 300 nm, which is needed to operate such a device with a reasonable driving voltage ($< 10 \text{ V}$). Such a thickness is comparable to the visible wavelength which results in an interference effect. Typically, we can consider the OLED structure (from anode to cathode) as a “microcavity.” For the case with ITO anode, reflection mainly comes from the reflective cathode (such as Al) and it is called “weak cavity.” On the other hand, a semitransparent anode (such as thin metal) can be employed to replace ITO, which is formed a “strong cavity” with reflective metal cathode. By suitable tuning of the organic layer thicknesses, optical interference effect can be optimized for achieving higher extraction efficiency.

Due to the thin organic layers, the emission dipole is physically close (typically $< 100 \text{ nm}$) to the metal electrode, which results in plasmonic loss. That is, the emission dipole excites the surface plasmon polariton (SPP), which relaxes the energy non-radiatively. A minor, but nonnegligible loss is the material coming from the organic materials and the electrodes. Fig. 1 shows the external quantum efficiency (EQE) versus the ETL thickness. The oscillatory behavior of the outcoupled mode results from the microcavity effect with ETL thickness modulation, and the substrate mode shows similar trend. Surface plasmon mode decreases with increasing the ETL thickness, because the distance between emission dipole and metal cathode increases. However, on the other hand, the substrate mode increases with thicker ETL because it supports more optical modes.

In the following, we will introduce various schemes for improving the light extraction in an OLED, starting from the enhancement of substrate mode. By applying some structures to destroy the TIR at the (glass) substrate and the air, extraction efficiency can be significantly improved. Note that OLED can be used for lighting and display applications. Image quality is an important parameter for a display, which may decrease due to these structures, and careful design should be made. Besides, enhancement ratio of the outside (substrate) mode cannot fully decoupled with the internal (waveguiding and plasmonic) modes,

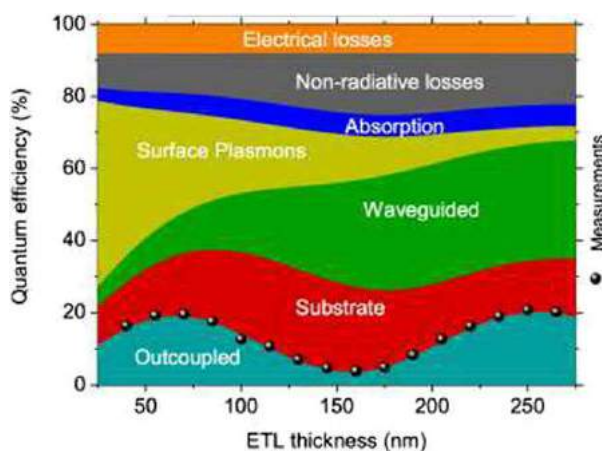


Fig. 1 Electron-transporting layer (ETL) thickness vs. distribution of different modes (outcoupled, substrate, waveguided, surface plasmons...) in an organic light-emitting diode (OLED). Black dots shows the measurement results. Maximum external quantum efficiency (EQE) (outcoupled mode) is 21% with ETL = 250 nm. Reproduced from Meerheim, R., Furno, M., Hofmann, S., Lüssem, B., Leo, K., 2010. Quantification of energy loss mechanisms in organic light-emitting diodes. *Applied Physics Letters* 97, 253305. Copyright (2010) with permission from American Institute of Physics.

which needs to be considered simultaneously. And we will move to the discussions of the internal modes. Again, at the ITO/organic and (glass) substrate interfaces, we can do some modulations of refractive index and structures for reducing the TIR. To reduce the plasmonic loss, a straightforward method is to keep the emission dipole away from the metal surface. And an alternative is to introduce nanostructure to couple the SPP back to the air mode.

Light Extraction of the Substrate Mode

To extract the light rays in the planar glass substrate, a straightforward method is to use a hemisphere lens structure, which has been used in semiconductor light-emitting diodes (LEDs).

Note there is a basic difference between semiconductor LED and OLED. LED was cut into a small piece, called “LED die” with the dimension $\sim 1\text{ mm} \times 1\text{ mm}$ for the following lens-like encapsulation process. On the other hand, a unique advantage for the OLED is the large-sized fabrication with the flat emission feature. For an OLED lighting panel (such as $10\text{ cm} \times 10\text{ cm}$), a hemisphere macro lens with radius much larger than 10 cm is not practical. Instead, miniaturized microlens is an alternative to extract the substrate mode. One may use a microlens array film (MAF) attached on the glass substrate to improve the OLED efficiency. Fig. 2(a) shows the scanning electron microscope (SEM) picture of MAF. In Fig. 2(b), MAF was applied on the top of OLED, and a brighter region was clearly seen. EQE was improved by 50% from 9.5 to 14.5%. The function of the lens structure is to redirect the light beams out of the OLED, which can be also accomplished by scattering medium. By incorporation of the material with different refractive index into the matrix film, it can be applied to the OLED which improved the light extraction with the flat out-surface, which is important for display application with touch-panel function. As shown in Fig. 2(c), a polyimide (PI) film (high-index matrix) with air voids ($n=1$) increased the EQE by 65%.

The basic idea of microlens structure for increasing light extraction of the OLED is to couple out the light at the substrate at larger incident angle. Fig. 3(a) shows the lit-on picture of an OLED pixel with the size of $100\text{ }\mu\text{m} \times 100\text{ }\mu\text{m}$. When applying MAF (with the pitch of $50\text{ }\mu\text{m}$) on the OLED as shown in Fig. 3(b), photons were extracted from the surrounding of the pixel region with larger incident angle (larger than critical angle between glass substrate and the air) which increased the light outcoupling. At the same time, note that there were some dark regions at the area where the emission area and the microlenses overlapped. Because the photons with smaller incident angle can couple out of the OLED without microlenses, the incorporation of microstructure on the pixel region actually decreased the outcoupled photons at this region. Hence, a concept of patterned microlens array was proposed. By removing the pixel on the top of the pixel region, light extraction at this region was maintained, while increasing the EQE from the contribution of surrounding areas, as shown in Fig. 3(c).

For using OLED in display application, not only light extraction, but also image quality should be taken into consideration. As shown in Fig. 3(b), light extraction from the region outside the pixel implied image blurring when using as an OLED display and it decreased the image quality. Fig. 4(a) shows the Lena image on the OLED display, and Fig. 4(d) shows the enlarged picture of the left eye region. Fig. 4(b) and (e) shows the image taken with MAF attachment. Although the luminance increased by 23%, the image quality index was decreased to 74%. Fig. 4(c) and (f) are the image with patterned-MAF. One can see that the image is better with patterned-MAF, with the image quality index of 94%. Besides, the luminance of the OLED display increased by 52%.

Light Extraction of the Waveguiding Mode

To extract the waveguiding mode, it is straightforward to use some structure inside the glass substrate, close to the OLED device. Fig. 5(a) shows a low-index structure which was inserted between the ITO substrate and the organic layers. The effective refractive

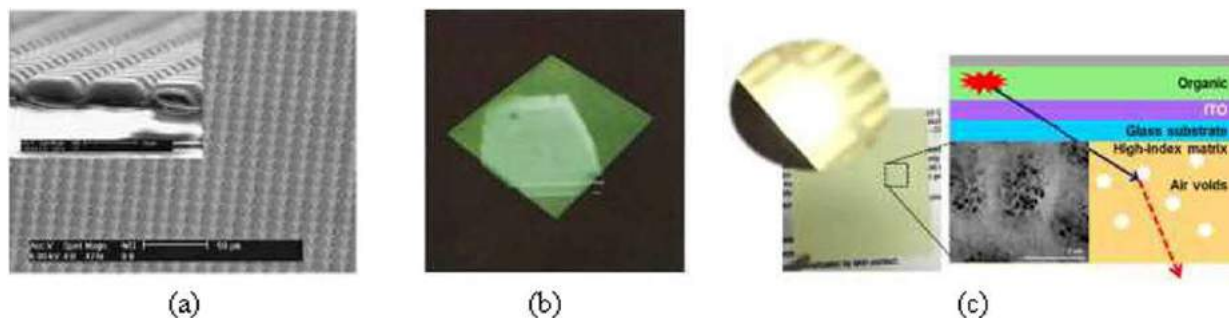


Fig. 2 (a) Scanning electron microscope (SEM) picture of the microlens array film (MAF), (b) picture of green organic light-emitting diode (OLED) emission with partially covered by MAF, and (c) light extraction by polyimide (PI) scattering film. (a,b) Reproduced from Moller, S., Forrest, S.R., 2002. Improved light out-coupling in organic light emitting diodes employing ordered microlens arrays. *Journal of Applied Physics* 91, 3324–3327. Copyright (2002) with permission from American Institute of Physics. (c) Reproduced from Koh, T.W., Spechler, J.A., Lee, K.M., Arnold, C.B., Rand, B.P., 2015. Enhanced outcoupling in organic light-emitting diodes via a high-index contrast scattering Layer. *ACS Photonics* 2, 1366 – 1372. Copyright (2015) with permission from American Chemistry Society.

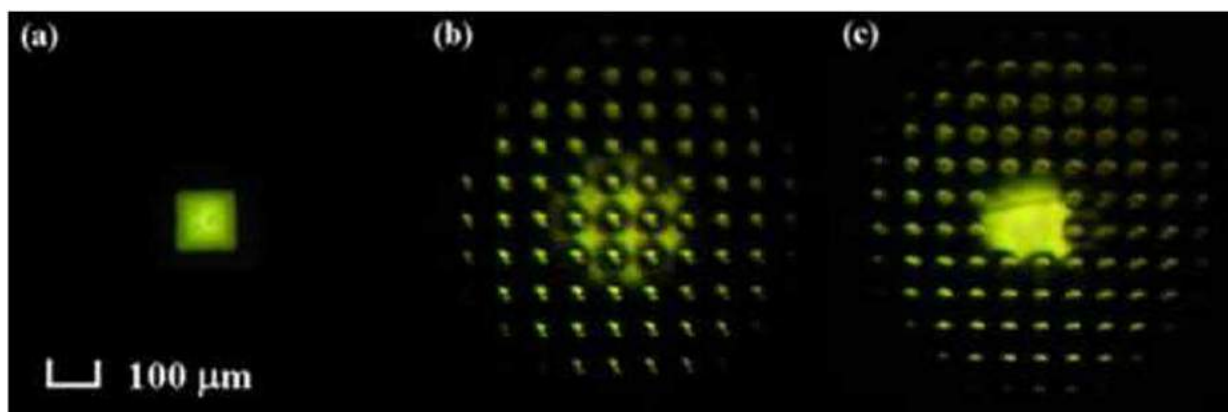


Fig. 3 (a) Optical microscope (OM) pictures of an organic light-emitting diode (OLED) emission. (b) and (c) shows the images with microlens array film (MAF) and patterned-MAF attachment on the OLED. Reproduced from Lin, H.Y., Chen, K.Y., Ho, Y.H., *et al.*, 2010. Luminance and image quality analysis of an organic electroluminescent panel with a patterned microlens array attachment. *Journal of Optics* 12, 085502. Copyright (2010) with permission from IOP Publishing.

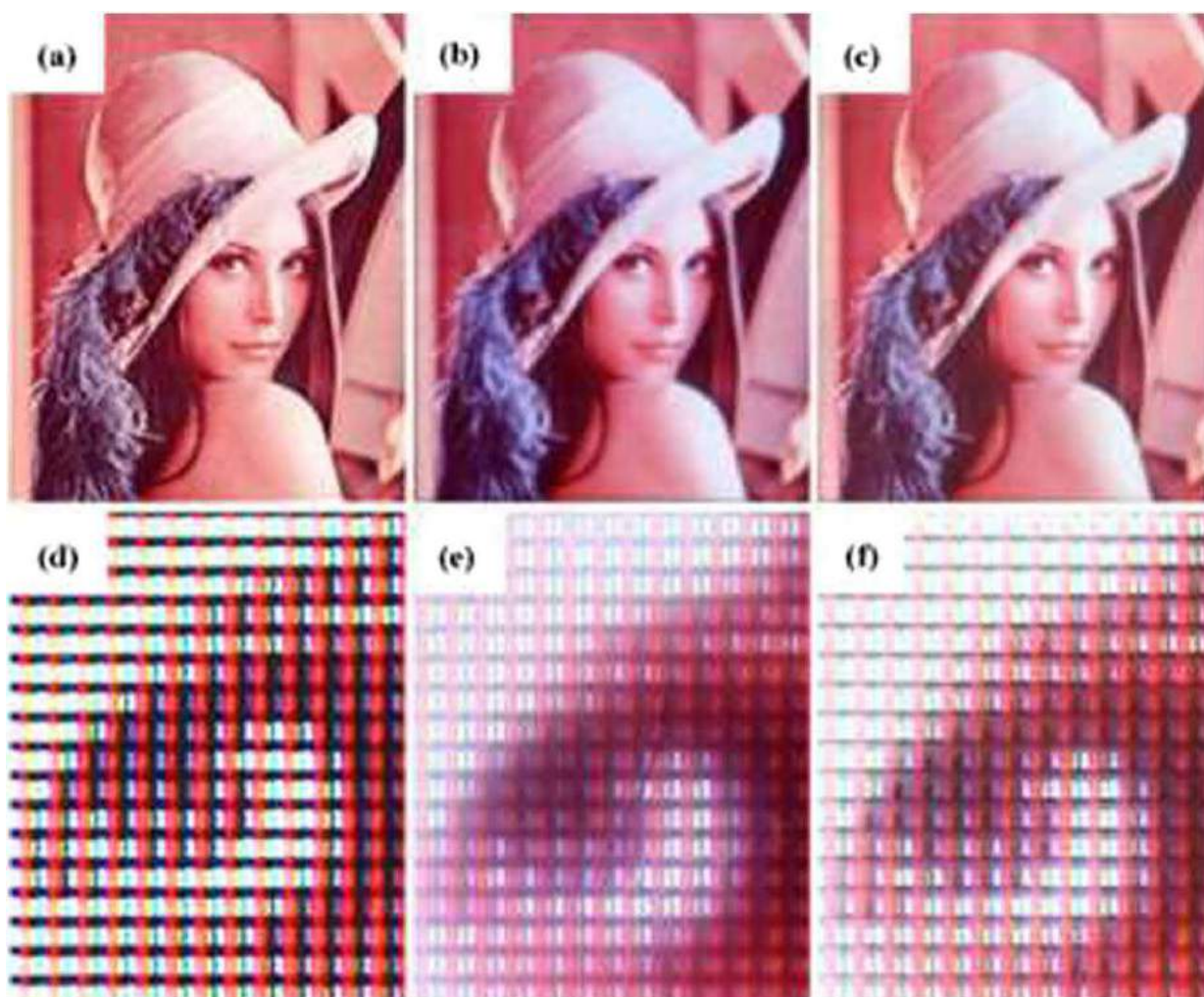


Fig. 4 (a) Lena picture on an organic light-emitting diode (OLED) display, with (b) microlens array film (MAF), and (c) patterned-MAF. (d)–(f) Are enlarged images of (a), (b), and (c), respectively. Reproduced from Lin, H.Y., Chen, K.Y., Ho, Y.H., *et al.*, 2010. Luminance and image quality analysis of an organic electroluminescent panel with a patterned microlens array attachment. *Journal of Optics* 12, 085502. Copyright (2010) with permission from IOP Publishing.

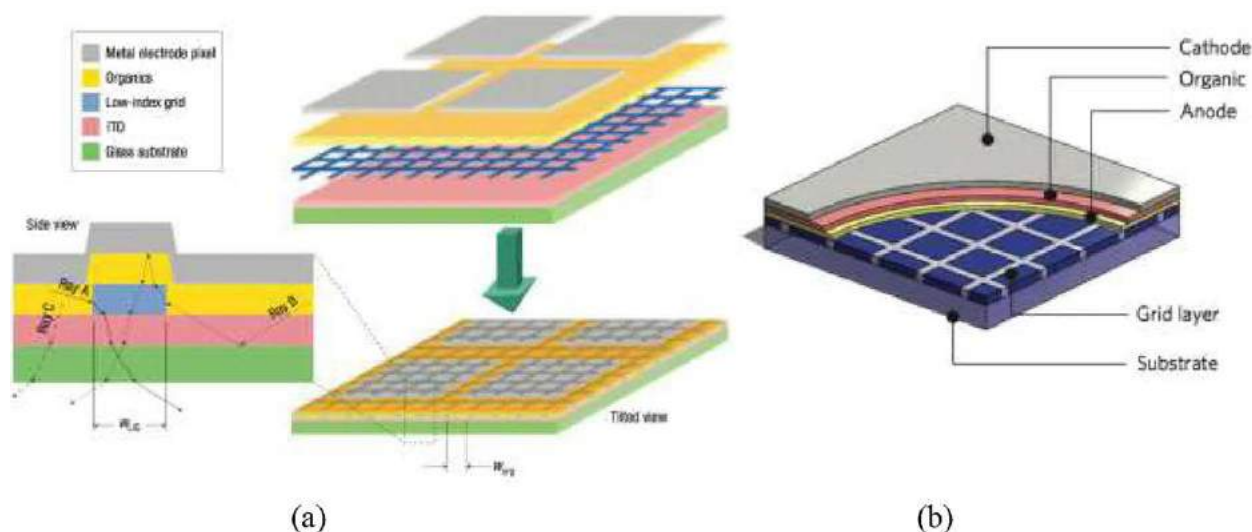


Fig. 5 (a) Schematic diagram of low-index grid (LIG) on the top of the indium tin oxide (ITO) layer and (b) planar grid layer structure between the glass substrate and the ITO. (a) Reproduced from Sun, Y., Forrest, S.R., 2008. Enhanced light out-coupling of organic light-emitting devices using embedded low-index grids. *Nature Photonics* 2, 483–487. Copyright (2008) with permission from Nature Publishing Group. (b) Reproduced from Qu, Y., Sloatsky, M., Forrest, S.R., 2015. Enhanced light extraction from organic light-emitting devices using a sub-anode grid. *Nature Photonics* 9, 758–763. Copyright (2015) with permission from Nature Publishing Group.

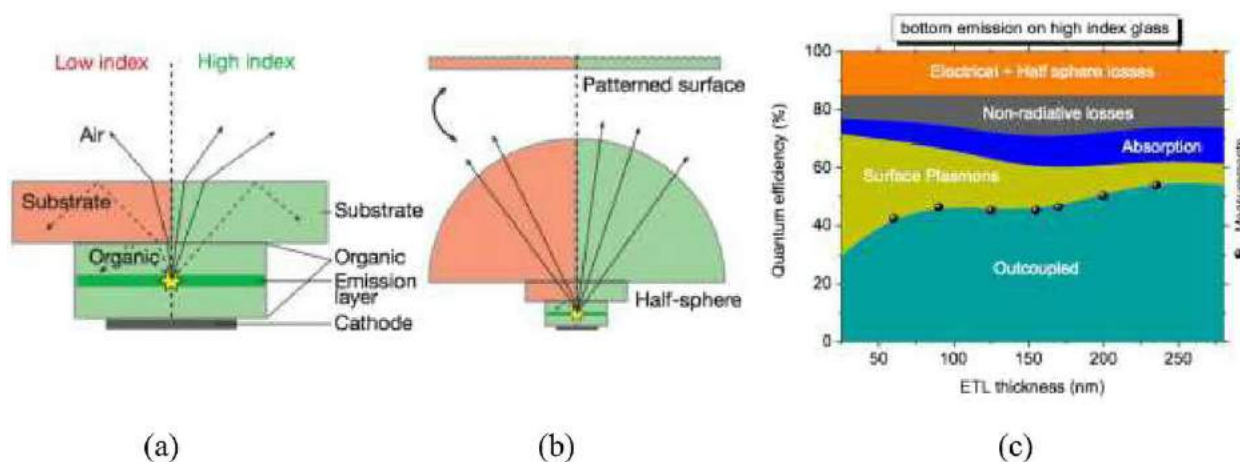


Fig. 6 (a) Schematic diagram of the organic light-emitting diode (OLED) on low- and high-index substrate, (b) with marolens and microlenses, and (c) electron-transporting layer (ETL) thickness vs. distribution of different modes in an OLED with high-index substrate and macro lens. (a,b) Reproduced from Reineke, S., Lindner, F., Schwartz, G., *et al.*, 2009. White organic light-emitting diodes with fluorescent tube efficiency. *Nature* 459, 234–238. Copyright (2009) with permission from Nature Publishing Group. (c) Reproduced from Meerheim, R., Furno, M., Hofmann, S., Lüssem, B., Leo, K., 2010. Quantification of energy loss mechanisms in organic light-emitting diodes. *Applied Physics Letters* 97, 253305. Copyright (2010) with permission from American Institute of Physics.

index of the waveguiding mode is decreased, and photons trapped in the waveguiding mode (light rays A and B in Fig. 5(a) with larger incident angle) can be effectively coupled out. By applying such low-index grid (LIG) in an OLED, 32% enhancement in EQE was observed. By using LIG and microlenses simultaneously, 130% improvement in EQE was achieved. However, the loss of contact area between ITO and organic layer reduced the luminance. Hence, one may insert such a grid structure beneath the ITO layer, as shown in Fig. 5(b). With sequential processes of deposition, etching, and planarization, embedded structure on the top of glass substrate was fabricated for extracting the waveguiding mode. Again, EQE was enhanced from 15 to 40% when using the grid structure to couple out the waveguiding mode, combined with the immersion of the OLED and photodetector inside the index-matching oil to extract all the glass substrate mode.

As discussed before, glass substrate mode and waveguiding mode both come from the TIR, with the refractive index of 1.5 and ~ 1.6 –2.0, respectively. By using a glass substrate with high refractive index ($n=1.8$), the waveguiding mode no longer exists. As

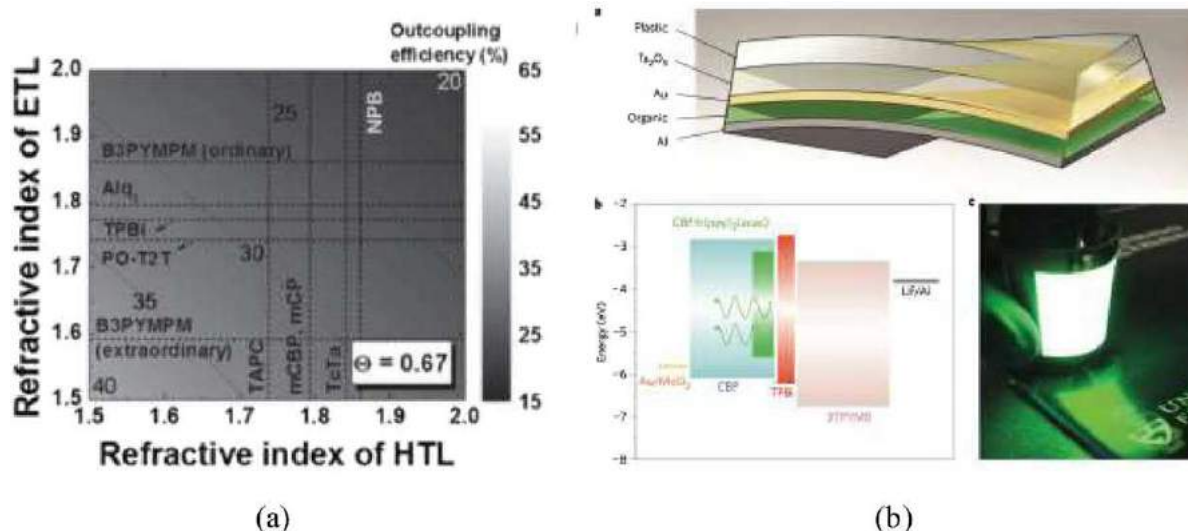


Fig. 7 (a) External quantum efficiency (EQE) values with different hole-transporting layer (HTL) and electron-transporting layer (ETL) refractive indexes and (b) indium tin oxide (ITO)-free flexible organic light-emitting diode (OLED) with Au as the anode and Ta_2O_5 to tune the optical field, together with the device structure and OLED emission pictures. (a) Reproduced from Shin, H., Lee, J.H., Moon, C.K., *et al.*, 2016. Sky-blue phosphorescent OLEDs with 34.1% external quantum efficiency using a low refractive index electron transporting layer. *Advanced Materials* 28, 4920–4925. Copyright (2016) with permission from Wiley-VCH Verlag GmbH & Co. (b) Reproduced from Wang, Z.B., Helander, M.G., Qiu, J., *et al.*, 2011. Unlocking the full potential of organic light-emitting diodes on flexible plastic. *Nature Photonics* 5, 753–757. Copyright (2011) with permission from Nature Publishing Group.

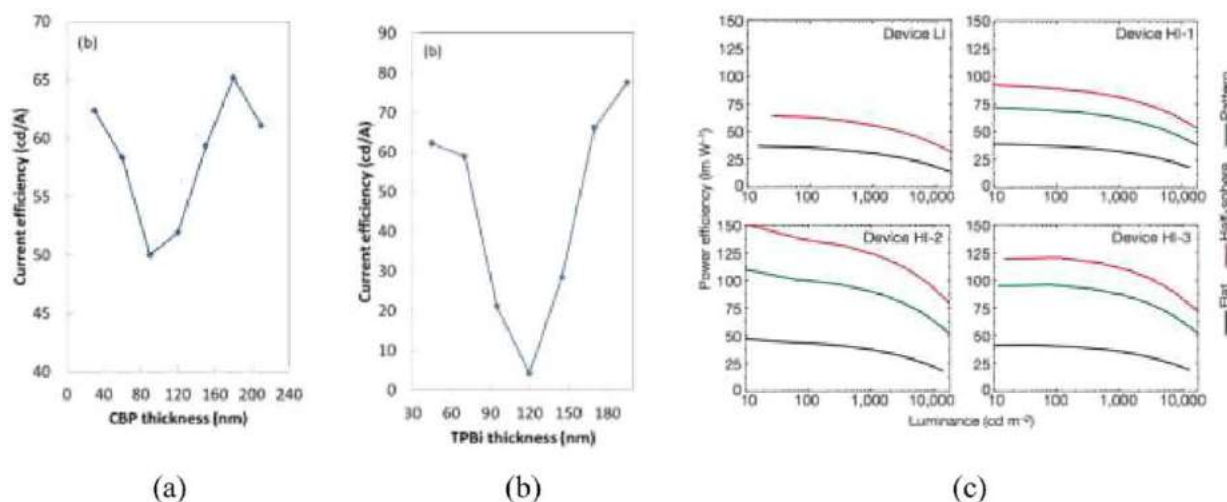


Fig. 8 Current efficiency vs. (a) hole-transporting layer (HTL) and (b) electron-transporting layer (ETL) thickness. (c) Power efficiency vs. luminance for different devices without and with outcoupling lenses. (a,b) Reproduced from Zhang, Y., Aziz, H., 2016. Insights into charge balance and its limitations in simplified phosphorescent organic light-emitting devices. *Organic Electronics* 30, 76–82. Copyright (2016) with permission from Elsevier. (c) Reproduced from Reineke, S., Lindner, F., Schwartz, G., *et al.*, 2009. White organic light-emitting diodes with fluorescent tube efficiency. *Nature* 459, 234–238. Copyright (2009) with permission from Nature Publishing Group.

shown in Fig. 6(a), one can see that when using a high-index substrate, there is no TIR at the OLED device and glass substrate. Then, with the incorporation of macro lens or patterned surface (such as microlenses), all the contributions from waveguiding and substrate modes can be extracted, as shown in Fig. 6(b). EQE=54% can be achieved when ETL=235 nm, as shown in Fig. 6(c).

Thinking from the other side, it is possible to reduce the effective refractive index of the organic materials in the OLED device for reducing the waveguiding mode. As shown in Fig. 7(a), by decreasing the refractive indexes of the HTL and ETL from 2.0 to 1.4, the EQE can be improved from 20 to 40%. One may also use a complex anode structure $\text{Ta}_2\text{O}_5/\text{Au}$ on the flexible substrate ($n=1.6$) to couple out the waveguiding light into the substrate mode. Then, with the use of the lens structure, 63% EQE was achieved.

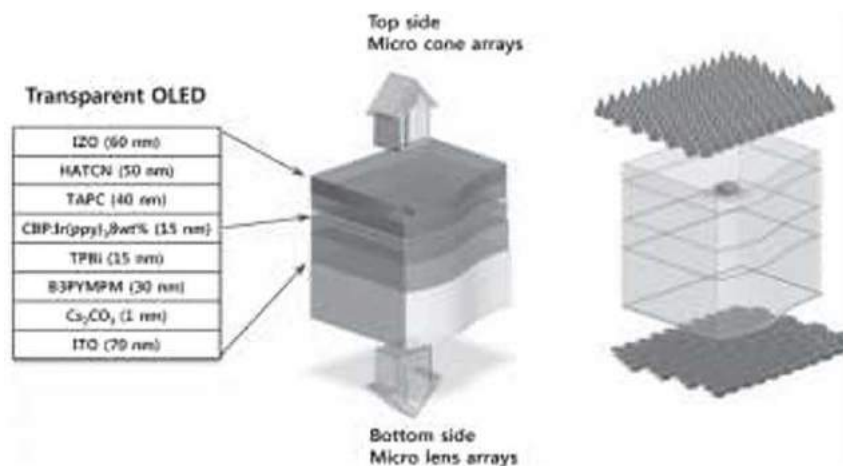


Fig. 9 Device structure of a transparent organic light-emitting diode (OLED) with metal oxide electrodes and micro-structures. Reproduced from Kim, J.B., Lee, J.H., Moon, C.K., Kim, S.Y., Kim, J.J., 2013. Highly enhanced light extraction from surface plasmonic loss minimized organic light-emitting diodes. *Advanced Materials* 25, 3571–3577. Copyright (2013) with permission from Wiley-VCH Verlag GmbH & Co.

Light Extraction of the Plasmonic Mode

Let us go back to the microcavity effect in an OLED in **Fig. 1**. For the bottom emission OLED on the glass substrate with ITO anode, the distance of the emission dipole to the reflective cathode is crucial to the light extraction efficiency. And hence modulation of ETL thickness changes the efficiency more effectively than that of HTL thickness. As shown in **Fig. 8(a)**, when varying the thickness of HTL (CBP, 4,4'-bis(carbazol-9-yl)biphenyl), the current efficiency changed between ~ 50 and ~ 65 cd/A. On the other hand, current efficiency changed from < 10 cd/A to ~ 80 cd/A with increasing the ETL (TPBi, 2,2',2''-(1,3,5-benzinetriyl)-tris(1-phenyl-1-H-benzimidazole)) thickness, in **Fig. 8(b)**. Besides, one can also notice that with increasing the ETL thickness, the efficiency can be higher than the first peak which was due to the reduction of plasmonic loss, as shown in **Fig. 1**. **Fig. 8(c)** shows the power efficiency versus luminance for different OLEDs. Device L1 is a common OLED structure on glass substrate. With the incorporation of half-sphere macro lens, power efficiency over 50 lm/W can be achieved. By applying a high-index substrate for coupling out the substrate mode, the efficiency can be boosted up over 80 lm/W, in device HI-1. For devices HI-2 and HI-3, thick ETL was employed, which resulted in a much higher power efficiency > 150 lm/W and > 110 lm/W, respectively.

Because the plasmonic loss mainly came from the metal electrode, it is possible to reduce this mode by using transparent conductive oxides as both the anode and cathode materials. **Fig. 9(a)** shows the device structure. ITO and indium zinc oxide (IZO) was used as the cathode and anode, respectively. Light emission from both sides and it is called transparent OLED. With the incorporation of micro- and macrostructures for extracting the substrate mode, EQE was increased from 18.1 to 47.3% and 62.9%, respectively, due to the reduction of plasmonic and substrate modes.

The origin of the non-radiative behavior of the plasmonic mode because the dispersion curve (w - k) of surface plasmon mode has no intersection with the air light cone, as shown in **Fig. 10(a)**. However, with the introduction of the periodic nanostructure such as grating, it is possible to provide an extra in-plane wave-vector which can convert the plasmonic mode radiatively. Note that the nanostructure extracts not only the plasmonic mode, but also the waveguiding mode. To distinguish these two effects, one can use polarization dependent measurement. Emission from the surface plasmonic mode is only p-polarized. On the other hand, enhanced emission from s-polarization resulted from the outcoupling of the waveguiding mode by the nanostructure grating. Hence, a polarizer can be put in front of the detector for separating these two effects. As shown in **Fig. 10(c)** and **(d)**, p- and s-polarized emission at different viewing angles were obtained, which came from the light extraction of plasmonic and waveguiding mode, respectively.

Although the uniform nanostructure provides an efficient way to extract the plasmonic and waveguiding mode, one can note that wavelength dependence is pretty strong, which means the color of this OLED is varied from different viewing angles and it is not suitable for display and lighting applications. A quasi-periodic structure was proposed by buckling fabrication. Due to different thermal expansion coefficient between poly(dimethylsiloxane) (PDMS) and Al, a quasi-periodic structure can be obtained, as shown in **Fig. 11**. With one-, two-, and three-times buckling, the structure with longer period (saying 900 nm) can be enhanced, with the major peak at 410 nm. 2.2- and 2.9-times improvement in power efficiency was achieved with two- and three-time buckling structure on an OLED without significantly changing the emission profile.

Thinking from the molecule viewpoint, it acts like an oscillation dipole for generating electromagnetic wave, as shown in **(Fig. 12a)**. Here, the upper medium is metal, and surface plasmon was excited due to the electrical dipole. Depending on the Θ value, one can divide the contribution from vertical ($\Theta = 0$ degree) and horizontal ($\Theta = 90$ degree) dipoles. Note that SP mode was excited by TM waves, and hence radiation of horizontal dipoles coupled much less into plasmonic mode compared

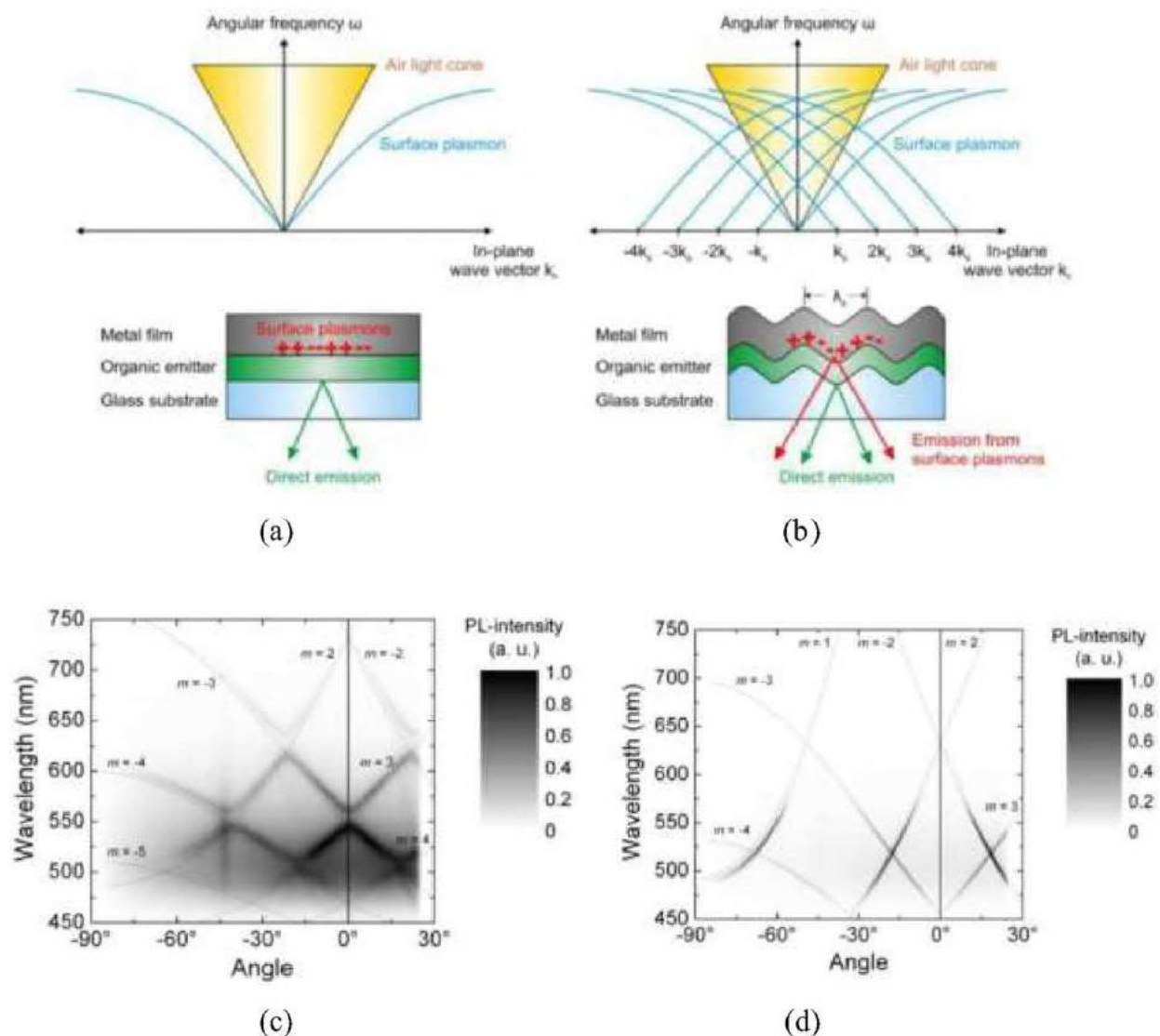


Fig. 10 Dispersion relation and device structure of a (a) planar and (b) grating structure. (c) and (d) shows the photoluminescence (PL) spectra at different viewing angles of p- and s-polarization. Reproduced from Frischeisen, J., Niu, Q., Abdellah, A., *et al.*, 2011. Light extraction from surface plasmons and waveguide modes in an organic light-emitting layer by nanoimprinted gratings. *Optics Express* 19, A7–A19. Copyright (2011) with permission from Optical Society of America.

to vertical dipole. By suitable choosing the organic material, it is possible to “align” the organic emitter horizontally and achieve the high EQE. Ratio between horizontal and vertical dipoles can be obtained from photoluminescence (PL) intensity at different viewing angles, as shown in Fig. 12(b). Black and yellow dashed lines show the cases with pure horizontal and isotropic dipole emission, respectively. With higher horizontal dipole ratio (93% for Pt(fppz)₂, Pt(II) bis(3-(trifluoromethyl)-5-(2-pyridyl)-pyrazolate)), the EQE can be as high as 38.8%, as shown in Fig. 12(c). Note that it was a planar structure without any lens structure. Substrate was a typical glass. And the thin-film thickness from the emitter to the metal electrode was less than 100 nm. Hence, such a high EQE came from the reduction of plasmonic mode due to the alignment of horizontal dipole.

Summary

Here, we introduced the problem of light extraction in an OLED which limits the outcoupling efficiency $\sim 20\%$ in a conventional bottom emission configuration, mainly due to TIR and plasmonic loss. Various methods were introduced to improve the light extraction, with different structures and refractive index modulation. For the recent stage, experimental and simulation results show that $\text{EQE} = \sim 60\%$ is achievable, and further improvement is still on-going.

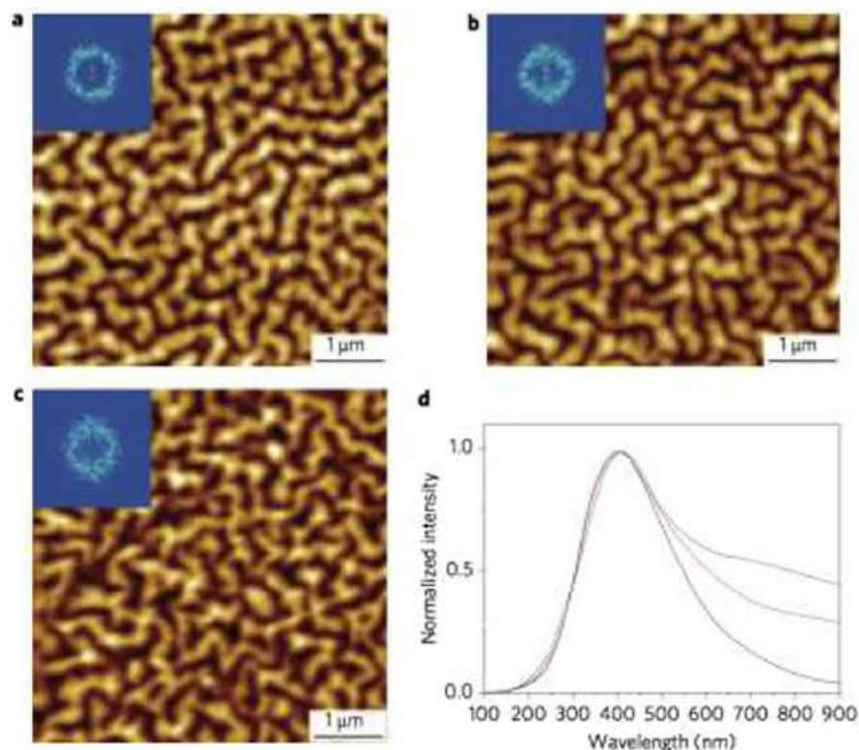


Fig. 11 Atomic-force microscopy (AFM) images for (a) one-, (b) two-, and (c) three-times buckling structure. (d) Power spectra of different buckling structures. Reproduced from Koo, W.H., Jeong, S.M., Araoka, F., *et al.*, 2010. Light extraction from organic light-emitting diodes enhanced by spontaneously formed buckles. *Nature Photonics* 4, 222–226. Copyright (2010) with permission from Nature Publishing Group.

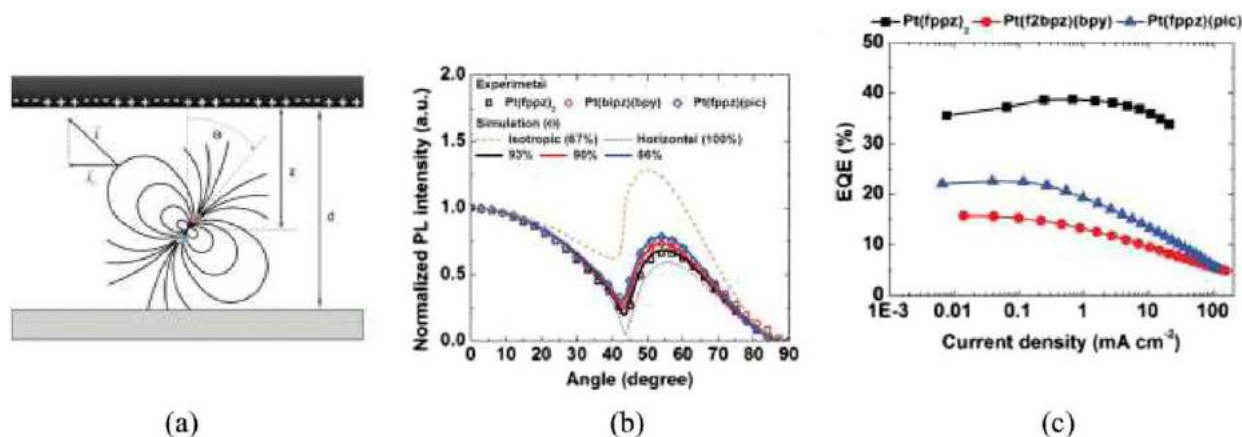


Fig. 12 (a) Schematic diagram of emission dipole in an environment. (b) Photoluminescence (PL) intensity at different viewing angles for different emitters. (c) External quantum efficiency (EQE) vs. current density of organic light-emitting diodes (OLEDs) with different emitters. (a) Reproduced from Brütting, W., Frischeisen, J., Schmidt, T.D., Scholz, B.J., Mayr, C., 2013. Device efficiency of organic light-emitting diodes: Progress by improved light outcoupling. *Physica Status Solidi A* 210, 44–65. Copyright (2013) with permission from Wiley-VCH Verlag GmbH & Co. (b,c) Reproduced from Kim, K.H., Liao, J.L., Lee, S.W., *et al.*, 2016. Crystal organic light-emitting diodes with perfectly oriented non-doped Pt-based emitting layer. *Advanced Materials* 28, 2526–2532. Copyright (2016) with permission from Wiley-VCH Verlag GmbH & Co.

See also: OLED/Introduction

Further Reading

Brütting, W., Frischeisen, J., Schmidt, T.D., Scholz, B.J., Mayr, C., 2013. Device efficiency of organic light-emitting diodes: Progress by improved light outcoupling. *Physica Status Solidi A* 210, 44–65.

- Frischeisen, J., Niu, Q., Abdellah, A., *et al.*, 2011. Light extraction from surface plasmons and waveguide modes in an organic light-emitting layer by nanoimprinted gratings. *Optics Express* 19, A7–A19.
- Kim, J.B., Lee, J.H., Moon, C.K., Kim, S.Y., Kim, J.J., 2013. Highly enhanced light extraction from surface plasmonic loss minimized organic light-emitting diodes. *Advanced Materials* 25, 3571–3577.
- Kim, K.H., Liao, J.L., Lee, S.W., *et al.*, 2016. Crystal organic light-emitting diodes with perfectly oriented non-doped Pt-based emitting layer. *Advanced Materials* 28, 2526–2532.
- Koo, W.H., Jeong, S.M., Araoka, F., *et al.*, 2010. Light extraction from organic light-emitting diodes enhanced by spontaneously formed buckles. *Nature Photonics* 4, 222–226.
- Lin, H.Y., Chen, K.Y., Ho, Y.H., *et al.*, 2010. Luminance and image quality analysis of an organic electroluminescent panel with a patterned microlens array attachment. *Journal of Optics* 12, 085502.
- Meerheim, R., Furno, M., Hofmann, S., Lüssem, B., Leo, K., 2010. Quantification of energy loss mechanisms in organic light-emitting diodes. *Applied Physics Letters* 97, 253305.
- Moller, S., Forrest, S.R., 2002. Improved light out-coupling in organic light emitting diodes employing ordered microlens arrays. *Journal of Applied Physics* 91, 3324–3327.
- Qu, Y., Slightsky, M., Forrest, S.R., 2015. Enhanced light extraction from organic light-emitting devices using a sub-anode grid. *Nature Photonics* 9, 758–763.
- Reineke, S., Lindner, F., Schwartz, G., *et al.*, 2009. White organic light-emitting diodes with fluorescent tube efficiency. *Nature* 459, 234–238.
- Shin, H., Lee, J.H., Moon, C.K., *et al.*, 2016. Sky-blue phosphorescent OLEDs with 34.1% external quantum efficiency using a low refractive index electron transporting layer. *Advanced Materials* 28, 4920–4925.
- Sun, Y., Forrest, S.R., 2008. Enhanced light out-coupling of organic light-emitting devices using embedded low-index grids. *Nature Photonics* 2, 483–487.
- Wang, Z.B., Helander, M.G., Qiu, J., *et al.*, 2011. Unlocking the full potential of organic light-emitting diodes on flexible plastic. *Nature Photonics* 5, 753–757.
- Zhang, Y., Aziz, H., 2016. Insights into charge balance and its limitations in simplified phosphorescent organic light-emitting devices. *Organic Electronics* 30, 76–82.

Conjugated Polymer-Based Solar Cells

Gang Li, The Hong Kong Polytechnic University, Hong Kong, China

Widhya Budiawan, Research Center for Applied Sciences, Academia Sinica, Taipei, Taiwan (ROC); National Tsing-Hua University, Hsinchu, Taiwan (ROC); and Academia Sinica and National Tsing-Hua University, Hsinchu, Taiwan (ROC)

Pen-Cheng Wang, National Tsing-Hua University, Hsinchu, Taiwan (ROC)

Chih Wei Chu, Research Center for Applied Sciences, Academia Sinica, Taipei, Taiwan (ROC)

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

Carrier mobility [$\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$]

Light intensity/power density [W m^{-2}]

Photon flux [$\text{m}^{-2} \text{s}^{-1} \text{nm}^{-1}$]

Short circuit current density [mA cm^{-2}]

Solar irradiance [$\text{W m}^{-2} \text{nm}^{-1}$]

Introduction

Photovoltaic, which directly harvest energy from sunlight into electricity, is being increasingly recognized as an important component for future global energy production. Organic semiconductor and particularly conjugated polymers-based solar cells (OPVs) have attracted world-wide attention in the past decade, because of their distinct features such as low cost production, lightweight, flexibility, electronic functional versatility from chemical synthesis, roll-to-roll compatible large area fabrication, (semi)transparency and potentially environmental friendly in fabrication, etc. The performance of OPV has been greatly improved in last decade, now reaching over 12% as shown in [Fig. 1\(a\)](#) (Source: National Renewable Energy Laboratory (NREL)). These achievements come from comprehensive research – from materials discovery, nanostructure morphology control in photon absorbing active layer, to interface materials innovation and device architecture engineering, device physics, to fabrication technologies. Strictly speaking, OPV can be divided into polymer and small molecule type. This article only focus on polymer solar cells and use the term OPV to represent them.

Polymer OPVs are based on organic conjugated polymers, which are carbon-based materials whose backbones are comprised of alternating C–C and C=C bonds (carbon–carbon triple bonds may also be possible in rare cases), shown in [Fig. 1\(b\)](#). In the alternating C–C and C=C bonds molecular structure, through sp^2 hybridization, it leads to a linear series of overlapping p_z orbitals, and creates a conjugated chain of delocalized electron density. The π electron delocalization along the conjugated backbone – i.e., the interaction of the π electrons – is responsible for their semiconducting properties. As the conjugation length (or the electron delocalization length) increases, the intrinsic material energy levels becomes closely spaced, and similar to inorganic semiconductors, a quasi “band” structure will form in organic semiconductors, including conjugated polymers. In organic semiconductors, the lowest unoccupied molecular orbital (LUMO) is equivalent to the conduction band minimum (CBM) in inorganic semiconductors, while the highest occupied molecular orbital (HOMO) is equivalent to the valence band maximum (VBM) in inorganic semiconductors ([Fig. 1\(c\)](#)). The primary photoexcitation in the organic semiconductors are bounded Frenkel excitons (electron–hole pair), which is very different from the free electrons/holes in crystalline inorganic semiconductors. The excitons have a typical lifetime of ~ 100 ps to 1 ns, decay either radiatively or non-radiatively. These differences make the working mechanism in organic polymer solar cells very different from the inorganic solar cells.

The high performance polymer solar cells are now dominated by BHJ structure, with conjugated polymer as electron donors, mixed with fullerene or nonfullerene electron acceptors.

This article aims at a concise review of polymer organics solar cell (OSC) development, and discuss the possible strategy for future improvement in polymer solar cell.

Basic Principles of OSC

The main structure of OPV is an organic material(s)-based active layer(s) sandwiched between a cathode and an anode. The active layers consist of material with low ionization potential (IP) as electron donor and material with high electron affinity (EA) as electron acceptor material. Various materials are used as either donor or acceptor including polymer, fullerene derivatives and small molecules. Furthermore, donor material is hole rich which ideally make ohmic contact with anode, whereas acceptor material contact with the cathode.

In the simplest picture of organic solar cell operation principle, the process of the photon-to-electron conversion can be divided into four basic steps ([Fig. 2\(a\)](#)) each with efficiencies – photon absorption (η_A), exciton diffusion (η_{ED}), charge carrier transfer (η_{CT}), and charge carrier collection (η_{CC}). The detail of every step is discussed as below.

1. Photon absorption

The organic material semiconductor only absorbs photon that higher or equal to its bandgap. As consequence, only material with low bandgap is suitable for active layer. In this step, photon was absorbed then electrons in donor (/acceptor) material are

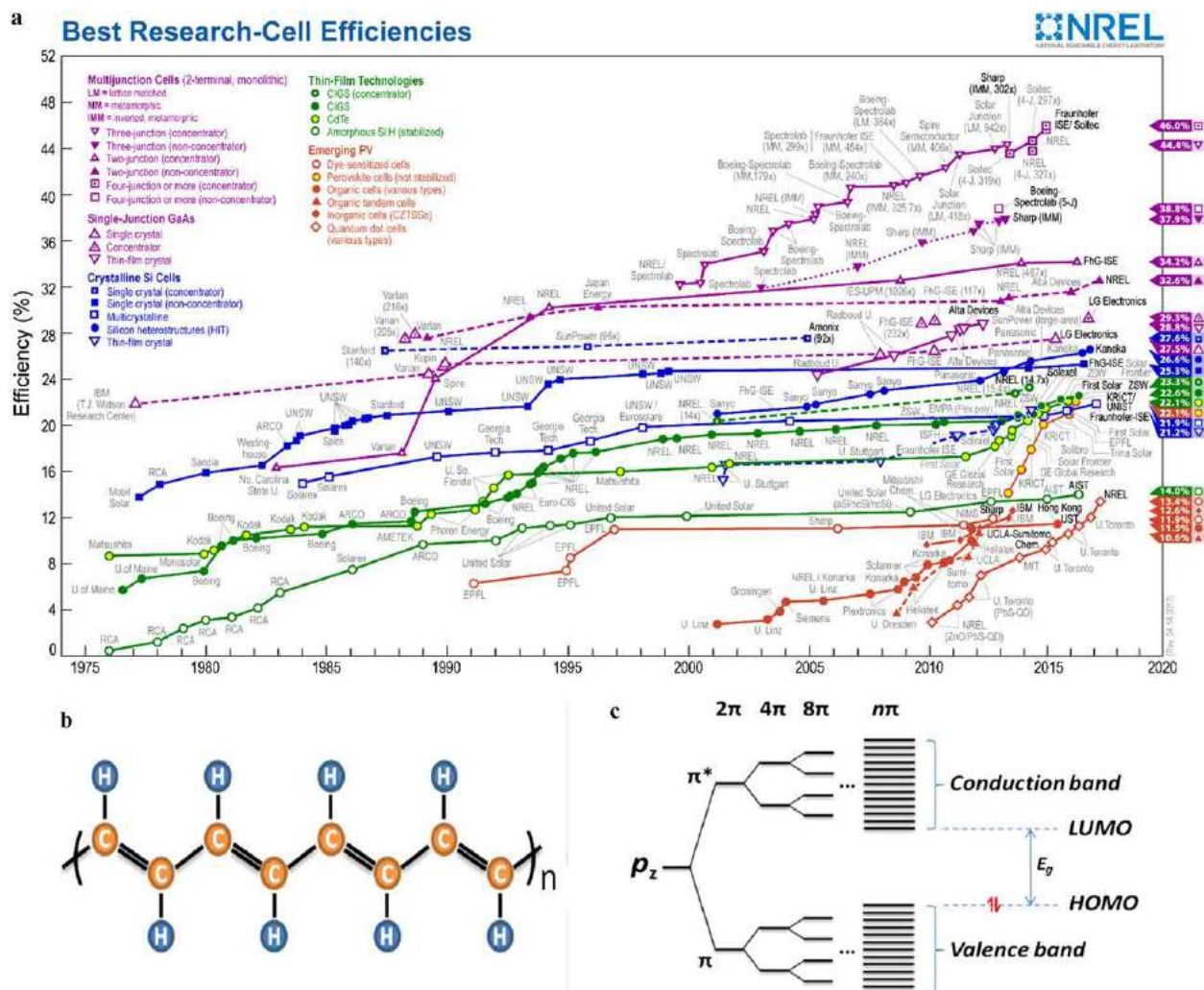


Fig. 1 (a) Efficiency-chart of best solar cell summarized by National Renewable Energy Laboratory (<https://www.nrel.gov/pv/assets/images/efficiency-chart.png>); (b) alternating C–C and C=C bonds in conjugated polymer backbone; (c) energy diagram in organic semiconductors via π electron delocalization – π conjugation. HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbital.

excited to the LUMO level and created holes in the HOMO level. These electron and hole pair are bound together by attractive coulomb interaction and are called excitons. Furthermore, the absorption efficiency (η_A) of active layer is determined by the absorption spectra of the organic material, their thickness, as well as the device architecture.

2. Exciton diffusion

The generated excitons in photon absorption process move via diffusion. This exciton should transport to the donor–acceptor interface within its lifetime to convert to charge carrier. The morphology plays significant role on exciton diffusion (η_{ED}).

3. Charge transfer

The room temperature thermal energy of 26 meV is an order smaller than the binding energy of tightly bound Frenkel excitons in organic semiconductors. Therefore to generate “free” electron and hole polarons as charge carriers, an exciton needs to reach the donor–acceptor interface where the molecular energy level difference will provide the driving force for exciton dissociation into a geminate pair first and overcome the monomolecular process called geminate recombination. Only after this process that free charge carriers can be generated which ultimately contribute to photocurrent.

4. Charge collection

From the D/A interface, the free charge carrier needs be transported (hole in donor, electron in acceptor material) to the organic/electrode interfaces. Again nano-morphology in BHJ configuration is critical. After the charge carriers transport to the active layer/electrode interface, they can finally be extracted from the device in order to yield photocurrent. Buffer layers (hole transport layer (HTL) and electron transport layer (ETL)) are usually necessary and critical to aid hole or electron collection by reducing the energy barriers required to transport carriers and increasing the built in electric field to help electron collections.

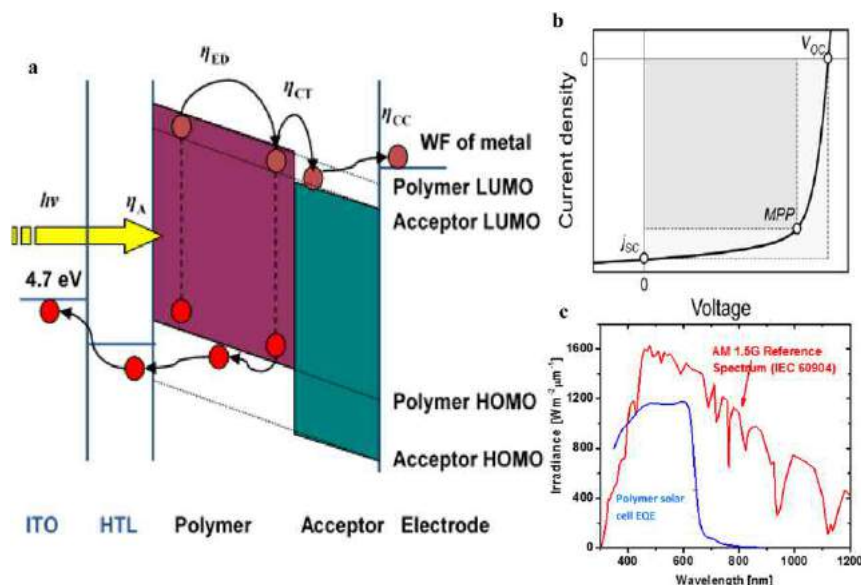


Fig. 2 (a) Simplified schematic mechanism of photocurrent generation in organic solar cells which consist of four sequence steps photon absorption (η_A), exciton diffusion (η_{ED}), charge transfer (η_{CT}), and charge collection (η_{CC}); (b) J - V characteristic curve of solar cell under illumination; (c) standard terrestrial AM1.5G solar spectrum (irradiance) and a typical external quantum efficiency (EQE) spectrum of polymer solar cell. HTL, hole transport layer; HOMO, highest occupied molecular orbital; ITO, indium-tin-oxide; LUMO, lowest unoccupied molecular orbital.

The overall quantum efficiency (a function of wavelength) can be calculated by following equation:

$$\eta_{EQE}(\lambda) = \eta_A(\lambda)\eta_{ED}(\lambda)\eta_{CT}(\lambda)\eta_{CC}(\lambda) \quad (1)$$

Beyond the simple model, one should realize that there are different electronic states in OPV devices – excitons, the charge transfer (CT) states/excitons, the mobile polaron transport states and localized trap states – each play a crucial role in the device operation. In particular, the CT state is an intermediate state between exciton and “free” polaron state. It comprises an electron and hole on either side of the BHJ interface, the electron and hole separation reduces the Coulomb interaction. The CT state is tightly connected with the V_{oc} limit of the solar cell.

Characterization of OSC

The J - V characteristics of a solar cell under illumination are shown in Fig. 2(b). There are three important parameters that combined to give the power conversion efficiency (PCE): short-circuit current density J_{sc} , open circuit voltage V_{oc} , and fill factor (FF).

1. Short circuit current density

Short circuit current density, J_{sc} , is defined as the current when the solar cell has no external load. The J_{sc} represents the number of charge carrier that generated in active layer can collected at electrode at short circuit condition. However, not all incident light is absorbed in active layer and not all generated carriers can reach the electrodes. The magnitude of the J_{sc} depends on the absorption spectrum of the material, the morphology, and the mobility of charge carrier in the material. The high photocurrent can be achieved by selecting material with absorption spectra that overlap the photon flux density.

The J_{sc} also depends on solar spectrum. The extraterrestrial solar spectrum resembles that of a blackbody at 5760K. For terrestrial solar cell application, the standard solar spectrum is AM1.5G, where the sunlight goes through 1.5 times the atmosphere, or incide from an elevation angle of 42 degree, and the label G means the spectrum includes both direct light and diffused light as there is scattering by the atmosphere. Fig. 2(c) shows the standard terrestrial AM1.5G solar spectrum (irradiance) and a typical external quantum efficiency (EQE) spectrum of polymer solar cell. In principle, the integration of the EQE with solar irradiance should be equal to the J_{sc} .

2. Open circuit voltage (V_{oc})

Besides J_{sc} , another key parameter that determine the efficiency of PSC is the open-circuit voltage V_{oc} , which represents the maximum voltage a solar cell can provide to an external circuit. Light harvesting materials employed in OPVs have an optical bandgap of around 1.3–2.1 eV and yet the V_{oc} barely exceeds 1.0 V, i.e., significant loss of photon's original energy. In organic solar cell, the V_{oc} mainly proportional to the energy level difference between HOMO level of donor and LUMO level of acceptor, and more detailed analysis will involve the CT states tightly linked to these levels. As consequence, the selection of the

donor material and acceptor will largely determine and limit the V_{OC} value. In addition, secondary parameters affecting the V_{OC} include recombination, temperature (and thus temperature dependent mobility), work function of electrode, and morphology of active layer.

For a long period of time, fullerene derivatives such as PC₆₁BM and [6,6]-phenyl-C71-butyric acid methyl ester (PC₇₁BM) are dominating acceptor materials in OPV research. In the BHJ system with fullerene derivatives acceptors, as the acceptor LUMO level is fixed, significant efforts were given to donor materials. To enhance the J_{SC} , lower bandgap polymers are the choice, and at the same time, to avoid the typical photovoltage loss, the low bandgap polymers have to have deep HOMO level.

3. FF

FF is the ratio of the maximum power to the open and short circuit values (J_{SC} times V_{OC}) which represents the difference between a real solar cell and the ideal condition. FF shows the dependence of the photocurrent output on the internal field of device, and it's related to the series resistance (R_s) and parallel resistance (R_p). While lower R_s normally results in the higher FF value, FF is much more complex, and potential dependent recombination has to be investigated.

4. Maximum power point

The performance of a solar cell is evaluated from an experimentally measured current–voltage characteristic curve. Maximum power point (MPP) is the point on the J – V curve that the area of resulting rectangle is largest. The MPP represent the maximum of power that can be generated by a solar cell.

5. PCE

The PCE is define as the efficiency of the solar cell and it can be calculated as the maximum output power divided by the input power. The PCE can be calculated as follows:

$$\text{PCE (\%)} = \frac{P_{\text{maz}}}{P_{\text{in}}} = \frac{FE \times J_{SC} \times V_{OC}}{PP_{\text{in}}} \quad (2)$$

The current in the solar cell at any voltage is most readily understood as the sum of the dark current $J_D(V)$ and the photocurrent $J_P(V)$, where they have opposite sign in PV region.

$$J(V) = J_D(V) + J_P(V) \quad (3)$$

From $J(V_{OC})=0$, we know the carrier excitation rate is balanced by an equal carrier recombination rate. The photon generation rate G under 1-sun illumination in a typical organic cell is about 10^{22} cm^{-3} , and the recombination lifetime τ is about 10^{-6} s , thus the carrier concentration $N_C=G\tau$ is about 10^{16} cm^{-3} .

The dark forward current in solar cells obey the diode equation,

$$J_D(V) = J_0 \exp \left[\left[\frac{e(V - R_s J_D)}{nkT} \right] - 1 \right] + \frac{V - R_s J_D}{R_p} \quad (4)$$

where n is the ideality factor, R_p is the parallel shunt resistance, and R_s is the series resistance. The J_0 is related to the barrier height and decreases with larger built-in potential V_{BI} . The ideality factor n is typically in the range 1.3–2 in good organic cells. In classic semiconductor theory, $n=1$ is associated with band-to-band recombination, while $n=2$ is associated with recombination through trap states by the Shockley–Read–Hall mechanism. Modeling shows that the intermediate values arise from recombination through band tails states, possibly in combination with deeper states.

OPV Device Structures

The organic polymer solar cell structure evolution is mainly for improve the interface between donor and acceptor materials, and also the polarity of the device. The development of device structure historically starts from single layer structure, to double layer structure, and bulk heterojunction (BHJ). The structure of the device discussed in sequence.

1. Single polymer layer device

The single polymer layer OPV is the simplest structure. Polyacetylene (PA) were used in single layer organic solar cell that producing V_{OC} 0.3 V and PCE 0.3%. And poly(3-nethyl-thiophene) film sandwiched between aluminum and platinum had external quantum of 0.17% and V_{OC} of 0.4 V and FF 30%.

Organic polymers intrinsically have large exciton binding energy. Using external electrical potential alone is very hard to provide excitation dissociation, and thus the photocurrent will be very low. It has to be extremely thin to dissociate exciton, but this leads to very low light absorption – a dilemma to conquer.

2. Bilayer polymer/acceptor structure.

In bilayer heterojunction (also known as planar heterojunction) solar cell device, film of a polymer as the electron donor (p-type), and a film of electron acceptor (n-type) are deposited on top, as shown in Fig. 3. The selected active layer should realize an offset in HOMO and LUMO energy level at the donor–acceptor (D/A) interface to drive exciton dissociation by CT. In the first successful of two layer organic photovoltaic device by Tang *et al.* a copper phthalocyanine (CuPC) as donor material and a perylene tetracarboxylic derivative as acceptor material were vacuum deposited to make active layer and achieved

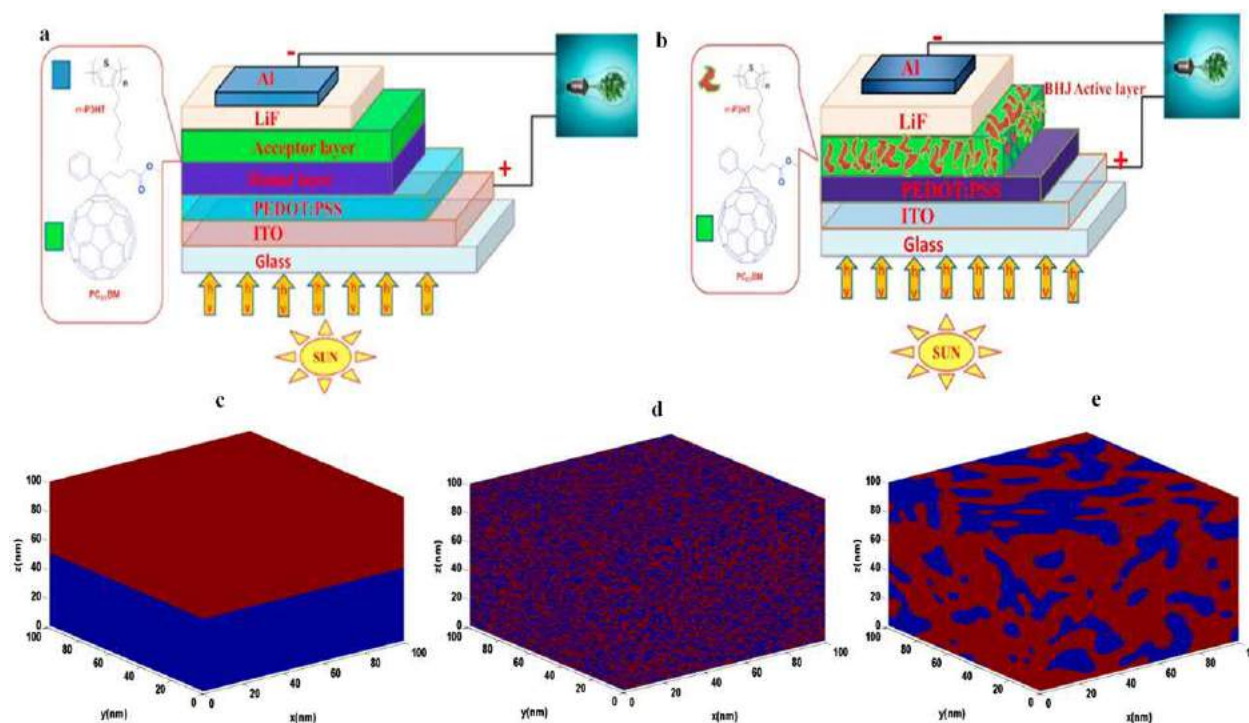


Fig. 3 Polymer solar cell in (a) bilayer structure and (b) bulk heterojunction (BHJ) structure (reproduced from Ganesamoorthy, R., Sathiyar, G., Sakthivel, P., 2017. *Solar Energy Materials & Solar Cells* 161, 102–148); schematic morphology in polymer solar cell (c) pure bilayer structure, (d) “uniform” BHJ structure with very small phase separation domain size, and (e) BHJ with moderately phase separation domain size (reproduced from Maqsood, I., Cundy, L.D., Biesevker, M., *et al.*, 2013. Monte Carlo simulation of Förster resonance energy transfer in 3D nanoscale organic bulk heterojunction morphologies. *Journal of Physical Chemistry C* 117 (41), 21086–21095). ITO, indium-tin-oxide.

efficiency around 1%. However, the drawback of bilayer structure is that the layer thickness is limited to the exciton diffusion length. Also, the bilayer structure is limited by small donor–acceptor interface area.

Since the lifetime of the exciton in most polymer films is short, the diffusion lengths are limited to less than 20 nm, which is shorter than the optical absorption pass length (~ 100 – 200 nm). Therefore only a limited part of exciton can diffuse to the interface in bilayer heterojunction. If the polymer is very thick, the excitons cannot reach the donor acceptor interface then recombine. In order to realize efficient photon conversion, this architecture must strike a balance between the relative long optical absorption length ($L_A \sim 50$ nm) and shorter characteristics length scale for exciton diffusion ($L_D \sim 10$ nm).

In evaporated small molecule solar cell, the complicated layered solar structure can be realized. IMEC group developed bilayer nonfullerene solar cell with PCE more than 6% using α -6T as donor and SubNc as acceptor. The device performance is better than using fullerene as acceptor due to broaden the absorption spectrum. The multilayer planar heterojunction to led PCE up to over 8%. The multiplayer structure, however, is difficult to realize in solution process OPV.

3. BHJ polymer solar cell structure.

Bilayer structure problem can be addressed by BHJ structure. In the BHJ device a blend or mixture of donor and acceptor molecules were used to form a bi-continual layer as shown in **Fig. 3(b)**. Bicontinuous three dimensional interpenetrating network yielded an increased D–A interface, which is the necessary criteria for the effective exciton dissociation.

The first BHJ concept was introduced by Hiramoto *et al.* in evaporated small molecule through co-deposited of phthalocyanine (Pc) and perylene derivatives (PTC) as interlayer between Pc and PTC. The photocurrent enhanced due to a large number of molecular contact while in the bilayer the donor and acceptor are completely separated. The much more attractive solution processed BHJ organic polymer solar cell was realized by Heeger’s group using polymer–fullerene and Friend’s group in polymer–polymer blend in 1995. The BHJ concept has significantly increased the number of interfacial area of donor and acceptor phase.

Strategy to Improve BHJ Device Performance

The high performance polymer solar cell is dominated by BHJ structure. We review the progress in three key research areas: morphology, active materials, and interface and structural engineering.

BHJ Morphology Control Through Process Engineering

Bicontinuous interpenetrating active layer morphology is a pre-requisition for high organic solar cell performance. The morphology controlling is affected by many factors including crystallinity, molecular arrangement, and drying rate. In bilayer structure, interfacial area between donor and acceptor is on the planar surface (Fig. 3(c)), thus exciton dissociation efficiency into free carrier is very low. The “uniform” BHJ structure has very large donor–acceptor interface area, but the morphology become more important. As shown in Fig. 3(d), the small domain phase separation of BHJ provides a large number D–A interface, but it is not providing continuous pathway. As consequence, large number of charge generation produced but electron and hole trapped in the active layer then recombine, and cannot effectively generate J_{sc} . The right BHJ morphology will look like the excessively D–A interface in Fig. 3(e) – it has the right balance of D/A interface (for efficient exciton dissociation) and nanoscale bicontinuous pathway to enable the separated electron and hole polarons reaching the electrodes.

The reported effective approaches in tuning morphologies of the active layer in OPVs including solvent annealing (SA), thermal annealing, solvent additives, solvent (vapor) annealing, and solvent mixtures, which would be presented in the following section.

Thermal Annealing

Thermal annealing is an effective method to improve PCE discovered in the early years of OPV, when the workhorse material changed from poly[2-methoxy,5-(2'-ethyl-hexyloxy)-p-phenylene vinylene] (MEH-PPV) to soluble polythiophenes, especially regioregular poly(3-hexylthiophene) (RR-P3HT), which has lower bandgap of 1.9 eV. The PCE of P3HT: fulleropyridine was enhanced by 3–4-fold with mild thermal treatment (50°C for 30 min); thermal annealing temperature, annealing time, and annealing sequence – either before (preannealing) or after (postannealing) electrode deposition, all play role in determining the performance of classical P3HT:phenyl-C61-butyric acid methyl ester (PCBM) solar cells.

The absorption of the P3HT in the P3HT:PCBM active layer, when processed in typical fast spin-coating approach, is very different from that of the pristine P3HT film. It's generally observed the P3HT absorption red shifted after thermal annealing, indicating the enhanced (or partially recovered) polymer crystallinity. Grazing incidence X-ray scattering (GIXS) become a powerful approach to investigate the polymer blend. GIXS technique gives clearly evidence of the π – π stacking orientation (orientated parallel to the substrate in P3HT:PCBM case), and enhanced crystallinity after thermal annealing above transition glass temperature (T_g). Upon thermal annealing, the PCBM molecules can freely diffuse out of these mixed regions to aggregate into PCBM cluster while enhancing P3HT crystallinity. The PCBM diffusion may induce simultaneously more ordered packing of P3HT and form bicontinuous morphology (Fig. 4(a)). In semi crystalline polythiophenes, high charge mobility occurs along the polymer main chain and parallel to the π – π stacking direction of the thiophenes rings. The increased crystallization via thermal annealing also reduces density of the defects at the interface and in the bulk. Thermal annealing is effective beyond P3HT:PCBM. Annealing at 100°C for 20 min improves PBDTS-DTBT (as donor) and nonfullerene ITIC (as acceptor) solar cell from 7.80 to 9.09%. Increased interchain packing with significantly intensified face-on π – π stacking indicates enhanced exciton dissociation and charge transport, thus results in enhanced current density and FF.

Solvent (Vapor) Annealing

SA is another method to control morphology of the active layer by slowing down the evaporation rate of the solvent, which slows the film formation/drying process. In P3HT:PC₆₁BM system using high boiling point solvent ODCB, the slow solvent removing process provide the polythiophene chains enough time to demonstrate the self-organization capability embedded in the regioregular design of the polymer. The SA approach results in a much increased hole mobility and balanced charge transport, which lead to a more than doubled PCE of 4.4%. The thermal annealing and SA approaches improves the morphology in different way in P3HT system. Regioregular P3HT itself has excellent absorption up to 650 nm and the π – π stacking of polymer chains ensured high mobility. However the long wavelength absorption is severely suppressed with the present of acceptor PCBM when the solvent was removed quickly in traditional way. The thermal annealing approach can partially recover the polymer ordering and absorption. The SA takes another route by preserving the high degree of ordering rooted in the self-organization during the initial film formation period instead. Fig. 4(b) shows the GIXS results of the RR-P3HT:PCBM film with and without SA. This shows the polymer chain packing is dominated by edge-on orientation, and the SA not only leads to higher crystallinity, but also smaller mainchain spacing (a -axis). Phase image of atomic force microscopy (AFM) shows very clear polymer fibril structure with ~20–30 nm fiber diameter (Fig. 4(c)).

Dynamically, during the SA process, the BHJ blend was found swollen by solvent, which allowed PCBM acceptor to diffuse out of P3HT domains. This also improves mobility of fullerene, increases the crystallinity of P3HT, and improves phase separation between domains.

Mixed-solvent and vapor annealing can be combined to improve polymer crystallinity and inhibit intercalation into polymer side chains, shown in a poly[N-9'-heptadecanyle-2,7-carbazole-alt-5,5'-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT): PC₇₀BM blend system. After solvent vapor annealing in mixed solvent tetrahydrofuran (THF) and carbon disulfide (CS₂), the out-of-plane GIXS results show that PCDTBT crystals with larger size and higher crystallinity were formed due to enhanced π – π interaction of polymer chain as shown in Fig. 4(d). Moreover, the PC₇₀BM molecules diffused out of polymer chain and formed

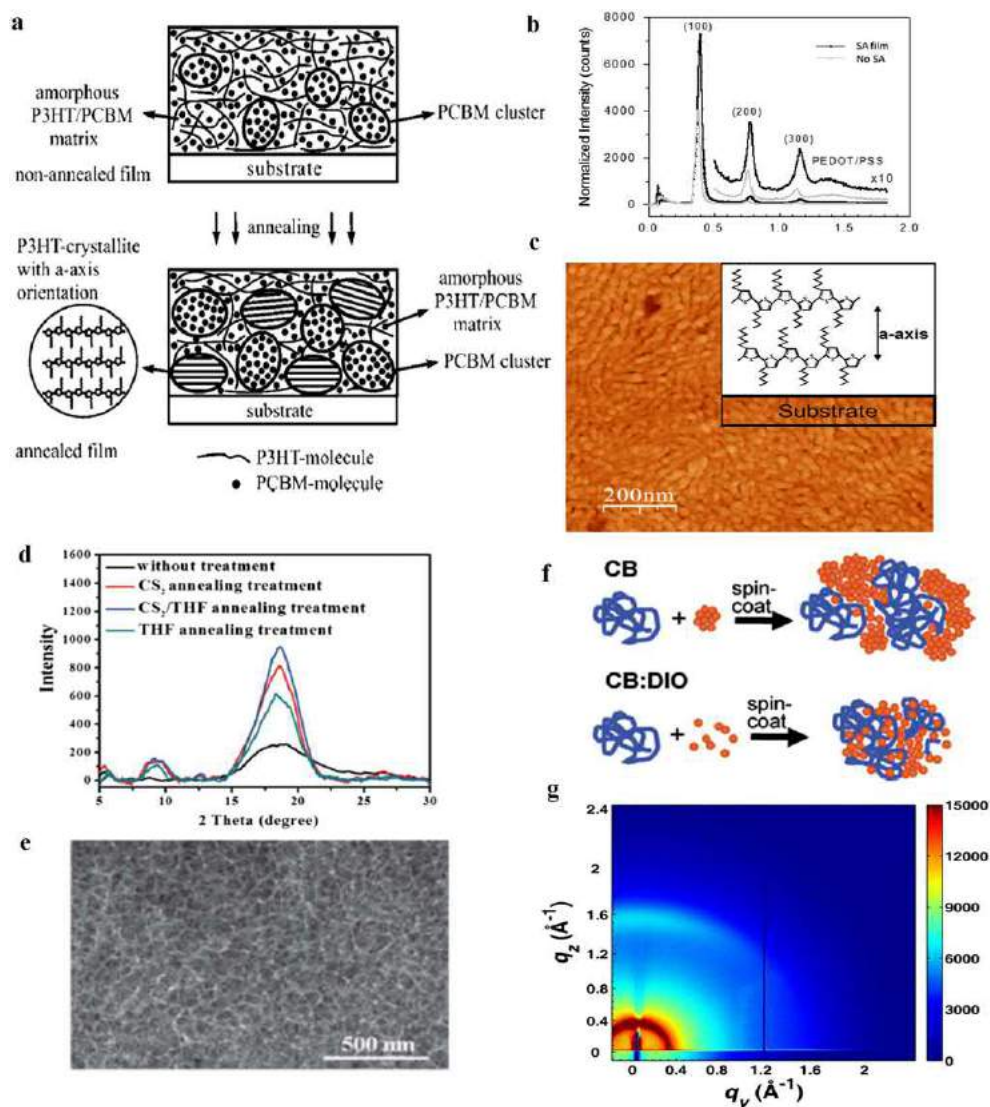


Fig. 4 (a) Schematic of structural change of P3HT:phenyl-C61-butyric acid methyl ester (PCBM) film upon thermal annealing (reproduced from Erb, T., Zhokhavets, U., Gobsch, G., *et al.*, 2005. *Advanced Functional Materials* 15, 1193–1196); (b) the grazing incidence X-ray scattering (GIXS) result of P3HT:PCBM films formed with and without solvent annealing (SA); (c) atomic force microscopy (AFM) image of solvent annealed P3HT:PCBM film ((b) and (c) 2007. *Advanced Functional Materials* 17, 1636–1644); (d) the out of plane GIXD profile shows the crystallinity of poly[*N*-9'-heptadecanyl-2,7-carbazole-alt-5,5-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT):PC₇₀BM films and (e) transmission electron microscopy (TEM) images of PCDTBT:PC₇₀BM blend films with mixed solvent vapor annealing ((d) and (e) 2013. *Journal of Materials Chemistry A* 1, 6216). (f) Working mechanism of poly(4,8-bis-alkyloxybenzo(1,2-b:4,5-b')dithiophene-2,6-diyl-alt-(alkylthieno(3,4-b)thiophene-2)) (PBDTTT):PCBM morphology formation with and without additives (reproduced from Lou, S.J., Szarko, J.M., Xu, T., *et al.*, 2011. Effects of additives on the morphology of solution phase aggregates formed by active layer components of high-efficiency organic solar cells. *Journal of the American Chemical Society* 133, 20661–20663); and (g) the GIXS data of PBDTTT polymer film, showing the face-on orientation of the polymer chain.

large aggregates due to the THF component, which induced phase separation. In addition, owing to the existence of the CS₂ vapor component, the migration ability of PCDTBT chains was improved, which enhanced its self-organization ability and facilitated the formation of nanofibrils (Fig. 4(e)).

Solvent Additive

The addition of small amount solvent additive during film casting is a simple and effective strategy to control morphology of polymer solar cell in many OPV systems.

The addition small amount of 1,8-octanedithiol (ODT) in poly[2,6-(4,4-bis-(2-ethylhexyl)-4H-cyclopenta[2,1-b;3,4-b']dithiophene)-alt-4,7-(2,1,3-benzothiadiazole)] (PCPDTTT):PCBM system effectively improve PCE from 2.8 to 5.5%. It's an important and surprising discovery. Alkanedithiol additives do not react with either the polymer or the fullerene and removed

while film drying. The alkanedithiol selectively dissolves the PCBM, but not the donor polymer. The differential solubility and boiling point provide powerful tool to achieve better ordering of donor polymer and control the phase separation. Devices without ODT addition has predominately amorphous structure, while addition 3 wt% of ODT resulted in the nucleation of crystalline PCPDTBT. Time resolved in situ wide angle (WA) GIXS experiments revealed that small amount of ODT simultaneously reduces the nucleation barrier for polymer crystallization, which facilitated reorientation and crystal growth during a prolonged film drying process.

DIO (1,8-diiodooctane) additive in main solvent chlorobenzene (CB) improves the PCE of PTB7:PC₇₁BM polymer solar cell from 3.92 to 7.40%. It was found that the dramatic enhancement of the device performance due to the change in the morphology more uniform, showing good miscibility between PTB7 and PC₇₁BM and form interpenetration networks. The DIO-processed blend films results more homogenous phase separation distance of 20–40 nm. The finer domain size is beneficial for charge separation and transport which enhancing FF. X-ray scattering technique shows that without DIO, the large PC₇₁BM aggregates hinder PC₇₁BM intercalation into the PTB7 network during film formation. DIO selectively dissolves PC₇₁BM aggregates, facilitates integration of PC₇₁BM into PTB7 domain and optimizing both domain sizes with improved donor/acceptor interface (Fig. 4(f)).

1-Chloronaphthalene (CN), diphenyl ether (DPE), 1-phenylnaphthalene (PN), 1-naphthalenethiol (SH-NA), diiodohexane (DIH) are also examples of effective additives in polymer solar cell. In fact, PN additive has enabled 11.7% PCE in OPV cell, which promotes face-on polymer backbone orientation, reduces domain size, and increases domain purity. While solvent additives show great performance in efficiency enhancement, one must try to avoid the possible drawbacks. Additives are typically high boiling point. They need to be fully removed, otherwise the film will be too weak mechanically to be used in production, and there are evidences that the remaining additives like DIO accelerates photodegradation of the active layer due to nano-morphology change. This actually leads to a new research direction of high boiling point additive free OPV toward high performance. Mixed solvent of common major organic solvents in OPV such as DCB, CB, CF, etc. sometimes can give better performance in solar cell. This approach actually already exists for over a decade. It is also a straight forward way of control the solvent remove rate.

Active Materials in Polymer Solar Cells

Polymer Active Materials

Polymer active materials are at the core of the polymer solar cell technology. The optimal morphology formation in the device is highly material system dependent. A brief review of some of the key polymer solar cells will be presented below. For detail OPV materials development, there are many excellent review publications on the topic.

One of the earliest PSCs polymers is MEH-PPV, Fig. 5(b)), and excellent representative PPV derivative, which is also a good light emitting polymer. MEH-PPV was further improved into poly(2-methoxy-5-(3',7'-dimethyl-octyloxy))-*p*-phenylene vinylene (MDMO-PPV) (or OC₁C₁₀-PPV) through side chain modification, and efficiency up to 3.5% was realized in MDMO-PPV:PC₇₁BM. This is the best before the society figured our ways to conquer the morphology problem in regioregular RR-P3HT:PCBM in ~ 2005, a decade after the invention of RR-P3HT. Since then, the pace of OPV progress became much faster, and many excellent materials were invented for OPV.

Model polymer RR-P3HT has bandgap of ~ 1.9 eV, which only allowed about 30% of photon was absorbed. Compared to the silicon band of 1.1 eV and absorption up to 77% of the sunlight, lower bandgap material is a continuously pursuing target of the OPV society. Hsu *et al.* reviewed the low bandgap polymer design strategy for OPV shown in Fig. 5(a), which still largely holds.

RR-P3HT (Fig. 5(c)) is an excellent example of the first strategy – stereo regularity induced interchain coupling. The interchain π - π interaction leads to increased ordering and crystallinity in the materials and red shift in absorption. Beyond RR-P3HT, several very effective synthetic strategies have been developed – (1) using alternating electron-rich (donor) and electron-deficient (acceptor) units in the backbone to form the D-A copolymers; (2) tuning the aromaticity and bond length alternation (BLA), which stabilize the lower bandgap quinoid resonance structure over aromatic structure; (3) incorporate strong electron withdrawing substitutes such as carbonyl group or fluorine atoms; (4) incorporating chemically rigid (e.g., larger planarity) unit. Thousands of new compounds have been designed and synthesized for OPV applications. Here are some representative OPV materials in the last decade.

Electron-donating carbazole unit and electron-withdrawing benzothiadiazole (BT) unit-based polymer (PCDTBT) with 1.88 eV bandgap and low lying HOMO (5.5 eV) shows promising PCE of 3.6%, which was later improved to ~ 7% level. Cyclopenta[2,1-b;3,4-b']dithiophene (CPDT) and dithieno[3,2-b:2',3'-d]silole (DTS) units with stronger electron-donating properties were copolymerized with BT unit, and two much lower bandgap polymers (Fig. 5(d)) – PCPDTBT (response up to 850 nm) and poly[(4,4'-bis(2-ethylhexyl)dithieno[3,2-b:2',3'-d]silole)-2,6-diyl-alt-(2,1,3-benzothiadiazole)-4,7-diyl] (PSBTBT) (response up to 820 nm) were formed due to the strong intramolecular donor-acceptor interactions. The substitution of C-atom to heavier Si-atom in PSBTBT started the heavy atom substitution approach in the OPV field. The C-Si bond is longer than the C-C bond and less steric hindrance is created in the DTS core by the side chains, leading to a better π - π stacking and enhanced crystallinity. Follow-up work shows replacing Si atom with even heavier Ge atom in carbazole unit can further improve the performance, due to the improved molecular ordering.

A beautiful example of using larger planar unit to improve carrier transport and thus OPV performance is the benzo[1,2-b;4,5-b']dithiophene (BDT) unit. When copolymerized with thieno[3,4-b]thiophene (TT) series of building blocks, star polymers

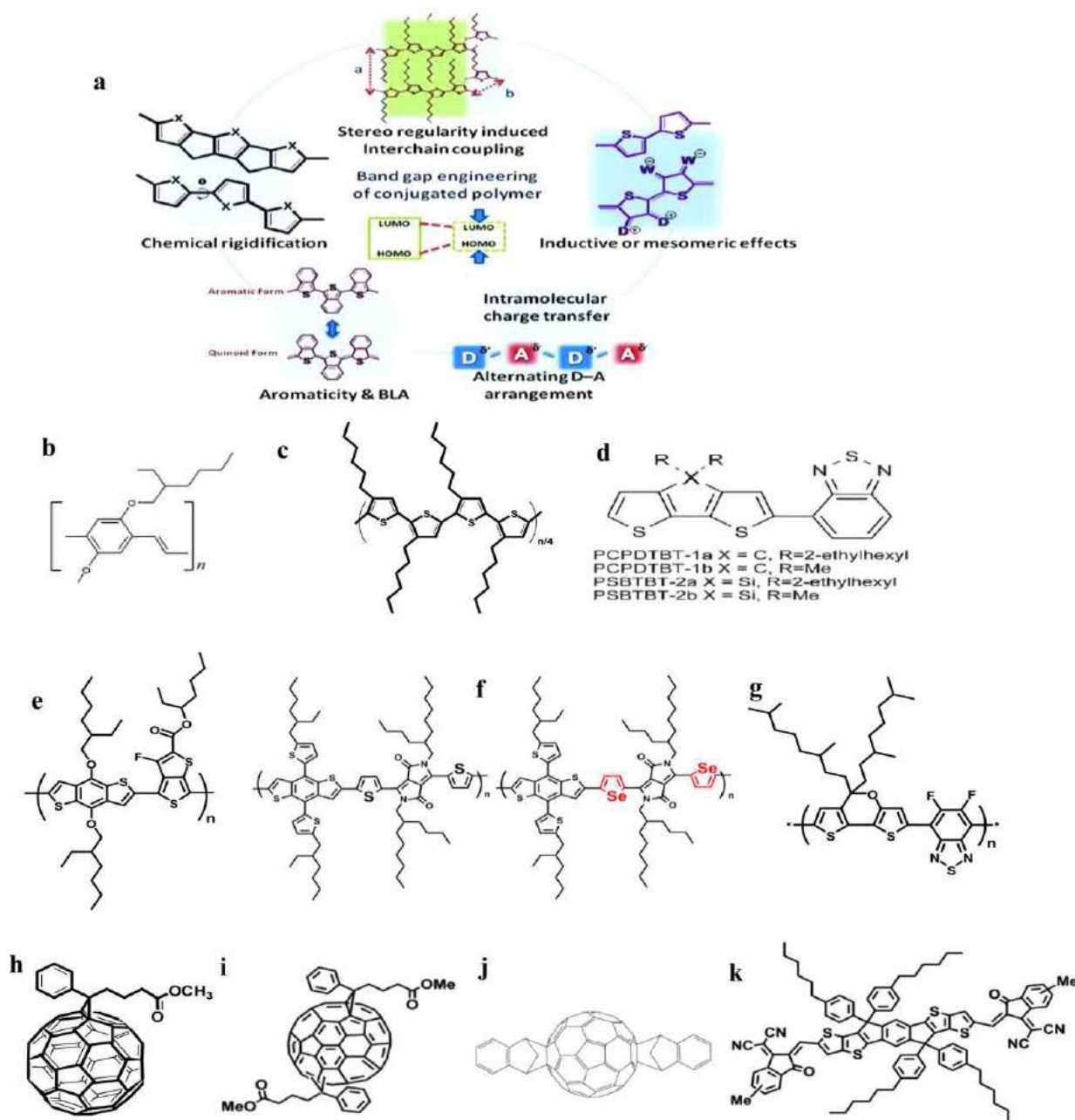


Fig. 5 (a) Design principles for energy level and bandgap tuning in conjugated polymers (reproduced from Cheng, Y.-J., Yang, S.-H., Hsu, C.-S., 2009. Synthesis of conjugated polymers for organic solar cell applications. *Chemical Reviews* 109 (11), 5868–5923); molecular structures of representative donor polymers; (b) poly[2-methoxy,5-(2'-ethyl-hexyloxy)-p-phenylene vinylene] (MEH-PPV); (c) regioregular poly(3-hexylthiophene) (RR-P3HT); (d) poly[2,6-(4,4-bis-(2-ethylhexyl)-4H-cyclopenta[2,1-b;3,4-b'] dithiophene)-alt-4,7-(2,1,3-benzothiadiazole)] (PCPDTBT) and poly[(4,4'-bis(2-ethylhexyl) dithieno[3,2-b:2',3'-d] silole)-2,6-diyl-alt-(2,1,3-benzothiadiazole)-4,7-diyl] (PSBTBT); (e) PTB-7; (f) poly(2,6'-4,8-di(5-ethylhexylthienyl) benzo[1,2-b;3,4-b]dithiophene-alt-5-dibutyltolyl-3,6-bis(5-bromothiophen-2-yl)pyrrolo[3,4-c]pyrrole-1,4-dione) (PBDDT-DPP) and poly(2,6'-4,8-di(5-ethylhexylthienyl)benzo[1,2-b;3,4-b]dithiophene-alt-2,5-bis(2-butyltolyl)-3,6-bis(selenophene-2-yl)pyrrolo[3,4-c]pyrrole-1,4-dione) (PBDDT-SeDPP) (g) PDTP-DFBT molecular structures of representative acceptors; (h) PC₆₁BM; (i) Bis-PC₇₁BM; (j) IC₆₁BA and (k) IT-M. HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbital.

PTBx, also called PBDT-TT (**Fig. 5(e)**) series of polymers were formed. Further materials improvements include adding a fluorine atom on the TT unit to lower the HOMO level, side chain engineering to tune the solubility, crystallinity, etc. This series of star polymers set the 7, 8, and 9% PCEs milestones of OPV, and still are among the best. It is also a great example that all four synthetic strategies have been used jointly to get the best result. Another example is a high V_{oc} polymer PBDT-TPD formed by copolymerizing BDT unit with strong electron accepting thieno[3,4-c]pyrrole-4,6-dione (TPD) unit.

Diketopyrrolopyrrole (DPP) unit is a strong electron-withdrawing unit, and shows its effectiveness of low bandgap OPV polymer below 1.5 eV. poly(2,6'-4,8-di(5-ethylhexylthienyl) benzo[1,2-b;3,4-b]dithiophene-alt-5-dibutyloctyl-3,6-bis(5-bromothiophen-2-yl)pyrrolo[3,4-c]pyrrole-1,4-dione) (PBDTT-DPP) polymers, a series of D-A copolymers with BDT and DPP unit, show small bandgap of ~ 1.46 eV, deep HOMO level, and high charge carrier mobility. The reduction of the bandgap and the enhancement of the charge transport properties of PBDTT-DPP can be accomplished by substituting the sulfur atoms on the DPP unit with selenium atoms. The resulted poly{2,6'-4,8-di(5-ethylhexylthienyl)benzo[1,2-b;3,4-b]dithiophene-alt-2,5-bis(2-butyloctyl)-3,6-bis(selenophene-2-yl)pyrrolo[3,4-c]pyrrole-1,4-dione} (PBDTT-SeDPP) ($E_g \sim 1.38$ eV, Fig. 5(f)) polymer gives single junction devices with PCEs of 7.2%.

Adding two fluorine atom on BT unit leads to strong electron deficient unit difluoro-benzothiadiazole (DFBT) unit. Copolymerized with less electron donating BDT unit (through side chain substitution with alkyl group), PBDT-DTffBT polymer of 1.7 eV bandgap gives higher V_{oc} and PCE. Asymmetric electron rich dithieno[3,2-b:2',3'-d]pyran (DTP) unit shows even stronger electron-donating property than BDT unit, and when copolymerized with electron deficient DFBT unit, a regiorandom polymer (PDTP-DFBT) with bandgap 1.38 eV was formed. Even with this low bandgap, high V_{oc} of 0.69V, EQE over 60 and 8.0% PCE single junction device was achieved. In polymer PffBT4T, where DFBT was copolymerized with four thiophene rings, the side chains on two of the thiophene rings were found very critical for the intermolecular interaction between different polymer chains and polymer-fullerene interaction. Generally, larger side chains improve the solubility of polymer by reducing π - π interchain interaction. However, longer alkyl chains were found to be detrimental to the domain size and domain purity. This is a great example of side chain engineering in polymer design. Record 11.7% PCE was achieved in PffBT4T- C_{80} :PCBM system, assisted by the sophisticated morphology control.

Acceptor Materials in OPV

Donor polymer and the acceptor together determine the device performance. The donor polymers are dominantly designed to be compatible with the fullerene derivatives ($PC_{61}BM$ (Fig. 5(h)) or $PC_{71}BM$) for a long time.

Fullerene bisadducts based on $PC_{61}BM$ (Bis- $PC_{61}BM$, Fig. 5(i)) has ~ 0.1 eV higher LUMO than that of PCBM, and gives higher V_{oc} and PCE in P33HT system. Indene- C_{60} or Indene- C_{60} bisadduct ($IC_{60}BA$ (Fig. 5(j)) or $IC_{70}BA$), via a simple one-pot reaction of indene and fullerene molecules, can achieve higher V_{oc} by over 0.2V, and PCE up to 7.4% in P3HT:ICBA system, which is due to the higher LUMO (by about 0.2 eV).

In addition, n-type polymer-based electron acceptors have also been explored such as F8TBT, CN-PPV and perylene diimide (PDI)-DTT. PDI-based acceptor perylene bisimide (PBI) has lead to 8.4% PCE polymer solar cell. Small molecule nonfullerene acceptor is a new star in the OPV research in the past a few years. Indacenodithieno[3,2-b]thiophene-based nonfullerene acceptor (ITIC), and methyl-modified nonfullerene acceptor (IT-M, Fig. 5(k)) are great examples. Poly[(2,6-(4,8-bis(5-(2-ethylhexyl)thiophen-2-yl) benzo[1,2-b:4,5-b']dithiophene))-co-(1,3-di(5-thiophene-2-yl)-5,7-bis(2-ethylhexyl) benzo[1,2-c:4,5-c']dithiophene-4,8-dione)] (PBDB-T):IT-M polymer cell reached 11.6% efficiency, while adding Bis-PCBM leads to ternary OPV with 12.2% efficiency – a record, from much preferred nonfullerene acceptor.

Interface and Architectural Engineering

The active materials and the BHJ morphology determines how efficient the active layer harvest solar irradiation, exciton conversion into free carrier, and the free carrier transport to the organic/electrodes interface. The organic/electrode contacts of OPV device determines the charge carrier extraction efficiency into the external circuit. An Ohmic electrical contact is highly desired, as a barrier height of few tens of millielectron volt can result in significant charge accumulation and thus inferior photovoltaic performance. This is true for both the anode and cathode sides of contacts for efficient charge extraction with high selectivity.

In the simplest sandwich OPV structure described by the rigid band metal-semiconductor-metal model, the difference of work function between the anode and cathode provides the driving force for the charge transport and extraction process. Sufficient metal electrodes work function difference is required for the efficient and selective charge collection in the device. Pure metal electrodes, however, are not desired, as the metal surfaces have a high density of defects causing surface recombination or strong surface dipoles at metal/organic interfaces that may harm selective charge extraction.

HTLs

Transparent conductive indium-tin-oxide (ITO) with a high work function of ~ 4.7 eV is the most widely used transparent electrode in organic electronics. Naturally a solution processible, water-based p-type interface layer poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) (PEDOT:PSS) (work function ~ 5.0 eV) was applied to form the ohmic contact with both the ITO and the organic photoactive layer for effective charge collection. Other functions including smoothing the ITO surface and removing potential pinholes, which improves the yield of the device fabrication. The PEDOT:PSS's organic, water absorbing and acidic nature are not preferred for device long-term stability.

p-Type transition metal oxide – NiOx is a wide bandgap (direct bandgap, $E_g \sim 3.7$ eV), transparent semiconductor with weak absorption bands due to d-d transitions of $3d^8$ electron configuration. Both vacuum sputtered and solution processed forms of

NiOx have been demonstrated to be effective HTL materials. With O₂-plasma treatment, the WF of NiOx can be increased from 4.7 to 5.3 eV, indicating a doping effect, and very suitable for HTL application. P-type copper oxides CuOx, and binary copper oxide such as CuGaOx, CuCrOx are also recently emerging as effective HTL materials. The binary copper oxide has larger bandgap of over 3 eV – is more suitable for window layer, as compared with visible absorbing pure CuOx.

V₂O₅, MoO₃, WO₃ and recently CrOx, are also experimental proven to be effective HTL materials in OPV. Surprisingly, it is demonstrated with direct and inverse photoemission spectroscopy measurements that Cr₂O₃, MoO₃, WO₃, and V₂O₅ are n-type materials exhibiting very deep lying electronic states. Ultrahigh vacuum (UHV) prepared MoO₃ by evaporation possesses a 6.86 eV Fermi level, a 6.70 eV EA and 9.68 eV IP. Evaporated transition metal oxides – WO₃ and V₂O₅ with even higher WFs of a ~6.7 eV, and ~7.0 eV, respectively. The fact that they function as HTL may be associated with the multivalence state of metal cation. It worth mentioning that the very different work function from UHV and normal high vacuum evaporation is due to the existence of oxygen in the evaporation case.

Graphene oxide (GO), also water soluble, can also function as HTL in OPVs. In addition to find the interface materials with right work function, the solution processability is a much preferred research topic.

ETLs

Low work function metals are very sensitive to oxygen and moisture – good for device efficiency, but very poor stability and cost perspective. N-type transition metal oxides, such as ZnO and TiOx, are proven ETLs in both OLEDs and OPV devices. They have good stability, proper work function, smooth film morphology and solution processability. TiOx and ZnO were also shown to function as a hole blocking layer and optical spacer to modulate the optical field inside thin device to enhance absorption, and thus may help efficiency.

Many compounds with low work-function metal such as Cs₂CO₃, CsF, LiF, Ca(acac)₂, etc., which have been proven to be effective ETL in OLED research, are also useful in OPV. Another class of ETL are polyelectrolyte, either conjugated (such as poly [(9,9-bis(3'-(N,N-dimethylamino)propyl)-2,7-fluorene)-alt-2,7-(9,9-dicetylfluorene)] (PFN) or nonconjugated (e.g., ethoxylated polyethylenimine (PEIE)), can enhance the electron extraction significantly in the OPV devices.

Combining the different ETLs has shown surprising performance improvement. Anatase TiOx with Cs₂CO₃ leads to Cs-doping of TiOx, and this leads more reliable electron buffer material for OPV and OLED, and it removes the UV photo-catalytic degradation disadvantage of TiOx alone in the device. ZnO and PFN double layer ETL was also shown to be of enhanced efficiency in OPV device.

Inverted OPV Device

Some ETL materials can convert the given high work function ITO transparent electrode into low work function contact. This provides the versatility in device polarity. Fig. 6(a) shows the 2006 UCLA work on the formation of inverted OPV by using Cs₂CO₃ ETL to convert ITO into low work function contact, and use transition metal oxide (V₂O₅) as HTL.

Different forms of TiOx, ZnO (sol-gel, nanoparticle), PFN, etc. were also shown to be able to enable low WF contact on transparent electrode. Ultraviolet photoemission spectroscopy gives clear evidence of the large WF reduction after these ETL applied on ITO, for example. ETLs like LiF does not have such effect. On the other side, MoOx is the most popular HTL used in inverted polymer solar cell.

Fig. 6(b) shows the device structure, the molecular structure of PFN ETL, and active layer polymer PTB7 in the over 9% PCE inverted polymer solar cell. Fig. 6(c) is the energy diagram of the inverted OPV device.

Tandem OPV Devices

Single junction solar cell has its theoretical limit. Shockley and Queisser proposed detailed balance model to predict the efficiency limit for p-n junction solar cells. The biggest two major losses under the S-Q assumptions (Fig. 7(a)) are (1) below bandgap loss – photons with energy higher than bandgap of absorbers can be completely absorbed, while photons below the bandgap just pass through without any absorption; (2) the thermalization loss – photons of higher energy than the bandgap energy will thermally relax, and the excess energy loses. While lower bandgap materials absorb more sunlight, and potentially higher photocurrent, the larger bandgap materials will win in the photovoltage side. Therefore a delicate balance exists in the theoretically optimization of bandgap, and ~1.4 eV is the optimal bandgap of solar cell material under terrestrial illumination, with theoretical limit of ~31%. The most effective way in overcoming S-Q limit is the multijunction or tandem structure, which consists of several semiconductors with different and complement bandgaps to have better coverage of the solar spectrum. The carefully designed system will use the solar photons much more effectively by reducing the thermal loss. For example, theoretically three-junction solar cell design raises the ceiling to ~50%.

For disordered solar cell system with lower efficiency like OPV, the tandem structure is a clear go for higher efficiency. The preferred way is the solution process tandem solar cell, as it has the potential of higher efficiency without losing the low-cost edge. The structure of tandem solar cell is more complex. Besides the two photoactive subcells, the interconnection layer connecting the two subcells is very important. In inorganic multijunction solar cells, the interconnection layer is heavily doped p-n

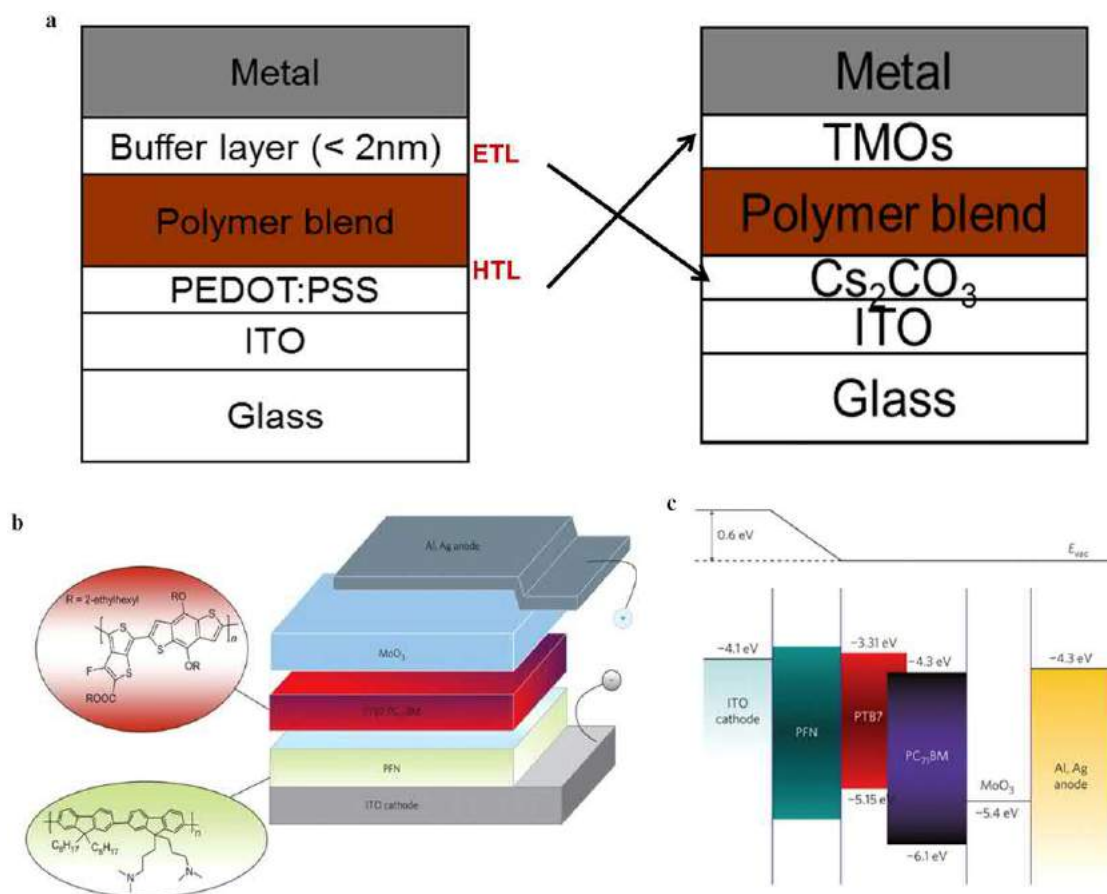


Fig. 6 (a) The formation of inverted organic conjugated polymers-based solar cells (OPVs) by switching position of electron transport layer (ETL) and hole transport layer (HTL); (b) device structure of inverted OPV using poly[(9,9-bis(3'-(*N,N*-dimethylamino)propyl)-2,7-fluorene)-alt-2,7-(9,9-dicyclopentafluorene)] (PFN) as ETL and PTB7 as active polymer; (c) energy diagram of the inverted OPV device. ITO, indium-tin-oxide; PEDOT:PSS, poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate). ((b) and (c)) Reproduced from He, Z., Zhong, C., Su, S., *et al.*, 2012. Enhanced power-conversion efficiency in polymer solar cells using an inverted device structure. *Nature Photonics* 6, 591–595.

($p^+ + n^+ +$) tunneling junction, which has the opposite direction than that of the two subcells to produce photovoltage of the same direction.

In organic polymer solar cell, the extensive interface engineering research provide many high quality p-type and n-type interface layers to choose. In polymer tandem solar cells, the interfacial layers should also provide physical protection to the underlying subcells. Here we only provide a brief introduction on the topic.

In the beginning, the simple way of making tandem device was using semitransparent solar cell, which differ from the normal solar cell by a thin layer of thermally evaporated metal as semitransparent electrode. Tandem solar cell by stacking two or more semitransparent solar cells (allow one to be normal nontransparent cell) has four terminals, and can be connected either in series or in parallel. The configuration has clear disadvantage in optical loss, materials (substrate/TCO) cost, and was used mainly for proof of concept. Early efforts typically use subcells with same active materials, for example, P3HT:PC₆₁BM, PCDTBT:PC₆₁BM, and the efficiency was not really improved. Monolithic integrated tandem solar cell on one substrate is the more preferred way.

Heterogeneous polymer tandem solar cells with complemented spectrum coverage is the main stream research direction. In principle the complemented spectrum coverage is required for the current matching of two subcells, the high and low bandgap combination also leave the opportunity of reducing thermalization loss.

The two most important components in heterogeneous tandem polymer cells are (1) the spectrum complemented active layers in two subcells, and (2) the interconnection layer. In the spectrum complemented active layer, the classical, process well controlled P3HT is a model high bandgap polymer subcell. The 1.9 eV bandgap is similar to the high bandgap component in inorganic multijunction solar cell. It's thus a long time trend in tandem OPV field to pursue high performance polymer at ~1.4 eV bandgap, which is lower than the optimal bandgap in OPV research. In interconnection layer, there are also two major device architectures just like in single junction OPV device, i.e., conventional tandem and inverted tandem OPV device (Fig. 7(b)). A few milestones in the tandem OPV devices are the following.

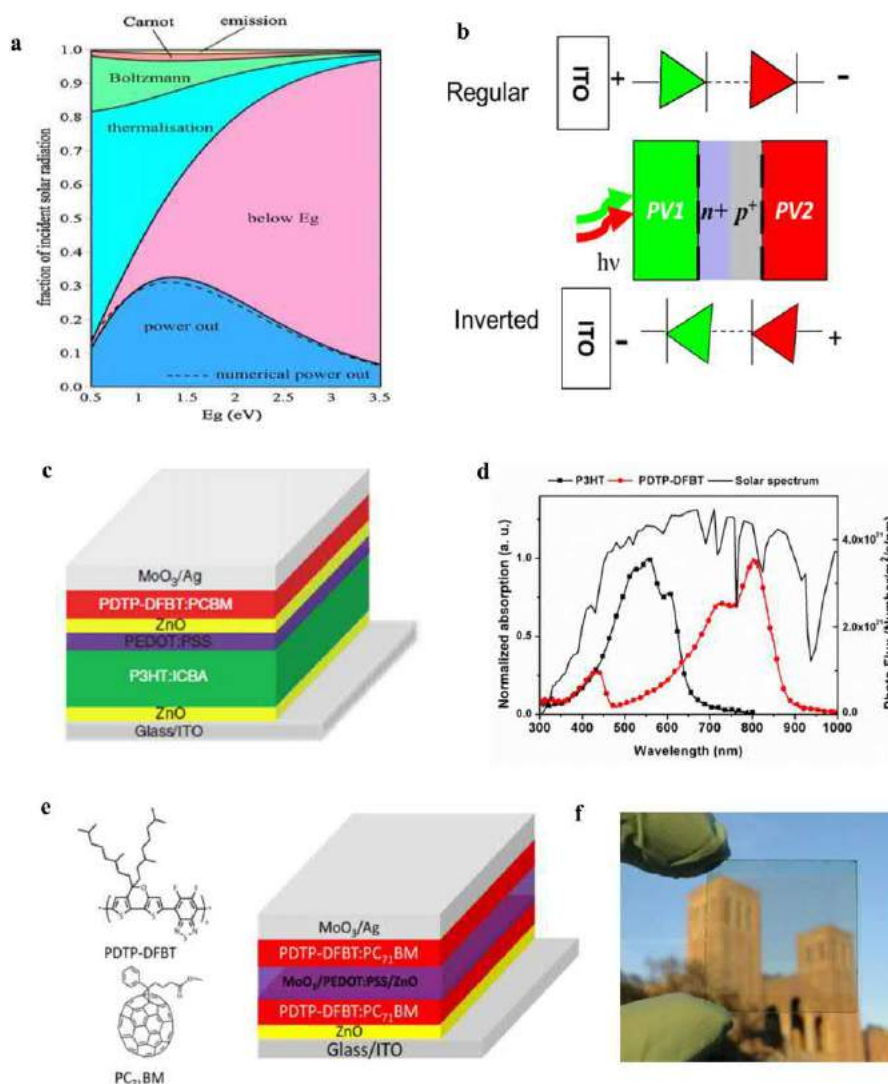


Fig. 7 (a) Loss mechanism and efficiency limitation of single junction solar cell as a function of semiconductor bandgap (reproduced from Hirst L.C., Ekins-Daukes N.J., 2011. Fundamental losses in solar cells. *Progress in Photovoltaics: Research and Applications* 19, 286–293); (b) diagram of conventional tandem and inverted tandem organic conjugated polymers-based solar cell (OPV) devices; (c) device structure of, and (d) wide and low bandgap polymers in the 10.6% certified tandem OPV device ((c) and (d) reproduced from You, J., Dou, L., Yoshimura, K., 2013. *Nature Communications* 4, 1446); (e) molecule structures and device structure of a homogeneous tandem OPV device (reproduced from You, J., Chen, C.-C., Hong, Z., *et al.*, 2013. 10.2% Power conversion efficiency polymer tandem solar cells consisting of two identical sub-cells. *Advanced Materials* 25, 3973–3978); and (f) picture of an all solution process visibly transparent OPV device. ITO, indium-tin-oxide.

Conventional tandem polymer solar cell – UCSB group reported solution processed TiO₂(n-type)/PEDOT:PSS (p-type) as interconnecting layer, and subcells are based on (1) low bandgap PCPDTBT:PC₆₁BM front cell, and (2) wide bandgap P3HT and PC₇₁BM rear cell. With the front and rear cell of 3.0% and 4.7% efficiency, respectively, a 6.5% tandem device was achieved. Improving both cells to P3HT:ICBA and PSBTBT:PCBM, and traducing MoO₃ into interconnection layer, UCLA improved the tandem efficiency to 7%. Another popular interconnection layer is based on ZnO(n)/pH-neutral PEDOT:PSS (p) from a Netherland group.

Inverted tandem polymer solar cells – UCLA used low bandgap polymer PBDTT-DPP (~ 1.44 eV bandgap):PC₇₁BM as rear cell and P3HT:ICBA as front cell, modified PEDOT:PSS/ZnO as interconnection layer, 8.62% NREL certified tandem efficiency was achieved. Further reducing the bandgap of LBG polymer by replacing S to heavier atom Se, polymer PBDTT-SeDPP has bandgap of 1.38eV. The better spectrum complementation and higher hole mobility lead to 9.5% tandem PCE. PDTP-DFBT:PC₇₁BM as LBG subcell gives even better performance, and the first double digit certified tandem polymer solar cell PCE of 10.6% (Fig. 7(c) and (d)).

Homogeneous tandem solar cell, in which the subcells adopt the same active materials, is not supposed to reduce thermalization loss in principle. But homogeneous tandem polymer solar cells were shown to be able to enhanced the solar cell efficiency

over single junction cell. For example, tandem cell with two identical PDTP-DFBT:PC₇₁BM subcells showed a PCE enhancement of over 25% – from 8 to 10.2% (Fig. 7(e)). It is worth mentioning that when the HOMO level of polymer is deeper, such as PDTP-DFBT, PEDOT:PSS will not be able to provide good enough energy alignment and hole transport function. Higher work function material MoO₃ was used to form a three layer MoO₃/PEDOT:PSS/ZnO interconnection layer for this efficient homogeneous tandem cell. Tandem cell with two identical PTB7-Th:PC₇₁BM subcells showed a PCE enhancement from ~9% to over 11%. The enhancement originates from the weakness of typical polymer solar cell materials (even champion materials like PDTP-DFBT and PTB7-Th) – the limited thickness for optimal device due to the low carrier mobility. Therefore using two thinner subcells to form tandem cell can overcome the limitation and achieve higher performance.

Triple junction tandem polymer solar cell was also successfully demonstrated. Based on the 10.6% heterogeneous tandem cell, a midbandgap PTB7-Th:PCBM cell was added into the device, and the efficiency was enhanced to 11.6%. The higher work function component in interconnection layer was further improved from thermally evaporated MoO₃ into solution processed WO₃. The success of the all solution processed triple junction polymer solar cell represents an important technology progress.

The donor-acceptor type low bandgap polymers, when goes below 1.4 eV bandgap, many have the absorption spectrum that absorb in UV and near infrared (NIR), but very low absorption in visible range. This makes possible high performance visibly transparent solar cells. Fig. 7(f) shows such a cell, where the top transparent electrode is from silver nanowire ink, and the whole process is free from vacuum. Other transparent electrode techniques include thin metal layer(s), dielectric/metal/dielectric multilayer, carbon nanotube, conducting polymer, polymer-CNT, polymer-metal nanowire composites etc. Transparent tandem solar cell can be used to tune the color, and achieve higher efficiency. Currently 7% transparent tandem solar cell has been realized with relatively high visible transparency. This shows how versatile the technology is, and broad applications the technology can offer.

Polymer solar cell research is very active with continuous discoveries and breakthroughs coming out. Future new polymer donor and acceptor materials, deep understanding and control of morphology, device structure innovations will continue improve OPV performance, establish it as a nontoxic, low cost, flexible and transparent form factor clean energy solution for novel PV applications.

See also: Light Harvesting in Organic Solar cells

Further Reading

- Bae, S.-H., *et al.*, 2016. Printable solar cells from advanced solution-processible materials. *Chem* 1, 197–219. [dx.doi.org/10.1016/j.chenpr.2016.07.010](https://doi.org/10.1016/j.chenpr.2016.07.010).
- Chen, H.Y., *et al.*, 2010. Silicon atom substitution enhances interchain packing in a thiophene-based polymer system. *Advanced Materials* 22, 371–375.
- Cheng, Y.-J., Yang, S.-H., Hsu, C.-S., 2009. Synthesis of conjugated polymers for organic solar cell applications. *Chemical Reviews* 109 (11), 5868–5923.
- Dou, L., *et al.*, 2015. Low-bandgap near-IR conjugated polymers/molecules for organic electronics. *Chemical Reviews* 15 (23), 12633–12665.
- Guo, X., *et al.*, 2012. High efficiency polymer solar cells based on poly(3-hexylthiophene)/indene-C70 bisadduct with solvent additive. *Energy & Environmental Science* 5, 7943.
- He, Z., *et al.*, 2015. Single-junction polymer solar cells with high efficiency and photovoltage. *Nature Photonics* 9, 174–179.
- Hirst, L.C., Ekins-Daukes, N.J., 2011. Fundamental losses in solar cells. *Progress in Photovoltaics: Research and Applications* 19, 286–293.
- Liang, Y., *et al.*, 2010. For the bright future-bulk heterojunction polymer solar cells with power conversion efficiency of 7.4%. *Advanced Materials* 22, 135–138.
- Li, G., *et al.*, 2005. High-efficiency solution processable polymer photovoltaic cells by self-organization of polymer blends. *Nature Materials* 4, 864–868.
- Li, N., Brabec, C.J., 2015. Air-processed polymer tandem solar cells with power conversion efficiency exceeding 10%. *Energy & Environmental Science* 8, 2902–2909.
- Peet, J., *et al.*, 2007. Efficiency enhancement in low-bandgap polymer solar cells by processing with alkane dithiols. *Nature Materials* 6, 497–500.
- Pope, M., Swenberg, C.E., 1999. *Electronic Processes in Organic Crystals and Polymers*. New York: Oxford University Press.
- Shockley, W., Queisser, H.J., 1961. Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics* 32, 510–519.
- Yu, G., *et al.*, 1995. Polymer photovoltaic cells – Enhanced efficiencies via a network of internal donor-acceptor heterojunctions. *Science* 270, 1789–1791.
- Zhao, W., *et al.*, 2016. Fullerene-free polymer solar cells with over 11% efficiency and excellent thermal stability. *Advanced Materials*. 4734–4739.

Dye-Sensitized Solar Cells

Lu-Yin Lin, National Taipei University of Technology (Taipei Tech), Taipei, Taiwan

Kuo-Chuan Ho, National Taiwan University, Taipei, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

History

The concept of dye-sensitized solar cells (DSSCs) based on a TiO_2 thin film originated from the year of 1972 (Fujishima and Honda, 1972). By constructing a chlorophyll-sensitized zinc oxide electrode, the photons were converted into electricity via the charge injection from the excited dye molecules into a wide band gap semiconductor for the first time (Tributsch, 1972). Numerous researchers started to study zinc oxide-based DSSCs since then. However, the relevant solar-to-electricity conversion efficiency is poor, probably due to flat surface of the semiconductor to provide only monolayer adsorption of the sensitizer. A breakthrough development was realized in the year of 1991 by Grätzel and his coworkers at the Laboratory of Photonic and Interfaces in the École Polytechnique Fédérale de Lausanne in Switzerland. They introduced nanoporous TiO_2 electrodes with rough surface to increase the light harvesting efficiency and successfully achieved a solar-to-electricity conversion efficiency (η) of 7% (O'Regan and Gratzel, 1991). Afterwards DSSCs experience a rapid growth since 2009, as shown in the annual publication number summarized in Fig. 1. In the year of 2011, a recorded η value of 12.3% was obtained for the DSSC composed of a zinc porphyrin dye (YD2-o-C8) co-sensitized with another organic dye (Y123) (Fig. 2), and $\text{Co}^{(\text{II/III})}$ tris(bipyridyl)-based redox electrolyte (Yella *et al.*, 2011). In the year of 2014, a DSSC coupled with a zinc porphyrin dye and $\text{Co}^{(\text{II/III})}$ tris(bipyridine)-based redox electrolyte achieved an impressive η value of 13.00% (Mathew *et al.*, 2014). Right after this achievement, a recorded η value of 14.30% was attained for a DSSC composed of the metal-free organic co-sensitizers and the redox electrolyte based on $\text{Co}^{(\text{II/III})}$ tris(phenanthroline) in the year of 2015 (Kakiage *et al.*, 2015). These breakthroughs inspire the researchers to pursue economical photovoltaic devices based on DSSCs more eagerly for the years to come.

Mechanism and Parameters

When the light illuminated the DSSC, the incident photons are absorbed by sensitizers adsorbed on the semiconductor surface. The sensitizers are excited from the ground state (S) to the excited state (S^*). Then the excited electrons would inject into the conduction band of the semiconductor, generating the oxidation of the sensitizer (S^+). These processes can be represented as Eqs. (1) and (2).



The injected electrons in the conduction band of the semiconductor would transport toward the conducting substrate, and then the electrons would reach the counter electrode (CE). Then the oxidized photosensitizer (S^+) would accept the electrons transported from the I^- ion redox mediator, and then be regenerated to the ground state (S), as shown in Eq. (3).



Simultaneously the I^- would be oxidized to I_3^- and diffuse toward the CE. Finally, it is reduced to I^- at the CE, as shown in Eq. (4).



The energy level diagram of a DSSC and the electron transfer routes are shown in Fig. 3.

Like other solar cells, the performance of the DSSC can be evaluated by using the open-circuit voltage (V_{OC}), the short-circuit current density (J_{SC}), the fill factor (FF), the solar-to-electricity conversion efficiency (η), the incident photon-to-current conversion efficiency (IPCE), and the charge transfer resistances measured by using the electrochemical impedance spectroscopy (EIS).

V_{OC} is determined by the difference in energy of the electron at the semiconductor and the CE for a typical photoelectrochemical cell, or the difference in energy level of the quasi-Fermi level of the semiconductor under illumination and the Nernst potential of the electrolyte solution. V_{OC} is measured with no current flow and is the maximum voltage which the DSSC can provide to an external circuit. Eq. (5) shows the relation of photovoltaic parameters to V_{OC} for regenerative photoelectrochemical systems.

$$V_{\text{OC}} = V_{\text{T}} \ln \left(1 + \frac{J_{\text{SC}}}{J_0} \right) \quad (5)$$

where V_{T} is the thermal voltage and J_0 is the reverse saturation current density.

J_{SC} is the maximum output current of a DSSC and is measured under illumination at the zero voltage across the cell. The magnitude of J_{SC} is determined by the product of the optical absorption efficiency, charge collection, and the solar irradiance,

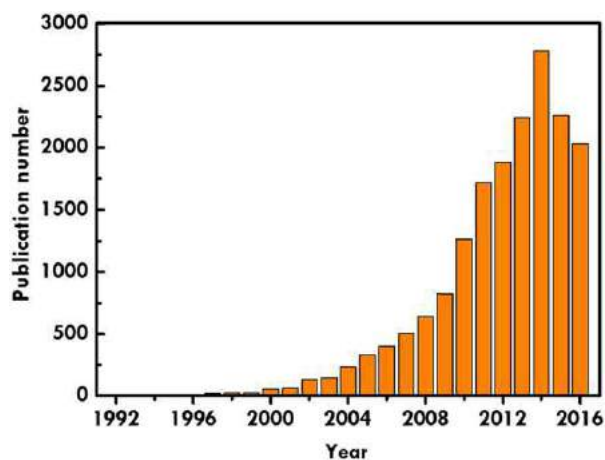


Fig. 1 Number of publications on the subject of dye-sensitized solar cells (DSSCs) from 1991 to December 2016. *Source:* Scopus.

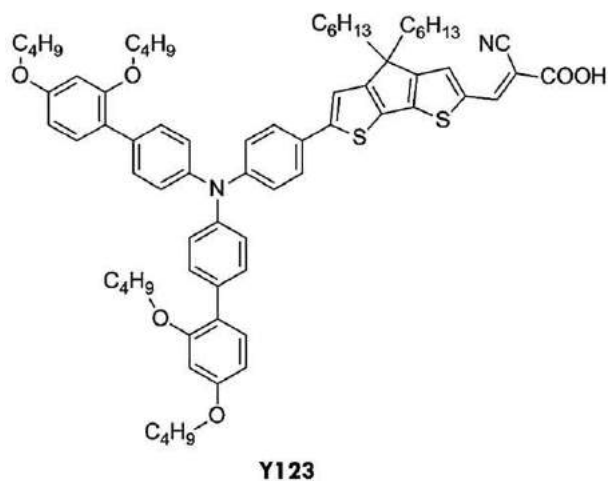
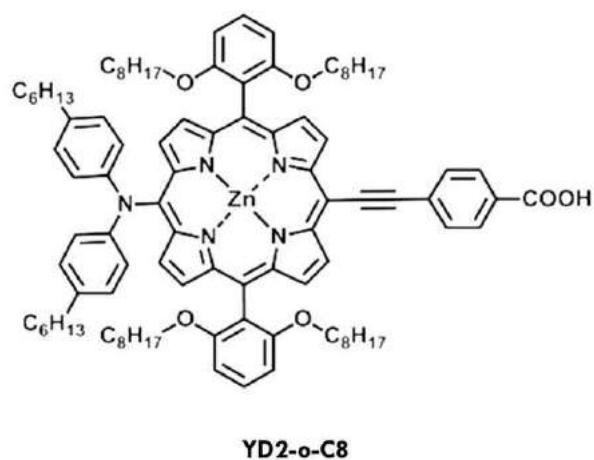


Fig. 2 The chemical structures of YD2-o-C8 and Y123 dyes.

integrated over the whole wavelengths of the incident light spectrum, as shown in Eq. (6).

$$J_{SC} = \int (1 - 10^{-Abs(\lambda)}) \times IQE(\lambda) \times e \times \Phi(\lambda) d\lambda \quad (6)$$

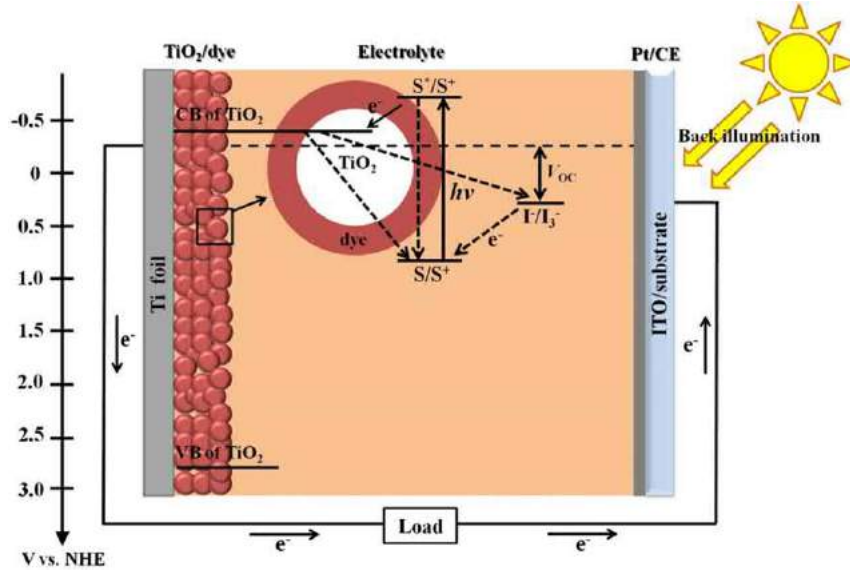


Fig. 3 The energy level diagram of a dye-sensitized solar cell (DSSC) and the illustration of the electron transfer routes in the DSSC.

where $\text{Abs}(\lambda)$ is the absorbance of the sensitizer as a function of the wavelength λ , e is the elementary electron charge, and $\Phi(\lambda)$ is the standard AM 1.5 irradiance spectrum. $\text{IQE}(\lambda)$ is the internal quantum efficiency for the hot electron injection, dye regeneration, and then charge collection.

FF is defined as the ratio of the maximum power output ($P_{\max}$) from a DSSC to the product of V_{OC} and J_{SC} . The fill factor can be calculated by using Eq. (7).

$$\text{FF} = \frac{P_{\max}}{V_{\text{OC}} \times J_{\text{SC}}} = \frac{V_{\max} \times J_{\max}}{V_{\text{OC}} \times J_{\text{SC}}} \quad (7)$$

where $V_{\max}$ and $J_{\max}$ respectively represent the photovoltage and photocurrent density corresponding to $P_{\max}$.

η depends on the total light intensity, the intensity spectral distribution, and the DSSC temperature. The η of a DSSC is calculated as the ratio between the maximum generated power ($P_{\max}$) and the incident power (P_{in}). The incident power is equal to the irradiance of the AM1.5 spectrum normalized to 1000 W/m^2 . Hence, the η of a DSSC can be determined by the V_{OC} , J_{SC} , FF, and P_{in} as shown in Eq. (8) as follows.

$$\eta = \frac{P_{\max}}{P_{\text{in}}} \times 100\% = \frac{V_{\text{OC}} \times J_{\text{SC}} \times \text{FF}}{P_{\text{in}}} \times 100\% \quad (8)$$

A typical current density–voltage curve is shown in Fig. 4 with the V_{OC} , J_{SC} and $P_{\max}$ indicated in the figure.

IPCE is defined as the ratio of the number of photo-generated electrons ($N_{\text{electrons}}$) flowing in the external circuit to the number of incident photons (N_{photons}) with a given wavelength, as shown in Eq. (9).

$$\text{IPCE}(\lambda) = \frac{N_{\text{electrons}}}{N_{\text{photons}}} \quad (9)$$

IPCE is also referred to as the external quantum efficiency (EQE). IPCE or EQE value is most often expressed as the percentage of the absorbed incident photons and the converted electrons as a function of the wavelength. The internal quantum efficiency (IQE) is also reported which represents the number of absorbed photons converted to electrons and indicates that how efficient the numbers of absorbed photons are converted into current. The IPCE value can also be calculated using Eqs. (10) and (11) as follows.

$$\text{IPCE}(\lambda) = \text{LHE}(\lambda) \times \text{IQE}(\lambda) \quad (10)$$

$$\text{LHE}(\lambda) = 1 - 10^{-\text{Abs}(\lambda)} \quad (11)$$

where $\text{LHE}(\lambda)$ is the light-harvesting efficiency.

Thus, the J_{SC} can also be represented as Eq. (12) as follows.

$$J_{\text{SC}} = \int \text{IPCE}(\lambda) \times e \times \Phi(\lambda) d\lambda \quad (12)$$

where e is the elementary charge and $\Phi(\lambda)$ is the photon flux.

It is important and useful to examine the $\text{IPCE}(\lambda)$ value not only in getting further information of $\text{IQE}(\lambda)$ or other parameters, but also in examining the validity of the measured photocurrent. If the measured J_{SC} value showed a discrepancy from the integral IPCE results in Eq. (12), there would be a spectral mismatch or variance between the solar simulator and the true solar emission.

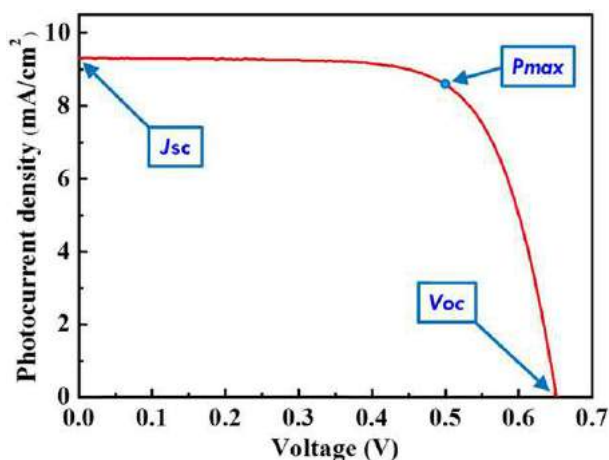


Fig. 4 A typical current density–voltage (J – V) curve with the V_{OC} , J_{SC} , and P_{max} indicated.

The steady state J – V measurements can only provide limited information about the electron transport and recombination rates. Hence, a dynamic technique like the EIS is required to study the charge transfer kinetics and therefore to get the information about the resistance and capacitance of different components in DSSCs. The primary investigated resistance at the interfaces of DSSCs from the charge transfer processes are the series resistance (R_s), the recombination resistance at the semiconductor/dye/electrolyte interface (R_{ct1}), the electron transfer resistance in the CE (R_{ct2}), and the diffusion resistance of the electrolyte causes (Z_w), which can be fitted from the Nyquist plot measured using the EIS technique. **Fig. 5** shows a typical Nyquist plot for a DSSC, in which the corresponding portions for fitting to get various resistances are also labeled in this figure.

Constructions

A DSSC is composed of a photoanode, a CE, and in-between embedded an electrolyte.

Photoanode

The photoanode is one of the important components of DSSC because of its functions in adsorbing dye molecules and transferring electrons. A high electron transport rate is required to reduce electron–hole recombination rate and enhance the conversion efficiency. The photoanode is consisted of a conducting substrate, a semiconductor layer deposited on the top of the substrate, and a sensitizer adsorbed on the semiconductor layer.

Conducting substrates

The selection of suitable conducting substrates is the key factor to determine the cost and the fabrication process of DSSCs, and therefore to influence the cell performance and long-term stability. Traditionally, transparent conducting oxide (TCO) glasses were utilized as the substrate of DSSCs, for example, fluorine-doped tin oxide ($\text{SnO}_2\text{:F}$) (FTO), tin-doped indium oxide ($\text{In}_2\text{O}_3\text{:Sn}$) (ITO), and aluminum-doped zinc oxide (ZnO:Al) (AZO) glasses. Despite of their good stability against oxygen and water, the shortages of high cost, fragility, rigidity, and shape limitations restrict their practical applications. In addition, the roll-to-roll production process cannot be used for this system. Therefore, flexible substrates with light weight and easy transportation properties have become an alternative choice.

One kind of the flexible substrates is conducting plastics, for example, ITO coated polyethyleneterephthalate (ITO/PET) and ITO coated polyethylenenaphthalate (ITO/PEN). The critical challenge for conducting plastics lies in the low temperature preparation of TiO_2 layer since their maximum endurable temperature is around 150°C . Normally the TiO_2 layer is sintered at higher temperature of around 400 – 500°C to get good electrical contact among particles. Additionally, the typical TiO_2 paste contains organic binder and viscous solvents which should be removed in a high temperature sintering process. Mechanical compression of nanocrystalline semiconductor powder and hydrothermal synthesis of metal oxide nanocrystals have been proposed to avoid applying the high temperature sintering process. Other kind of the flexible substrates is metal foils, which possess inherent high conductivity, light scattering, and high-temperature sintering abilities, for example, stainless steel (StSt), titanium foils, and zinc foils. Among them, the best is the Ti foil-based DSSC because of its characteristics of higher conductivity and superior corrosion resistance resulted from the existence of a passive film of TiO_2 on the surface of a Ti foil. However, there remain two major challenges to overcome for metal-based DSSCs. First, metal substrates are non-transparent and the pertinent DSSCs have to be illuminated from the back side, i.e., the Pt CE side. A loss of about 25% incident solar energy is encountered because the Pt CE partially reflects light and iodine in the electrolyte absorbs photons in the near UV region. Secondly, there would be serious formation of metal oxides after sintering, leading to the increase in the interfacial resistance and therefore the cell performance is reduced.

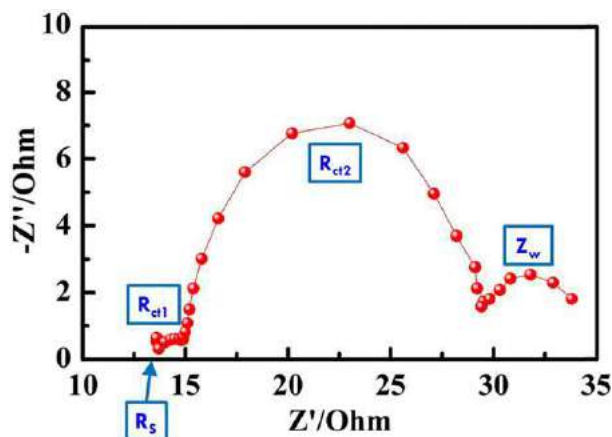


Fig. 5 A typical Nyquist plot for a dye-sensitized solar cell (DSSC).

Table 1 A partial list of the photovoltaic performance for the dye-sensitized solar cells (DSSCs) using different semiconductors on the photoanode

Semiconductor	Dye	V_{OC} (V)	J_{SC} (mA/cm ²)	FF (%)	η (%)
TiO ₂ ^a NP	N719	0.85	17.7	75	11.18
TiO ₂ ^b NR	N719	0.74	8.9	67	4.40
TiO ₂ ^c NF	N719	0.77	8.7	60	4.01
TiO ₂ ^d NT	Orange G	0.68	7.2	61	2.91
TiO ₂ ^a NP/TiO ₂ ^d NT	CYC-B1	0.71	12.9	73	6.68
ZnO ^a NP	N719	0.56	5.1	46	1.30
ZnO ^b NR	D149	0.58	2.8	40	0.66
ZnO ^c NF	N719	0.60	3.6	62	1.34
ZnO ^a NP/ZnO ^b NR	N719	0.59	7.0	52	2.17
SnO ₂ hollow spheres	N719	0.77	14.6	54	6.02
SnO ₂ flowers	N3	0.63	7.2	55	2.50
SnO ₂ ^e NC	N719	0.56	6.7	34	1.29
SnO ₂ ^f NS	N719	0.65	11.5	45	3.34
Nb ₂ O ₅ ^g NF	N3	0.77	6.7	59	3.05
Nb ₂ O ₅ ^e NC	N3	0.64	17.6	40	4.48
Nb ₂ O ₅ ^g NW	N3	0.77	6.7	59	3.05
Nb ₂ O ₅ mountains	N719	0.59	10.7	53	3.35

^aNP = nanoparticles

^bNR = nanorods

^cNF = nanofibers

^dNT = nanotubes

^eNC = nanochanneal

^fNS = nanosheet

^gNW = nanowire.

Semiconductor layers

The semiconductor layer is a junction with a large contact area between two continuous networks for dye adsorption and allows electron transferring through the DSSC.

Mesoporous oxides are the most commonly used materials as the semiconductor layer, including TiO₂, ZnO, SnO₂, and Nb₂O₅. **Table 1** lists the photovoltaic performance for the DSSCs using different semiconductors on the photoanode. The nanocrystalline TiO₂ is the most applied material because of its cheap, abundant, and nontoxic characteristics. The band gap of anatase TiO₂ is 3.2 eV at an absorption edge of 388 nm and that of rutile TiO₂ is 3.0 eV at an absorption edge of 413 nm. The crystalline of TiO₂ is classified into anatase, brookite, and rutile. The illustration for the crystal structures of TiO₂ is shown in **Fig. 6** (Zhang *et al.*, 2014). Anatase and rutile phases are the main TiO₂ crystallines but the brookite TiO₂ is rarely existed, which remains stable only at the temperature higher than 600°C. Moreover, rutile TiO₂ absorbs 4% of incident light in the near-UV region, resulting in generation of holes which would reduce the long-term stability of DSSCs (Yuvapragasam *et al.*, 2015). Therefore, anatase TiO₂ is most suitable as the semiconductor layer on the photoanode of DSSCs. The high dielectric constant of TiO₂ ($\epsilon = 80$

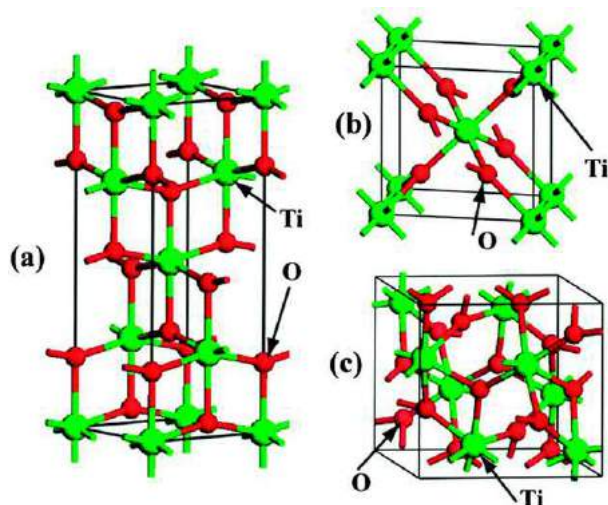


Fig. 6 TiO_2 crystallographic phases of (a) anatase, (b) brookite and (c) rutile. Reproduced with permission from Zhang, J., Zhou, P., Liu, J., Yu, J., 2014. New understanding of the difference of photocatalytic activity among anatase, rutile and brookite TiO_2 . *Physical Chemistry Chemical Physics* 16, 20382. Copyright (2014) by Royal Society of Chemistry.

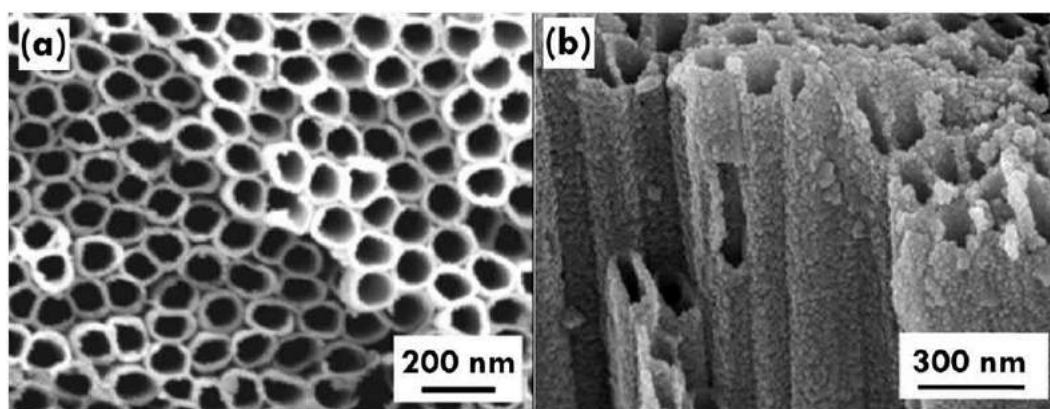


Fig. 7 (a) The top-view and (b) the side-view scanning electron microscope (SEM) images for the 1D TNT array. Reproduced with permission from Lin, L.Y., Yeh, M.H., Lee, C.P., Chen, Y.H., Vittal, R., Ho, K.C., 2011. Metal-based flexible TiO_2 photoanode with titanium oxide nanotubes as the underlayer for enhancement of performance of a dye-sensitized solar cell. *Electrochimica Acta* 57, 270. Copyright (2011) by Elsevier.

for anatase) provides good electrostatic shielding of the injected electron from the oxidized dye molecule, thus preventing the electrons from recombination before the reduction of the dye molecule by the electrolyte. High refractive index of TiO_2 ($n=2.5$ for anatase) results in efficient diffuse scattering of light inside the porous photoanode and hence significantly enhances the light absorption.

Most photoanodes of DSSCs were based on a nanocrystalline TiO_2 film with a particle size of about 20 nm. However, the performance of the pertinent DSSCs was limited by its slow electron transport, which is in the range of a few milliseconds. The random distribution of the TiO_2 particles and the associated structural disorder of the film lead to an enhancement on the dye-injected free electrons scattering, and thereby to a reduction of electron mobility. Therefore, altering the morphology and structure of the TiO_2 film is one of the strategies for improving the efficiency of a DSSC. Several nanotubular architectures have been investigated for potential enhancement of electron percolation pathways as well as for improved ion diffusion at the semiconductor/electrolyte interface. Ordered and strongly interconnected nanoscale architecture of a TiO_2 film enables a great improvement in the electron transport in the film, for example, a vertically oriented array of one-dimensional (1D) TNTs (**Fig. 7**), which have been proved to exhibit better electron transport and therefore to reduce the loss of electrons by recombination (Lin *et al.*, 2011).

Sensitizers

The sensitizer carries out the light harvesting function and influences the electron transfer processes occurring at the semiconductor/dye/electrolyte interfaces. An efficient sensitizer has a broad and strong absorption, excellent chemical stability,

reduction potential sufficiently higher than the conduction band edge of the semiconductor, and an oxidation potential sufficiently lower than the redox potential of the electrolyte. The primary classifications of sensitizers are metal-based sensitizers and metal-free organic sensitizers.

The metal-based sensitizers comprising the chromophores of ruthenium complexes have been widely studied for DSSCs, which can achieve solar-to-electricity conversion efficiencies of around 11% under AM1.5 conditions, due to the favorable photoelectrochemical properties as well as the excellent chemical stability and intense light absorption over a wide visible range. Several efficient Ru-based complex photosensitizers were developed by Grätzel *et al.* The first famous Ru-based complex photosensitizer is *cis*-bis(isothiocyanato)bis-(4,40-dicarboxylic acid-2,20-bipyridine) Ru (II), denoted as N3 or red dye (Nazeeruddin *et al.*, 1993). Afterwards, the other two efficient dyes were synthesized by the same group. One is the black dye or denoted as N749 dye with the formula of tris(isothiocyanato)-2,20,2''-terpyridyl-4,40,4''-tricarboxylate Ru(II) complex, and the other is denoted as N719 dye, with the formula of di(tetrabutylammonium)*cis*-bis(isothiocyanato)bis-(4-carboxylic acid-40-carboxylate-2,20-bipyridine)Ru(II). The structures for these efficient sensitizers are shown in Fig. 8. Table 2 gives a partial list of the photovoltaic parameters for the DSSCs using Ru-based complex photodyes as the sensitizer. The disadvantage of the ruthenium complex dyes is that they contain a

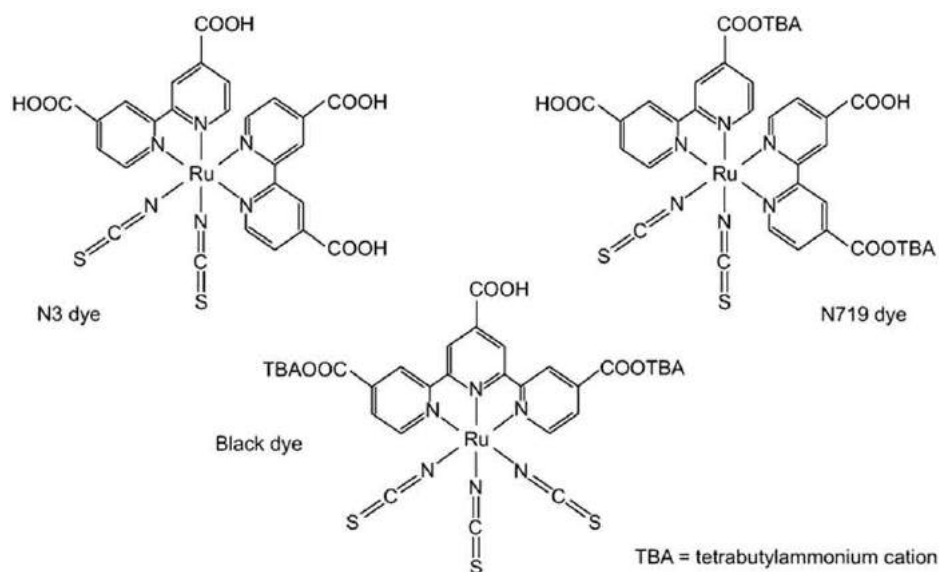


Fig. 8 The chemical structures of N3 dye, black dye, and N719 dye.

Table 2 A partial list of the photovoltaic parameters for the dye-sensitized solar cells (DSSCs) using Ru-based complex dyes as the sensitizer

Ru complex sensitizer	V_{OC} (V)	J_{SC} (mA/cm ²)	FF (%)	η (%)
<i>cis</i> -[Ru(^a dcBH ₂) ₂ (NCS) ₂]	0.64	5.0	76	10.4
N719	0.85	17.7	75	11.1
N621	0.78	16.8	73	9.6
Black dye	0.74	20.9	72	11.1
<i>cis</i> -[Ru(^a dcBH ₂) ₂ (NCS) ₂]	0.70	15.0	74	7.8
[Ru(^a dcBH ₂) ₂ (^b acac)]Cl	0.67	13.2	68	6.0
Ru(^c bmipy)(^d bpy-PO ₃ H ₂)(NCS)	0.60	8.6	70	4.1
[Ru(^a dcBH ₂) ₂ (^e phdmp)]Cl ₂	0.42	0.60	38	3.8
Ru(bpy-(CH ₂ OH) ₂)(NCS) ₂	0.51	10.0	48	3.8
Ru(^a dcBH ₂) ₂ (^f qdt)	0.60	11.1	70	3.7
[Ru(^a dcBH ₂) ₂ Cl ₂]	0.57	2.6	38	2.1
[Ru(^a dcBH ₂) ₂ (^g aphb)]	0.58	5.5	48	1.5

^adcBH₂ = 4,4'-(CO₂H)₂-2,2'-bipyridine

^bacac = acetoacetate

^cbmipy = benzimidazol

^dbpy = bipyridine

^ephdmp = pteridinedione

^fqdt = quinoxaline-2,3-dithiolate

^gaphb = phenanthroline-5-yliminomethyl-phenol.

Table 3 A partial list of the photovoltaic parameters for the dye-sensitized solar cells (DSSCs) using metal-free organic dyes as the sensitizer

Metal-free organic dyes	V_{oc} (V)	J_{sc} (mA/cm ²)	FF (%)	η (%)
C219	0.77	17.94	73	10.10
C229	0.85	15.3	73	9.40
C218	0.95	13.3	74	9.30
WS-9	0.70	18.0	72	9.04
WS-2	0.65	17.7	76	8.70
IQ2	0.77	15.7	70	8.50
WS-5	0.78	13.2	78	8.02
WS-6	0.67	15.0	77	7.76
TTZ5	0.71	18.3	59	7.71
OHxDPTP	0.77	13.8	72	7.64
BT-I	0.78	15.7	61	7.51
YCD01	0.76	13.4	73	7.43
R4	0.67	15.2	66	6.72
VP3	0.79	12.62	63	6.29
DTInDT	0.69	13.54	71	6.70
CR204	0.67	14.9	66	6.49
C278	0.84	19.6	73	12.00
C275	0.96	17.0	77	12.50

heavy metal which is expensive and harmful to the environment. Therefore, the metal-free organic dyes have been proposed as an alternative for DSSCs.

There are some advantages for the metal-free organic dyes to be used in DSSCs. First, the metal-free organic dyes are relatively inexpensive, since they can be easily synthesized and their resources are more abundant than that of the noble metals. Secondly, the extinction coefficients and the absorption coefficients of organic dyes are relatively higher than those of the noble metal-based ones. Organic dyes based on porphyrin, perylene, cyanine, xanthenes, merocyanine, coumarin, hemicyanine, indoline, diketopyrrolopyrrole, and ethylenedioxythiophene have been widely investigated for DSSCs due to their phototherapeutic and photochemical applications. **Table 3** is a partial list of the photovoltaic parameters for the metal-free organic dye-based DSSCs repeated in literatures.

Electrolytes

Electrolyte is one of the crucial parts overtaking the dye regeneration after electron injection and the electron transfer between the photoanode and the CE. The electrolyte can be classified into liquid electrolyte, quasi-solid-state electrolyte, and solid-state electrolyte.

The liquid electrolyte is composed of redox couples, additives, and solvents. The redox couple is in charge of transferring charges between the photoanode and the CE, as well as the dye regeneration. The redox couple possesses a redox potential less negative than the oxidized level of the sensitizer. Also, fast electron transfer kinetics at the CE, slow electron recombination kinetics at the sensitizer, small absorption in the visible region, as well as the noncorrosive, reasonable stability, and good diffusion properties are indispensable factors for the redox couples. The redox couple of lithium iodide/iodine was first used in DSSC by Hagfeldt and coauthors in an organic liquid electrolyte (Hagfeldt and Graetzel, 1995). However, the I_3^- is extremely corrosive toward metals such as copper or silver, and a significant part of visible light would be absorbed by the I^-/I_3^- based electrolyte. Hence, many other redox couples were proposed to replace this redox couple nowadays, for example, cobalt complexes and inorganic redox couples such as Br^-/Br_3^- , $SCN^-/(SCN)_2$, $SeCN^-/(SeCN)_2$, and S^{2-}/S , or other organic compounds. A recorded η of 12.3% was attained for the DSSC with $Co^{(II/III)}$ tris(bipyridyl)-based redox couple (**Fig. 9**) in its electrolyte. **Table 4** is a partial list of the photovoltaic parameters for the DSSCs using different redox couples in the liquid electrolytes. On the other hand, the additives were used in the electrolyte to enhance the photovoltaic performances. Specific cations like guanidium or alkali cations are typically used to increase the photocurrent density. The cations in the electrolyte would adsorb on the surface of the semiconductor layer to render the conduction band of the semiconductor shift toward more positive potentials, thus affecting the electron injection dynamics of the dye molecules in the excited state. Also, nitrogen-containing heterocyclic compounds like 4-tert-butylpyridine (4-TBP) and pyridine analogs of the N-alkylbenzimidazole derivatives are purposely used to improve the V_{oc} . Lastly, the organic solvents and ionic liquids are primarily selected as the solvents for the liquid electrolyte. The organic solvents like acetonitrile (ACN), ethylene carbonate (EC), 3-methoxypropionitrile (MPN), propylene carbonate (PC), N-methylpyrrolidone (NMP) and water have been widely investigated for liquid electrolyte-based DSSCs. The ACN is the first choice organic solvent in considering of the solar-to-electricity conversion efficiency of DSSC, due to its low viscosity (0.34 cp at 25°C) and good solubility to dissolve salts and organic components. However, the encapsulation of volatile organic solvents is a critical issue owing to their high vapor pressures. Therefore, the quasi-solid-state electrolyte and solid-state electrolyte were used to enhance the encapsulation of the DSSCs.

The quasi-solid-state electrolyte is consisted of a polymer network swollen with liquid electrolytes or ionic liquid electrolytes. The quasi-solid-state electrolyte possesses both cohesive property of solid and diffusive transport property of liquid, and therefore

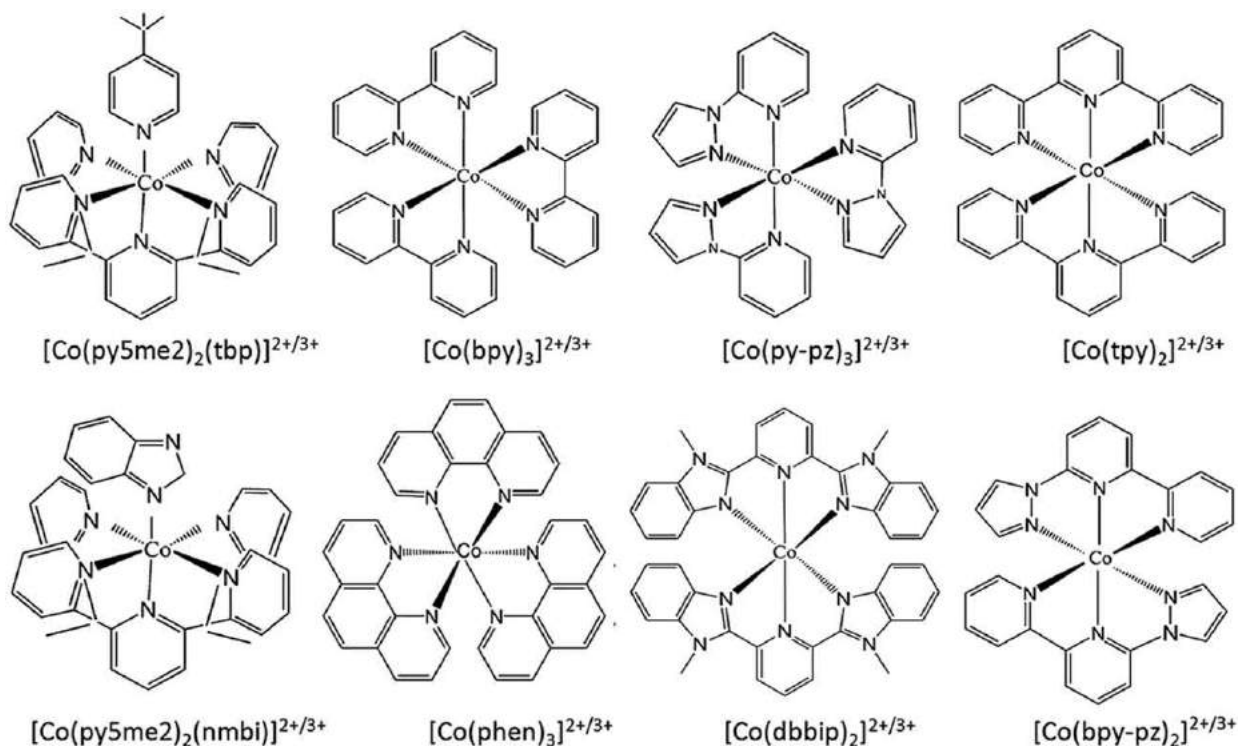


Fig. 9 The structures of the $\text{Co}(\text{II/III})$ tris(bipyridyl)-based redox couple. py5me2 = 2,6-bis(1,1-bis(2-pyridyl)ethyl)pyridine; tbp = 4-tert-butylpyridine; bpy = 4,4'-bipyridine; py-pz = 2-pyridyl pyrazolate; tpy = 2,2':6',2''-terpyridine; nmbi = N-methylbenzimidazole; phen = 1,10-phenanthroline; dbbip = 2,6-bis(1'-butylbenzimidazol-2'-yl)pyridine; bpy-pz = 6-(1H-pyrazol-1-yl) 2,20-bipyridine.

Table 4 A partial list of the photovoltaic parameters for the dye-sensitized solar cells (DSSCs) using different redox couples in the liquid electrolytes

Redox couple	Sensitizer	V_{oc} (V)	J_{sc} (mA/cm^2)	FF (%)	η (%)
I^-/I_3^-	N719	0.79	17.0	77	10.29
I^-/I_3^-	C219	0.77	17.94	73	10.10
$\text{Co}^{2+}/\text{Co}^{3+}$	C275	0.96	17.0	77	12.50
$\text{Co}^{2+}/\text{Co}^{3+}$	C229	0.85	15.3	73	9.40
$[\text{Co}(\text{bpy})_3](\text{B}(\text{CN})_4)_2/[\text{Co}(\text{bpy})_3](\text{B}(\text{CN})_4)_3$	YD2-o-C8	0.97	17.3	71	11.9
$[\text{Co}(\text{bpy})_3](\text{B}(\text{CN})_4)_2/[\text{Co}(\text{bpy})_3](\text{B}(\text{CN})_4)_3$	YD2-o-C8/Y123	0.94	17.7	74	12.3
^a TEMPO/TEMPO ⁺	D149	0.83	9.4	70	5.4
^b T ⁻ / ^c T ₂	Z907Na	0.68	16.2	58	6.4
^d Fc/FcPF ₆	Carbz-PAHTDDT	0.84	12.2	73	7.5
^e TMTU/TMFDS ²⁺	N3	0.66	9.45	49	3.1

^aTEMPO = 2,2,6,6-tetramethyl-1-piperidinyloxy radical

^bT = 1-methyl-1-H-tetrazole-5-thiolate anion

^cT₂ = dimer of T⁻

^dFc = Ferrocene

^eTMTU/TMFDS²⁺ = tetramethylthiourea/tetramethylformammonium disulfide.

the quasi-solid-state electrolytes offer better long-term stability than that of the liquid electrolytes. The quasi-solid-state electrolytes can be made by different methods and being classified into various sorts, for example, composite polymer electrolytes, quasi-solid ionic liquid electrolytes, and thermosetting polymer electrolytes. Table 5 lists the used sensitizers and the photovoltaic parameters for the DSSCs based on the three types of quasi-solid-state electrolytes mentioned above.

The solid-state DSSC applied the hole transport materials (HTMs) or the solid-state electrolyte containing I^-/I_3^- redox couple as the medium to solve the solvent leakage and corrosion issues. Major research interests focus on the organic hole conductor, 2,2',7,7'-tetrakis(*N,N*-di-*p*-methoxyphenyl-amine)-9,9'-spiro-bifluorene (spiro-MeOTAD), which was originated in 1998 by Bach *et al.* (1998). To enhance the cell performance of the solid-state DSSC, Kojima *et al.* used two organolead halide perovskite

Table 5 A partial list of the photovoltaic parameters for the dye-sensitized solar cells (DSSCs) based on the quasi-solid-state electrolyte

Electrolyte	Dye	V_{oc} (V)	J_{sc} (mA/cm ²)	FF (%)	η (%)
Composites polymer electrolyte					
Acetamide/ ^a PEO/LiI/I ₂	N3	0.74	19.81	52	9.0
^b PEO ₁₃ Blm-I/ ^c MPN	N719	0.73	11.61	66	5.6
^d MPII/MPII-modified Al ₂ O ₃	N719	0.72	17.70	60	7.6
^e PAN/LiI/activated carbon	N719	0.77	16.82	64	8.4
^f PMMA-EA/TiO ₂	N719	0.79	12.68	68	6.9
Quasi-solid ionic liquid electrolyte					
^g PEG/ ^h EMImI	N719	0.69	20.02	52	7.1
ⁱ PMImI/PVDF-HFP	Z907	0.67	11.29	71	5.3
^k POEI-II/MWCNT	N719	0.78	14.50	67	7.7
^j PVDF-HFP/ ^a PEO	N719	0.81	5.45	74	3.3
^m IL/ ⁿ [PBVIIm][TFSI]	N719	0.61	10.57	68	4.4
Thermosetting polymer electrolytes					
^j PVDF-HFP	N719	0.75	16.26	68	8.4
Paa/PEG/ ^a PPY	N719	0.71	13.34	71	6.3
^o P(VA-co-MMA)/ ^c MPN	N719	0.74	16.66	74	9.1
^p PAA-PEG/graphene	N719	0.73	15.00	68	7.7

^aPEO = poly(ethylene oxide)^bPEO₁₃Blm-I = dicationic bis-imidazolium diiodide salts^cMPN = 3-methoxypropionitrile^dMPII = 1-methyl-3-propylimidazolium iodide^ePAN = polyacrylonitrile^fPMMA-EA = poly(methyl methacrylate-co-ethyl acrylate)^gPEG = polyethylene glycol^hEMImI = 1-ethyl-3-methylimidazolium iodideⁱPMImI = 1-propyl-3-methylimidazolium iodide^jPVDF-HFP = poly(vinylidene fluoride-co-hexafluoropropylene)^kPOEI-II = poly(oxyethylene)-imide-imidazole^lMWCNT = multi-walled carbon nanotubes^mIL = ionic liquidⁿ[PBVIIm][TFSI] = poly(1-butyl-3-vinylimidazolium bis(trifluoromethanesulfonyl)imide)^oP(VA-co-MMA) = polyvinyl(acetate-co-methyl methacrylate)^pPAA-PEG = poly(acrylic acid)-poly(ethylene glycol)^qPPy = polypyrrole.

nanocrystals, CH₃NH₃PbBr₃ and CH₃NH₃PbI₃ to sensitize TiO₂ for the visible-light conversion, and a η value of 3.8% was achieved for the photoelectrochemical cell (Kojima *et al.*, 2009). Afterwards, another type of the solid-state TiO₂ solar cell, the lead iodide perovskite (CH₃NH₃PbX₃) solar cell, was proposed to have promising photovoltaic performance. A recorded η value of 23% was attained for the perovskite solar cell in the year of 2016 by Chen *et al.* (2016).

Counter electrodes

The CE plays a role of transferring electrons from the external circuit to the electrolyte to reduce the oxidized redox couple used as a mediator in regenerating the sensitizer after electron injection, so as to generate the photocurrent. High electrocatalytic activity, excellent electrical and thermal conductivities, high corrosion resistance and large electrochemical surface area, as well as high exchange current densities and low charge transfer resistances are important requirements for a CE of DSSCs.

Platinum is the most commonly used catalyst in the CE, due to the high conductivity, high stability, and high catalytic activity toward I₃[−] reduction. Pt can be deposited on the transparent conductive oxide (TCO) glasses by various methods, for example, spin-coating, screen-printing or sputtering. Modifying the structural characteristics and facets of Pt were verified to be useful in enhancing its catalytic activity. The high surface area and therefore the high amount of active sites available for reducing I₃ can be obtained by varying the morphology of Pt, like the Pt nanoflowers as shown in Fig. 10. The catalytic and photovoltaic behavior of the CE toward the I₃[−] reduction also varied with different facets of Pt (Zhang *et al.*, 2013). However, since Pt is one of the expensive and rare metals in the earth, it is not suitable for the commercial production of DSSCs. Developing alternative materials to Pt as the catalyst on the CE of DSSCs is expected to decrease the cost of DSSCs.

Other cheaper materials were proposed to reduce the production cost of the cells. Conducting polymers like poly(3,4-ethylenedioxythiophene) (PEDOT), PEDOT doped with poly(4-styrene sulfonate) (PEDOT:PSS), polypyrrole (PPy), polyaniline (PANI), and poly(3,4-propylenedioxythiophene) (PProDOT) have been used as the catalyst on the CE of

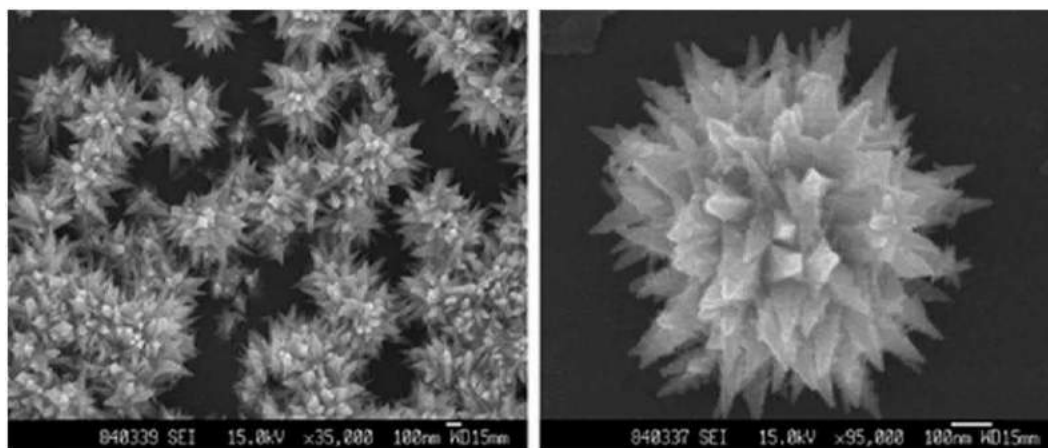


Fig. 10 The scanning electron microscope (SEM) images for the Pt nanoflowers. Reproduced with permission from He, Y.B., Li, G.R., Wang, Z. L., Ou, Y.N., Tong, Y.X., 2010. Pt nanorods aggregates with enhanced electrocatalytic activity toward methanol oxidation. *The Journal of Physical Chemistry C* 114 (45), 19175 Copyright (2010) by Royal Society of Chemistry.

Table 6 A partial list for the photovoltaic parameters for the dye-sensitized solar cells (DSSCs) using different catalysts on the CE

Counter electrode (CE)	Sensitizer	V_{oc} (V)	J_{sc} (mA/cm ²)	FF (%)	η (%)
PEDOT	N719	0.76	17.5	64	8.5
PEDOT	Y123	1.02	12.2	69	8.6
PPy	N719	0.74	15.0	68	7.7
PANI	N719	0.71	14.6	69	7.2
PANI	N719	0.71	17.2	59	7.2
PProDOT	Y123	1.00	13.1	77	10.1
PEDOT:PSS/TiN	CYC-B1	0.68	14.2	69	6.67
Glassy carbon	N719	0.71	16.5	65	7.6
Carbon nanotube (CNT)	N719	0.80	13.3	65	7.0
C60 (fullerene)	N719	0.76	11.6	32	2.8
N-doped carbon	N3	0.71	15.5	64	7.0
Graphene	Perovskite	0.70	13.1	63	5.9
^a SWCNT	Perovskite	0.72	13.0	52	5.0
Boron-doped CNT	N719	0.70	17.1	67	8.0

^aSWCNT = single-walled carbon nanotube.

DSSCs, due to their unique properties such as high conductivity, low cost, good stability, and high catalytic activity for I_3^- reduction. Among the conducting polymers, PANI has been one of the most intensively studied conducting polymers due to its easy synthesis, high conductivity, good environmental stability, and interesting redox properties in recent years. The performance of the conducting polymer can be enhanced by designing the morphology to achieve optimized surface roughness or doping foreign anions such as SO_4^{2-} , ClO_4^- , BF_4^- , Cl^- , and $tosyl^-$ into the conducting polymer film to increase the conductivity. However, the major problem of the conducting polymer is that the thick layer is required to achieve sufficient catalytic activities.

On the other hand, carbon materials like glassy carbon, graphene, and carbon nanotubes (CNT), are interesting low-cost alternatives to Pt. Carbonaceous materials have the advantages of high heat resistance, high electronic conductivity, high electrocatalytic activity toward I_3^- reduction, and high corrosion resistance. The power conversion efficiency and the fill factor of DSSCs are strongly dependent on the thickness of the carbon layer, which can be increased up to 10 μm to improve the fill factor of DSSCs. However, the main drawback of carbon-based CE is the requirement of large dosage to attain the targeted catalytic activity and the poor connection between the carbon film and the substrate. The photovoltaic parameters for the DSSCs using different catalysts on the CE are shown in Table 6.

Future Prospects

DSSCs have emerged as one of the cheap alternatives to the traditional silicon solar cells for harnessing the immense solar energy. Great progresses have been made for DSSCs since it was originated in 1972. The major aims are to enhance its solar-to-electricity

conversion efficiency and long-term durability. Developing efficient and novel sensitizers for converting light with broader wavelength especially in the near IR region into multiple electrons, constructing well-defined structures of semiconductors for providing more dye-adsorbed sites and the enhancing charge transfer among the photoanode, and designing suitable catalysts on the CE for accelerating the redox reactions in the electrolyte are necessary to achieve a highly efficient DSSC. Also, gel-type or solid-state electrolytes are needed to exploit for enhancing the long-term operating stability. Eventually, the detailed studies on the charge transfer resistance at the interfaces and the electron lifetime in the electrodes should be understood for selecting more effective routes to improve the performance of DSSCs.

See also: Organic Semiconductors

References

- Bach, U., Lupo, D., Comte, P., *et al.*, 1998. Solid-state dye-sensitized mesoporous TiO₂ solar cells with high photon-to-electron conversion efficiencies. *Nature* 395, 583–585.
- Chen, B., Bai, Y., Yu, Z., *et al.*, 2016. Efficient semitransparent perovskite solar cells for 23.0%-efficiency perovskite/silicon four-terminal tandem cells. *Advanced Energy Materials* 6, 1601128.
- Fujishima, A., Honda, K., 1972. Electrochemical photolysis of water at a semiconductor electrode. *Nature* 238, 37–38.
- Hagfeldt, A., Graetzel, M., 1995. Light-induced redox reactions in nanocrystalline systems. *Chemical Reviews* 95, 49–68.
- Kakiage, K., Aoyama, Y., Yano, T., *et al.*, 2015. Highly-efficient dye-sensitized solar cells with collaborative sensitization by silyl-anchor and carboxy-anchor dyes. *Chemical Communications* 51, 15894–15897.
- Kojima, A., Teshima, K., Shirai, Y., Miyasaka, T., 2009. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the American Chemical Society* 131, 6050–6051.
- Lin, L.Y., Yeh, M.H., Lee, C.P., *et al.*, 2011. Metal-based flexible TiO₂ photoanode with titanium oxide nanotubes as the underlayer for enhancement of performance of a dye-sensitized solar cell. *Electrochimica Acta* 57, 270–276.
- Mathew, S., Yella, A., Gao, P., *et al.*, 2014. Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers. *Nature Chemistry* 6, 242–247.
- Nazeeruddin, M.K., Kay, A., Rodicio, I., *et al.*, 1993. Conversion of light to electricity by cis-X₂bis(2,2'-bipyridyl-4,4'-dicarboxylate)ruthenium(II) charge-transfer sensitizers (X = Cl⁻, Br⁻, I⁻, CN⁻, and SCN⁻) on nanocrystalline titanium dioxide electrodes. *Journal of the American Chemical Society* 115, 6382–6390.
- O'Regan, B., Gratzel, M., 1991. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO₂ films. *Nature* 353, 737–740.
- Tributsch, H., 1972. Reaction of excited chlorophyll molecules at electrodes and in photosynthesis. *Photochemistry and Photobiology* 16, 261–269.
- Yella, A., Lee, H.-W., Tsao, H.N., *et al.*, 2011. Porphyrin-sensitized solar cells with cobalt (II/III)-based redox electrolyte exceed 12 percent efficiency. *Science* 334, 629–634.
- Yuvapragasam, A., Muthukumarasamy, N., Agilan, S., *et al.*, 2015. Natural dye sensitized TiO₂ nanorods assembly of broccoli shape based solar cells. *Journal of Photochemistry and Photobiology B* 148, 223–231.
- Zhang, B., Wang, D., Hou, Y., *et al.*, 2013. Facet-dependent catalytic activity of platinum nanocrystals for triiodide reduction in dye-sensitized solar cells. *Scientific Reports* 3, 1836.
- Zhang, J., Zhou, P., Liu, J., Yu, J., 2014. New understanding of the difference of photocatalytic activity among anatase, rutile and brookite TiO₂. *Physical Chemistry Chemical Physics* 16, 20382–20386.

Organic Inorganic Hybrid Perovskite Materials and Devices

Yihua Chen, Ning Zhou, and Huanping Zhou, Peking University, Beijing, P.R. China

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic–inorganic hybrid halide perovskites have received increasing attention in the photovoltaic research community due to its remarkable optoelectronic properties, such as high carrier mobility, tunable band gap, long diffusion lengths and long lifetime. The efficiency record of perovskite solar cells has experienced a dramatic progress from 3.8% in 2009 to more than 20%, which makes perovskite one of the most competitive and promising candidate materials for solar cells. Besides, the perovskite materials have also proven to be feasible in a large variety of other optoelectronic applications, for example, light emitting diode, laser, detector, and transistors as well. In this section, the perovskite materials and the corresponding devices are reviewed regarding the progress in the crystal structure, optoelectronic property, and device working mechanism.

Crystal Structure

Organic–inorganic hybrid halide perovskite is a category of materials with the chemical formula of ABX_3 , where A and B sites are often occupied with two different cations with different ionic radii. A site occupant is generally methylammonium ($CH_3NH_3^+$), formamidinium ($NH_2CH=NH_2^+$) or cation cesium (Cs^+), rubidium (Rb^+), and X is halide anion, such as I^- , Br^- and Cl^- (Green *et al.*, 2014). The crystal structure of perovskites is depicted in Fig. 1(a). Their crystallographic structure and pertaining stability can be guided by a tolerance factor t proposed by Goldschmidt and an octahedral factor μ . In theory, $t = (R_A + R_X) / \{\sqrt{2} (R_B + R_X)\}$ (Yin *et al.*, 2015) where R_A , R_B and R_X are the ionic radii of the corresponding ions, and μ is defined as the ratio R_B/R_X . To achieve stable halide perovskite, it typically requires $0.81 < t < 1.11$ and $0.44 < \mu < 0.90$. If t is lower, it is intended to form the less symmetric structures such as tetragonal (β phase) or orthorhombic (γ phase) perovskite (Lu *et al.*, 2016). In contrast, larger t ($t > 1$) could result in destabilized three dimensional (3D) B–X framework, leading to a two-dimensional (2D) layered structure (Yin *et al.*, 2015). μ is directly correlated to the BX_6 octahedral structure. When $\mu > 0.442$, we often observe BX_6 octahedron in a stable assembly. However, lower μ value will destabilize BX_6 octahedral assembly (Li *et al.*, 2008). A summary of both calculated tolerance factors and octahedral factors for commonly seen halide perovskites are provided in Fig. 1(b) (Chen *et al.*, 2015). Along with different ions and structural stabilities, different perovskites possess different transition temperatures. As the archetypal halide perovskite of $CH_3NH_3PbI_3$, the α to β , and β to γ phase transitions occur at around 330 and 160K, respectively (Ball *et al.*, 2013). Some relevant crystallographic data of different perovskite materials with phases in high, intermediate or low temperatures (e.g., α , β , and γ phases, respectively), are summarized in Table 1 (Chen *et al.*, 2015).

It is generally believed that the cation at A site has shown no apparent effects on the electronic configurations and thus the band structure of the resultant materials, but plays the role of making charge compensation within the lattice. The A cation has been observed to influence the optical properties of perovskite by deforming the MX_6^{4-} octahedron framework when intercalating into the crystal structures. A larger (smaller) A cation can cause the entire lattice to expand (contract), and lead to the change of B–X bond lengths, which has been demonstrated to be important in determining the bandgap (Borriello *et al.*, 2008; Kulkarni *et al.*, 2014). For example, the use of formamidinium ($HC(NH_2)_2^+$) instead of $CH_3NH_3^+$ as A cation of the perovskite can shift the band

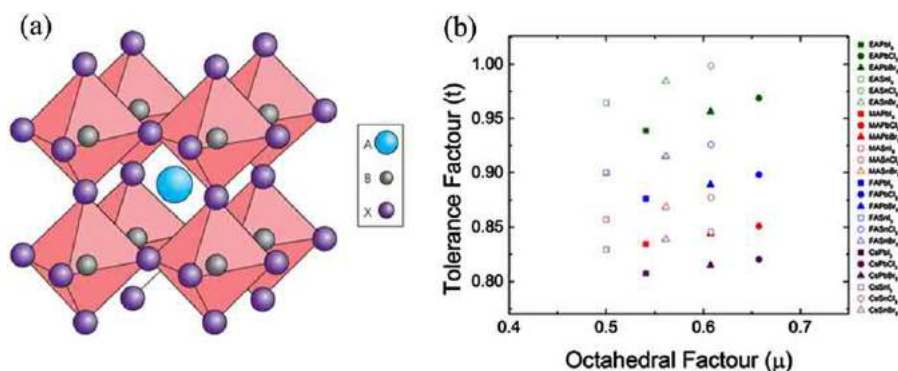


Fig. 1 (a) Cubic perovskite crystal structure; (b) calculated octahedral and tolerance factors for various combinations of commonly employed organic and inorganic hybrid perovskite components.

Table 1 Structural phase transformations for commonly employed hybrid perovskites

PVSK	Phase	Temperature (K)	Structure	Space group	Lattice parameters(Å)	Volume (Å ³)
MAPbI ₃	α	400	Tetragonal	P4mm	$a=6.3115; b=6.3115; c=6.3161$	251.6
	β	293	Tetragonal	I4cm	$a=8.849; b=8.849; c=12.642$	990
	γ	162–172	Orthorhombic	Pna2 ₁	$a=5.673; b=5.628; c=11.182$	959.6
MAPbCl ₃	α	> 178.8	Cubic	Pm3m	$a=5.675$	182.2
	β	172.9–178.9	Tetragonal	P4/mmm	$a=5.655; c=5.630$	180.1
	γ	< 172.9	Orthorhombic	P222 ₁	$a=5.673; b=5.628; c=11.182$	375
MAPbBr ₃	α	> 236.9	Cubic	Pm3m	$a=5.901$	206.3
	β	155.1–236.9	Tetragonal	I4/mcm	$a=8.322; c=11.833$	819.4
	γ	149.5–155.1	Tetragonal	P4/mmm	$a=5.8942; c=5.8612$	
	δ	< 144.5	Orthorhombic	Pna2 ₁	$a=7.979; b=8.580; c=11.849$	811.1
MASnI ₃	α	293	Tetragonal	P4mm	$a=6.2302; b=6.2302; c=6.2316$	241.88
	β	200	Tetragonal	I4cm	$a=8.7577; b=8.7577; c=12.429$	953.2
FAPbI ₃	α	293	Trigonal	P3m1	$a=8.9817; b=8.9817; c=11.006$	768.9
	β	150	Trigonal	P3	$a=17.791; b=17.791; c=10.091$	2988.4
FASnI ₃	α	340	Orthorhombic	Amm2	$a=6.3286; b=8.9554; c=8.9463$	507.03
	β	180	Orthorhombic	Imm2	$a=12.512; b=12.512; c=12.509$	1959.2

gap from 1.55 to 1.45 eV (Eperon *et al.*, 2014). Following FA⁺, the ions such as Cs⁺, Rb⁺ with smaller ion radii have also been reported to be the applicable substitution for MA⁺ at A site, which results in the tailored bandgap and carrier behavior in the resultant materials (Choi *et al.*, 2014; Saliba *et al.*, 2016).

The most widely reported B cations ever reported in hybrid perovskite are Pb²⁺ and Sn²⁺, and the most efficient devices have been obtained always using Pb²⁺ based perovskite absorber. However, the toxicity of Pb²⁺ is seriously concerned which hinders the development of large-scale industrialization. To address this issue, the replacement of this heavy metal cation by an environmental friendly metal is under intense research, whereas Sn²⁺ or Ge²⁺ are investigated to be used as absorber without significant detriment to the device performance. Although CH₃NH₃SnX₃ possesses a more optimal band gap (1.2–1.4 eV) than CH₃NH₃PbX₃ (1.6–1.8 eV) does, the devices based on Sn perovskite do not appear comparable power conversion efficiency (PCE). It is mainly because the Sn based perovskite has shown a poor stability in air compared with the Pb based perovskite due to the oxidation of Sn²⁺ to Sn⁴⁺, which destroys the crystal structure and thus deteriorates the device performance (Kumar *et al.*, 2014).

The halide anion exhibit the substantial influence to the optoelectronic properties in the resultant hybrid perovskites. It has been widely reported the halogen elements I, Cl and Br, and their combinations affects the performance of the perovskite solar cells significantly. The Br[−] introduction into the perovskite lattice to replace I[−] will produce a higher absorption coefficient in the visible range of the spectrum, and upshift the conduction band to result in a wider bandgap perovskite suitable for tandem perovskite solar cells (Noh *et al.*, 2013). However, the Br[−] incorporation in MAPbI₃ often causes significant phase segregation upon illumination, which causes the instability issue. Introduction of Cl[−] into the perovskite can extend the charge diffusion length without changing the spectrum response, but only a small percentage of Cl atoms can be tolerated in the perovskite lattice (Chen *et al.*, 2015).

Optoelectronic Properties

Hybrid perovskites demonstrate strong optical absorption, adjustable band gap, long diffusion length, high carrier mobility, ambipolar charge transport and a high tolerance of defects. We thus provide a thorough discussion regarding the origin of the above excellent optoelectronic properties in the hybrid perovskites.

Hybrid perovskites exhibit large optical density with absorption coefficient of 10⁴–10⁵ cm^{−1}, which allows for a reduced film thickness as the absorber in solar cells to facilitate the light harvest efficiently. Devices based on less than 500 nm perovskite film can absorb sufficient sunlight and achieve high efficiency over 15%. In comparison, the commercially available crystalline Si or other inorganic thin film GaAs or CdTe photovoltaic technology are based on the absorber with thickness of few to hundreds of μm, much thicker than those of hybrid perovskites ones. The schematic illustration for optical transition over the bandedge that contributes to the optical absorption in Si, GaAs and perovskite materials are depicted in Fig. 2(a). The direct bandgaps of the halide perovskite (e.g., CH₃NH₃PbI₃) often exhibit optical absorption much stronger than that of silicon which possess an indirect bandgap. Compared with the GaAs semiconductor, the CB edge of the CH₃NH₃PbI₃ is mainly composed of degenerate Pb p orbitals. The atomic p orbital is less dispersive than s orbitals. Additionally, the edge transition in CH₃NH₃PbI₃ is mainly contributed from mixed-(Pb s, I p) to Pb p orbitals. It is generally believed the probability of intra-atomic Pb s to Pb p transition is high, which makes the VBM–CBM transition probability of CH₃NH₃PbI₃ comparable to that of GaAs. As is observed, halide perovskites show stronger optical absorption than GaAs in the visible range (Yin *et al.*, 2015). The calculated optical absorption coefficients of CH₃NH₃PbI₃ and GaAs are shown in Fig. 2(b) (Yin *et al.*, 2014).

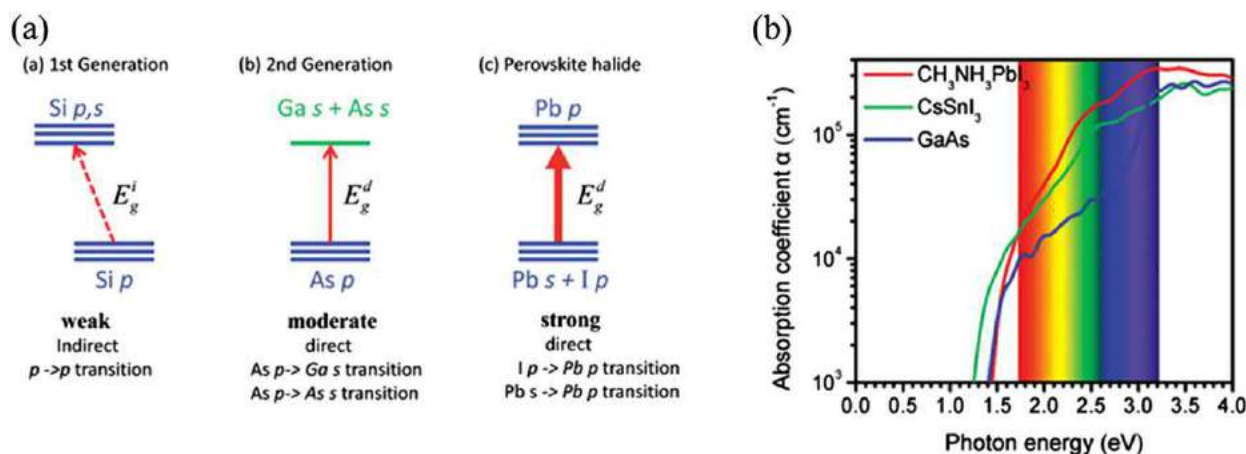


Fig. 2 (a) The schematic optical absorption of the three kinds of solar cell absorber. (b) The optical absorptions of $\text{CH}_3\text{NH}_3\text{PbI}_3$, CsSnI_3 and GaAs.

Table 2 DFT–PBE calculated effective masses for electrons and holes

	Electron effective mass	Hole effective mass
$\text{CH}_3\text{NH}_3\text{PbI}_3$	0.35, 0.32	0.31, 0.36
$\text{CH}_3\text{NH}_3\text{PbI}_3$ (SOC)	0.18, 0.23	0.21, 0.29
CsSnI_3	0.19	0.09
CsSnI_3 (SOC)	0.16	0.07
Silicon	0.26	0.29
GaAs	0.07	0.34
CuInSe_2	0.09	0.75
$\text{Cu}_2\text{ZnSnSe}_4$	0.1	0.26
CdTe	0.09	0.28

The electronic structure of $\text{CH}_3\text{NH}_3\text{PbI}_3$ is different from the conventional III–V (p–s) semiconductors. Its CBM is derived from Pb p orbitals, and the VBM is a combination of Pb s and I p orbitals. Even though the CBM comes from p orbitals, the p orbital of Pb^{2+} has a much higher energy level than anion p orbitals, as is usually observed in III–V (p–s) semiconductors. It leads to a more dispersive conduction band of $\text{CH}_3\text{NH}_3\text{PbI}_3$ which is lower than that of the p–s semiconductors with a higher valence band. Furthermore, the valence band of $\text{CH}_3\text{NH}_3\text{PbI}_3$ is dispersive due to strong s–p coupling over the VBM. The calculated electron and hole effective masses of $\text{CH}_3\text{NH}_3\text{PbI}_3$ via DFT–PBE, are compared with typical first and second-generation solar cell absorbers, which are listed in Table 2. The hole effective mass of $\text{CH}_3\text{NH}_3\text{PbI}_3$ is close to the electron effective mass, leading to ambipolar conductivity in perovskites, which facilitates to construct the p–i–n junction solar cells with versatile architectures (Yin *et al.*, 2015; Giorgi *et al.*, 2013). The carrier diffusion lengths of $\text{CH}_3\text{NH}_3\text{PbI}_3$ is up to 100 nm for both electrons and holes, and exceeding 1 μm in the mixed halide $\text{MAPbI}_{3-x}\text{Cl}_x$ have been reported via transient photoluminescence (TRPL) measurements (Xing *et al.*, 2013; Stranks *et al.*, 2013). Through the formation of hybrid perovskite single crystal, it demonstrated diffusion lengths over 175 μm in single crystals of $\text{CH}_3\text{NH}_3\text{PbI}_3$ under 1 sun illumination, owing to the reduced density of defects in the single crystal (Dong *et al.*, 2015). The defects with the relatively low formation energies are Pb^{2+} vacancies and interstitial MA^+ . Consequently, both of them are able to create shallow charge-carrier trap states near the perovskite band edges as shown in Fig. 3 (Yin *et al.*, 2014). Fortunately, the shallow defects are not responsible to nonradiative recombination centers that hampers the carrier diffusion length, which has been experimentally observed (Duan *et al.*, 2015).

Device Structures of Perovskite Solar Cell

The unprecedented rise in power conversion efficiency of perovskite solar cell, combined with the potential of cost-effective fabrication technique have stimulated the further understanding of perovskite devices (Green *et al.*, 2014; Stranks and Snaith, 2015; Park *et al.*, 2016). Two commonly employed structures for perovskite photovoltaic devices are shown in Figs. 4(a) and (b). Both device architectures consist of a perovskite layer as the light absorber layer which is sandwiched between electron transport layer (ETL) and hole transport layer (HTL). The main difference between the two configurations is whether there is mesoporous scaffold layer or not, and accordingly they were categorized as mesoporous PSC and planar heterojunction PSC. The first adoption

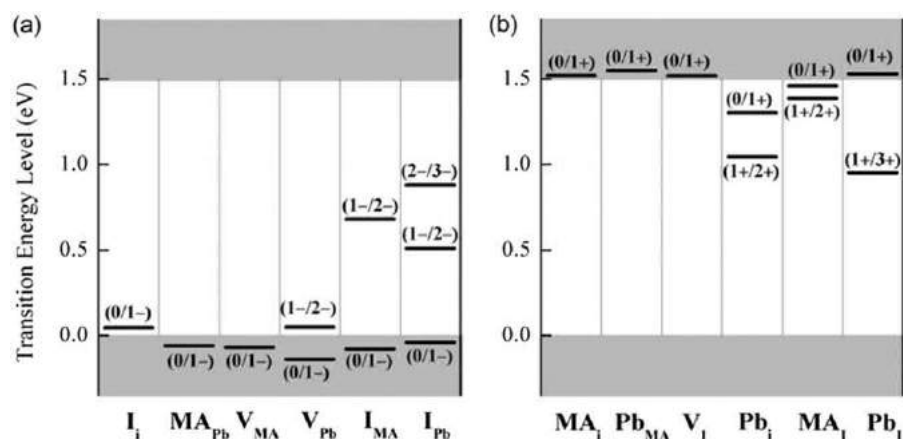


Fig. 3 The transition energy levels of (a) intrinsic acceptors and (b) intrinsic donors in $\text{CH}_3\text{NH}_3\text{PbI}_3$.

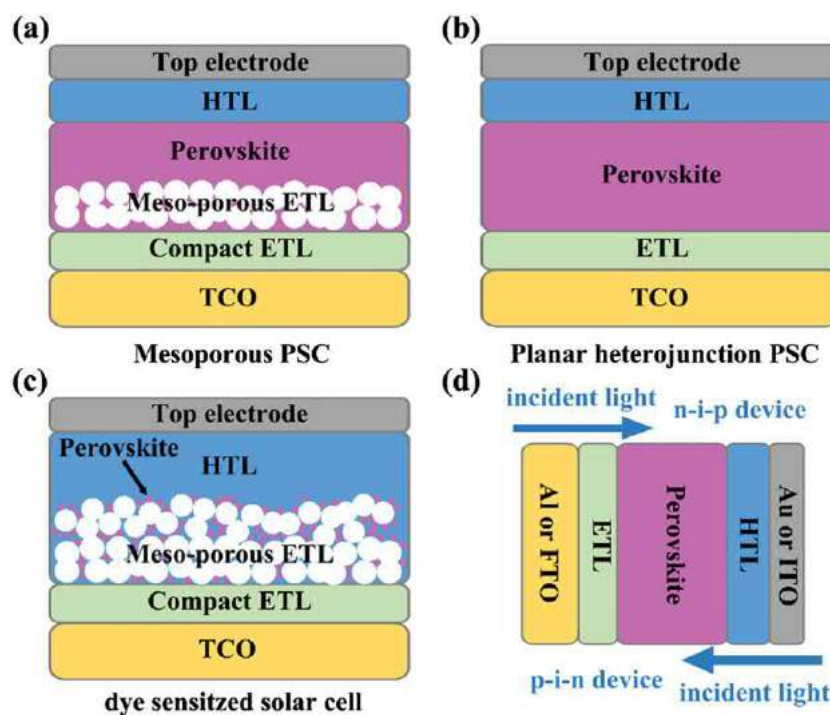


Fig. 4 The device structures of (a) mesoporous perovskite solar cell, (b) planar heterojunction perovskite solar cell, and (c) dye sensitized solar cell. (d) The n-i-p and p-i-n cell architectures.

of perovskite absorber $\text{CH}_3\text{NH}_3\text{PbI}_3$ (MAPbI_3) in photovoltaic device was to replace conventional dye in dye sensitized solar cell (DSSC) (Kojima *et al.*, 2009), whose structure was shown in Fig. 4(c). Generally, it is considered as the archetype of mesoporous PSC. The presence of TiO_2 mesoporous scaffold layer as ETL greatly increase the contact area between perovskite and TiO_2 material, giving rise to the faster and more efficient electron transport. Further investigation has demonstrated that perovskite absorbers exhibit superior ambipolar transport ability (both electron and hole, which means that mesoporous scaffold is not necessary for perovskite photovoltaic devices, leading to the emergence of planar heterojunction PSC (Leijtens *et al.*, 2014). Compared to the mesoporous PSC, the fabrication of planar structured PSC possessed the added advantage of more simple and low-temperature processing that are attractive for mass production and flexibility. Additionally, a great deal of excellent carrier conducting layers including inorganic materials, as well as organic materials have been successfully demonstrated in the efficient planar heterojunction PSC. Depending on the relative position of the electron/hole conducting layer and incident sunlight, the planar heterojunction PSC can be simply divided into n-i-p structure devices (regular configurations) and p-i-n structure devices (inverted configurations), as shown in Fig. 4(d).

To this day, the highest PCE of perovskite solar cell have reached a certified 22.1%, which is reported from National Renewable Energy Laboratory (NREL), but the detailed device structures and experiment process have not come forward yet. Moreover, Grätzel and his coworkers have realized the mesoporous PSCs with good reproducibility showing a certified PCE of 21.02% and a champion PCE of up to 21.6% under standard test condition (Bi *et al.*, 2016). The typical mesoporous PSCs device structure is FTO/compact TiO₂/mesoporous TiO₂/perovskite/spiro-OMeTAD/Au. And recently, Sargent and his coworkers reported top-performing planar heterojunction PSC using chlorine-capped TiO₂ colloidal nanocrystal film. The optimized planar photovoltaic devices have been demonstrated a certified PCE of 20.1%, which is the highest efficiency of planar heterojunction devices for now (Tan *et al.*, 2017). The reported device structure is ITO/TiO₂-Cl/Perovskite/Sprio-OMeTAD/Au.

Working Mechanism of Perovskite Solar Cell

When two types of semiconductors contact, the majority carriers in the material bulk will diffuse to each other. The movement of majority carriers will leave positive ion cores in the n-type semiconductor and negative ion cores in the p-type semiconductor, respectively, resulting in a charge depletion region with a built-in electric field, which is usually defined as a junction. When the sunlight incidence happens, the photogenerated carriers are created and the equilibrium is broken in the semiconductor absorber and in the contact region. During the subsequent diffusion process, part of the photogenerated carriers will recombine via direct recombination, Shockley-Read-Hall recombination or Auger recombination. The others will eventually diffuse to the junction and get separated assisted by the built-in electric field. The carriers will then be extracted selectively through the carrier-transport layers and reach the external circuit to generate photocurrent. When a photovoltaic devices is illuminated under sunlight, the key processes are shown in Fig. 5, including that (1) the incident photo with energy higher than the bandgap of light absorber layer can be absorbed in perovskite layer to generate holes and electrons, (2) the electrons transfer from perovskite to ETL, (3) the holes transfer from perovskite to HTL, which is equivalent to that of electron, (4) the electrons are recombined with photogenerated holes, (5) and (6) back charge carriers transport may occur at the interface of ETL and HTL with perovskite, (7) the electrons and holes recombine between ETL and HTL, which may occur if the quality of perovskite film is inferior so that many pinholes exist. In order to pursue high performance, process (4)–(7) must be restrained as strictly as possible, to ensure sufficient charge generation and extraction in operated perovskite photovoltaic devices.

Charge Carriers Transport Materials

Apart from the perovskite absorber, charge carriers transport layers, including ETL and HTL, also play a significant role in perovskite photovoltaic devices (Fan *et al.*, 2016; Hossain *et al.*, 2015; Zhang *et al.*, 2017; Ke *et al.*, 2015). The relative position of the valence band maximum (VBM) and conductive band minimum (CBM) of the charge carriers transport materials and perovskite provides the intrinsic motivation force for carrier separation and extraction at charge carriers transport layers/perovskite junction, which also determines the device's built-in potential of cell operation at equilibrium upon illumination. The design principles for choosing charge carriers transport materials must take two into account of two aspects. On one hand, the VBM and CBM of charge carriers transport materials are critical for appropriate energy level alignment between perovskite and charge carriers transport layers, which is essential for efficient charge carrier separation, extraction and transport to minimize open-circuit voltage loss (Cha *et al.*, 2016). On the other hand, the processibility of each functional layers should also be taken into consideration. For instance, the substrate wetting characteristics and chemical compatibility of carrier transport layers, particularly the deposition of subsequent layer upon perovskite. It has been widely reported that the processing approach can directly influence the film quality

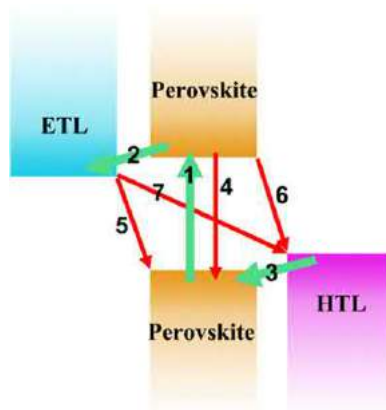


Fig. 5 Schematic of electron-transfer processes in an operated perovskite solar cell.

of perovskite solar cell, in terms of the grain size, morphology, crystallinity, surface coverage, *etc* (Song *et al.*, 2015). Therefore, rational design of carrier selective layers and the relevant processing methods determines the performance for photovoltaic devices, which has been carefully investigated in the research community.

Holes Transport Materials

Originated from the other third generation photovoltaic technologies, such as organic photovoltaics (OPV) and DSSC, 2,2',7,7'-Tetrakis[N,N-di(4-methoxyphenyl)amino]-9,9'-spirobifluorene (spiro-OMeTAD) and Poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) (PEDOT:PSS) have become two of the most popular hole transport materials (HTM) in perovskite photovoltaic devices, whose molecular structures are shown in Fig. 6(a). Owing to the excellent hole carriers mobility and appropriate VBM, CBM and Fermi energy level (Fig. 6(b)), both spiro-OMeTAD and PEDOT:PSS based PSC exhibit fascinating power conversion efficiency (Tan *et al.*, 2017; Meng *et al.*, 2016). However, neither of them is perfect, and they are facing a few critical challenges in way of commercialization of PSCs. Spiro-OMeTAD is too expensive to use on a large scale. In addition, it is reported pure spiro-OMeTAD exhibits weak electric conductivity, whereas chemical doping has been employed to greatly improve the carrier concentration and conductivity of the materials. However, the additive induced for chemical doping will damage the perovskite film in the solar cells. PEDOT:PSS shows acidic characteristics, which is not compatible to the perovskite film in devices and leads to long-term stability issues. Apart from that, both materials are organic species which are sensitive to moisture and oxygen. Therefore, lots of efforts are invested by research community aiming at feasible substitutions to address the current concern of inferior long-term stability in PSC devices (Bella *et al.*, 2016; Saliba *et al.*, 2016). Recently, lots of attention have been attracted to a variety of inorganic new candidates, such as NiOx, Cu₂O, CuSCN, CuI, and some cheaper or more stable organic HTM, whose energy level structures are shown in Fig. 6(b) (Hossain *et al.*, 2015). Taken as the example, NiOx is now under the spotlight as an inorganic p-type metal oxide hole transport material for PSCs. With the regard of valence band edge, it is demonstrated that NiOx can match with MAPbI₃ absorber (−5.4 eV) better than that of PEDOT:PSS (5.26 eV) or spiro-OMeTAD (5.22 eV), which leads to greater open-circuit voltage and more efficient hole carriers separation and extraction. Moreover, its intrinsic stability and low-cost processing has brought forth to apparent advantages of NiOx among other potential HTM candidate material (Xu *et al.*, 2015).

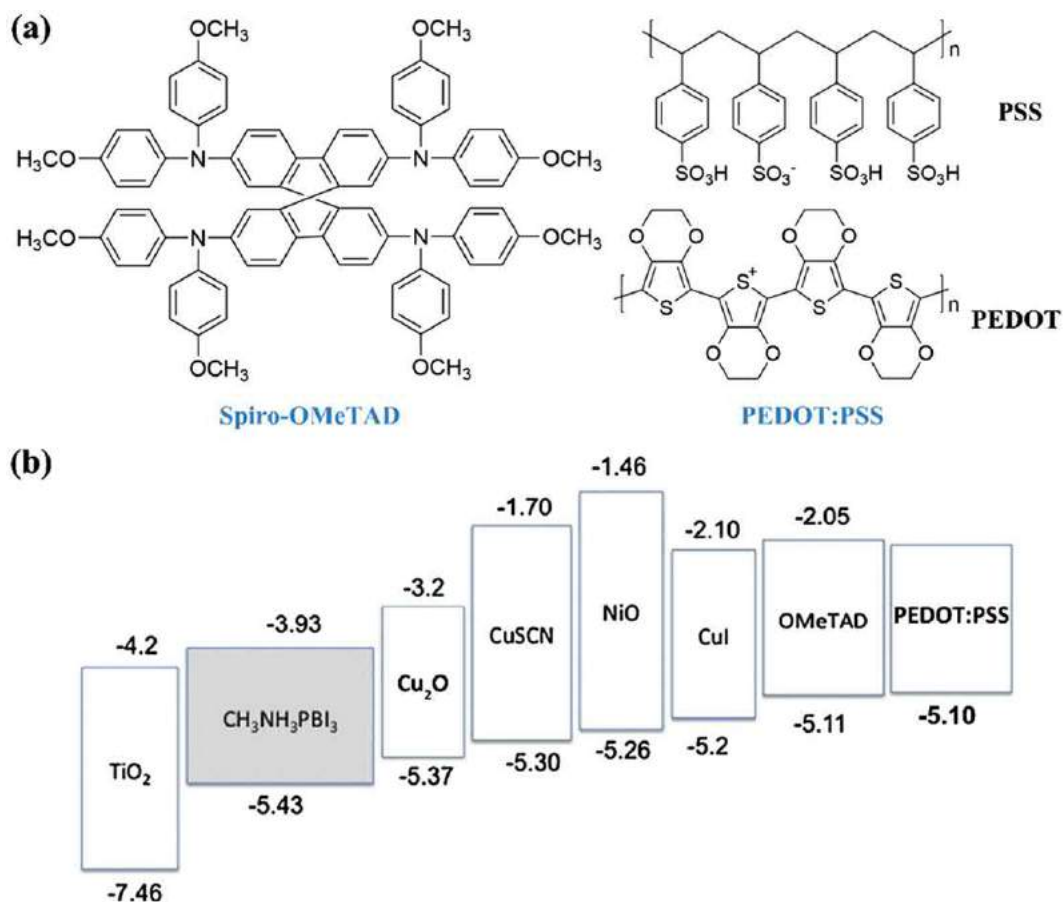


Fig. 6 (a) The molecular structures of spiro-OMeTAD and PEDOT:PSS. (b) Schematic energy level diagram of holes transport materials.

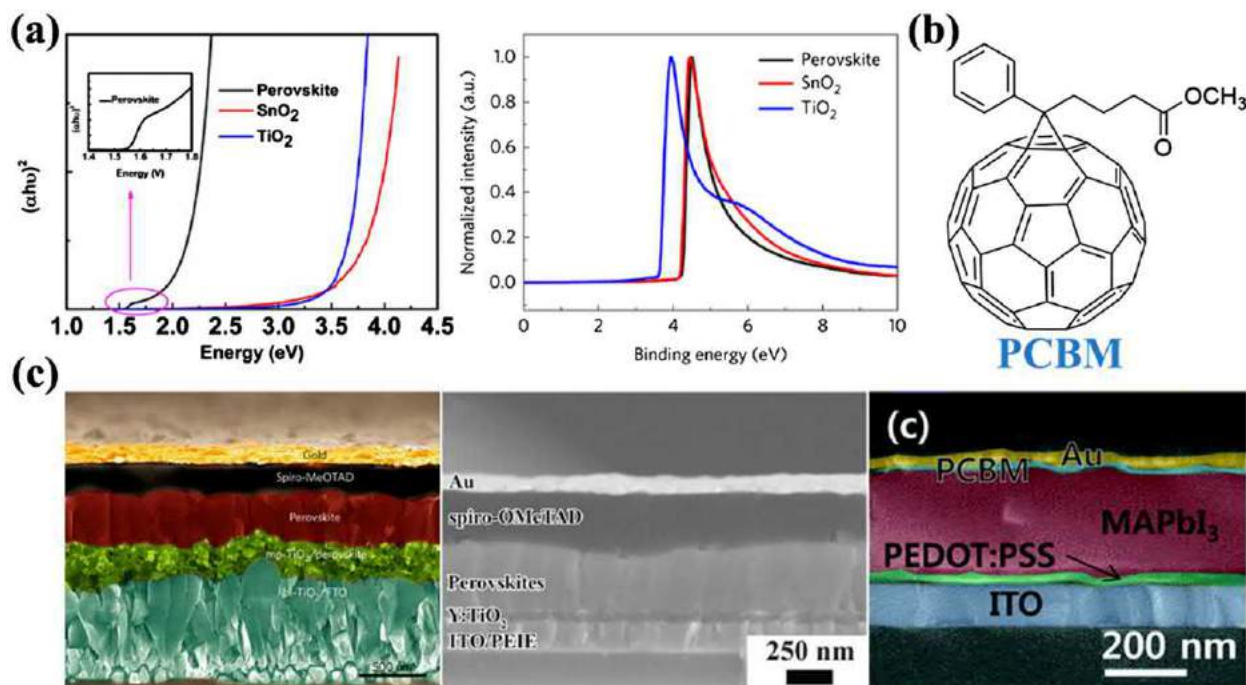


Fig. 7 (a) The left curves is the relationship of $(\alpha h\nu)^2$ vs. energy for perovskite, SnO_2 and TiO_2 , respectively. The bandgap of materials can be determined by the linear extrapolation of the leading edges of the $(\alpha h\nu)^2$ curve to the base lines. The right curves are ultraviolet photoelectron spectroscopy (UPS) cutoff edges of SnO_2 , TiO_2 and perovskite layer. (b) The molecular structures of PCBM. (c) The scan electron micrographs of three kinds of typical perovskite solar cells.

Electrons Transport Materials

Regarding electron transport materials (ETM), TiO_2 is the most popular choice as the electrode buffer to transfer electron from the absorber to electrode often employed in both mesoporous and planar heterojunction architectures (Tan *et al.*, 2017). This is mainly attributed to its well-aligned energy level to the perovskite absorber, as well as its well spread in DSSC/OPV. However, both compact and mesoporous TiO_2 suffer from low electron mobility and relatively high density of electronic trap states, giving rise to a high electron recombination rates. Moreover, it is reported that the TiO_2 layer is sensitive to ultraviolet exposure and oxygen, and its catalytic effects to decompose perovskite absorber is still under investigation, which may severely restrict the long-term application of perovskite photovoltaic devices. Right now, tin dioxide (SnO_2) is the most promising candidate to replace TiO_2 as electron transport material in planar heterojunction PSC with superiority of the much higher electron mobility and wider band gap than TiO_2 (Fig. 7(a)) (Jiang *et al.*, 2016). So with a wider bandgap of the SnO_2 as ETL, it can greatly minimize the photocurrent loss. SnO_2 also exhibits considerable stability upon ultraviolet illumination, which satisfies the requirements for commercialization. Besides these inorganic ETM, [6,6]-phenyl-C61-butyric acid methyl ester (PCBM, the molecular structure is shown in Fig. 7(b)) is one of the most popular ETM usually used in inverted planar heterojunction PSC to deliver outstanding performance due to its good mobility, Fermi energy level, and particularly, the electron extraction capability. What attracts researchers most for PCBM-based perovskite devices is its negligible J-V hysteresis behavior during cell efficiency measurement.

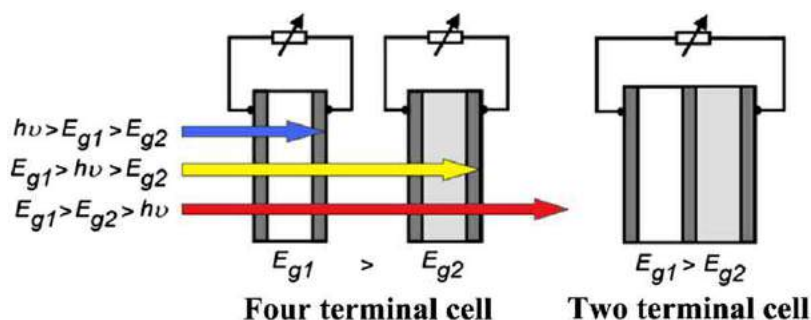
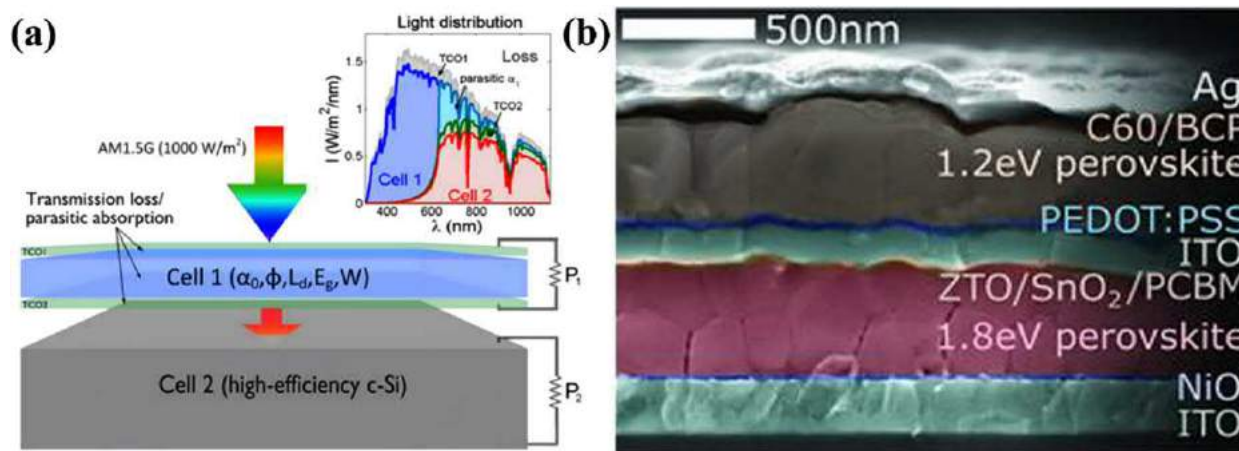
Based on various charge carriers transport materials mentioned above, the typical devices structure of mesoporous PSC is FTO/compact TiO_2 /mesoporous TiO_2 /perovskite/spiro-OMeTAD/Au. As shown in Fig. 7(c) (Bi *et al.*, 2016), the device contains a FTO glass substrate covered by a compact TiO_2 ETL (~ 30 nm). A mesoporous TiO_2 scaffold (~ 150 nm) is deposited on the top of compact TiO_2 layer, followed with the perovskite absorber (~ 450 nm). Subsequently, a ~ 200 nm thick spiro-OMeTAD as HTL is deposited by spin-coating. Finally, a thin layer of gold (~ 100 nm) is generally evaporated as the top electrode. As the counterpart, the typical devices structures of regular planar PSC indium tin oxide (ITO)/compact TiO_2 /MAPbI₃/spiro-OMeTAD/Au (Zhou *et al.*, 2014), And inverted planar PSC often adopt the configuration of ITO/PEDOT:PSS/ MAPbI₃/PCBM/Ag (Heo *et al.*, 2015), which are shown in Fig. 7(c). Moreover, the devices performance of PSC using various HTLs and ETLs are summarized in Table 3.

Tandem Solar Cell

For a single-junction solar cell, it achieves the limited power conversion efficiency due to the thermalization and transmission losses according to the Shockley–Queisser theory, leaving the maximum achievable efficiency close to 31% (Meillaud *et al.*, 2006).

Table 3 The device performance of perovskite solar cell with different charge carriers transport materials

Device structure	V_{oc} (V)	J_{sc} (mA/cm ²)	FF (%)	PCE (%)	References
ITO/CuSCN/perovskite/C60/BCP/Ag	1.00	21.90	75.8	16.6	Ye <i>et al.</i> (2015)
Mesoscopic NiO/perovskite/PCBM	1.04	13.24	69.0	9.51	Wang <i>et al.</i> (2014)
FTO/NiO/CH ₃ NH ₃ PbI ₃ /PCBM/Ag	0.88	16.27	63.5	9.11	Zhu <i>et al.</i> (2014)
ITO/Cu: NiO _x /perovskite/C60	1.05	22.32	76.0	17.74	Kim <i>et al.</i> (2015)
FTO/TiO ₂ /perovskite/X60	1.14	24.20	71.0	19.84	Xu <i>et al.</i> (2016)
ITO/PEIE/Y:TiO ₂ /perovskite/spiro-OMeTAD/Au	1.13	22.75	75.1	19.30	Zhou <i>et al.</i> (2014)
FTO/NiMgLiO/perovskite/PCBM/TiNbO _x /Ag	1.07	20.62	74.8	16.20	Chen <i>et al.</i> (2015)
FTO/SnO ₂ /perovskite/Spiro-OMeTAD/Au	1.11	23.27	67.0	17.21	Ke <i>et al.</i> (2015)
ITO/PEDOT:PSS/CH ₃ NH ₃ PbI ₃ (MAPbI ₃)/PCBM/Au	1.10	20.90	79.0	18.10	Heo <i>et al.</i> (2015)
FTO/WO _x /perovskite/spiro-OMeTAD/Ag	0.71	21.77	58.0	8.99	Wang <i>et al.</i> (2015)
ITO/ZnSnO ₄ /perovskite/PTAA/Au	1.05	21.6	67.0	15.3	Shin <i>et al.</i> (2015)

**Fig. 8** Schematics of four terminal tandem solar cell and two terminal tandem solar cell.**Fig. 9** (a) Schematic of a four-terminal tandem cell on silicon. (Inset) Optical distribution of absorbed sunlight (AM1.5G) through the complete tandem cell. (b) Scanning electron micrograph of the two terminal perovskite-perovskite tandem solar cell.

Different absorbers with different bandgaps will meet different theoretical PCE limits. So far, the traditional silicon solar cells are approaching their theoretical efficiency limit of $\sim 25\%$. To further improve the PCE of solar cells, tandem solar cells that consists of multiple absorbers stacked in a horizontal fashion have attracted many researchers' interests. The simplest tandem solar cell, that two sub-cells are working together, can be divided into "Four terminal device" and "Two terminal device" depending on how the power output is extracted. As shown in Fig. 8, in a four terminal device, the photocurrents extracted from the two sub-cells are collected independently. In a two terminal device however, the two sub-cells are connected in series for power output (Chen *et al.*, 2015). When the incident sunlight photons reach the tandem devices, the top cell will absorb short-wavelength photons with energy higher than its bandgap E_{g1} , and the bottom cell will utilize the remaining long-wavelength photos with energy higher than its bandgap E_{g2} . In this scenario, a smart tandem design is expected to include the top photo-active layer and junction layer with

negligible absorption for long-wavelength photons to reduce the parasitic absorption as much as possible. This is essential for tandem solar cell to achieve higher efficiency. The perovskite materials are reported to exhibit negligible mid-band-gap states, which is evidenced by its low Urbach energy. Therefore, the hybrid perovskites not only exhibit very high absorption in short-wavelength spectrum range, but also display excellent transmittance over the long-wavelength spectrum range, indicating the promising potential of perovskite materials in tandem cells.

Another challenge for light activated materials in tandem solar cells application is their bandgaps. According to the theoretical calculation, the silicon-perovskite tandem solar cell can achieve an efficiency over 35%, only when the absorber with an optimized bandgap of 1.7 eV is employed in the front cell, as shown in Fig. 9(a) (Lal *et al.*, 2014). For this reason, the perovskite materials with tunable bandgap from 1.48 to 2.3 eV are carefully investigated, regarded as the excellent candidate as the top absorber materials. Moreover, perovskite-perovskite tandem solar cells are born sequentially and reported 20.3% efficiency for four terminal tandem cells, 17.0% efficiency for two terminal tandem cells, whose device structure is shown in Fig. 9(b) (Eperon *et al.*, 2016). It is believed the perovskite/c-silicon solar cell is one of the most potential technical approach toward commercialization by capitalization the well-developed crystalline silicon industry.

Summary

In summary, organic–inorganic hybrid perovskite materials as the star candidate employed in photovoltaics have made tremendous progress in decade years thanks to the intrinsic characteristics, such as high absorption coefficients, ambipolar carrier transport, long carrier lifetimes and diffusion lengths and shallow defect levels. However, there is still a long way for practical use of perovskites in photovoltaic market. In this article, we briefly review optical and electrical property for perovskite, and present the fundamental device structure and working mechanism of perovskite solar cell, whereas the relevant charge carriers transport materials are also included. Moreover, the perovskite-based tandem solar cell as a promising developing direction have been concisely discussed. Substantial efforts and researches toward the further development of perovskites photovoltaics are in progress. We confidently believe that low-cost, high performance, long-term stable and environmental friendly perovskite photovoltaic technology will be achieved in the coming years.

References

- Ball, J.M., Lee, M.M., Hey, A., Snaith, H.J., 2013. Low-temperature processed meso-superstructured to thin-film perovskite solar cells. *Energy Environ Sci* 6, 1739.
- Bella, F., Griffini, G., Correa-Baena, J.-P., *et al.*, 2016. Improving efficiency and stability of perovskite solar cells with photocurable fluoropolymers. *Science* aah4046.
- Bi, D., Yi, C., Luo, J., *et al.*, 2016. Polymer-templated nucleation and crystal growth of perovskite films for solar cells with efficiency greater than 21%. *Nat. Energy* 1, 16142.
- Borriello, I., Cantele, G., Ninno, D., 2008. Ab initio investigation of hybrid organic–inorganic perovskites based on tin halides. *Phys. Rev. B* 77, 235214.
- Cha, M., Da, P., Wang, J., *et al.*, 2016. Enhancing perovskite solar cell performance by interface engineering using $\text{CH}_3\text{NH}_3\text{PbBr}_2$. 912. 1: Quantum dots. *J. Am. Chem. Soc.* 138, 8581–8587.
- Chen, Q., De Marco, N., Yang, Y., *et al.*, 2015. Under the spotlight: The organic–inorganic hybrid halide perovskite for optoelectronic applications. *Nano Today* 10, 355–396.
- Chen, Q., Zhou, H., Fang, Y., *et al.*, 2015. The optoelectronic role of chlorine in $\text{CH}_3\text{NH}_3\text{PbI}_3(\text{Cl})$ -based perovskite solar cells. *Nat. Commun.* 6, 7269.
- Chen, W., Wu, Y., Yue, Y., *et al.*, 2015. Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers. *Science* 350, 944–948.
- Choi, H., Jeong, J., Kim, H.-B., *et al.*, 2014. Cesium-doped methylammonium lead iodide perovskite light absorber for hybrid solar cells. *Nano Energy* 7, 80–85.
- Dong, Q., Fang, Y., Shao, Y., *et al.*, 2015. Solar cells. Electron-hole diffusion lengths $> 175 \mu\text{m}$ in solution-grown $\text{CH}_3\text{NH}_3\text{PbI}_3$ single crystals. *Science* 347, 967–970.
- Duan, H.S., Zhou, H., Chen, Q., *et al.*, 2015. The identification and characterization of defect states in hybrid organic–inorganic perovskite photovoltaics. *Phys. Chem. Chem. Phys.* 17, 112–116.
- Eperon, G.E., Leijtens, T., Bush, K.A., *et al.*, 2016. Perovskite–perovskite tandem photovoltaics with optimized band gaps. *Science* 354, 861–865.
- Eperon, G.E., Stranks, S.D., Menelaou, C., *et al.*, 2014. Formamidinium lead trihalide: A broadly tunable perovskite for efficient planar heterojunction solar cells. *Energy Environ. Sci.* 7, 982–988.
- Fan, R., Huang, Y., Wang, L., *et al.*, 2016. The progress of interface design in perovskite-based solar cells. *Adv. Energy Mater.* 6, 1600460.
- Giorgi, G., Fujisawa, J., Segawa, H., Yamashita, K., 2013. Small photocarrier effective masses featuring ambipolar transport in methylammonium lead iodide perovskite: A density functional analysis. *J. Phys. Chem. Lett.* 4, 4213–4216.
- Green, M.A., Ho-Baillie, A., Snaith, H.J., 2014. The emergence of perovskite solar cells. *Nat. Photonics* 8, 506–514.
- Heo, J.H., Han, H.J., Kim, D., Ahn, T.K., Im, S.H., 2015. Hysteresis-less inverted $\text{CH}_3\text{NH}_3\text{PbI}_3$ planar perovskite hybrid solar cells with 18.1% power conversion efficiency. *Energy Environ. Sci.* 8, 1602–1608.
- Hossain, M.I., Alharbi, F.H., Tabet, N., 2015. Copper oxide as inorganic hole transport material for lead halide perovskite based solar cells. *Sol. Energy* 120, 370–380.
- Jiang, Q., Zhang, L., Wang, H., *et al.*, 2016. Enhanced electron extraction using SnO_2 for high-efficiency planar-structure $\text{HC}(\text{NH}_2)_2\text{PbI}_3$ -based perovskite solar cells. *Nat. Energy* 1, 16177.
- Ke, W., Fang, G., Liu, Q., *et al.*, 2015. Low-temperature solution-processed tin oxide as an alternative electron transporting layer for efficient perovskite solar cells. *J. Am. Chem. Soc.* 137, 6730–6733.
- Kim, J.H., Liang, P.W., Williams, S.T., *et al.*, 2015. High-performance and environmentally stable planar heterojunction perovskite solar cells based on a solution-processed copper-doped nickel oxide hole-transporting layer. *Adv. Mater.* 27, 695–701.
- Kojima, A., Teshima, K., Shirai, Y., Miyasaka, T., 2009. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* 131, 6050–6051.
- Kulkarni, S.A., Baikie, T., Boix, P.P., *et al.*, 2014. Band-gap tuning of lead halide perovskites using a sequential deposition process. *J. Mater. Chem. A* 2, 9221–9225.
- Kumar, M.H., Dharani, S., Leong, W.L., *et al.*, 2014. Lead-free halide perovskite solar cells with high photocurrents realized through vacancy modulation. *Adv. Mater.* 26, 7122–7127.
- Lal, N.N., White, T.P., Catchpole, K.R., 2014. Optics and light trapping for tandem solar cells on silicon. *IEEE J. Photovolt.* 4, 1380–1386.
- Leijtens, T., Lauber, B., Eperon, G.E., Stranks, S.D., Snaith, H.J., 2014. The importance of perovskite pore filling in organometal mixed halide sensitized TiO_2 -based solar cells. *J. Phys. Chem. Lett.* 5, 1096–1102.

- Li, C., Lu, X., Ding, W., *et al.*, 2008. Formability of ABX_3 ($X=F, Cl, Br, I$) halide perovskites. *Acta Crystallogr B: Struct Sci* 64, 702–707.
- Lu, X., Wang, Y., Stoumpos, C.C., *et al.*, 2016. Enhanced structural stability and photo responsiveness of $CH_3NH_3SnI_3$ perovskite via pressure-induced amorphization and recrystallization. *Adv. Mater.* 28, 8663–8668.
- Meillaud, F., Shah, A., Droz, C., Vallat-Sauvain, E., Miazza, C., 2006. Efficiency limits for single-junction and tandem solar cells. *Sol. Energ. Mat. Sol. C.* 90, 2952–2959.
- Meng, L., You, J., Guo, T.F., Yang, Y., 2016. Recent advances in the inverted planar structure of perovskite solar cells. *Acc. Chem. Res.* 49, 155–165.
- Noh, J.H., Im, S.H., Heo, J.H., Mandal, T.N., Seok, S.I., 2013. Chemical management for colorful, efficient, and stable inorganic–organic hybrid nanostructured solar cells. *Nano Lett.* 13, 1764–1769.
- Park, N.-G., Grätzel, M., Miyasaka, T., Zhu, K., Emery, K., 2016. Towards stable and commercially available perovskite solar cells. *Nat. Energy* 1, 16152.
- Saliba, M., Matsui, T., Domanski, K., *et al.*, 2016. Incorporation of rubidium cations into perovskite solar cells improves photovoltaic performance. *Science* 354, 206–209.
- Shin, S.S., Yang, W.S., Noh, J.H., *et al.*, 2015. High-performance flexible perovskite solar cells exploiting Zn_2SnO_4 prepared in solution below 100 degrees C. *Nat. Commun.* 6, 7410.
- Song, T.-B., Chen, Q., Zhou, H., *et al.*, 2015. Perovskite solar cells: Film formation and properties. *J. Mater. Chem. A* 3, 9032–9050.
- Stranks, S.D., Eperon, G.E., Grancini, G., *et al.*, 2013. Electron-hole diffusion lengths exceeding 1 Micrometer in an organometal trihalide perovskite absorber. *Science* 342, 341–344.
- Stranks, S.D., Snaith, H.J., 2015. Metal-halide perovskites for photovoltaic and light-emitting devices. *Nat. Nanotechnol.* 10, 391–402.
- Tan, H., Jain, A., Voznyy, O., *et al.*, 2017. Efficient and stable solution-processed planar perovskite solar cells via contact passivation. *Science* 355, 722–726.
- Wang, K.C., Jeng, J.Y., Shen, P.S., *et al.*, 2014. p-Type mesoscopic nickel oxide/organometallic perovskite heterojunction solar cells. *Sci. Rep.* 4, 4756.
- Wang, K., Shi, Y., Dong, Q., *et al.*, 2015. Low-temperature and solution-processed amorphous $WO(x)$ as electron-selective layer for perovskite solar cells. *J. Phys. Chem. Lett.* 6, 755–759.
- Xing, G., Mathews, N., Sun, S., *et al.*, 2013. Long-range balanced electron- and hole-transport lengths in organic–inorganic $CH_3NH_3PbI_3$. *Science* 342, 344–347.
- Xu, B., Bi, D., Hua, Y., *et al.*, 2016. A low-cost spiro[fluorene-9,9'-xanthene]-based hole transport material for highly efficient solid-state dye-sensitized solar cells and perovskite solar cells. *Energy Environ. Sci.* 9, 873–877.
- Xu, X.B., Liu, Z.H., Zuo, Z.X., *et al.*, 2015. Hole selective NiO contact for efficient perovskite solar cells with carbon electrode. *Nano Lett.* 15, 2402–2408.
- Ye, S., Sun, W., Li, Y., *et al.*, 2015. CuSCN-based inverted planar perovskite solar cell with an average PCE of 15.6%. *Nano Lett.* 15, 3723–3728.
- Yin, W.-J., Shi, T., Yan, Y., 2014. Unique properties of halide perovskites as possible origins of the superior solar cell performance. *Adv. Mater.* 26, 4653–4658.
- Yin, W.-J., Shi, T., Yan, Y., 2014. Unusual defect physics in $CH_3NH_3PbI_3$ perovskite solar cell absorber. *Appl. Phys. Lett.* 104, 063903.
- Yin, W.-J., Yang, J.-H., Kang, J., Yan, Y., Wei, S.-H., 2015. Halide perovskite materials for solar cells: A theoretical review. *J. Mater. Chem. A* 3, 8926–8942.
- Zhang, H., Wang, H., Chen, W., Jen, A.K., 2017. $CuGaO_2$: A promising inorganic hole-transporting material for highly efficient and stable perovskite solar cells. *Adv. Mater.* 29, 1604984.
- Zhou, H.P., Chen, Q., Li, G., *et al.*, 2014. Interface engineering of highly efficient perovskite solar cells. *Science* 345, 542–546.
- Zhu, Z.L., Bai, Y., Zhang, T., *et al.*, 2014. High-performance hole-extraction layer of sol–gel-processed NiO nanocrystals for inverted planar perovskite solar cells. *Angew. Chem. Int. Edit.* 53, 12571–12575.

Light Harvesting in Organic Solar cells

Supriya Pillai, Matthew Wright, and Kah H Chan, University of New South Wales, Sydney, NSW, Australia

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Considerable interest has evolved in the development of technologies to address the complex fabrication process and high cost of solar cells. Organic solar cells (OSCs) belong to a class of photovoltaic technology that are fabricated with organic, carbon-based semiconductors as the light absorbing active layer. OSCs offer a competitive path to realize cost-effective as well as environmentally friendly solar cells for future photovoltaic applications. They have the potential to be ultra-low cost, ultrathin, light weight and flexible (Krebs, 2009). They also offer the promise of high performance with opportunities not only in photovoltaics, but also in electronics and lighting applications. The ultrathin nature of OSC makes it transparent and allows it to be integrated into window glass or used as design elements for lighting installations. The polymers used in OSC can be easily printed and can be made in different colors.

The working principle of an organic solar cell has been discussed in the previous articles. However, to bring the need for light harvesting to context, it is important to understand that the exciton diffusion length L_D (the distance charge carriers can travel before recombining) is very short in organic semiconductors. Furthermore, L_D is significantly less than the absorption length $1/\alpha$ (distance into a material where the light intensity has dropped to $1/e$ or 36% of its original incident flux) for an organic solar cell. This means there is an inherent trade-off between the absorption in the organic absorber layer and the exciton diffusion length, thus forming an exciton diffusion bottleneck whereby the excitons cannot reach the donor–acceptor interface prior to dissociation into free carriers. This limits the thickness of the absorber layer d used to ~ 100 nm so that more charge carriers can be successfully extracted before recombination occurs (Yu *et al.*, 1995; Ma *et al.*, 2005) and significantly affects the cell efficiency. The absorption efficiency within the absorber layer $\eta_A = 1 - e^{-\alpha d}$ and the exciton diffusion efficiency can be calculated as $\eta_{ED} = e_D^{-d/L_D}$. Fig. 1 shows the dependence of the exciton diffusion bottleneck and the ratio of the layer thickness d to the exciton diffusion length L_D (Forrest, 2005).

The low efficiencies of OPV have hindered the commercial success of this technology. While different strategies are in place to improve either the absorption length or the diffusion length by design variations, blending polymers or improved material properties (Vezie *et al.*, 2016), this article will briefly discuss light harvesting schemes that have demonstrated or could potentially be used for increased light absorption through optical means. Conventional light trapping in silicon (Si) solar cells make use of surface textures that are on the micron scale (wavelength of light) to increase the optical pathlength. However, this is not feasible on very thin absorber layers like OSC. Moreover, texturing can induce roughness which increases the density of surface and bulk defect states, with the potential of degrading the electrical performance of cells. Hence, alternate light harnessing schemes need to be employed. They include antireflection (AR) schemes that direct more light into the absorber layer or light trapping schemes that aid the light to bounce back and forth in the active layer, thus increasing the optical path length thereby providing multiple opportunities for the light to be absorbed in the process. Other techniques include resonance/light concentration/near-field effects

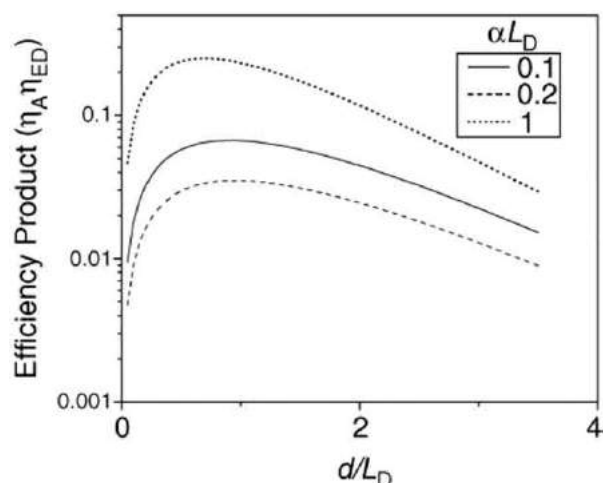


Fig. 1 Plot showing the dependence of the “exciton diffusion bottleneck” on the ratio of layer thickness d to exciton diffusion length L_D . Here α is the optical absorption coefficient within a particular layer. Taken from Forrest, S.R., 2005. The limits to organic photovoltaic cell efficiency. MRS Bulletin 30 (1), 28–32.

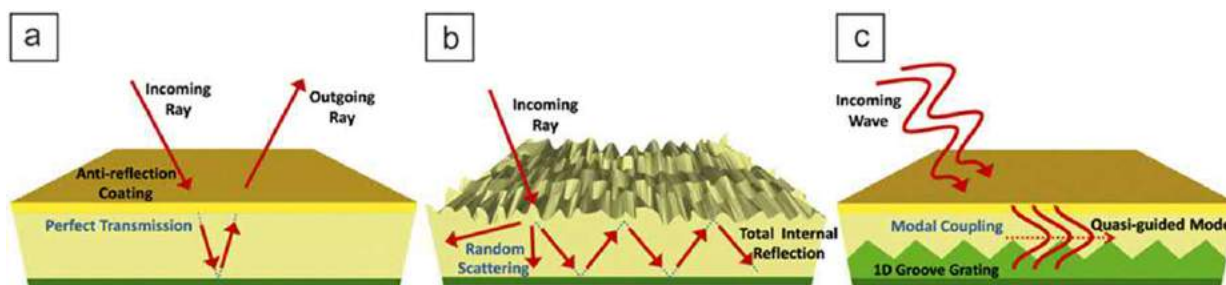


Fig. 2 Different ways of harnessing light (a) antireflection (AR) to maximize absorption, (b) light trapping where random scattering increase the path length, and (c) coherent light trapping where diffraction or electric field concentration is exploited. Taken from Mallick, S.B., Sergeant, N.P., Agrawal, M., Lee, J.-Y., Peumans, P., 2011. Coherent light trapping in thin-film photovoltaics. *MRS Bulletin* 36 (6), 453–460.

that can increase the optical density of states or confine the light in the absorber layer where photovoltaic conversion occurs. Approaches include graded index structures, substrate texturing, plasmonics, diffraction grating, photonic crystals, microcavities (Chen *et al.*, 2014a,b; Sergeant *et al.*, 2013) which makes use of one or more schemes. A schematic of the different light harnessing schemes is shown in Fig. 2. These light harvesting schemes are targeted for optical enhancement, however, some designs like plasmonics can benefit from both optical and electronic enhancement. Other approaches involve manipulating the light spectrum to be absorbed in different layers such as tandem configurations. In this work, we will discuss each of these approaches in more detail.

AR Effects

The realization of an efficient cell comes with minimization of optical losses. Since OSCs consist of stacks of different layers, the chances of reflective losses at each of the interfaces have to be eliminated to maximize in-coupling of incident light in the active layer. Hence the mitigation of optical losses has to start at the first point of incidence, which is at the front interface. This ensures that all the light that reaches the cell can be harvested in some way.

One way to achieve broadband AR properties is by using techniques seen in nature. For instance, the moth eye structures consist of nanostructured hexagonal cone shaped patterns that can substantially suppress the reflections of light at an interface of light over a broad spectral range (Bernhard, 1967) thereby increasing transmittance into the absorbing substrate over a wide angle of incidence (Huang *et al.*, 2007). Photonic structures are fabricated by mimicking moth eyes where the sub-wavelength features in the moth eye provide an effective refractive index that gradually varies and can suppress the reflection of light. The effective refractive index depends on the refractive indices of the nanostructured material and incident medium, and the angle of incidence. By tweaking these parameters the photonic structures can be optimized. Large scale fabrication of such structures is possible by interference lithography deeming it suitable for large scale production in solar cells.

Moth eye AR structures have been successfully incorporated to improve the performance of OSCs with an increase in the peak EQE by 3.5% along with making it angularly independent to oblique light (Forberich *et al.*, 2008). Another study achieved hydrophobic surfaces that enable self-cleaning properties for solar cells using a polydimethylsiloxane (PDMS) film as the AR layer on glass substrate with an encapsulated poly[(4,8-bis-(2-ethylhexyloxy)-benzo[1,2-b:4,5-b']dithiophene)-2,6-diyl-alt-(4-(2-ethylhexanoyl)-thieno[3,4-b]thiophene)-2,6-diyl]:[6,6]-phenyl-C61 butyric acid methyl ester (PBDTTT-C:PCBM) cell increasing photocurrent efficiency. An improvement of 6.19% was reported where the nanostructures were fabricated by soft imprint lithography and the patterns formed by laser interference lithography and dry etching (Leem *et al.*, 2014). Another study using such dual-side bio-inspired nanostructures fabricated using nanoimprint lithography improved the light in-coupling efficiency of OSCs by 20%, yielding an enhanced power conversion efficiency of 9.33% (Zhou *et al.*, 2014).

Another light harnessing inspired by nature is called “flower power” that makes use of the epidermal structures of plants. Plants living in low light intensity regions make use of nanostructured surfaces need to harness light for metabolism or flowers for color saturation. These structures can not only reduce reflections like the moth eye structures, but also, in addition, can redirect incoming photons to induce light focussing effects thereby facilitating efficient light collection as well as light trapping (Hünig *et al.*, 2016). The densely packed combination of microstructures increase the chance of reflected light to bounce onto the neighboring feature especially at large angles of incidences and show only minor dependence to wavelengths. These structures also act as an imperfect lens diverging the incoupled light thereby increasing the optical path length resulting in light trapping. This double contribution has increased the power conversion efficiency of OSCs by 13% at normal incidence and 44% at an angle of incidence of 80 degree (Hünig *et al.*, 2016). The structures were fabricated by transferring the features directly from the rose petals to a PDMS stamp and then imprinting it to the cells as shown in the inset of Fig. 3(a) and (b) shows the smoothening of the Fabry–Perot resonances and reduction in the reflection on employing the rose structure on a ZnO:Al layer.

Nanowire (NW) arrays are another type of structure which has demonstrated AR and light trapping effects (Muskens *et al.*, 2008). Like in moth eye and flower structures they also exhibit broadband and omnidirectional AR properties due to the gradual

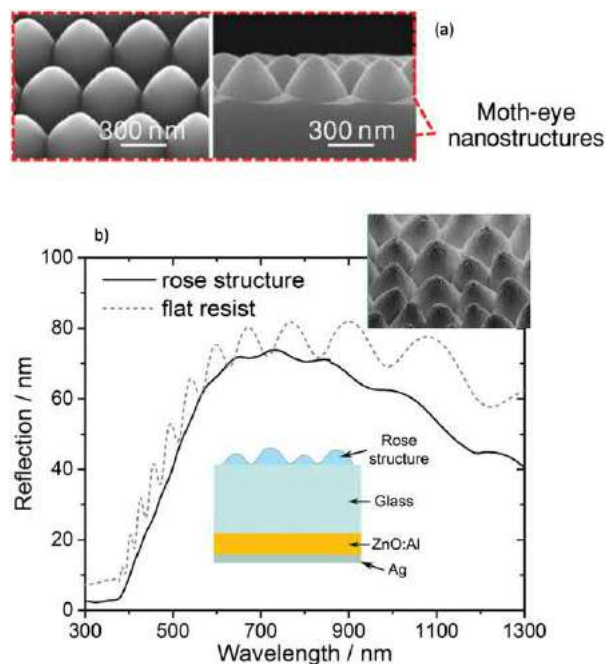


Fig. 3 (a) Scanning electron microscope (SEM) images of moth eye structures on a Si master that can be transferred to the substrate by soft imprint lithography to form inverted moth-eye structures (Reproduced from Leem, J.W., Kim, S., Lee, S.H., *et al.*, 2014. Efficiency enhancement of organic solar cells using hydrophobic antireflective inverted moth-eye nanopatterned PDMS films. *Advanced Energy Materials* 4 (8), 1301315) and (b) reduction in reflection and smoothening interference upon employing the “flower structure.” Inset shows a 55 degree tilted SEM picture of one such structure. Taken from Hünig, R., Mertens, A., Stephan, M., *et al.*, 2016. Flower power: Exploiting plants’ epidermal structures for enhanced light harvesting in thin-film solar cells. *Advanced Optical Materials* 4 (10), 1487–1493.

index profile. The tapered shape of the nanorod arrays leads to impedance matching between absorber and air through a gradual reduction of the effective refractive index away from the surface, resulting in low reflection particularly at longer wavelengths and eliminating interference fringes through roughening of the top interface (Lee *et al.*, 2008). Following its success on Si solar cells, ZnO NWs grown using metalorganic chemical vapor deposition have been applied on OSCs following its success on Si solar cells resulting in a 36% boost in optical absorption compared to planar devices (Hiralal *et al.*, 2014). These AR structures are particularly interesting due to the fact that the AR effect becomes more pronounced at large angles of incidence and they can exhibit self-cleaning properties due to the hydrophobic surfaces created through relevant surface treatment.

Light Trapping

While optical losses in cells can be reduced by proper choice of materials and good AR layers, inefficient light absorption in thin OSCs is a critical issue that limits cell efficiency. Traditional light trapping in solar cells consist of surface textures such as pyramids which allows randomized light paths in the cell, thus enhancing the optical path length and providing more opportunities for the light to be absorbed. However, the fragile nature and small physical dimensions of OSCs preclude such trapping schemes used in inorganic solar cells like Si which makes low cost fabrication challenging. Hence light trapping schemes have to be modified in OSCs using coherent light trapping approaches that make use of wave nature of light to make significant contributions. A good discussion on coherent light trapping can be found in Mallick *et al.* (2011).

The fundamental light trapping limit of light is $4n^2$ where n is the refractive index of the thick absorber layer. This was proposed by Yablonovitch and is known as the ray-optics limit (Yablonovitch, 1982) where the light scattered is considered to be ergodic, i.e., light will be coupled equally to all available modes. However, for thin layers where the thickness of the absorber layer is smaller or comparable to the wavelength of light, this limit does not apply as the number of modes available reduces to a set of discrete modes. Based on a rigorous electromagnetic wave approach, Yu *et al.* proposed that the standard limit can be surpassed to an absorption enhancement factor of $12 \times 4n^2$ using nanophotonic light trapping approaches (Yu *et al.*, 2011). Two such nanophotonic structures were proposed by Yu *et al.*, (2011) and (Green, 2011) where both proposed low index absorber layer sandwiched between high index layers that is ideal for OSCs. In both cases the index contrast between the cladding layer and the absorber layer provides strong nanoscale field confinement in the absorber layer with the potential to improve cell performance dramatically. Yu’s structure considered a 5-nm absorber layer sandwiched between a 60 nm thick high refractive index cladding layer and a mirror. An 80-nm scattering layer consisting of groove patterns was used to couple light into the guided modes from all

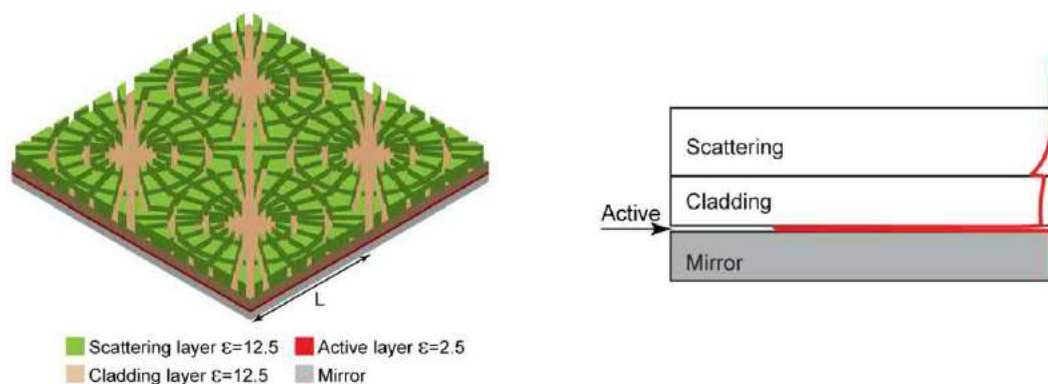


Fig. 4 (left) Schematic of a nanophotonic light trapping structure with an 80 nm thick scattering layer comprising square lattice of air grooves of periodicity 1200 nm, 60 nm cladding layer and 5 nm active layer on a mirror; (right) electric field intensity profile of the fundamental waveguide mode. Taken from Yu, Z., Raman, A., Fan, S., 2011. Nanophotonic light-trapping theory for solar cells. *Applied Physics A*. doi: 10.1007/s00339-011-6617-4.

angles. Green's work used a similar sandwiched structure but with the absorber layer between two high index cladding layers leading to an enhanced field in the absorber layer due to tunneling evanescent waves. While organic materials with low index are suitable and could highly benefit with active layers in these two configurations, it is crucial to identify high index cladding materials with low parasitic absorption. **Fig. 4** shows nanophotonic light trapping structure along with the electric-field intensity for the waveguide mode that is strongly confined to the active layer.

Light trapping by scattering and randomizing of light in such a way that light travels more in the plane of the absorber is difficult to realize in OSCs as the feature sizes should be larger or comparable to the thickness of the organic layer (which is the order of wavelength of light). Moreover, as mentioned earlier, nano-textures can degrade the performance of the device due to recombination losses. Hence different engineering solutions to trap light have been investigated without increasing the surface recombination. One such case is where the scattering layer is isolated from the absorber layer. The structure comprises of low absorption micro-lenses on top of reflecting mirrors with small transmitting openings for the focused light to be incident onto an OSC absorber as shown in **Fig. 5(a)**. The lens/mirror combination collimates the incident light in the forward direction and is highly reflective to both collimated and scattered light in the opposite direction. The light in this case bounces back and forth between the mirror and the back electrode increasing the probability of light absorption. More chances of photon recycling exist where absorption is weak with these light trapping elements enhancing photocurrent by as much as 25% (Tvingstedt *et al.*, 2008). One drawback of such structures is the need for accurate focusing as it is angle sensitive which may otherwise result in shading when the holes are not properly aligned.

Micro-lens array of closely packed hemispherical shape has been fabricated earlier using soft lithographic molding technique as shown in **Fig. 5(b)** (Myers *et al.*, 2012). Both AR and light trapping resulted in a 15–60% relative increase in efficiency. This structure did not have any mirrors, unlike the patterned mirror-and-lens light trap structure in **Fig. 5(a)**, hence a more economical structure considering that it is attached to the outer layer of glass without affecting the surface morphology of the active layer. Similar enhancements and 2.5 times longer pathlength using pyramidal rear reflectors fabricated on glass for semitransparent OSC have also been reported (Cao *et al.*, 2011).

Attempts have also been made to realize surface textures that are on a length scale much larger than the active layer thickness and that requires no modification of the device structure. The active layer and reflective metal electrode are deposited on a V or W-shaped transparent substrate coated with a transparent electrode such as indium tin oxide (ITO) (**Fig. 5(c)**) that increases photocurrent efficiency for all angles of incidence (Rim *et al.*, 2007). A 52% increase in power conversion efficiency was demonstrated using poly(3-hexylthiophene-2,5-diyl):[6,6]-phenyl-C61 butyric acid methyl ester (P3HT:PCBM) polymer blend cells that occupied the complete V shaped area. The viability of the V-shaped trap depends on the balance of the added cost of the structure and increase in active material with the reduction in installation cost of modules with high efficiencies. **Fig. 5(d)** shows a photo of the V and W-geometry cell taken from Zhou *et al.* (2008).

A photonic crystal structure is another method used for light management in OSCs. They comprise of dielectric or metallic structures that can manipulate the flow of photons. A variation to the micro-lens structure consist of polymer micro-lens on the exterior glass coupled with a photonic-plasmonic crystal at the metal cathode on the back of the cell but without mirrors (Peer and Biswas, 2014) (plasmonic structures are metal-based nanostructures with the potential to manipulate the flow of light in the adjoining absorber layer and will be discussed in more detail in a separate section). These micro-lens arrays focus and diffract light within the OSC which then is coupled to a periodically patterned organic layer facilitated by the periodically patterned metal cathode. Strong diffraction, waveguide modes and plasmonic light concentration effects contribute to the performance enhancement. This work demonstrated experimentally realizable solar cell structures with simulations predicting a photocurrent enhancement of ~58% relative to a flat cell. The absorption approached the Lambertian limit with enhanced absorption at longer wavelengths.

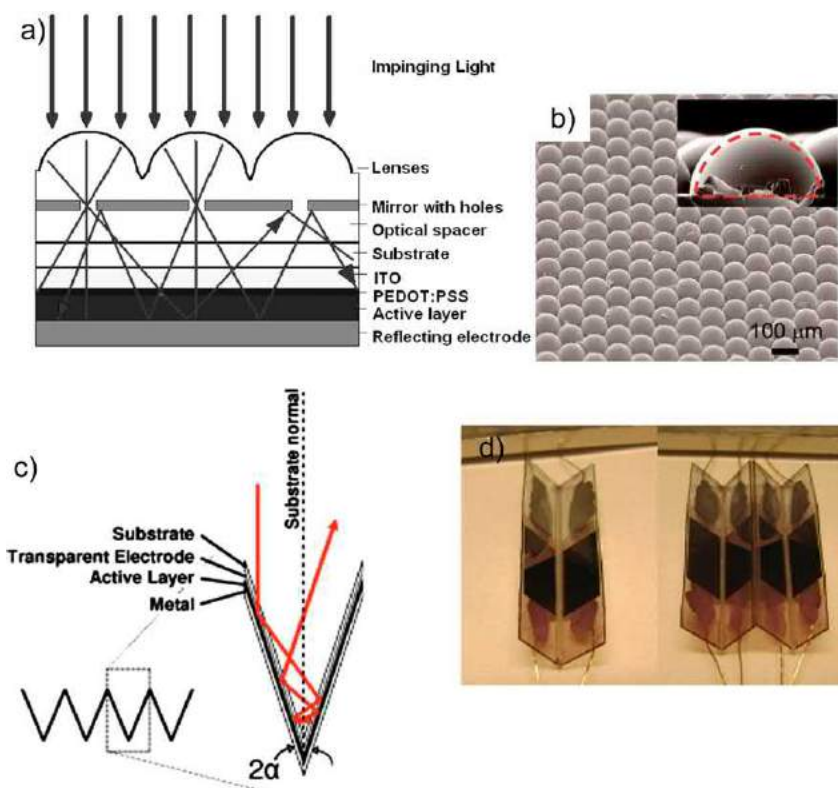


Fig. 5 (a) Schematic of an array of micro-lenses as a light trapping layer for an organic solar cell (Taken from Tvingstedt, K., Zilio, S.D., Inganäs, O., Tormen, M., 2008. Trapping light with micro lenses in thin film organic photovoltaic cells. *Optics Express* 16 (26), 21608–21615); (b) an scanning electron microscope (SEM) image of a microlens array (Reproduced from Myers, J.D., Cao, W., Cassidy, V., *et al.*, 2012. A universal optical approach to enhancing efficiency of organic-based photovoltaic devices. *Energy & Environmental Science* 5 (5), 6900–6904); (c) geometry of the V shaped light trapping structure with a very thin layer of absorber layer (Taken from Rim, S.-B., Zhao, S., Scully, S.R., McGehee, M.D., Peumans, P., 2007. An effective light trapping configuration for thin-film solar cells. *Applied Physics Letters* 91 (24), 243501). (d) Pictures of V and W-geometry cell (Taken from Zhou, Y., Zhang, F., Tvingstedt, K., Tian, W., Inganäs, O., 2008. Multifolded polymer solar cells on flexible substrates. *Applied Physics Letters* 93 (3), 033302. ITO, indium tin oxide; PEDOT:PSS, poly (3,4-ethylenedioxythiophene):polystyrene sulfonate).

A photonic substrate texturing technique, making use of folds and wrinkles, has shown to significantly enhance the incoupling of the light in OSCs (Gregg and van de Lagemaat, 2012) taking inspiration from the structure of leaves which has high light-harvesting efficiency. Fig. 6(a)–(e) shows the scanning electron microscope (SEM) image of the wrinkles developing into folds. Here an optically curable adhesive layer on the substrate is subject to compressive stress with a sequence of ultraviolet pulses and subsequent plasma treatment creating folds and wrinkles (Kim *et al.*, 2012). The entire folded layer was then coated with gold (Au) (anode) followed by the organic polymer P3HT:PCBM as shown in the inset of Fig. 6(e). These combination of wrinkles and deep folds act as photonic structures; they increase light coupling into and trapping light by waveguiding effects. In this work although there was a 79% increase in the external quantum efficiency (EQE) in the visible region compared to devices on flat surfaces, the EQE of the polymer material was interestingly improved by more than 600% in the near-infrared (where absorption is weak), where the useful range of solar energy conversion is extended by more than 200 nm (beyond the absorption edge of P3HT:PCBM > 650 nm). Fig. 6(e) and (f) show the IV and EQE of the cells with the wrinkles and wrinkle and fold combination compared to a reference cell.

Metal Nanostructures (Plasmonics)

Metal nanostructures in the form of metal nanoparticles (NPs), NWs or nanogratings have attracted a lot of attention as a means of light trapping in solar cells. Metal nanostructures in the nanoscale support plasmons which are collective oscillations of conduction electrons at a metal–dielectric interface. They support resonances that can be tuned by changing the geometry (size, shape, aspect ratio), the choice of metal or the dielectric environment around it (Raether, 1988). An entirely new field called Plasmonics is credited to the unique properties of these metallic nanostructures that can scatter, concentrate or manipulate the flow of light. Inside the device, the scattered light effectively increases the optical path length, while the near-field effect increases the electric field in the vicinity of the plasmons through the excitation of localized surface plasmon resonance (LSPR). Both effects enhance

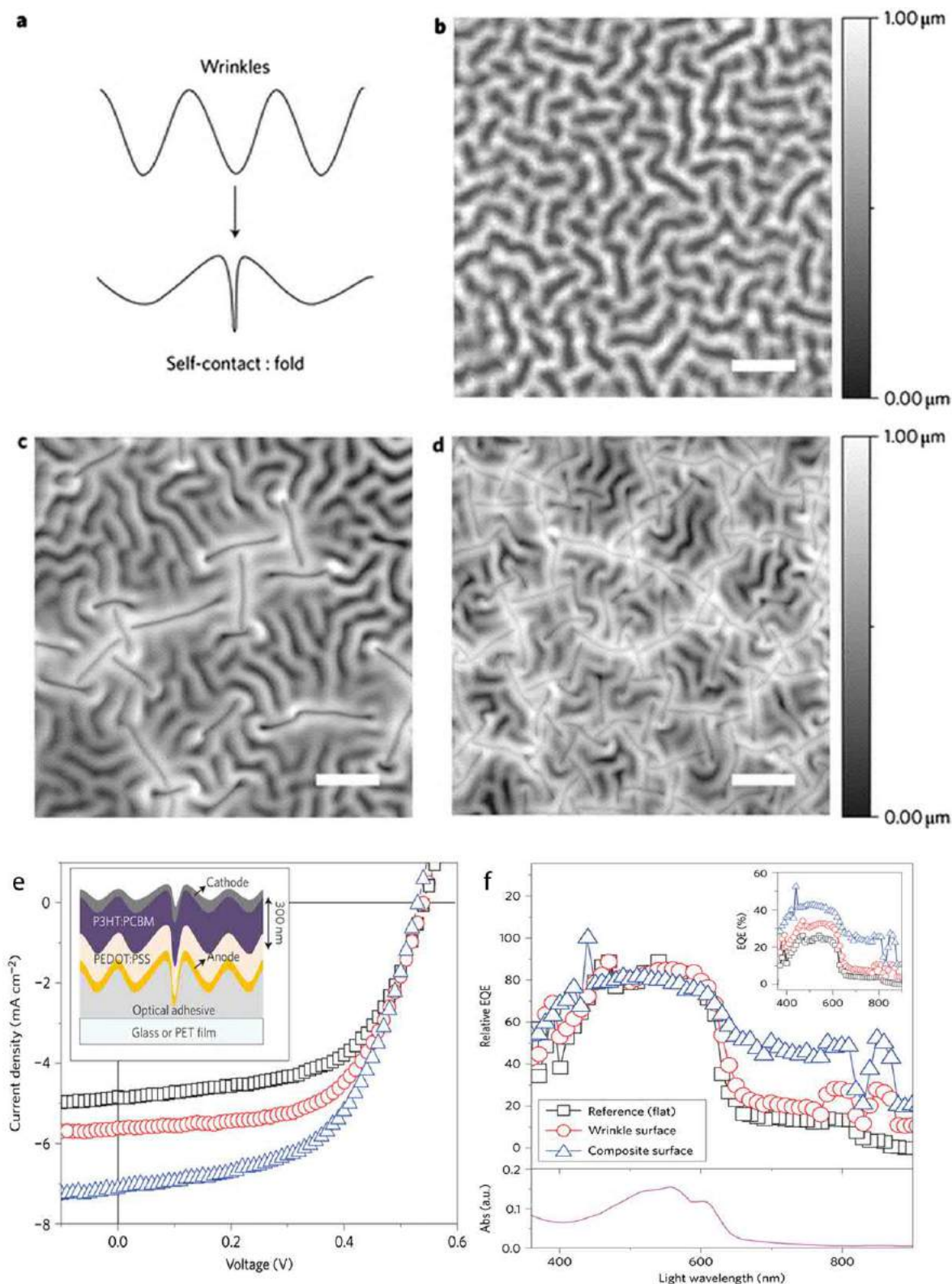


Fig. 6 (a)–(d) Scanning electron microscope (SEM) image of the morphology scan showing wrinkles developing into folds. (e) I–V characteristic of the cells with wrinkles and wrinkle and fold combination (inset) schematic of the structure with wrinkles and fold. (f) External quantum efficiency (EQE) of the polymer solar cell. The light is folded into the active layer by the folds of the adhesive and the wrinkles help localize the light through waveguiding effects. Taken from Kim, J.B., Kim, P., Pegard, N.C., *et al.*, 2012. Wrinkles and deep folds as photonic structures in photovoltaics. *Nature Photonics* 6 (5), 327–332.

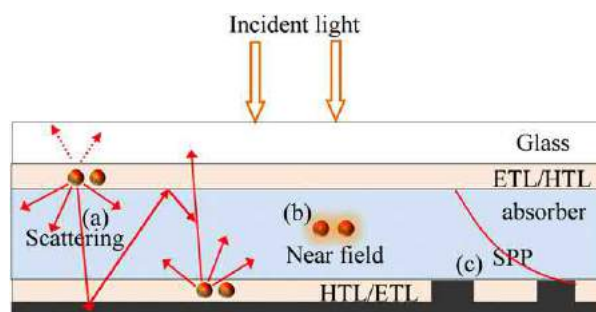


Fig. 7 Light trapping mechanisms using plasmonic effects: (a) scattering due to plasmons embedded in one of the hole/electron transport layers, (b) near field often with embedded particles in the active layer, and (c) propagating plasmons at a structured interface between metal electrode and active layer. ETL, electron transport layer; HTL, hole transport layer; SPP, surface plasmon polaritons.

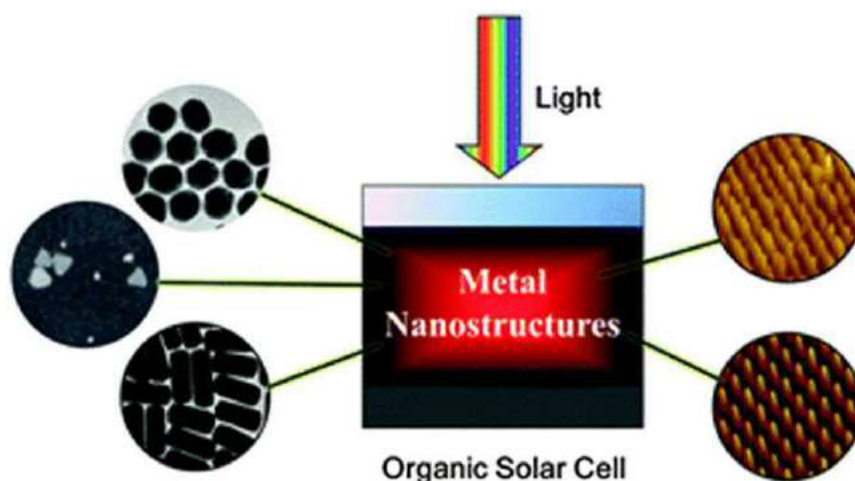


Fig. 8 Different plasmonic metal structures that can contribute to the light trapping in organic solar cells (OSCs) which includes nanoparticles (NPs), nanorods, (NRs) and grating structures. Available online only with abstract for reference Choy, W.C.H., 2014. The emerging multiple metal nanostructures for enhancing the light trapping of thin film organic photovoltaic cells. *Chemical Communications* 50 (81), 11984–11993.

absorption of light in the active absorber layer in a solar cell (Pillai and Green, 2010). Structured metal interfaces can also help couple the light using surface plasmon polaritons (SPP) that are propagating plasmons at a metal dielectric interface. Fig. 7 shows a schematic of the various plasmonic effects contributing to enhanced power conversion efficiency.

Among the metal NPs used to improve light absorption, silver (Ag) and Au are traditionally the focus of research, even though other metals also exhibit plasmonic properties (Pillai and Green 2010). In particular, aluminum (Al) NPs have been known to effectively reduce reflection losses due to their higher resonant frequency, and hence have limited metal absorption (Green and Pillai 2012). While an interband absorption in Al at 800 nm limits its use in Si solar cells (Temple and Bagnall, 2011), potential applications in organic photovoltaics (OPV) devices can be considered due to the larger bandgap of organic semiconductors.

Various plasmonic device architectures have been proposed to take advantage of the improved optical properties when plasmonic nanostructures are integrated into OSCs (Gan *et al.*, 2013; Chou and Chen, 2014; Stratakis and Kymakis, 2013) as shown in Fig. 8. The device structure can incorporate NPs, nanorods (NRs), nanoplatelets, nanoprisms, NWs, or nanogratings, which can be located in the active layer, the transport layers, buffer layer or at the interface between two layers (Chou and Chen, 2014). The improved optical properties in these metal nanostructures can manifest in the near field or far field, with various effects observed depending on the configuration of the plasmonic organic solar cell. A common methodology for fabricating plasmonic OSCs involves integration of plasmonic NPs (or discrete particles such as NRs) in one of the charge transport layer of the device as shown in Fig. 7(a).

Rand *et al.* (2004) reported one of the earliest device architectures incorporating plasmonic nanostructures, where enhanced optical field intensity in the near-field was observed when Ag nanoclusters were utilized in OSCs. This broadband enhanced field allowed increased light absorption up to 10 nm from the clusters. More recently, Au NPs and NRs were incorporated into the poly (3,4-ethylenedioxythiophene):polystyrene sulfonate (PEDOT:PSS) layer of an organic solar cell to trigger various LSPR bands. Improvements in light absorption due to an enhanced LSPR-induced local field resulted in an increase in the device efficiency by up to 24% (Hsiao *et al.*, 2012). Transmission electron microscopy (TEM) images of the plasmonic nanostructures and the improved EQE in the 450–600 nm wavelength range are shown in Fig. 9. Another example of near field enhancement can be seen

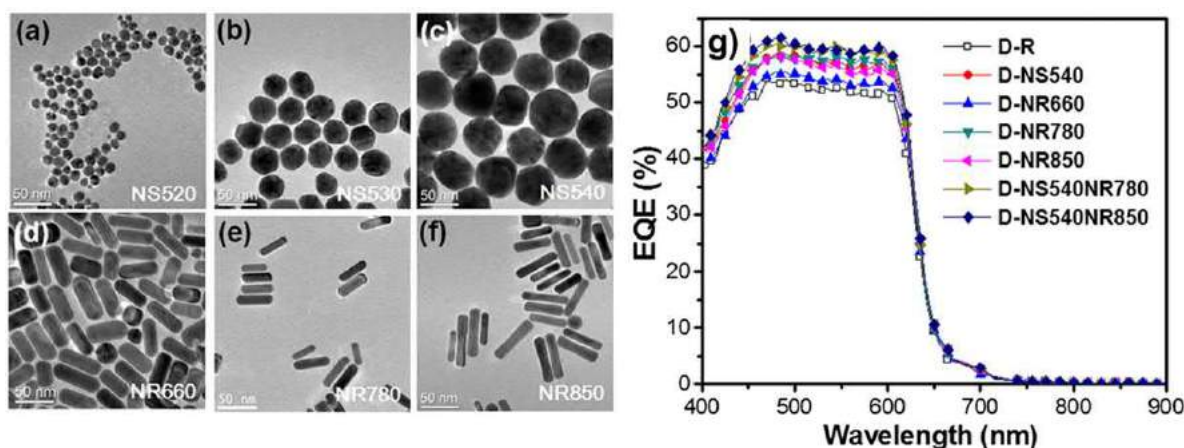


Fig. 9 (a)–(f) Transmission electron microscopy (TEM) images of different size/shape Au nanoparticle (NPs) with tunable localized surface plasmon resonance (LSPR) and (g) external quantum efficiency (EQE) spectra of poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS) layers incorporating the Au NPs at 100 mW/cm² illumination and EQE of devices containing various plasmonic nanostructures. Taken from Hsiao, Y.-S., Charan, S., Wu, F.-Y., *et al.*, 2012. Improving the light trapping efficiency of plasmonic polymer solar cells through photon management. *The Journal of Physical Chemistry C* 116 (39), 20731–20737.

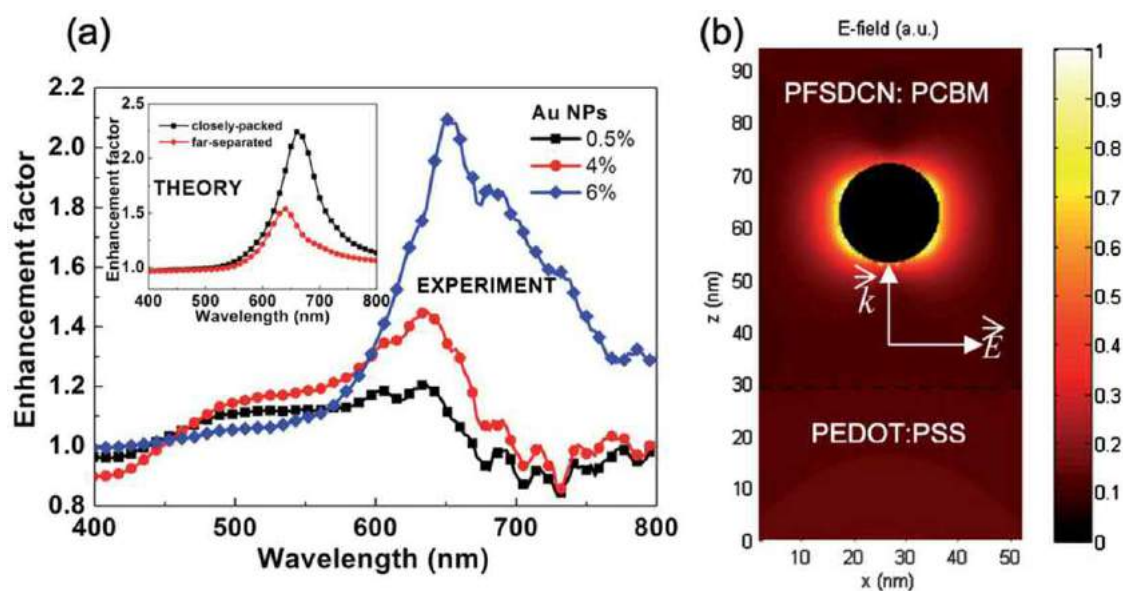


Fig. 10 (a) Experimental and theoretical demonstration of absorbance enhancement factor when Au nanoparticles (NPs) are embedded in the active layer and (b) the calculated near field enhancement around the Au NP in the active layer. PEDOT:PSS, poly(3,4-ethylenedioxythiophene) polystyrene sulfonate; PFSDCN:PCBM, poly[2,7-(9,9-dioctylfluorene)-alt-2-((4-(diphenylamino) phenyl)thiophen-2-yl)malononitrile]:[6,6]-phenyl-C61 butyric acid methyl ester. Taken from Wang, C.C.D., Choy, W.C.H., Duan, C., *et al.*, 2012. Optical and electrical effects of gold nanoparticles in the active layer of polymer solar cells. *Journal of Materials Chemistry* 22 (3), 1206–1211.

when Ag NPs are incorporated into the PEDOT:PSS using a pulse-current electrodeposition technique (Kim *et al.*, 2008). It was found that stronger light absorption is due to the high electromagnetic field strength in the vicinity of excited surface plasmons. While NPs can affect device performance when inadvertently contacting the active layer, there have been successful demonstrations of plasmonic nanostructures integrated in the active layer as represented in Fig. 7(b). In one report, Au NPs embedded into the active layer of the polymer solar cells showed improved performance both experimentally and theoretically due to excitation of LSPR (Wang *et al.*, 2012). The strong lateral distribution of the near field in this case is cited as the reason for enhancement as shown in the results in Fig. 10. Plasmonic nanostructures have also been used in a dual plasmon configuration. Liu *et al.* demonstrated increased J_{sc} from a device containing Au nano-dots and Au NPs in a blend with PEDOT:PSS respectively at each interface of the active layer (Liu *et al.*, 2013). The effective optical path length was increased due to the light scattering and LSPR effects of the Au NPs. It was also reported that increased light absorption from the Au nanodots was mainly derived from near-field

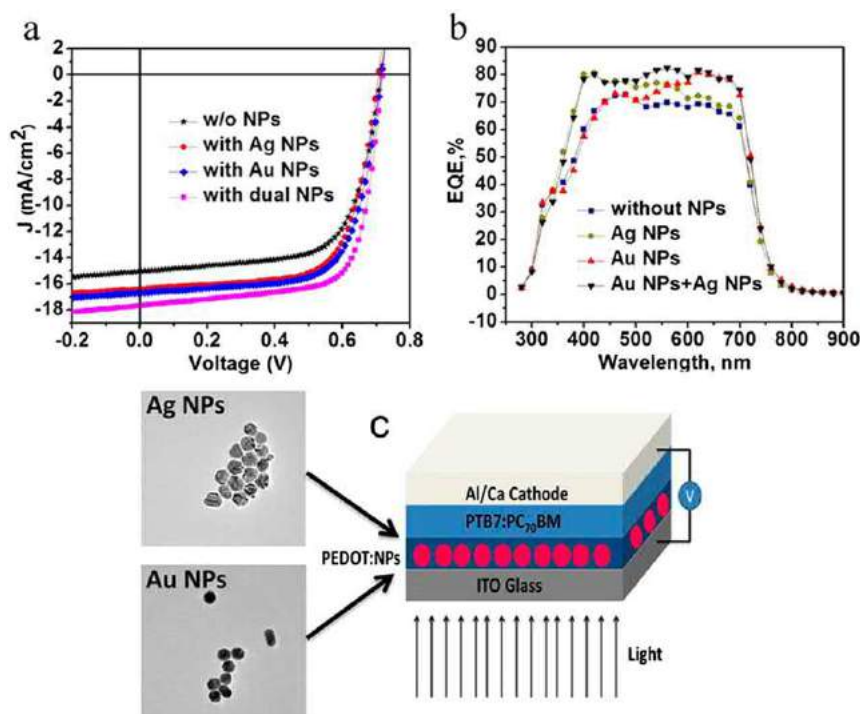


Fig. 11 (a) I–V characteristics of solar cells without and with Au and Ag nanoparticles (NPs), (b) external quantum efficiency (EQE) spectra without and with Au and Ag NPs, (c) schematic of the placement of the NP in the poly(3,4-ethylenedioxythiophene) (PEDOT) layer and the corresponding transmission electron microscope (TEM) images of the Au and Ag NPs. ITO, indium tin oxide; PTB7:PC₇₀BM, poly[[4,8-bis[(2-ethylhexyl)oxy]-benzo[1,2-b:4,5-b']dithiophene-2,6-diyl][3-fluoro-2-[(2-ethylhexyl)-carbonyl]thieno[3,4-b]thiophenediyl]]:[6,6]-phenyl-C70-butyric acid methyl ester. Taken from Lu, L., Luo, Z., Xu, T., Yu, L., 2013. Cooperative plasmonic effect of Ag and Au nanoparticles on enhancing performance of polymer solar cells. *Nano Letters* 13 (1), 59–64.

effects. Cooperative plasmonic effect using Au and Ag NP of ~ 40 – 50 nm diameter size embedded in the anode buffer layer PEDOT resulted in an 8.67% bulk heterojunction polymer cell accounting for 20% enhancement (Lu *et al.*, 2013). The observed effects were greater than any one type of the single NPs alone and were attributed to the dual resonance peaks of the two different NPs. The IV curve and the EQE of the comparative performance of poly[[4,8-bis[(2-ethylhexyl)oxy]-benzo[1,2-b:4,5-b']dithiophene-2,6-diyl][3-fluoro-2-[(2-ethylhexyl)-carbonyl]thieno[3,4-b]thiophenediyl]] (PTB7) and [6,6]-Phenyl-C70-butyric acid methyl ester (PC₇₀BM) blend polymer cell is shown in Fig. 11(a) and (b), respectively, along with a schematic and TEM images of the NPs in Fig. 11(c). In this work, it was concluded that the LSPR-induced local field not only helped to improve light absorption, but also aided the electrical properties like charge separation and transport thereby improving carrier density and lifetime.

There have been many observations of the near-field in plasmonic OSCs, however, far-field effects or light scattering can also significantly contribute to improved device performances. By varying the size of Ag NPs in PEDOT:PSS, a plasmonic forward scattering effect was demonstrated in PCDTBT:PC₇₀BM and PTB7:PC₇₀BM OSCs (Baek *et al.*, 2013). Enhanced light absorption due to plasmonic scattering by the Ag NPs led to improved EQE. Near-field scanning optical microscopy was used to probe the higher degree of light scattering in plasmonic devices. Simulations with these Ag NPs also found a high ratio of forward light scattering to total scattering, which indicates minimum backward scattered power loss.

Advanced plasmonic nanostructures, such as NWs, nanogratings and nanocomposites, have also been successfully integrated into OSCs. Kang *et al.* (2010) used a Ag NW electrode configuration to enhance power conversion efficiency by SPR and waveguide effects. Further improvements can be made to raise the efficiency by adjusting the period of the Ag grating such that the enhanced spectral range matches the peak absorption in organic semiconductors, thus demonstrating the flexibility and potential when integrating plasmonic nanostructures. Improvements in device performance extends to the application of Ag nanogratings and Au NPs in a dual plasmonic nanostructured device, which revealed broadband absorption enhancement (Li *et al.*, 2012). The improved broadband light absorption was ascribed to the simultaneous effects of LSPR (Au NPs) and hybridized SPR (Ag nanogratings).

As discussed previously, far-field effects also exist when these advanced plasmonic nanostructures are incorporated into OSCs. In one study, hole-mask colloidal lithography was used to pattern the Ag rear electrode and MoO_x hole transport layer (Niesen *et al.*, 2013). The resultant nanostructured Ag electrode as shown in Fig. 12(a) was tuned by varying the size of the protrusions in the MoO_x layer to maximize light scattering due to the dipole LSPR at the red absorption tail of the poly(3-hexylthiophene-2,5-diyl):indene-C60 bisadduct (P3HT:ICBA) active layer. Systematic optimization of the optical interference pattern, when changing the dimensions of the optical spacer and plasmonic nanostructure, were essential in achieving the observed enhanced absorption

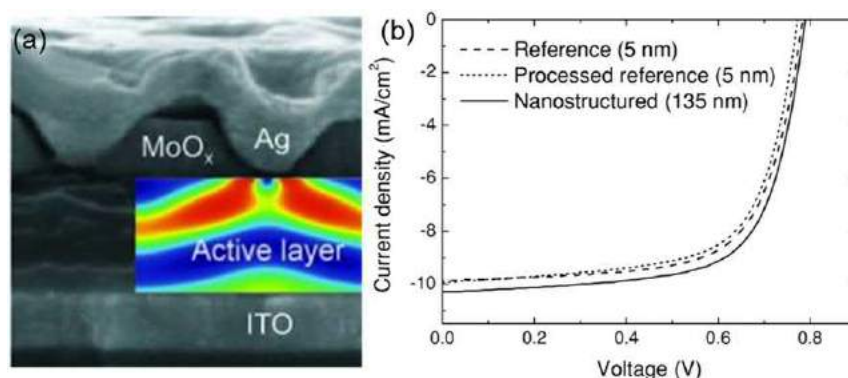


Fig. 12 (a) Scanning electron microscope (SEM) image of a nano-textured rear Ag electrode showing the different layers. Inset is the field profile in the active layer overlaid onto a section of the active layer and (b) I-V characteristics of flat reference devices with a 5-nm-thick MoO_x layer and the nanostructured device with a total MoO_x layer thickness of 135 nm with the increase attributed to light scattering.

(Fig. 12(b)). Importantly, it was noted that the near-field enhancement exists mainly within the MoO_x layer and, in that case, was not the cause of enhanced light absorption, rather the enhancement was the result of light scattering.

In general, it is often important to isolate the plasmonic nanostructures from the active layer. This is to avoid any metal assisted recombination or defects impacting the electrical properties of the cell. Another important consideration is also to avoid any parasitic losses in the metal that could compromise useful absorption. Plasmonic applications that rely on the near-field effect would need to address the trade-off between proximity of plasmon to the active absorber layer and electric field strength caused by SPR. The trade-off arises because it is advantageous to physically separate plasmonic nanostructures from the active layer to decrease surface recombination. However, the electric field enhancement provided by the plasmons decays exponentially with distance. Hence, there is an optimal physical separation to achieve high efficiency devices, which usually is ~ 20 nm in the vicinity of the particle. A systematic study of the near-field interaction between excitons and plasmonic metal NPs in OSCs found that far-field light scattering was not a significant consideration for the Ag NPs used in their study (Niesen *et al.*, 2012). As expected, the enhanced absorption from the near-field rapidly decreased when a spacer layer was introduced between the active layer and the NPs. In another experiment, Morfa *et al.* (2008) found that varying Ag film thickness affects the average J_{sc} and V_{oc} . It was suggested that the film thickness of the PEDOT:PSS layer should be limited as the electric field decays with distance.

Recent efforts in utilizing core-shell NPs have found promising results that both preserves the near-field and reduce surface recombination effects (Chen *et al.*, 2013; Janković *et al.*, 2013). A unique device structure for plasmonic OPV devices, that used large sized Au@SiO₂ NPs (50 nm SiO₂ shell and 70 nm Au core) which penetrated into all layers in the device had been successfully demonstrated (Chen *et al.*, 2013). The main contribution of the Au@SiO₂ NPs was reported to be LSPR effects from the Au NP core. Au-SiO₂ core-shell NRs can also be blended within the active layer to improve performance in poly[2,6-(4,4-bis-(2-ethylhexyl)-4H-cyclopenta [2,1-b;3,4-b']dithiophene)-alt-4,7(2,1,3-benzothia-diazole)]:[6,6]-phenyl-C70-butyric acid methyl ester (PCPDTBT:PC₇₀BM) OPV devices (Xu *et al.*, 2013).

The challenges of having the metal nanostructures incorporated into the device and the extent of understanding the physical mechanisms behind these effects have prompted many simulation studies. The studies attributed enhanced performance in plasmonic device to light concentration and redistribution of plasmons (Zhu *et al.*, 2012; Dunbar *et al.*, 2012), which can be further categorized depending on the positioning of plasmonic nanostructures: NPs dispersed in the active layer, nanostructured metallic electrode and in-scattering NPs. Simulations can be used to model light absorption and compared to a reference flat solar cell. The presence of the NPs provided an overall increase in absorption despite the reduction in absorption due to the omitted parts of the active layer. It is predicted that both scattering and near-field enhancement due to SPR are responsible for the broadband enhancement in light absorption. A simulation study has also analyzed the trade-off between absorption in the active layer as opposed to parasitic absorption in the near-field of the metal for the case of embedded spheres (Lee and Peumans 2010). The study showed that enhanced optical absorption results from strong scattering by a 10 nm Ag metal nanoparticle which locally enhances the optical electric fields. The enhancement extends to wavelengths longer than the plasmon resonance wavelength for which the losses in the metallic particles become insignificant.

A three-dimensional finite-difference time domain (FDTD) model is also commonly employed for simulation purposes. Using FDTD software, changes in metal NPs scattering cross-section, emission efficiency and far-field scattering pattern can be investigated by varying properties of a thin dielectric film around the NPs (Powell *et al.*, 2013). It is also imperative that the refractive indexes of materials within the device be considered when integrating metal NPs into OSCs. In particular, transparent conducting oxides, that frequently form the front contact in OPV devices, typically have large refractive indexes compared to the active layer. Hence, losses around the dipolar peak are expected due to the preferential light scattering into the higher index front contact.

Discussions on the progress in incorporating metallic nanostructures in OSCs and categorization of the physical mechanisms that led to improvements can be found in the literature (Choy, 2014). These improvements were found to originate from modifications of the optical or electrical properties in the OPV device. Changes in electrical properties, such as exciton dissociation,

charge carrier transport and charge extraction can also contribute to enhanced performance in plasmonic OSCs as mentioned earlier. Further discussions on the electrical effects of plasmonic OSCs can be found in the literature (Choy and Xingang, 2016; Chan *et al.*, 2016) and will not be discussed here.

Spacer Layer and Cavity Effects

Light harvesting by manipulating and optimizing the optical field can also be achieved via cavity effects or optical spacers. OSCs consist of many thin layers, each with a unique refractive index and thickness. (Burkhard *et al.*, 2010). Fig. 13 provides a schematic device structure, which includes a glass substrate, conducting oxide ITO electrode, thin electron transport layer (e.g., TiO_2), organic active layer (like P3HT:PCBM), a hole transport layer (e.g., MoO_3) and a reflective metal electrode (Ag).

As shown in Fig. 13, light enters the device through the glass substrate. Ideally, all the light would be absorbed in the active layer, however, in practice, this is not the case as mentioned before due to various losses. The light passes through the transparent electrode, some light is absorbed in the active layer, however, some light passes through and is reflected by the back electrode. Interference between the incoming and reflected light causes the optical field to go to zero at the reflective electrode. This leads to the formation of an optical field standing wave, of which the maximum and minimum will lie somewhere in the device. The nature of this optical field is determined by the refractive index and thickness of each layer.

It seems intuitive to simply increase the thickness of the active layer to maximize the absorption, however, the nature of the organic semiconductors places limitations on the film thickness. The optimum active layer film thickness is usually very thin, on the order of 100 nm, limited by the diffusion length of the excitons. Due to the limited film thickness, it is crucial to understand and optimize this optical interference effect.

As the maximum in the optical field is governed by the thickness and refractive index of the layers, it is possible to control the position of the maximum field by altering the device architecture. Including an additional layer to redistribute the optical field is called an optical spacer (Kim *et al.*, 2006). Fig. 14 describes the influence of an optical spacer. As shown, the insertion of a thin layer can shift the absorption maximum to better align with the active layer. It was shown that the insertion of a TiO_2 optical

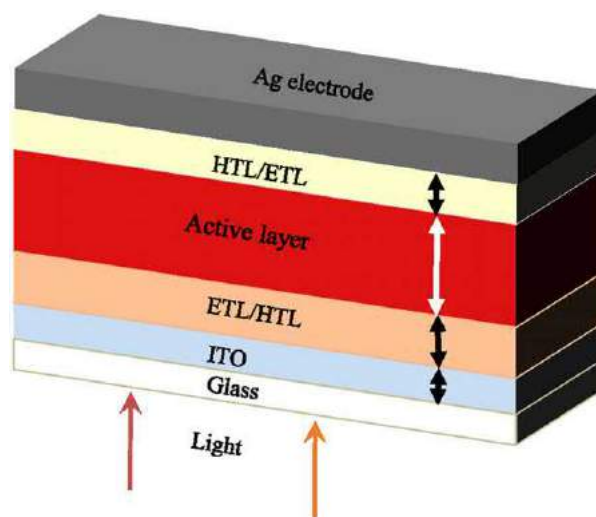


Fig. 13 Schematic diagram of a device structure showing the many layers involved. ETL, electron transport layer; HTL, hole transport layer; ITO, indium tin oxide.

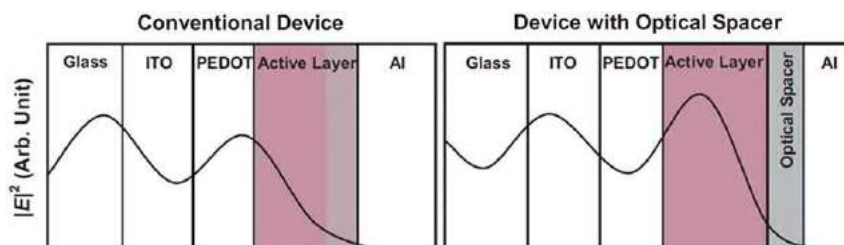


Fig. 14 Schematic representation of the optical field in an organic solar cell with and without an optical spacer. ITO, indium tin oxide; PEDOT, poly(3,4-ethylenedioxythiophene). Taken from Kim, J.Y., Kim, S.H., Lee, H.H., *et al.*, 2006. New architecture for high-efficiency polymer photovoltaic cells using solution-based titanium oxide as an optical spacer. *Advanced Materials* 18 (5), 572–576.

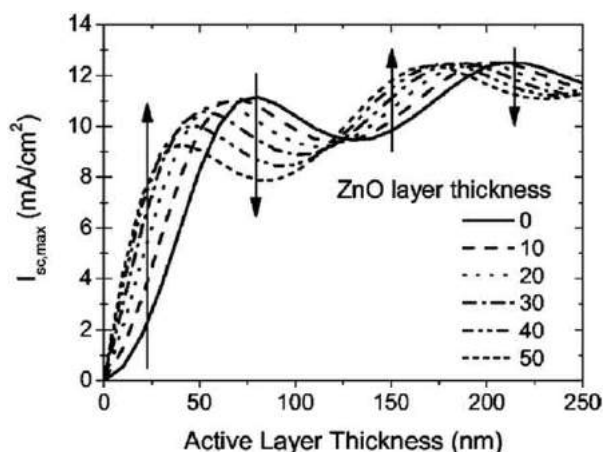


Fig. 15 Calculated short circuit current density (I_{sc}) as a function of active layer thickness for varying ZnO thicknesses. An assumption of 100% internal quantum efficiency (IQE) for all wavelengths was used. Reproduced from Gilot, J., Barbu, I., Wienk, M.M., Janssen, R.A.J., 2007. The use of ZnO as optical spacer in polymer solar cells: Theoretical and experimental study. *Applied Physics Letters* 91 (11), 113520.

spacer caused the power conversion efficiency to increase by 50% (relative) (Kim *et al.*, 2006). It is important to note that the optical spacer material must also serve a purpose electronically, as well as optically. For example, due to the spacer layer being part of the device, the photocatalytic property of TiO_2 under ultraviolet light can result in the breakdown of organic molecules limiting the lifetime of the device (Hiralal *et al.*, 2014). Hence the choice of the materials becomes important.

Gilot *et al.* (2007) used a ZnO layer as an optical spacer. They used transfer matrix modeling to calculate the optimum ZnO thickness, which would most effectively shift the optical field maximum into the active layer. They also calculated the maximum generated photocurrent (I_{sc}) as a function of active layer thickness, for varying ZnO layer thickness. This is displayed in Fig. 15. This clearly shows that the I_{sc} is not directly proportional to active layer thickness. The oscillations, causing maxima and minima, are due to the interference effects. By incorporating the ZnO optical spacer, the maxima are shifted to lower thicknesses. This positive influence of the optical spacer is more pronounced for thinner active layers.

Betancur *et al.* extend on this argument. They suggest that the position of the optical field maximum is dependent on the thickness of all layers in the device stack, not just the optical spacer (Betancur *et al.*, 2012). The device architecture, meaning the types of buffer layers used, which determines the direction of current flow, may also influence the optical field distribution. In another study it was suggested that an optical spacer will have much less impact on the photon absorption in an inverted structure device, when compared to conventional structure devices, which contain a PEDOT:PSS hole selective contact (Ameri *et al.*, 2008). In a study using ITO as spacer layer, it was found that although the degree of light absorption in inverted OPVs was not increased in such inverted structures, the favorable distribution of photogenerated excitons is expected to have decreased the level of exciton quenching near the electrodes (Chen *et al.*, 2010). Their results indicate that the inverted OPVs could still benefit from such optical effects when they had a sufficiently thick active layer. Subsequently there were studies that completely dismissed the idea of using spacer layer for increasing efficiency although the potential for increase was not ruled out for suboptimal thicknesses (Andersson *et al.*, 2009).

The nature of the interference effects also plays a role in the correct determination of the internal quantum efficiency (IQE) of an organic solar cell (Burkhard *et al.*, 2010). The IQE represents the wavelength dependent solar response for photons which are absorbed by the active layer; this parameter should be independent of the absorption profiles. Armin *et al.* showed that due to the nature of OSCs, optical interference and parasitic absorption in non-active layers must be considered to accurately determine the IQE (Armin *et al.*, 2014).

Another strategy adopted to help enhance the absorption in the thin photoactive layer is to incorporate a microcavity array into the device architecture. The concept of a microcavity design involves a resonator structure which consists of a spacer layer sandwiched between two reflectors (Chen *et al.*, 2014a,b). This allows for optical confinement due to the multiple reflections, significantly increasing the absorption path length of light with resonant frequencies. A normal OPV cell has a reflective back electrode; however, the front electrode is required to be transparent to allow light in. Such a structure behaves as an inefficient microcavity device, due to the limited reflection from the transparent electrode, which is most commonly ITO. This is made more complex as the two sandwiching layers must play a role in the electronic operation of the device, not only the optical confinement. To successfully harness this optical confinement technique, a precise trade-off between the transmission and reflection of the bottom electrode, to allow for light to enter the active layer and to enhance the multi-reflection optical confinement, respectively, must be understood.

To do so, the ITO layer was replaced. Studies have demonstrated an efficient device containing a microcavity structure by replacing the ITO with an ultrathin metal film (UTMF) and a capping layer which acted as an optical spacer (Chen *et al.*, 2014a,b; Sergeant *et al.*, 2013). The UTMF behaves as a semitransparent mirror enhancing lateral conductivity and reduce sheet resistance. One such modified device structure is shown schematically in Fig. 16. As shown, the light shines through the top, semitransparent

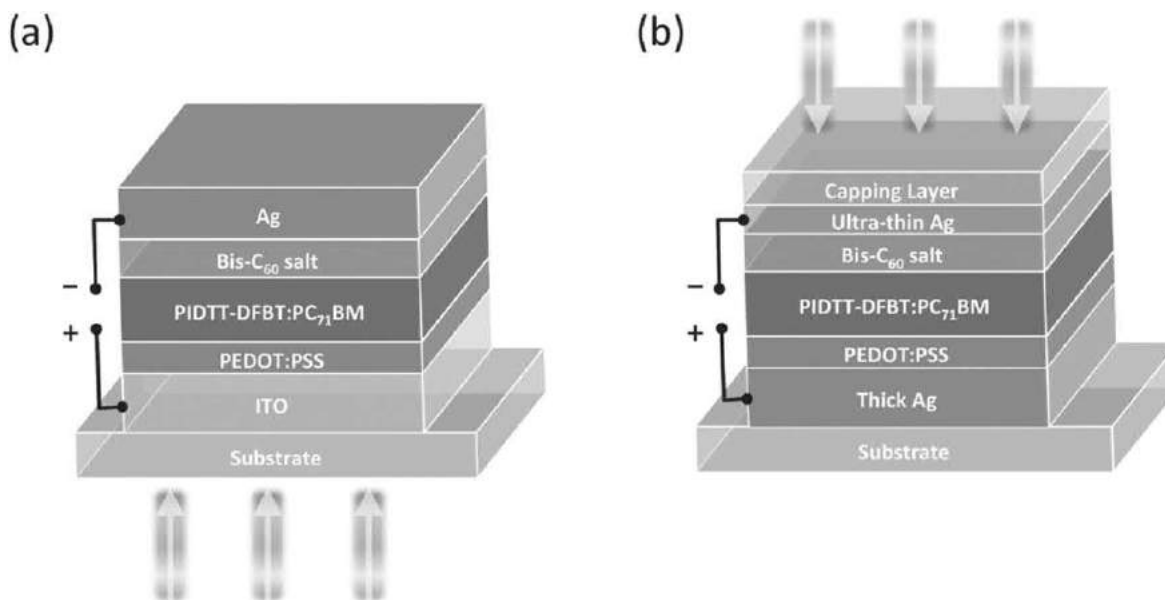


Fig. 16 Schematic depiction of (a) the traditional device structure, containing indium tin oxide (ITO), and (b) the microcavity configuration containing an ultrathin metal film (UTMF) and capping layer to increase the incoupling of light. In this device structure, the light is illuminated through the top electrode. ITO, indium tin oxide; PEDOT:PSS, poly(3,4-ethylenedioxythiophene):polystyrene sulfonate, PIDTT-DFBT, poly[(indacenodithieno[3,2-b]thiophene)-alt-difluorobenzothiadiazole]. Reproduced from Chen, K.-S., Yip, H.-L., Salinas, J.-F., *et al.*, 2014b. Strong photocurrent enhancements in highly efficient flexible organic solar cells by adopting a microcavity configuration. *Advanced Materials* 26 (20), 3349–3354.

electrode. The device contained an Ag UTMF and a TeO₂ capping layer. Compared to the reference, with an ITO electrode, the average efficiency increased from 7.21 to 8.47%, which was due mainly to the J_{sc} increase. This was attributed to the increased intensity in the EQE curve, caused by the resonance effect. In addition, a device employing a variation to this architecture was fabricated using a flexible PET substrate, with an efficiency of 8.37%, only slightly lower than that for the glass substrate. Huang *et al.* achieved higher performance by coupling a microcavity structure with a high performance bulk heterojunction (BHJ) blend of poly({4,8-bis[(2-ethylhexyl)oxy]benzo[1,2-b:4,5-b']dithiophene-2,6-diyl}{3-fluoro-2-[(2-ethylhexyl)-carbonyl]-thieno[3,4-b]thiophenediyl}) (PBDDT-F-TT):PC₇₁BM (Huang *et al.*, 2015). They also used an UTMF with a TeO₂ buffer layer to achieve the microcavity resonance. The best performing device achieved an efficiency of 10.4%. The enhancement mechanism was similar to that in reference (Chen *et al.*, 2014a,b), a broadband improvement in the EQE profile correlated with improved J_{sc} . The structure was also tested on large area devices. An efficiency of more than 7% was achieved for large area devices, due to a novel semitransparent electrode containing the UTMF. This electrode was shown to be much more applicable to large device areas, when compared with ITO. Additionally, it is advantageous to replace the ITO layer, as it represents the highest cost and embodied energy in thin film solar cells.

Other Approaches for Light Harvesting

The narrow absorption band of OPV limits the use of the wider solar spectrum for energy conversion. One way to circumvent this issue is the use of multijunction or tandem solar cells in stacks with each of the cells absorbing different parts of the spectrum. The concept of tandem cells was proposed as a third generation solar cell alternative to increase light harvesting for solar applications (Green, 2001). The structure consists of a high bandgap top cell absorbing the high energy photons and thereby transmitting the low energy photons to the subsequent bottom solar cells which is arranged in the order of decreasing bandgap through a transparent interconnecting layer. For series connected cells (i.e., contact is made between the top and bottom cell), the open circuit voltage (V_{oc}) of the tandem cell is the sum of V_{oc} of the subcells, while the overall current is limited by the subcell with the smaller current. This means that the cells need to be current matched. Another approach is a three (Sista *et al.*, 2010a,b) or four terminal configuration in a two cell stack that allows separate connections, which does not require current matching. In this device configuration the two subcells are connected in parallel through a common/separate semitransparent interlayer. Hence tandem cell approach allows light harvesting of a much wider spectrum of the solar radiation than what would be possible by a single junction solar cell.

With this strategy, tremendous efforts have been devoted to the development of desirable device configurations. The lack of high performance low bandgap polymer materials along with good interconnection layer has limited the performance of tandem cells (Kim *et al.*, 2007; You *et al.*, 2013; Sista *et al.*, 2010a,b). The interconnection layers mostly used semitransparent metal layers

that added to the complexity of device fabrication and caused undesirable parasitic absorption. One of the earlier reports was that of 6.5% performance conversion efficiency at 200 W/cm^2 from fully solution processed tandem cells comprising a transparent TiO_x layer between the front and bottom cell. It made use of an inverted structure with the low bandgap polymer-fullerene composite as the charge-separating layer in the front cell and the high band-gap polymer composite as the back cell (Kim *et al.*, 2007). While tandem solar cells benefit from complementary absorption spectra by utilizing different polymer materials for the sub cells, more recently an efficiency of 6.7% was reported from identical subcells based on P2F-DO:N2200, which the authors claim is among the highest efficiencies for an all polymer tandem solar cell and attributed to the enhanced absorption of the tandem cell (Yuan *et al.*, 2016).

More understanding and development of the technology has made further improvements possible with a power conversion efficiencies exceeding 11% for a triple junction solar cell (see Fig. 17(a) and (b)) (Chen *et al.*, 2014a,b) which was greater than the first certified efficiency of 10.6% from a double junction tandem from the same group (You *et al.*, 2013).

Tremendous efforts have been devoted to the development of desirable device configurations using the tandem strategy. One such structure, unlike the sequential deposited layers, consists of a folded V-structure, like the one previously mentioned for light trapping (Rim *et al.*, 2007). It also allows an alternative solution to the problem of assembling tandem and multiple junction solar cells in the micron to centimeter range. Here the single cells are reflecting the non-absorbed light upon another adjacent cell and by folding the cells spectrally toward each other like in Fig. 5(c) wherein spectral broadening and light trapping are combined (Tvingstedt *et al.*, 2007).

Photon absorption and reemission in a quantum dot (QD) layer is another scattering technique capable of light harnessing in a polymer layer. Whilst only in its conceptual stage, this method has the potential to increase absorption by 28% (Xu and Munday, 2014). Here high luminescence QDs are used to absorb photons and reemit to the waveguide modes, thereby changing the directionality of the incident photons to improve coupling in the active layer. Another method of light harnessing is when

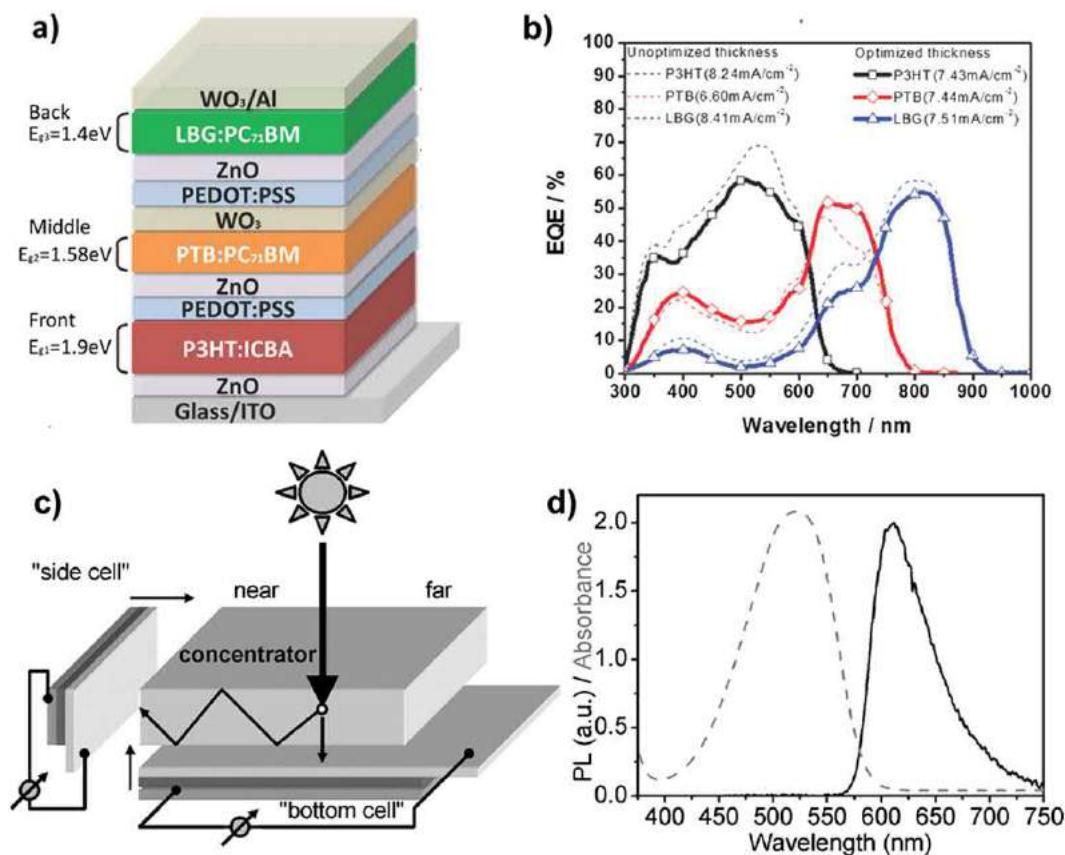


Fig. 17 (a) Layers of a triple junction organic tandem solar cell arranged in decreasing order of bandgap (light incident from glass side), (b) measured external quantum efficiency (EQE) of device in (a) showing a wider use of the solar spectrum, (c) schematic of a luminescent concentrator with cells placed either on the side or on the bottom of the concentrator plate, (d) absorbance (dashed) and photoluminescence (solid) spectra of the concentrator plate. PEDOT:PSS, poly(3,4-ethylenedioxythiophene) polystyrene sulfonate; ITO, indium tin oxide; LBG:PC₇₁BM, low band gap:phenyl-C61-butyric acid methyl ester; P3HT:ICBA, poly(3-hexylthiophene-2,5-diyl):indene-C60 bisadduct. Images (a) and (b) taken from Chen, C.C., Chang, W.H., Yoshimura, K., *et al.*, 2014a. An efficient triple-junction polymer solar cell having a power conversion efficiency exceeding 11%. *Advanced Materials* 26, 5670–5677; (c) and (d) from Koeppel, R., Sariciftci, N.S., Büchtemann, A., 2007. Enhancing photon harvesting in organic solar cells with luminescent concentrators. *Applied Physics Letters* 90 (18), 181126.

luminescence concentrators are optically coupled to OSCs as shown in Fig. 17(c) where the solar cell can be placed either on the side or bottom of the concentrator plate. Luminescence concentrators use organic dyes with high photoluminescence quantum yields that concentrate and red shift the incident light (also called down conversion) and couples it to a waveguide. OSCs fabricated using zinc-phthalocyanine and C_{60} with the absorption in the 600–800 nm range, demonstrated photocurrent density increase from 8.5 to 10 mA/cm² by spectrally shifting the blue and green light (below 600 nm) toward the red when using such concentrators fixed to the cell (Koeppel *et al.*, 2007). Here the luminescence spectrum of the dye must overlap with the maximum in the conversion efficiency of the solar cell for optimum performance. The absorbance and emission spectrum of the concentrator plate used in the study in (Koeppel *et al.*, 2007) is shown in Fig. 17(d).

The wider implementation of low-cost OSCs relies on increasing the efficiency of this technology through material improvements and optical gains. With myriad promising opportunities available for light harvesting for OSCs, the challenge is in optimizing one or more techniques into the cell without compromising on the electrical gains and at the same time keeping the fabrication cost low. In depth understanding the physical mechanisms for each of these techniques along with its dependence on the material properties is hence critical. For metal-based structures the optical properties should always be considered along with its electrical properties of OSCs. The ultimate goal of light harnessing structures should be to increase the absolute absorption in the thin film layers and not necessarily approaching the efficiency limits (Tang *et al.*, 2014). More detailed light harnessing reviews are available for those interested in further reading (Tang *et al.*, 2014; Ou *et al.*, 2016; Wang *et al.*, 2014; Fei Guo *et al.*, 2014).

Conclusion

OSCs, due to the exciton diffusion bottleneck that limits charge carrier transport in such cells, are severely limited by the thickness of the active layer. This comes with challenges in absorption of light and hence requires efficient light harvesting schemes to minimize optical losses. Different light harnessing schemes were discussed that includes AR coatings, light trapping using dielectric and metallic structures, light concentration using plasmonics, cavity effects and also the tandem structures that make use of different parts of the solar spectrum using cell stacks of different bandgaps. While the poor absorption efficiency in such cells have limited this technology to be a major player in the photovoltaics market compared to Si and other emerging thin film technologies, OSCs that can offer cheap, light weight and pliable cells may see a bright future as technology improves. It is expected that this technology may not need to be a replacement technology to Si but has the potential to find its own niche market for example, on windows or electric vehicles which require light weight structures. However, irrespective of the application, efficiency improvements will be key and it can be foreseen that light harnessing will play a major role in the development of OSCs given the potential for remarkable improvements possible as discussed in this article.

References

- Ameri, T., Dennler, G., Waldauf, C., *et al.*, 2008. Realization, characterization, and optical modeling of inverted bulk-heterojunction organic solar cells. *Journal of Applied Physics* 103 (8), 084506.
- Andersson, B.V., Huang, D.M., Moulé, A.J., Inganäs, O., 2009. An optical spacer is no panacea for light collection in organic solar cells. *Applied Physics Letters* 94 (4), 043302.
- Armin, A., Velusamy, M., Wolfer, P., *et al.*, 2014. Quantum efficiency of organic solar cells: Electro-optical cavity considerations. *ACS Photonics* 1 (3), 173–181.
- Baek, S.W., Noh, J., Lee, C.H., *et al.*, 2013. Plasmonic forward scattering effect in organic solar cells: A powerful optical engineering method. *Scientific Reports* 3 (1726), 1726.
- Bernhard, C.G., 1967. Structural and functional adaptation in a visual system. *Endeavour* 26 (79), 79–84.
- Betancur, R., Martínez-Otero, A., Elias, X., *et al.*, 2012. Optical interference for the matching of the external and internal quantum efficiencies in organic photovoltaic cells. *Solar Energy Materials & Solar Cells* 104, 87–91.
- Burkhard, G.F., Hoke, E.T., McGehee, M.D., 2010. Accounting for interference, scattering, and electrode absorption to make accurate internal quantum efficiency measurements in organic and other thin solar cells. *Advanced Materials* 22 (30), 3293–3297.
- Cao, W., Myers, J.D., Zheng, Y., *et al.*, 2011. Enhancing light harvesting in organic solar cells with pyramidal rear reflectors. *Applied Physics Letters* 99 (2), 023306.
- Chan, K., Wright, M., Elumalai, N., Uddin, A., Pillai, S., 2016. Plasmonics in organic and perovskite solar cells: Optical and electrical effects. *Advanced Optical Materials* 5 (6), 1600698.
- Chen, B., Zhang, W., Zhou, X., *et al.*, 2013. Surface plasmon enhancement of polymer solar cells by penetrating Au/SiO₂ core/shell nanoparticles into all organic layers. *Nano Energy* 2 (5), 906–915.
- Chen, C.C., Chang, W.H., Yoshimura, K., *et al.*, 2014a. An efficient triple-junction polymer solar cell having a power conversion efficiency exceeding 11%. *Advanced Materials* 26, 5670–5677.
- Chen, F.-C., Wu, J.-L., Hung, Y., 2010. Spatial redistribution of the optical field intensity in inverted polymer solar cells. *Applied Physics Letters* 96 (19), 193304.
- Chen, K.-S., Yip, H.-L., Salinas, J.-F., *et al.*, 2014b. Strong photocurrent enhancements in highly efficient flexible organic solar cells by adopting a microcavity configuration. *Advanced Materials* 26 (20), 3349–3354.
- Chou, C.H., Chen, F.C., 2014. Plasmonic nanostructures for light trapping in organic photovoltaic devices. *Nanoscale* 6 (15), 8444–8458.
- Choy, W.C.H., 2014. The emerging multiple metal nanostructures for enhancing the light trapping of thin film organic photovoltaic cells. *Chemical Communications* 50 (81), 11984–11993.
- Choy, W.C.H., Xingang, R., 2016. Plasmon-electrical effects on organic solar cells by incorporation of metal nanostructures. *IEEE Journal of Selected Topics in Quantum Electronics* 22 (1), 4100209.
- Dunbar, R.B., Pladler, T., Mende, L.S., 2012. Highly absorbing solar cells—a survey of plasmonic nanostructures. *Optics Express* 20, A177–A189.
- Fei Guo, C., Sun, T., Cao, F., Liu, Q., Ren, Z., 2014. Metallic nanostructures for light trapping in energy-harvesting devices. *Light: Science & Applications* 3, e161.

- Forberich, K., Dennler, G., Scharber, M.C., *et al.*, 2008. Performance improvement of organic solar cells with moth eye anti-reflection coating. *Thin Solid Films* 516 (20), 7167–7170.
- Forrest, S.R., 2005. The limits to organic photovoltaic cell efficiency. *MRS Bulletin* 30 (1), 28–32.
- Gan, Q., Bartoli, F.J., Kafafi, Z.H., 2013. Plasmonic-enhanced organic photovoltaics: Breaking the 10% efficiency barrier. *Advanced Materials* 25 (17), 2385–2396.
- Gilot, J., Barbu, I., Wien, M.M., Janssen, R.A.J., 2007. The use of ZnO as optical spacer in polymer solar cells: Theoretical and experimental study. *Applied Physics Letters* 91 (11), 113520.
- Green, M.A., 2001. Third generation photovoltaics: Ultra-high conversion efficiency at low cost. *Progress in Photovoltaics* 9, 123–135.
- Green, M.A., 2011. Enhanced evanescent mode light trapping in organic solar cells and other low index optoelectronic devices. *Progress in Photovoltaics: Research and Applications* 19 (4), 473–477.
- Green, M.A., Pillai, S., 2012. Harnessing plasmonics for solar cells. *Nature Photonics* 6 (3), 130–132.
- Gregg, B.A., van de Lagemaat, J., 2012. Solar cells: Folding photons. *Nature Photonics* 6 (5), 278–280.
- Hiralal, P., Chien, C., Lal, N.N., *et al.*, 2014. Nanowire-based multifunctional antireflection coatings for solar cells. *Nanoscale* 6 (23), 14555–14562.
- Hsiao, Y.-S., Charan, S., Wu, F.-Y., *et al.*, 2012. Improving the light trapping efficiency of plasmonic polymer solar cells through photon management. *The Journal of Physical Chemistry C* 116 (39), 20731–20737.
- Huang, J., Li, C.-Z., Chueh, C.-C., *et al.*, 2015. 10.4% Power conversion efficiency of ITO-free organic photovoltaics through enhanced light trapping configuration. *Advanced Energy Materials* 5 (15), 1500406.
- Huang, Y.-F., Chattopadhyay, S., Jen, Y.-J., *et al.*, 2007. Improved broadband and quasi-omnidirectional anti-reflection properties with biomimetic silicon nanostructures. *Nature Nanotechnology* 2 (12), 770–774.
- Hünig, R., Mertens, A., Stephan, M., *et al.*, 2016. Flower power: Exploiting plants' epidermal structures for enhanced light harvesting in thin-film solar cells. *Advanced Optical Materials* 4 (10), 1487–1493.
- Janković, V., Yeng, Y., You, J., *et al.*, 2013. Active layer-incorporated, spectrally tuned Au/SiO₂ core/shell nanorod-based light trapping for organic photovoltaics. *ACS Nano* 7 (5), 3815–3822.
- Kang, M.G., Xu, T., Park, H.J., Luo, X., Guo, L.J., 2010. Efficiency enhancement of organic solar cells using transparent plasmonic Ag nanowire electrodes. *Advanced Materials* 22 (39), 4378–4383.
- Kim, J.B., Kim, P., Pegard, N.C., *et al.*, 2012. Wrinkles and deep folds as photonic structures in photovoltaics. *Nature Photonics* 6 (5), 327–332.
- Kim, J.Y., Kim, S.H., Lee, H.H., *et al.*, 2006. New architecture for high-efficiency polymer photovoltaic cells using solution-based titanium oxide as an optical spacer. *Advanced Materials* 18 (5), 572–576.
- Kim, J.Y., Lee, K., Coates, N.E., *et al.*, 2007. Efficient tandem polymer solar cells fabricated by all-solution processing. *Science* 317 (5835), 222–225.
- Kim, S.S., Na, S.I., Jo, J., Kim, D.Y., Nah, Y.C., 2008. Plasmon enhanced performance of organic solar cells using electrodeposited Ag nanoparticles. *Applied Physics Letters* 93 (7), 073307.
- Koeppel, R., Sariciftci, N.S., Büchtemann, A., 2007. Enhancing photon harvesting in organic solar cells with luminescent concentrators. *Applied Physics Letters* 90 (18), 181126.
- Krebs, F.C., 2009. Fabrication and processing of polymer solar cells: A review of printing and coating techniques. *Solar Energy Materials & Solar Cells* 93 (4), 394–412.
- Lee, J.-Y., Peumans, P., 2010. The origin of enhanced optical absorption in solar cells with metal nanoparticles embedded in the active layer. *Optics Express* 18 (10), 10078–10087.
- Lee, Y.-J., Ruby, D.S., Peters, D.W., McKenzie, B.B., Hsu, J.W.P., 2008. ZnO nanostructures as efficient antireflection layers in solar cells. *Nano Letters* 8 (5), 1501–1505.
- Leem, J.W., Kim, S., Lee, S.H., *et al.*, 2014. Efficiency enhancement of organic solar cells using hydrophobic antireflective inverted moth-eye nanopatterned PDMS films. *Advanced Energy Materials* 4 (8), 1301315.
- Li, X., Choy, W.C.H., Huo, L., *et al.*, 2012. Dual plasmonic nanostructures for high performance inverted organic solar cells. *Advanced Materials* 24 (22), 3046–3052.
- Liu, C.M., Chen, C.M., Su, Y.W., Wang, S.M., Wei, K.H., 2013. The dual localized surface plasmonic effects of gold nanodots and gold nanoparticles enhance the performance of bulk heterojunction polymer solar cells. *Organic Electronics: Physics, Materials, Applications* 14 (10), 2476–2483.
- Lu, L., Luo, Z., Xu, T., Yu, L., 2013. Cooperative plasmonic effect of Ag and Au nanoparticles on enhancing performance of polymer solar cells. *Nano Letters* 13 (1), 59–64.
- Ma, W., Yang, C., Gong, X., Lee, K., Heeger, A.J., 2005. Thermally stable, efficient polymer solar cells with nanoscale control of the interpenetrating network morphology. *Advanced Functional Materials* 15 (10), 1617–1622.
- Mallick, S.B., Sergeant, N.P., Agrawal, M., Lee, J.-Y., Peumans, P., 2011. Coherent light trapping in thin-film photovoltaics. *MRS Bulletin* 36 (6), 453–460.
- Morfa, A.J., Rowlen, K.L., Reilly, T.H., Romero, M.J., Van De Lagemaat, J., 2008. Plasmon-enhanced solar energy conversion in organic bulk heterojunction photovoltaics. *Applied Physics Letters* 92 (1), 013504.
- Muskens, O.L., Rivas, J.G., Algra, R.E., Bakkers, E.P.A.M., Legendijk, A., 2008. Design of light scattering in nanowire materials for photovoltaic applications. *Nano Letters* 8 (9), 2638–2642.
- Myers, J.D., Cao, W., Cassidy, V., *et al.*, 2012. A universal optical approach to enhancing efficiency of organic-based photovoltaic devices. *Energy & Environmental Science* 5 (5), 6900–6904.
- Niesen, B., Rand, B.P., Van Dorpe, P., *et al.*, 2012. Near-field interactions between metal nanoparticle surface plasmons and molecular excitons in thin-films. Part I: Absorption. *The Journal of Physical Chemistry C* 116 (45), 24206–24214.
- Niesen, B., Rand, B.P., Van Dorpe, P., *et al.*, 2013. Plasmonic efficiency enhancement of high performance organic solar cells with a nanostructured rear electrode. *Advanced Energy Materials* 3 (2), 145–150.
- Ou, Q.-D., Li, Y.-Q., Tang, J.-X., 2016. Light manipulation in organic photovoltaics. *Advanced Science* 3 (7), 1600123. -n/a.
- Peer, A., Biswas, R., 2014. Nanophotonic organic solar cell architecture for advanced light trapping with dual photonic crystals. *ACS Photonics* 1 (9), 840–847.
- Pillai, S., Green, M.A., 2010. Plasmonics for photovoltaic applications. *Solar Energy Materials and Solar Cells* 94 (9), 1481–1486.
- Powell, A.W., Wincott, M.B., Watt, A.A.R., Assender, H.E., Smith, J.M., 2013. Controlling the optical scattering of plasmonic nanoparticles using a thin dielectric layer. *Journal of Applied Physics* 113 (18), 184311.
- Raether, H., 1988. *Surface Plasmons on Smooth and Rough Surface and on Gratings*. Berlin: Springer Verlag.
- Rand, B.P., Peumans, P., Forrest, S.R., 2004. Long-range absorption enhancement in organic tandem thin-film solar cells containing silver nanoclusters. *Journal of Applied Physics* 96 (12), 7519–7526.
- Rim, S.-B., Zhao, S., Scully, S.R., McGehee, M.D., Peumans, P., 2007. An effective light trapping configuration for thin-film solar cells. *Applied Physics Letters* 91 (24), 243501.
- Sergeant, N.P., Niesen, B., Liu, A.S., *et al.*, 2013. Resonant cavity enhanced light harvesting in flexible thin-film organic solar cells. *Optics Letters* 38 (9), 1431–1433.
- Sista, S., Hong, Z., Park, M.-H., Xu, Z., Yang, Y., 2010a. High-efficiency polymer tandem solar cells with three-terminal structure. *Advanced Materials* 22 (8), E77–E80.
- Sista, S., Park, M.-H., Hong, Z., *et al.*, 2010b. Highly efficient tandem polymer photovoltaic cells. *Advanced Materials* 22 (3), 380–383.
- Stratakis, E., Kymakis, E., 2013. Nanoparticle-based plasmonic organic photovoltaic devices. *Materials Today* 16 (4), 133–146.
- Tang, Z., Tress, W., Inganäs, O., 2014. Light trapping in thin film organic solar cells. *Materials Today* 17 (8), 389–396.
- Temple, T.L., Bagnall, D.M., 2011. Optical properties of gold and aluminium nanoparticles for silicon solar cell applications. *Journal of Applied Physics* 109, 084343.
- Tvingstedt, K., Andersson, V., Zhang, F., Inganäs, O., 2007. Folded reflective tandem polymer solar cell doubles efficiency. *Applied Physics Letters* 91 (12), 123514.
- Tvingstedt, K., Zilio, S.D., Inganäs, O., Tormen, M., 2008. Trapping light with micro lenses in thin film organic photovoltaic cells. *Optics Express* 16 (26), 21608–21615.
- Vezie, M.S., Few, S., Meager, I., *et al.*, 2016. Exploring the origin of high optical absorption in conjugated polymers. *Nature Materials* 15, 746–753.

- Wang, C.C.D., Choy, W.C.H., Duan, C., *et al.*, 2012. Optical and electrical effects of gold nanoparticles in the active layer of polymer solar cells. *Journal of Materials Chemistry* 22 (3), 1206–1211.
- Wang, H.-P., Lien, D.-H., Tsai, M.-L., *et al.*, 2014. Photon management in nanostructured solar cells. *Journal of Materials Chemistry C* 2 (17), 3144–3171.
- Xu, X., Kyaw, A.K.K., Peng, B., *et al.*, 2013. A plasmonically enhanced polymer solar cell with gold–silica core–shell nanorods. *Organic Electronics* 14 (9), 2360–2368.
- Xu, Y., Munday, J.N., 2014. Light trapping in a polymer solar cell by tailored quantum dot emission. *Optics Express* 22 (S2), A259–A267.
- Yablonovitch, E., 1982. Statistical ray optics. *Journal of the Optical Society of America* 72, 899.
- You, J., Dou, L., Yoshimura, K., *et al.*, 2013. A polymer tandem solar cell with 10.6% power conversion efficiency. *Nature Communications* 4, 1446.
- Yu, G., Gao, J., Hummelen, J.C., Wudl, F., Heeger, A.J., 1995. Polymer photovoltaic cells: Enhanced efficiencies via a network of internal donor-acceptor heterojunctions. *Science* 270 (5243), 1789–1791.
- Yu, Z., Raman, A., Fan, S., 2011. Nanophotonic light-trapping theory for solar cells. *Applied Physics A*. doi:10.1007/s00339-011-6617-4.
- Yuan, J., Gu, J., Shi, G., *et al.*, 2016. High efficiency all-polymer tandem solar cells. *Scientific Reports* 6, 26459.
- Zhou, L., Ou, Q.D., Chen, J.D., *et al.*, 2014. Light manipulation for organic optoelectronics using bio-inspired moth's eye nanostructures. *Scientific Reports* 4, 8.
- Zhou, Y., Zhang, F., Tvingstedt, K., Tian, W., Inganäs, O., 2008. Multifolded polymer solar cells on flexible substrates. *Applied Physics Letters* 93 (3), 033302.
- Zhu, J., Xue, M., Hoekstra, R., *et al.*, 2012. Light concentration and redistribution in polymer solar cells by plasmonic nanoparticles. *Nanoscale* 4, 1978–1981.

Organic Lasers

Graham A Turnbull, University of St Andrews, St Andrews, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

Charge mobility [$\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$]

Current density [A cm^{-2}]

Energy density [$\mu\text{J cm}^{-2}$]

Excited state densities [cm^{-3}]

Gain coefficient α ($\exp(\alpha z)$) [cm^{-1}]

Intensity/power density [kW cm^{-2}]

Optical gain [dB cm^{-1}]

Stimulated emission cross section [cm^2]

Waveguide propagation loss coefficient

γ ($\exp(-\gamma z)$) [cm^{-1}]

Wavelength [nm]

Conjugated Polymers

Plastics are widely used in low-cost optical components such as lenses, diffraction gratings, optical fibers, and adhesives. Their common use derives from a combination of qualities: good optical transmission, a wide range of mechanical properties and simple fabrication via molding, extrusion, or solution processing. Functional optical polymers, including liquid crystals and photorefractives, have also been widely applied to polarization control and for nonlinear- and electro-optics. Conjugated polymers are a remarkable class of functional plastics that exhibit some quite unusual properties. Firstly, they are plastics that can conduct electricity; and in fact are electrically semiconducting. Secondly, they are plastics that can be stimulated either optically or electrically to emit visible light.

These two unusual properties derive from a particular chemical structure specific to this family of plastics. In all conjugated polymers, the backbone of the molecule is a chain of carbon atoms joined by alternating single and double covalent bonds. The simplest example is illustrated in Fig. 1(a), which shows the chemical structure of the polymer poly(acetylene). Structural integrity of the polymer is provided by a string of single σ -bonds between adjacent carbon atoms. The additional bonding electrons in the double bonds are located in π -orbitals oriented perpendicular to the polymer chain. Adjacent π -orbitals overlap, and their electrons can be widely delocalized along the polymer backbone. The delocalized electrons can move relatively freely along the chain, giving the electronic properties of the material. The π -orbitals can be transformed between bonding and anti-bonding states by the absorption or emission of light, leading to strong optical transitions. While poly(acetylene) is a very inefficient light emitter, other conjugated polymers, notably those that include phenyl rings in the backbone (Fig. 1(b)–(d)), can be highly emissive. Photoluminescence quantum efficiencies approaching unity in solution and as high as 60% in the semiconducting solid state have been realised.

Organic Semiconductor Lasers

While organic dye lasers have been widely used since the mid-1960s, the history of organic semiconductor lasers is much more recent. Other than a few early observations of lasing in molecular crystals in the 1970s by Karl and others, the field of organic

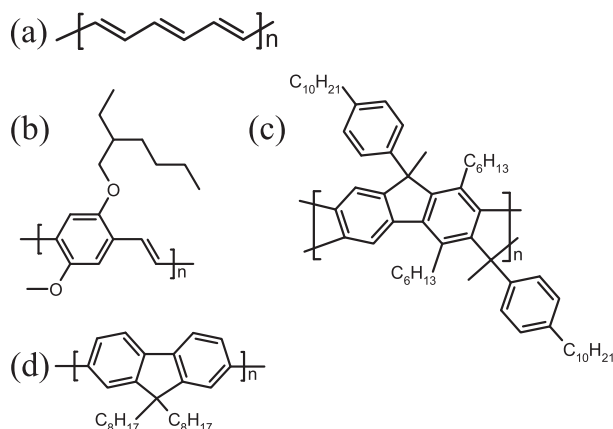


Fig. 1 Chemical structures of several conjugated polymers: (a) poly[acetylene]; (b) poly[2-methoxy-5-(2'-ethylhexyloxy)-1,4-phenylene vinylene] (MEH-PPV); (c) ladder-type poly[1,4-phenylene] (MeLPPP); (d) poly[9,9-dioctylfluorene] (PFO).

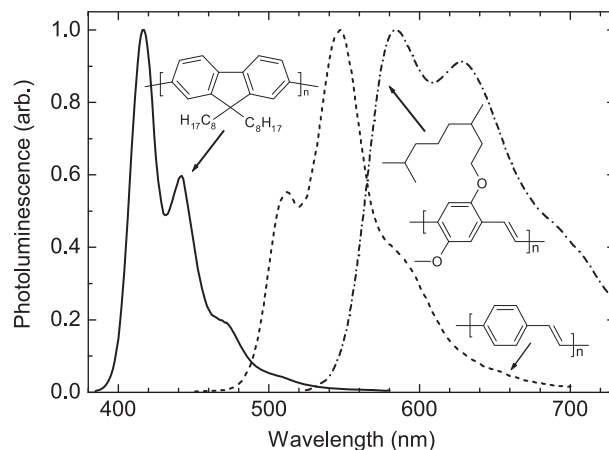


Fig. 2 Photoluminescence spectra of three common conjugated polymers. With chemical structures inset.

semiconductor lasers began in earnest following the related discovery in 1990 by Burroughes *et al.* of electroluminescence in a conjugated polymer. In 1992, Moses demonstrated lasing for the first time in a solution of the conjugated polymer poly[2-methoxy-5-(2'-ethylhexyloxy)-1,4-phenylene vinylene] (MEH-PPV) (**Fig. 1(b)**). However, it was another 4 years – once material synthesis had been sufficiently refined – before stimulated emission and then true laser action were observed in polymers in the semiconducting solid state. In the following year the first demonstrations were made of dye-doped small molecular semiconductor lasers. Since that time, there has been a great growth in interest in the field with an average of ~ 30 research papers per year. Further details of the development of the field of organic semiconductor lasers can be found in the bibliography.

So, what are the particular properties of conjugated polymers that have stimulated interest in developing organic semiconductor lasers? Firstly these materials are broadband visible emitters. Through appropriate chemical design, their emission can be tuned throughout the visible spectrum, from 400 to 700 nm (**Fig. 2**). The polymers exhibit large optical gain coefficients, because the optical transitions are dipole-allowed. Compared with small molecular dyes, polymers exhibit a relatively low degree of self-absorption of the emitted light, and much reduced quenching of the luminescence with increased concentration. This means that it is possible to work with much higher chromophore densities than conventional laser dyes, which is advantageous for achieving very low threshold lasing in extremely compact resonators. Furthermore, because these materials are semiconducting, they have the potential for direct electrical excitation. This could lead to semiconductor diode lasers operating throughout the visible spectrum and, in particular, in the blue and green spectral bands where currently there are few commercially-viable inorganic semiconductor sources. Finally, these materials are plastics and so are amenable to simple processing, and can be shaped and structured into complex feedback resonators.

Organic Semiconductor Gain Materials

A great advantage of organic laser media is that a very wide range of materials can be readily synthesized, and through minor changes in the molecular structure one can relatively simply tune the emission properties of the material. Many different organic semiconductors have been applied as lasers, though they fall into three major categories. These are conjugated polymers, dye-doped small molecular semiconductors and molecular single crystals. Conjugated polymers make up the most diverse materials set, most notably including derivatives of poly(paraphenylene-vinylene), poly(paraphenylene) and poly(fluorene) (**Fig. 1(b)–(d)**). The second category comprises small molecular semiconductors that are co-evaporated under vacuum with a dopant laser dye. In this blend of materials, the organic semiconductor functions simply as an energy transfer host for the emitting dye molecules. In the third category are organic single crystals, including materials, such as polyacenes and thiophene oligomers. These materials have the advantage of having the highest electrical mobilities of any organic semiconductors, which is important for electrical excitation. However, they can be demanding to grow into high quality single crystals, and so lack the key processing advantages of the other families of organic semiconductors.

Fig. 3(a) shows a typical energy level structure of a semiconducting polymer. Similar to a conventional laser dye, the energy levels comprise a manifold of singlet states and a manifold of triplet states, each with vibrational sublevels. The laser transition occurs from the bottom of the first excited singlet state S_1 to the vibronic overtones of the ground state S_0 , so forming a classic 4-level laser system. The optical transition is dipole-allowed and so has a high oscillator strength. As a result conjugated polymers typically exhibit simulated emission cross sections in the range 10^{-16} – 10^{-15} cm², while the population inversion lifetime is as short as a few hundred picoseconds.

There are, however, several other transitions that can compete with the stimulated emission process. There can be excited-state absorption from the S_1 singlet state into higher lying singlet states. Additionally intersystem crossing from the S_1 state to the lowest excited triplet state T_1 can subsequently lead to excited triplet absorption. Other processes in the polymer can also impede the

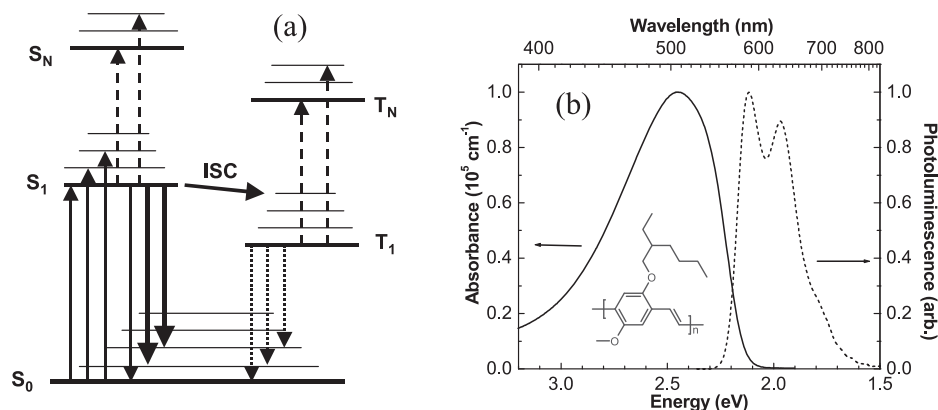


Fig. 3 (a) Generic energy level structure and optical transitions of a conjugated polymer. S_N refers to the N th singlet state, T_N refers to the N th triplet state; (b) absorption and photoluminescence spectra of the polymer poly[2-methoxy-5-(2'-ethylhexyloxy)-1,4-phenylene vinylene] (MEH-PPV). ISC, intersystem crossing.

stimulated emission. Firstly, charged excitations can exhibit absorption bands that overlap with the emission. Secondly, at high excitation densities, exciton-exciton annihilation can significantly deplete the excited states available for stimulated emission. Ultimately, singlet-singlet annihilation is the limiting factor for the maximum excitation density possible, and while polymer films can have a repeat-unit density of $\sim 10^{21} \text{ cm}^{-3}$ the excited state density is typically limited to around 10^{18} cm^{-3} . In many materials this is not restrictive on achieving optical gain because stimulated emission can occur at densities of 10^{17} cm^{-3} and below.

An important factor that distinguishes the photo-physics of semiconducting polymers from conventional laser dyes is an internal process of exciton migration between the absorption and emission of a photon. Excitations are confined to short straight sections of the polymer between kinks or twists in the chain. These straight sections, or "sites," are typically of a few repeat units in length, though there is a statistical spread of site lengths in any given material. When formed, excitons tend to hop to the longest, lowest energy, sites before emitting a photon. The exciton migration process has two effects. Firstly it means the emission takes place from a subset of the chain segments, resulting in a lower degree of inhomogeneous broadening of the emission spectrum (vibronic peaks are better resolved than in absorption – see Fig. 3(b)). Secondly the peaks of the absorption and emission spectra tend to be widely separated in wavelength, reducing the level of self-absorption at the lasing wavelength.

As mentioned before, blending materials is a particularly attractive approach toward further reducing self-absorption in organic semiconductor lasers. In such systems, a narrow band-gap dopant chromophore is blended with a concentration of 1–2% in a wider band-gap semiconducting host. Emission from the host is strongly quenched via nonradiative Förster energy transfer to the guest molecule. The blended material exhibits an absorption profile that is dominated by the host absorption, while the emission profile is almost completely that of the guest dopant. By strongly separating the emission peak from the absorption edge of the host it is possible to achieve a substantially reduced (by perhaps an order of magnitude or more) residual absorption coefficient at the lasing wavelength. In doing so, the combined material acts much more like an ideal 4-level laser medium and can thereby have substantially lower pump thresholds for stimulated emission.

Measuring Gain

The dynamics of the excited state population in semiconducting polymers have been widely studied using femtosecond transient absorption experiments. This is a pump-probe technique in which the transmission of the broadband probe pulse is compared in the presence and absence of a pump pulse that excites molecules into the first excited state. Such measurements have the time resolution of the pump pulses used and so can typically be ~ 100 fs. By systematically delaying the arrival of the broadband probe pulse it is possible to map out the evolution of the gain dynamics across the emission spectrum of the material with very high time resolution.

Fig. 4 shows a series of snapshots of the absorption induced by a pumping pulse in a poly(thiophene) derivative. This figure shows the changing absorption of the sample between 200 fs and 400 ps after excitation. Within the absorption band of the material, one finds that the level of absorption is slightly reduced because the pump pulse has partly depleted the ground state. In the emission band there is also a reduction in the absorption that mimics in shape the photoluminescence spectrum of the polymer. This negative absorption- or optical-gain shows direct evidence for stimulated emission through much of the emission band. In the low energy tail of the emission, however, there is an increase in absorption that arises due to the excited-state absorption process mentioned earlier. Clearly, for a good laser material, the spectra of excited-state and charge-induced absorptions should overlap as little as possible with the photoluminescence.

Many other investigations of gain in semiconducting polymer films have used a phenomenon known as spectral line narrowing to explore indirectly the amplifying properties of the medium. The experimental procedure for such studies is as follows. A thin

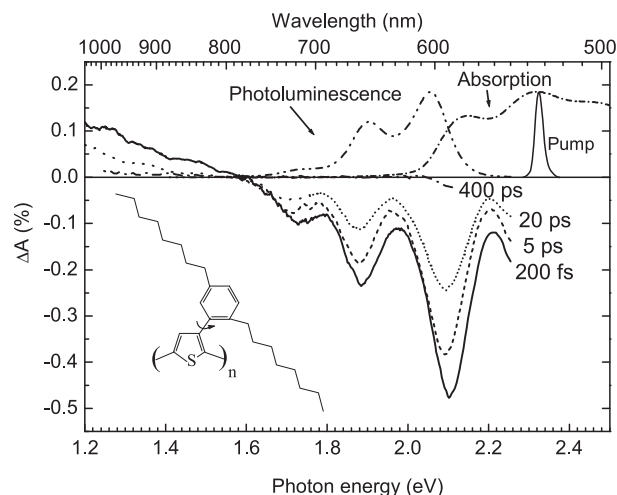


Fig. 4 Transient absorption spectra of poly[3-(2,5-dioctyl-phenyl)-thiophene]; chemical structure inset. Reproduced from Ruseckas, A., Theander, M., Valkunas, L., *et al.* 1998. Energy transfer in a conjugated polymer with reduced inter-chain coupling. *Journal of Luminescence* 76–77, 474–477. Copyright (1998), with permission from Elsevier.

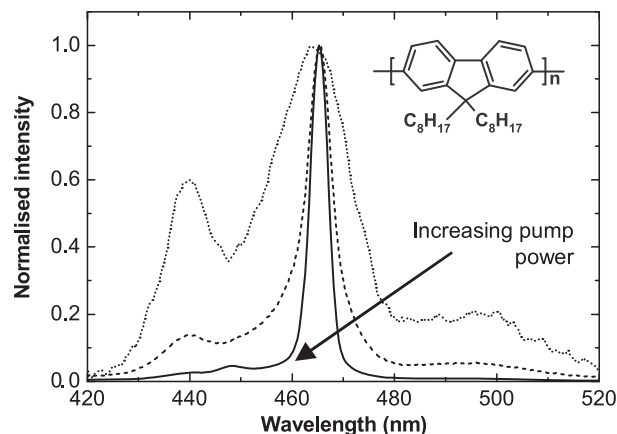


Fig. 5 The change in spectral shape with increasing excitation density (from 10 to 133 $\mu\text{J cm}^{-2}$) of the edge emission from a PFO film – the normalized spectra show a significant spectral line narrowing at high excitation densities. Adapted from Heliotis, G., Bradley, D.D.C., Turnbull, G.A., Samuel, I.D.W., 2002. Light amplification and gain in polyfluorene waveguides. *Applied Physics Letters* 81, 415–417. Copyright (2002), with permission from American Institute of Physics.

film of the semiconducting material, of typically 100 nm thickness, is excited by a pulsed pump laser focused to a stripe of dimensions $\sim 100 \mu\text{m}$ wide by a few millimeters in length. A significant fraction of the light emitted by the material is trapped in the film by total internal reflection at the polymer–air and polymer–substrate interfaces, and is thereby waveguided along the length of the excitation stripe. This waveguided spontaneous emission can be amplified by stimulated emission before being emitted out the edge of the film.

One finds that, above a particular threshold of pump energy, there is a superlinear increase in the output energy with pump intensity. Associated with this is a dramatic narrowing of the broad emission spectrum to a linewidth of typically less than 10 nm, as illustrated in **Fig. 5**. These two features are signatures of gain narrowing via the process of amplified spontaneous emission. The line-narrowed feature usually appears at the first vibronic overtone, of the optical transition, where the gain exceeds the losses in the material by the greatest amount. This particular wavelength experiences the greatest amplification and, above a particular pumping density, undergoes a runaway effect in which most of the light is stimulated to emit at that wavelength at the expense of the rest of the spectrum.

In the case of amplified spontaneous emission or mirrorless lasing as it is also known, the spontaneous emission guided along the length of the stripe acts as the “probe pulse.” In these experiments it is therefore necessary to extract information indirectly about the net gain in the material. By integrating the spontaneous and stimulated emission along the length of the stripe one finds that the wavelength-dependent output intensity $I(\lambda)$ of the spectral line narrowed light is given by the relationship

$$I(\lambda) \propto \frac{1}{g(\lambda)} [\exp[g(\lambda)l] - 1]$$

where I_p is pumping intensity, l is the length of the stripe, and $g(\lambda)$ is the net gain coefficient. Therefore, by monitoring the intensity of the line-narrowed emission as a function of the length of the pump stripe, one can extract the value of the net gain for the material. Similarly, by progressively moving the stripe further and further away from the edge of the film, it is possible to evaluate the waveguide losses of light propagating through an unpumped region of the waveguide. Waveguide losses in conjugated polymers typically lie in the range of 3–50 cm⁻¹. Net gains can be as high as 60 cm⁻¹ or 260 dB cm⁻¹ at modest pumping densities of 4 kW cm⁻². These substantial gains in very small propagation lengths therefore highlight the attraction of waveguide-based resonators for polymer lasers. In the next section we will discuss the most commonly used feedback structures that have been used for polymer lasers.

Polymer Laser Resonators

The first demonstration of cavity lasing in a conjugated polymer was made by Moses in 1992. The resonator consisted simply of Fresnel reflections from the walls of the glass cuvette containing the polymer in solution. Lasing occurred at the peak wavelength of the gain profile. Subsequent developments in solid-state polymer lasers have involved more sophisticated resonators that exploit the capacity of these materials for very simple processing. For example, solution-processing methods such as spin-casting or dip-coating, make it possible to fabricate very high quality optical waveguides. Wavelength-scale microstructures can also be readily formed allowing the straightforward fabrication of laser microresonators.

Many different resonators have been explored for solid-state polymer lasers. Some of the key types are shown in Fig. 6, including different geometries of both microcavities and distributed feedback (DFB) structures. The first resonator used, by Tessler *et al.*, was a planar microcavity (Fig. 6(a)) in which the polymer film was sandwiched between two mirrors. The structure supported standing-wave resonances perpendicular to the plane of the film, at three wavelengths in the emission band of the polymer. Above laser threshold, light was preferentially stimulated into only one of these modes. Microcavities of other geometries have also been studied, including annular and spherical microresonators (Fig. 6(b) and (c)) in which the optical field is confined in waveguide modes, whispering gallery modes, or a superposition of the two. Such structures provide a longer interaction length in the gain medium for stimulated emission; though tend to lase on many longitudinal modes and have no directional output beam.

The final key group of resonators are DFB structures, which researchers have progressively favoured to attain the low pump threshold densities that will be required to realize electrically pumped lasing. Fig. 6(d) shows a typical DFB laser structure. The device is based on a thin polymer film that is typically 100 nm in thickness deposited on top of a silica substrate. The silica–polymer–air structure forms an asymmetric slab waveguide that supports only the lowest order transverse electric mode TE₀. To provide the feedback, the substrate is modulated with a wavelength-scale periodic grating structure. These corrugations give rise to Bragg scattering of the waveguide mode. Through careful selection of the grating period one may arrange that optical wavelengths close to the peak of the polymer gain will be Bragg scattered from the propagating TE₀ mode to the counter-propagating TE₀ mode. This is achieved when Bragg equation

$$2n_{\text{eff}}\Lambda = m\lambda_B,$$

is satisfied, where Λ is the grating period, n_{eff} is the effective refractive index of the waveguide, m is an integer and λ_B is Bragg wavelength. In so doing one may arrange that two counter-propagating waveguide modes can be coupled together thereby providing the distributed resonant feedback for the laser process. Commonly in semiconducting polymer lasers, a second order (i.e., $m=2$) Bragg grating has been favoured to provide the feedback for the laser, so that the first order scattering provides a surface-emitted output coupling, perpendicular to the plane of the guide. This surface-emitting geometry is particularly attractive in polymer lasers because it is relatively difficult to achieve a high optical quality edge to the waveguide. DFB lasers have the

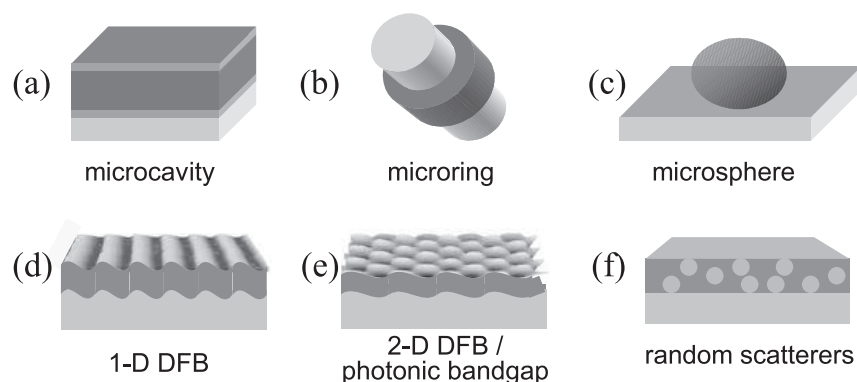


Fig. 6 Resonators used for conjugated polymer lasers. (a) Planar microcavity; (b) annular microcavity; (c) spherical microcavity; (d) distributed feedback (DFB) resonator; (e) 2-D DFB/photonic crystal resonator; (f) random laser.

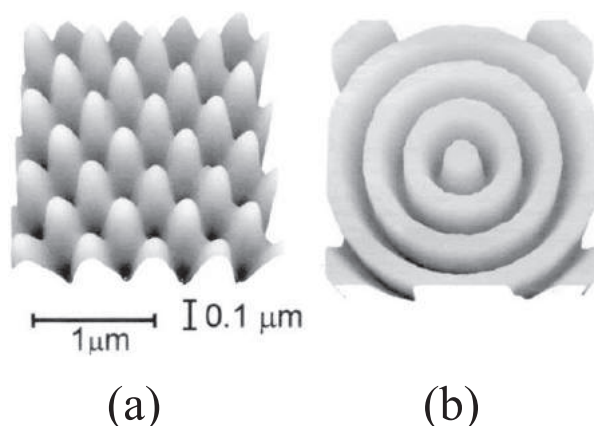


Fig. 7 Atomic force microscope images of (a) “egg-box” distributed feedback (DFB) grating and (b) circular DFB grating; grating period is 400 nm in each structure.

advantage over simple microcavity resonators in that they can combine long interaction lengths, limited by the scattering length of Bragg grating, with excellent wavelength selection, which is controlled by the grating period.

In addition to conventional 1-D DFB gratings, a number of more exotic 2-D and 3-D DFB, and photonic crystal resonators have been explored in semiconducting polymers. Even randomly located scattering sites in a polymer film have been used to provide a weak closed-loop cavity feedback. Examples of 2-D gratings are given in **Fig. 7**, which shows atomic force micrographs of an eggbox structure, formed from two orthogonal gratings, and a circular DFB grating. Such 2-D DFB structures can significantly improve the operation of surface-emitting polymer lasers compared with 1-D gratings. Improved feedback yields lower oscillation thresholds, combined with higher slope efficiencies and significantly improved beam quality. Indeed such surface emitting DFB lasers can produce nearly diffraction limited laser beams.

Typical operating characteristics of an egg-box DFB polymer laser are shown in **Fig. 8**. This laser, which is based on the material MEH-PPV, is transversely pumped with a pulsed pump laser focussed onto a small region of the waveguide; up to 90% of the incident pump light is absorbed in the 100 nm thick film. The laser has a threshold pump energy of 4 nJ and an external slope efficiency of $> 7\%$, and operates on a single frequency close to Bragg wavelength of the periodic waveguide. Lasing does not occur exactly at Bragg wavelength because at this wavelength the propagation of light is suppressed by the periodic structure. This leads to a photonic stopband within which the emission is reduced. At the edges of this stopband the microstructure can support standing wave electric field patterns of the same spatial period. At these wavelengths there can be a strong interaction between the electric field and excited states in the gain medium, which leads to low threshold band edge lasing as observed in the figure.

Although DFB resonators have advantages over microcavity devices, the complicated lithographic and etching steps involved in their fabrication can detract from the key material advantage of simple processing. To address this issue several different techniques for simply replicating the complicated microstructures have been explored. These include UV embossing of flexible plastic substrates, and hot embossing and micromoulding of the active polymer itself. Such techniques can very readily reproduce structures with feature sizes as small as 100 nm and could allow future mass production of photonic microstructures in these materials.

Towards Plastic Diode Lasers

A typical experimental configuration for characterising optically pumped polymer lasers is shown in **Fig. 9**. The laser is mounted in a vacuum chamber in order to isolate the polymer from oxygen and water. A pulsed visible or ultraviolet pump laser is attenuated and then focussed onto the waveguide to transversely pump the polymer laser. The emission is collected using a charge-coupled device (CCD) spectrograph for spectral measurements and energy meter and beam profilers for other characteristics. Typical pulsed pump lasers which have been used include nitrogen laser pumped dye lasers, and spectral harmonics of regeneratively amplified Ti^{3+} -sapphire lasers or flashlamp pumped Nd^{3+} :YAG lasers. Recent advances have shown the thresholds to be low enough to be successfully pumped with a modest Nd^{3+} pulsed microchip laser.

So while optically pumped polymer lasers are now proving to be compact, functional light sources, a major motivation for research in this field remains the prospect of electrically driven plastic diode lasers. Such lasers could be very low cost, flexible light sources that would access any wavelength throughout the visible spectrum. In particular, there are exciting prospects of producing blue and green diode lasers in spectral regions currently difficult to access with inorganic semiconductors. The lowest threshold densities reported for optically pumped organic semiconductor lasers are of the order of 100 W cm^{-2} . To achieve similar excitation densities electrically, one requires pump current densities of $\sim 200 \text{ A cm}^{-2}$. This is substantially more than the few milliampere per square centimeter commonly used in polymer light-emitting diodes (LEDs) for display applications. Presently, such high current densities are infeasible with continuous excitation, but can be exceeded with short current pulses of a few tens of nanoseconds duration.

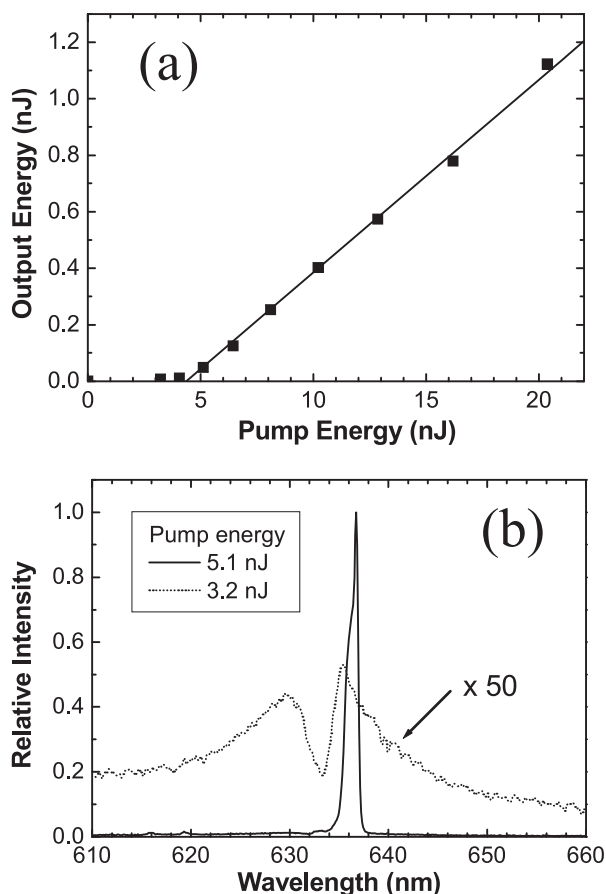


Fig. 8 (a) Energy characteristic of an egg-box distributed feedback (DFB) laser; (b) normalized emission spectra above and below lasing threshold, showing lasing in the edge of the photonic stopband. Reprinted from Turnbull, G.A., Andrew, P., Barnes, W.L., Samuel, I.D.W., 2002. Operating characteristics of a semiconducting polymer laser pumped by a microchip laser. *Applied Physics Letters* 82, 313–315, Copyright (2003), with permission from American Institute of Physics.

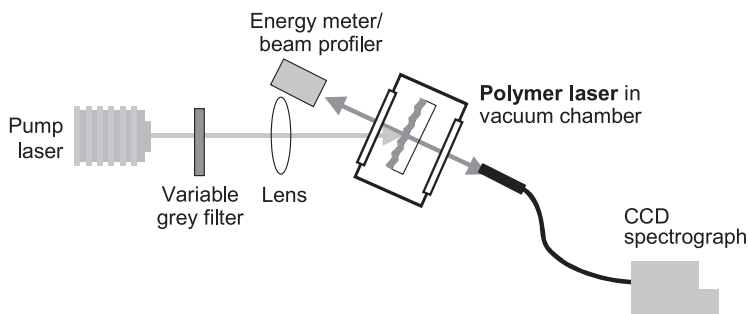


Fig. 9 Schematic of a typical experimental configuration for characterising polymer lasers. CCD, charge-coupled device.

Electroluminescence in conjugated polymers was first achieved in 1990 by Burroughes and coworkers. Since that time there has been remarkable progress made in the science and technology of organic LEDs. Commercial products are now on the market, and the feasibility of both flexible and full color light emitting displays, in which the individual pixels have been ink-jet printed, have been demonstrated in the laboratory. Despite such rapid progress in the field of organic LEDs, there have been no successful demonstrations to date of electrically pumped lasing in any organic semiconductor. In this section we highlight some of the challenges that must be faced if electrically pumped lasing is to be achieved. To understand the problems associated with electrical driving of organic semiconductor lasers it is useful first to review the generic structure of organic light emitting diodes.

A generic organic LED structure is shown in **Fig. 10**. The device consists of one or more organic layers sandwiched between two electrical contacts. Usually a transparent anode is used, commonly made from indium tin oxide (ITO), which allows light to be emitted from the front face of the LED. A low work-function metal such as calcium or aluminum is used as the cathode. The

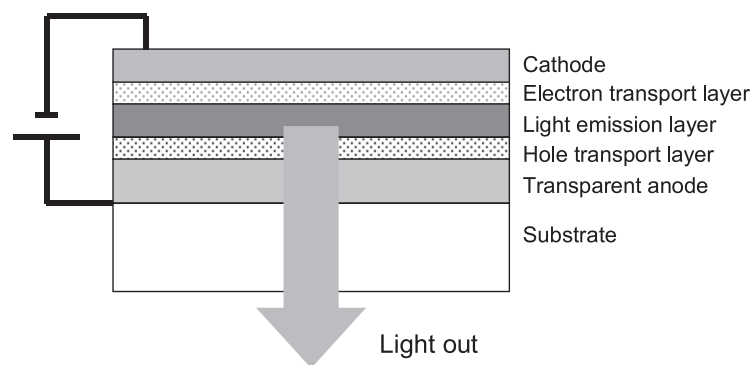


Fig. 10 Schematic structure of a generic organic light emitting diode.

organic part of the device may consist of several layers. These include one or more charge transport layers, plus a light-emitting layer in which the electrons and holes combine to form excitons before emitting light. Typically, the combined thickness of these organic layers is as little as 100 nm. Such thin devices are necessary to keep operating voltages low, to overcome potential barriers for injection and because charge mobilities in organic semiconductors are very low (10^{-5} – 10^{-2} cm² V⁻¹ s⁻¹).

The inclusion of electrical contacts into organic semiconductor laser structures can be problematic since they need to be situated very close to the emitting region. Metallic contacts can quench the luminescence and introduce losses to the waveguide mode, thereby increasing the oscillation thresholds. Furthermore, the ITO anode has a high refractive index and so can reduce the spatial overlap of the guided mode with the active medium. Such issues indicate that the design of organic LEDs needs to be somewhat modified for achieving electrically pumped lasing. In particular, there needs to be a trade-off between the optimum electronic and photonic designs. To address these issues various device structures have been proposed that distance the optical mode from metallic contacts, while reducing the thickness of the ITO layer to optimize the mode confinement in the active material.

A more demanding obstacle to electrically pumped lasing is the additional excited state absorption associated with the electrical pumping scheme. As many as three quarters of the generated excitons in an organic LED are formed in the triplet state which is non-emissive and can lead to excited-state absorption, albeit principally at lower energies than the lasing wavelength. Additionally, injected charges that have not yet formed excitons exhibit strong absorption, which overlaps significantly with the peak of the material gain. Indeed, this charge-induced absorption has so far prevented the demonstration of injection lasing in conjugated polymers, even in optimized device structures mentioned above. A possible alternative route for efficient pumping of polymer lasers could be through indirect electrical pumping. For example, an inorganic diode laser or LED, or even an organic LED, could be used as an electrically efficient optical pump for the semiconducting polymer laser. Recent advances in GaN-based LEDs and lasers provide interesting prospects for such alternative systems.

Research on semiconducting polymer lasers keenly continues, both to reduce the excitation densities needed for threshold and to investigate new material systems that could overcome the obstacle of charge-induced absorption. By building on recent developments in resonator geometries, material blends and compact pumping schemes, semiconducting polymer laser systems, whether optically or electrically pumped, should soon become practical, visible light sources for a range of applications.

See also: Organic Semiconductors

Further Reading

- Burroughes, J.H., Bradley, D.D.C., Brown, A.R., *et al.*, 1990. Light-emitting-diodes based on conjugated polymers. *Nature* 347, 539–541.
- Friend, R.H., Gymer, R.W., Holmes, A.B., *et al.*, 1999. Electroluminescence in conjugated polymers. *Nature* 397, 121–128.
- Karl, N., 1972. Laser emission from an organic molecular crystal. *Physica Status Solidi* 13, 651–655.
- Kranzelbinder, G., Leising, G., 2000. Organic solid-state lasers. *Reports on Progress in Physics* 63, 729–762.
- Kuehne, A.J.C., Gather, M.C., 2016. Organic lasers: Recent developments on materials, device geometries, and fabrication techniques. *Chemical Reviews* 116, 12823–12864.
- McGehee, M.D., Heeger, A.J., 2000. Semiconducting (conjugated) polymers as materials for solid-state lasers. *Advanced Materials* 12, 1655–1668.
- Miyata, S., Nalwa, H.S. (Eds.), 1997. *Organic Electroluminescent Materials and Devices*. Amsterdam: OPA.
- Moses, D., 1992. High quantum efficiency luminescence from a conducting polymer in solution – A novel polymer laser-dye. *Applied Physics Letters* 60, 3215–3216.
- Pope, M., Swenberg, C.E., 1999. *Electronic Processes in Organic Crystals and Polymers*. New York: Oxford University Press.
- Samuel, I.D.W., 2000. Polymer electronics. *Philosophical Transactions of the Royal Society of London A* 358, 193–210.
- Scherf, U., Riechel, S., Lemmer, U., Mahrt, R.F., 2001. Conjugated polymers: Lasing and stimulated emission. *Current Opinion in Solid State and Materials Science* 5, 143–154.
- Tessler, N., 1999. Lasers based on semiconducting organic materials. *Advanced Materials* 11, 363–370.
- Tessler, N., Denton, G.J., Friend, R.H., 1996. Lasing from conjugated-polymer microcavities. *Nature* 382, 695–697.
- Zhao, Z., Mhibik, O., Nafa, M., *et al.*, 2015. High brightness diode-pumped organic solid-state laser. *Applied Physics Letters* 106, 051112.

Organic Photodetectors

Jaehoon Jeong, Hwajeong Kim, and Youngkyoo Kim, Kyungpook National University, Daegu, Republic of Korea

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic photodetector is an optoelectronic device that can sense light with organic semiconducting materials, which typically feature pi-conjugated molecular structures, and convert it to electrical signals. The sensing range of organic photodetectors is basically dependent on the optical absorption spectra of organic semiconducting materials. Various organic semiconducting materials including small molecules and polymers can be used as a light-sensing layer in organic photodetectors.

Two types of device structures, diode and transistor, are generally employed for the fabrication of organic photodetectors. It is noted that transistors here mean “field-effect transistors.” The diode structure leading to organic photodiodes has organic sensing layers (OSLs) between two electrodes, while the transistor structure leading to organic phototransistors includes additional electrode for gate control in the presence of organic layers and two counter electrodes. Therefore, organic photodiodes are in principle much simpler than organic phototransistors in terms of device structure and fabrication process. In contrast to organic photodiodes, however, organic phototransistors benefit signal amplifications by adjusting the gate voltage and enable active-matrix sensor arrays for two-dimensional images.

Organic photodetectors, both organic photodiodes and organic phototransistors, can be fabricated using wet-coating processes such as ink-jet printing, screen printing, gravure printing, spray coating, slot-die coating, etc. at low temperatures. In addition, flexible plastic films, such as poly(ethylene naphthalate) (PEN) and poly(ethylene terephthalate) (PET), can be applied as a substrate for the fabrication of organic photodetectors. These benefits enable organic photodetectors to be manufactured with continuous roll-to-roll processes leading to low-cost advantages.

Device Structure and Mechanism

Organic Photodiodes

Organic photodiodes consist of two electrodes and OSLs between the two electrodes (Fig. 1(a)). As shown in Fig. 1(b), the charges (holes and electrons) generated by incoming photons transport to corresponding electrodes due to the built-in potential made at short circuit condition. The OSLs can be composed of single materials, either p-type or n-type organic semiconductors,

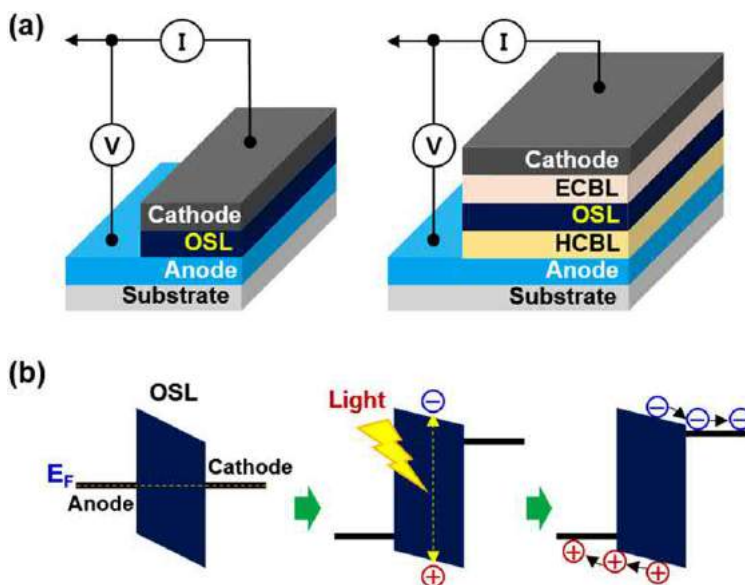


Fig. 1 (a) Typical device structures for organic photodiodes consisting of (left) single organic sensing layer (OSL) between anode and cathode and (right) with additional hole-collecting buffer layer (HCBL) and electron-collecting buffer layer (ECBL). (b) Simplified energy band diagrams for the sensing mechanism in organic photodiodes with single OSL (refer to the left structure in (a)).

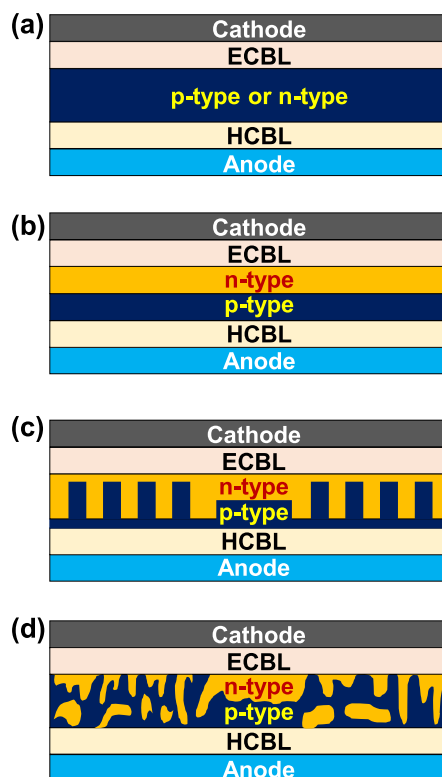


Fig. 2 Illustration for the cross-sections of organic photodiodes according to the types of organic sensing layers (OSLs): (a) single sensing layer, (b) bilayer, (c) interdigitated bilayer, and (d) bulk heterojunction layer. ECBL, electron-collecting buffer layer; HCBL, hole-collecting buffer layer.

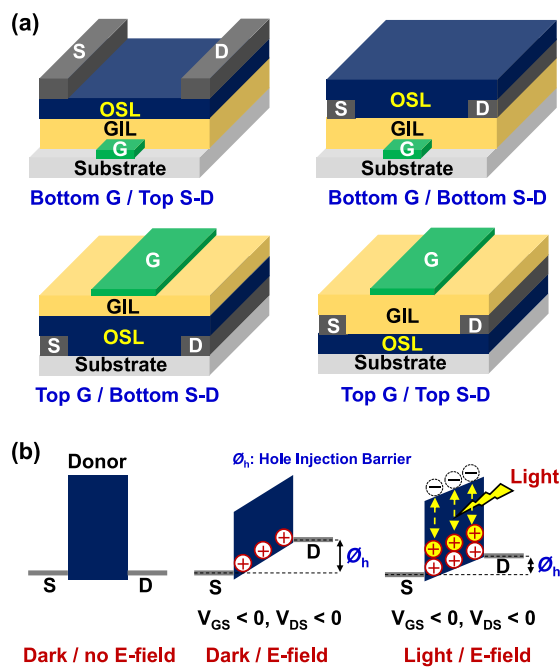


Fig. 3 (a) Representative geometry of organic phototransistors with organic sensing layer (OSL) and gate-insulating layer (GIL): (top-left) bottom gate/top source-drain, (top-right) bottom gate/bottom source-drain, (bottom-left) top gate/bottom source-drain, (bottom-right) top gate/top source-drain. (b) Energy band diagram for organic phototransistors (p-type): (left) no electric field condition in the dark, (middle) under a field-effect condition in the dark (ϕ_h : hole injection barrier), (right) light illumination under a field-effect condition. V_{GS} and V_{DS} denote gate voltage and drain voltage, respectively.

as shown in Fig. 2(a). However, p–n junction structures are required in order to enhance the photosensitivity of devices. The simple bilayer structure can be easily prepared by employing multilayer coatings even though orthogonal solvents are essential for preparing solutions for each layer in the case of wet-coating processes (Fig. 2(b)). Considering the contact area between p-type and n-type organic semiconducting layers, the simple bilayer structure has an intrinsic limitation in the number of p–n junctions. As shown in Fig. 2(c), an interdigitated bilayer structure is expected to provide ideal p–n junctions. Unfortunately, however, the interdigitated bilayer structure on a scale of a few nanometers cannot be easily fabricated with current coating technology. Alternatively, bulk heterojunction structures are considered to deliver high photosensitivity because numerous p–n junctions are made intimately in the mixture (blend) of p-type and n-type organic semiconducting materials (Fig. 2(d)).

Organic Phototransistors

Organic phototransistors are different from organic photodiodes in the point that gate electrodes are required for the control of devices (Fig. 3(a)). Four types of organic phototransistors can be fabricated according to the position of electrodes. In general, the source and drain electrodes exist in parallel with the film plane of organic sensing channel layers. As shown in Fig. 3(b), very low charge transport (current flow) in the dark are measured when voltages (biases) are applied between source and drain electrodes in the absence of applying voltages between source and gate electrodes. Applying voltages to the gate electrodes creates field-effect charges in organic channel layers, which contributes to the increased charge transport between source and drain electrodes. Here, when light shines to the channel part of organic sensing channel layers, photons generate excitons that in turn dissociate into individual charges (electrons and holes). Therefore, the photo-generated charges are added to the field-effect charges, which are measured at the same time.

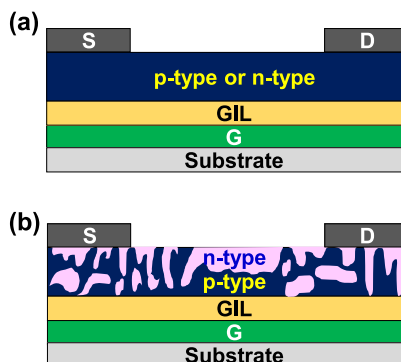


Fig. 4 Illustration for the cross-sections of organic phototransistors: (a) single sensing layer (either p-type or n-type organic semiconductors) and (b) bulk heterojunction layer (mixture of p-type and n-type organic semiconductors). GIL, gate-insulating layer.

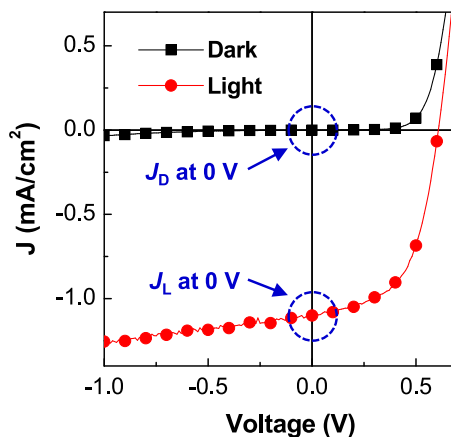


Fig. 5 Current density–voltage (J – V) curves for organic photodiodes in the dark and under light. J_D and J_L represent short circuit current density (0 V) in the dark and under light, respectively.

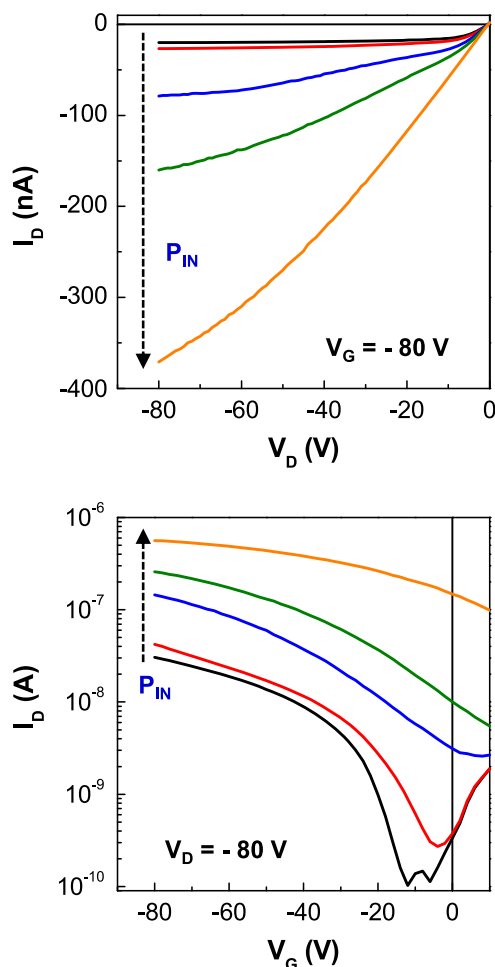


Fig. 6 Output (top) and transfer (bottom) curves for representative organic phototransistors with p-type organic sensing layers (OSLs) according to the incident light intensity (P_{IN}).

In terms of organic sensing channel layers, a variety of material combinations are possible but a single layer structure is basically applied as used for typical organic transistors. The single sensing channel layer is typically composed of one kind of organic semiconducting materials (either p-type or n-type). However, as explained for organic photodiodes, bulk heterojunction-type sensing channel layers can deliver enhanced photosensitivity because of the benefit of bulk heterojunctions leading to efficient charge separation. In addition, the bulk heterojunction-type sensing channel layers can take further advantages in tuning the sensing wavelengths by controlling the composition of p-type and n-type organic semiconducting materials with different absorption spectral ranges (Fig. 4).

Photosensitivity (Responsivity) of Organic Photodetectors

The photosensitivity (or responsivity, R) of organic photodetectors is defined by the ratio of net photocurrent density (J_P) to incident light intensity (P_{IN}), as given by Eq. (1). The net photocurrent density is obtained by subtracting the dark current density (J_D) from the whole current density under light (J_L) (see Fig. 5). It is noted that the unit of current density and incident light intensity is typically expressed as milliamperere per square centimeter and watt per square centimeter, respectively. Thus the unit of responsivity becomes A/W.

$$R = \frac{J_P}{P_{IN}} [\text{A/W}] \quad (1)$$

In addition, signal-to-noise (S/N) ratios are frequently used to understand the effect of background noises made in photodetectors. The S/N ratio is defined by the ratio of photocurrent density to dark current density. Further detailed approach is to

employ specific detectivity (D^*) that represents a figure of merit for photodetectors. As shown in Eq. (2), the specific detectivity is expressed by the ratio of responsivity (R and active area (A)) to noise spectral density (S_n).

$$D^* = \frac{R \cdot \sqrt{A}}{S_n} \text{ [Jones]} \quad (2)$$

The responsivity of organic photodiodes can be simply calculated using Eq. (1) from current density–voltage (J – V) curves as illustrated in Fig. 5. However, organic phototransistors provide two types of current–voltage curves, output and transfer curves, as shown in Fig. 6. Output curves are useful to get responsivity depending on drain voltage at a fixed gate voltage, while transfer curves can be employed to calculate the gate voltage-dependent responsivity at a fixed drain voltage. A particular attention needs to be paid for the measurement of photocurrent because the active area is too small to block with a shadow mask if the incident light comes directly to the sensing channel layers but the source/drain electrodes are unable to block the incident light owing to their positions.

Organic Photodiodes With Small Molecules

Organic Layer – Single Material

Various types of small molecules have been used for the fabrication of organic photodiodes with OSLs composed of single material. As shown in Fig. 7, pentacene and phthalocyanines were used as a p-type small molecule, while fullerene (C_{60}) and perylene derivatives were employed as an n-type small molecule. Most of small molecules have been deposited by thermal evaporation in vacuum, but soluble small molecules such as [6,6]-phenyl- C_{61} -butyric acid methyl ester ($PC_{61}BM$) can be coated from solutions. Lee *et al.* (2004) reported pentacene photodiodes (Schottky type) that could measure spectral photoresponses from 325 to 650 nm at zero bias condition. Voz *et al.* (2007) demonstrated fullerene devices of which the rectifying behavior could

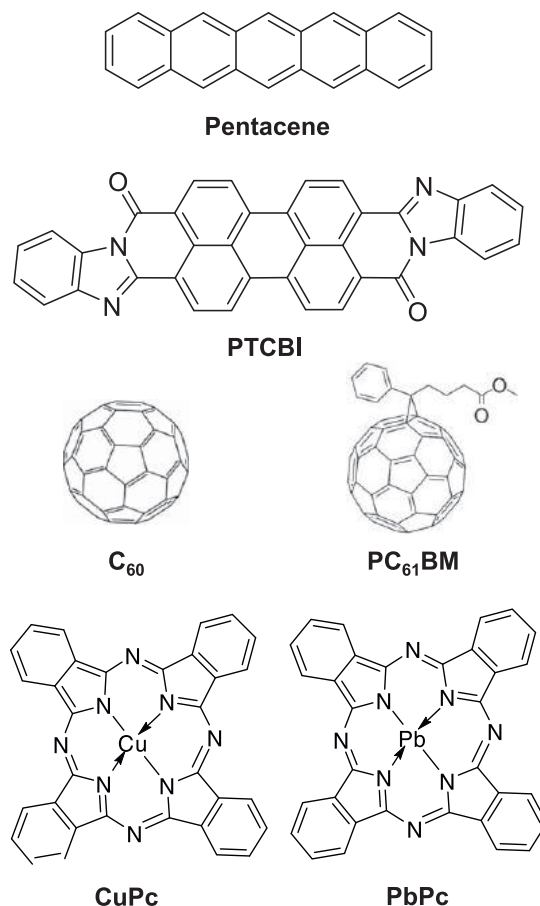


Fig. 7 p-type and n-type small molecules used for organic sensing layers (OSLs) in organic photodiodes: pentacene, 3,4,9,10-perylenetetracarboxylic bisbenzimidazole (PTCBI), fullerene (C_{60}), [6,6]-phenyl- C_{61} -butyric acid methyl ester ($PC_{61}BM$), copper phthalocyanine (CuPc), and lead phthalocyanine (PbPc).

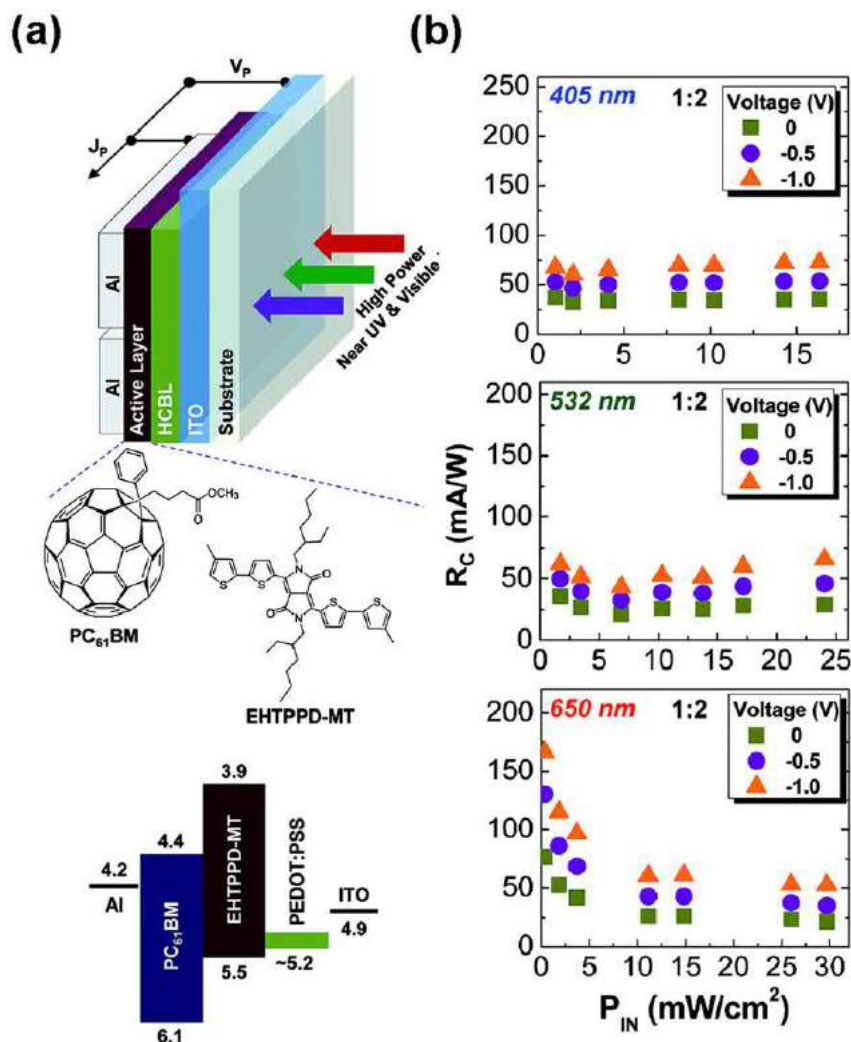


Fig. 8 (a) (top) Device structure and materials used for the fabrication of organic photodiodes with the EHTPPD-MT:PC₆₁BM (1:2) BHJ layers, (bottom) flat energy band diagram for the device (note that minus “-” sign and “eV” unit were omitted for clarity). (b) Corrected responsivity (R_c) as a function of the incident light intensity (P_{IN}) under illumination with near UV (405 nm), green (532 nm), and red (650 nm) lights. (Copyright 2015 Royal Society of Chemistry).

be improved by inserting a conducting polymer interlayer. The performance of organic photodiodes was noticeably improved by employing small molecular p–n junction structures. Peumans *et al.* (2000) reported bilayer-type organic photodiodes using copper phthalocyanine (CuPc) and 3,4,9,10-perylene-tetracarboxylic bisbenzimidazole (PTCBI) as a p-type and an n-type small molecule, respectively. However, the bilayer-type devices have a limitation in improving the device performances because of short exciton diffusion lengths in organic layers.

Organic Layer – Bulk Heterojunction

In this regard, organic bulk heterojunction structures were suggested as an alternative approach. Peumans *et al.* (2003) reported organic photodiodes based on CuPc:PTCBI bulk heterojunction layers, which showed controllable device performances by applying thermal annealing treatment leading to internal crystallization in bulk heterojunction films. Sullivan *et al.* (2004) demonstrated organic photodiodes based on CuPc:C₆₀ bulk heterojunction films in which exciton dissociations could be improved due to the intermolecular mixing between p-type (CuPc) and n-type (C₆₀) molecules. Panchromatic organic photodiodes based on the bulk heterojunction films of lead phthalocyanine (PbPc) and fullerene (C₇₀), which can detect broadband lights from 300 to 1000 nm, have been reported by Su *et al.* (2015). As shown in Fig. 8, Lee *et al.* (2015) reported solution-processed all-small molecular organic photodiodes, using solutions of 2,5-bis(2-ethylhexyl)-3,6-bis(4'-methyl-[2,2'-bithiophen]-5-yl)pyrrolo[3,4-c]pyrrole-1,4(2H,5H)-dione (EHTPPD-MT) and PC₆₁BM, which could sense near ultraviolet (UV, 405 nm) and broad visible light (up to 650 nm).

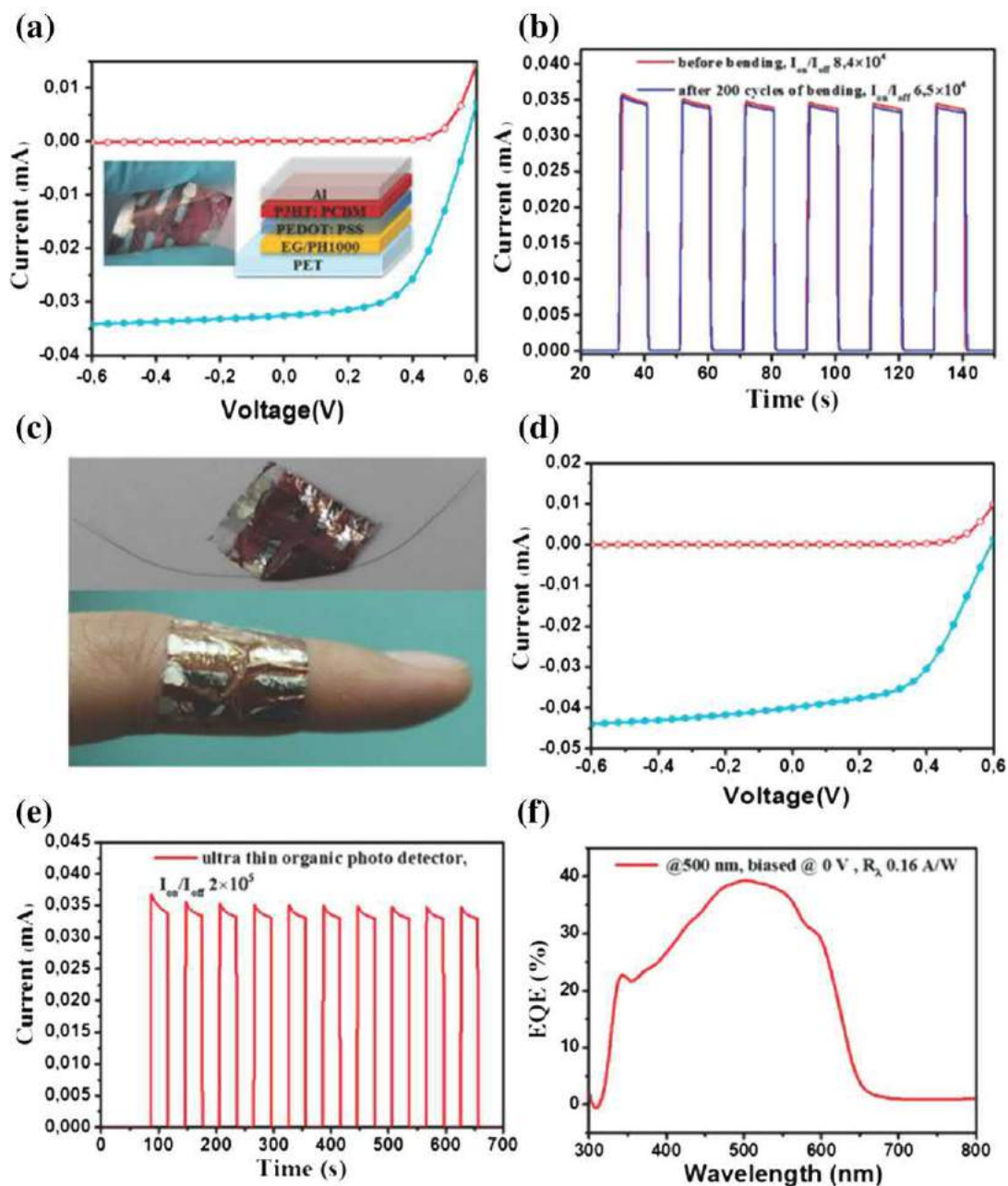


Fig. 9 (a) I - V characteristics of organic photodiodes fabricated on regular poly(ethylene terephthalate) (PET) (inset shows the fabricated device and schematic device structures of the organic photodiodes). (b) The on/off characteristics of the photodiodes at 0 V bias on regular PET before and after 200 cycles of bending test. (c) Ultrathin and flexibility demonstration of the fabricated photodiodes, thinner than a human hair, could be attached to the human body. (d) I - V characteristics of the ultrathin photodiodes. (e) The on/off characteristics of the ultrathin photodiodes. (f) External quantum efficiency (EQE) spectrum of the ultrathin photodiodes. (Copyright 2015 Wiley-VCH).

Organic Photodiodes With Polymers

Organic Layer – Single Material

Semiconducting polymers have been more extensively studied than small molecules because of their potentials for real plastic photodetectors with tough and rugged characteristics. Harrison *et al.* (1997) studied the thickness-dependent photo-generation characteristics for organic photodiodes with poly[2-methoxy-5-(2-ethylhexyloxy)-1,4-phenylenevinylene] (MEH-PPV) films.

Polymer nanowire photodiodes with poly[(9,9-dioctylfluorenyl-2,7-diyl)-*co*-(bi-thiophene)] (F8T2) films have been reported by O'Brien *et al.* (2006); Ray *et al.* (2015) tried to fabricate organic photodiodes with poly(3-hexylthiophene) (P3HT) films and tried to explain the collection-limited theory for extraordinary responses.

Organic Layer – Bulk Heterojunction

Nalwa *et al.* (2010) reported organic photodiodes based on the P3HT:PC₆₁BM bulk heterojunction films and applied for chemical sensors that could sense gas-phase oxygen and glucose. The similar P3HT:PC₆₁BM photodiodes, which were fabricated by Liu *et al.* (2015) who applied transparent conductive electrodes composed of graphene/PEDOT:PSS hybrid inks, exhibited excellent sensitivity in the wavelength range between 300 and 650 nm (see Fig. 9). Saracco *et al.* (2013) reported organic photodiodes with the bulk heterojunction layers of poly[(4,8-bis-(2-ethylhexyloxy)-benzo[1,2-b:4,5-b']dithiophene)-2,6-diyl-*alt*-(4-(2-ethylhexanoyl)-thieno[3,4-b]thiophene)-2,6-diyl] (PBDTTT-C) and PC₇₁BM, which exhibited the effective role of very thin polymeric interfacial layer for achieving homogenous interfaces leading to considerably low dark current by preventing undesired charge injection from cathodes in the devices. Armin *et al.* (2015) proposed new concept called charge collection narrowing (CCN)

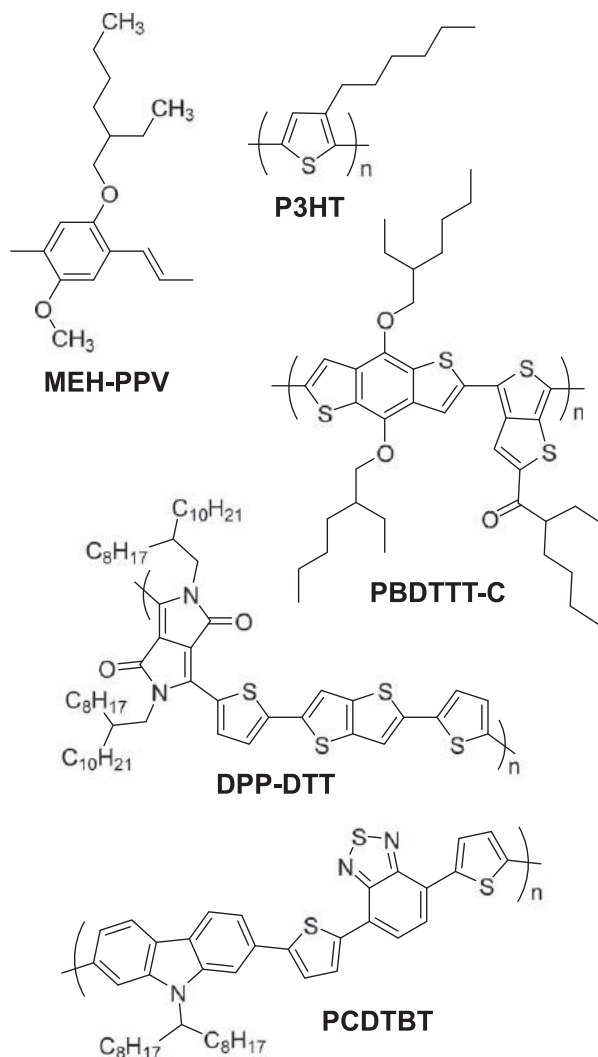


Fig. 10 Semiconducting (conjugated) polymers used for organic photodiodes: poly[2-methoxy-5-(2-ethylhexyloxy)-1,4-phenylenevinylene] (MEH-PPV), poly(3-hexylthiophene) (P3HT), poly[(4,8-bis-(2-ethylhexyloxy)-benzo[1,2-b:4,5-b']dithiophene)-2,6-diyl-*alt*-(4-(2-ethylhexanoyl)-thieno[3,4-b]thiophene)-2,6-diyl] (PBDTTT-C), poly[2,5-bis(2-octyldodecyl)-3,6-diketopyrrolopyrrole-*alt*-5,5'-(2,5-di(thien-2-yl)thieno[3,2-b]thiophene)] (DPP-DTT), and poly[N-9'-heptadecanyl-2,7-carbazole-*alt*-5,5'-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT).

in organic photodiodes fabricated with the bulk heterojunction layers of poly[2,5-(2-octyldodecyl)-3,6-diketopyrrolopyrrole-*alt*-5,5'-(2,5-di(thien-2-yl)thieno[3,2-b]thiophene)] (DPP-DTT) or poly[*N*-9'-heptadecanyl-2,7-carbazole-*alt*-5,5'-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT) and PC₇₁BM. Some examples of semiconducting (conjugated) polymers for organic photodiodes are listed in Fig. 10.

Organic Phototransistors With Small Molecules

Organic Layer – Single Material

Organic phototransistors with two types of organic thin-films (pentacene and tetracene), which exhibited photoresponses in both UV and visible ranges, were reported by Choi *et al.* (2006); Qi *et al.* (2013) showed n-type organic phototransistors using core expanded naphthalene diimides fused with two 2-(1,3-dithiol-2-ylidene)malononitrile moieties (NDI-DTYM2) bears one N-alkyl chain and one aromatic phenyl group, (NDI(2OD)(4tBuPh)-DTYM2). Organic phototransistors with highly aligned crystalline 2,7-dioctyl[1]benzothieno[3,2-b][1]benzothiophene (C8-BTBT) films, which exhibited strong responses under weak UV

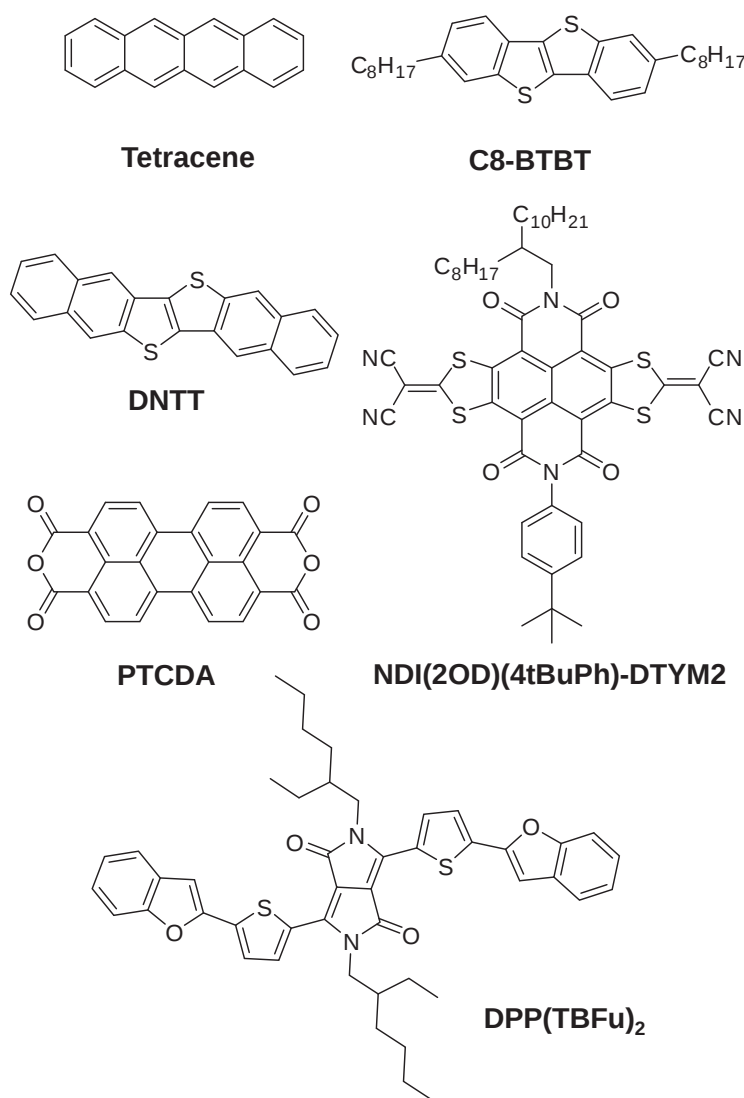


Fig. 11 Small molecules used for organic phototransistors: tetracene, 2,7-dioctyl[1]benzothieno[3,2-b][1]benzothiophene (C8-BTBT), dinaphtho[2,3-b:2',3'-f]thieno[3,2-b]thiophene (DNTT), perylene-3,4,9,10-tetracarboxylic dianhydride (PTCDA), core-expanded naphthalene diimides fused with two 2-(1,3-dithiol-2-ylidene)malononitrile moieties (NDI-DTYM2) bears one N-alkyl chain and one aromatic phenyl group, (NDI(2OD)(4tBuPh)-DTYM2), and 3,6-Bis(5-(benzofuran-2-yl)thiophen-2-yl)-2,5-bis(2-ethylhexyl)pyrrolo[3,4-c]pyrrole-1,4(2H,5H)-dione (DPP(TBFu)₂).

illumination, were reported by Yuan and Huang (2016); Chu *et al.* (2016) demonstrated high photosensitivity from flexible organic phototransistors using dinaphtho[2,3-b:2',3'-f]thieno[3,2-b]thiophene (DNTT) as an organic sensing channel layer. Fig. 11 shows various small molecules used for organic phototransistors.

Organic Layer – Bulk Heterojunction

Peng *et al.* (2013) reported organic phototransistors with the bulk heterojunction layers of PbPc and PTCDA, which showed superior performances compared to single component-based devices and high photo-responsivity in the near infrared region. Zhang *et al.* (2013) demonstrated organic phototransistors with solution-processable bulk heterojunction layers of 3,6-Bis(5-(benzofuran-2-yl)thiophen-2-yl)-2,5-bis(2-ethylhexyl)pyrrolo[3,4-c]pyrrole-1,4(2H,5H)-dione (DPP(TBFu)₂) and PC₇₁BM and studied field-assisted charge separation in small molecular bulk heterojunction structure. It has been also reported by Xu *et al.* (2015) that solution-processed ambipolar organic phototransistors can be fabricated with microphase-separated bulk heterojunction films from solutions of p-type (silylthynylated pentacene) and n-type (silylthynylated tetraazapentacene). This study was demonstrated that photo-effects were associated with a shift of threshold voltage because of photoinduced charge carriers.

Organic Phototransistors With Polymers

Organic Layer – Single Material

Narayan and Kumar (2001) demonstrated polymer-based phototransistors with regioregular poly(3-octylthiophene-2,5-diyl) (P3OT) films for a low intensity optical detection, while wavelength-selective phototransistors with dye-doped P3HT films were also reported by Meixner *et al.* (2006); Li *et al.* (2014) showed that organic phototransistors can be fabricated by employing a

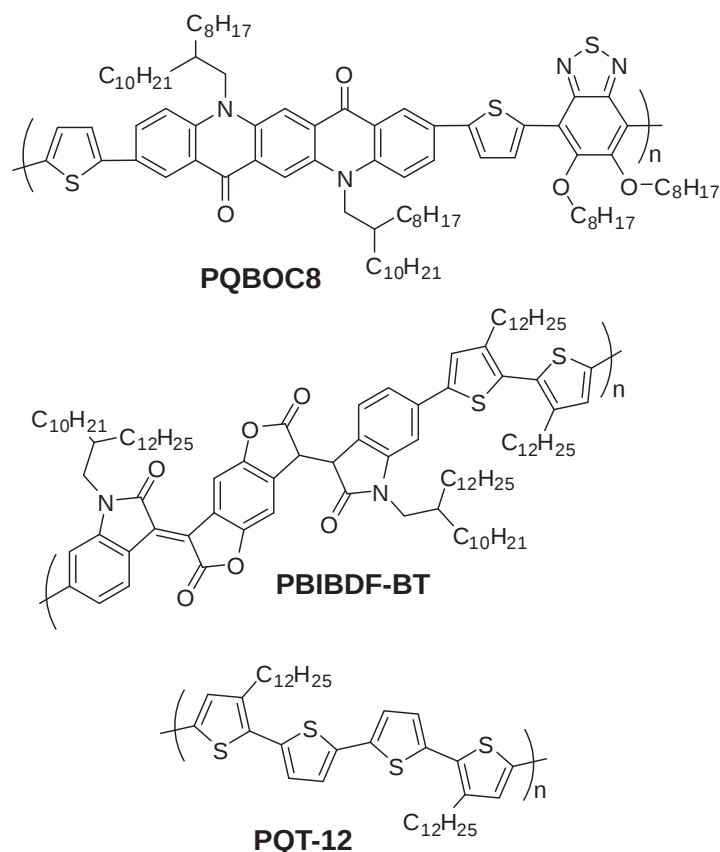


Fig. 12 Conjugated polymers used for organic phototransistors: PQBOC8 that was synthesized by Suzuki coupling reaction between 2,9-diboronic ester-N,N'-di(2-octyldodecyl)quinacridone and 4,7-bis(5-bromothiophen-2-yl)-5,6-bis(octyloxy)-benzothiadiazole, PBIBDF-BT that contains electron-rich bithiophene (BT) and electron deficient bis(2-oxindolin-3-ylidene)-benzodifuran-dione (BIBDF), and poly(3,3'-didodecylquaterthiophene) (PQT-12).

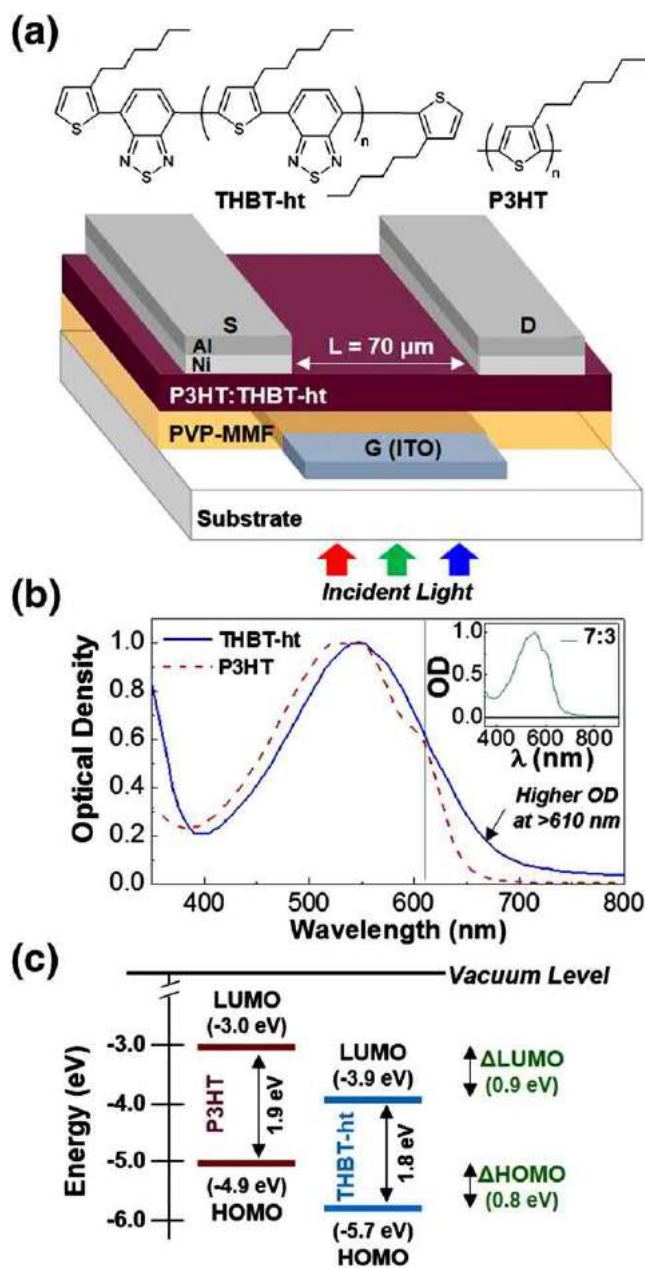


Fig. 13 (a) Materials and device structure for all-polymer phototransistors with the P3HT:THBT-ht layer. (b) Optical density (OD) as a function of wavelength (λ) for the pristine polymer and bulk heterojunction (P3HT:THBT-ht) layers. (c) Comparison of HOMO and LUMO levels for P3HT and THBT-ht. (Copyright 2016 IEEE).

simple self-assembly process in a large area at chloroform/water interface using ultrathin crystalline p-type semiconductor (PQBOC8) that was synthesized by Suzuki coupling reactions between 2,9-diboronate ester- N,N' -di(2-octyldodecyl)quinacridone and 4,7-bis(5-bromothiophen-2-yl)-5,6-bis(octyloxy)-benzothiadiazole. Lee *et al.* (2016) reported organic phototransistors with highly flexible nanofiber based on poly(3,3-didodecylquaterthiophene) (PQT-12) and poly(ethylene oxide) (PEO) as a semi-conducting polymer and a processing aid component, respectively. A couple of conjugated polymers for organic phototransistors are shown in Figs. 12 and 13.

Organic Layer – Bulk Heterojunction

Hwang *et al.* (2011) demonstrated organic phototransistors with the nanoscale phase-separated polymer:polymer bulk heterojunction films of p-type P3HT and n-type poly(9,9-dioctylfluorene-*co*-benzothiadiazole) (F8BT). Han *et al.* (2015)

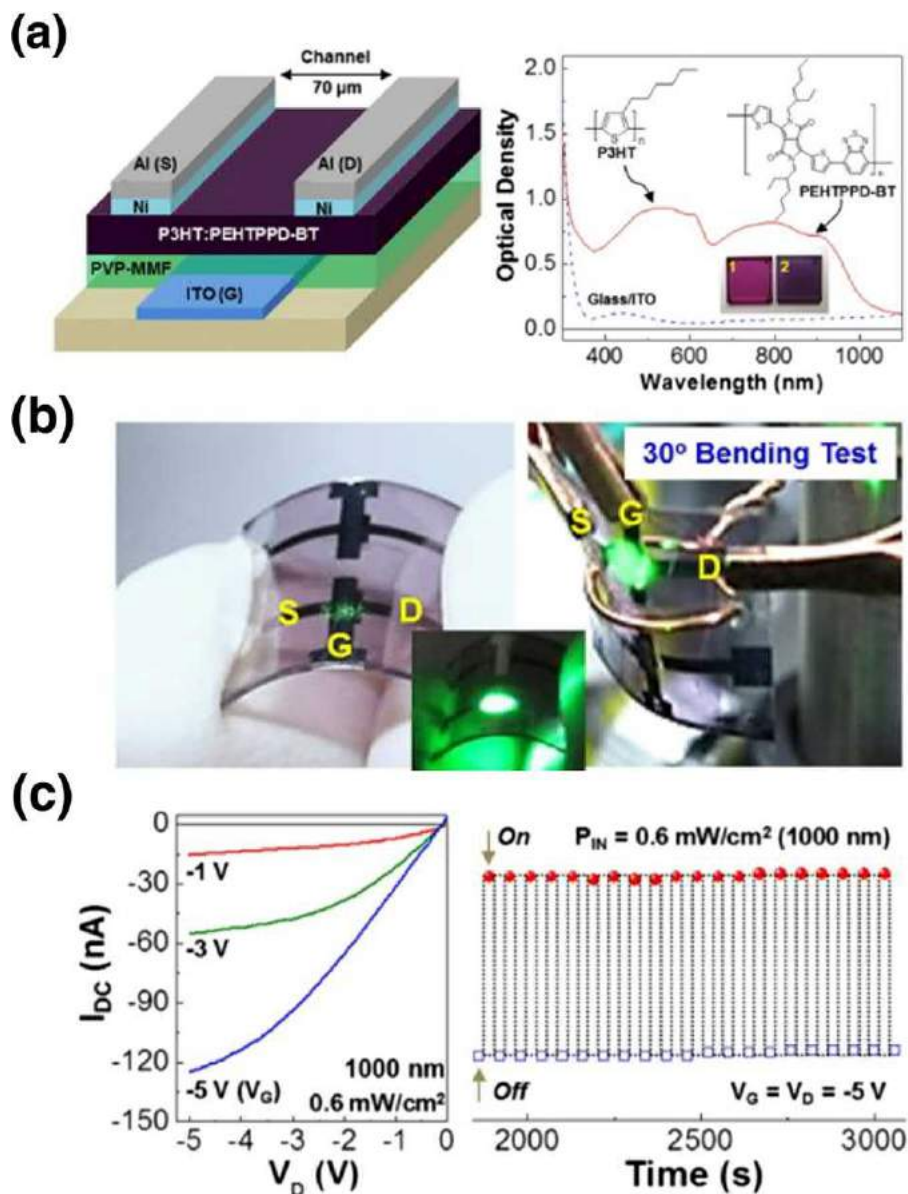


Fig. 14 (a) (left) Device structure for organic phototransistor with all-polymer bulk heterojunction (P3HT:PEHTPPD-BT) layer and (right) optical absorption spectra for the indium tin oxide (ITO) glass substrate and the P3HT:PEHTPPD-BT layer-coated ITO-glass substrate (see the photograph “2” compared with the photograph “1” of the P3HT layer-coated ITO-glass substrate). (b) Photographs for flexible plastic organic phototransistors with the P3HT:PEHTPPD-BT layers, which are bended at an angle of ~ 30 degree for the measurement of phototransistor performances (note that the green light was roughly illuminated as a guide for the channel part). (c) Low-voltage operation (upon bending at 30 degree) of the optimized flexible plastic organic phototransistors under illumination with near-infrared (NIR) (1000 nm) lights: (left) output curves and (right) repeated on/off modulation. (Copyright 2015 Nature Publishing Group).

demonstrated that all-polymer phototransistors, which were fabricated with the nanostructured bulk heterojunction layers of P3HT and poly[$\{2,5\text{-bis-(2-ethylhexyl)-3,6-bis-(thien-2-yl)-pyrrolo[3,4-c]pyrrole-1,4-diyl}\}-\text{co-}\{2,2'-(2,1,3\text{-benzothiadiazole})\}-5,5'\text{-diyl}\}$] (PEHTPPD-BT), can sense a broadband light from visible to near infrared region (see Fig. 14). Han *et al.* (2016) reported all-polymer phototransistors with the polymer:polymer bulk heterojunction layers of p-type P3HT and n-type hexylthiophene-end capped poly(3-hexylthiophene-co-benzothiadiazole) (THBT-ht) (see Fig. 13). All-polymer phototransistors with P3HT and poly[$\{4,8\text{-bis(2-ethylhexyloxy)-benzo[1,2-b:4,5-b']\text{-dithiophene}-2,6\text{-diyl-alt-(N-2-ethylhexylthieno[3,4-c]pyrrole-4,6-dione)-2,6\text{-diyl}\}$] (PBDITPD) were also demonstrated by Nam *et al.* (2016). These reports support that all-polymer (polymer:polymer) bulk heterojunctions play an efficient role in improving the photosensitivity of organic phototransistors (Fig. 15).

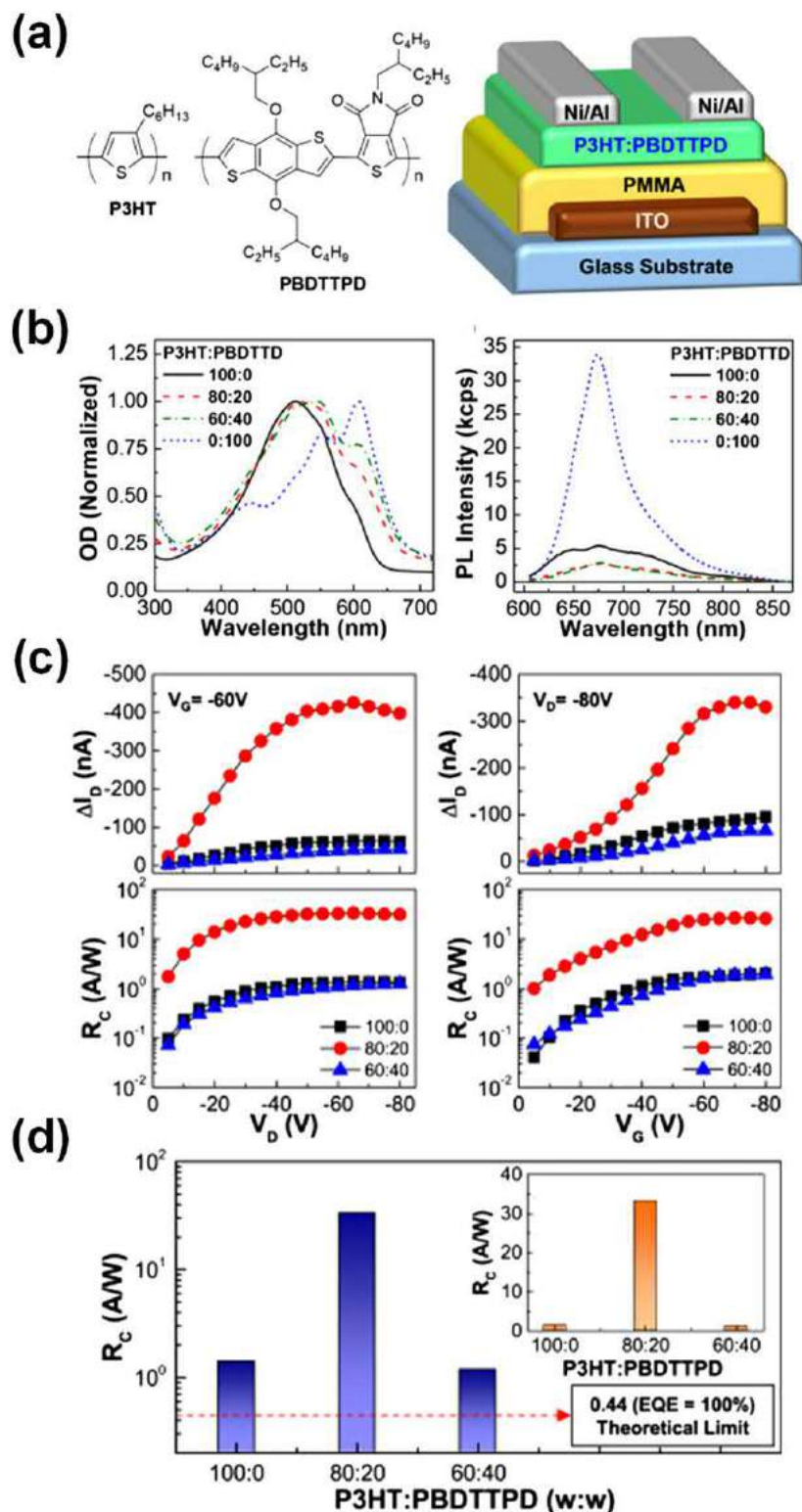


Fig. 15 (a) Chemical structure of materials and device structure for all-polymer phototransistors with the P3HT:PBDTTPD bulk heterojunction layer. (b) Normalized optical density (OD) (left) and photoluminescence spectra for P3HT:PBDTTPD bulk heterojunction layers coated on quartz substrates for different P3HT:PBDTTPD composition ratios (right). (c) Variation of drain current (ΔI_D) (top) and corrected responsivity (R_C) (bottom) as a function of drain voltage V_D , and gate voltage V_G , respectively. (d) Maximum R_C values for different P3HT:PBDTTPD composition ratios. (Copyright 2016 Elsevier).

See also: Organic Semiconductors

References

- Armin, A., Vuuren, R.D.J., Kopidakis, N., Burn, P.L., Meredith, P., 2015. Narrowband light detection via internal quantum efficiency manipulation of organic photodiodes. *Nature Communications* 6, 6343.
- Choi, J.-M., Lee, J., Hwang, D.K., *et al.*, 2006. Comparative study of the photoresponse from tetracene-based and pentacene-based thin-film transistors. *Applied Physics Letters* 88, 043508.
- Chu, Y., Wu, X., Lu, J., *et al.*, 2016. Photosensitive and flexible organic field-effect transistors based on interface trapping effect and their application in 2D imaging array. *Advanced Science* 3, 1500435.
- Han, H., Nam, S., Seo, J., *et al.*, 2015. Broadband all-polymer phototransistors with nanostructured bulk heterojunction layers of NIR-sensing n-type and visible light-sensing p-type polymers. *Scientific Reports* 5, 16457.
- Han, H., Nam, S., Seo, J., *et al.*, 2016. Organic phototransistors with all-polymer bulk heterojunction layers of p-type and n-type sulfur-containing conjugated polymers. *IEEE Journal of Selected Topics in Quantum Electronics* 22, 6000107.
- Harrison, M.G., Grüner, J., Spencer, G.C.W., 1997. Analysis of the photocurrent action spectra of MEH-PPV polymer photodiodes. *Physical Review B* 55, 7831–7849.
- Hwang, H., Kim, H., Nam, S., *et al.*, 2011. Organic phototransistors with nanoscale phase-separated polymer/polymer bulk heterojunction layers. *Nanoscale* 3, 2275–2279.
- Lee, H., Nam, S., Kwon, H., *et al.*, 2015. Solution-processable all-small molecular bulk heterojunction films for stable organic photodetectors: Near UV and visible light sensing. *Journal of Materials Chemistry C* 3, 1513–1520.
- Lee, J., Hwang, D.K., Park, C.H., Kim, S.S., Im, S., 2004. Pentacene-based photodiode with Schottky junction. *Thin Solid Films* 451–452, 12–15.
- Lee, M.Y., Hong, J., Lee, E.K., *et al.*, 2016. Highly flexible organic nanofiber phototransistors fabricated on a textile composite for wearable photosensors. *Advanced Functional Materials* 26, 1445–1453.
- Li, H., Wu, Y., Wang, X., Kong, Q., Fu, H., 2014. A self-assembled ultrathin crystalline polymer film for high performance phototransistors. *Chemical Communications* 50, 11000–11003.
- Liu, Z., Parvez, K., Li, R., *et al.*, 2015. Transparent conductive electrodes from graphene/PEDOT:PSS hybrid inks for ultrathin organic photodetectors. *Advanced Materials* 27, 669–675.
- Meixner, R.M., Göbel, H., Yildirim, F.A., Bauhofer, W., Krautschneider, W., 2006. Wavelength-selective organic field-effect phototransistors based on dye-doped poly-3-hexylthiophene. *Applied Physics Letters* 89, 092110.
- Nalwa, K.S., Cai, Y., Thoenig, A.L., *et al.*, 2010. Polythiophene-fullerene based photodetectors: tuning of spectral response and application in photoluminescence based (bio) chemical sensors. *Advanced Materials* 22, 4157–4161.
- Nam, S., Kim, H., Bradley, D.D.C., Kim, Y., 2016. All-polymer phototransistors with bulk heterojunction sensing layers of thiophene-based electron-donating and thienopyrroledione-based electron-accepting polymers. *Organic Electronics* 39, 199–206.
- Narayan, K.S., Kumar, N., 2001. Light responsive polymer field-effect transistor. *Applied Physics Letters* 79, 1891.
- O'Brien, G.A., Quinn, A.J., Tanner, D.A., Redmond, G., 2006. A single polymer nanowire photodetector. *Advanced Materials* 18, 2379–2383.
- Peng, Y., Lv, W., Yao, B., *et al.*, 2013. High performance near infrared photosensitive organic field-effect transistors realized by an organic hybrid planar-bulk heterojunction. *Organic Electronics* 14, 1045–1051.
- Peumans, P., Bulović, V., Forrest, S.R., 2000. Efficient, high-bandwidth organic multilayer photodetectors. *Applied Physics Letters* 76, 3855–3857.
- Peumans, P., Uchida, S., Forrest, S.R., 2003. Efficient bulk heterojunction photovoltaic cells using small-molecular-weight organic thin films. *Nature* 425, 158–162.
- Qi, Z., Liao, X., Zheng, J., *et al.*, 2013. High-performance n-type organic thin-film phototransistors based on a core-expanded naphthalene diimide. *Applied Physics Letters* 103, 053301.
- Ray, B., Baradwaj, A.G., Khan, M.R., Boudouris, B.W., Alam, M.A., 2015. Collection-limited theory interprets the extraordinary response of single semiconductor organic solar cells. *Proceedings of the National Academy of Sciences of the United States of America* 112, 11193–11198.
- Saracco, E., Bouthinon, B., Verilhac, J.-M., *et al.*, 2013. Work function tuning for high-performance solution-processed organic photodetectors with inverted structure. *Advanced Materials* 25, 6534–6538.
- Su, Z., Hou, F., Wang, X., *et al.*, 2015. High-performance organic small-molecule panchromatic photodetectors. *ACS Applied Materials & Interfaces* 7, 2529–2534.
- Sullivan, P., Heutz, S., Schultes, S.M., Jones, T.S., 2004. Influence of codeposition on the performance of CuPc–C₆₀ heterojunction photovoltaic devices. *Applied Physics Letters* 84, 1210–1212.
- Voz, C., Puigdollers, J., Cheylan, S., *et al.*, 2007. Photodiodes based on fullerene semiconductor. *Thin Solid Films* 515, 7675–7678.
- Xu, X., Xiao, T., Gu, X., *et al.*, 2015. Solution-processed ambipolar organic thin-film transistors by blending p- and n-type semiconductors: Solid solution versus microphase separation. *ACS Applied Materials & Interfaces* 7, 28019–28026.
- Yuan, Y., Huang, J., 2016. Ultrahigh gain, low noise, ultraviolet photodetectors with highly aligned organic crystals. *Advanced Optical Materials* 4, 264–270.
- Zhang, L., Wu, T., Guo, Y., *et al.*, 2013. Large-area, flexible imaging arrays constructed by light-charge organic memories. *Scientific Reports* 3, 1080.

Further Reading

- Azzellino, G., Grimaldi, A., Binda, M., *et al.*, 2013. Fully inkjet-printed organic photodetectors with high quantum yield. *Advanced Materials* 25, 6829–6833.
- Baeg, K.-J., Binda, M., Natali, D., Caironi, M., Noh, Y.-Y., 2013. Organic light detectors: Photodiodes and phototransistors. *Advanced Materials* 25, 4267–4295.
- Clark, J., Lanzani, G., 2010. Organic photonics for communications. *Nature Photonics* 4, 438–446.
- Falco, A., Cinà, L., Scarpa, G., Lugli, P., Abdellah, A., 2014. Fully-sprayed and flexible organic photodiodes with transparent carbon nanotube electrodes. *ACS Applied Materials & Interfaces* 6, 10593–10601.
- Heeger, A.J., 2010. Semiconducting polymers: The third generation. *Chemical Society Reviews* 39, 2354–2371.
- Vuuren, R.D.J., Armin, A., Pandey, A.K., Burn, P.L., Meredith, P., 2016. Organic photodiodes: The future of full color detection and image sensing. *Advanced Materials* 28, 4766–4802.
- Wakayama, Y., Hayakawa, R., Seo, H.-S., 2014. Recent progress in photoactive organic field-effect transistors. *Science and Technology of Advanced Materials* 15, 024202.
- Williams, G., Backhouse, C., Aziz, H., 2014. Integration of organic light emitting diodes and organic photodetectors for lab-on-a chip bio-detection systems. *Electronics* 3, 43–75.
- Zhang, H., Jenatsch, S., Jonghe, J.D., *et al.*, 2014. Transparent organic photodetector using a near-infrared absorbing cyanine dye. *Scientific Reports* 5, 9439.

Index

Note

This index is in letter-by-letter order, whereby hyphens and spaces within index headings are ignored in the alphabetization, and it is arranged in set-out style, with a maximum of three levels of heading.

Cross-reference terms in *italics> are general cross-references, or refer to subentry terms within the main entry (the main entry is not repeated to save space).*

Location references refer to the volume number, in bold, followed by the page number.

The page numbers suffixed by 'F' and 'T' refer to Figures and Tables, respectively.

A

- a*-axes [2:271](#)
- a*-plane [2:271](#)
- A549 human lung carcinoma cells [5:182](#)
- AA stacking [4:286](#)
- AB stacking [4:286](#)
- Abbe theory of microscope imaging [5:187–188](#)
- Abbe/Rayleigh limit of two-point resolution [5:15–16](#)
- Abbott medical optics (AMO) [4:531](#)
- aberrations [5:55–58](#)
 - aberration-free
 - diffraction limited systems [5:15–16](#)
 - system [5:185](#)
 - astigmatism [1:118–119](#)
 - chromatic aberrations [5:58–59](#)
 - coma [1:116–118](#)
 - contributions
 - calculation at image [4:211–214](#)
 - in freeform prism [4:214–215](#)
 - curvature of field and distortion [1:119–121](#)
 - monochromatic aberrations [5:55–58](#)
 - nonspherical surfaces [1:121–122](#)
 - ray tracing in meridional plane [1:114F](#)
 - theories [4:205](#)
 - vector [4:207, 4:211](#)
- aberrometry [5:112–113](#)
- above threshold ionization (ATI) [2:313–314](#)
- abruptly autofocusing optical beam (AAF optical beam) [1:415–416](#)
- absolute horizontal accuracy [4:402](#)
- absolute measure of SHG [4:287–289, 4:288T](#)
 - models for sample [4:287–289](#)
 - role of excitonic resonances [4:289](#)
 - role of substrate [4:289](#)
- absolute vertical accuracy [4:401–402](#)
- absorbers in rare gas halide lasers [2:363T](#)
- absorption [2:107, 4:67, 4:310, 4:339, 5:168, 5:207](#)
 - acoustic [1:135–136](#)
 - band [4:364](#)
 - coefficient [3:221, 5:160–161, 5:161F](#)
 - cross-section [3:221](#)
 - edge [2:94](#)
 - efficiency [5:256–257](#)
 - experiment [1:392–393](#)
 - measurements [2:73](#)
 - spectroscopy with attosecond pulses [2:242–243](#)
 - attosecond four-wave mixing spectroscopy [2:243](#)
 - attosecond transient absorption spectroscopy [2:242–243](#)
 - spectrum [1:459, 2:107F, 4:88](#)
 - 2D spectroscopy [2:161](#)
 - "absorptive" 2DES spectrum [2:185–186](#)
- abstract imaging system [3:156](#)
- "AC apodized" Gaussian profile [1:233](#)
- ac Stark effect [4:320](#)
- ac Stark shift [4:346](#)
- Acanthamoeba* keratitis [5:103–104](#)
 - two-photon imaging [5:105F](#)
- accelerated free electrons [1:405](#)
- acceptor
 - acceptor mode [1:187](#)
 - materials in OPV [5:265](#)
- accidental cell death [3:107](#)
- accommodation [5:62F, 5:141, 5:150–151](#)
 - amplitude of [5:47](#)
 - restoration techniques [5:120](#)
- acetonitrile (ACN) [5:277](#)
- achievable bit-error rates [1:276–277](#)
 - comparison of modulation formats [1:280](#)
 - intensity modulation [1:277](#)
 - partial response signaling [1:278–280](#)
 - phase-shift keying [1:277](#)
 - quadrature amplitude modulation [1:277–278](#)
- achievable SNR [1:273](#)
- achromatic doublet [5:188–189](#)
- acoustic
 - absorption [1:135–136](#)
 - attenuation [1:136–137](#)
 - density grating [2:4](#)
 - diffraction by acoustic waves [1:131–133](#)
 - disturbance [1:130](#)
 - frequency [1:131](#)
 - power [1:134–135](#)
 - propagation [1:131–132](#)
 - properties [2:7](#)
 - radiation [5:192](#)
 - strain amplitude [1:130–131](#)
 - wavelength [1:132–133](#)
 - wavevector [1:137](#)
- acoustic optical modulator (AOM) [3:46](#)
- acoustic radiation force (ARF) [5:147](#)
- acoustical studies and vibrometry [4:528](#)
 - laser Doppler vibrometry [4:528](#)
 - successive pulse illumination in acoustical studies [4:528](#)
 - vibration measurement of vocal folds [4:528–530](#)
- acousto-optic modulator (AOM) [1:139, 2:190, 4:98–99, 4:115, 4:435–436, 4:501–502, 4:502](#)
- acousto-optic programmable dispersive filter (AOIPDF) [2:297–298](#)
- acousto-optic tunable filter (AOTF) [1:133–134, 1:137–139](#)
 - wavevector diagram for [1:138F](#)
- acousto-optics (AO) [1:130](#)
 - see also adaptive optics (AO)
 - anisotropic bragg diffraction [1:133–134](#)
 - AOITs [1:137–139](#)
 - beam deflectors [5:194](#)
 - birefringent scanners [1:136–137, 1:137F](#)
 - bragg diffraction [1:133](#)
 - devices [1:130, 1:131, 1:135, 1:135–136](#)
 - diffraction
 - by acoustic waves [1:131–133](#)
 - cell [1:134F](#)
 - efficiency [1:134–135](#)
 - of light in Raman-Nath limit [1:132F](#)
 - imaging [3:226, 3:226–228, 3:227F](#)
 - interaction of light waves [1:131F](#)
 - laser cavity, acousto-optic Q-switch in [1:140F](#)
 - materials [1:135, 1:135T](#)
 - modulators [1:139–140](#)
 - momentum diagram for diffraction of light [1:133F](#)
 - noncollinear AOIF orientation [1:138F](#)
 - photo-elastic effect [1:130–131](#)
 - Q-switch [1:140](#)

- acousto-optics (AO) (*continued*)
 scanners [1:135–136](#), [1:136F](#)
 spectrum of phase-modulated wave [1:132F](#)
 wavevector diagram for AOTF [1:138F](#)
- acousto-optics-optical coherence
 tomography (AO-OCT) [5:73](#)
- activator [2:324](#)
- activator protein-1 (AP1) [3:115–116](#)
- active cavity interferometer [2:309](#), [2:309F](#)
- active chirp laser stabilization for FMCW
 lidar [4:436](#)
[80](#) GHz distributed feedback based chirp
 stabilization [4:436–438](#)
 broadband [5](#) THz chirp stabilization
[4:436](#)
- active components [1:223–224](#)
- active dynamic holographic display devices
 (ADHDDs) [4:114](#), [4:114–115](#)
- active fiber compatible components [1:221](#),
[1:223–226](#)
- active gain medium [2:329](#)
- active lasing materials [2:268](#)
- active locking techniques [2:453](#)
- active materials in polymer solar cells
 acceptor materials in OPV [5:265](#)
 polymer active materials [5:263–265](#)
- active matrix (AM) [3:9](#), [3:12](#), [5:239](#)
 driving method [3:30F](#), [3:31](#)
 TFT driving [3:66–67](#)
- active matrix liquid crystal displays
 (AMLCDs) [3:1](#), [3:2](#), [3:5–6](#), [3:14](#)
- active matrix organic light emitting diode
 (AMOLED) [3:14](#), [5:239](#)
- active medium [2:368](#)
- active mode locked lasers (MLL) [4:38](#)
- active mode-locking of laser diodes [4:62](#)
- active optical devices [1:214](#)
- active optics [3:197](#), [3:198](#), [3:198](#)
- active Q switching [2:418–419](#)
- active regions [1:382–383](#), [1:382F](#)
- active remote sensing [2:435](#)
- active ring resonators [4:294](#)
- Active Tiling display system (AT display
 system) [4:116](#)
- active-passive axial waveguide integration
[4:243–246](#), [4:244F](#)
- acuity [3:235–237](#)
- adaptive control [5:42](#)
- adaptive immunity [3:112–114](#)
- adaptive optics (AO) [3:197](#), [3:198–200](#),
[3:278](#), [3:279](#), [5:67](#), [5:72](#), [5:72–73](#),
[5:91](#), [5:119](#)
see also acousto-optics (AO)
- adaptive system design [3:201–202](#)
- applications [3:202–203](#), [5:81–82](#)
 correction [5:23F](#)
 detection schemes [5:26F](#)
 fast corrections [3:198–200](#)
 flood-illuminated ophthalmoscopy
[5:73–74](#)
 functional assessment in conjunction with
 AO imaging [5:82](#)
 future development [3:202](#)
 imaging through atmosphere [3:200–201](#),
[3:200F](#)
 metrics to analyze cell mosaics [5:79–81](#)
 post processing and AO image analysis
[5:78–79](#)
 and post-processing [3:278–279](#)
 retinal imaging [5:73](#)
 slow corrections [3:198](#)
 in vivo multiphoton retinal imaging with
[5:86–87](#)
- adaptive optics scanning laser
 ophthalmoscope (AOSLO) [5:64–65](#),
[5:65F](#), [5:85](#), [5:86–87](#)
 erythrocyte imaging with [5:67–68](#)
 leukocyte imaging with [5:67](#)
- adaptive optics-optical coherence
 tomography (AO-OCT) [5:77–78](#)
- adaptive signal processing [4:333](#)
- add noise [1:263](#)
- "add-drop filter" configuration [4:216](#)
- add/drop multiplexer (ADM) [4:65](#)
- Added White Gaussian Noise (AWGN)
[1:276](#)
- Addition de Photons par Transferts
 d'Energie' (APTE) [2:395](#)
- Additive White Gaussian Noise (AWGN)
[1:299](#)
- adequate organic solvent [2:337](#)
- adiabatic pulse pairs [4:345](#)
- adjacent-channel isolation [1:266](#)
- advance cruise missile (ACM) [4:450](#)
- advance super in-plane-switching (AS-IPS)
[3:60–61](#)
- advanced coherent imaging with FMCW
 lidar [4:439–440](#)
- advanced delayed fluorescence [5:244–246](#)
- advanced EDFA functions [1:362–363](#)
- advanced LIGO [1:432](#), [1:434](#), [1:435](#), [1:436F](#),
[1:440F](#), [1:441F](#)
- advanced plasmonic nanostructures [5:300](#)
- Advanced Scientific Concepts (ASC) [4:446](#)
- advanced Virgo [1:432](#)
- aerial cables [1:333](#)
- aero-optics [3:199](#), [3:203](#)
see also adaptive optics (AO); acousto-
 optics (AO)
- age, changes with [5:60](#)
 accommodation [5:62F](#)
 aperture stop and iris [5:60](#)
 cornea [5:60](#)
 lens [5:60–61](#)
 light loss at retina [5:61](#)
 presbyopia [5:61](#)
 refraction [5:61](#)
- age related macular degeneration (AMD)
[5:164](#), [5:166F](#)
- agency for science, technology and research
 system (A*STAR system) [4:117–118](#),
[4:118F](#)
- aggregate nominal pulse density (ANPD)
[4:401](#)
- aggregate nominal pulse spacing (ANPS)
[4:401](#)
- air bands [1:171](#), [1:185](#)
- air periodic system [1:185](#)
- air puff corneal dynamic imaging [5:142](#),
[5:143F](#)
 air puff OCT imaging [5:142–143](#)
 air puff scheimpflug imaging [5:142](#)
- clinical and experimental applications
[5:145–146](#)
- corneal deformation parameters
[5:143–144](#)
- corneal mechanical properties from
[5:144–145](#)
- air-biased-coherent-detection matrix-optics
 (ABCD matrix-optics) [4:205–206](#)
- air-clad fiber [1:200–203](#)
- air-cooled argon ion lasers [4:361](#)
- air-cooled ion lasers [2:376](#)
- Air-guiding Photonic Bandgap fibers (PBG
 fibers) [1:263F](#)
- air-plasma
 single-color method [1:411–412](#), [1:411F](#),
[1:412F](#)
 two-color method [1:412–416](#)
- airborne targets [4:479–482](#)
- airborne topographic lidar systems, 3D
 metrics for [4:401](#)
- current lidar metric standards [4:401](#)
- emerging capabilities and challenges
[4:402–403](#)
 accounting for relative horizontal
 accuracy [4:403–404](#)
 density and completeness in 3D
[4:404–405](#)
 foliage penetration performance
 characterization [4:404](#)
 lidar error model [4:402–403](#)
 point cloud resolution measurement
[4:403](#)
- airy disk [5:186](#)
- airy pattern [1:72](#), [5:15](#)
- albumin and other heavy proteins [3:110](#)
- Aldis calculus [4:205](#)
- Aldis theorem [4:212](#)
- AlGaAs/GaAs quantum wells, optical
 properties of [2:105–107](#)
 general principles [2:105–107](#)
 optical absorption [2:107–108](#)
 photoluminescence and spontaneous
 emission [2:108](#)
- AlGaN-based laser diodes [2:271](#), [2:273–277](#),
[2:274F](#), [2:277F](#)
- algebra theorem [3:163](#)
- alkoxides [2:346](#)
- "all pass" arrangement [4:216](#), [4:216F](#)
- all-optical
 data regeneration [1:251–252](#)
 pump-probe spectroscopy of correlated
 electron materials [2:133–135](#)
 switching [1:142](#), [4:224](#)
- all-optical signal regeneration [1:364–365](#),
[1:366F](#), [1:370F](#)
- all-optical 2R/3R regeneration using
 optical nonlinear gates [1:367–369](#),
[1:367F](#)
 optical clock recovery [1:369](#)
 optical decision element [1:367–369](#)
 optical regeneration by saturable
 absorbers [1:369–370](#)
 generalities on [1:365–366](#)
 principles [1:365–366](#)
 qualification of signal regenerator
 performance [1:366–367](#)

- synchronous modulation technique [1:370–371](#)
- allocation ID (Alloc-ID) [2:75](#)
- alternating current signal (AC signal) [3:91–92](#)
- alternating projection algorithms [3:147–148](#), [3:147F](#), [3:152–153](#)
- alternative coatings [1:33](#)
- alternative plasmonic materials
- applications [2:261–262](#)
 - case for [2:252–254](#)
 - chemical sensing [2:262](#)
 - conventional plasmonic materials [2:255–256](#)
 - finding alternatives [2:256–257](#)
 - high temperature plasmonics [2:262–263](#)
 - ideal plasmonic material [2:254–255](#)
 - integrated plasmonics [2:261–262](#)
 - metal alloys [2:258–259](#)
 - metals to “less-metals” [2:259–261](#)
 - metamaterials and transformation optics [2:263–264](#)
 - optical properties of materials [2:254](#)
 - semiconductors to metals [2:256–257](#), [2:256F](#)
 - two-dimensional plasmonic materials [2:261](#)
- altitude-resolved atmospheric Raman water spectrum signal [4:465](#), [4:468F](#)
- alumina (Al_2O_3) [1:28](#), [2:379](#)
- ceramic tube [2:369](#)
 - thermal insulator [2:369](#)
- aluminosilicate ($\text{Al}_2\text{O}_3\text{-SiO}_2$) [2:347](#)
- aluminum (Al) [2:256](#), [2:271](#), [5:298](#)
- aluminum nanoparticle (Al NP) [5:298](#)
- aluminum zinc oxide (AZO) [2:257](#)
- aluminum-doped zinc oxide (AZO) [2:258](#), [5:273](#)
- Alzheimer's disease (AD) [3:124](#), [3:124–125](#)
- ambient contrast ratio (ACR) [3:73](#)
- ambient illumination and reflection [3:72–73](#)
- ambiguity function (AF) [4:164](#), [4:174–175](#)
- approach to AF reconstruction [4:170–171](#)
 - of image of incoherent source [4:168–169](#)
 - inoptics [4:164](#)
- imensity spectrum of Fresnel diffraction pattern [4:165–166](#)
- mutual intensity [4:164–165](#)
- phase-space tomography [4:169–170](#)
- propagation through paraxial optical system [4:167](#)
- propagation-based holographic phase retrieval [4:171–172](#)
- in space-invariant (isoplanatic) imaging [4:168](#)
- Ambiguity Zone Detectors (AZD) [1:279–280](#)
- American National Standards Institute (ANSI) [5:106–107](#)
- contrast ratio [3:72](#)
- American Society for Photogrammetry and Remote Sensing (ASPRS) [4:401](#)
- lidar calibration/validation Working Group [4:403–404](#)
- Amici prism [1:126](#), [1:126F](#)
- ammonia (NH_3) [1:389](#), [2:273](#)
- amorphous glass [2:265](#)
- amorphous oxide semiconductors (AOSs) [3:12–13](#), [3:13F](#)
- TFT [3:14–15](#)
 - thin film [3:14–15](#)
- amorphous Si TFT (a-TFT) [3:53](#)
- amorphous silicon (a-Si) [3:14](#)
- amplification [1:97](#), [1:98](#), [1:223–224](#)
- factors [2:412](#)
 - regenerative and multipass amplification [2:412](#)
 - of spectrally narrow nanosecond or sub-nanosecond pulses [2:412](#)
- amplified laser emission [2:338](#)
- amplified spontaneous emission (ASE) [1:242–243](#), [1:263](#), [1:298–299](#), [1:353](#), [1:361–362](#), [1:364](#), [1:369–370](#), [2:82](#), [2:286](#), [2:410–411](#), [2:426–427](#), [4:46](#), [5:312–313](#)
- amplified stimulated emission (ASE) [4:362](#)
- amplifier
- add noise [1:273](#)
 - configurations [1:353–354](#), [1:354F](#)
 - gain [1:246](#), [1:351–352](#), [1:353](#)
 - noise figure [1:353](#)
 - technologies [1:263F](#)
- amplitude [1:139](#), [1:231–232](#)
- of accommodation [5:47](#)
 - deterministic [4:180](#)
 - of diffraction wave from rectangular slit [1:70](#), [1:70F](#)
 - distribution [4:182](#)
 - transmittance of hologram [4:30](#)
- amplitude modulation (AM) [1:146–147](#), [1:367–368](#), [1:373](#), [1:392–393](#)
- radio [1:373–374](#)
- amyloid fibers [2:174](#)
- amyloid proteins [2:174–175](#)
- kinetics of amyloid formation by human amylin [2:175–176](#)
 - structure of polyglutamine fibrils [2:174–175](#)
- anaglyph [3:45–46](#)
- analog signals [1:337](#)
- analog to digital converters (ADCs) [1:143](#), [1:278](#), [1:295–296](#), [3:175](#), [4:434–435](#), [4:509–510](#), [5:8](#)
- analogue holographic recording and replay [4:106–107](#)
- in-line hologram recording [4:107F](#)
 - in-line hologram replay [4:107F](#)
 - off-axis hologram recording [4:108F](#)
 - real image reconstruction from off-axis hologram [4:108F](#)
- analyzer [2:211](#), [3:59](#)
- angle [1:158](#)
- analyzer-based imaging (ABI) [4:171](#)
- Anderson localization [4:278](#)
- Kerr nonlinearity in theory and experiment [4:279–281](#)
 - of light [4:278](#)
 - other forms of nonlinearity [4:282](#)
 - theoretical considerations [4:281–282](#)
 - transverse Anderson localization of light [4:278–279](#)
 - in theory and experiment [4:279–281](#)
- androgenetic alopecia (AGA) [3:120](#)
- anemometer [4:497](#)
- angle of refraction [1:153](#)
- angle-resolved photoemission spectroscopy (ARPES) [2:144–145](#)
- angled physical contact technique (APC technique) [1:337–338](#)
- angular aperture [1:133](#)
- angular dependence [1:83](#)
- angular field [1:139](#)
- angular isoplanatism [3:201](#)
- angular magnification [1:108–109](#)
- angular streaking [2:240](#)
- angular-spectrum expansion [3:140](#)
- anharmonic oscillator spectra interpretation [2:160–161](#), [2:161F](#)
- anharmonic potential [2:161](#)
- anisometropia [5:54](#)
- anisotropic Bragg diffraction [1:133–134](#), [1:136](#), [1:137](#)
- anisotropic scanner [1:136](#)
- anisotropic scattering [3:221](#)
- anisotropic waveguide-based modulators (AWM) [4:117](#)
- holographic video system [4:117F](#)
- annealing process [3:15](#)
- anode [5:232](#)
- anomalous dispersion [1:5](#), [1:6](#), [1:195](#), [1:315](#), [4:1–2](#)
- antenna theory [3:137–138](#)
- anterior corneal surfaces [5:45](#)
- anti-aliasing filters (AAFs) [1:299](#)
- anti-distorted beam [4:322](#)
- anti-glare surfaces (AG surfaces) [3:73](#)
- anti-reflecting surfaces (AR surfaces) [3:73](#)
- anti-Stokes waves [2:265](#)
- anti-tumor immune response induction after PDT [3:114](#)
- anti-vascular effect of PDT [3:111](#)
- anti-vascular endothelial growth factor (anti-VEGF) [5:164](#)
- antibiotic resistance, rise of [3:102–103](#)
- anticorrelation [1:453](#)
- antiferromagnetic order (AFM) [2:123](#)
- antimicrobial photodynamic inactivation (aPDI) [3:103](#), [3:103](#), [3:105](#), [3:106](#)
- localized infections [3:105F](#)
 - pathogen classification and PS structure [3:103–105](#)
 - PDT for infectious disease [3:105–106](#)
 - rise of antibiotic resistance [3:102–103](#)
 - virulence [3:105](#)
- antimissile defense (AMD) [4:445](#)
- antireflection (AR) [2:329](#), [2:408–409](#), [5:292–293](#)
- coating [2:457](#)
- AO scanning laser/light ophthalmoscopy (AOSLO) [5:73](#), [5:74–77](#), [5:77F](#), [5:78–79](#), [5:80F](#), [5:81F](#)
- aperture [3:247](#)
- mask [3:3](#)
 - size [1:139](#)
 - stop [5:45–46](#), [5:46F](#)
 - and iris [5:60](#)
 - synthesis [3:241](#)
 - transmission function [1:73](#)
- aplanatic front element [5:188](#), [5:188F](#)

- aplanatic system [5:185](#)
 - apochromats [5:188–189](#)
 - apodization [5:16](#)
 - apoptosis, PDT and [3:107](#)
 - applanation-free femtosecond laser [5:133](#)
 - application-specific integrated circuits (ASICs) [1:295–296](#), [4:418](#)
 - approximate model [5:14](#)
 - aquatic holography [4:106](#)
 - aqueous humor [5:43](#)
 - Ar II laser blue-green wavelengths [2:376–377](#), [2:377T](#)
 - aramid yarns [1:331](#)
 - arbitrary angle [4:178–179](#)
 - arbitrary polarization [1:162](#)
 - Arc gap [3:32](#)
 - argon ion [2:376](#)
 - laser [2:381](#), [4:360–361](#), [4:360T](#)
 - oscillation [2:382](#)
 - argon laser trabeculoplasty (ALT) [5:136](#)
 - array illuminators *see* multiple beamsplitters
 - array of N identical apertures [1:74](#)
 - array theorem [1:72–73](#)
 - arrayed waveguide grating (AWG) [1:252](#), [1:266](#), [2:74](#), [4:222](#)
 - arrayed waveguide grating routing (AWGR) [4:241](#)
 - arsenic triselenide (As_2Se_3) [1:210](#)
 - arsenic trisulfide [1:210](#)
 - artificial reference stars [3:202](#)
 - aspherical optical metrology [4:438–439](#)
 - asphericity [5:44–45](#)
 - astigmatism [1:118–119](#), [1:120F](#), [1:122F](#), [5:52–54](#), [5:53F](#)
 - astronomical telescope [1:122](#)
 - astronomy, high resolution imaging in [3:253–254](#)
 - astrophysical molecular spectroscopy [1:396–399](#)
 - Asus X-tion Pro [4:512](#)
 - asymmetric cladding layers [1:186](#)
 - asymmetric double quantum well (a-DQW) [2:85–86](#)
 - asymmetric twin guides (ATG) [4:244](#)
 - asymmetrical devices [1:340](#)
 - asynchronous disk mirroring [1:281](#)
 - asynchronous transfer mode protocol (ATM protocol) [2:74](#)
 - Atacama Large Millimeter Array (ALMA) [1:397](#), [1:397T](#), [1:398F](#)
 - athletic performance, PBM for [3:126](#)
 - atmospheric/atmosphere
 - coherence length
 - components and contaminants [4:463–469](#)
 - profiling [4:483](#)
 - turbulence [3:272–274](#), [3:273F](#), [3:278](#)
 - atom optics
 - atom interferometry [5:217–218](#)
 - atom-standing wave interactions [5:214–217](#)
 - Kapitza-Dirac scattering [5:213–214](#)
 - nanofabricated atom optics [5:204–207](#)
 - standing waves of light [5:207–213](#)
 - atom-standing wave interactions [5:214–217](#)
 - atom(s) [1:224](#)
 - condensation [1:458–459](#)
 - flux in n th diffraction order [5:205](#)
 - interferometry [5:217–218](#), [5:217F](#)
 - atomic beam diffraction from optical standing wave [5:212](#)
 - atomic Coulomb potentials [2:33](#)
 - atomic force microscopy (AFM) [4:419](#), [4:419F](#), [5:175](#), [5:185](#), [5:190](#), [5:261](#)
 - atomic layer deposition (ALD) [2:260](#), [3:56](#)
 - atomic orbitals (AOs) [5:220–221](#)
 - atomic physics [2:40](#)
 - EIT and cross-coupling of photons in doped PCs [2:15–18](#)
 - radiative decay and photon-atom binding in PC [2:12–15](#)
 - atomic QED approaches [1:458–459](#)
 - atomic quantum optics [1:458–459](#)
 - atomic systems [1:462](#)
 - atomic vapors [2:17](#)
 - atomic-like situations [1:488](#)
 - attention bias modification (ABM) [3:126](#)
 - attenuation [1:18](#), [1:255](#), [1:255–256](#), [1:308–309](#), [1:308F](#), [1:309F](#)
 - of silica material [4:1](#)
 - test methods for [1:309](#)
 - backscattering method [1:309–311](#), [1:310F](#)
 - cut-back method [1:309](#)
 - insertion loss method [1:311](#)
 - vs.* wavelength approximation for silica fibers [4:14–15](#), [4:14F](#)
 - attenuators [1:336](#), [1:342](#), [1:360](#)
 - attosecond
 - beating reconstruction by interference of two-photon transitions [2:241–242](#)
 - four-wave mixing spectroscopy [2:243](#)
 - light house [2:238](#), [2:239F](#)
 - optical pulses [2:233–234](#)
 - electric laser field in time and frequency domain [2:234](#)
 - formation of attosecond pulses [2:237](#)
 - high-order harmonic generation [2:234–235](#)
 - streak camera [2:241](#), [2:241F](#), [2:322](#), [2:323F](#)
 - technology [2:316–319](#)
 - gating techniques [2:319–321](#)
 - III-V [2:316–319](#)
 - transient absorption spectroscopy [2:242–243](#)
- attosecond pulses [2:311](#)
 - absorption spectroscopy with [2:242–243](#)
 - characterization [2:321–323](#)
 - formation [2:232](#)
 - attosecond pulse trains [2:237](#)
 - isolated attosecond pulses [2:237](#)
 - generation [2:234](#)
 - in water-window region [2:321](#)
 - photoelectron and ion spectroscopy with [2:240–241](#)
 - attosecond streak camera [2:241](#)
 - ion spectroscopy [2:242](#)
 - reconstruction of attosecond beating [2:241–242](#)
 - technology [2:233](#)
 - trains [2:237](#)
- attosecond spectroscopy
 - see also* 2D electronic spectroscopy (2DES); multidimensional THz spectroscopy
- and applications [2:238–240](#)
 - with attosecond pulses [2:242–243](#)
 - with femtosecond pulses [2:239–240](#)
 - generation of attosecond optical pulses [2:233–234](#)
- photoelectron and ion spectroscopy with attosecond pulses [2:240–241](#)
- attractive Coulomb force [2:40](#)
- AOI Optonics [3:61](#)
- Auger carrier [2:115](#)
- Auger effects [2:114–115](#), [2:116–117](#), [2:117F](#)
- Auger recombination [2:110](#), [2:463–464](#)
- augmented reality (AR) [3:48](#), [4:124](#)
 - devices [3:75](#)
 - display [3:48–49](#)
 - HMD [3:51](#), [3:51F](#)
- Autler-Townes effect [2:24–25](#)
- Autler-Townes splitting [4:342](#)
- auto-discovery [2:76](#), [2:77F](#)
- Auto-Recording Instrumented Environmental Sampler (ARIES) [4:112](#)
- auto-stereoscopic display [3:45](#)
 - anaglyph [3:45–46](#)
 - E-holography [3:46](#)
 - head mounted display [3:46](#)
 - integral imaging display [3:48](#)
 - LC lens for 2D/3D switchable display [3:48](#)
 - multi-planar [3:46–47](#)
 - polarizing glass [3:45](#)
 - shutter glasses [3:46](#)
 - SMV display [3:48](#)
 - spatial-multiplexed [3:48](#)
 - time-multiplexed [3:47–48](#)
 - volumetric display [3:46](#)
- autocorrelation [2:316](#), [3:251](#), [4:49–50](#)
- autocorrelator [2:316](#)
- autofluorescence [5:86](#), [5:97](#), [5:192](#)
- automatic gain control amplifier (AGC amplifier) [4:6](#)
- automatic power control (APC) [1:369–370](#)
- Autonomous Landing and Hazard Avoidance Technology project (ALHAT project) [4:450–451](#)
- autonomous underwater vehicles (AUVs) [4:110](#)
- autophagosome [3:108](#)
- autophagy, PDT and [3:108](#)
- auxiliary optical system [4:107](#)
- avalanche gain [4:423](#)
- avalanche photodiodes (APDs) [1:271](#), [3:98](#), [4:415](#), [4:416F](#), [4:431](#), [4:446–447](#)
 - comparison of false alarm rate *vs.* detection threshold profiles [4:427F](#)
 - dark current [4:423](#)
 - signal and noise distributions [4:427F](#)
- average DGD [1:257](#)
- average index [2:303–304](#)
- average modulation at optimum phase (AMOP) [3:239](#)
- average successive reflections (ASR) [3:148](#)
- axes and angles [5:47–48](#), [5:47F](#)
- axial Airy disc function [4:355](#)
- axial eye length measurement [4:500](#), [4:500F](#)
- axial shift variance [3:263](#)
- axial strain [1:216](#)

- axially symmetric imaging system [5:14](#), [5:14F](#)
- Axially Symmetric Vertical Alignment (ASVA) [3:61](#)
- Axis Symmetric Micro-Cell (ASM) [3:61](#)
- azaquinolone derivatives [2:346](#)
- azimuth invariant case [5:19](#)
- azimuthal acceptance angles [1:136](#)
- azimuthal circular ring of supercritical illumination [5:169](#)
- azobenzene dyes [4:95–96](#)
- azobenzene-based devices [4:114](#)
- B**
- Babinet and Soleil compensators [1:156–157](#), [1:158F](#), [1:159F](#)
- back focal plane (BFP) [5:168](#), [5:173](#)
- back reflection [1:337–338](#), [1:338F](#)
- back-end-of-line (BEOL) [4:219](#)
- back-projection operation [3:194](#), [3:195F](#)
- back-propagating conjugate wave [3:207](#)
- back-scattered Raman radiation [4:355](#)
- backchannel etched structure (BCE structure) [3:13](#)
- backlight module
- light sources [3:27–28](#)
 - optical films [3:28](#)
 - QD film [3:28–29](#)
- backlight reflector film [3:28](#)
- backscattering method [1:259](#), [1:309–311](#), [1:310F](#)
- backward phase conjugator [3:205](#)
- backward wave oscillators (BWO) [1:391](#)
- bacteriochlorophylls (BChl) [2:191–192](#), [2:191F](#)
- bacteriorhodopsin [4:93](#), [4:94F](#)
- balanced detector (BD) [1:295](#)
- ballast chamber [2:374](#)
- ballistic regime, imaging in [3:222–223](#)
- ‘ballistic’ photons [3:219](#)
- shallow tissue imaging [3:222–223](#)
- Balmer series of hydrogen [2:46–47](#)
- band filling effects [2:464](#)
- band gap materials [2:40](#)
- band structure [1:120](#)
- bloch states [2:87](#)
 - density of states [2:90](#)
 - effective mass [2:89–90](#)
 - electron–photon interaction [2:90–91](#)
 - $k \cdot p$ theory [2:87–88](#)
 - NGSs [2:88–89](#)
 - quantum wells [2:91](#)
 - selection rules for intersub-band optical transitions [2:91–92](#)
 - three-level model of band structure [2:89F](#)
- for multilayer fibers for external reflection applications [1:211–213](#)
- and optical properties
- three-level model of [2:89F](#)
- band-bending regions [2:104](#)
- band-diagrams [1:233](#)
- band-band processes [2:115](#)
- bandgap [1:169](#), [2:269](#)
- energy [2:33–34](#), [2:351](#)
 - fibers [1:203](#), [1:203F](#)
 - guided fibers [1:203](#)
 - renormalization [2:465](#)
- bandpass filter (BPF) [1:378](#), [4:46](#)
- bandpass wavelength-division multiplexor (BWDW) [1:181–183](#), [1:182F](#)
- bandwidth [1:135–136](#), [1:351–352](#)
- bandwidth-limited frontends [4:261–263](#)
 - requirements [4:230](#)
- bare fiber, connectors with direct alignment of [1:338](#)
- Bargmann transform [4:203](#)
- Bargmann’s discrete representation series [4:203](#)
- barium nitrate [2:266](#)
- barium titanate (BaTiO_3) [4:96](#), [4:324](#)
- barium vapor laser [2:368–369](#)
- barrel distortion [1:120–121](#), [1:121F](#)
- basal cell carcinoma (BCC) [3:114](#), [5:162](#)
- baseband units (BBU) [1:280–281](#)
- basic fibroblast growth factors (bFGF) [3:118–119](#)
- BBU Hotels (BBUH) [1:280–281](#)
- beam coherence
- analysis [4:177](#)
 - state [4:176](#)
- beam divergence [4:358](#)
- beam intensity distribution [4:175](#)
- beam propagation method [1:197](#)
- beam pump-probe techniques [2:82](#)
- beam quality [2:407](#), [4:356–358](#)
- beam scanning confocal configuration [5:194](#)
- beam splitter (BS) [1:376](#), [4:181](#), [4:500](#), [4:507](#), [4:523](#)
- gratings [1:57F](#), [1:58F](#), [1:59](#), [1:60F](#)
- beamsplitters [3:242](#)
- Beer’s law [4:351–352](#), [4:352](#)
- Bell basis [1:482](#)
- bend insensitive fibers [1:195](#)
- bending lens [1:114–116](#), [1:116F](#)
- Benton’s technique [4:104](#)
- benzalkonium chloride-ethylenediaminetetraacetic acid (BKC-EDTA) [5:148](#)
- benzene [5:221](#), [5:223F](#)
- benzocyclobutene derivative (BCB) [5:230–231](#)
- benzophenone (BP) [2:6–7](#)
- Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung (BES Y) [1:405](#)
- Bertrand lens [5:187](#)
- beryllia (BeO) [2:379](#), [2:381](#)
- Bessel functions [5:16](#)
- best spectacle-corrected visual acuity (BCVA) [5:133](#)
- beta-barium borate (BBO) [1:177](#), [1:479](#), [4:366](#)
- β -barium borate crystal (β -BBO crystal) [1:412](#), [1:413–414](#), [2:166](#), [2:296](#), [2:296F](#), [2:299F](#)
- bi-stable switches [1:214](#)
- bias-dependent collection efficiency of primary photocarriers by junction [4:423–425](#)
- bicontinuous interpenetrating active layer morphology [5:261](#)
- bidirectional reflectance distribution function (BRDF) [3:73–74](#), [3:74F](#), [4:477](#)
- biexcitons [2:97](#), [2:195–196](#)
- bifocal spectacles [5:120](#)
- bifunctional crystals [2:399](#)
- up-conversion lasers based on [2:403–404](#)
 - self-frequency doubling laser [2:404](#), [2:404F](#), [2:405T](#)
 - self-sum frequency mixing laser [2:404–405](#), [2:405F](#), [2:405T](#)
- bilayer heterojunction [5:259–260](#)
- bilayer polymer/acceptor structure [5:259](#)
- bilayer-type organic photodiodes [5:321–322](#)
- bilinear operation [4:199](#)
- bilirubinometer [3:92](#)
- binary grating [1:51F](#), [1:52](#)
- binary modulation (OOK) [1:277](#)
- binary phase shift key (BPSK) [1:292](#)
- binocular
- adaptive optics vision simulator [5:120](#)
 - approaches to presbyopia correction [5:125](#)
 - disparity [3:208](#)
 - parallax [3:44–45](#), [3:45](#)
 - summation [5:125](#)
 - vision [5:44](#)
 - visual functions [5:125](#)
- biological tissues [3:95](#), [3:219](#), [3:219F](#)
- biomedical applications [3:224](#)
- of Raman lasers [2:269](#)
- biometric parameters of lens [5:47](#)
- biophotonics [5:39](#)
- biosensing, optical techniques for [3:95](#)
- biperiodic grating [1:52](#), [1:53F](#), [1:59](#)
- bipolarons [5:223](#)
- birefringent/birefringence [1:33](#), [1:141–142](#), [5:192](#)
- diffraction [1:133](#)
 - materials [1:150–151](#)
 - modulator functions [1:146–147](#)
 - in PCF [1:179–180](#)
 - scanners [1:136–137](#), [1:137F](#)
- bis(2-oxoindolin-3-ylidene)-benzodifurandione (BIBDF) [5:326–327](#), [5:326F](#)
- bispectrum algorithms [3:275–276](#)
- bistability [3:79](#)
- bistable LCDs [3:2](#)
- bit interleaved OTDM [4:35–36](#), [4:36F](#)
- bit interleaving *see* bit level
- bit level [4:35](#)
- bit rate product [4:1](#), [4:1F](#)
- bit-error-rate (BER) [1:248–249](#), [1:273](#), [1:276](#), [1:291](#), [1:299](#), [1:364](#), [1:366F](#), [4:31–33](#), [4:428](#)
- OTDM [4:45–47](#)
 - performance [4:46–47](#)
- bit-interleaving process [4:61](#)
- black box approach [4:321](#)
- black box imager [3:157](#)
- black box optical regeneration (BBOR) [1:370](#)

black dye [5:276–277](#), [5:276F](#)
 blazed gratings [1:75–76](#), [1:78–79](#), [1:79F](#)
 bleaching process [4:89](#), [4:89F](#)
 blending materials [5:311](#)
 blind deconvolution techniques [3:276–277](#)
 blind spot [3:17](#)
 Bloch equations [2:83](#)
 Bloch oscillations [2:36](#)
 Bloch states [2:87](#)
 Bloch vector approach [2:249](#)
 Bloch waves [1:170](#)
 Bloch–Floquet theorem [1:169](#)
 blocked impurity band detectors (BIB detectors) [1:26](#), [1:426F](#)
 blocker [4:233–234](#)
 blood flow measurement
 in capillaries [5:65](#)
 in retinal vessels [4:525](#)
 blown fiber [1:334–335](#)
 blue field entoptic phenomenon [5:65–66](#)
 blue InGaN laser diodes [2:271](#)
 blue lasers [2:356](#)
 blue phases (BPs) [3:10](#)
 blue-green dyes [2:345–346](#)
 blur strength [5:110](#)
 Bohr atom [2:99](#)
 Bohr radius [2:41](#), [2:93](#)
 Bohr-like models [2:40](#)
 bolometers [1:419](#)
 metal films [1:427](#)
 Boltzmann distribution [2:265–266](#), [2:372](#)
 Boltzmann kinetic approach [2:83](#), [2:84](#)
 Boltzmann transport equation (BTE) [2:327](#), [3:257](#)
 bond length alternation (BLA) [5:263](#)
 bookkeeping method [4:350](#)
 booster amplifier [1:248](#)
 Born approximation [2:22–23](#), [3:137](#), [3:143](#)
 boron [2:271](#)
 boron oxide (B_2O_3) [1:28](#)
 Bose condensation [1:461](#)
 Bose–Einstein condensate (BEC) [4:278](#), [5:212](#)
 Bosonic analysis stresses [1:461](#)
 Bosonic operators [2:96–97](#)
 bottom-gate back channel etch (BCE) [3:12–13](#)
 bound electrons [2:118](#)
 bound exciton [2:93](#)
 bound-to-continuum designs (BTC designs) [1:382](#)
 box geometry [2:56](#), [2:156](#)
 BOXCARS geometry [2:185](#), [2:185–188](#), [2:186F](#)
 “phasing” 2DES signal [2:188–189](#)
 pulse-shapers [2:188](#)
 2DES in pump–probe geometry [2:189](#)
 Bragg angles [1:133](#)
 Bragg cell [4:98](#)
 Bragg condition [2:356–357](#)
 Bragg configuration [1:139](#)
 Bragg diffraction [1:133](#), [5:207](#), [5:208F](#)
 Bragg fibers [1:263](#)
 Bragg grating [5:313–314](#)
 Bragg imaging filter [3:267](#)
 Bragg pinhole [3:270](#)

Bragg reflection [1:229–230](#), [1:230F](#)
 condition for [1:230](#)
 Bragg reflectors [2:12](#), [2:466](#)
 Bragg regime [4:87](#)
 Bragg scattering [1:203](#), [1:384](#), [5:207](#), [5:210](#), [5:210F](#)
 Bragg slit [3:270](#)
 Bragg spectroscopy [5:213](#)
 Bragg structure [4:266–267](#)
 Bragg wavelength [2:455–456](#)
 brain derived neurotrophic factor (BDNF) [3:123–124](#)
 brain research [3:93–94](#)
 branching devices [1:336](#), [1:339–341](#)
 nonwavelength-selective [1:340–341](#)
 WDM [1:341–342](#)
bremsstrahlung [1:405](#)
 Brewster angle [1:153](#), [1:404](#)
 brief-pulse, direct-detect lidar [4:431](#)
 bright field microscope [5:192](#)
 bright-field imaging [5:198](#)
 brightness enhancement film (BEF) [3:28](#)
 Brillouin frequency shift [5:148](#)
 Brillouin microscopy [5:142](#), [5:148–149](#)
 applications
 in cornea [5:149–150](#), [5:149F](#)
 in crystalline lens [5:150–151](#), [5:150F](#)
 Brillouin modulus and relationship with standard mechanical properties [5:149](#)
 technique [5:148–149](#), [5:148F](#)
 towards clinical Brillouin microscopy system [5:151](#)
 Brillouin scattering [1:180](#), [1:259](#), [3:210](#), [5:140](#)
 Brillouin zone (BZ) [1:170](#), [1:170](#), [1:173](#), [2:36](#), [2:87](#)
 boundary [1:185](#)
 Brillouin-assisted FWM [4:324](#)
 broad bandwidth [2:120](#), [2:121](#)
 broad-area striped AlGaIn laser diodes [2:275–276](#)
 broadband
 chirp stabilization [4:436](#)
 generation [2:299](#)
 light sources [2:178](#)
 resonator spectrometer [1:395–396](#)
 spectroscopic ellipsometer [1:210](#)
 spectroscopy [1:393–395](#)
 THz detection methods [1:414–415](#)
 THz sources [1:403](#)
 accelerated free electrons [1:405](#)
 air-plasma [1:411–412](#)
 nonlinear crystals [1:405–409](#)
 photo-Dember effect [1:404–405](#)
 photo-induced carriers in solid materials [1:403–404](#)
 tunability of laser sources [2:435](#)
 Broadband Passive Optical Access Networks (BPONs) [2:73](#), [2:74](#), [2:75F](#), [2:78F](#), [2:84F](#)
 Broadband Remote Access Servers (BRAS) [1:281](#)
 Bryngdahl method [4:182–183](#)
 buffer gas [2:372](#)
 building block [4:17](#)
 built-in field in semiconductor [1:404](#)

bulk heterojunction (BHJ) [5:259](#), [5:260F](#), [5:303–304](#)
 active materials in polymer solar cells [5:263–265](#)
 ETIs [5:266](#)
 HTIs [5:265–266](#)
 interface and architectural engineering [5:265–266](#)
 inverted OPV device [5:266](#)
 morphology control through process engineering [5:260–261](#)
 polymer solar cell structure [5:260](#)
 SA [5:261–262](#)
 solvent additive [5:262–263](#)
 strategy to improving BHJ device performance [5:260](#)
 tandem OPV devices [5:266–269](#)
 thermal annealing [5:261](#)
 bulk materials [2:107–108](#)
 electron states in [2:99–100](#)
 bulk saturable absorbers, Q switching with [2:419–421](#)
 bulk semiconductors, continuum states in [1:491](#)
 bulk SOAs [1:243–244](#), [1:244F](#), [1:245](#), [1:247–248](#)
 MZI logic gates [1:251](#)
 Burgess variance theorem [4:428](#)

C

c-axis [2:271](#)
 C-band tunable PIC *see* integrated laser
 Mach-Zehnder modulator PIC (ILMZ PIC)
 c-plane [2:271](#)
 cable core [1:330](#)
 cable cut-off wavelength [1:312](#)
 calcite [1:150–151](#), [1:150F](#)
 calcium ion
 lasers [2:373](#)
 recombination lasers [2:372–373](#)
 California Verbal Learning Test (CVLT) [3:124](#)
 cancer
 diagnostics [3:92–93](#)
 PDT for [3:106–107](#)
 adaptive immunity [3:112–114](#)
 anti-vascular effect [3:111](#)
 cellular mechanisms [3:106–107](#)
 direct destruction of tumors [3:111](#)
 induction of anti-tumor immune response [3:114](#)
 innate immunity [3:111–112](#)
 PS biodistribution [3:109–111](#)
 PS pharmacokinetics [3:108–109](#)
 canonical evolution [4:199](#)
 canonical Fourier transform [4:202](#)
 canonical integral transforms [4:174](#), [4:200–201](#), [4:201](#)
 canonical partition function [2:111](#)
 canonical transformations [4:180](#), [4:199](#)
 complex extension [4:202–203](#)
 cantilever beam [5:190](#)
 capacitance–voltage (C–V) [4:420](#), [4:420F](#)

- capacitive coupling [1:474–475](#)
 capacitive-feedback transimpedance amplifier (CFIA) [4:429](#)
 capillary blood flow
 blue field entoptic phenomenon [5:65–66](#), [5:66F](#)
 erythrocyte imaging
 with AOSLO [5:67–68](#)
 with flood illuminated adaptive optics [5:68–69](#)
 fluorescein angiography [5:66–67](#)
 fluorescently stained leukocytes [5:67](#)
 leukocyte imaging with AOSLO [5:67](#)
 measurement [5:65](#), [5:65F](#)
 capital expenditure (CAPEX) [2:74](#)
 carbon dioxide (CO₂) [1:217](#), [2:327](#), [2:442](#)
 carbon dioxide laser [1:234–235](#), [2:325](#)
 see also metal vapor lasers
 characteristics [2:326](#)
 diffusion cooled laser [2:334](#)
 fast axial flow lasers [2:332–334](#), [2:332F](#)
 fast transverse flow laser [2:334](#)
 laser configuration [2:331](#)
 molecular dynamics [2:326–327](#)
 optical resonator [2:329–331](#), [2:330F](#)
 sealed-off and waveguide lasers [2:331](#), [2:331F](#), [2:332F](#)
 six level model [2:326F](#)
 slow axial flow lasers [2:331–332](#), [2:332F](#)
 TEA pressure [2:334](#)
 carbon disulfide (CS₂) [5:261–262](#)
 carbon materials [5:280](#)
 carbon nanotubes (CNT) [2:72–73](#), [2:308](#), [5:280](#)
 cardinal plane locations [1:105–106](#), [1:105F](#), [1:107F](#)
 cardinal points and planes of optical system [1:101–102](#), [1:102F](#)
 carotenoids (Car) [2:192](#)
 carrier heating (CH) [1:250–251](#)
 carrier phase estimation (CPE) [1:278](#), [1:284](#)
 carrier suppressed RZ (CSRZ) [1:294–295](#), [4:6–7](#)
 carrier to envelope phase (CEP) [2:302](#), [2:302F](#)
 carrier to frequency offset (CEO) [2:302](#), [2:302F](#)
 carrier-envelope-phase (CEP) [2:311](#), [2:313F](#), [2:314F](#)
 stability [2:312–314](#)
 carrier-injection modulator [4:220](#), [4:220F](#)
 carrier-phonon scattering [2:84](#)
 carriers [1:489–490](#)
 density
 coherence control [1:490–491](#), [1:490F](#)
 dependence [2:21–23](#), [2:22F](#)
 injection effect [4:236](#)
 phase compensation and decoding [1:305](#)
 Cartesian coordinates [1:162](#), [5:14](#)
 cascade ionization [5:92](#)
 cascaded amplifiers [1:273](#)
 cascaded multiplication spectrometer [1:391–393](#), [1:393F](#)
 cascaded optical amplifiers, noise in [1:273](#)
 cascaded Raman generation [2:268](#)
 cascaded transformations [4:183–184](#)
 cat conjugator [4:324](#)
 cataphoresis process [2:374](#)
 cathode [5:232](#)
 cathode ray tube (CRT) [3:1](#), [3:2](#)
 color [3:2–3](#), [3:3F](#)
 cathodoluminescence microscopy [5:193](#)
 Cauchy principal value of integral [4:317–318](#)
 causality [4:317–318](#), [4:320–321](#)
 caustic surface [1:114–116](#)
 cavity [1:383–384](#), [1:383F](#), [2:268](#)
 cavity configurations, OPSLs [2:285](#)
 dumping [2:412](#)
 effects [5:302–304](#), [5:304F](#)
 in OLEDs [1:447–448](#)
 experiments [1:455](#)
 lasing [5:313](#)
 length [1:459](#)
 change [2:306](#)
 linewidth equation [2:344](#)
 mode energy [1:460](#)
 quantum electrodynamics [1:479](#)
 effects [1:455](#)
 cavity QED [1:451](#)
 cavity experiments [1:455](#)
 confined space
 energy shift in [1:451–452](#)
 modification of spontaneous transition rate in [1:451](#)
 effects in molecular systems
 cavity effects in OLEDs [1:447–448](#)
 modification of molecular spontaneous emission [1:445–446](#)
 molecules in plasmonic resonators [1:446–447](#)
 strong light-matter coupling [1:448–450](#)
 generation of number states (Fock states) of radiation field [1:454–455](#)
 maser operation [1:452–454](#)
 microlasers [1:455–456](#)
 setup of rubidium Rydberg atoms [1:453F](#)
 one-atom maser or micromaser
 characteristics [1:454F](#)
 trapping states [1:454](#), [1:455F](#)
 CD/DVD pickup head [4:507](#)
 CdTe bulk [2:58](#)
 cell death [5:91–92](#)
 cellular mechanisms of PDT [3:106–107](#)
 and apoptosis [3:107](#)
 and autophagy [3:108](#)
 and necrosis [3:107–108](#)
 cellular mosaics [5:86–87](#)
 analysis [5:100](#)
 cellulite treatment, PBM for [3:120–121](#)
 cemented doublet [1:109](#), [1:110F](#)
 cemented Glan-Thompson prism [1:131](#), [1:131F](#)
 central nervous system (CNS) [5:64](#)
 central offices (COs) [1:280–281](#), [2:73](#), [2:82F](#)
 centrosymmetric crystal [1:142](#)
 centrosymmetric materials [1:489–490](#)
 cerium (Ce) [4:91](#)
 cerium-doped laser host crystals [2:446](#)
 cesium (Cs) [2:361–362](#)
 cesium lithium borate (CLBO) [2:372](#)
 chalcogenide fibers [2:433](#)
 chalcogenide glass family [1:210](#)
 channel spacing [4:68](#)
 channelrhodopsins (ChRs) [3:116](#)
 charge carriers transport materials [5:286–287](#), [5:286F](#)
 charge collection narrowing (CCN) [5:324–325](#)
 charge diffusion [2:7](#)
 charge dynamics of dye-sensitized solar cells [2:168](#)
 charge neutrality [2:350](#)
 charge ordered states (CO states) [2:135](#)
 charge oscillations [2:118](#)
 charge qubits [1:471–472](#)
 charge transfer (CT) [5:223–224](#), [5:246](#), [5:257](#), [5:258](#)
 in OPVs [2:194](#)
 organic chromophores [1:144](#)
 charge-coupled device (CCD) [2:168](#), [3:267](#), [4:106](#), [4:110](#), [4:111](#), [5:314](#)
 detector [5:189](#)
 charge-coupled device/complementary metal-oxide semiconductor imaging technique (CCD/CMOS imaging technique) [4:510](#)
 charge-coupled devices (CCDs) [3:86](#), [3:211–212](#), [3:214F](#), [4:98](#), [4:193](#)
 charge-generation layer (CGL) [3:67](#)
 CHCC process [2:115](#)
 Chebyshev functions [5:16](#)
 chemical sensing [2:262](#)
 chemical vapor deposition (CVD) [1:28–29](#), [2:260](#), [2:280](#), [4:284](#)
 Cherenkov radiation [1:409](#)
 CHHS process [2:115](#)
 $\chi^{(2)}$, $\chi^{(3)}$ cascaded nonlinearity [4:321](#)
 Chilla-Martinez method [4:52](#)
 chiral-nematic LCs [3:10](#)
 chiral-smectic C phase (SmC* phase) [3:10](#)
 chirp [1:142](#), [2:355–356](#)
 modulation [4:499–500](#)
 nonlinearity [4:435–436](#), [4:435F](#)
 rate [4:434–435](#)
 chirped fiber Bragg gratings [4:27–28](#), [4:27F](#)
 chirped gratings [1:239](#)
 chirped mirrors (CMs) [2:314](#)
 chirped pulse amplification (CPA) [1:239](#), [1:239F](#), [2:299](#), [2:311](#), [2:324](#)
 chirped pulse THz spectrometer [1:393–395](#), [1:394F](#)
 chirped RZ (CRZ) [4:6–7](#), [4:8F](#)
 chloranil (CA) [2:7–8](#)
 chlorine [2:382](#)
 chlorobenzene (CB) [5:263](#)
 1-chloronaphthalene (CN) [5:263](#)
 chlorophyll-sensitized zinc oxide electrode [5:270](#)
 cholesteric liquid crystal display (ChLCD) [3:82](#), [3:84](#)
 cholesteric phase [3:10](#)
 choroid [5:130](#)
 choroidal neovascularization (CNV) [5:164](#)
 chromatic aberrations [5:58–59](#), [5:120](#)
 chromatic difference of refraction [5:58](#)

- chromatic dispersion (CD) [1:178–179](#), [1:255](#), [1:256](#), [1:294–295](#), [1:313–316](#), [1:315F](#), [1:363](#), [4:2](#), [4:9](#), [4:10F](#), [4:20](#), [4:25](#)
- control [1:238–239](#)
- management
- corrections to linear dispersion maps [4:25–26](#)
 - dispersion map [4:24–25](#), [4:24F](#), [4:24T](#), [4:25F](#)
 - optical nonlinearities as factors in dispersion compensation [4:22–23](#)
 - monitoring [4:31–33](#), [4:32F](#)
 - in optical fiber communication systems [4:21–22](#), [4:21F](#), [4:22F](#), [4:23F](#)
 - test methods for determination [1:315–316](#)
- chromatic microencapsulated EPD [3:80](#), [3:80F](#)
- chromatic-dispersion compensation [1:265](#)
- chromaticity diagram (CIE 1931) [3:40–41](#)
- chromatin condensation [3:107](#)
- chromium (Cr) [2:413](#), [2:442](#)
- doping [2:413](#)
 - laser advances [2:440–441](#), [2:441T](#)
 - revolution [2:440–441](#)
- chromophore cytochrome c oxidase (CCO) [3:123](#)
- chromophores [1:144](#), [3:115](#)
- chronic mild stress [3:125](#)
- chronocyclic tomography [4:179](#), [4:180](#)
- chronocyclic tomography for optical pulse characterization [4:179](#)
- ciliary body [5:130](#)
- ciliary neuro-trophic factor (CNTF) [5:82](#)
- circle of least confusion [1:114–116](#)
- circular aperture [1:93–94](#)
- diffraction from [1:70–72](#), [1:72F](#)
- circular error at 90th percentile (CE90) [4:402](#)
- circular polarization [3:45](#)
- circulators [1:336](#), [1:343](#)
- cladding [1:322](#)
- center [1:322](#)
 - diameter [1:322](#)
 - of large holes [1:197](#)
 - modes [1:229](#)
 - noncircularity [1:322](#)
- cladding pumped amplifier or laser see double-clad amplifier or laser
- classical analogue holography [4:106](#)
- classical Bohr-like description of electron [2:40](#)
- classical electromagnetism [1:458](#)
- classical motion equation [2:316–317](#), [2:318F](#)
- classical-holography [4:30](#), [4:109](#)
- clinical standard of care [3:91–92](#)
- clock
- multiplexing [4:45](#)
 - in OTDM [4:44–45](#)
- clock data recovery (CDR) [1:278](#), [4:6](#)
- clock recovery (CR) [1:301–302](#), [1:367](#), [1:369](#)
- closed form MIMO [4:408–412](#)
- closed-head injury (CHI) [3:123](#)
- closed-loop
- control methods [5:41–42](#)
 - learning control methods [5:41](#)
 - quantum learning control algorithm [5:35](#)
 - techniques [5:39](#)
- clothes-integratable-type [3:52](#)
- cloud [1:227](#)
- mapping [4:483](#)
- CNOT gate [1:484](#)
- coarse wavelength division multiplexing (CWDM) [1:255](#), [4:18](#), [4:68](#)
- coat-debond method [3:57](#)
- coating [1:322](#)
- code division multiple access techniques (CDMA techniques) [5:6–7](#)
- FPA read-out technique [5:8](#)
- Code vs. ABCD-linear system analysis [4:214](#)
- coexistence [2:78](#)
- cognitive performance enhancement, PBM for [3:126–127](#)
- coherent anti-Stokes Raman scattering (CARS) [2:244–245](#), [2:247](#), [4:315](#), [4:315F](#), [5:193](#)
- coherent anti-stokes Raman spectroscopy see coherent anti-Stokes Raman scattering (CARS)
- coherent detection [1:295–296](#), [4:432](#)
- heterodyne detection [1:296–297](#)
 - homodyne and intradyne detection [1:297](#)
 - phase and polarization diversity receiver [1:298](#)
 - phase diversity receiver [1:297–298](#), [1:298F](#)
 - principles [1:296](#)
 - SF and BER [1:299](#)
 - SNR [1:298–299](#)
- coherent diffraction radiation (CDR) [1:405](#)
- coherent diffractive imaging technique (CDI technique) [3:146](#)
- defocus-based techniques [3:149–150](#)
 - Fourier measurement-based techniques [3:146–148](#), [3:146F](#)
 - transverse translation-based techniques [3:151–153](#)
- coherent edge radiation (CER) [1:405](#)
- Coherent Lidar Airborne Shear Sensor (CLASS) [4:457](#)
- coherent light [1:374](#), [3:96](#)
- emission [2:63](#)
 - trapping approaches [5:294](#)
- coherent light-matter interaction [2:53–56](#)
- double-sided Feynman diagrams [2:54F](#)
 - heterogeneous ensemble of two-level systems [2:56F](#)
 - three-level V-system and rephasing 2D spectrum [2:55F](#)
 - time-ordering for three-pulse FWM [2:54F](#)
- coherent lightwave systems
- coherent detection [1:295–296](#)
 - differential detection [1:295](#)
 - digital coherent receiver [1:299–300](#)
 - optical modulation [1:291–293](#)
- coherent mode-locked lasers [4:432](#)
- dual-comb interferometry based distance measurement [4:432–434](#)
 - mode-locked laser based multi-wavelength interferometry [4:432](#)
- coherent multidimensional spectroscopy
- theory [2:150–151](#)
 - double-sided FDs [2:151–152](#)
 - Fourier transform spectroscopy [2:151](#)
 - theoretical signal origin [2:152–153](#)
- coherent population oscillations (CPOs) [1:252–253](#)
- coherent ray tracing method (CRT method) [4:123](#)
- coherent synchrotron radiation (CSR) [1:405](#)
- coherent terahertz sources [2:121–122](#)
- coherent THz radiation [2:118](#)
 - free electron THz laser [2:119](#)
 - gas THz laser [2:118–119](#)
 - optical rectification [2:121](#)
 - semiconductor THz laser [2:119–120](#)
 - THz PCA [2:120–121](#)
- coherent transients [4:264](#), [4:264–265](#)
- coherent control techniques [4:269](#)
- coherent photocurrents [4:269–270](#), [4:272F](#)
- destructive interference [4:267](#)
- multidimensional fourier transform spectroscopy [4:275–276](#)
- optical Bloch equations [4:265](#)
- photon echo [4:272–275](#), [4:275F](#)
- quantum beats [4:267–269](#)
- radiative decay [4:266](#)
- semiconductor Bloch equations [4:265–266](#)
- spectral oscillations [4:271–272](#), [4:274F](#)
- superradiance [4:266–267](#)
- transient absorption changes [4:270–271](#), [4:272F](#), [4:273F](#)
- coherent transition radiation (CTR) [1:405](#)
- coherent/coherence [1:34](#), [2:192–193](#), [2:194F](#), [2:198–199](#), [2:199T](#)
- area of radiation [1:35–36](#)
 - beam [2:247](#)
 - coherent features, Rydberg excitons [2:47–49](#)
 - coherent illumination, PSF for [3:263](#)
 - correlation-induced spectral changes [1:45–46](#)
 - detector [5:190](#)
 - dynamics [2:82–83](#)
 - effects [4:264](#)
 - energy transfer [2:59–61](#)
 - excitons [2:37–38](#)
 - higher-order coherence [1:48–49](#)
 - homogeneity [4:176](#)
 - Innova™ ion laser [2:381](#)
 - length [1:34](#)
 - mathematical description [1:36](#)
 - analytical field representation [1:36](#)
 - degree of coherence and visibility of interference fringes [1:37](#)
 - mutual coherence [1:36–37](#)
 - spectral representation of mutual coherence [1:38](#)
 - temporal and spatial coherence [1:37–38](#)
- mode decomposition model [3:148](#)
- non-linear optical process [2:164](#)
- optical
- energy [2:336](#)
 - signals [3:167](#)
 - spectroscopy [2:52](#)
- oscillations [4:272](#)
- parameter [5:187](#)
- photocurrents [4:269–270](#), [4:272F](#)
- polarization sources [2:118](#)
- properties [3:222–223](#)
- radiation [2:118](#)

- regime [2:82](#)
 self-imaging *see* Talbot effect
 spectroscopy [4:264](#)
 systems [1:278](#)
 time of radiation [1:34](#)
 types of fields [1:40–41](#)
 cross-spectrally pure fields [1:41–42](#)
 perfectly coherent fields [1:41](#)
 quasi-monochromatic fields [1:41](#)
 types of sources [1:42–43](#)
 equivalence theorems [1:45](#)
 primary sources [1:42–43](#)
 quasi-homogeneous sources [1:44–45, 1:44F](#)
 Schell-model sources [1:43–44](#)
 secondary sources [1:43](#)
 volume [1:36](#)
 waves [4:189](#)
- coherent/coherence control [1:487, 5:39, 5:40](#)
 of carrier density, spin population, and spin current [1:490–491, 1:490F](#)
 of electrical current
 using single color beams [1:491](#)
 using two color beams [1:488–490, 1:489F](#)
 experimental setup to measure steady-state voltage [1:489F](#)
 of quantum systems [2:54](#)
 in semiconductors [1:487–490](#)
 techniques [4:269](#)
- cold cathode fluorescence lamps (CCFLs) [3:25, 3:27–28](#)
- collagen [3:118–119, 5:104–105](#)
- collection-mode NSOM [5:190–191, 5:190F](#)
- colliding-pulse mode locking (CPM) [2:307–308, 2:337, 2:340](#)
- collimator lens (CL) [4:500](#)
- collinear devices [1:139](#)
- collinear multidimensional THz spectroscopy [2:197–199](#)
- collision [1:20–21, 5:40](#)
- collision-induced vibrational relaxation [2:328](#)
- colloidal self-assembly [1:175](#)
- Cologne Database for Molecular Spectroscopy (CDMS) [1:401](#)
- color encoding of OCT data [4:506, 4:506F](#)
- color filter [3:26](#)
- color formation of LCD—spatial and temporal
 CMFs and tristimulus [3:17](#)
 colors in LCDs [3:17–20](#)
 dynamic color breakup [3:21–23](#)
 LPD [3:23–24](#)
 spatial color synthesis [3:19–20](#)
 static color breakup [3:21](#)
 stencil-FSC [3:22–23](#)
 structure of human eye [3:17](#)
 temporal color synthesis [3:20–21](#)
- color matching functions (CMFs) [3:17, 3:18F, 3:19F](#)
- colorless
 directionless ROADMs [1:269](#)
 ROADM [4:227](#)
- colorless directionless (CD) [4:227](#)
- colorless-directionless-contentionless (CDC) [4:227](#)
- colossal magnetoresistance (CMR) [2:125](#)
- colour [3:71, 4:107](#)
 gamut and HDR [3:40–43, 3:42F](#)
 holography [4:102, 4:103, 4:103–104](#)
 pseudo colour and manipulative artist [4:104–105](#)
 red, green, blue [4:103–104](#)
 in liquid crystal displays [3:17–20](#)
 temperature [3:75–76](#)
 triangle [3:76, 3:77F](#)
- columnar mesophases [3:10–11](#)
- coma [1:116–118, 1:118F, 1:119F](#)
- comb function [4:385](#)
- commercial Hartmann–Shack aberrometers [4:530–531](#)
- commercial iTrace visual function analyzer [4:532–533, 4:532F](#)
- commercial OCT instruments for
 ophthalmology [4:505–506, 4:506F](#)
 color encoding of OCT data [4:506, 4:506F](#)
 multiple reference OCT for dermoscopy [4:507](#)
 OCT in cerebrovascular and cardiovascular research [4:506–507](#)
- Commission Internationale de l'Éclairage (CIE) [3:17, 3:71](#)
- common-gate amplifier [4:260, 4:260F](#)
- communication
 components [4:3–6](#)
 mechanisms [4:13](#)
 OTDM [4:66F](#)
 demultiplexing of OTDM signal at receiver [4:63–65](#)
 transmission of OTDM signal over fiber [4:63](#)
 ultra-short optical pulse sources [4:62–63](#)
- commutators [4:199](#)
- compensators [1:149, 1:156–159](#)
- complementary metal-oxide-semiconductor (CMOS) [1:469, 2:168, 3:86, 4:98, 4:99, 4:106, 4:110, 4:193](#)
- cameras [3:267, 4:431](#)
- chips [4:140](#)
- CMOS-based detectors [4:422–423, 1:424F](#)
- process technology [1:419](#)
- sensor [4:112](#)
- transceiver circuits for optical interconnects
 optical interconnects [4:254](#)
 receiver circuitry [4:259–261](#)
 transmitter circuitry [4:254–255](#)
- completeness [4:401](#)
 in 3D [4:404–405](#)
- complex coherence factor [4:174](#)
- complex degree of coherence [1:37](#)
- complex DSP algorithm [2:83, 2:83–84](#)
- complex field amplitude [4:176–177](#)
- complex IV *see* cytochrome C oxidase (Cox)
- complex quantities [4:350](#)
- complex refractive index profile [4:291](#)
- complex-valued distributions [1:458](#)
- composite cavities [2:416](#)
 microchip lasers [2:416](#)
- compound lenses [1:109](#)
- compound microscope [5:185](#)
- compressed gases [4:322](#)
- Compton scattering [4:354](#)
- computational imaging [3:136–137, 3:136F, 3:262, 3:278](#)
- computed tomography (CT) [3:156, 3:157, 3:192, 3:256, 4:174](#)
- computer generated hologram (CGH) [4:31, 4:120–122](#)
 see also digital holography
 with 64 x64 cells [4:32F](#)
 algorithms [4:33–135](#)
 from classical hologram to [4:30–131](#)
 from diffraction grating to Fourier [4:31–133, 4:31F](#)
 digital holography and [4:113–114](#)
 error reduction in [4:124](#)
 hardware aspects [4:35](#)
 ninth-order vortex [4:32F](#)
 optical filtering [4:35–136](#)
 optical vortices [4:36–137](#)
 polarization [4:37, 4:37F, 4:38F](#)
 terminology [4:37–138](#)
- computer guided treatments [5:138](#)
- computer holography [4:30](#)
- computer-based algorithms [4:108](#)
- computerized corneal topography [5:106](#)
- concentric rings [1:197](#)
- condensed matter systems [2:213](#)
- condenser lens [5:189](#)
- condition number [3:158](#)
- conducting polymers [5:279–280](#)
- conducting substrates [5:273](#)
- conduction band (CB) [1:171, 1:467, 2:33–34, 2:115, 2:116F, 4:304](#)
- conduction band minimum (CBM) [5:256, 5:286–287](#)
- cone-effect [3:202](#)
- cones [3:17, 5:43](#)
- confined space
 energy shift in confined space [1:451–452](#)
 modification of spontaneous transition rate in [1:451](#)
- confocal AOSLO imaging [5:74–75, 5:74F](#)
- confocal arrangement [5:190–191](#)
- confocal diaphragms [4:355](#)
- confocal fluorescence microscopy [5:198](#)
- confocal laser scanning microscopy (CLSM) [5:98](#)
- confocal microscopy [3:270, 4:194, 5:97–98, 5:99F, 5:189–190, 5:189F, 5:194, 5:194F](#)
 applications of depth discrimination [5:196–198](#)
 of cornea [5:99–100, 5:100](#)
 in corneal pathologies [5:100–101, 5:101F](#)
 fluorescence microscopy [5:198–201](#)
 HRT-RCM [5:100, 5:100F](#)
 image formation in scanning microscopes [5:194–196](#)
 structured illumination to optical sectioning [5:201–202](#)
- confocal optical systems [5:195–196](#)
- confocal single-photon microscope [5:90](#)
- confocal system [5:185](#)
- conjugated hydrocarbons [2:345](#)
- conjugated plane (CPL) [3:202, 3:203F](#)

- conjugated polymer-based solar cells
[5:256](#)
 BI-II morphology control through process engineering [5:260–261](#)
 characterization of OSC [5:258–259](#)
 OPV device structures [5:259–260](#)
 principles of OSC [5:256–258](#)
- conjugated polymers [5:220](#), [5:309](#), [5:309F](#), [5:310F](#)
- conjugated wave [3:205](#)
- conjugated π -electron organic materials
[1:142](#)
- connectors with alignment by means ferrule
[1:338](#)
- connexin [43](#) ($C \times 43$) [3:118–119](#)
- conservation of momentum [1:133](#)
- consolidation lowers capital expenditure (CAPEX) [2:82](#)
- constant bit-rate (CBR) [1:287](#)
- constant lens index [5:48](#)
- constant modulus algorithm (CMA) [1:278](#), [1:303–304](#)
- constructive interference [3:244](#)
- constructive quantum wave interferences
[5:30](#)
- contact hyperfine interaction [1:473–474](#)
- contact lenses [5:116–119](#), [5:118F](#)
 material [5:126–127](#)
- contact recording technique [4:103](#)
- contentionless, ROADM [4:227](#)
- continuity condition [4:183](#)
- continuous HIO (CHOI) [3:148](#)
- continuous part of phase function [4:187](#)
- continuous phase functions [4:182](#), [4:188](#)
- continuous sensing [3:87](#)
- continuous variables [1:478](#)
- continuous wave (CW) [1:20–21](#), [1:234–235](#), [1:249–250](#), [1:352](#), [2:119](#), [2:276–277](#), [2:283](#), [2:325–326](#), [2:334](#), [2:336](#), [2:377–378](#), [2:411](#), [2:415](#), [2:440–441](#), [2:462](#), [4:108–109](#), [4:254](#), [4:257](#), [4:347](#), [5:1](#), [5:41](#), [5:212](#)
- amplifiers [4:434](#)
 frequency-converted [2:411–412](#)
 lasers [2:447](#), [2:451](#), [4:110–111](#), [4:282](#), [4:302](#), [5:91–92](#)
 diode lasers [5:3](#)
 dye lasers [2:337](#), [2:339–340](#)
- LIOT [4:302](#)
- light [2:82](#)
 radiation [4:496](#)
- metal ion lasers [2:368](#), [2:373–374](#)
- mode [1:382](#)
- operation [1:454](#)
- phase difference mode [4:497–500](#)
 CW modulated signals [4:497–500](#)
 frequency difference techniques [4:500](#)
- pumped supercontinua [2:431–432](#)
- Raman fiber lasers [2:268](#)
- signal [1:367](#)
- single-frequency [2:411](#)
 source [4:421](#)
- continuous-director-rotation mode (CDR mode) [3:10](#)
- continuous-time linear amplifier (CTLF)
[4:263](#)
- continuous-to-discrete transformation in image gathering [3:175](#)
- continuous-wave frequency modulation (CWFM) [4:499–500](#)
- contrast ratio [3:72](#)
- contrast sensitivity [5:120](#)
- contrast transfer function (CTF) [3:151](#), [4:166](#), [4:403](#)
- control field-system cooperativity principle
[5:33](#)
- control knobs for frequency comb
[2:305–306](#), [2:307F](#)
 cavity length change [2:306](#)
 comb control with intracavity etalon [2:306–307](#)
 locking repetition rate [2:307–308](#)
 repetition rate and group velocity [2:306](#)
- control laser technology [5:34](#)
- control mechanism [1:471–472](#), [1:471F](#)
 charge qubits [1:471–472](#)
 spin qubits [1:472–473](#)
- control system [3:202](#)
- controllers [4:222](#)
- controlling polarization of high-order harmonics [2:237](#)
- conventional “fundamental” resolution
[3:137–138](#)
- conventional 1-D DFB gratings [5:314](#)
- conventional band (C-band) [1:357](#)
- conventional broad-area weakly-guiding single-mode lasers [4:298](#)
- conventional cold cathode fluorescence lamps (CCFLs) [3:19–20](#), [3:20F](#)
- conventional dielectric waveguides [1:383](#)
- conventional microscope [5:185–186](#), [5:189](#)
see also probe microscopes
 Abbe theory of microscope imaging [5:187–188](#)
 depth of focus [5:188](#)
 illumination [5:186–187](#)
 image formation [5:187](#)
 microscope objectives [5:188–189](#)
 resolution [5:185–186](#), [5:186F](#)
- conventional noninterferometric imaging methods [3:241](#)
- conventional optical imaging [3:262](#)
 transmission systems [1:248–249](#)
- conventional semiconductor device processing method [2:351](#)
 processing methods [2:351–352](#)
- conventional single-frequency ring cavity fiber laser [2:455](#)
- conventional single-photon microscope [5:90](#)
- conventional wide-field optical microscopes
[5:195](#)
- convergence [3:34](#)
- conversion efficiency (η) [1:250–251](#)
- “convex hull” method [4:488](#)
- convolution approach, numerical reconstruction by [4:144F](#), [4:145–146](#)
- convolution theorem [3:139–140](#)
- cool disk technology [2:381](#)
- cooled silicon bolometers [1:427](#)
- “Cooper pairs” [2:126](#)
- cooperative plasmonic effect [5:298–300](#)
- cooperativity principle [5:33–34](#)
- coordinate transformations [4:182–183](#)
 Bryngdahl method [4:182–183](#)
 continuity condition [4:183](#)
- copper [2:256](#)
 twisted-pair [1:280–281](#)
- copper bromide laser (CuBr laser) [2:370](#)
 tube construction [2:370F](#)
- copper phthalocyanine (CuPc) [5:259–260](#), [5:321–322](#), [5:321F](#), [5:322](#)
- copper vapor lasers (CVLs) [2:369–372](#), [2:370](#), [2:371–372](#)
 copper bromide laser tube construction [2:370F](#)
 excitation circuits [2:371F](#)
 moderate peak power of [2:372](#)
 partial energy level scheme for [2:369F](#)
 pumped dye laser system [2:338](#), [2:339F](#)
 tube construction [2:370F](#)
- core concentricity error [1:322](#)
- core noncircularity [1:322](#)
- core router growth [4:230](#)
- core-cladding index difference [1:9](#)
- cornea [4:500](#), [5:44–45](#), [5:60](#), [5:130](#)
 anatomical structure [5:97](#), [5:97F](#)
 confocal imaging [5:100](#)
 confocal microscopy [5:99–100](#)
 innervation after refractive surgery [5:102F](#)
 Langerhans cells [5:101](#), [5:101F](#)
 laser and light interaction [5:133–134](#)
 multiphoton microscopy [5:99–100](#)
 second harmonic imaging [5:104–105](#)
 two-photon imaging [5:102–103](#)
 two-photon laser scanning microscopy [5:102](#)
- corneal biomechanics [5:148](#)
- corneal birefringence [4:523](#)
- corneal collagen crosslinking [5:126](#)
- corneal cross-linking (CXL) [5:106](#), [5:145](#)
 for corneal infections [5:136–138](#)
 for refractive procedures [5:136](#)
 two-photon imaging [5:105–106](#)
- corneal deformation parameters [5:143–144](#)
- corneal disease [5:97](#)
- corneal endothelium [5:100](#)
- corneal refractive ablations [4:533](#)
- corneal refractive surgery [5:140](#)
- corneal stroma [5:100](#)
- corneal tissue [5:142](#)
- corneal videokeratography [5:140](#)
- corneo-scleral lenses [5:119](#)
- Cornu spiral [1:88–89](#), [1:88F](#)
- correlated color temperature (CCT) [3:75](#)
- correlated electron materials (CEM)
[2:123](#)
 classes [2:124–126](#)
 HF materials [2:129–130](#)
 time-integrated optical spectroscopy [2:130–131](#)
 time-resolved optical spectroscopy [2:132–135](#)
 transition metal oxides [2:124–126](#)
 typical phase diagram for hypothetical CEM [2:123F](#)
- correlated electron systems, hidden phases in [2:212–214](#), [2:213F](#)

- correlation [3:249](#)
 coefficient [4:491F](#), [4:493](#)
 correlation-induced spectral changes
 [1:45–46](#)
 experimental confirmations [1:46](#)
 scaling law [1:45–46](#)
 in Young's interference experiment [1:46](#)
 cortex velocity map [4:526](#), [4:527F](#)
 Corvis processing software [5:143](#)
 cosmetic applications, PBM for
 for fat reduction and cellulite treatment
 [3:120–121](#)
 for hair regrowth [3:119–120](#)
 skin rejuvenation [3:118–119](#), [3:118F](#)
 Coulomb attraction [2:93](#)
 Coulomb force [1:458–459](#), [2:40](#)
 Coulomb interaction [1:460](#), [1:461](#), [2:34](#),
 [2:83](#), [2:394](#), [2:465](#), [4:265–266](#)
 Coumarin [153](#) [2:345–346](#)
 Coumarin [307](#) [2:345–346](#)
 Coumarin [503](#) *see* Coumarin [307](#)
 Coumarin 540A *see* Coumarin [153](#)
 coumarins [2:345–346](#)
 counter electrodes (CE) [5:270](#), [5:279–280](#),
 [5:280F](#), [5:280T](#)
 coupled mode theory (CMT) [1:231–232](#)
 coupled-wave equations [4:329–330](#), [4:329F](#)
 coupled-wave electro-optically Q-switched
 microchip laser [2:419](#), [2:419F](#)
 coupler devices [1:341](#)
 couplers [1:339–340](#)
 coupling [5:40–41](#)
 light and sound [3:226](#)
 mechanisms [1:475](#)
 optics [3:97–98](#)
 course integral holographic display system
 (CIH system) [4:118–119](#), [4:119F](#)
 course-WDM (CWDM) [4:8](#)
 CR-calculus [3:150](#)
 Cramer–Rao bound (CRB) [3:137](#)
 creatine kinase (CK) [3:126](#)
 critical angle [1:124–125](#)
 cross dichroic prism (X-prism) [3:33](#)
 cross linking of corneal collagen (CXL)
 [5:131](#), [5:133](#), [5:136](#)
 cross peaks [2:164](#), [2:165–166](#)
 cross spectrum algorithms *see*
 Knox–Thompson algorithms
 cross-correlation technique [2:19–20](#)
 cross-coupling of photons in doped PCs
 [2:15–18](#)
 cross-gain modulation (XGM) [1:249–250](#),
 [1:251](#)
 cross-linking [5:145–146](#)
 cross-phase modulation (CPM) [1:16](#),
 [1:19–21](#), [1:21F](#), [1:22F](#), [1:180](#), [1:181](#),
 [1:251](#), [1:320](#), [1:321](#), [1:348](#), [1:364](#),
 [1:367–368](#), [4:3](#), [4:20–21](#), [4:23](#), [4:34](#)
 cross-spectral density [4:174](#), [4:180](#)
 cross-spectrally pure fields [1:41–42](#), [1:41–42](#)
 crossbar fabric [4:234](#)
 crossed grating geometry [2:9](#)
 crossover resonances [2:230](#)
 crosstalk attack [1:272](#)
 cryogenically cooled SBDs [1:421–422](#)
 crystal(s) [2:265](#), [4:328–329](#)
 detectors [1:422](#)
 field theory [2:436–437](#), [2:436F](#)
 flakes [4:286](#)
 laurice [2:102–103](#)
 optic axis [1:138–139](#)
 orientation [1:488](#)
 second harmonic generation in [2:207–209](#)
 SHG in [1:12–16](#), [1:13F](#)
 Crystallens [5:121](#)
 crystalline Raman lasers, advancements in
 [2:266–267](#)
 crystalline semiconductors [1:488](#), [2:33–34](#)
 CT-scanners [3:192–193](#)
 CT26 [3:114](#)
 cubic perovskite crystal structure [5:282](#),
 [5:282F](#)
 cuprous oxide (Cu₂O) [2:40–41](#), [2:45–46](#)
 basic symmetries [2:46T](#)
 Rydberg excitons [2:41](#), [2:42F](#), [2:43F](#)
 current lidar metric standards [4:401](#)
 absolute and relative horizontal accuracy
 [4:402](#)
 absolute and relative vertical accuracy
 [4:401–402](#)
 density and completeness [4:401](#)
 current nanofabrication technology [1:185](#)
 current–voltage (I–V) [4:416–417](#), [4:420](#),
 [4:420F](#)
 curvature sensing wavefront sensors
 [3:277–278](#)
 cushion layers [1:332](#)
 cut-back method [1:309](#)
 cut-off wavelength [1:311–312](#)
 Cut-off-Shifted single-mode Fibers (CSF)
 [1:261](#), [1:262](#)
 cutoff-smoothing [2:14](#)
 cyanamide [1:390](#)
 cyanines [2:344](#), [2:345F](#)
 cyclic adenosine monophosphate levels
 (cAMP levels) [3:115–116](#)
 cyclic boundary conditions [2:99](#)
 cyclic prefix (CP) [2:83](#)
 cyclic pulsed metal vapor lasers [2:368](#)
 cyclooxygenase inhibitor indomethacin [3:111](#)
 cyclotron energy [2:96](#)
 cyclotron resonance (CR) [2:75](#)
 cylinder function [4:385](#)
 cylindrical lens [4:175–176](#)
 cytochrome C oxidase (Cox) [3:115](#)
 Czokhralski method [2:404](#)
- ## D
- D-dimensional system [4:199](#)
 'd-effective' coefficient [4:350](#)
 D-ILA display device [3:37](#)
 d-shell electrons [2:436](#)
 D_{3h} symmetry point-group [4:284](#)
 D_{6h} symmetry point-group [4:284](#)
 damage associated molecular patterns
 (DAMPs) [3:108–109](#), [3:112](#)
 damping rate of relaxation oscillations [4:69](#)
 Daniel Palanker Department of
 Ophthalmology [5:131–132](#)
 dark current [4:428](#)
 dark fiber [4:31](#)
 dark states [4:339–340](#)
 dark-field AOSLO [5:76–77](#), [5:76F](#)
 darkroom principle [3:247](#), [3:247F](#)
 data acquisition system (DAS) [5:6–7](#)
 data center interconnects, WDM in
 [1:281–282](#)
 data communications [1:226–227](#)
 PIC [4:216](#)
 data modulation [4:21](#)
 formats [4:6–7](#)
 data multiplexing [4:7–8](#)
 data processing algorithm for optoelectronic
 sensor [3:89–91](#)
 quasi-spectral analysis [3:91](#)
 ratiometric analysis [3:90–91](#)
 data quality measures (DQM) [4:403–404](#)
 data rate [3:182](#)
 data retrieval [4:35](#)
 data synchronization in OTDM [4:44–45](#)
 data voids [4:401](#)
 DBPSK [1:275](#)
 "DC apodized" Gaussian profile [1:233](#)
 DCI-P3 [3:40–41](#)
 de Broglie wavelength [2:98–99](#), [2:99](#),
 [5:203](#)
 Debye–Sears diffraction [1:131–132](#)
 Debye–Sears effect [1:133](#)
 decision feedback equalizer (DFE) [4:263](#)
 decision flip-flop (DFF) [1:366–367](#)
 decoherence [5:40–41](#)
 effects [1:487](#)
 decoherence free subspaces (DFS) [1:472](#)
 decoherence free subsystems *see* decoherence
 free subspaces (DFS)
 deconvolution from wavefront sensing
 (DWFS) [3:277–278](#)
 deconvolution processing by optical
 methods [3:251](#)
 deep anterior lamellar keratoplasty (DAALK)
 [5:149–150](#)
 deep ultraviolet (DUV) [2:446](#)
 deep-ridge waveguide [4:246–247](#), [4:247F](#)
 defect-induced damage [4:305–306](#), [4:306](#)
 defects [1:203](#), [1:214](#)
 case [2:112](#)
 and dislocations [2:103–104](#)
 states [1:488](#)
 deflection angle [1:147](#)
 defocus-based techniques [3:149–150](#), [3:149F](#)
 direct-inversion methods [3:151](#)
 iterative algorithms [3:150](#)
 recursive algorithms [3:150–151](#)
 degeneracy [2:290](#), [4:298–299](#)
 degenerate four wave mixing spectroscopy
 (DFWM spectroscopy) [1:349](#), [2:248](#)
 and phase conjugation [4:313–314](#)
 degradation of OLEDs [5:238](#)
 degree, ROADMs [4:227](#)
 degree of coherence [1:37](#)
 degree-4 ROADMs [1:269](#)
 degrees of freedom (DOF) [2:123](#)
 delay [4:53–54](#)
 lines [1:193](#)
 delay interferometers (DIs) [1:295](#)
 delayed fluorescence (DF) [5:245](#)

- delayed match-to-sample memory task (DMS memory task) [3:126](#)
 delayed self-heterodyne detection [2:460–461](#), [2:460F](#)
 δ -function band shape [2:440](#)
 demultiplexers [4:39–41](#)
 M–Z interferometers [4:40–41](#), [4:40F](#)
 Sagnac interferometers [4:41–42](#)
 demultiplexing [4:35](#)
 of OTDM signal at receiver [4:63–65](#), [4:64F](#)
 dendritic cells (DCs) [3:112](#), [3:113](#)
 Denisyuk's technique [4:103](#)
 dense wavelength division multiplexing (DWDM) [1:255](#), [2:81](#), [2:391](#), [4:6](#), [4:17–18](#), [4:17](#), [4:68](#), [4:221–222](#)
 optical fiber communication systems [1:25](#)
 transceivers [4:68](#)
 density [4:401](#)
 density operator formalism [4:180](#)
 phase grating [2:6](#), [2:7](#)
 of states function [2:90](#), [2:102–103](#), [2:103F](#), [2:109](#)
 in 3D [4:404–405](#)
 density functional calculations (DFT) [1:144–146](#), [1:145F](#)
 density matrix [1:164](#), [1:478](#), [2:153](#), [4:174](#)
 approach [4:339](#)
 formalism [2:56](#)
 operator [4:180](#)
 density of modes (DOM) [2:12](#)
 density recovery profile algorithm (DRP algorithm) [5:79](#)
 dentate gyrus (DG) [3:123–124](#)
 Department of Defense (DoD) [5:3](#)
 depletion layer [1:404](#)
 depletion-mode modulators [4:258](#)
 depolarization [4:369](#)
 deposition schemes [2:258](#)
 depth discrimination applications [5:196–198](#), [5:197F](#)
 depth of focus (DOF) [4:34](#), [5:158](#), [5:188](#)
 depth perception [3:44](#)
 derivatives of phase function [4:186](#)
 dermoscopy, MR-OCT system for [4:507](#)
 Descemet's stripping endothelial keratoplasty (DSEK) [5:149–150](#)
 destructive interference [3:244](#), [4:267](#)
 destructive quantum wave interferences [5:30](#)
 detected laser power [4:475](#)
 detection
 of control action [5:34](#)
 systems [3:98](#)
 detector [1:224](#), [3:187](#), [3:259](#)
 elements [4:417–418](#)
 sensitivity [1:419](#)
 detour phase [4:32](#)
 detuning [4:243](#)
 deuteroporphyrin (DP) [3:104](#)
 Deutsches Elektronen Synchrotron (DESY) [1:405](#)
 dexter energy transfer [5:237F](#), [5:238](#)
 dextrose [1:160](#)
 DFB based chirp stabilization, [80](#) GHz [4:436–438](#)
 diabetic macular edema (DME) [5:138](#)
 diabetic retinopathy (DR) [5:61](#)
 Diabetic Retinopathy Vitrectomy Study Research Group trial [5:134](#)
 diabolic point (DP) [4:299](#)
 diagonal matrix [3:157](#), [3:159](#)
 diamagnetic term [2:65](#)
 diameter measurement [1:32](#)
 diamond-like carbon coated emitters (DLC coated emitters) [3:1](#)
 diamonds [2:267](#), [2:271](#)
 Raman lasers [2:267](#)
 diatomic molecular systems [5:41](#)
 1,4-diazobicyclo[2.2.2]octane (DABCO) [2:347](#)
 dichroic absorbers [1:151–153](#), [1:154](#)
 dichroic materials [1:151–153](#)
 dichromated gelatin (DCG) [4:90](#)
 die-attached material, die-level chip
 assembled on fabricated board with [3:89](#)
 die-level chip assembled on fabricated board with die-attached material [3:89](#)
 dielectric
 bands [1:171](#), [1:185](#)
 breakdown [4:303–304](#), [4:304](#)
 cavity mirrors [2:415](#)
 composition [4:13–14](#)
 material [1:185](#)
 media [1:346](#), [1:346F](#)
 mirror fibers [1:213](#)
 multi-layer mirrors [2:12](#)
 difference frequency generation (DFG) [1:385](#), [1:406](#), [2:121–122](#), [2:166](#), [2:208](#), [2:280](#), [2:291](#), [2:314](#)
 OPSLs [2:287](#)
 different map (DM) [3:148](#)
 differential absorption glucose content measurement [4:522](#), [4:523F](#)
 differential absorption lidar (DIAL) [4:463](#), [4:467F](#), [4:469](#)
 differential dispersion parameter *see* dispersion slope
 differential external quantum efficiency (DEQE) [2:274–275](#)
 differential group delay (DGD) [1:256](#), [1:257F](#), [1:302–303](#), [1:316](#), [4:2–3](#), [4:226](#)
 differential mode delay (DMD) [1:313](#)
 differential operation mode (DOM) [1:368](#)
 differential PSK (DPSK) [1:293](#)
 differential quadrature DPSK (DQPSK) [1:293](#), [1:294](#)
 differential quantum efficiency [2:467](#)
 differentiation filter [4:35](#), [4:36F](#)
 differentiation output [4:35](#), [4:36F](#)
 diffracted intensity [2:3](#)
 diffracted surface [1:136](#)
 diffracting aperture [1:89](#)
 diffraction [1:62](#), [1:65F](#), [3:264–265](#)
 angle [5:205](#)
 calculation of 3D PSF [3:264–265](#), [3:266F](#)
 from circular aperture [1:70–72](#)
 diffraction intensity [3:148](#)
 diffraction-based scanning 3D colour video holographic display [4:119](#), [4:120F](#)
 diffraction-propagation from spherical emitter to spherical receiver [4:157](#)
 effects [1:10–11](#), [1:131](#), [3:205–206](#)
 efficiency [1:134–135](#)
 formulae [1:38–39](#)
 Kirchhoff theory [1:62–67](#)
 of light, theory [1:133–134](#)
 limit [3:136–137](#)
 of imaging system [3:264–265](#), [3:272](#)
 by rectangular aperture [1:69–70](#)
 diffraction grating [1:51](#), [1:75–77](#), [1:131](#), [1:341](#)
 applications [1:54–57](#)
 beam splitter gratings [1:57F](#), [1:58F](#), [1:59](#), [1:60F](#)
 diffractive lenses [1:56](#), [1:57–59](#), [1:57F](#)
 inductive grid filters [1:59–61](#), [1:61F](#)
 spectroscopic gratings [1:55–57](#), [1:56F](#)
 equations [1:51–52](#), [1:51F](#), [1:52F](#), [1:53F](#)
 fabrication [1:54](#), [1:55F](#)
 to Fourier CGH [4:31–133](#), [4:31F](#)
 in optics [1:74](#)
 theory [1:52–54](#)
 diffractive intraocular lens technology [5:120](#)
 diffractive lenses [1:56F](#), [1:57–59](#), [1:57F](#)
 diffractive multifocals design [5:123](#)
 diffractive optic (D.O.) [2:186](#)
 diffractive optical elements [4:191](#)
 diffuse correlation spectroscopy (DCS) [3:95–96](#)
 diffuse reflectance, methods for analyzing [3:89–90](#)
 diffuser [3:28](#)
 diffusing wave spectroscopy (DWS) [3:96](#)
 diffusion
 approximation equation [3:222](#)
 coefficient [3:222](#)
 cooled laser [2:334](#)
 diffusion-cooled slab laser [2:331](#)
 regime [3:221](#)
 diffusion equation (DE) [3:257](#)
 diffusive density grating [2:4](#)
 diffusive photon correlation transport [3:97](#)
 difluoro-benzothiadiazole (DFBT) [5:265](#)
 digital camera [4:174](#), [4:178](#)
 display [4:177](#)
 digital coherent receiver (DCR) [1:291](#), [1:299–300](#), [1:301F](#)
 carrier phase compensation and decoding [1:305](#)
 clock recovery [1:301–302](#)
 FO compensation [1:304–305](#)
 group velocity dispersion equalization [1:300–301](#)
 nonlinear impairments [1:305–306](#)
 PMD and PDI equalization and polarization demultiplexing [1:302–304](#)
 sampling [1:299–300](#)
 digital elevation models (DEMs) [4:401](#)
 digital Fresnel holography [4:140](#)
 digital geospatial data (2015) [4:401](#)
 digital holographic display
 see also computer generated holograms (CGH)
 and CGH [4:113–114](#), [4:120–122](#)
 hardware for [4:114](#)
 ADHDDs [4:114–115](#)

- bandwidth requirements [4:115–116](#)
 PDHDDs [4:114](#)
 selected holographic display systems [4:116](#)
 holographic display VR/AR [4:124–125](#)
 holographic stereogram for [2:5D](#) images [4:122–123](#)
 key challenges [4:125–126](#)
 light sources [4:120](#)
 phase retrieval approach for 2D images [4:121–122](#)
 3D images [4:123–124](#)
 digital holographic microscopy (DHM) [4:106](#)
 digital holography (DH) [1:378](#), [4:106](#), [4:139](#), [4:407](#)
see also computer generated holograms (CGH)
 advantages [4:148–149](#)
 digital hologram setup [4:141F](#)
 direct phase reconstruction [4:140–143](#), [4:142F](#)
 interferometry
 phase reconstruction by [4:146F](#)
 principle of [4:145F](#)
 influences of discretization [4:147–148](#)
 reconstruction principles [4:143–145](#)
 recording [4:106](#)
 and reconstruction [4:142F](#)
 and replay [4:107–109](#)
 temporal unwrapping [4:148F](#)
 digital lensless Fourier holography [4:147](#), [4:147F](#)
 digital light processing (DLP) [3:22](#), [3:75](#), [4:99](#)
 projecting engines [3:2](#)
 digital micro-mirror devices (DMDs) [3:32](#), [3:42–43](#), [4:99](#), [4:100F](#), [4:115](#), [4:120](#), [4:195–196](#)
 digital optical communications [1:293](#)
 digital phase-locked loop [4:498–499](#)
 digital projector *see* projection display; system
 digital recording [1:109](#)
 digital scene matching area correlator (DSMAC) [4:450](#)
 digital signal processing (DSP) [1:265](#), [1:278](#), [1:291](#), [2:83](#)
 digital speckle pattern interferometry (DSPI) [4:140](#)
 digital subsea holocameras [4:110–112](#)
 appendicularia [4:110F](#)
 calenoid copepod reconstruction [4:111F](#)
 eHolocam [4:100F](#)
 jellyfish larvae reconstruction [4:111F](#)
 digital terrain models (DTMs) [4:401](#)
 digital-signal processing [4:369](#)
 digital-to-analog converters (DACs) [1:143](#), [1:224](#)
 diiodohexane (DIH) [5:263](#)
 diketopyrrolopyrrole (DPP) [5:265](#)
 dilute metals [2:259](#)
 dimension combinations [4:470–471](#)
 repetitively pulsed hypersonic flow [4:471F](#)
 velocity and temperature [4:470–471](#)
 diode laser locking, saturated absorption spectroscopy for
 crossover resonances [2:230](#)
 Doppler free spectroscopy [2:229–230](#)
 ECDL [2:227–228](#)
 experimental setup [2:230](#), [2:231F](#), [2:232F](#)
 principles of saturated absorption spectroscopy [2:228–230](#)
 diode lasers [2:119](#), [2:356](#), [5:3](#)
 diode pumped solid state (DPSS) [4:363–364](#)
 diodes [1:422](#), [5:317](#)
 diopters [1:110](#), [5:44](#), [5:109](#)
 dioptric length of eye [5:109](#)
 diphenyl ether (DPE) [5:263](#)
 dipolar molecules [1:144](#)
 dipole-dipole effects [1:467](#)
 Dirac delta function [3:151](#), [4:202](#), [4:317](#)
 direct buried cables [1:333](#)
 direct cell-free light-mediated effects on molecules [3:117](#)
 direct current (DC) [3:91–92](#), [4:421](#)
 direct detection (DD) [1:248–249](#), [1:291](#), [4:474–475](#)
 direct digital synthesizer (DDS) [4:436–438](#), [5:9](#)
 direct energy gap material [2:350](#)
 direct excitation
 and de-excitation [2:327](#)
 paths [2:35](#)
 direct phase reconstruction in digital holography [4:140–143](#), [4:142F](#)
 direct radiative recombination [2:41](#)
 direct transitions [2:88](#)
 direct transmitted beam [3:207](#)
 direct-inversion methods [3:151](#)
 direction of polarization [1:162](#)
 directionless, ROADM [4:227](#)
 directly modulated lasers (DMLs) [4:68–69](#), [4:247](#), [4:251–252](#)
 large-signal modulation [4:69–70](#)
 small-signal modulation [4:69](#)
 discharge technologies [2:379](#)
 discotic liquid crystals (DLCs) [3:10–11](#)
 discrepancy principle [3:189](#)
 discrete eigenvalue [5:32](#)
 discrete Fourier transform (DFT) [5:2](#), [5:6F](#)
 discrete nonlinear Schrödinger equation [4:281](#)
 discrete phase distribution [4:144](#)
 discrete sampling in digital holography [4:144](#)
 discrete states [1:488](#)
 discrete systems [1:478](#)
 discretization influences in digital holography [4:147–148](#)
 discrimination mechanism of SBS-PC [4:322–323](#)
 disk mirroring [1:281](#)
 disk module [2:409](#)
 dislocations [2:464](#)
 dismutation [3:101](#)
 disordered 1D waveguide lattices [4:280](#), [4:280F](#)
 dispersion [1:1](#), [1:8F](#), [1:18](#), [1:23–24](#), [1:25](#), [1:170](#), [2:304–305](#), [2:305F](#), [4:20](#), [4:25](#), [5:160](#), [5:160F](#)
 characteristic [4:15–16](#), [4:16F](#)
 compensation [1:195](#), [4:24–25](#), [4:29](#), [4:63](#)
 optical nonlinearities as factors [4:22–23](#)
 effect on optical signal [1:1–5](#)
 flattened fibers [1:33](#)
 map [4:24–25](#), [4:24F](#), [4:24T](#), [4:25F](#)
 material group velocity dispersion [1:5–6](#)
 monitoring techniques [4:20–21](#)
 in PCF [1:178–179](#)
 profiles [1:195](#)
 relation [2:88](#)
 waveguide group velocity dispersion [1:6–9](#)
 dispersion compensating fiber (DCF) [1:265](#), [4:10](#), [4:25](#), [4:26–27](#), [4:26F](#), [4:27F](#), [4:31](#)
 dispersion management (DM) [1:23–24](#), [1:370](#), [4:24–25](#)
 chromatic [4:22–23](#)
 chromatic dispersion
 monitoring [4:31–33](#), [4:32F](#)
 in optical fiber communication systems [4:21–22](#)
 and compensation [4:8–11](#)
 slope mismatch [4:30–31](#), [4:31F](#), [4:32F](#)
 solutions
 fixed dispersion compensation [4:26–27](#)
 tunable dispersion compensation [4:28](#)
 dispersion slope [1:195](#), [1:315](#)
 effects [4:20–21](#)
 management [4:20–21](#)
 mismatch [4:30–31](#), [4:31F](#), [4:32F](#)
 dispersion-compensating gratings [1:265](#)
 dispersion-shifted fiber (DSF) [1:9](#), [1:33](#), [4:2](#), [4:2F](#), [4:16](#), [4:20–21](#)
 dispersion-shifted single-mode Fibers (DSF) [1:261](#)
 dispersive medium [4:179](#)
 dispersive tunable oscillator [2:338](#)
 displacement error [3:210](#)
 display(s) [3:1](#), [3:70](#)
 see also projection display
 ambient illumination and reflection [3:72–73](#)
 BRDF [3:73–74](#), [3:74F](#)
 color [3:71](#)
 CRT [3:2–3](#), [3:3F](#)
 temperature [3:75–76](#)
 triangle [3:76](#), [3:77F](#)
 contrast ratio [3:72](#)
 CRT [3:2](#)
 electroluminescence [3:5](#)
 FED [3:4](#)
 gloss [3:71](#)
 illuminance [3:70–71](#)
 inorganic LED Displays [3:5](#)
 Lambertian surface [3:74](#)
 luminance [3:70](#)
 medium [3:53](#)
 non-Emissive Displays [3:5–6](#)
 OLEDs [3:5](#), [3:5F](#)
 optical characteristics [3:70](#)
 plasma displays [3:4–5](#), [3:4F](#)
 PPI [3:75](#)
 radiometric and photometric units [3:70](#)
 reflective LCD [3:6–7](#), [3:6F](#)
 reflectivity [3:71](#)

- display(s) (*continued*)
 refresh rate and frame rate [3:77–78](#)
 resolution and image sharpness [3:75](#)
 response time [3:76–77](#)
 techniques [4:506](#)
 transreflective LCD [3:6F](#), [3:7](#)
 uniformity [3:71–72](#)
 VFD [3:3–4](#)
 viewing angle [3:74–75](#)
 viewing cone [3:74–75](#)
- dissipative processes [5:204](#)
- distance measurement techniques [4:497](#)
 CW phase difference mode [4:497–500](#)
 optical ranging measurement of cornea of rabbit eye performed in vivo [4:498F](#)
 pulse mode [4:497](#)
- distortion [1:2](#), [1:119–121](#)
 analysis [4:213F](#), [4:215](#)
 distortion compensation by phase conjugation [4:331–332](#)
- distributed Bragg mirror [2:280](#), [2:283](#)
- distributed Bragg reflector (DBR) [2:280](#), [3:88](#), [4:18](#), [4:68](#), [4:70](#), [4:255](#), [4:293](#), [4:362](#), [4:419](#)
 lasers [2:454](#), [2:456–457](#), [2:466](#), [4:362](#), [4:362F](#)
 resonator [2:356–357](#), [2:357F](#)
 PWL [2:388–389](#), [2:389F](#), [2:390F](#)
- distributed Bragg reflector laser diodes (DBR-LDs) [2:81–82](#)
- distributed feedback (DFB) [1:186](#), [1:237–238](#), [1:260](#), [1:296–297](#), [1:384](#), [2:356–357](#), [2:357F](#), [4:68](#), [4:70](#), [4:293](#), [4:362](#), [5:313](#)
- dye laser [2:337](#)
 laser [1:369](#), [2:454](#), [2:455–456](#), [2:466](#), [4:3](#), [4:38](#), [4:436–438](#), [4:437F](#)
 cavity [1:233](#)
 PWL [2:389–391](#)
 solid-state dye lasers [2:337](#), [2:344](#)
- distributed Raman amplification (DRA) [2:83](#)
- Distributed Undercompensation (DUICs) [1:265](#), [1:266F](#)
- distribution losses [1:354](#)
- dither [1:252](#)
- DMD™ chip [3:36](#), [3:36F](#)
- donor [2:394](#)
 mode [1:187](#)
- donor–acceptor (D/A) [5:259–260](#)
 pair recombination [2:113](#), [2:113F](#)
- dopants [1:28](#)
- doped fused silica [1:224–225](#)
- doped silica fibers [1:195](#)
- doped TMD monolayers [2:59–61](#)
- doped transport layer [5:234–235](#), [5:236F](#)
- doping [2:139](#), [2:257](#), [2:412–413](#), [5:226–227](#)
 compensation [2:257](#)
 density [2:104](#)
 efficiency [2:257](#)
- Doppler broadened medium, propagation and wave-mixing in [4:344](#)
- Doppler broadening [2:378](#), [4:267](#), [4:342](#), [4:343F](#)
- Doppler components [5:5](#)
- Doppler effect [2:228–229](#), [4:98](#), [4:452](#), [4:528](#)
- Doppler free spectroscopy [2:229–230](#)
- Doppler lidar [4:450–451](#)
- Doppler OCT [5:64](#)
- Doppler principle [4:497](#)
- Doppler shift [4:344](#), [4:369](#), [4:435](#), [5:64–65](#)
- Doppler-free geometries [4:344](#)
- Doppler-free medium [4:345](#)
- double brightness enhancement film (DBEF) [3:28](#)
- double clad fiber [1:201](#)
- double convex lens [1:101–102](#), [1:101F](#), [1:102F](#), [1:103F](#)
- double heterostructure laser diode [2:462](#), [2:463F](#)
- double optical gating (DOG) [2:320](#)
- double prism reflectors [1:126](#)
- double quantum 2D spectra [2:155](#), [2:155F](#)
- double-clad amplifier or laser [4:504](#)
- double-grating polychromator approach [4:469](#)
- double-heterostructure Fabry–Perot laser [2:465](#)
- double-sideband (DSB) [1:429](#)
- double-sided Feynman diagrams [2:151–152](#), [2:153](#)
- doubled Nd:YAG laser microchip [4:447](#)
- down-conversion [1:429](#)
 cross-relaxation [2:395](#), [2:395F](#)
- DQPSK [1:275](#)
 demodulator [1:295F](#)
 transmitter structure [1:294F](#)
- Dragonfly series [2:287–288](#)
- drawing tower [1:30](#), [1:31F](#)
- Dresden protocol [5:126](#), [5:131](#), [5:133](#)
- drilling [1:196](#)
- drive power density [1:135](#)
- Drude relaxation rate (γ) [2:257](#)
- Drude–Lorentz model [2:131–132](#)
- drum vibration [4:453](#), [4:455F](#)
- dry etching [5:293](#)
- dry eye [5:126](#)
- DST see time division multiplexing (TDM); T3 TDM
- DSP satellite [4:488](#), [4:488F](#)
- dual coating [1:33](#)
- dual polarization (DP) [1:291](#)
 IQ transmitter PIC [4:250–251](#), [4:251F](#)
 modulation [1:225](#)
 transmission [1:293–294](#)
- dual-beam double frequency configuration [4:500–502](#), [4:501F](#)
- dual-comb approach [4:433–434](#), [4:433F](#)
- dual-comb interferometry [4:431](#)
 distance measurement [4:432–434](#)
- dual-optic IOL designs [5:121–122](#)
- dual-stage EDFA [1:273F](#)
- dual-view liquid crystal display [3:62](#)
- dual-wavelength optical polarimeter [4:523](#), [4:524F](#)
- duct cables [1:333](#)
- Duffendack reactions [2:373](#)
- duo-binary [4:6–7](#)
- Dyadic Green's function [1:445](#)
- dye lasers [2:336](#), [2:336–337](#), [4:347](#)
 applications [2:348](#)
 classes of [2:336F](#)
- emission characteristics of liquid dye lasers [2:336F](#)
 future of [2:348–349](#)
 laser dye survey [2:344](#)
 liquid dye lasers [2:337–338](#)
 molecular energy levels [2:337](#), [2:338F](#)
 solid-state dye laser oscillators [2:340–342](#)
 solid-state laser dye matrices [2:346–347](#)
- dye methods [5:64](#)
- dye-sensitized solar cells (DSSCs) [2:168](#), [5:270](#), [5:271F](#), [5:284–285](#)
 charge dynamics [2:168](#)
 chemical structures of YD2-o-C8 and Y123 dyes [5:271F](#)
 constructions [5:273](#)
 counter electrodes [5:279–280](#)
 electrolytes [5:277–279](#)
 photoanode [5:273](#)
 future prospects [5:280–281](#)
 mechanism and parameters [5:270–273](#)
 Nyquist plot for [5:274F](#)
- dynamic bandwidth report upstream (DBRu) [2:75–76](#)
- dynamic gain equalizing filter (DGEF) [1:362](#)
- dynamic light scattering (DLS) [3:96](#)
- dynamic spectrum analyzer (DSA) [2:459](#)
- dynamic(s)
 color breakup [3:21–23](#), [3:21F](#)
 contrast ratio [3:72](#)
 dynamic range, illumination [4:88](#)
 of homogeneous catalysts [2:172](#)
 irises [3:42](#)
 linear optical effects [4:351](#)
 LPCs [1:236](#)
 optical misalignment [3:197–198](#)
 of water [2:176–177](#)
- dynamical Bloch oscillations [2:36–37](#), [2:37F](#)
- dynamical wavefront aberration [3:197–198](#), [3:197F](#)

E

- e-holography [3:46](#)
- e-paper [3:56](#)
- e-shift
 device [3:37](#), [3:37F](#)
 by IVC [3:37](#)
- early supercontinuum sources [2:424–425](#)
- early transition metal lasers [2:435–436](#)
- Early Treatment Diabetic Retinopathy Study (ETDRS) [5:134](#)
- Earth Observing System (EOS) [1:379](#)
- ectasia [5:126](#), [5:142](#)
- edge emitters *see* edge emitting laser diode
- edge emitting laser diode [2:350](#), [2:350F](#)
 density of states *vs.* energy for conduction and valence bands [2:351F](#)
 DFB and DBR [2:357F](#)
 direct current modulation frequency response for [2:355F](#)
 emission spectra from Fabry–Perot edge emitting laser structure [2:353F](#)
 energy band diagram for quantum well heterostructure [2:356F](#)

- energy versus momentum diagram [2:350F](#)
- Fabry-Perot resonator [2:352](#)
- gain vs. emission energy [2:353F](#)
- P - I curve for edge emitting laser diode [2:354F](#)
- quantum well lasers [2:354](#)
- real index guided ridge waveguide diode laser [2:352F](#)
- SCM laser [2:352F](#)
- semiconductor [2:350](#)
- edge spread function (ESF) [4:403](#)
- edge-emitting lasers (EELs) [2:466](#), [4:67](#)
- effective index [1:191-192](#)
- effective length [1:18](#)
- effective mass [2:87](#), [2:89-90](#)
- effective Poisson ratio [1:215-216](#)
- effective third-order nonlinearities [4:351](#)
- effective-index model [1:178](#)
- eHoloCam [4:110F](#), [4:111](#), [4:112](#)
- eicosanoids [3:111](#)
- eigenfunctions [1:170](#)
- eigenstates [1:458](#)
- elastic regime [1:215](#)
- elastic strain [1:131](#)
- elastography [5:140](#)
- electric dipole (ED) [2:207](#), [2:208](#)
- approximation [1:448-449](#)
- coupling [1:474-475](#)
- electric dipole spin resonance (EDSR) [1:472](#)
- electric field (E-field) [1:172](#), [1:229](#), [4:335](#)
- amplitude [5:182](#)
- of emitting fluorophore [5:178](#)
- enhancement [5:301](#)
- influence of [2:96](#)
- poling technique [1:14](#)
- electric laser field in time and frequency domain [2:234](#)
- electric multipolar expansions [2:212](#)
- electrical current, coherence control of [1:488-490](#), [1:491](#)
- electrical dispersion compensation (EDC) [2:79](#)
- electrical impedance [3:196](#)
- electrical networks [4:317](#)
- electrical TDM [4:34](#), [4:60](#), [4:61F](#)
- electrically addressed SLM (EASLM) [4:116](#)
- electrically based time division multiplexing systems [4:34](#)
- electrically pumped lasing [5:316](#)
- electrically pumped VCSELs (EP-VCSELs) [2:284-285](#)
- electro absorption modulator (EAM) [4:3](#)
- electro optic effect [1:225-226](#)
- electro-absorption (EA) [4:243](#)
- electro-absorption modulated lasers (EMLs) [4:247](#), [4:251-253](#), [4:252F](#)
- electro-absorption modulators (EAM) [1:142](#), [1:260](#), [1:261F](#), [1:367](#), [4:62](#), [4:254](#), [4:256F](#)
- transmitters [4:255-257](#), [4:256F](#)
- electro-absorptive modulation [1:142](#)
- electro-generated excitons [3:64](#)
- electro-kinetic phenomenon [3:79](#)
- electro-luminescent display (EL display) [3:1](#)
- electro-optic detection (EOD) [2:25](#)
- electro-optic modulator (EOM) [1:141-142](#), [1:156-157](#), [1:261](#), [2:306](#), [4:3](#), [4:4F](#)
- electro-optic(s) (EO) [1:141](#), [1:141-142](#), [2:199](#), [3:232](#)
- see also nonlinear optics
- design [3:181-182](#)
- devices [1:143](#), [1:146-147](#), [1:159-160](#)
- effect [1:141](#), [4:235](#)
- materials [1:142](#), [1:144-146](#), [1:145F](#), [1:146F](#), [1:146T](#)
- modulator [4:216](#)
- receiver [4:446-447](#)
- relationships [1:143-144](#)
- switching [1:142](#)
- technology [2:244](#)
- tensor [1:143](#)
- electro-wetting display (EWD) [3:80-82](#), [3:81F](#), [3:82F](#)
- electrochemical capacitance-voltage profiling (ECV profiling) [4:416-417](#)
- electrochemical doping [5:226](#)
- electrochemical impedance spectroscopy (EIS) [5:270](#)
- electrode array direction (EAD) [3:48](#)
- electrode structures [1:403](#)
- electrode-less discharge [2:382](#)
- electrode/organic interface [5:232](#)
- electrodynamics [3:257](#)
- electroluminescence (EL) [2:105-106](#), [3:5](#), [3:66](#), [5:220](#), [5:240](#), [5:315](#)
- electrolytes [5:277-279](#), [5:278F](#)
- electromagnetic (EM) [1:212-213](#), [1:217-218](#)
- field [1:451](#), [1:452](#), [2:384](#), [5:203-204](#)
- flux [3:229](#)
- irradiation [4:264](#)
- radiation [1:432](#), [1:432-433](#)
- IR and microwave radiation [5:191](#)
- UV radiation and X-rays [5:191](#), [5:191F](#)
- signal generators [1:143](#)
- spectrum [2:118](#), [2:118F](#)
- for electrical and optical communications [4:13F](#)
- wave [1:118](#), [2:118](#), [2:207-208](#), [3:241](#), [3:262](#), [4:13](#), [4:113](#), [4:180](#)
- interaction [4:310](#)
- propagation [1:169](#), [3:248-249](#), [3:248F](#)
- electromagnetic theory
- block waves and Brillouin zones [1:170](#)
- computational techniques [1:172](#)
- Maxwell's equations in periodic media [1:169-170](#)
- origin of photonic bandgap [1:170-172](#)
- photonic-crystal slabs [1:173-175](#)
- semi-analytical methods [1:172-173](#)
- three-dimensional photonic crystals [1:175](#)
- two-dimensional photonic crystals [1:173](#)
- electromagnetically induced transparency (EIT) [2:15-18](#), [2:17](#), [4:339](#)
- enhancement technique [4:344](#)
- nonlinear optical processes [4:342-344](#), [4:343F](#)
- nonlinear optics with pair of strong coupling fields [4:344-345](#)
- propagation and wave-mixing in Doppler broadened medium [4:344](#)
- pulse propagation and nonlinear optics [4:345-346](#), [4:345F](#)
- theoretical treatment of [4:340-342](#), [4:340F](#)
- three-level atomic system [4:339F](#)
- electron affinity (EA) [5:256](#)
- electron beam induced current (EBIC) [5:193](#)
- electron block layer (EBL) [5:232](#), [5:240](#)
- electron bombarded active pixel sensor technology (EBAPS technology) [4:445-446](#), [4:446F](#)
- electron energy
- diagram [2:98F](#), [2:101F](#), [2:102F](#)
- distribution function [2:327F](#)
- electron injection layer (EIL) [3:64](#), [5:232](#), [5:240](#)
- electron phonon relaxation (e-ph relaxation) [2:133](#), [2:134F](#)
- electron spin resonance (ESR) [1:467](#)
- electron transfer (ET) [2:6-7](#)
- electron transport layer (ETL) [5:232](#), [5:236F](#), [5:240](#), [5:257](#), [5:266](#), [5:267F](#), [5:284-285](#), [5:286-287](#)
- electron transport materials (ETM) [5:287-288](#), [5:289T](#)
- electron-acoustic phonon scattering [2:464-465](#)
- electron-injection layer (EIL) [5:247](#)
- electron-transport layer (ETL) [5:227](#), [5:247](#), [5:247F](#)
- electron/hole transporting layers (ETL/HTL) [3:64](#)
- electron(s) [2:40](#), [2:101](#), [2:207](#), [2:464](#), [5:175](#), [5:185](#), [5:192](#)
- beam lithography [5:204](#)
- correlation [4:290](#)
- in large cubic sample, properties of [2:100F](#)
- microscope [4:105](#), [5:191-192](#), [5:191F](#)
- spin [1:473-474](#)
- wave [4:113](#)
- packet [2:239](#)
- electron-boson interactions effect [2:110](#)
- electron-electron
- correlations [2:124](#)
- interaction
- for electrons in bands [2:110](#)
- in traps [2:110-111](#)
- scattering [2:102-103](#)
- electron-hole systems (e-h systems) [1:451](#)
- polarization [2:37](#)
- recollisions [2:37-38](#), [2:38F](#)
- scattering [2:84-85](#)
- electronic
- collimation [3:196](#)
- controller [4:224](#)
- degree of freedom [5:208](#)
- devices [4:97-98](#)
- AOM [4:98-99](#)
- FPAD [4:98](#)
- SLM [4:99](#)
- infrared imager [3:232](#)
- Kerr effect [4:53-54](#)
- recording [4:108](#)
- regime [2:119](#)
- sensor array [4:106](#)
- structures [5:220-223](#), [5:222F](#)
- calculations [4:289-290](#)

- electronic (*continued*)
 temperature [2:164–465](#)
 wave packet [2:37](#)
- electronic spectrum analyzer (ESA) [2:458](#)
- Electronics Industries Alliance (EIA) [3:70](#)
- electron–phonon interactions, control of [1:488](#)
- electron–photon interaction [2:90–91](#)
- electrophoretic display (EPD) [3:79–80](#), [3:79F](#)
 microcup® EPD [3:80](#)
 microencapsulated [3:79–80](#)
- electrophotographic device [5:227](#), [5:228F](#)
- electroplated metal film [1:54](#)
- electroplating [1:54](#), [4:91](#)
- electroretinogram [5:87](#)
- electrostatic units (esu) [4:350](#)
- electrostatics [3:80–81](#)
- elementary transient holograms [2:1](#)
- ellipsometry [1:149](#)
- embossed holograms [4:91](#), [4:92F](#)
- emission
 band [4:364](#)
 linewidth [4:358–359](#)
 processes [5:207](#)
 spectra [3:217](#)
- emission computerized tomography (ECT) [3:195](#)
- emission layer (EML) [5:232](#)
- emissive displays [3:1](#), [3:79](#)
- emissive layer (EML) [5:240](#)
- emissive OLED displays [3:72](#)
- emitted sideband radiation [2:38](#)
- emitting layer (EML) [5:240](#), [5:247](#)
- emmetropia [5:51](#)
- emmetropization [5:61](#)
- empty cavity [1:459](#)
- empty-lattice approximation [2:87](#)
- emulsion swelling [4:105](#)
- enantiotropic LC mesophase [3:8](#)
- Encyclopedia [5:169–170](#)
- end-to-end WDM transmission [1:260](#)
- endlessly single-mode fiber [1:197–198](#), [1:198F](#)
- endogenous agents [3:86–87](#)
- endogenous fluorophores [5:87](#)
- energy
 conservation [2:105–106](#), [2:115–116](#)
 conversion [2:350](#)
 dissipation rate of dipole [1:445](#)
 gap [2:87](#)
 law [2:397](#), [2:398F](#)
 level diagram [2:101](#)
 shift in confined space [1:451–452](#)
 transfers [2:394–395](#), [2:395F](#), [2:400](#), [2:401F](#)
- energy vs. momentum diagram (E – k diagram) [2:350](#), [2:350F](#)
- energy-efficient dynamic bandwidth allocation [2:79–80](#)
- energy-efficient processes [2:447](#)
- energy-recovered linac (ERL) [1:405](#)
- energy-saving techniques [2:79–80](#)
- “enhanced permeability and retention” effect (EPR effect) [3:110–111](#)
- enhanced specular reflector (ESR) [3:28](#)
- entangled state [1:479](#), [1:479–480](#)
- entangled wavefunctions [1:462](#)
- entanglement [1:478](#), [1:478](#)
 basic notions [1:478](#)
 creating entangled states experimentally [1:479–480](#)
- entangled-photon source [1:480F](#)
- multiqubit states and entanglement [1:478–479](#)
- optical entanglement, application of [1:484](#)
- polarization entanglement-based quantum cryptography protocol [1:482F](#)
- quantum key distribution [1:480–482](#)
- quantum superdense coding [1:182](#)
- quantum teleportation [1:482–484](#)
- qubits [1:478](#)
- entero-hepatic recycling [3:110](#)
- EPI-fluorescence [5:170](#)
- epi-illumination [5:187](#)
- epitaxial lateral overgrowth method (ELO method) [2:273](#)
- epsilon-near-zero (ENZ) [2:263](#)
- equi-omnidirectional ranges [1:209](#)
- equilibrium [2:103–104](#)
 equations [1:215–216](#)
- equivalence theorems [1:45](#)
- equivalently wavefunctions [1:458](#)
- erbium [1:264](#)
- erbium ions [1:362](#), [1:363](#)
- erbium-doped fiber (EDF) [1:355](#), [1:356](#), [4:38](#)
- erbium-doped fiber amplifier (EDFA) [1:19](#), [1:181–183](#), [1:182F](#), [1:263](#), [1:264F](#), [1:291](#), [1:331](#), [1:355](#), [2:83](#), [4:1](#), [4:4F](#), [4:5F](#), [4:17](#), [4:18](#), [4:35](#), [4:46](#), [4:68](#), [4:419](#)
 advanced EDFA functions [1:362–363](#)
 components [1:355–356](#)
 low-noise design [1:361–362](#)
 multiple-stage [1:358–361](#)
 single-stage amplifier [1:356–358](#), [1:358F](#)
- ergodic process [4:174](#)
- error-reduction algorithm (ER algorithm) [3:147](#)
- erythrocyte imaging
 with AOSLO [5:67–68](#)
 with flood illuminated adaptive optics [5:68–69](#)
- esthetic applications, PBM for
 for fat reduction and cellulite treatment [3:120–121](#)
 for hair regrowth [3:119–120](#)
 skin rejuvenation [3:118–119](#), [3:118F](#)
- etalons [2:340](#), [5:140](#)
- etch stop (ES) [3:13](#)
- etch stopper structure (ES structure) [3:12–13](#)
- ethylene carbonate (EC) [5:277](#)
- ethylenediamine tetra-acetic acid (EDTA) [3:104](#)
- European Society of Cataract and Refractive Surgeons (ESCRS) [5:136](#)
- europium (Eu) [5:244](#)
- eutectic LC mixtures [3:9](#)
- evanescent field
 coupling technique [1:341](#)
 of THz [5:168](#)
- evanescent spectra [3:140–142](#)
- evanescent wave coupling [1:340](#)
- evolution associated spectra (EAS) [2:192](#)
- evolutionary algorithms [5:41](#)
- Ewald limiting sphere [3:161](#)
- Ewald sphere [3:139](#), [3:139F](#)
 of reflection [3:161](#)
- ex situ approach [2:321](#)
- ex vivo techniques [5:142](#)
- exceptional points (EP) [4:291](#), [4:291–293](#)
 sensing at [4:298–300](#)
- exchange [1:474](#)
 gate [1:469](#)
- excimer laser crystallization (ELC) [3:14](#)
- excimer lasers [2:358–360](#), [2:358F](#), [5:135](#)
 absorbers in rare gas halide lasers [2:363F](#)
 decay paths and absorber [2:364F](#)
 discharge technology [2:365–367](#)
 energy and pulse rate for applications and technologies [2:359F](#)
 excimer wavelengths [2:360F](#)
 fundamentals [2:360–365](#)
 potential energy curves for rare gas halides [2:362F](#)
 typical formation and quenching reactions and rate coefficients [2:363F](#)
 typical ground and excited state densities [2:364F](#)
 variants on magnetic switching circuits [2:366F](#)
- excitation-induced dephasing (EID) [2:58](#)
- excitation(s) [4:289](#)
 excitation light, nomenclature with [4:347–349](#)
 mechanisms [2:376](#)
 properties [2:118](#)
 transport [2:7](#)
- excited ions [1:201](#)
- excited state absorption (ESA) [2:55](#), [2:435–436](#), [2:442](#)
- excited valley-orbit state [1:474](#)
- excited-state emission (ESE) [2:55](#)
- exciton lines [2:75](#)
- exciton-polaritons [1:445](#), [2:94](#)
- exciton(s) [2:63](#), [2:93](#), [5:223–226](#), [5:225F](#)
 binding energy [2:94–95](#)
 diffusion [5:257](#)
 formation [1:461](#), [5:238](#)
 and decay [5:235–237](#)
 influence of electric and magnetic fields [2:96](#)
 influence of quantum confinement [2:94–96](#)
 in magnetic fields
 exciton states in magnetic fields [2:63–66](#)
 interband magneto-optical absorption [2:68–69](#)
 intraexciton magneto-optical absorption [2:73–75](#)
 many-body effects of high-density excitons in high magnetic fields [2:77–79](#)
 migration process [5:311](#)
 optical properties [2:93–94](#), [2:94F](#)
 scattering [2:96–97](#)
- exciton–exciton annihilation [5:310–311](#)

excitonic

- absorption [1:460](#)
- Auger effect [2:114–115](#)
- electronic structure and dynamics in
 - I-aggregates [2:193–194](#)
- molecules [2:97](#)
- Mou transition [2:72–78](#)
- peak [2:107](#)
- resonance [1:459, 4:289](#)
- system [1:488](#)
- exogenous agents [3:86–87](#)
- “expanding Fourier modulus” method [3:148](#)
- experimental control of quantum phenomena [5:39](#)
- experimental implementation of coherent control [5:39](#)
 - atomic systems control [5:39–40](#)
 - closed-loop control methods [5:41–42](#)
 - time-dependent methods [5:40–41](#)
 - two-path interference methods [5:41](#)
- experimental time window [2:6–7](#)
- extended-focus method [5:197](#)
- extended-reach PONs (ER-PONs) [2:74, 2:82–83, 2:82F](#)
- external cavity diode lasers (ECDLs) [2:457](#)
- external cavity lasers [1:260, 2:457–458, 2:457F](#)
- external cavity semiconductor laser (ECSL) [4:362, 4:362–363](#)
- external modulators [1:260, 1:260–261](#)
- external optical cavity [2:280](#)
- external perturbations [1:487](#)
- external quantum efficiency (EQE) [1:448, 5:235–237, 5:240, 5:242–243, 5:247, 5:251F, 5:254F, 5:258, 5:296](#)
- external resonator [2:447](#)
- external tunability [1:172](#)
- external-cavity diode lasers (ECDLs) [2:227, 2:227–228](#)
- externally modulated lasers, requirements for [4:70](#)
- extinction ratio (ER) [1:261, 4:218, 4:218–219, 4:226](#)
 - of polarizer [1:149–150](#)
- extra-terrestrial THz spectra [1:397](#)
- extracellular matrix [3:86–87](#)
- extraordinary polarization directions [2:293, 2:293F](#)
- extraordinary ray [1:150](#)
- extrema counting (EC) [1:318](#)
- extreme ultraviolet spectral region (XUV spectral region) [2:311](#)
- extrinsic detectors [1:418](#)
- extrinsic infrared detector arrays [1:426](#)
- extrinsic IR photodetectors [1:424](#)
- extrinsic semiconductor detectors [1:423–426, 1:425F](#)
- extrusion [1:196](#)
- eye
 - depth of focus, extending [5:122](#)
 - diagrams [4:252–253, 4:252F](#)
 - optical limitations for imaging [5:72](#)
 - optical system [5:108](#)
 - refractive error [5:116](#)
 - refractive state [5:111](#)

F

- fabrication [1:181–183, 1:210, 1:210F, 1:342, 4:217, 4:218F](#)
 - air-clad fibers [1:196](#)
 - disorder [1:172](#)
 - of gratings [1:54](#)
 - strategies [1:175, 1:214](#)
- Fabry-Perot laser diode (FP-LD) [2:74, 2:82](#)
- Fabry-Perot cavity [1:384, 1:433–434, 1:435](#)
 - resonator [2:356](#)
 - storage time [1:437–439](#)
- Fabry-Perot etalon [2:386–387, 4:320–321](#)
- Fabry-Perot fibers
 - simulation of opto-mechanical behavior [1:215–216](#)
 - structure and optical properties [1:214–215, 1:214F](#)
- Fabry-Perot interferometers (FP interferometers) [4:29, 5:148–149](#)
- Fabry-Perot lasers (FP lasers) [2:386–387, 2:387F, 4:3](#)
- Fabry-Perot resonances [5:293](#)
- Fabry-Perot resonator [2:351, 2:352, 2:465](#)
- Fabry-Perot spectroscopy [1:99](#)
- face-centered cubic lattice (fcc lattices) [1:175](#)
- Facility for Advanced Accelerator Experimental Tests (FACET) [1:405](#)
- failures integrated in time (FITs) [4:70](#)
- fanning [4:324–325](#)
 - effect [4:332–333](#)
- far-field patterns (FFPs) [2:275](#)
- far-infrared (FIR) [2:73, 3:229](#)
 - gas lasers [1:379, 1:379F](#)
 - magnetoabsorption in photoexcited bulk semiconductors [2:73–75](#)
- far-infrared lasers *see* terahertz lasers
- Faraday compensator (FC) [4:523](#)
- Faraday effect [1:159–160](#)
- Faraday modulator (FM) [4:523](#)
 - ranging theory [5:3–6](#)
- Faraday/Kerr rotation [2:124](#)
- fast axial flow lasers [2:332–334, 2:332F](#)
- fast Fourier transform (FFT) [4:121–122, 4:432, 5:8](#)
 - algorithm [1:300–301, 3:151](#)
 - libraries [4:109](#)
- fast learning algorithms [5:34](#)
- fast pulsed electric discharge [2:368](#)
- Fast Scan Submillimeter Spectroscopic Technique (FASST) [1:395](#)
- fast transverse flow laser [2:334](#)
- fat reduction, PBM for [3:120–121](#)
- Fe²⁺ laser advances [2:441, 2:441F](#)
- Federal Emergency Management Agency (FEMA) [4:404](#)
- feed-forward equalizers (FFE) [1:265, 1:278](#)
- feed-forward mode [1:369–370](#)
- feedback algorithm [5:37–38](#)
- feedback TIA [4:261, 4:261F](#)
- femtosecond (fs) [2:124](#)
 - dye laser cavities [2:340, 2:341F](#)
 - excitation laser [5:39–40](#)
 - IR lasers [1:234–235](#)
 - metrology [2:321](#)
 - optical pulses [1:403](#)
- femtosecond laser [1:404, 5:135–136](#)
 - effects on retina [5:91–92](#)
 - induced micro incisions [5:120](#)
 - IRAF reduction [5:93](#)
 - mechanical damage [5:92](#)
 - plasma formation [5:92](#)
 - self-focusing [5:92](#)
 - photochemical damage [5:92–93](#)
 - in retinal imaging
 - artifacts and implementation of TPE fluorescence ophthalmoscopy [5:89–91](#)
 - fluorophores [5:85F](#)
 - multiphoton retinal imaging theory [5:87–89](#)
 - in vivo multiphoton retinal imaging with adaptive optics [5:86–87](#)
 - thermal damage [5:91–92](#)
- femtosecond laser-assisted LASIK (FS-LASIK) [5:135](#)
- femtosecond pulses [2:19–20, 2:233, 4:497, 4:497F, 5:39–40](#)
 - attosecond spectroscopy with attoclock [2:240](#)
 - recollision spectroscopy [2:239–240](#)
 - pumped supercontinua [2:428–429](#)
- femtosecond-assisted cataract surgery (FLACS) [5:134](#)
- Fenna-Mathews-Olson (FMO) [2:191–192](#)
- Fermat's principle [1:96, 3:262](#)
- Fermi edge singularity [2:204–205](#)
- Fermi energy [2:103–104](#)
- Fermi functions [2:351](#)
- Fermi occupancy function [2:350–351](#)
- Fermi's Golden Rule [1:491](#)
- Fermi-Dirac statistics [2:102–103](#)
- Fermionic features [1:461](#)
- Fermions [1:460](#)
- ferroelectric LCOSs [4:115](#)
- ferroelectric materials (FE materials) [1:14, 2:128](#)
- ferroelectricity [2:123](#)
- ferroelectrics [2:128](#)
- ferrotoroidal order in LiCoPO₄ [2:214–216, 2:214F, 2:215F, 2:216F](#)
- ferrule [1:338](#)
- Feussner prism [1:151](#)
- few cycle THz spectroscopy
 - see also* terahertz lasers
 - of QC heterostructures [2:25–31, 2:26F](#)
 - amplification factor [2:31F](#)
 - dispersion in THz broadband QCL [2:30F](#)
 - dynamics of gain and THz pulse train formation [2:28F](#)
 - field of THz pulse [2:30F](#)
 - gain switching [2:27F](#)
 - modeling time response of THz QC heterostructure [2:29F](#)
 - of semiconductor quantum structures [2:19–21](#)
 - carrier density dependence [2:21–23, 2:22F](#)
 - linear response [2:19–21, 2:20F](#)
 - nonlinear response [2:23–25, 2:23F](#)

- few cycle THz spectroscopy (*continued*)
- few-cycle lasers [2:311–312](#)
- carrier-envelope-phase stability [2:312–314](#)
- hollow-core fiber compression [2:311–312](#)
- OPA and OPCPA technology [2:314](#)
- pulse [2:237](#)
- pulse synthesis [2:314–316](#)
- ultrashort pulse characterization [2:316](#)
- Few-Mode Fibers (FMF) [1:262](#)
- Feynman diagrams (FDs) [2:54–55](#), [2:151](#)
- double-sided [2:151–152](#), [2:153](#)
- Feynman path integral [5:217–218](#)
- fiber [4:14–15](#)
- coating [1:32–33](#), [1:221–222](#)
- communications [4:14](#)
- configuration [2:445](#)
- cut-off wavelength [1:312](#)
- diameter [1:32](#)
- during draw [1:211](#)
- drawing [1:30](#), [1:196–197](#), [1:211](#)
- lasers [2:455](#), [4:347](#)
- loss [1:27](#)
- material dispersion [4:1–2](#)
- nonlinearities [4:3](#), [4:4F](#)
- point-to-point [1:280–281](#)
- preform [1:218](#)
- proof stressing [1:326](#)
- property requirements [1:210](#)
- sensors [1:237](#)
- types [1:33](#)
- fiber array measurements [1:213](#)
- fiber Bragg grating (FBG) [1:230–231](#), [1:237F](#), [1:238–239](#), [1:238F](#), [1:265](#), [1:267](#), [1:267F](#), [1:341](#), [1:342F](#), [2:265](#), [2:268](#), [2:457](#), [4:11](#), [4:27](#)
- mirrors [2:262](#)
- transmission characteristic [1:342](#), [1:343F](#)
- fiber cable [1:221–222](#)
- fiber channel (FC) [1:282](#), [1:287](#)
- fiber characteristics measurement [1:308](#)
- environmental characteristics [1:326](#)
- hydrogen aging for low-water-peak single-mode fiber [1:326](#)
- nuclear gamma irradiation [1:326–327](#)
- fiber dimension characteristics and tests
- methods [1:322](#)
- effective area [1:325–326](#)
- fiber geometry characteristics [1:322](#)
- mode field diameter [1:325](#)
- NA [1:324–325](#)
- fiber optical characteristics and tests
- methods [1:308–309](#)
- attenuation [1:308–309](#)
- chromatic dispersion [1:313–316](#)
- cut-off wavelength [1:311–312](#)
- DMD for multimode fibers [1:313](#)
- macrobending sensitivity [1:311](#)
- microbending sensitivity [1:311](#)
- multimode fiber bandwidth for multimode fibers [1:312–313](#)
- nonlinear effects [1:320–321](#)
- PMD [1:316–318](#)
- polarization crosstalk [1:318–320](#)
- mechanical measurement and test methods [1:326](#)
- proof stressing [1:326](#)
- residual stress [1:326](#)
- stress corrosion susceptibility [1:326](#)
- fiber delay lines (FDL) [4:36](#)
- fiber distributed feedback (DFB) [1:240](#)
- fiber geometry characteristics [1:322](#)
- cladding [1:322](#)
- coating [1:322](#)
- core center [1:322](#)
- measurement of fiber geometrical attributes [1:322–323](#)
- mechanical diameter [1:324](#)
- refracted near-field technique [1:323](#)
- side-view technique/transverse interference [1:323](#)
- TNF image technique [1:323–324](#)
- transmitted near-field technique [1:323](#)
- fiber gratings [1:229](#), [4:29](#)
- fabrication
- grating inscription [1:236](#)
- photosensitivity [1:234–236](#)
- nonlinear applications [1:239–240](#)
- spectra [1:232–233](#), [1:232F](#)
- stabilized diode lasers and fiber lasers [1:237–238](#)
- theory [1:229–234](#)
- fiber optics [1:221–223](#), [1:221F](#)
- communication systems [4:13–14](#)
- reliability of lasers in [4:70–71](#)
- fiber polarization controller (FPC) [1:181–183](#), [1:182F](#)
- fiber Raman amplifiers (FRA) [1:19](#), [1:263](#), [1:264](#)
- fiber Raman lasers [2:265](#), [2:269](#)
- advancements in [2:268–269](#)
- fiber under test (FUT) [1:309](#)
- fiber-based amplifier [2:324](#)
- fiber-based DCS measurements [3:98–99](#)
- fiber-coupled 266-nm frequency-quadrupled passively Q-switched microchip laser [2:422F](#)
- fiber-coupled diode lasers [2:416–417](#)
- Fiber-To-The-Home networks (FTTH networks) [1:227](#), [1:252](#)
- Fiber-To-The-X networks [2:73](#)
- fiberoptic telecommunications systems [2:356](#)
- fidelity [3:182–183](#)
- field [4:310–311](#)
- control layer [4:416–417](#)
- effect [5:227–229](#)
- emission [5:233](#)
- evolution dynamics [4:291–292](#)
- fielded systems [4:16–17](#)
- model [4:176](#)
- phases aberration [3:278](#)
- field effect transistors (FETs) [1:419](#), [1:422–423](#), [1:424F](#)
- detectors [1:420](#)
- field emissive display (FED) [3:1](#), [3:4](#), [3:4F](#)
- Field Imaging Far-Infrared Line Spectrometer (FIFILS) [1:425](#)
- field of view (FOV) [3:51](#), [3:152](#), [3:214](#), [3:216](#), [4:125](#), [4:475](#), [5:8](#)
- field sequential color (FSC) [3:20](#)
- field sequential color LCDs (FSC-LCDs) [3:20–21](#)
- field-effect transistors [5:317](#)
- Fienup algorithm [4:122](#)
- Fienup with don't-care method (Fidoc method) [4:122](#)
- fifth harmonic generation (5HG) [2:446–447](#)
- Figures of Merit (FoM) [1:286](#)
- filamentation [1:411](#)
- fill factor (FF) [4:147–148](#), [5:258](#), [5:270](#), [5:272](#)
- filling compounds [1:332](#)
- film casting [5:262](#)
- film patterned retarder (FPR) [3:60](#)
- filter function [4:486–488](#)
- filter polychromator [4:469](#)
- filtered back-projection (FBP) [3:194](#), [4:486–488](#)
- filtered back propagation algorithm [4:177](#)
- Filtered Optical Duobinary [1:279](#)
- filters [1:336](#), [1:342](#)
- final-state stimulation [1:461](#)
- finite difference method [1:197](#)
- finite element modeling (FEM) [1:197](#), [5:142](#)
- finite impulse response (FIR) [4:255](#)
- filter [1:300–301](#)
- finite schematic eyes [5:48](#)
- finite size incoherent radiators [3:244](#)
- finite-difference time domain model (FDTD model) [5:301](#)
- finite-dimensional quantum systems [5:30–31](#)
- Finite-Element-Models (FEM) [5:144](#)
- Flrpic [3:64–65](#)
- first-order Bragg transition [5:216](#)
- first-order CD methods [3:150](#)
- first-order optical system [4:175](#)
- fixed analyzer (FA) [1:317](#), [1:318](#), [1:319F](#)
- fixed dispersion compensation [4:26–27](#)
- see also* tunable dispersion compensation
- chirped fiber Bragg gratings [4:27–28](#), [4:27F](#)
- DCF [4:26–27](#), [4:26F](#), [4:27F](#)
- HOM dispersion compensation fiber [4:28](#)
- fixed frequency shifter [1:140](#)
- fixed grid [4:227](#)
- flash ladar [4:446](#), [4:447F](#)
- flashlamp-pumped dye laser [2:337](#)
- flashlamps [2:336](#)
- flat panel displays (FPD) [3:2](#), [3:2F](#)
- flat photographic transparency [4:102](#)
- flat-flat cavity design [2:417](#)
- flavin adenine dinucleotide (FAD) [3:86–87](#)
- flavins [3:117](#)
- flavoproteins [5:86](#)
- flex transparent conducting film (Flex-TCF) [3:57](#)
- flex-grid [4:227](#)
- flexible display
- integrating with clothes [3:52](#)
- integrating with human body [3:52–53](#)
- practical processes for [3:57](#)
- flexible E-paper [3:56](#)
- flexible electrophoretic displays (EPD) [3:56](#)
- flexible liquid crystal display [3:56–57](#)
- flexible OLED (FOLED) [3:5](#), [3:69](#)
- flexible organic light emitting diode [3:54–56](#)
- flexible substrates [3:53](#)
- flexible TFT [3:53–54](#), [3:55F](#), [3:557](#)
- flexible transparent conductive film [3:57](#)
- flood illuminated adaptive optics
- erythrocyte imaging with [5:68–69](#)

- flow cytometers [4:516](#)
 flow cytometry [4:514–516](#)
 flow measurement techniques [4:524–525](#),
[4:524F](#)
 flower power [5:293](#)
 flowmetry [5:64–65](#)
 fluctuation as dimension [4:469–470](#)
 fluctuation vision [4:469–470](#)
 fluorescein angiography [5:66–67](#)
 fluorescein-5-isothiocyanate (FITC) [3:211](#)
 fluorescence [2:248](#), [4:369](#), [4:456](#), [5:241](#),
[5:241F](#)
 imaging [3:225–226](#), [5:85](#)
 lifetime [5:181](#)
 microscopy [5:175](#), [5:187](#), [5:192](#),
[5:198–201](#), [5:198F](#)
 multispectral imaging approaches [3:213](#)
 rate [1:446](#), [1:446F](#)
 spectroscopy [4:518](#)
 fluorescence correlation spectroscopy (FCS)
[5:172](#)
 fluorescence lifetime imaging microscopy
 (FLIM) [5:172](#), [5:181–182](#)
 fluorescent dyes [3:211](#), [4:516](#), [5:192](#)
 fluorescent molecules [2:249](#), [5:88](#)
 fluorescently stained leukocytes [5:67](#)
 fluorescence recovery after photobleaching
 (FRAP) [5:172](#)
 fluorine (F) [1:28](#), [4:91](#)
 fluorine-doped tin oxide (FTO) [5:273](#)
 fluorochromes *see* fluorescent dyes
 fluorophores [5:85F](#), [5:86–87](#), [5:169](#), [5:178](#)
 fluorophore-metal interaction [5:180–181](#)
 radiation [5:181](#)
 flux [3:234](#)
 flux-doubling model [4:304](#)
 “fly-scan” ptychography [3:153](#)
 FMCW lidar [4:434–435](#)
 focal length [2:329–331](#)
 focal plane array (FPA) [1:419](#), [2:167–168](#),
[3:239](#), [4:369](#), [4:372](#), [4:373F](#), [4:446](#),
[5:1](#)
 focal plane array detector (FPAD) [4:98](#)
 Fock states [1:454–455](#), [4:181](#)
 FOL-ladarSim sensor model [4:477](#)
 folding dynamics, protein structure and
[2:173–175](#)
 amyloid proteins [2:174–175](#)
 membrane proteins [2:176](#)
 foliage penetration performance
 characterization [4:404](#)
 forced coherent population oscillations
 (FCPO) [1:252–253](#)
 forced swimming test (FST) [3:125](#)
 form birefringence [1:179–180](#)
 formalism [4:310–311](#)
 properties of nonlinear susceptibilities
[4:311](#)
 relation between field and polarization
[4:310–311](#)
 wave equation in nonlinear regime
[4:311–312](#)
 formamidineum ($\text{NH}_2\text{CH}=\text{NH}_2^+$) [5:282](#)
 Förster energy transfer [5:238](#)
 Förster resonance energy transfer (FRET)
[5:180–181](#)
 forward error correction (FEC) [1:294](#),
[2:77–78](#)
 forward looking infrared (FLIRs) [3:229](#),
[3:232](#)
 forward problem, PDWI [3:256–258](#)
 Foscan [3:103F](#)
 4 f imaging [2:409](#)
 four transmitter and five aperture array
[4:411–412](#), [4:412F](#)
 four-wave mixing (FWM) [1:320](#), [1:321–322](#),
[1:412](#), [3:208](#)
 four-bucket algorithm [4:509–510](#), [4:510F](#)
 four-dimensional imaging (4D imaging)
[3:270](#), [4:526](#)
 four-level pulse amplitude modulation
 (PAM-4) [1:275](#), [1:277](#), [2:80](#), [4:242](#)
 four-wave mixing (FWM) [1:10](#), [1:16](#), [1:180](#),
[1:250–251](#), [1:348](#), [1:348–349](#),
[1:349F](#), [2:82](#), [2:156](#), [2:249](#), [4:3](#), [4:16](#),
[4:20–21](#), [4:23–24](#), [4:24F](#), [4:34](#),
[4:42–44](#), [4:44F](#), [4:45F](#), [4:64](#), [4:323](#),
[4:342](#)
 degenerate four-wave mixing and phase
 conjugation [4:313–314](#)
 example and selected applications
[4:312–313](#)
 field [2:53](#)
 formalism [4:310–311](#)
 nondegenerate four-wave mixing
[4:314–315](#)
 optical Kerr effect [4:312–313](#)
 spectroscopy [2:53](#)
 third harmonic generation [4:313](#)
 4d phase space, extension to [4:209–211](#)
 4D-WD projection set [4:177](#)
 Fourier CCH [4:34](#), [4:35F](#)
 diffraction grating to [4:31–133](#), [4:31F](#)
 reconstruction setup [4:33F](#)
 Fourier domain OCT (FD OCT) [5:156](#)
 Fourier group [4:200](#)
 Fourier measurement-based techniques
[3:146–148](#), [3:146F](#), [3:150](#)
 alternating projection algorithms
[3:147–148](#), [3:147F](#)
 partial coherence algorithms [3:148–149](#)
 three-dimensional recovery problems
[3:149](#)
 Fourier modal method (FMM) [1:52–53](#),
[1:53](#)
 Fourier optics [4:174–175](#), [5:16](#), [5:187](#)
 “Fourier Ping-Pong” [4:34](#)
 Fourier plane
 off-axis IPRC [4:374–375](#)
 off-axis PPRC [4:377–378](#)
 Fourier projection [3:152](#)
 Fourier ptychography (FP) [3:154](#), [3:155F](#)
 mode multiplexing algorithms [3:155](#)
 multi-slice algorithm [3:155](#)
 Fourier space [3:137–138](#)
 Fourier sum times incident wave-function
[5:204–205](#)
 Fourier theory [3:162](#)
 Fourier transform (FT) [1:69](#), [1:69–70](#), [1:318](#),
[2:121](#), [3:188](#), [3:245](#), [3:249](#), [4:123](#),
[4:164](#), [4:174–175](#), [4:182](#), [4:267](#),
[4:272](#), [4:317](#), [4:500](#)
 of aperture function [1:73](#)
 domain [3:176](#)
 methods of optical MDCS [2:53](#)
 spectroscopy [2:151](#), [4:276](#)
 theory [1:19](#)
 2D spectroscopy [2:184](#)
 Fourier transform infrared (FTIR) [1:218](#),
[2:131](#), [4:433–434](#)
 interferometers [1:391](#)
 Fourier Transform Profilometry (FTP) [4:453](#)
 Fourier Transform Raman (FTR) [4:355](#)
 4K × 2K DMD [3:37](#)
 fourth harmonic generation (4HG)
[2:446–447](#)
 fourth-order correlation function [1:48–49](#)
 “fovea centralis” [3:17](#)
 Fowler–Nordheim equation [5:233](#)
 fractional Fourier transform (FrFT) [4:176](#),
[4:200](#), [4:201](#)
 fractional order Fourier transform [4:157](#)
 see also Hough transform
 diffraction-propagation from spherical
 emitter to spherical receiver [4:157](#)
 mathematical expression of diffraction
 through [4:157–158](#)
 working with fractional orders [4:159–160](#)
 fractional orders
 Gaussian beams and Gouy phase [4:162](#)
 image formation by lens [4:160–161](#)
 optical resonator stability [4:161–162](#)
 uncoupling variables [4:162–163](#)
 working with [4:159–160](#)
 Fractional Talbot effect [4:166](#)
 frame rate [3:77–78](#)
 Franck–Condon principle [2:439](#), [5:246](#)
 Franz–Keldysh effect [2:96](#), [4:255–256](#)
 Fraunhofer approximation [1:67–69](#)
 Fraunhofer diffraction [1:67](#), [1:67F](#), [1:71F](#),
[4:158](#)
 array theorem [1:72–73](#)
 blazed gratings [1:78–79](#)
 diffraction
 from circular aperture [1:70–72](#)
 grating [1:75–77](#)
 by rectangular aperture [1:69–70](#)
 Fourier transform [1:69](#)
 Fraunhofer approximation [1:67–69](#)
 grating spectrometer [1:77–78](#)
 Kirchhoff theory of diffraction [1:62–67](#)
 N rectangular slits [1:73–74](#)
 Young’s double slit [1:74–75](#)
 free carriers [1:403](#), [2:19](#)
 free electron laser (FEL) [2:119](#), [2:119F](#)
 free electron THz laser [2:119](#)
 free induction decay (FID) [1:393–395](#)
 free radicals [5:131](#)
 free space [2:12](#)
 ECDIs [2:457–458](#)
 free-space-based tunable dispersion
 compensation device [4:29](#)
 optical communications [4:325](#)
 optical switch technologies [4:231](#)
 LCoS WSS [4:231](#)
 piezoelectric moving-fiber optical switch
[4:233](#)
 2D-MEMS optical switch [4:232–233](#)

- free space (*continued*)
 3D-MEMS optical switches [4:231–232](#)
 propagation in [4:167](#)
 free spectral range (FSR) [1:266](#), [2:81](#)
 free-carrier
 plasma [4:67](#)
 theory [2:464](#)
 free-electron lasers [1:405](#), [1:405F](#)
 free-induction decay [4:266](#)
 “free” carrier absorption [4:303–304](#)
 “free” polaron state [5:258](#)
 freeform prism, aberration contributions in [4:214–215](#)
 freeform systems, visualizing aberrations in [4:208–209](#)
 French curves [1:160](#)
 Frenkel exciton [2:93](#)
 frequency [4:53–54](#), [4:180](#)
 chirp [1:3–4](#), [4:38](#)
 conversion and amplification [2:421–422](#)
 difference techniques [4:500](#)
 3D laser microvision [4:502](#)
 axial eye length measurement [4:500](#)
 dual-beam double frequency
 configuration [4:500–502](#)
 wavelength tuning interferometry [4:500](#)
 domain [3:87](#), [4:179](#)
 electric laser field in [2:234](#)
 instruments [3:259](#)
 measurement method [1:313](#)
 method [1:172](#), [5:147](#)
 methods [5:39](#)
 frequency-converted continuous wave
[2:411–412](#), [2:412](#)
 frequency-dependent refractive index [1:5–6](#)
 gating [3:223–225](#)
 mode-locked laser as frequency comb
 generator [2:303](#)
 shifter [1:140](#)
 tuning [2:417–418](#)
 up conversions [2:398](#)
 vectors [2:198](#)
 frequency counter *see* RF spectrum analyzer
 frequency division multiplexing (FDM)
[1:428](#), [4:17](#)
 frequency domain equalizer (FDE)
[1:300–301](#)
 frequency domain phase measurement
 (FDPM) [4:52](#)
 frequency mixing *see* heterodyning
 frequency modulation (FM) [1:392–393](#), [5:1](#)
 frequency noise *see* phase noise
 frequency offset (FO) [1:297](#)
 frequency-doubled argon ion laser [4:366](#)
 frequency-doubled krypton ion laser [4:366](#)
 frequency-doubled Nd:YAG laser [2:347](#),
[4:111](#)
 frequency-resolved optical gating (FROG)
[2:189](#), [2:316](#), [4:48](#), [4:52](#), [4:52–54](#),
[4:53F](#), [4:56F](#)
 inversion algorithms [4:54–57](#), [4:55](#)
 measurements [4:57](#)
 simplified FROG-GRENOUILLE [4:54](#)
 trace [4:54](#)
 trace errors [4:57](#)
 frequency-resolved optical gating [2:36](#)
 frequency-shift keying [1:275](#)
 frequency-tagging [2:153–154](#)
 Fresnel approximation [4:143–145](#)
 Fresnel CDI [3:150](#)
 Fresnel diffraction [1:85–86](#), [1:86F](#), [1:376](#),
[4:157](#), [4:157F](#)
 circular aperture [1:93–94](#)
 Cornu spiral [1:88–89](#)
 diffraction-propagation from spherical
 emitter to spherical receiver [4:157](#)
 Fermat’s principle [1:96](#)
 Fresnel zones [1:89–93](#)
 images as in-line holograms [4:171–172](#)
 integral [4:371](#)
 mathematical expression of diffraction
 through fractional order Fourier
 transform [4:157–158](#)
 obliquity factor [1:82–87](#)
 opaque screen [1:94–95](#)
 pattern under coherent illumination
[4:165–166](#)
 rectangular apertures [1:87–88](#)
 working with fractional orders [4:159–160](#)
 zone plate [1:95–96](#)
 Fresnel effect [1:337](#)
 Fresnel hologram [4:34](#)
 Fresnel integral [1:84–85](#), [1:84F](#), [1:90T](#)
 Fresnel reflection [1:337](#), [2:371–372](#)
 Fresnel transforms [4:175–176](#), [4:201](#)
 Fresnel zone plates (FZP) [1:377](#), [1:378](#),
[3:251–252](#), [3:251F](#), [3:252F](#), [4:177](#)
 Fresnel zones [1:89–93](#), [1:90F](#), [1:92F](#), [1:94F](#),
[1:96F](#)
 “Fresnel-type” CGH [4:31](#)
 Fresnel–Kirchhoff diffraction formula
[1:66](#)
 Fresnel–Kirchhoff equation [4:109](#)
 Fresnel–Kirchhoff integral [1:67](#)
 Friis’ formula [1:273](#)
 Fringe Field Switching (FFS) [3:60–61](#)
 FROG for Complete Reconstruction of
 Attosecond Bursts (FROG-CRAB)
[2:322–323](#)
 fructose [1:160](#)
 Fubini theorem [4:162](#)
 full service access network (FSAN) [2:74](#), [2:77](#)
 full width at half maximum (f-WHM)
[1:18–19](#), [1:137–138](#), [1:188](#), [1:215](#),
[1:248](#), [1:266](#), [2:277](#), [2:338](#),
[2:459–460](#), [3:66–67](#), [4:218](#), [4:347](#),
[4:363–364](#), [5:90](#), [5:156](#), [5:186](#)
 full-colour reflection hologram [4:103](#)
 full-response pulse [1:275](#)
 full-spectrum fibers [4:15](#), [4:18](#)
 full-width at half height (FWHH) [1:390](#)
 fullerene (C₆₀) [5:321–322](#), [5:321F](#)
 fully collinear geometry [2:189–190](#)
 functional in vivo multiphoton
 ophthalmoscopy [5:87](#), [5:87F](#), [5:88F](#)
 functional near-infrared spectroscopy
 (fNIRS) [3:94](#)
 functional optical polymers [5:309](#)
 fundamental linewidth [2:417](#)
 fundamental pulse [2:313–314](#)
 fundamental wavelength (FW) [2:295](#)
 fundus [5:50–51](#)
 camera with photoelectronic system [4:520](#),
[4:520F](#)
 reflectance [5:51](#)
 furnaces [1:30–32](#)
 fused fiber WDMs [1:356](#)
 fusion
 splices [1:336](#)
 technique [1:340–341](#)
- ## G
- GaAlAs diode laser [2:400](#)
 Gaboroscopy [4:103](#)
 gadolinium scandium gallium garnet
 (GSGG) [2:436](#)
 gain [2:328](#), [2:410](#)
 coefficient [1:351–352](#), [1:352F](#)
 dependence [1:264](#)
 medium physics [2:464–465](#)
 OPSIs gain structure [2:283](#)
 saturation [1:352–353](#), [4:421–422](#), [4:422F](#)
 spectrum of Fiber Raman Amplifiers [1:264](#)
 gain switched modes (GS modes) [2:440–441](#)
 gain-flattening filters (GFFs) [1:358–359](#),
[1:360](#), [1:361F](#)
 gain-switch DFB [4:38](#)
 gallium (Ga) [2:271](#)
 gallium arsenide (GaAs) [1:142](#), [1:489–490](#),
[2:58](#), [2:350](#), [4:502](#), [4:503F](#)
 GaAs/AlAs multiple-quantum wells [2:68](#),
[2:69F](#)
 GaAs/AlGaAs quantum well structures
[2:105](#)
 quantum well
 optical properties of [2:105–107](#)
 structures [2:105](#)
 Schottky barrier diode [1:422](#), [1:422F](#)
 gallium-doped ZnO (GZO) [2:258](#)
 gamma reference [3:2](#), [3:2F](#)
 GaN lasers [2:271](#)
see also optically pumped semiconductor
 lasers (OPSLs)
 AlGaN-based laser diodes [2:273–277](#),
[2:274F](#), [2:277F](#)
 GaN-based materials [1:383](#)
 group III-nitrides [2:271–273](#)
 InGaN-based laser diodes [2:277–279](#),
[2:278F](#)
 Gardner method [1:301–302](#)
 gas
 dynamic lasers [2:326](#)
 lasers [1:379](#), [2:337](#)
 mixture [2:326](#)
 THz laser [2:118–119](#)
 gating techniques [2:319–321](#), [2:320F](#)
 Gaussian approximation [4:428](#)
 Gaussian beams [4:162](#), [4:348](#), [4:357](#)
 Gaussian constants [1:102–104](#), [1:105–107](#)
 Gaussian distribution [1:2–3](#), [1:3](#), [2:55–56](#),
[4:344](#)
 Gaussian process regression-based method
[3:151](#)
 Gaussian pulse [1:3F](#), [1:1–5](#)
 Gaussian statistics [4:174](#)

- Gaussian units [4:335](#)
 Gaussian-distributed random process [4:373](#)
 Gaussian-shaped spectrum resolution [5:158](#)
 Gd-Sc-Al garnet (GSAG) [2:441–442](#)
 $\text{Cd}_3\text{Sc}_2\text{Ga}_3\text{O}_{12}$ *see* gadolinium scandium gallium garnet (GSGG)
 IOGE-PON [2:77–78](#), [2:78F](#), [2:78T](#)
 Ge:Ga photoconductors [1:425](#)
 Geiger discharge process [3:89](#)
 Geiger-mode APD (GM/APD) [4:475](#)
 gelatin [4:90](#)
 geminate recombination [5:257](#)
 general analysis interferometry (GINTY) [1:318](#)
 general phase function [4:186](#)
 generalized Gauss lens equation [1:105–106](#)
 generalized lens [4:175–176](#)
 generalized multiple-prism dispersion theory [2:337](#)
 generalized nonlinear Schrödinger equation (GNISE) [1:258](#), [2:425–426](#)
 generalized projections [4:55](#)
 algorithm [4:55](#)
 generalized propagation [1:39](#)
 generalized Rayleigh criterion [5:187](#)
 generalized version of DOG (GDOG) [2:320](#)
 generic framing procedure (GFP) [1:287](#)
 generic mapping procedure (GMP) [1:288](#)
 generic organic LED structure [5:315–316](#), [5:316F](#)
 generic phase function [4:186](#)
 genetically encoded calcium indicators (GECIs) [5:87](#)
 geometric canonical transforms [4:199–200](#)
 geometrical laws [3:262](#)
 geometrical optics [1:62](#)
 approximation [3:262](#)
 transformations [3:262](#)
 Gerchberg-Saxton algorithm (GS algorithm) [3:147](#), [3:147F](#), [4:122](#)
 germania (GeO_2) [1:28](#)
 germanium
 intra-valence band lasers [1:380–381](#), [1:380F](#)
 photoconductors [1:425](#)
 ghosts [1:54](#)
 Gibbs effect [3:191](#)
 Gibbs phenomenon [3:184–185](#)
 Gigabit Ethernet PON (GE-PON) [2:73](#), [2:76](#), [2:78T](#)
 auto-discovery [2:76](#), [2:77F](#)
 normal operation [2:76](#), [2:77F](#)
 Gigabit PON (GPON) [2:73](#), [2:74–76](#), [2:75F](#), [2:78T](#)
 downstream GPON packet structure [2:75F](#)
 upstream
 frame structure [2:75–76](#)
 GPON packet structure [2:76F](#)
 glass [1:131](#), [1:195](#), [2:369](#), [3:25](#), [4:14](#)
 casting of glass powder [1:196](#)
 fiber transmission systems [2:468](#)
 optical fiber preform [1:28](#)
 glass-fiber reinforced plastic (G-FRP) [1:331](#)
 glassy carbon [5:280](#)
 glaucoma [5:110](#)
 global energy production [5:256](#)
 global positioning system survey (GPS survey) [4:401–402](#)
 gloss [3:71](#)
 glucose lidars [4:522](#)
 differential absorption glucose content measurement [4:522](#)
 polarization-based glucose content measurement [4:522–523](#)
 glue(ing) [1:336](#), [2:410](#)
 Golay cells [1:420](#)
 gold (Au) [2:256](#), [5:296](#), [5:299F](#)
 gonioscopy lens [5:140](#)
 Gouy phase [4:162](#)
 GPON Transmission Control layer (GTC layer) [2:75](#)
 graded index fiber (GRIN fiber) [1:223](#), [1:356](#)
 gradient descent methods (GD methods) [3:150](#)
 gradient of singular phase function [4:187–188](#)
 gradient-rapid-assisted-pulse-echo spectroscopy (GRAPEES) [2:190](#)
 granularity [4:225](#)
 graphene [5:280](#)
 graphene oxide (GO) [5:266](#)
 grating
 equations [1:51–52](#), [1:51F](#), [1:52](#), [1:52F](#), [1:53F](#)
 factor [1:74](#)
 grating-dispersion hyperspectral imaging [3:215–216](#)
 inscription [1:236](#)
 lattice vector [5:204–205](#)
 spectra [1:234](#)
 spectrometer [1:77–78](#), [1:77F](#)
 theory [1:52–54](#)
 grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE) [4:54](#)
 grating light valve (GLV) [4:100](#)
 gravitational wave detection [1:432](#)
 gravitational wave detection GW150914 [1:442–443](#)
 GW150914 [1:442–443](#), [1:442F](#)
 interferometer configurations [1:434–437](#)
 noise sources and interferometer sensitivity [1:437–442](#)
 gray-scale inversion [3:59](#)
 grazing incidence X-ray scattering (GIXS) [5:261](#)
 grazing-incidence grating cavities [2:337](#)
 green problem [2:411–412](#)
 Green's function [1:62–64](#), [1:63](#), [1:65](#), [3:138](#)
 Green's theorem [1:63](#)
 grey-to-grey (GTG) [3:77](#)
 Gross-Pitaevskii equation [4:278](#)
 ground sample distance (GSD) [4:401](#)
 ground state [5:224–225](#)
 ground state bleaching (GSB) [2:55](#)
 group conditions [4:199](#)
 group delay [1:314](#)
 group delay dispersion (GDD) [2:297](#)
 group III-nitrides [2:271–273](#)
 group index of refraction [1:315](#)
 group velocity [2:306](#)
 group velocity dispersion (GVD) [1:2](#), [1:4–5](#), [1:4F](#), [1:5](#), [1:5–6](#), [1:178](#), [1:179](#), [1:180](#), [1:299](#), [2:28](#), [2:294](#), [4:325](#)
 equalization [1:300–301](#)
 group velocity mismatch (GVM) [2:294](#)
 CTC encapsulation method frames (GEM frames) [2:75](#)
 guided rays [1:222–223](#), [1:222F](#), [1:223](#)
 guided wave optics
 active fiber compatible components [1:223–226](#)
 data communications [1:226–227](#)
 fiber optics [1:221–223](#), [1:221F](#)
 telecommunications technology [1:226](#)
 guided-wave optical channel [4:13](#), [4:14](#)
- ## H
- Hadamard pattern [4:195–196](#), [4:197F](#)
 hair regrowth, PBM for [3:119–120](#)
 half-shade devices [1:158–159](#)
 half-V shape mode *see* continuous-director-rotation mode (CDR mode)
 half-wave plate (HWP) [1:156](#), [1:479–480](#), [1:480F](#)
 half-width, half-maximum (HWHM) [4:347](#)
 halide anion [5:282](#)
 Hall effect measurements [4:416–417](#)
 halo photonics Doppler velocimeter [4:459–461](#), [4:461F](#)
 halogen elements [5:283](#)
 Hamilton anxiety rating scale (HAM-A) [3:126](#)
 Hamilton depression rating scale (HAM-D) [3:126](#)
 Hamiltonian(s) [2:96–97](#)
 conservation of Hamiltonian structure [4:199](#)
 function [4:199](#)
 systems [4:199](#)
 Hanbury-Brown and Twiss experiment [1:49](#)
 handheld holographic video projector [4:122](#), [4:122F](#)
 Hanning window function [4:190](#)
 Hansen Experimental Physics Laboratory [5:131–132](#)
 “hard-knock” permeation [2:176](#)
 harmonic
 beams [1:490](#)
 broadening [1:258](#)
 conversion [2:246](#), [2:447](#)
 emission [2:236](#), [2:317](#), [2:318F](#)
 functions [1:2](#)
 generation [2:234](#)
 mode-locking EDF laser [4:38](#)
 oscillator [2:161](#)
 hartmanngram [4:530](#)
 Hartmann-Shack ocular wavefront sensing [4:530–531](#)
 commercial Hartmann-Shack aberrometers [4:530–531](#)
 Hartmann-Shack sensor [4:530](#), [4:530F](#)
 Hartree atomic units [2:233](#)

- harvesting triplet excitons in OLED
 advanced delayed fluorescence [5:214–246](#)
 phosphorescent EL [5:241–244](#)
 HDR, color gamut and [3:40–43](#), [3:42F](#)
 head mounted display (HMD) [3:46](#), [3:48](#), [3:51](#)
 health issue–visual fatigue [3:45](#)
 hearing problems, PBM for [3:121](#)
 heat shock proteins (HSPs) [3:119](#)
 heat-assisted magnetic recording (HAMR)
 [2:262–263](#)
 Heaviside step function [4:317–318](#)
 heavy atom effect [5:242](#)
 heavy fermion materials (Hf materials)
 [2:129–130](#)
 HeCd laser [4:361](#), [4:366–367](#)
 Heidelberg retina tomograph (HRT)
 [5:99–100](#)
 Heidelberg retinal tomograph-Rostock
 cornea module (HRT-RCM) [5:100](#),
 [5:100F](#), [5:102](#)
 Heisenberg uncertainty principle [1:458](#),
 [1:463–465](#)
 helium [2:326](#)
 buffer gas [2:373–374](#)
 cadmium laser [2:373–374](#), [2:374F](#)
 helium-neon laser (He-Ne laser) [2:325–326](#),
 [4:359–360](#), [4:359F](#)
 hellman–feynman expression [1:172–173](#)
 Helmholtz equation [1:81](#)
 hematoporphyrin derivative (HPD) [3:110](#)
 hemoglobin (Hb) [3:91](#)
 hepatocyte growth factor (HGF) [3:120](#)
 Hermite–Gauss wavefunctions [4:161](#), [4:202](#),
 [4:203](#)
 Hermitian matrix [3:157](#)
 Herschel condition [5:185](#)
 Hessian matrix [3:150](#)
 heterodyne(ing) [1:373](#)
 becomes homodyning [1:373F](#)
 current [1:374–375](#)
 in demodulation stage [1:374F](#)
 detection [1:296–297](#), [1:374](#), [1:374F](#), [2:460](#),
 [2:460F](#)
 holographic recording of point source
 object [1:376F](#)
 optical direction detection [1:375F](#)
 optical heterodyne detection [1:375F](#)
 optical scanning holography [1:377–378](#),
 [1:378F](#)
 point-object hologram [1:377](#), [1:377F](#)
 TDFZP [1:378](#)
 of two sinusoidal signals [1:373F](#)
 heterogeneous integration [4:242](#)
 hexacarbonitrile (HAT-CN) [5:233](#)
 3HG *see* third harmonic generation imaging
 (THG imaging)
 HgCdTe detectors (MCT detectors) [2:167](#)
 III-CLASS laser system [4:488](#)
 hidden order (HO) [2:145](#)
 hidden phases
 in correlated electron systems [2:212–214](#),
 [2:213F](#)
 revealing by SHG-based techniques
 [2:214–216](#)
 ferrotoroidal order in LiCoPO₄
 [2:214–216](#)
 magnetic quadrupole order in
 pseudogap of YBa₂Cu₃O₇ [2:220–223](#)
 odd-parity hidden order in Sr₂IrO₄
 [2:217–220](#)
 parity-broken nematic order in
 Cd₂Re₂O₇ [2:216–217](#)
 hidden symmetry of 2D magnetoexcitons
 [2:78–79](#)
 high aperture theory [5:195–196](#)
 high bandwidth detector arrays [4:413](#)
 high capacity TWDM-PONs [2:80–81](#)
 high delta core fiber [1:198–199](#), [1:199F](#)
 high delta core microstructure fibers
 (HDCMF) [1:198](#)
 high density lipoprotein (HDL) [3:110](#)
 high frequency enhancement in images [3:16](#)
 high mobility group box protein [1](#)
 (HMGCB1) [3:112](#)
 high performance computer (HPC) [4:230](#)
 switched intracorenet [4:230–231](#), [4:230F](#)
 high power
 narrow linewidth CW OPSL sources [2:286](#)
 wavelength tunable, high brightness CW
 OPSLs [2:285–286](#)
 high reflector (HR) [1:237–238](#)
 high resolution imaging
 in astronomy [3:253–254](#)
 principles of Stellar interferometry
 [3:253–254](#)
 two telescopes interferometer of Labeyrie
 [3:254](#)
 VLT [3:254–255](#)
 methods [5:93](#)
 high resolution range (HRR) [4:483](#)
 high resolution range profiles (HRRPs)
 [4:474](#)
 high temperature plasmonics [2:262–263](#)
 high temperature poly-silicon (HTPS) [3:33](#)
 HTPS LCD panel [3:33](#), [3:33F](#)
 high-bandwidth detectors [4:413](#)
 high-capacity transmission [1:363](#)
 high-density 2D magnetoexcitons [2:79–81](#)
 high-harmonic generation (HHG) [2:33](#),
 [2:142](#), [2:234–235](#), [2:234](#), [2:235F](#),
 [2:311](#), [2:316–319](#), [2:317F](#), [2:319F](#),
 [4:337](#)
 controlling polarization [2:237](#)
 macroscopic phase-matching aspects
 [2:236–237](#)
 microscopic description [2:234–235](#)
 microscopic quantum many-body theory
 [2:34](#)
 three-step model [2:235–236](#)
 time-resolved [2:36](#), [2:36F](#), [2:37F](#)
 in two vs. three bands [2:35](#)
 high-intensity discharge lamps (HID lamps)
 [3:32](#)
 high-magnification lenses [5:186](#)
 high-order harmonics controlling
 polarization [2:237](#)
 high-order modulation [1:293–295](#)
 high-order photonic bandgaps [1:218–219](#)
 high-order sideband generation (HSG)
 [2:37–38](#)
 high-power channels [1:283–284](#)
 high-power lasers [3:199](#), [3:203](#)
 high-powered femtosecond laser systems
 [4:335](#)
 high-precision dual-comb spectroscopy
 [1:384](#)
 high-prf narrow-linewidth dye lasers [2:348](#)
 high-Q Fabry–Perot cavities [1:456](#)
 high-quality 3D imaging [4:510](#)
 high-resolution electronic sensors [4:110](#)
 high-resolution transmission molecular
 absorption database (HITRAN) [1:401](#)
 high-resolution underwater holographic
 imaging
 analogue holographic recording and replay
 [4:106–107](#)
 digital holographic recording and replay
 [4:107–109](#)
 digital subsea holocameras [4:110–112](#)
 subsea holography and holocameras
 [4:109–110](#)
 high-speed
 ADC/DAC algorithm [2:83](#), [2:83–84](#)
 intravascular C7XR optical coherence
 tomography [4:507](#), [4:507F](#)
 phase and/or amplitude modulation [4:243](#)
 high-T_c superconductors (HTSC) [2:125](#),
 [2:126](#)
 higher-order coherence [1:48–49](#)
 Hanbury-Brown and Twiss experiment [1:49](#)
 Stellar intensity interferometry [1:49](#)
 higher-order mode (HOM) [1:28](#)
 dispersion compensation fiber [4:28](#)
 higher-order nonlinear effects [1:10](#)
 higher-order spectroscopy [2:155–156](#), [2:249](#)
 higher-power argon ion lasers [2:381](#), [2:381F](#)
 highest occupied molecular orbital (HOMO)
 [3:64](#), [5:221](#), [5:244](#), [5:256](#)
 highly inclined and laminated optical sheet
 (HILO) [5:170](#)
 highly intense beams [5:39](#)
 highly nonlinear fibers (HNL fibers) [1:368](#)
 highly reflective (HR) [2:408–409](#)
 highly stabilized mode-locked lasers [4:434](#)
 Hilbert filter [4:35](#), [4:36](#), [4:36F](#)
 Hilbert setup [4:35](#), [4:36F](#)
 Hilbert space of square-integrable
 wavefunctions [4:200](#)
 Ho doping [2:413](#)
 hole block layer (HBL) [5:232](#), [5:240](#)
 hole burning [2:229–230](#), [2:230F](#), [2:231F](#)
 experiment [2:167](#)
 hole injection layer (HIL) [3:64](#), [5:232](#),
 [5:234F](#), [5:240](#), [5:247](#)
 hole transport materials (HTMs) [5:278–279](#),
 [5:287](#), [5:287F](#)
 hole-mask colloidal lithography [5:300–301](#)
 hole-transport layer (HTL) [5:227](#), [5:232](#),
 [5:235F](#), [5:240](#), [5:247](#), [5:251F](#), [5:257](#),
 [5:265–266](#), [5:284–285](#), [5:286–287](#)
 hollow all-dielectric fibers [1:217–218](#)
 hollow core fiber [2:269](#)
 hollow optical transmission fibers [1:217](#)
 hollow-core fiber
 approach [2:311](#)
 compression [2:311–312](#)
 hollow-core PCFs [1:263](#)
 holmium [2:413](#)

- holocameras [4:108–109](#), [4:109–110](#)
 hologram(s) [4:30](#), [4:103](#), [4:106](#), [4:120](#), [4:151](#)
 amplitude transmittance [4:30](#)
 computer generated [4:113–114](#)
 frames [4:112](#)
 holographic image formation [4:30F](#)
 irradiance [4:372–373](#)
 mean number of photoelectrons [4:373](#)
 of object [1:376–377](#)
 photoelectron density [4:373–374](#)
 of point object [4:152](#)
 for three-dimensional object [4:34F](#)
 types [4:153–154](#)
 holographic interferometry (HI) [4:154–155](#)
 holographic optical elements (HOEs) [4:105](#), [4:113](#)
 holographic/holography [1:376](#), [3:206](#), [3:223](#), [4:30](#), [4:105](#), [4:106](#), [4:109](#), [4:113](#), [4:139](#), [4:151](#)
 application [4:154–155](#)
 camera [4:140](#)
 colour recording techniques [4:104](#)
 data storage [4:331](#)
 display systems [4:116](#), [4:121F](#)
 display VR/AR [4:124–125](#)
 holographic 3D display in VR/AR HMD [4:125](#)
 holographic 3D head-mounted display [4:125F](#)
 holographic waveguide/lightguide combiner in VR/AR HMD [4:124–125](#)
 imaging [3:168–169](#), [3:169F](#)
 interference patterns [4:113–114](#)
 lithography [1:175](#)
 mechanism of PC [4:323](#)
 perception [4:102](#)
 phase conjugation [3:206–208](#)
 principle [4:113](#), [4:151–152](#)
 process [3:205–206](#), [3:206](#), [3:206F](#)
 reconstruction [1:378](#)
 recording [1:376–377](#)
 electronic devices [4:97–98](#)
 holography terminology [4:87–88](#)
 permanent materials [4:89–90](#)
 process in photopolymer [4:90](#), [4:91F](#)
 and reconstruction of 3D scene [4:114F](#)
 refreshable materials [4:93](#)
 sensors [4:92–93](#)
 simple mathematical description of [4:152–153](#)
 stereogram for 2.5D images [4:122–123](#)
 techniques [2:1](#), [4:103–104](#)
 terminology [4:87–88](#)
 illumination dynamic range [4:88](#)
 illumination linearity response [4:88–89](#)
 holoscopy [4:103](#)
 holovideos [4:106](#), [4:109](#)
 homodyne detection [1:297](#)
 homogeneous broadening [2:52](#), [2:58–59](#), [2:93](#), [2:265](#)
 homogeneous catalysts, dynamics of [2:172](#)
 homogeneous dynamics [2:164–165](#)
 homogeneous Helmholtz equation [3:263–264](#)
 homogeneous lineshapes interpretation [2:157](#)
 homogeneous linewidths measurement [2:160](#)
 homogeneous medium [1:207](#), [4:175](#)
 homogeneous tandem solar cell [5:268–269](#)
 homogenous spectral lineshapes [2:164](#)
 homojunction device [2:462](#)
 Hooke's law [1:215–216](#)
 hopping process [5:233](#)
 horizontal depletion modulator [4:216–217](#), [4:216F](#)
 horizontal dipole orientation [5:235–237](#)
 "horizontal parallax only" systems (HPO systems) [4:118](#)
 horizontal resolution of lidar point cloud [4:403](#)
 host-guest system [5:237–238](#), [5:238F](#)
 hot carrier
 distribution [2:85](#)
 regime [2:82](#), [2:84–85](#)
 hot electron bolometer (HEB) [1:421–422](#), [1:428](#)
 superconducting HEB detectors [1:428–430](#)
 hot phonons [2:85](#)
 Hough transform [4:182](#), [4:189](#), [4:190F](#)
 see also fractional order Fourier transform
 implicit transformation [4:189](#)
 optical implementation [4:189–190](#)
 HSP70 [3:112](#)
 Hubble Space Telescope (HST) [3:188](#)
 human amylin [2:175](#)
 kinetics of amyloid formation by [2:175–176](#)
 human eye [5:130](#)
 aberrations [5:55–58](#)
 aperture stop [5:45–46](#), [5:46F](#)
 axes and angles [5:47–48](#), [5:47F](#)
 basic optical structure [5:43–44](#)
 changes with age [5:60](#)
 cornea [5:44–45](#)
 horizontal section of right eye [5:43F](#)
 Le Grand four-refracting surface schematic eye [5:44F](#)
 lens [5:46–47](#)
 light and eye [5:49–50](#)
 pupils [5:45–46](#), [5:46F](#)
 refractive anomalies [5:51–52](#)
 retinal image quality [5:59–60](#)
 schematic eyes [5:48–49](#)
 structure [3:17](#), [3:17F](#)
 human reconstructed skin (HRS) [3:119](#)
 human retina imaging [5:163–166](#), [5:164F](#), [5:165F](#)
 human skin imaging [5:162–163](#)
 human-in-the-loop (HITL) [3:233](#)
 Huntington's disease [2:174–175](#)
 Huygens source [1:89](#)
 Huygens-Fresnel principle [4:141–143](#), [4:158](#), [4:159F](#)
 Huygens' principle [1:81](#), [1:82F](#), [4:141](#)
 Huygens' wavelets [1:62](#), [1:92](#)
 Huygens-Fresnel integral [1:81](#)
 integral for Fraunhofer diffraction [1:71](#)
 hybrid aligned nematic cell (HAN cell) [3:62](#)
 hybrid designs [1:382](#)
 hybrid EH₁₁ mode [2:311–312](#)
 hybrid fiber-coaxial (HFC) [1:280–281](#)
 hybrid input-output algorithm (HIO algorithm) [3:148](#)
 hybrid multiple-prism near-grazing-incidence grating dye laser oscillator (HMPCI grating dye laser oscillator) [2:338](#), [2:343](#)
 hybrid perovskites [5:283](#)
 hybrid projection-reflection (HPR) [3:148](#)
 hybridizations [1:473–474](#)
 hydrogel [5:116–117](#)
 hydrogen [2:44–45](#)
 aging for low-water-peak single-mode fiber [1:326](#)
 atom [1:458](#)
 loading [1:235](#)
 model [2:44–45](#)
 hydrogen bromide in discharge (CII HyBrID) [2:370](#)
 hydrogen cyanide (HCN) [4:436](#), [4:437F](#)
 hydrogen peroxide (H₂O₂) [3:101](#), [5:131](#)
 hydrogenated amorphous silicon (a-Si:H) [3:12–13](#)
 TFT [3:13](#)
 thin film [3:13](#)
 2-hydroxyethyl methacrylate (HEMA) [2:347](#)
 hyperfine energy level diagrams [2:228](#), [2:228F](#)
 hyperfluorescence [3:65](#)
 hyperlens [3:145](#)
 hypermetropic eye *see* hyperopic eye
 hyperopic eye [5:52](#)
 hyperspectral imaging techniques [3:211](#)
 acquisition [3:217–218](#)
 field of view [3:214](#), [3:216](#)
 grating-dispersion hyperspectral imaging [3:215–216](#)
 image parameters [3:211](#), [3:213–214](#)
 interferometric hyperspectral imaging [3:213–214](#)
 LVI [3:216](#)
 microarray scanning and customized excitation [3:218](#)
 multispectral imaging [3:211–213](#)
 software [3:216–217](#)
 spatial resolution [3:214](#), [3:216](#)
 spectral resolution [3:214–215](#), [3:216](#)
I
 Ianus *see* Janus
 Ianus Geminus [3:100](#)
 ideal coma [1:117–118](#)
 ideal TEM₀₀ Gaussian [4:357–358](#)
 iDesign instrument [4:531](#)
 III-nitride-based semiconductors [2:221](#)
 III-V semiconductors [2:257](#)
 ill-conditioning [3:158](#)
 ill-posedness [3:162](#)
 illuminance [3:70–71](#)
 illumination
 angle scanning-based Fourier ptychography [3:154–155](#)
 illumination-mode NSOM [5:190–191](#), [5:190F](#)

- illumination (*continued*)
 level [5:45](#)
 optics [3:32](#)
- IM-DD links, capacity of [1:291](#)
- image deblurring *see* image deconvolution
- image deconvolution [3:175](#), [3:179–180](#),
[3:187](#), [3:192](#)
- image filtering [4:332](#)
- image formation [5:187](#)
 by lens [4:160–161](#), [4:160F](#)
 in scanning microscopes [5:194–196](#),
[5:195F](#), [5:196F](#)
- image gathering [3:175](#), [3:176–178](#), [3:177F](#)
- image intensification [3:232](#)
- image magnification [5:189](#)
- image plane [4:372](#)
- image plane recording geometry (IPRC)
[4:369](#)
- image processing [5:189](#)
- image restoration *see* image deconvolution
- image rotator (IR) [4:176](#)
- image sensors [3:86](#)
- image space [3:158](#), [3:262](#)
- image-based approach [4:123–124](#)
- “image-plane” CGH’s [4:31](#)
- imaging
 aberrations [3:167](#)
 algorithm, PDWI [3:258–259](#)
 through atmosphere [3:200–201](#), [3:200F](#)
 in Ballistic regime [3:222–223](#)
 biological tissues [3:219F](#)
 computer-simulated diffraction-limited
 image
 of binary star [3:275F](#)
 of point source [3:273F](#)
 computer-simulated long-exposure image
 of binary star [3:276F](#)
 image of point source [3:274F](#)
 computer-simulated short-exposure image
 of point source [3:274F](#)
 development [4:174](#)
 using fluorescence light [5:202](#)
 Fourier amplitudes of diffraction-limited
 transfer function [3:275F](#)
 image gathering [3:176–178](#)
 image restoration [3:179–180](#)
 information theory in model [3:175–176](#),
[3:175F](#)
 matrix formulation [3:156–157](#)
 measurements [5:6](#), [5:7F](#)
 in multiple scattering regime [3:225–226](#)
 optical parameters [3:219–221](#)
 process [3:136–137](#)
 radiance field [3:176](#)
 resolution [3:136–137](#), [3:137](#)
 through scattering media [3:219](#)
 signal coding [3:178–179](#)
 in snake-like regime [3:223–225](#)
 systems [3:262](#)
 technique [3:272–274](#), [4:106](#)
 between ultrafast light source and detector
[3:220F](#)
- imbedded isolator EDFA designs [1:358](#),
[1:359F](#)
- “Immediate Early Response” gene X-1 (IE-X-1)
[3:124](#)
- immune stimulation effect of PDT
 adaptive immunity [3:112–114](#)
 innate immunity [3:111–112](#)
 “immune synapse” [3:113](#)
- impact avalanche transit time (IMPATT)
[2:114](#)
- impact ionization process [2:115–116](#), [4:304](#),
[4:420–421](#), [4:421F](#)
- implicit transformation [4:189](#)
- impulse function [4:385](#)
- impulse response [5:11–12](#), [5:11F](#)
- in situ approach [2:321](#), [2:323](#)
- in situ optical tissue diagnostics/laser
 speckle-based spectroscopy
 techniques
 applications to in vivo tissue sensing
[3:98–99](#)
 coupling optics [3:97–98](#)
 detection systems [3:98](#)
 diffusive photon correlation transport [3:97](#)
 experimental considerations [3:97](#)
 historical overview and background
[3:96–97](#)
 light sources [3:97](#)
 modeling measured intensity correlations
[3:98](#)
- in situ optical tissue diagnostics/
 miniaturized optoelectronic sensors
 applications [3:91–92](#)
 assembly techniques [3:89](#)
 brain research [3:93–94](#)
 cancer diagnostics [3:92–93](#)
 clinical standard of care [3:91–92](#)
 data processing algorithm for
 optoelectronic sensor [3:89–91](#)
 die-level chip assembled on fabricated
 board with die-attached material [3:89](#)
 monolithic fabrication of integrated sensor
[3:89](#)
 optoelectronic components [3:87–88](#)
 optoelectronic sensing system [3:86–88](#),
[3:87F](#)
 small footprint components soldered on
 conventional PCB [3:89](#)
- in vivo
 applications to in vivo tissue sensing
[3:98–99](#)
 functional in vivo multiphoton
 ophthalmoscopy [5:87](#), [5:87F](#), [5:88F](#)
 multiphoton fluorescence ophthalmoscopy
[5:93](#)
 ophthalmoscopy [5:90](#)
 structural in vivo multiphoton
 ophthalmoscopy [5:86–87](#)
 techniques [5:142](#)
 in vivo multiphoton retinal imaging with
 adaptive optics [5:86–87](#)
- in-line amplifier [1:248](#)
- in-line amplifiers [1:354](#)
- in-line holograms, Fresnel diffraction images
 as [4:171–172](#)
- in-line holography [4:151](#)
- in-line method [1:320](#)
- in-line reference beam hologram (ILRH)
[4:106](#), [4:107F](#), [4:109](#)
- in-line SOAs [1:249](#)
- in-line recording process [4:102](#)
- in-phase I.O [2:248](#)
- in-plane switching LCDs (IPS LCDs) [3:72](#)
- in-plane-switching mode (IPS mode) [3:9](#),
[3:59](#), [3:60–61](#)
- (In₂O₃:Sn) *see* tin-doped indium oxide
 (ITO)
- incident photon-to-current conversion
 efficiency (IPCE) [5:270](#), [5:272](#)
- incident wave [1:66](#), [1:84–85](#), [3:142–143](#)
- incident wavevector [1:137](#)
- incoherence parameter [5:187](#)
- incoherent energy transfer [2:59–61](#)
- incoherent FM/CW FPA breadboard
 architecture [5:7–8](#)
- incoherent FM/CW ladar architecture [5:1–2](#),
[5:2F](#)
 advantages [5:2–3](#)
 FM ranging theory [5:3–6](#)
 ladar block diagram [5:1–2](#)
 ladar hardware [5:6](#)
 mathematical analysis [5:3–6](#)
- incoherent illumination, PSF for [3:263](#)
- incoherent image [3:242](#)
- incoherent light holograms [3:252–253](#),
[3:253F](#)
- incoherent relaxation dynamics [2:83–84](#)
- incoherent source image [4:168–169](#)
 derivation of Zernike-Van Citteri theorem
 from propagation of AF [4:169](#)
 partial coherence properties [4:169](#)
 pupil-AF and OTF with defocusing [4:169](#)
 pupil-AF as generalization of OTF [4:169](#)
- incubation [4:304](#), [4:304–305](#), [4:306](#)
- index dielectric slab [1:186](#)
- index guided laser [2:351–352](#)
- index guiding [1:174](#)
- index of refraction [1:141](#)
 of material medium [1:102–104](#)
 of materials [4:48](#)
- index-guided fibers [1:197–198](#), [1:263](#)
 air-clad fiber [1:200–203](#)
 endlessly single-mode fiber [1:197–198](#),
[1:198F](#)
 high delta core fiber [1:198–199](#), [1:199F](#)
 TMF [1:199–200](#), [1:201F](#)
- index-weighted direction tangents [4:209–210](#)
- indicatrix [1:143](#), [4:512](#), [4:513F](#)
- indirect energy gap material [2:350](#)
- indium (In) [2:271](#)
- indium aluminum gallium arsenide
 (InAlGaAs) [4:243](#)
- indium antimonide (InSb) [1:392](#), [3:233](#)
 three-pulse 2D THz spectroscopy on
[2:203–205](#), [2:203F](#)
- indium gallium arsenide phosphide
 (InGaAsP) [1:224–225](#), [4:243](#)
- indium phosphide (InP) [1:142](#), [4:415–416](#)
 active-passive axial waveguide integration
[4:243–246](#)
- dual polarization IQ transmitter PIC
[4:250–251](#)
- electro-absorption modulated lasers
[4:251–253](#)
- functional requirements [4:242–243](#),
[4:243F](#)

- ILMZ PIC [4:248–249](#)
 InP waveguides [4:246–248](#)
 integration technologies [4:243–246](#)
 low loss InP waveguides [4:248](#)
 narrow linewidth high power SG-DBR laser [4:249–250](#)
 photonic integrated circuits
 photonic integration [4:242](#)
 waveguides [4:246–248](#)
 indium tin oxide (ITO) [1:447, 2:257, 2:258, 3:26, 3:35, 3:57, 5:227, 5:240–241, 5:247, 5:265, 5:288, 5:295, 5:303–304, 5:304E, 5:315–316](#)
 indium zinc oxide (IZO) [5:252](#)
 indium–gold (In–Au) [3:89](#)
 individual Fourier transforms [4:145–146](#)
 indocyanine green (ICG) [5:75–76](#)
 induction furnace [1:32](#)
 inductive grid filters [1:59–61, 1:61F](#)
 inelastic light scattering [4:354](#)
 inert buffer gas [2:368](#)
 inertial navigation system (INS) [4:450](#)
 infectious disease, antimicrobial
 photodynamic inactivation and PDT for
 history of aPDI [3:103](#)
 localized infections [3:105F](#)
 pathogen classification and PS structure [3:103–105](#)
 PDT for infectious disease [3:105–106](#)
 rise of antibiotic resistance [3:102–103](#)
 virulence [3:105](#)
 infectious keratitis [5:100–101](#)
 two-photon imaging [5:103–104](#)
 inflammation, PBM for [3:118](#)
 influenza [3:102–103](#)
 role of water in influenza M2 channel [2:176](#)
 Information and Communications Technologies (ICT) [1:285](#)
 information efficiency [3:180–181](#)
 information handling [1:130](#)
 information rate [3:180, 3:182, 3:182–183, 3:183–185, 3:183F](#)
 information theory [3:137–138, 3:175](#)
 data rate [3:182](#)
 electro-optical design [3:181–182](#)
 figures of merit [3:180](#)
 information efficiency [3:180–181](#)
 information rate [3:180, 3:182, 3:182–183, 3:183–185](#)
 maximum-realizable fidelity [3:181](#)
 model of imaging [3:175–176](#)
 performance and design trade-offs [3:181–182](#)
 theoretical minimum data rate [3:180](#)
 throughput SFR [3:182](#)
 visual quality [3:183–185](#)
 information-entropy [3:182](#)
 infrared (IR) [2:124, 2:280, 3:91–92](#)
 absorption [1:255](#)
 arrays [1:426](#)
 bolometers [1:419](#)
 light [5:85, 5:86](#)
 photon counter [2:396](#)
 radiation [3:230, 5:191](#)
 spectroscopy [2:184](#)
 spectrum [3:229](#)
 technology [1:418](#)
 wavelengths [1:217–218](#)
 Infrared Astronomical Satellite (IRAS) [1:425](#)
 infrared autofluorescence (IRAF) [5:93](#)
 reduction [5:93](#)
 infrared imager(s)
 characteristics [3:234–235, 3:235F](#)
 performance [3:233–234, 3:234F, 3:235F](#)
 types [3:232–233](#)
 infrared imaging [3:233T](#)
 see also photon density wave imaging (PDWI); planar waveguide lasers (PWL)
 infrared imaging [3:233T](#)
 characteristics of infrared imagers [3:234–235](#)
 history [3:232](#)
 infrared imager performance [3:233–234, 3:234F](#)
 infrared image–self-emitted energy [3:230F](#)
 Johnson's criteria [3:238T](#)
 modeling infrared imagers [3:238–239](#)
 Moser's data [3:239T](#)
 signal/dynamic range limitation in IR bands [3:232T](#)
 testing infrared imagers [3:235–237](#)
 types of infrared imagers [3:232–233](#)
 typical infrared imaging scenario [3:232F](#)
 visible and infrared spectra [3:229F](#)
 visible image–reflected energy [3:230F](#)
 infrared search and track sensor (IRST sensor) [4:447](#)
 Infrared Space Observatory (ISO) [1:425](#)
 infrared transition metal solid-state lasers
 see also microchip lasers; thin disk laser (TDL)
 Cr²⁺ revolution [2:440–441](#)
 early transition metal lasers [2:435–436](#)
 transition metal laser advantages and issues [2:441–442](#)
 transition metal spectroscopy [2:136–437](#)
 InGaAs linear-mode avalanche photodiodes (InGaAs LM APD) [4:415, 4:424F](#)
 device characteristics [4:420–429, 4:425F](#)
 etched mesa InGaAs APD pixel [4:418F](#)
 InP- and InAlAs-multiplier InGaAs avalanche photodiodes [4:417F](#)
 output count distributions [4:426F](#)
 room temperature band gap vs. crystal lattice parameter [4:416F](#)
 structure and manufacturing [4:415–420](#)
 InGaAs QW [2:284](#)
 InGaN-based laser diodes [2:271, 2:277–279, 2:278F](#)
 InGaZnO (IGZO) [3:27](#)
 inhomogeneity [2:159, 3:207](#)
 inhomogeneous broadening [2:52, 2:58–59, 2:93, 4:342, 4:343F](#)
 inhomogeneous dynamics [2:161–165](#)
 inhomogeneous Helmholtz equation [3:138](#)
 inhomogeneous linewidths [2:160](#)
 inhomogeneous spectral lineshapes [2:164](#)
 inhomogeneously-broadened systems, 2D spectra interpretation of [2:158–160](#)
 measuring homogeneous and inhomogeneous linewidths [2:160](#)
 spectral diffusion [2:160, 2:160F](#)
 injection diode laser see laser diode (LD)
 injection laser diodes (ILDs) [1:141–142](#)
 innate immunity [3:111–112](#)
 inorganic
 crystalline materials [1:142](#)
 dielectrics [3:15](#)
 glasses [2:346](#)
 and hybrid hosts [2:347–348](#)
 inorganic-based optoelectronics [3:89](#)
 inorganic–organic hybrid matrices [2:346](#)
 LED Displays [3:5](#)
 materials [4:330–331](#)
 redox couples [5:277](#)
 semiconductors [5:226](#)
 solar cells [5:294](#)
 input power (IP) [2:327](#)
 input saturation power [1:352–353](#)
 insertion loss (IL) [4:226](#)
 method [1:311](#)
 instantaneous frequency [1:3, 1:4F](#)
 instantaneous power [4:179](#)
 insulated copper conductors [1:332](#)
 integral imaging display [3:48](#)
 integral theorem of Kirchhoff [1:64](#)
 integral transform
 complex extension of canonical transforms [4:202–203](#)
 exponentials of second-order differential operators [4:202](#)
 issues pertaining [4:201–202](#)
 metaplectic sign [4:201–202](#)
 operators [4:201](#)
 integrated circuits (ICs) [4:254](#)
 integrated laser Mach-Zehnder modulator PIC (ILMZ PIC) [4:248–249, 4:248F](#)
 integrated optical circuit (IOC) [1:141–142](#)
 integrated photonic circuitry [1:189–190](#)
 integrated plasmonics [2:261–262](#)
 integrating-bucket technique [4:509](#)
 intense broadband THz generation [1:411](#)
 intense pulsed laser (IPL) [3:119](#)
 intensified photodiode [4:446](#)
 intensity [4:348](#)
 autocorrelation [4:49, 4:49F, 4:50F](#)
 by axis around image/focal plane [5:22–27, 5:25F, 5:26F](#)
 of cross peaks [2:165–166](#)
 detector array [3:267](#)
 distribution [1:486, 4:164, 4:174](#)
 division multiplexing [4:45](#)
 intensity-based OCE [5:146](#)
 intensity–intensity correlation [1:48–49](#)
 of interference [3:245](#)
 on neighboring transverse planes [5:27](#)
 pattern [1:72, 5:15](#)
 reflection coefficients [1:153](#)
 transport derivation of intensity equation [4:166](#)
 on transverse image/focal plane [5:19–22](#)
 intensity and phase modulator (IM/PM) [1:370](#)
 intensity modulation (IM) [1:247–248, 1:275, 1:277, 1:291](#)

- Intensity Modulation with Direct Detection (IM/DD) [1:275](#)
- intensity spectrum of Fresnel diffraction pattern under coherent illumination application to simple objects [4:166](#)
- contrast transfer functions [4:166](#)
- derivation of transport of intensity equation [4:166](#)
- general formulation [4:165–166](#)
- partial Talbot effect [4:166–167](#)
- inter-aperture
- phase aberrations [4:409](#)
 - tip-tilt-piston aberrations [4:411](#)
- inter-channel crosstalk [1:271](#)
- inter-diffusion process [4:244–245](#), [4:245F](#)
- Inter-Division Interface (IrDI) [1:287](#)
- inter-pulse configuration [2:314](#)
- interaction length [1:139](#)
- interaction types in OPA [2:293](#)
- interband magneto-optical absorption [2:68–69](#)
- carbon nanotubes [2:72–73](#)
 - perovskites [2:71](#)
 - quantum wells [2:68–69](#)
 - quantum wires and dots [2:71–72](#)
 - TMDs [2:69–71](#)
- interband transitions [2:252](#)
- interband tunnelling [1:422](#)
- intercellular adhesion molecule 1 (ICAM-1) [3:118–119](#)
- interface inversion symmetry [2:246–247](#)
- interface-sensitive 2D vibrational spectroscopy [2:177–178](#)
- interfacial two-dimensional electron gases (2DEGs) [2:125](#)
- interference [1:36–37](#), [1:62](#), [1:223](#)
- effects [1:486](#), [1:486–487](#), [1:486F](#)
 - methods [5:41](#)
 - microscopes [5:190](#)
 - phenomena [1:486](#)
 - polarizers [1:153–154](#)
 - spectroscopy [1:47–48](#)
- interference filter (IF) [2:211](#)
- waveform [5:4](#), [5:4F](#)
- interference fringes [5:217–218](#), [5:218F](#)
- visibility [1:37](#)
- interferometer [4:233](#)
- configurations [1:434–437](#)
 - sensitivity [1:437–442](#)
- interferometric “null test” [4:35](#)
- interferometric crosstalk [1:271](#), [1:272](#)
- interferometric filter [2:453](#)
- interferometric hyperspectral imaging [3:213–214](#)
- interferometric imaging [3:241](#)
- see also multiplex imaging
 - distance between incoherent sources and two field points [3:242F](#)
 - image of two interfering wavefronts [3:243F](#)
 - two telescopes receive radiation from stellar object [3:243F](#)
- interferometric method [1:318](#), [4:496](#), [4:502–504](#), [5:190](#)
- commercial OCT Instruments for ophthalmology [4:505–506](#)
- optical interferometric micro-lidar [4:502–504](#)
- optical low-coherence tomography [4:504](#)
- interferometric modulator display (IMOD) [3:82](#), [3:84](#)
- interferometric multi-interferer crosstalk [1:272](#)
- interferometric multispectral system [3:213](#)
- interferometric optical switch
- Mach-Zehnder interferometer optical switch [4:236](#)
 - microring resonator optical switch [4:236](#)
 - optical phase shifters [4:236–239](#)
- interferometric PALM (iPALM) [5:176](#)
- interferometric technique [1:54](#), [3:146](#), [4:174](#)
- interferometry (INTV) [1:318](#), [1:318](#), [1:319F](#), [4:139](#)
- interferon gamma (IFN- γ) [3:120](#)
- interleavers [1:267](#)
- interleukin-1 beta (IL-1 β) [3:118–119](#)
- intermediate frequency (IF) [5:2](#)
- intermodal dispersion see modal bandwidth
- intermolecular donor-acceptor system [5:246](#)
- internal conversion process [2:337](#)
- internal quantum efficiency (IQE) [2:278–279](#), [5:235–237](#), [5:241](#), [5:272](#), [5:303](#)
- internal reflection tomography [3:144](#)
- internal rotor [1:388–389](#)
- International Organization for Standardization (ISO) [3:70](#)
- International System of Units (SI unit) [3:70](#)
- International Telecommunications Union (ITU) [1:226](#), [1:387](#), [2:74](#), [4:8](#), [4:15](#)
- ITU-T recommendation G. 694. [1:417–18](#)
- internet of things (IoT) [3:52](#)
- ‘interrogating’ beam [4:325](#)
- interstitial doping [2:257](#)
- intersubband (IS) [2:92](#), [2:201](#)
- selection rules for intersub-band optical transitions [2:91–92](#)
- intersymbol interference (ISI) [1:275](#), [1:299](#), [4:218–219](#), [4:262–263](#)
- intersystem-crossing (ISC) [2:6–7](#), [3:101](#), [5:225–226](#), [5:235–237](#), [5:237F](#), [5:245](#)
- intervalley biexcitons [2:61–62](#)
- intra-aperture phase aberrations [4:409](#)
- intra-cavity self-Raman lasers [2:266](#)
- intra-cavity SHG [2:287](#)
- intra-cavity SHG/DFG [2:283](#)
- intra-channel crosstalk [1:271](#)
- Intra-Division Interface (IrDI) [1:287](#)
- intra-pulse schemes [2:314](#)
- intradband [2:92](#)
- “intracapsular mechanism of accommodation” [5:47](#)
- intracavity etalon, comb control with [2:306–307](#)
- intracavity Fabry-Perot [2:306–307](#)
- intracavity laser-absorption spectroscopy (ICLAS) [2:288](#)
- intracavity nonlinear spectroscopy [2:308](#)
- intracavity phase interferometry (IPI) [2:308–309](#), [2:309F](#)
- intradyne detection [1:278](#), [1:297](#)
- intraexciton magneto-optical absorption [2:73–75](#)
- FIR magnetoabsorption in photoexcited bulk semiconductors [2:73–75](#)
- semiconductor quantum wells
- optical-pump/THz-probe magnetospectroscopy [2:76–77](#)
 - optically detected FIR resonance spectroscopy [2:75–76](#)
- intraocular lens (IOL) [5:121](#), [5:126–127](#)
- intraocular pressure (IOP) [5:141](#), [5:143](#), [5:144](#)
- intraström corneal ring segments (ICRS) [5:136](#), [5:140](#)
- intraström ring segment (ICRS) [5:141](#)
- intravital microscopy of rat cremaster muscle [3:111](#)
- intrinsic
- anisotropy [2:348–349](#)
 - detectors [1:418](#)
 - factors [2:406](#)
 - laser-induced damage mechanism of wide bandgap dielectrics [4:303](#)
- INVAR [3:3](#)
- inverse Bremsstrahlung see “free” carrier absorption
- Inverse HFI (IHFI) [1:300–301](#)
- inverse filter [3:188](#), [3:189F](#)
- inverse Fourier transform (IFT) [3:139–140](#), [3:245](#)
- inverse matrix [4:200](#)
- inverse piezoelectric effect [1:143](#)
- inverse problem [3:156](#), [3:187](#)
- matrix formulation [3:156–157](#)
 - PDWI [3:258–259](#)
- inverse Radon transform [4:174](#), [4:176](#)
- inverse scattering [1:233](#), [1:234](#), [3:161](#)
- inverse source [3:137](#)
- problem [3:159–161](#)
- inverse synthetic aperture radar imaging (ISAR imaging) [4:474](#)
- inversion
- algorithms [3:136–137](#)
 - symmetry [2:218](#)
- inverted OPV device [5:266](#)
- “invisible region” [3:141](#)
- Ioffe-Regel condition [4:278](#)
- ion
- ion permeation through K⁺ ion channel [2:176](#)
 - neutralization [2:370](#)
 - spectroscopy [2:240–241](#), [2:242](#)
 - structure and dynamics of electrolytes [2:170–172](#)
 - traps [1:479](#)
- ionic AOS (IAOS) [3:14](#)
- ionic liquid [2:171–172](#)
- ionization gating (IG) [2:320–321](#)
- ionization potential (IP) [5:256](#)
- ionized plasma [2:372](#)
- iProfiler [4:531](#), [4:531F](#)
- IQ Modulator (IQM) [1:292–293](#)
- Ir(ppy)₃*, organometallic-based coordination complex [3:64–65](#)
- iridium [5:243](#)
- iris [5:130](#)

irradiance-dependent refractive index [4:350–351](#)
 ischemic diseases [5:132](#)
 is iterative Fourier transform algorithm (IFTA) [4:34](#), [4:35F](#)
 ISO LIDT standard [4:307F](#), [4:308](#)
 isolated attosecond pulses [2:237](#)
 attosecond light house [2:238](#)
 phase matching gating [2:238](#)
 spectral selection [2:237](#)
 temporal gating [2:237–238](#)

isolation loss [1:266](#)
 isolators [1:336](#), [1:343](#)
 isomorphic relations [4:291](#), [4:292T](#)
 isoplanatic angle [3:202](#)
 isothermal regime [2:82](#), [2:85](#)
 isotropic case [1:133](#)
 isotropic material [1:131](#)
 isotropic media [2:9](#)
 iterative algorithms [3:150](#)
 iterative inversion algorithms [2:322–323](#)
 iterative method [3:191](#)
 iterative techniques [4:174](#)
 ITO coated polyethylenenaphthalate (ITO/PEN) [5:273](#)
 ITO coated polyethyleneterephthalate (ITO/PEI) [5:273](#)
 IULT study group [15](#) (SG15) [2:78–79](#)

J

J-aggregates, excitonic electronic structure and dynamics in [2:193–194](#)
 Jablonski diagram [3:101](#), [5:98F](#)
 Jackson crossed-cylinder [5:109](#)
 Jacobi identity [4:199](#)
 Jahn–Teller effect (JT effect) [2:126](#), [2:441](#)
 Janus [3:100](#), [3:100F](#)
 face of photomedicine [3:100](#)
 Janus kinase (JAK) [3:100](#)
 Jaynes–Cummings model [1:452](#), [1:460–461](#)
 Jet Propulsion Laboratory (JPL) [1:391](#), [1:392F](#)
 JET amplifiers [1:127](#)
 Johnson–Nyquist noise [4:425](#)
 joint devices [1:336](#)
 joint time-frequency transforms [3:164](#)
 Joint-Density-of-States (JDOS) [2:254](#)
 Jones calculus [1:161](#), [1:168](#)
 Jones matrix Eigen analysis (JME analysis) [1:317](#)
 Jones vector [1:166–168](#)
 Josephson-junction-based qubits [1:473](#)
 Ludd–Ofelt theory [2:397](#)
 jumper cable cut-off wavelength [1:312](#)

K

K⁺ ion channel, ion permeation through [2:176](#)
 Kagomé photonic crystal fibers [2:269](#)
 KACRA, Japanese detector [1:440](#)
 Kalman filter [3:151](#)

Kalman filter phase retrieval (KF-PR) [3:151](#)
 Kane model [2:88–89](#)
 Kapitza–Dirac diffraction [5:214](#), [5:214F](#), [5:215F](#)
 Kapitza–Dirac scattering [5:213–214](#)
 kappa (κ) [5:48](#)
 Kepler-type inverse-square law central force problem [2:44–45](#)
 keratoconus (KC) [5:100–101](#), [5:106](#), [5:118](#), [5:119F](#), [5:141](#)
 keratomileusis [5:126](#)
 Kerr effect [1:141](#), [1:255](#), [1:257](#), [1:258](#), [1:258F](#), [1:320](#), [1:368](#), [4:64](#), [5:92](#)
 Kerr lens mode-locking [5:41](#)
 Kerr nonlinear index [1:320](#)
 Kerr nonlinearity [1:239](#), [4:279–281](#)
 Kerr type nonlinearity [4:346](#)
 Kerr-lensing [2:412](#)
 kinetic energy [1:171](#)
 kinetic exchange interaction [1:469](#)
 kinetics of amyloid formation by human amylin [2:175–176](#)
 Kirchhoff formulation [3:265](#)
 Kirchhoff Green's function [1:63](#)
 Kirchhoff theory of diffraction [1:62–67](#), [1:63F](#)
 “knock-on” permeation [2:176](#)
 Knox–Thompson algorithms [3:275–276](#)
 Köhler illumination system [5:187](#)
 “Kondo liquid” state [2:129](#)
 k.p perturbation theory [2:87–88](#)
 k.p theory [2:87](#), [2:87–88](#)
 Kramers–Krönig dispersion relations (KK dispersion relations) [4:317](#)
 “black box” approach [4:321](#)
 causal linear system [4:318F](#)
 implication of causality [4:320–321](#)
 nonlinear absorption coefficient [4:319F](#)
 two-photon absorption coefficient in semiconductors [4:321F](#)
 Kramers–Kronig relations [1:5](#)
 KrF⁺ formation kinetics [2:362](#)
 Kronecker delta [5:17](#)
 krypton ion laser [2:379](#), [4:361](#), [4:361T](#)

L

label-free sensing [3:86–87](#)
 Labeyrie's method [3:274–275](#)
 two telescopes interferometer [3:254](#)
 laboratory molecular spectroscopy [1:390–391](#)
 broadband resonator spectrometer [1:395–396](#)
 chirped pulse THz spectrometer [1:393–395](#)
 detection and cascaded multiplication spectrometer [1:391–393](#)
 EASST [1:395](#)
 sources [1:390–391](#)
 THz lamb-dip spectroscopy [1:396](#)
 LACE satellite [4:488](#), [4:489F](#)
 ladar block diagram [5:1–2](#)
 ladar hardware [5:6](#)
 imaging measurements [5:6](#)
 point-target measurements [5:6](#)
 ladar impulse response [5:10–11](#), [5:11F](#)
 ladder-type poly[1, 4-phenylene] (MeLPPP) [5:309](#), [5:309F](#)
 Lagrange invariant [4:206](#)
 Laguerre–Gauss functions [4:203](#)
 Laguerre–Gaussian mode [4:180](#)
 “lamb dip” [2:451](#)
 lamb shift [1:452](#)
 lambda (λ) [5:48](#)
 Lambertian surface [3:74](#)
 Lamé's equations [1:215–216](#)
 laminated aluminum polyethylene (LAP) [1:331](#)
 land targets [4:482–483](#)
 LANDSAT satellite program [3:211](#)
 Langerhans cells [5:101](#), [5:101F](#)
 Large Binocular Telescope (LBT) [3:192](#), [3:192F](#)
 large mirrors working [3:247–248](#)
 large photonic bandgaps for CO₂ laser transmission [1:216–220](#)
 large-area (LA) [1:262](#)
 large-effective-area fibers (LEAF) [1:262](#)
 large-particle optics [4:512](#)
 large-signal modulation [4:69–70](#)
 laser [1:54](#), [1:99–100](#), [1:100](#), [1:201](#), [2:384](#), [2:385](#), [2:386F](#), [4:106](#)
 ablation [2:258](#)
 assembly [2:415](#)
 beam [1:140](#), [1:487](#), [5:39](#)
 divergence [4:358](#)
 quality [4:356–358](#)
 cavity [1:139](#), [1:140F](#), [2:331–332](#)
 ceramics [2:411](#)
 characteristics [4:356–358](#)
 configuration [2:331](#)
 cooled atoms [4:344](#)
 cooling [1:458–459](#)
 detection and ranging [4:444](#)
 diffusion cooled [2:334](#)
 Doppler flowmetry [3:95–96](#)
 Doppler vibrometry [4:528](#)
 dual beam method [4:523–524](#)
 electric field [5:30](#)
 emission linewidth [4:358–359](#)
 fast axial flow [2:332–334](#), [2:332F](#)
 fast transverse flow [2:334](#)
 invention [1:10](#), [2:324](#)
 isotope separation [2:372](#)
 laser induced refractive index correction technique [5:126–127](#)
 laser interference lithography [5:293](#)
 laser-induced grating scattering [2:248](#)
 laser-induced interference [1:487](#), [4:340](#)
 laser-induced photo dissociation [5:39](#)
 laser-pumped pulsed dye lasers [2:337–338](#)
 laser-type amplifier [4:322](#)
 light [3:203](#), [4:355](#)
 material [4:325](#)
 micro-radar [4:496](#)
 phosphor [3:42](#)
 phosphor illumination projector [3:39–40](#), [3:40F](#), [3:41F](#)
 polarization [4:358](#)
 pulses [2:337](#)

- laser (*continued*)
- safety considerations of two-photon microscopy [5:106–107](#)
 - scanners [4:483](#)
 - scanning microscopy [5:99–100](#)
 - sealed-off and waveguide [2:331](#), [2:331F](#), [2:332F](#)
 - slow axial flow [2:331–332](#), [2:332F](#)
 - sources [2:311](#), [4:222](#)
 - spectrum [2:302](#)
 - TEA pressure [2:334](#)
 - transverse mode filtering [4:294–298](#)
 - tube construction [2:374](#)
 - ultrasonic inspection [4:332–333](#)
 - velocimeter [4:497](#)
 - vibrometers [4:452](#)
- laser and light interaction
- application [5:137F](#)
 - computer guided treatments [5:138](#)
 - cornea [5:133–134](#)
 - corneal cross linking
 - for corneal infections [5:136–138](#)
 - for refractive procedures [5:136](#)
 - lens [5:134](#)
 - refractive errors [5:135](#)
 - SIT and ALT [5:136](#)
 - SMILE procedure [5:138](#)
 - vitreoretinal disorders [5:134–135](#)
- laser crystallization (LC) [3:14](#)
- improvement from LC modes [3:60–61](#)
 - lens for 2D/3D switchable display [3:48](#)
- laser diodes (LDs) [1:224](#), [1:260](#), [1:277](#), [2:98–99](#), [2:271](#), [2:350](#), [2:409](#), [2:446](#), [4:16](#), [4:500](#)
- laser Doppler Anemometer (LDA) [4:525](#)
- laser Doppler velocimetry (LDV) [4:525](#), [4:525F](#), [5:64](#)
- laser dye survey
- azaquinolone derivatives [2:346](#)
 - conjugated hydrocarbons [2:345](#)
 - coumarins [2:345–346](#)
 - cyanines [2:344](#), [2:345F](#)
 - oxadiazole derivatives [2:346](#)
 - pyromethenes [2:344–345](#)
 - xanthene [2:344](#)
- Laser Geodynamics Satellite (LAGEOS) [4:488](#), [4:489F](#)
- laser guide stars (LGS) *see* artificial reference stars
- Laser Interferometer Gravitational-Wave Observatory (LIGO) [1:432](#), [1:438F](#), [1:439F](#)
- laser interferometer space antenna (LISA) [1:434](#)
- laser radar (ladar) [4:430](#), [4:474](#)
- laser range profiler (LRP) [4:447](#), [4:448F](#), [4:449F](#)
- laser range profiling (LRP) [4:474](#), [4:474–476](#)
- range performance for typical profiling system [4:475F](#)
 - range profiles studies [4:477–479](#)
 - signal processing techniques for range profiling and tomography [4:489–494](#)
 - simulated laser radar waveforms [4:476F](#)
 - simulating laser profiling data [4:476–477](#)
- tomography based on laser profiling [4:483–489](#)
- laser velocimetry [4:523–525](#)
- blood flow measurement in retinal vessels [4:525](#)
 - two-directional velocity measurement [4:525–526](#)
- laser-assisted in situ keratomileusis technique (LASIK technique) [5:100–101](#), [5:126](#), [5:135](#)
- postsurgical assessment [5:101–102](#)
- procedures [5:142](#)
- surgery [5:150](#)
- laser-induced breakdown (LIB) [5:92](#)
- laser-induced damage (LID) [4:302](#)
- defect-induced damage [4:305–306](#)
 - energy diagram of optical excitation and relaxation of dielectric materials [4:305F](#)
 - fundamental processes [4:303–305](#)
 - fundamental LIDTs [4:303F](#)
 - in optical materials [4:304F](#)
 - response of dielectric material [4:302F](#)
- laser-induced damage threshold (LIDT) [4:302](#), [4:305F](#)
- ISO standard for [4:307F](#)
- LIDT F_{th} [4:302](#)
- measuring [4:306–308](#)
- lasing cavity [2:119](#)
- last mile fibers (LMFs) [2:82–83](#)
- lateral motions [3:202](#)
- lateral spatial resolution [4:355](#)
- lateral waveguide component [2:351](#)
- latter effect [1:451](#)
- lattice
- DOF [2:126](#)
 - lattice-matched quantum well system [2:105](#)
 - parameter [2:105](#)
 - vibration [2:118](#)
- law of refraction [1:222–223](#)
- layer-based methods [4:123–124](#)
- layer-by-layer fabrication [1:175](#)
- LC alignment direction (LCAD) [3:48](#)
- 3-LCD projector [3:33](#)
- LCOS *see* liquid crystal on silicon (LCoS)
- Le Grand four-refracting surface schematic eye [5:44F](#)
- Le Grand full theoretical eye [5:48](#), [5:48F](#)
- lead iodide perovskite solar cell ($\text{CH}_3\text{NH}_3\text{PbX}_3$) [5:278–279](#)
- lead lanthanum zirconium titanate (PLZT) [4:235](#)
- lead phthalocyanine (PbPc) [5:321–322](#), [5:321F](#)
- leaky-mode coupling [4:98](#)
- least-squares sense [3:188](#)
- Lebesgue square-integrable 'wave'-functions [4:199](#)
- Leibniz rule [4:199](#)
- Leith and Upatniek's technique [4:103](#)
- lens(es) [3:247](#), [3:248F](#), [5:16–47](#), [5:60–61](#), [5:134](#), [5:203](#)
- approximations for thin lenses [1:109–110](#)
 - cardinal points and planes of optical system [1:101–102](#)
 - combination [4:128](#)
 - compound lenses [1:109](#)
 - using Gaussian constants [1:105–107](#)
 - image formation by [4:160–161](#), [4:160F](#)
 - lens-scanning method [5:194](#)
 - and mirrors [1:101](#)
 - nodal points and planes [1:107–109](#)
 - paraxial matrices [1:102–104](#)
 - paraxial system for design or analysis [1:110–111](#)
 - reflectors [1:111–113](#)
- lensless Fourier approach
- numerical reconstruction by [4:146–147](#)
 - setup [4:147F](#)
- lensmakers equation [1:110](#)
- lenticular lens [3:48](#)
- "less-metals", metals to [2:259–261](#)
- leukocyte
- imaging with AOSLO [5:67](#)
 - velocity [5:67](#)
- levulose [1:160](#)
- Lewenstein model [2:318](#)
- LiCoPO_4 , ferrotoroidal order in [2:214–216](#), [2:214F](#), [2:215F](#), [2:216F](#)
- Lie brackets [4:199](#)
- light [1:221](#), [1:222](#), [1:223–224](#), [3:225–226](#)
- beam characterization [4:174–175](#)
 - crystal [5:203–204](#)
 - detecting devices [3:86](#)
 - emitted from pinholes [1:81–82](#)
 - emitting resonant cavity [2:386](#)
 - harnessing [5:293](#)
 - incident on aperture [3:247](#)
 - interferometers [5:217](#)
 - light-generation and amplification [4:243](#)
 - light-guiding methodology [1:217–218](#)
 - light-sensitive tissue [5:43](#)
 - light-sword wavefront [5:122](#)
 - light/heat sensitive ion channels [3:116–117](#)
 - line [1:186](#)
 - loss at retina [5:61](#)
 - optics [5:203](#)
 - polarization [1:188](#)
 - propagation [3:219](#)
 - properties [4:139](#)
 - pulses [4:347](#)
 - ray region [1:192](#)
 - reflected from mirrors [3:205–206](#), [3:205F](#)
 - scattering [1:237](#), [2:82](#)
 - sheet microscopy [5:170](#)
 - source [1:148](#), [3:27–28](#), [3:32](#), [3:97](#), [3:259](#), [3:267](#), [4:16](#), [4:120](#), [4:500](#)
 - squeezing of light emission [1:463–465](#)
 - substantial fraction [1:233](#)
 - therapy [3:100](#)
 - trapping [5:294–296](#), [5:295F](#), [5:296F](#)
 - wave [1:131](#), [1:131F](#), [4:102](#), [4:113](#)
- light amplification
- process [4:103](#)
 - by stimulated emission of radiation (LASER) [2:329](#)
- light and eye [5:49–50](#)
- light at retina [5:50](#)
 - light interaction with back of eye [5:50–51](#)
 - passage of light into eye [5:49–50](#)
- light and laser therapy [5:130–131](#)

- interaction between laser and biological tissue [5:130F](#)
 photochemical interaction [5:131](#)
 photochemical relationship between oxygen, riboflavin and UVA light [5:132F](#)
 photomechanical interactions [5:132–133](#)
 photothermal interactions [5:131–132](#)
 light detection and ranging (LIDAR) [2:271](#), [2:451](#), [2:461](#), [4:401](#), [4:430](#)
see also micro-lidars
 configuration [4:512](#), [4:512F](#)
 error model [4:402–403](#)
 Lidar Base Specification (2014) [4:401](#)
 light emitting diodes (LEDs) [1:141–142](#), [1:142](#), [1:224](#), [2:82](#), [2:194–195](#), [2:271](#), [2:446](#), [3:1](#), [3:25](#), [3:88](#), [3:118–119](#), [4:3](#), [4:16](#), [4:99](#), [4:509–510](#), [5:1](#), [5:248](#), [5:314](#)
 light extraction
 of plasmonic mode [5:251–253](#)
 of substrate mode [5:248](#)
 of waveguiding mode [5:248–251](#)
 light harvesting (LH) [5:292](#), [5:302](#)
 AR effects [5:293–294](#)
 exciton diffusion bottleneck [5:292F](#)
 LH2 [2:192](#)
 light trapping [5:294–296](#)
 metal nanostructures [5:296–302](#)
 in OSCs [5:292](#), [5:293](#)
 other approaches for [5:304–306](#)
 spacer layer and cavity effects [5:302–304](#), [5:304F](#)
 ways of harnessing light [5:293F](#)
 lightfield imaging [3:169–170](#), [3:170F](#)
 light-matter
 interaction [1:459](#)
 observation via strong light-matter interaction [1:459–461](#)
 system [1:462](#)
 lightpath [4:233](#)
 lightwave [2:384](#)
 lightwave transmitters
 directly modulated lasers [4:68–69](#)
 DWDM transceivers [4:68](#)
 reliability of lasers in fiber optic systems [4:70–71](#)
 requirements for externally modulated lasers [4:70](#)
 semiconductor laser principles [4:67](#)
 structure and requirements of optical fiber communication systems [4:67–68](#)
 transmitter requirements [4:68](#)
 limbal stem cells, two-photon imaging of [5:103](#)
 limited-memory: Broyden–Fletcher–Goldfarb–Shanno method (L-BFGS) [3:150](#)
 limiting amplifier (LA) [4:6](#), [4:261](#)
 Linear Coherent Light Source (LCLS) [1:405](#), [2:142](#)
 LiNbO₃ *see* lithium niobate (LN)
 line broadening effect [2:328](#)
 line spread function (LSF) [4:403](#)
 linear absorption spectrum [2:52](#)
 linear band [2:89](#)
 linear canonical transforms [4:199](#)
 canonical integral transforms [4:200–201](#)
 conservation of Hamiltonian structure [4:199](#)
 geometric canonical transforms [4:199–200](#)
 issues pertaining integral transforms [4:201–202](#)
 radial canonical transforms [4:203](#)
 linear crosstalk [1:271–272](#)
 linear defect [1:189–190](#)
 linear dispersion maps, corrections to [4:25–26](#)
 linear effects [1:255](#)
 attenuation [1:255–256](#)
 chromatic dispersion [1:256](#)
 polarization mode dispersion [1:256–257](#)
 linear electro-optic coefficient [4:330](#)
 linear electro-optic effect [4:3](#), [4:324](#), [4:351](#)
 linear error at 90th percentile (LE90) [4:402](#)
 linear feedback shift register (LFSR) [5:9](#)
 linear frequency chirp (LFC) [4:434](#)
 linear gratings [1:51](#), [1:51F](#)
 linear IR [2:164](#)
 linear Kramers–Krönig relations [4:317–318](#)
 linear matrix [4:210](#)
 linear optics [4:350](#)
 implementation [4:182](#)
 spectroscopy [2:82](#), [2:246–247](#)
 system [1:39](#)
 linear particle accelerator (LINAC) [1:405](#)
 linear perturbation analysis of rate equations [4:69](#)
 linear polarization [2:418](#), [4:268](#)
 optical beam [1:408](#)
 linear polarizer (LP) [1:149](#), [2:211](#)
 linear radiation transfer equation [3:257](#)
 linear regularization method [3:191](#)
 linear response [2:19–21](#), [2:20F](#)
 linear signatures [1:345–346](#), [1:345F](#)
 linear spectroscopy [2:244](#)
 linear susceptibility [4:339](#)
 linear variable filter (LVF) [3:216](#)
 linear-algebraic theorems [1:170](#)
 linear-mode (LM) [4:415](#)
 linearity response, illumination [4:88–89](#)
 linearly chirped grating [4:29](#)
 linearly modulated laser frequency [4:451](#), [4:452F](#)
 linearly polarized (LP) [1:229](#)
 optical beam [4:174](#)
 linewidth
 enhancement factor [2:467–468](#)
 narrowing [2:341](#)
 Linewidth Enhancement Factor (LEF) [1:245](#)
 link layer ID (LLID) [2:76](#)
 link-level considerations [4:222](#)
 Liouville equation [4:341](#)
 Liouville space pathway [2:204–205](#), [2:204F](#)
 lipid peroxidation [3:102](#)
 lipofuscin [5:86](#)
 liquid and solid state systems [2:164–165](#)
 liquid crystal displays (LCDs) [3:1](#), [3:2](#), [3:18–19](#), [3:19F](#), [3:25](#), [3:25F](#), [3:31](#), [3:56–57](#), [4:99](#), [5:240](#)
 backlight module [3:27–28](#)
 colors in [3:17–20](#)
 controlled viewing angle [3:62](#)
 LC panel [3:25](#)
 periphery component [3:29–31](#)
 projection display [3:32–34](#)
 stripe half-wave-plate [3:34F](#)
 liquid crystal on silicon (LCOS) [1:269](#), [1:269F](#), [3:32](#), [3:34](#), [3:34F](#), [4:99](#), [4:99](#), [4:231](#)
 devices [4:115](#), [4:115F](#)
 display [3:2](#)
 projection display [3:34–35](#)
 liquid crystal technologies (LCT) [1:268](#), [1:269](#)
 liquid crystals (LCs) [3:8](#), [3:25](#), [3:26–27](#), [4:99](#)
 chiral smectic and nematic phases [3:10](#)
 color filter [3:26](#)
 discotic phases [3:10–11](#)
 glass [3:25](#)
 nematic phase [3:8–9](#)
 panel [3:25](#)
 polarizer [3:25–26](#)
 smectic phases [3:9–10](#)
 TFT [3:27](#)
 thermotropic LC phases [3:8](#)
 transparent electrode [3:26](#)
 liquid dye lasers
 see also solid-state dye laser
 CW dye lasers [2:339–340](#), [2:340F](#)
 flashlamp-pumped dye lasers [2:338–339](#), [2:339F](#)
 laser-pumped pulsed dye lasers [2:337–338](#), [2:339F](#)
 ultrashort-pulse dye lasers [2:340](#), [2:341F](#)
 liquid electrolyte [5:277](#)
 liquid nitrogen temperature [2:120](#)
 liquid-crystal on silicon wavelength-selective switch (LCOS WSS) [4:231](#)
 Lissajous figure [1:162](#)
 lithium niobate (LN) [1:14](#), [1:16](#), [1:144–146](#), [1:177](#), [2:294](#), [4:3](#), [4:96](#), [4:331](#)
 based pulse front tilting setup [1:409](#)
 technology [4:39](#)
 lithium quinolate (LiQ) [5:233](#)
 lithium tantalate [1:16](#)
 lithography [1:54](#), [1:55F](#), [1:57](#), [1:484](#), [4:91](#)
 Littman configurations [4:363](#)
 Littman–Metcalf configuration [2:227](#), [2:228](#)
 Littrow configuration [2:227](#), [2:227F](#)
 Littrow type-I design [4:363](#)
 Littrow type-II design [4:363](#)
 living marine plankton recording [4:110](#)
 LiYF₄ *see* yttrium lithium fluoride (YLF)
 LN *see* lithium niobate (LiNbO₃)
 local area networks (LANs) [1:354](#), [2:76](#), [4:3](#)
 local density of optical states (LDOS) [1:445](#)
 Local eXchanges (LX) [1:280–281](#)
 local oscillator (LO) [1:250](#), [1:374](#), [1:428](#), [2:56–57](#), [2:164](#), [2:197](#), [2:248](#), [4:45](#), [4:250](#), [4:432–433](#), [5:2](#)
 field [2:184](#)
 laser [1:291](#)
 local-primary-desaturation (LPD) [3:22](#), [3:23–24](#), [3:23F](#)
 localized surface plasmon applications (Q_{LSP}) [2:255](#)
 localized surface plasmon resonance (LSPR) [2:262](#), [5:296–298](#)

lock-in amplifiers (LIA) [4:523](#)
 Lockheed Martin Coherent Technologies [4:458–459](#)
 long pass filter (LPF) [2:211](#)
 Long Period Grating (LPG) [1:230–231](#), [1:234F](#), [1:236](#)
 long trajectories [2:316–317](#)
 long-haul optical transmission systems [1:249](#)
 long-haul WDM systems [1:282–284](#)
 long-period FBGs [1:267](#)
 long-range WindScanner (LRWS) [4:462](#)
 long-term fluctuations [1:32](#)
 long-wave IR (LWIR) [3:232](#)
 long-wavelength-cone (L-cone) [3:17](#), [3:18F](#)
 longitudinal aberration [5:55](#), [5:56F](#)
 longitudinal acoustic waves [1:131](#)
 longitudinal chromatic aberration of eye [5:58](#)
 longitudinal modes [2:329–331](#), [2:386–387](#)
 longitudinal optical phonons (LO phonons) [1:382](#), [2:84–85](#)
 longitudinal spherical aberration (LSA) [1:114–116](#), [1:115F](#)
 longitudinal-optical and-acoustical differences (LO-LA differences) [1:408–409](#)
 look-up-table method (LUT method) [4:123](#)
 looping mechanism [2:396](#)
 loose jacket [1:328–330](#), [1:330F](#)
 Lorentz model [1:346](#)
 Lorentz oscillator [2:254](#)
 Lorentzian profile [1:352](#)
 loss factor [2:254](#)
 loss mechanisms, OPSIs [2:283–284](#)
 low bending-loss single-mode fibers (C.657 low bending-loss fibers) [1:261](#), [1:262](#)
 Low Cost Autonomous Attack System (LOCAAS) [4:483](#)
 low density lipoprotein (LDL) [3:110](#)
 low extinction ratio (ER) [1:252](#)
 low frequency response [1:252–253](#)
 low level light (laser) therapy (LLLT) [3:114–115](#)
 low loss InP waveguides [4:248](#)
 low noise multi-gigahertz repetition rate femtosecond pulsed OPSIs [2:286–287](#)
 low temperature poly-silicon (LTPS) [3:14](#), [3:27](#), [3:66–67](#)
 low-bandwidth feedback TIA [4:262–263](#), [4:262F](#)
 low-coherence interferometry [4:503–504](#)
 low-index grid (LIG) [5:248–250](#), [5:250F](#)
 low-index polymers [1:210–211](#)
 low-loss optical fiber [1:28](#)
 low-noise design of EDFAs [1:361–362](#)
 low-pass filter (LPF) [1:279](#), [1:373](#), [1:374](#), [1:378](#), [3:194](#), [4:423](#)
 low-power argon ion lasers [2:374](#)
 low-water-content fibers *see* full-spectrum fibers
 low-water-peak fibers [4:15](#)
 lower exciton polaritons (LP) [1:448–449](#), [1:449F](#)
 lowest eigenfrequencies [1:172](#)
 lowest order mode [2:329–331](#)

lowest unoccupied molecular orbital (LUMO) [3:64](#), [5:221](#), [5:244](#), [5:256](#)
 lowest-order light-matter interaction [1:458](#)
 LTPS-TFT AMOLED smartphones [3:68–69](#)
 luminance [3:70](#)
 luminance uniformity [3:72](#)
 luminescence
 process [2:105–106](#)
 spectra [2:85](#)
 up-conversion [2:82](#)
 luminescence mechanisms
 energy transfers [2:394–395](#), [2:395F](#)
 materials [2:397](#)
 photon avalanche [2:396–397](#), [2:396F](#)
 sequential two-photon absorption [2:395–396](#), [2:396F](#)
 up-conversion lasers based on [2:399–400](#)
 up-conversion mechanisms from [2:394–395](#)
 lutetium aluminum garnet (Yb:LuAG) [2:413](#)
 Luttinger and Kohn representation (LK representation) [2:87–88](#)

M

M-aryPSK (*m*-PSK) [1:293](#)
 M-ary Quadrature Amplitude Modulation (*m*-QAM) [1:293](#)
m-axes [2:271](#)
m-plane [2:271](#)
 Mach-Zehnder Interferometer (MZI) [1:146–147](#), [1:251](#), [1:193](#), [1:225](#), [1:225–226](#), [1:261](#), [1:316](#), [1:368](#), [1:68F](#), [4:3](#), [4:40–41](#), [4:50–51](#), [4:51F](#), [4:505](#)
 optical switch [4:236](#), [4:236F](#), [4:242](#), [4:250](#)
 Mach-Zehnder Modulator (MZM) [1:144–146](#), [1:261](#), [1:261F](#), [1:292](#), [1:292–293](#), [1:293F](#), [2:80](#), [4:39–40](#), [4:259](#), [4:260F](#)
 transmitters [4:258–259](#)
 macro-bending loss [1:256](#)
 macrobending sensitivity [1:311](#), [1:311F](#), [1:312F](#)
 macroscopic parameters [2:244](#)
 macroscopic phase handle [1:487](#)
 macroscopic phase-matching aspects [2:236–237](#)
 macroscopic polarization [4:264–265](#), [4:342](#), [4:349](#)
 macula [3:17](#)
 Madison Darby canine kidney II cells (MDCK II) [3:108](#)
 magnetic dipole [1:475](#)
 coupling [1:475](#)
 magnetic dipole-dipole interaction [1:475](#)
 magnetic fields
 exciton states [2:63–66](#)
 influence [2:96](#)
 three-dimensional case [2:64–66](#)
 two-dimensional case [2:66–68](#)
 magnetic flux quantum [2:72](#)
 magnetic lenses [5:192](#)
 magnetic quadrupole order in pseudogap of YBa₂Cu₃O₇ [2:220–223](#)

magnetic resonance imaging (MRI) [3:124](#), [3:192](#), [3:259](#)
 magnetically ordered oxides [2:127–128](#)
 magnetism [2:123](#)
 magneto-optic devices [1:159–160](#)
 magneto-optical Kerr effect microscopy (MOKE) [2:147](#)
 magnetoelectric coupling [2:125](#)
 magnetoelectric multiferroics (ME multiferroics) [2:128–129](#)
 magnetoresistive pyrochlores [2:135–136](#)
 magnification [1:105–106](#)
 main central office (MCO) [2:82–83](#), [2:82F](#)
 major histocompatibility complex class I molecules (MHC-I) [3:113](#), [3:114](#)
 malaria [3:102–103](#)
 manipulative artist, pseudo colour and [4:104–105](#)
 many-body effects of high-density excitons in high magnetic fields [2:77–79](#)
 hidden symmetry of 2D magnetoexcitons [2:78–79](#)
 superfluorescence from high-density 2D magnetoexcitons [2:79–81](#)
 mapping [4:482–483](#)
 marginal rays [1:114–116](#)
 Mark systems [4:115](#), [4:116–117](#)
 maser operation [1:452–454](#)
 maser patent [1:99](#)
 Massachusetts General Hospital (MGH) [4:507](#)
 mast modes measurement [4:453](#), [4:453F](#)
 master oscillator power fiber amplifier (MOPFA) [2:429](#)
 master oscillator/fiber power amplifier systems (MOPA systems) [4:431](#)
 flared planar amplifier [2:286](#)
 master-oscillator laser (MO laser) [4:369](#)
 material absorption [1:172](#)
 material dispersion [1:5](#), [1:217–218](#), [1:313](#), [1:314F](#)
 material effective index [4:220](#)
 material group velocity dispersion [1:5–6](#)
 material sciences
 molecular electrolytes [2:170–172](#)
 organometallic chemical catalysts [2:172](#)
 solid state materials [2:168](#)
 materials selection criteria [1:210–211](#)
 mathematical expression of diffraction through fractional order Fourier transform [4:157–158](#)
 mathematical model of imaging [3:176](#)
 mathematical representation of optical pulse [4:48–49](#)
 matrices [4:199](#)
 matricial active triangulation [4:511–512](#)
 matricial detectors, time-of-flight principle with [4:510–511](#)
 matrix algebra [4:202](#)
 matrix analysis
 Jones calculus [1:168](#)
 Jones vector [1:166–168](#)
 Mueller calculus [1:165–166](#)
 polarization ellipse [1:162–163](#)
 Stokes parameters [1:163–166](#)
 matrix element [2:106](#)

- matrix formulation of imaging and inverse problems [3:156–157](#)
- matrix methods [1:149](#)
- for computing polarization [1:160–161](#)
- matter interferometer [1:487](#)
- matter wave optics [5:203](#)
- maximum efficiency [4:88](#)
- Maximum Entropy Method (MEM) [3:191–192](#)
- maximum modulation [4:88](#)
- maximum permissible exposures (MPE) [5:106–107](#)
- maximum power point (MPP) [5:259](#)
- maximum-realizable fidelity [3:181](#)
- Maxwell electromagnetic theory [2:398](#)
- Maxwell's equation [1:148](#), [1:177](#), [1:458](#), [2:53](#), [2:118](#), [4:264](#), [4:266](#), [4:312](#), [4:321](#), [4:342–343](#)
- in periodic media [1:169–170](#)
- Maxwell's theorem [3:262](#)
- Maxwell's wave equation [4:335](#)
- Maxwell-Boltzmann distribution [2:82](#)
- McIntyre distributions [4:426](#), [4:428](#)
- MDA-MB-231 human mammary gland adenocarcinoma cells [5:182](#)
- mean free paths (MFP) [3:222–223](#)
- mean time to failure (MTTF) [4:71](#), [4:255](#)
- mechanical diameter [1:324](#)
- mechanical flexibility [1:218](#)
- mechanical scanners [1:136](#)
- mechanical scanning [4:509–510](#)
- mechanical splices [1:336](#), [1:339](#), [1:339F](#)
- mechanical tuning experiment and discussion [1:216](#), [1:216F](#)
- media access control (MAC) [2:73](#)
- media parameters as dimensions [4:455–463](#)
- atmosphere components and contaminants [4:463–469](#)
- direct detection elastic lidar with three laser transceivers [4:466F](#)
- temperature [4:469](#)
- three components of lidar wind speed [4:466F](#)
- wind and flow velocity [4:456–463](#), [4:457F](#), [4:458F](#)
- medical physics [3:199](#)
- megasample per second (MS/s) [4:434–435](#)
- MEH-PPV films [5:261](#), [5:309–310](#), [5:309F](#), [5:323–324](#), [5:324–325](#), [5:324F](#)
- membrane proteins [2:176](#)
- ion permeation through K⁺ ion channel [2:176](#)
- role of water in influenza M2 channel [2:176](#)
- memory LCD [3:84](#)
- memory mechanism [1:471–472](#), [1:471F](#)
- charge qubits [1:471–472](#)
- spin qubits [1:472–473](#)
- memory-in-pixel (MIP) [3:84](#)
- mercury ion laser [2:376](#)
- MESA Imaging SR4000 emitters [4:510–511](#), [4:511F](#)
- meso-tetra(4-sulfonatophenyl)porphine (TPPS4) [3:108](#)
- mesogen [3:8](#)
- mesoporous oxides [5:274–275](#)
- mesostructured fibers for external reflection applications [1:209–210](#)
- fabrication approach [1:210](#), [1:210F](#)
- materials selection criteria [1:210–211](#)
- Metal Induced Energy Transfer (MIET) [5:176](#), [5:182](#)
- metal organic chemical vapor deposition (MOCVD) [2:462](#)
- metal vapor lasers [2:368](#)
- see also* carbon dioxide laser
- afterglow recombination [2:372–373](#)
- afterglow recombination metal vapor lasers [2:372–373](#)
- continuous-wave metal ion lasers [2:373–374](#)
- copper bromide laser tube construction [2:370F](#)
- CVIs [2:369–372](#)
- excitation circuits [2:371F](#)
- tube construction [2:370F](#)
- partial energy level scheme for copper vapor laser [2:369F](#)
- principal self-terminating resonance-metastable metal vapor lasers [2:369F](#)
- resonance-metastable energy levels [2:368F](#)
- self-terminating resonance-metastable pulsed metal vapor lasers [2:368–369](#)
- unstable resonator configuration [2:372F](#)
- metal-induced energy transfer microscopy (MIET microscopy) [5:180–183](#), [5:181F](#)
- metal-induced lateral crystallization (MILC) [3:14](#)
- metal-metal waveguide (MM waveguide) [1:384](#)
- metal-organic chemical vapor deposition (MOCVD) [4:245–246](#), [4:416–417](#), [5:293–294](#)
- metal-organic-frameworks (MOFs) [2:170](#)
- metal-oxide-semiconductor (MOS) [1:468](#)
- metal-oxide-semiconductor field effect transistor (MOSFET) [3:12–13](#)
- metal-semiconductor-metal detectors (MSM detectors) [5:8](#), [5:10](#), [5:10F](#)
- metal(s) [5:277](#)
- alloys [2:258–259](#)
- electrode [5:242](#)
- III-V semiconductors [2:257](#)
- to “less-metals” [2:259–261](#)
- metal-based sensitizers [5:276–277](#)
- metal-free cables [1:331](#)
- metal-free organic dyes [5:277](#), [5:277F](#)
- nanostructures [5:296–302](#), [5:297F](#), [5:298F](#)
- nitrides [2:260](#)
- NPs [5:298](#)
- semiconductors to [2:256–257](#), [2:256F](#)
- TCOs [2:257–258](#)
- vapor [2:368–369](#), [2:373](#)
- metal-insulator transition (MIT) [2:123](#), [2:126](#)
- metallic armor [1:332](#)
- metallic atoms [4:96](#)
- metallic barrier [1:331](#)
- metallic coatings [1:33](#)
- metallic counterparts [1:207](#)
- metallic silver particles [4:93](#)
- metal-ligand charge transfer (MLCT) [5:242](#)
- metallo-dielectric waveguides [1:217–218](#)
- metallo-organic compounds [2:346](#)
- metalorganic vapor phase epitaxy (MOVPE) [2:98–99](#), [2:273](#)
- metamaterials [2:263–264](#)
- metaplectic group [4:202](#)
- metastable lower laser level [2:368](#)
- metasurface VCSEL [1:384](#)
- methanol [2:119](#)
- 3-methoxypropionitrile (MPN) [5:277](#)
- methyl methacrylate (MMA) [2:347](#)
- 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) [3:125](#)
- methylammonium (CH₃NH₃⁺) [5:282](#)
- methylene blue [3:103F](#)
- methylnaphthalene (MNA) [2:7–8](#)
- metro and access networks, WDM in [1:280–281](#)
- metro area network (MAN) [4:17–18](#)
- metropolitan-area networks [1:281](#)
- MF concentricity error [1:322](#)
- Michelson interference rule [4:432](#)
- Michelson interferometer (MI) [1:34](#), [1:35F](#), [1:432](#), [1:433](#), [1:434](#), [1:434F](#), [1:435F](#), [4:500](#), [4:502–503](#)
- Michelson stellar interferometer [1:47](#), [1:47F](#)
- microchip laser [2:454](#)
- micro electro-mechanical systems (MEMS) [1:268](#), [3:36](#), [3:84](#), [4:99–100](#)
- projection display [3:35–37](#)
- Micro-Electromechanical Vertical Cavity Surface Emitting Lasers (MEM-VCSEL) [1:260](#)
- micro-interferences [4:140](#)
- micro-lidars [4:496](#)
- see also* light detection and ranging (LIDAR)
- acoustical studies and vibrometry [4:528](#)
- devices with photonic arrays [4:507–509](#)
- distance measurement techniques [4:497](#)
- flow cytometry [4:514–516](#)
- glucose lidars [4:522](#)
- interferometric methods [4:502–504](#)
- laser velocimetry [4:523–525](#)
- measuring media parameters [4:512–513](#)
- nephelometers [4:512–513](#)
- normal and abnormal cell with fused genes [4:516F](#)
- with optical doppler tomography [4:526](#)
- oxymetry lidars [4:519–522](#)
- Raman micro-lidars [4:516–518](#)
- smoke detectors [4:513–514](#)
- time-resolved fluorescence spectroscopy [4:518–519](#)
- for velocity measurement [4:523–525](#)
- wavefront sensing micro-lidars [4:530–531](#)
- micro-optic technologies [1:341](#)
- micro-opto-electro-mechanical systems (MOEMS) [4:99](#)
- microarray scanning and customized excitation [3:218](#)
- microbending [1:328](#)
- loss [1:256](#)
- sensitivity [1:311](#)
- microbolometers [1:423](#), [1:424F](#)

- microcavity [2:107](#), [5:313](#)
 luminescence intensities [1:462F](#)
 photoluminescence spectrum [1:461](#)
- microchip lasers [2:415](#)
see also thin disk laser (TDL)
 applications [2:422–423](#)
 background [2:415–416](#)
 composite-cavity microchip lasers [2:416](#)
 frequency conversion and amplification [2:421–422](#)
 monolithic microchip laser [2:415F](#), [2:416](#)
 polarization control [2:418](#)
 pulsed operation [2:418–419](#)
 pumping [2:416–417](#)
 spectral properties [2:417](#)
 transverse mode definition [2:417](#)
- microcup[®] EPD [3:80](#), [3:80F](#), [3:81F](#)
- microdisplay [3:32](#)
- microencapsulated EPD [3:79–80](#)
- microlasers [1:455–456](#)
- microlens array [5:235–237](#), [5:295](#), [5:296F](#)
- microlens array film (MAF) [5:248](#)
- microlenses [5:250–251](#)
- micromaser [1:452](#), [1:454F](#), [1:455](#)
- micron scale [5:292–293](#)
- microoptical coherence tomography (μ OCT) [4:507](#)
- microplasma [1:412](#)
- microring resonator (MRR) [4:29](#), [4:236](#), [4:237F](#), [4:293–294](#), [4:297F](#), [4:298F](#)
 optical switch [4:236](#), [4:237F](#)
- microscopic/microscopy [1:484](#), [3:241](#), [4:103](#), [5:185](#)
 confocal microscope [5:189–190](#), [5:189F](#)
 contrast mechanisms [5:192](#)
 conventional microscope [5:185–186](#), [5:189](#)
 corneal deformation [5:146](#)
 description [2:234–235](#)
 fluorescence microscopy [5:192](#)
 images [5:192](#)
 interaction processes [4:264](#)
 nonlinear microscopy [5:193](#)
 objectives [2:249](#), [5:188–189](#)
 other contrast mechanisms [5:193](#)
 parameters [2:244](#)
 perturbation [5:146](#)
 phase contrast microscopy [5:192](#)
 polarization [2:34](#)
 microscopy [5:192](#)
 probe microscopes [5:190](#)
 quantum many-body theory of FHG [2:34](#)
 quasi-particle dynamics [2:38](#)
 radiation types [5:191](#)
 Raman microscopy [5:192–193](#)
 scanning microscopes [5:189](#)
- microsecond pulse emission [2:336](#)
- Microsoft Kinect play station [4:511–512](#), [4:512F](#)
- microstructure fiber(s) [1:195](#)
see also photonic crystal fibers (PCFs)
 air-clad fiber [1:200–203](#)
 bandgap guided fibers [1:203](#)
 endlessly single-mode fiber [1:197–198](#), [1:198F](#)
 fabrication [1:195–196](#)
 fiber draw [1:196–197](#)
 high delta core fiber [1:198–199](#), [1:199F](#)
 index guided fibers [1:197–198](#)
 preform fabrication [1:195–196](#)
 TMF [1:199–200](#), [1:201F](#)
 tunable microstructure fiber devices [1:203–205](#)
 types [1:197–198](#)
- microstructure tunable filter [1:205](#)
- microwave
 holography [4:155](#)
 radiation [5:191](#)
 spectroscopy [1:100](#)
 tomography [3:196](#)
- microwave amplification by stimulated emission of radiation (MASER) [2:325–326](#)
- microwave kinetic inductance detectors (MKIDs) [1:428](#)
- Microwave Photonics (MWP) [1:252–253](#)
- mid infrared (MIR) [2:166](#), [2:197](#), [2:260](#), [2:290](#)
 CO₂ laser [1:381](#)
 nonlinear generation of THz radiation in mid-IR QC-lasers [1:384–385](#)
 pulse [2:166](#)
 THz to [2:135–139](#)
- mid-IR *see* mid infrared (MIR)
- mid-to-far-infrared (IR) frequencies [2:124](#)
- mid-wavelength infrared (MWIR) [3:229](#), [3:232](#)
- middle cerebral artery occlusion (MCAO) [3:121–122](#)
- Middle East (ME) [4:18](#)
- middle-wavelength-cone (M-cone) [3:17](#), [3:18F](#)
- mini-scleral lenses [5:119](#)
- minimum Gibbs free energy [3:81](#)
- minimum resolvable temperature difference (MRTD) [3:234](#), [3:235–237](#), [3:237F](#)
- minimum resolvable temperature (MRT) [3:234](#)
- minimum temperature difference perceived (MTDP) [3:234](#)
- mirrorless lasing [5:312–313](#)
- mirrors [1:376](#), [5:203](#)
 array [4:232](#)
 fiber reflectivity [1:213](#)
 optical characterization of mirror fibers [1:213](#)
 planes [1:170](#)
- mitochondrial membrane potential (MMP) [3:115](#)
- modal bandwidth [1:312](#), [1:313F](#)
- modal control methods [3:202](#)
- modal dispersion [1:222–223](#)
- mode field diameter (MFD) [1:322](#), [1:325](#)
- mode locked fiber laser (MLFL) [4:45](#)
- mode locking [2:421](#)
 of TDLs [2:412](#)
- mode management in PT-symmetric lasers [4:293–294](#), [4:295F](#), [4:296F](#)
 laser transverse mode filtering [4:294–298](#)
 single longitudinal mode lasing [4:294](#)
- mode multiplexing algorithms [3:153](#)
- mode suppression, single mode operation with [2:453–454](#)
- mode-propagation constant [1:178–179](#)
- mode-selective excitation [2:139–141](#), [2:141F](#)
- model-based methods [4:490](#)
- modeling infrared imagers [3:238–239](#)
- modelocking (ML) [2:435](#), [2:440–441](#), [2:468](#)
 based multi-wavelength interferometry [4:432](#), [4:432F](#)
 control knobs for frequency comb [2:305–306](#)
 as frequency comb generator [2:303](#)
 IPI [2:308–309](#)
 lasers [2:302](#)
 mode-locking as orthodoxy [2:303–304](#)
 phase modulation and dispersion [2:304–305](#)
 solid state lasers [2:424](#)
 typical laser configurations [2:303](#), [2:303F](#)
- modern high-resolution CCD [4:140](#)
- modern imaging [3:144](#)
- modern optical systems [3:262](#)
- modern THz TDS [2:19](#)
- modes [1:6–7](#), [1:177](#)
- modest pulse energies [2:424](#)
- modified chemical vapor deposition process (MCVD process) [1:28](#), [1:28–29](#), [1:29F](#)
- modified Gauss–Newton methods [3:150](#)
- modified monovision technique [5:125](#)
- modified poly(methyl methacrylate) (MPMMA) [2:347](#)
- modified Schrödinger equations [2:266](#)
- modulated laser [1:141–142](#), [1:142](#)
- modulation [1:373](#), [3:242](#), [4:95](#)
- modulation bandwidth [1:139–140](#)
- modulation transfer function (MTF) [3:233](#), [3:235–237](#), [3:236](#), [3:237F](#), [4:412](#), [4:413F](#), [4:533](#), [5:16](#), [5:122](#)
- modulational instability (MI) [1:348](#), [2:426–427](#), [2:427](#)
- modulator [1:139–140](#), [4:216–217](#), [4:217F](#)
- molecular beam epitaxy (MBE) [1:422](#), [2:94](#), [2:98–99](#), [2:462](#), [4:416–417](#)
- molecular beam intensity [1:98](#)
- molecular dynamics [2:326–327](#)
 collision-induced vibrational relaxation of upper and lower laser levels [2:328](#)
 direct excitation and de-excitation [2:327](#)
 electron energy distribution function [2:327F](#)
 gain [2:328](#)
 radiative relaxation [2:328](#)
 resonant energy transfer [2:327–328](#)
 stimulated emission [2:328–329](#)
- molecular electrolytes
 ion structure and dynamics of electrolytes [2:170–172](#)
 time dependent 2D IR spectral series [2:171F](#)
- molecular energy levels [2:337](#), [2:338F](#)
- molecular gas lasers [1:380](#)
- molecular moment of inertia [1:387](#)
- molecular orbital (MO) [5:220–221](#)
- molecular rate equations [2:329](#)
- molecular spectra [1:387](#)

- molecular spontaneous emission
modification [1:445–446](#)
- molecular vibrational fingerprints [2:164](#)
[1:446–447](#)
- Mollow triplet [2:47–48](#)
- momentum [1:487](#), [5:207](#), [5:211F](#)
conservation [2:115–116](#)
diagram for diffraction of light [1:133F](#)
- mono-mode optical fiber [4:38](#)
- monochromatic/monochrome
aberrations [5:55–58](#), [5:91](#), [5:120](#)
light wave field [1:291](#)
material [4:88](#)
microencapsulated EPD [3:80](#), [3:80F](#)
- monocular vision [5:44](#)
- monolayer TMDs
coherem energy transfer [2:59–61](#)
homogeneous broadening [2:58–59](#)
incoherent energy transfer [2:59–61](#)
inhomogeneous broadening [2:58–59](#)
intervalley biexcitons [2:61–62](#)
optical properties [2:57–58](#)
2DCS [2:58–59](#)
- monolayer transition metal dichalcogenides [2:53](#)
- monolithic fabrication of integrated sensor [3:89](#)
- monolithic InP PIC [4:250](#), [4:251F](#)
- monolithic integration [4:242](#)
- monolithic microchip lasers [2:415F](#), [2:416](#)
- monolithic microwave integrated circuits (MMICs) [5:1](#)
- monolithically-integrated detector arrays [1:419](#)
- monomolecular process [5:257](#)
- monotropic LC mesophase [3:8](#)
- monovision [5:54](#), [5:125](#)
- Monte Carlo simulations [4:422–423](#), [4:423F](#)
- Morpheus lander [4:450–451](#), [4:451F](#)
- MOS Field Effect Transistor (MOSFET) [1:468](#)
- moth eye AR structures [5:293](#), [5:294F](#)
- motion blur [3:77](#)
- motion parallax [3:44–45](#)
- motorized instruments [3:197](#)
- mounting [2:410](#), [2:411](#)
- movable digital mirror display (DMD) [3:2](#)
- moving waveguide coupler [4:233](#)
- Mueller calculus [1:160](#), [1:161](#), [1:165–166](#),
[1:165F](#), [1:166F](#)
- Multi-Core Fibers (MCF) [1:262](#)
- multi-dimensional coherent spectroscopy (MDCS) [2:53](#)
interpreting 2D spectra [2:157](#)
power of 2DCS [2:150](#)
quantum pathway selection in 2D coherent spectra [2:153–154](#)
theory of coherent multidimensional spectroscopy [2:150–151](#)
3-pulse transient four-wave mixing [2:156–157](#)
- multi-dimensional laser radars [4:444](#),
[4:444F](#)
adding velocity as dimension [4:449–452](#)
dimension combinations [4:470–471](#)
fluctuation as dimension [4:469–470](#)
- histogram of range image [4:445F](#)
- media parameters as dimensions
[4:455–463](#)
- ranging supplemented by two-dimensional scanning [4:444–445](#)
- stepping up dimensionality [4:444–445](#)
- three-dimensional description of parameters [4:444F](#)
- 2D imaging supplemented by scanning in z direction [4:445–449](#)
- 2D projections of objects [4:446F](#)
- vibrations [4:452–455](#)
- multi-domain vertical alignment (MVA) [3:59](#), [3:61](#)
- multi-electrode driven (MeD) [3:48](#)
- multi-fluorescence in situ hybridization system (M-FISH system) [3:213](#)
- multi-frame blind deconvolution algorithm (MFB algorithm) [3:276–277](#)
- multi-modal dispersion [1:178–179](#)
- multi-mode interferometers (MMI) [4:236](#)
- multi-mode THz QC-lasers [1:384](#)
- multi-order Raman sideband generation [4:345](#)
- multi-path interference [1:272](#)
- multi-planar [3:46–47](#)
- multi-point control protocol (MPCP) [2:76](#),
[2:77F](#)
- multi-slice algorithms [3:153–154](#)
- multi-spin encodings enable initialization and measurement [1:473](#)
- multi-stage OPA system [2:166](#)
- multi-transmitter, single receive aperture system [4:409](#)
- multi-wavelength laser source [4:222](#)
- multicomponent glasses [1:27](#)
- multiconjugate adaptive optics (MCAO) [3:202](#)
- multidimensional fourier transform spectroscopy [4:275–276](#)
- multidimensional nuclear magnetic resonance (multidimensional NMR) [2:150](#)
- multidimensional spectroscopy [2:150](#), [2:249](#)
- multidimensional THz spectroscopy
see also 2D electronic spectroscopy (2DES); attosecond spectroscopy
collinear multidimensional THz spectroscopy [2:197–199](#)
three-pulse 2D THz spectroscopy on InSb [2:203–205](#)
THz-driven AC stark effect of intersubband transitions [2:200–203](#)
two-dimensional THz spectroscopy with three phase-locked THz pulses [2:199–200](#)
- multidrug-resistant (MDR) [3:102–103](#)
- multifaceted transmission function [4:184–185](#)
- multifiber connectors [1:338–339](#), [1:339F](#)
- multifocal wavefront aberrations [5:122–125](#),
[5:124F](#)
- multilayer dielectric mirrors [1:207](#)
- multilayer preform
fabrication sequence [1:211](#), [1:212F](#)
rod [1:211](#)
- multimode core [1:203](#)
- multimode fiber(s) [1:333](#), [1:223](#), [1:336](#),
[1:337](#)
bandwidth for multimode fibers [1:312–313](#)
- multimode interference switches (MMI switches) [1:143](#)
- multimode operation [2:451–452](#)
- multipass amplification [2:112](#)
- multipass unstable resonator [2:334](#)
- multipath interference (MPI) [1:234](#), [4:226](#)
- multiphoton
absorption [4:336](#), [5:92](#)
excitation [5:89](#)
- multiphoton intrapulse interference phase scan (MIIPS) [5:91](#)
- multiphoton ionization process [4:337](#)
rate [4:304](#)
- multiphoton microscopy [5:98](#), [5:99F](#)
of cornea [5:99–100](#)
future directions [5:107](#)
- multiphoton retinal imaging [5:87](#)
theory [5:87–89](#)
- multiple beamsplitters [1:59](#)
- multiple input, multiple output lidar (MIMO lidar) [4:407–408](#)
closed form MIMO [4:408–412](#)
multiple sub-aperture overlap using 3 transmitters [4:409F](#)
single receive aperture, light green, with three transmitter apertures [4:410F](#)
trade space of current MIMO [4:412–413](#)
- multiple path interference (MPI) [1:358](#)
- multiple quantum well (MQW) [2:23](#),
[2:273–274](#)
- multiple receive apertures [4:407](#)
- multiple reference OCT system (MR-OCT system) [4:507](#)
- multiple scattering [3:136–137](#), [3:144](#)
acousto-optic [3:226–228](#)
coupling light and sound [3:226](#)
imaging in multiple scattering regime [3:225–226](#)
inverse source and scattering problems [3:137](#), [3:137F](#)
light [3:225–226](#)
photo-acoustic [3:226](#)
resolution in scatterer imaging [3:142–144](#)
- multiple transmitter apertures [4:407](#)
- multiple-access mechanism [1:255](#)
- multiple-inputs multiple outputs (MIMO) [1:262](#), [1:278](#)
- multiple-prism array [2:343](#)
- multiple-prism dispersion grating theory [2:341–342](#), [2:342F](#)
- multiple-prism grating cavities [2:337](#)
- multiple-prism Littrow grating laser oscillator (MPI grating laser oscillator) [2:343](#), [2:343F](#)
- multiple-quantum-well (MQW) [1:245](#),
[4:243](#), [4:250](#), [4:255–256](#)
- multiple-section diode lasers [2:468](#)
- multiple-stage EDFAs [1:358–361](#), [1:359F](#),
[1:360F](#)

multiplex imaging [3:250–252](#)
see also interferometric imaging
 electromagnetic waves propagation [3:248–249](#)
 high resolution imaging in astronomy [3:253–254](#)
 spatially coherent light [3:249](#)
 spatially incoherent light [3:249–250](#)
 synthetic apertures with spatially incoherent objects [3:250–252](#)
 theta rotating interferometer [3:252–253](#)
 multiplexers [4:38–39](#), [4:39F](#)
 multiplexing techniques [4:34](#)
 multiplied shot noise [4:425](#)
 multipole method [1:197](#)
 multispectral imager [3:211](#)
 multispectral imaging systems [3:211–213](#), [3:212F](#)
 Munich interferometers [1:433–434](#)
 muscle performance, PBM for [3:126](#)
 mutual coherence [1:36–37](#)
 spectral representation [1:38](#)
 mutual intensity (MI) [1:37–38](#), [3:263](#), [4:164–165](#), [4:174](#)
 ‘mutual PC’ *see* ‘double PC’
 Mux/DMux stages [1:370](#)
 myopia [5:51](#), [5:141](#)
 myopic eye [5:116](#)

N

N rectangular slits [1:73–74](#)
N-alkylbenzimidazole derivatives [5:277](#)
N-atom system [1:460–461](#)
n-butyl benzyl phthalate (BBP) [4:96–97](#)
n-dimensional (nD) [4:174](#)
n-doped semiconductor [3:87–88](#)
N-ethylcarbazole (ECZ) [4:96–97](#)
n-FFS [3:60–61](#)
N-methylpyrrolidone (NMP) [5:277](#)
 ti-type double heterostructure [2:103–104](#)
 ti-type OSCs [5:226–227](#)
 N3 dye [5:276–277](#), [5:276F](#)
 N719 dye [5:276–277](#), [5:276F](#)
 NAD-phosphate, reduced form (NADPH) [3:101](#)
 nano-textures [5:295](#)
 nanocavity systems [1:460–461](#)
 nanocomposites [5:300](#)
 nanocrystalline TiO₂ [5:274–275](#)
 nanofabricated atom optics [5:204–207](#), [5:204F](#), [5:205F](#)
 diffraction data from beam of sodium formed [5:206F](#)
 diffraction data using mixed beam of sodium atoms [5:206F](#)
 nanofabrication techniques [5:204](#)
 nanogratings [5:296–298](#), [5:300](#)
 nanoimprint lithography [5:293](#)
 nanoparticles (NPs) [5:296–298](#)
 nanophotonics [2:252](#)
 light trapping approaches [5:294–295](#), [5:295F](#)
 nanoporous TiO₂ electrodes [5:270](#)
 nanorods (NRs) [5:298](#)

nanoscale materials [2:168–170](#)
 nanostructures [2:86](#)
 Nanowire arrays (NW arrays) [5:293–294](#), [5:296–298](#), [5:300](#)
 naphthalene diimides fused with two 2-(1,3-dithiol-2-ylidene)malononitrile moieties (NDI-DITYM2) [5:325–326](#)
 1-naphthalenethiol (SH-NA) [5:263](#)
 narrow gap semiconductors (NGSs) [2:88–89](#), [2:88–89](#)
 narrow laser beam [5:112–113](#)
 narrow linewidth high power SG-DBR laser [4:249–250](#), [4:250F](#)
 narrow localization [5:99](#)
 narrow viewing angle (NVA) [3:62](#)
 narrow-linewidth solid-state dye lasers [2:348](#)
 oscillator [2:332](#)
 National Geospatial Program (NGP) [4:401](#)
 National Geospatial-Intelligence Agency (NGA) [4:402](#)
 national institute of information and communications technology (NICT) [4:117](#), [4:117F](#)
 National Renewable Energy Laboratory (NREL) [5:256](#), [5:257F](#), [5:286](#)
 National Television Standards Committee (NTSC) [5:243–244](#)
 “NATO Maritime Target ID” [4:478–479](#)
 natural guide star (NGS) *see* off-axis reference star
 natural killer cells (NK cells) [3:113](#)
 NbN HEB mixer chip [1:429](#), [1:430F](#)
 Nd doping [2:413](#)
 Nd:YLF laser [4:364](#)
 Nd:YVO₄ laser [4:364](#)
 near infrared (NIR) [2:75](#), [2:252](#), [2:394](#), [3:97](#), [3:124](#), [3:226](#), [3:229](#), [5:269](#)
 DBR lasers [4:362](#)
 ECSL [4:362–363](#)
 imaging systems [3:272](#)
 lasers [4:361–362](#)
 light [5:93](#)
 Nd:YAG laser [4:363–364](#)
 Nd:YVO₄ and Nd:YLF lasers [4:364](#)
 pump beam [2:121–122](#)
 radiation [5:106–107](#)
 Ti:Sapphire laser [4:364](#)
 near-field CGH [4:35](#)
 near-field light distribution *see* TNF image technique
 near-field optics techniques [3:144](#)
 near-field scanning microscopy [3:144](#)
 near-field scanning optical microscope (NSOM) [5:190–191](#), [5:190F](#), [5:300](#)
 nearest-neighbor periodicity [1:173](#)
 necrosis, PDT and [3:107–108](#)
 necrotic cell death [3:107](#)
 negative dielectric liquid crystal in ASM (N-type ASM) [3:61](#)
 negative magnification [3:166](#)
 nematic LC modes [3:60–61](#)
 nematic liquid crystals [3:8–9](#)
 neodymium [2:413](#)
 neodymium:yttrium-aluminum-garnet (Nd:YAG laser) [2:337](#), [4:363–364](#), [4:364F](#), [5:134](#), [5:134–135](#)

neon
 buffer gas [2:370](#)
 ion lasers [2:379](#)
 nephelometers [4:512–513](#)
 networking issues, OTDM [4:65–66](#), [4:65F](#)
 neurodegenerative diseases, PBM for [3:124–125](#)
 Alzheimer’s disease [3:124–125](#)
 Parkinson’s disease [3:125](#)
 neurological severity score (NSS) [3:123](#)
 Neurothera Effectiveness and Safety Trials (NEST) [3:122–123](#)
 neurovascular coupling [5:64](#)
 neutral atom microscopes [5:192](#)
 neural particles [5:192](#)
 neutrons [2:207](#), [5:192](#)
 neutrophil with fluorescent markers [4:516](#), [4:517F](#)
 New European Wind Atlas project (NEWA project) [4:461–462](#)
 next-generation PONs (NG-PON) [2:73](#), [2:74](#)
 coexistence [2:78](#)
 energy-efficient dynamic bandwidth allocation [2:79–80](#)
 10GE-PON [2:77–78](#), [2:78F](#)
 high capacity [2:80–81](#)
 NG-PON1 [2:73](#), [2:74](#)
 NG-PON2 [2:74](#)
 traffic flow of OFF-DWBA algorithm [2:80F](#)
 TWDM-PON system [2:78–79](#), [2:79F](#)
 wavelength allocation [2:78](#)
 XG-PON [2:77–78](#), [2:78F](#)
 NF-κB [3:115–116](#)
 Ni cathode [2:331](#)
 Nicol types [1:150–151](#)
 Nicolet SpectraTech NicPlan [1:215](#)
 nicotinamide adenine dinucleotide (NADH) [3:101](#)
 nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) [3:86–87](#), [5:86](#)
 ninth-order vortex CGH [4:37F](#)
 NIST [4:436](#)
 nitric oxide (NO) [3:115](#), [3:120](#)
 nitrogen (N) [2:271](#), [2:327](#)
 laser [2:426](#)
 nitrogen-containing heterocyclic compounds [5:277](#)
 noble gas [2:373](#)
 noble gas ion lasers
 see also up-conversion lasers
 Ar II laser blue-green wavelengths [2:377F](#)
 energy levels in neutral and singly ionized argon [2:377F](#)
 4p and 4s doublet levels in singly ionized argon [2:376F](#)
 Krypton ion laser wavelengths [2:379F](#)
 operating characteristics [2:379](#), [2:380F](#)
 technology [2:379–382](#), [2:380F](#), [2:381F](#)
 theory of operation [2:376–379](#)
 ultraviolet argon ion laser wavelengths [2:378F](#)
 nodal points and planes [1:107–109](#)
 noise
 in cascaded optical amplifiers [1:273](#)
 noise-free image [3:188](#)

- in optical transmission systems [1:272–275](#)
sources [1:437–442](#)
spectral intensity theorem [4:428](#)
suppression [2:190](#)
- noise equivalent power (NEP) [1:420](#),
[4:475](#)
- noise equivalent temperature difference
(NE Δ T) [1:419](#), [3:233](#), [3:235–237](#),
[3:236F](#)
- noise factor (F) [1:242–243](#), [4:428](#)
- noise figure (NF) [1:242](#), [1:357](#), [1:361–362](#),
[4:5](#), [4:431](#)
- nominal pulse density (NPD) [4:401](#)
- nominal pulse spacing (NPS) [4:401](#)
- non return to zero (NRZ) [1:365](#)
- non-confocal “split detection” [5:76–77](#)
- non-Doppler incoherent lidars [4:463](#)
- non-emissive display [3:1](#), [3:5–6](#)
- non-fiber-related effects [1:271–272](#)
linear crosstalk [1:271–272](#)
noise in optical transmission systems
[1:272–275](#)
- non-Hermitian operators [4:291](#)
- non-integrable situation [2:65](#)
- “non-interferometric” techniques [3:146](#)
- non-invasive vision correction [5:126–127](#)
- non-orthogonal functions [4:194–195](#)
- non-parabolic bands [2:46–47](#), [2:88–89](#)
- non-perturbative phenomena [2:54](#)
- non-perturbative quantum interference
[2:35–36](#), [2:35F](#)
- non-polarizing beam splitter (NBS) [4:502](#)
- non-radiative (NR) [5:241](#)
Auger effects [2:114–115](#)
decay processes [5:225–226](#)
identification of Auger effects [2:116–117](#)
impact ionization [2:115–116](#)
processes [2:114–115](#)
recombination [2:463–464](#), [2:464](#)
relaxation [2:442–444](#)
transition [2:3](#)
- non-relativistic classical mechanics [4:199](#)
- non-resonant
excitations [1:449](#), [2:33–34](#), [2:33](#)
phenomena [1:452](#)
THz detection [1:423](#)
- non-return-to-zero (NRZ) [1:247](#), [1:275](#),
[1:294](#), [4:6](#), [4:7F](#), [4:68](#)
data [4:21–22](#)
modulation format [2:74](#)
pulses [1:20–21](#)
- non-Telecentric architectures [3:36](#)
- non-zero dispersion shifted fiber (NZDSF)
[1:9](#), [1:21](#), [1:23](#), [1:261](#), [1:262](#), [4:16](#),
[4:24–25](#)
- noncentrosymmetric materials [1:489–490](#),
[1:490](#)
- noncollinear AOTF orientation [1:138F](#)
- noncollinear devices [1:139](#)
- noncollinear optical gating (NOC) [2:321](#)
- noncollinear optical parametric amplifier
(NOPA) [2:298–299](#), [2:299F](#)
- noncontact 3D metrology [4:439](#)
- noncontinuity [3:158](#)
- nondegenerate four-wave mixing [4:314–315](#),
[4:315F](#)
- nondegenerate FWM [1:250–251](#), [1:349](#),
[1:349F](#)
- nonequilibrium statistics [2:111](#)
- nonexistence [3:158](#)
- nonhyperspectral method [3:213](#)
- nonideal amplifier [1:352](#)
- nonlinear (NL) [4:311](#)
absorption [4:351–352](#), [4:352F](#)
applications of fiber gratings [1:239–240](#)
equations [4:329–330](#)
generation of THz radiation in mid-IR QCL
lasers [1:384–385](#), [1:385F](#)
impairments [1:305–306](#)
luminescence transition [1:461](#)
microscopy [5:193](#)
nonlinear-fiber optics [1:177](#)
ophthalmoscopy theory [5:85–86](#)
phenomena [1:180–181](#)
polarization [2:244](#), [4:335–336](#), [4:349](#)
process [1:407](#)
quantum electrodynamics [4:338](#)
refraction [4:350–351](#)
refractive index [1:177](#), [4:335–336](#)
response [1:345–346](#), [1:368](#), [2:23–25](#),
[2:23F](#)
second-order parametric interaction
[2:291](#)
signatures [1:345–346](#), [1:345F](#)
spectroscopy techniques [2:53](#)
systems [3:144](#)
technique [2:82](#)
THz spectroscopy [2:197](#)
- nonlinear coefficient [1:320–321](#), [2:312](#)
methods for measuring [1:321](#)
- Nonlinear Compensation (NLC) [1:278](#)
- nonlinear crystals [4:287](#)
OR [1:405–409](#), [1:406F](#), [1:407F](#)
organic crystals [1:409–411](#)
tilted pulse front [1:409](#), [1:410F](#)
- nonlinear effect [1:255](#), [1:320–321](#), [2:15](#),
[4:54](#), [4:220–221](#), [4:221F](#), [4:337–338](#)
- FWM [1:321–322](#)
linear and nonlinear signatures [1:345–346](#)
- nonlinear coefficient [1:320–321](#)
- parametric phenomena in optical fibers
[1:347–348](#)
- physical origin of optical nonlinearity
[1:346–347](#)
- SBS [1:321](#)
- SPM [1:321](#)
- SRS [1:321](#)
- XPM [1:321](#)
- nonlinear fiber effects [1:257–259](#)
dominance [1:258F](#)
values of nonlinearity for various types of
transmission fiber [1:258F](#)
- nonlinear frequency
conversion methods [2:145–146](#)
generation [2:421–422](#)
- nonlinear optical
effects [1:491](#), [2:424](#), [4:348](#)
frequency mixing schemes [4:339](#)
interactions [1:10–11](#)
phenomena [4:347](#)
processes [4:342–344](#), [4:343F](#)
techniques [2:348](#)
- nonlinear optical loop mirror (NOLM)
[1:368](#), [4:41](#), [4:41–42](#), [4:41F](#), [4:42F](#),
[4:64](#)
- nonlinear optical medium, nomenclature
with [4:349–350](#)
complex quantities [4:350](#)
nonlinear absorption [4:351–352](#), [4:352F](#)
nonlinear refraction [4:350–351](#)
nonlinear susceptibility [4:349–350](#)
- nonlinear optical PC
collimated beam [4:322F](#)
mechanism of double PC [4:325F](#)
recording of hologram [4:323F](#)
SBS [4:322F](#)
tail-biting scheme of phase conjugation
[4:324F](#)
- nonlinear optical spectroscopy [2:245](#)
in chemistry [2:245–246](#)
harmonic conversion [2:246](#)
higher-order and multidimensional
spectroscopy [2:249](#)
second-order spectroscopy [2:245–246](#)
spatially resolved spectroscopy [2:249–250](#)
surface second-order spectroscopy
[2:246–247](#)
techniques [2:53](#)
third-order spectroscopy [2:247–248](#)
ultrafast time resolved spectroscopy
[2:248–249](#)
- nonlinear optics [1:10](#), [1:345](#), [1:321](#), [4:335](#),
[4:350](#)
see also electro-optics (EO)
“black box” approach [4:321](#)
causal linear system [4:318F](#)
implication of causality [4:320–321](#)
KK relations [4:317](#)
nonlinear absorption coefficient [4:319F](#)
nonlinear polarization [1:11–12](#)
planar waveguide [1:11F](#)
in Raman resonance [4:344–345](#)
SHG in crystals [1:12–16](#), [1:13F](#)
texts [4:317](#)
third-order nonlinear effects [1:16–18](#)
two-photon absorption coefficient in
semiconductors [4:321F](#)
for weak CW fields [4:345–346](#), [4:345F](#)
- Nonlinear Phase Noise (NLPN) [1:306](#)
- nonlinear Schrödinger equation (NLSE)
[1:18–19](#), [1:258](#), [2:425–426](#),
[4:279–280](#)
- nonlinear susceptibility [4:342](#), [4:349–350](#)
properties [4:311](#)
- nonlinear tolerance (NLT) [1:283](#)
- nonlinear-optical FWM [4:323](#)
- nonlinearity [1:18](#), [1:172](#), [2:394](#)
forms of [4:282](#)
- nonplanar ring oscillators (NPRO)
[1:440–441](#), [2:454](#), [2:454–455](#), [2:454F](#)
- nonpolar InGaIn layers [2:279](#)
- nonradiating sources [3:139](#), [3:160](#)
- nonreciprocal two-port optical device [1:343](#)
- nonrelativistic Schrödinger equation
[2:234–235](#)
- nonrephasing
complex signal fields [2:184–185](#)
signals [2:185–186](#)

nonrephasing (*continued*)

spectrum [2:56](#)

2D spectra [2:154–155](#)

nonspherical surfaces [1:121–122](#)

nonthermal regime [2:82, 2:84](#)

nonuniqueness in imaging [3:156](#)

inverse source problem [3:159–161, 3:162F](#)

matrix formulation of imaging and inverse problems [3:156–157](#)

nonuniqueness and prior knowledge [3:157–159](#)

phase retrieval [3:161–163](#)

nonvisible radiation [5:185](#)

nonwavelength-selective branching devices [1:336, 1:340–341](#)

normal dispersion [1:5, 1:195, 1:315](#)

normal operation [2:76, 2:77F](#)

normal-mode coupling [1:459, 1:462](#)

normalization [5:56](#)

normalized shift [1:216, 1:217F](#)

Normative References (N.R.) [1:286](#)

novelty filter [4:332–333](#)

nuclear

gamma irradiation [1:326–327](#)

motion control [5:39](#)

shrinkage [3:107](#)

spin of phosphorus atom [1:474](#)

nuclear magnetic resonance (NMR) [2:197](#)

spectroscopy [2:249](#)

spectroscopy studies [2:174](#)

number of degrees of freedom (NDF) [3:137](#)

numerical aperture (NA) [1:201, 1:213, 1:215, 1:223, 1:323, 1:324–325, 3:155, 5:177–178, 5:186](#)

numerical correction (NC) [4:500](#)

numerical reconstruction

by convolution approach [4:144F, 4:145–146](#)

by lensless Fourier approach [4:146–147](#)

NV-THERM model [3:239](#)

Nyquist condition [4:432–433](#)

Nyquist interval [4:376](#)

Nyquist WDM [1:275](#)

O

object plane [4:369–371](#)

imaging system [4:370F](#)

object signal (OS) [4:500](#)

object space [3:158](#)

focal length [1:101–102](#)

object wave [1:376](#)

object-image matrix [1:105–106](#)

obliquity factor [1:82–87, 1:84F](#)

OCT angiography (OCTA) [5:64–65, 5:161–162](#)

1,8-octanedithiol (ODT) [5:262–263](#)

ocular aberrations [5:108](#)

ocular coherence tomography (OCT) [5:142](#)

capillary velocimetry [4:526, 4:526F](#)

ocular target tissue [5:133–134](#)

odd-parity hidden order in Sr_2IrO_4 [2:217–220, 2:218F, 2:219F](#)

OFDMA [1:255](#)

off-axis

geometry [4:102](#)

hologram [4:107, 4:108F](#)

method [3:207](#)

reference star [3:202](#)

system [4:102–103](#)

technique [4:151](#)

off-axis JPRG [4:369, 4:374–375, 4:381, 4:382F](#)

estimation [4:376–377](#)

Fourier plane [4:374–375](#)

MATLAB[®] code to exploring [4:386–395](#)

sampling quotient [4:375–376](#)

shot-noise limit [4:377](#)

SNRs [4:376–377](#)

off-axis PPRC [4:369, 4:377–378, 4:381–385, 4:383F](#)

estimation [4:378–379](#)

Fourier plane [4:377–378](#)

MATLAB[®] code to exploring [4:386–395](#)

sampling quotient [4:378](#)

shot-noise limit [4:379](#)

SNRs [4:379](#)

off-axis reference beam hologram (OAIB) [4:106, 4:107, 4:109, 2:80F](#)

OFF-DWBA algorithm, traffic flow of [2:80, 2:80F](#)

off-resonant interaction sequences [2:203](#)

“off-the-shell” components of Fourier spectrum [3:139](#)

“off-set quantum-well” technology [4:244](#)

OHF-RIK S method [2:248](#)

omnidirectional range [1:209](#)

omnidirectional reflectance [1:207](#)

omnidirectional reflection [1:209](#)

omnidirectional surfaces and fibers

bandstructure for multilayer fibers for external reflection applications [1:211–213](#)

fabrication approach [1:210, 1:210F](#)

materials selection criteria [1:210–211](#)

mechanical tuning experiment and discussion [1:216, 1:216F](#)

mesostructured fibers for external reflection applications [1:209–210](#)

mirror fibers, optical characterization of [1:213](#)

omnidirectional reflecting mirrors [1:207–209](#)

preform construction process and fiber draw [1:211](#)

simulation of opto-mechanical behavior of fabry-perot fibers [1:215–216](#)

structure and optical properties of fabry-perot fibers [1:214–215, 1:214F](#)

“tunable fabry-perot” fibers [1:213–214](#)

wavelength-scalable hollow optical fibers [1:216–220, 1:220F](#)

on-axis PSRC [4:369, 4:379–380, 4:384F, 4:385](#)

estimation [4:380](#)

MATLAB[®] code to exploring [4:386–395](#)

shot-noise limit [4:380–381](#)

SNRs [4:380](#)

on-chip Raman lasers [2:267–268](#)

Feynman diagram [2:268F](#)

on-off keying (OOK) [1:265, 1:275, 4:242](#)

“on-the-shell” components of Fourier spectrum [3:139](#)

On/Off Keying (OOK) [1:291](#)

one dimension (1D) [2:184](#)

artificial molecules [1:382](#)

beam characterization [4:176–177](#)

Doppler-resolved projections [4:488](#)

Fourier transform spectroscopy [2:151](#)

laser techniques [4:474](#)

1D TINTs [5:275](#)

nonlinear spectroscopy [2:53](#)

photonic crystals [1:169](#)

PST for [4:176–177](#)

system [1:170](#)

unbounded disordered systems [4:278](#)

Walsh functions [5:17–18](#)

one way phase conjugation [3:209–210, 3:209F](#)

one-atom maser [1:452, 1:454F](#)

one-time pad [1:481](#)

opaque screen [1:94–95](#)

open-circuit voltage (V_{oc}) [5:258, 5:258–259, 5:270](#)

open-loop techniques [5:39](#)

operational expenditure (OPEx) [2:74, 2:82](#)

operational functionality [1:459](#)

ophthalmic AO systems [5:72](#)

ophthalmic lenses [5:54, 5:118](#)

ophthalmology [3:199, 3:203](#)

laser and light interaction and ocular target tissue [5:133–134](#)

light and laser therapy [5:130–131](#)

ophthalmoscopy [5:72](#)

OPScan series [2:287–288](#)

OPSilent series [2:287–288](#)

optic axis [1:150](#)

optical 2DCS [2:62](#)

coherent light-matter interaction [2:53–56](#)

experimental implementation of 2DCS [2:56–57, 2:57F](#)

optical 3R regenerator [1:368](#)

optical absorption [1:142, 2:105–106, 2:106, 2:107–108](#)

spectrum [2:107](#)

optical addressed spatial light modulators (OASLMs) [4:114](#)

optical amplifier (OA) [1:263–265, 1:343, 1:362](#)

optical axis [2:329, 3:198, 5:43, 5:47–48](#)

Optical Band-Pass Filter (OBPF) [1:277](#)

optical beam induced current (OBIC) [5:193](#)

optical bench metrology of IOIs [5:124–125](#)

optical biopsy [4:516–517](#)

optical Bloch equations (OBEs) [2:27–28, 2:53, 2:54, 2:82, 2:152, 4:265](#)

optical carrier-level N signal (OC-N signal) [4:17](#)

optical cavity [1:214, 2:280, 2:351](#)

Optical Channel Data Unit (ODU) [1:287](#)

Optical Channel Payload Unit (OPLU) [1:287](#)

Optical Channel Transport Unit (OTU) [1:287](#)

Optical Channels (OCh) [1:287](#)

optical characteristics of display [3:70](#)

- ambient illumination and reflection [3:72–73](#)
- BRDF [3:73–74](#), [3:74F](#)
- color [3:71](#)
 - temperature [3:75–76](#)
 - triangle [3:76](#), [3:77F](#)
- contrast ratio [3:72](#)
- gloss [3:71](#)
- illuminance [3:70–71](#)
- Lambertian surface [3:74](#)
- luminance [3:70](#)
- PPI [3:75](#)
- radiometric and photometric units [3:70](#)
- reflectivity [3:71](#)
- refresh rate and frame rate [3:77–78](#)
- resolution and image sharpness [3:75](#)
- response time [3:76–77](#)
- uniformity [3:71–72](#)
- viewing angle [3:74–75](#)
- viewing cone [3:74–75](#)
- optical characterization [2:258](#)
- optical circulator, principle scheme of [1:342](#), [1:343F](#)
- optical clock recovery [1:369](#)
- optical coherence
 - see also* coherence
 - applications [1:46–47](#)
 - interference spectroscopy [1:47–48](#)
 - Stellar interferometry [1:47](#)
 - theory [1:34](#)
- optical coherence elastography (OCE) [5:146](#)
- optical coherence tomography (OCT)
 - [3:222–223](#), [3:223F](#), [3:224F](#), [4:174](#), [4:504](#), [5:64](#), [5:77](#), [5:78](#), [5:155](#), [5:156F](#), [5:159F](#), [5:163F](#)
- dispersion [5:160](#)
- FD OCT [5:156–158](#)
- imaging of human retina [5:163–166](#)
- imaging of human skin [5:162–163](#)
- OCTA [5:161–162](#)
- principle [5:155–156](#)
- resolution [5:158–159](#)
- sensitivity [5:159–160](#)
- TD-OCT [5:155–156](#)
- wavelength range [5:160–161](#)
- optical coherent detection [1:374–375](#)
- optical communications [2:269](#)
 - systems [1:291](#), [4:20](#)
- optical compensation film [3:59–60](#)
- optical components [5:309](#)
- optical conductivity [2:131](#)
- optical confinement technique [5:303](#)
- optical connectors [1:336](#), [1:338](#)
 - with alignment by means ferrule [1:338](#)
 - with direct alignment of bare fiber [1:338](#)
 - multifiber connectors [1:338–339](#), [1:339F](#)
- optical coordinate [5:188](#)
 - transformation [4:182](#)
- optical cycle [2:15–16](#)
- optical damage [4:336–337](#)
- optical decision element [1:367–369](#)
- optical detection devices [3:146](#)
- optical devices [4:290](#)
- optical diffraction tomography [4:174](#)
- optical disk drives [4:193](#)
- optical distribution network (ODN) [2:73](#)
- optical Doppler tomography, velocity measurement with [4:526](#)
- OCT capillary velocimetry [4:526](#)
- speckle contrast tomography [4:526](#)
- time dimension in velocity measurement [4:526](#)
- triple-beam spectral-domain OCT [4:526–528](#)
- Optical Duobinary (ODB) [1:275](#), [1:279](#)
 - coding [1:280F](#)
 - format [2:80](#)
- optical elements [4:178](#)
- optical emission [2:1](#), [8](#)
- optical entanglement application [1:484](#)
 - lithography [1:484](#)
 - quantum computing [1:484](#)
 - two-photon imaging and microscopy [1:484](#)
- optical fiber [1:10–11](#), [1:178–179](#), [1:221](#), [1:308](#), [1:336](#), [2:269](#), [4:1–3](#), [4:13](#)
 - development [4:14–16](#)
 - diameter measurement [1:32](#)
 - fabrication [1:27](#), [1:28F](#)
 - fiber coating [1:32–33](#)
 - fiber drawing [1:30](#)
 - furnaces [1:30–32](#)
 - FWM [1:348–349](#), [1:349F](#)
 - linear effects [1:255–256](#)
 - nonlinear fiber effects [1:257–259](#)
 - optical Kerr effect [1:348](#)
 - parametric phenomena [1:347–348](#)
 - preform fabrication [1:28](#)
 - structure [4:14–15](#), [4:14F](#)
 - supercontinuum generation [2:426–427](#), [2:426F](#)
 - telecommunications [4:67](#)
 - transmission effects [1:255–256](#)
 - types of fiber [1:33](#)
- optical fiber amplifiers (OFAs) [1:242](#)
- optical fiber cables [1:328](#)
 - blown fiber [1:334–335](#)
 - cable core [1:330](#)
 - components [1:330](#)
 - filling compounds and other components [1:332](#)
 - secondary fiber coating [1:328](#)
 - sheaths [1:331–332](#)
 - special cables [1:334](#)
 - strength member [1:330–331](#)
 - structures [1:332–333](#)
 - submarine cables [1:334](#)
 - types and applications [1:333–334](#)
- optical fiber communication systems [1:23](#), [4:67–68](#)
 - chromatic dispersion [4:21–22](#)
- optical fiber gratings
 - applications [1:236–237](#)
 - chromatic dispersion control [1:238–239](#)
 - fabrication [1:234–236](#)
 - fiber modes [1:229](#), [1:230F](#)
 - fiber sensors [1:237](#)
 - grating enhanced continuum [1:240F](#)
 - nonlinear applications of fiber gratings [1:239–240](#)
- optical filtering [1:238](#)
- stabilized diode lasers and fiber lasers [1:237–238](#)
- theory [1:229–234](#)
- optical fill factor [4:419–420](#)
- optical films [3:28](#), [3:28F](#)
- optical filtering [1:238](#), [4:35–136](#)
- optical frequency [2:329–331](#), [4:1–2](#), [4:17](#)
 - oscillations [4:341–342](#)
- optical frequency combs *see* highly stabilized mode-locked lasers
- optical frequency domain imaging (OFDI) [4:504–505](#)
- optical gain [2:417](#)
- optical heterodyne detection (OHD) [2:248](#)
- optical heterodyning [1:374–375](#)
- optical holography [4:102](#)
- optical hybrid (OH) [1:297](#)
- optical imaging
 - systems [3:142](#), [3:272](#), [5:14](#)
 - techniques [3:156](#)
- optical implementation [4:189–190](#)
- optical information [1:374](#)
- optical insulator [1:169](#)
- optical intensity [2:244](#)
- optical interconnects [4:254](#)
- optical interferometers [1:368](#)
- optical interferometric micro-lidar [4:502–504](#)
- optical isolator (OI) [4:500](#)
- optical Kerr effect (OKE) *see* self-phase modulation
- optical limiting [4:332–333](#)
- optical line terminals (OLTs) [2:73](#)
- optical logic gates [1:251](#)
- optical low-coherence tomography [4:504](#)
 - in cerebrovascular and cardiovascular research [4:506–507](#)
 - SD-OCT [4:504–505](#)
 - SS-OCT [4:505](#)
 - TD-OCT [4:504](#)
- optical low-coherent tomography (OCT) [4:496](#)
- optical markers [2:82](#), [2:85](#)
- optical maser [1:99](#), [1:99–100](#), [2:451](#)
- optical masks [5:16](#)
 - intensity distribution
 - along axis around image/focal plane [5:22–27](#)
 - on neighboring transverse planes [5:27](#)
 - on transverse image/focal plane [5:19–22](#)
 - Walsh filters as [5:19](#)
 - and self-similarity [5:27–28](#)
 - Walsh functions [5:16–17](#)
- optical materials [5:91](#)
- optical measurements [5:64](#)
 - locating eyes far point with [5:111](#)
- optical methods [5:190](#)
- optical microcavity [1:456](#)
- optical microscopy [5:175](#), [5:185](#)
- optical mode density [2:107](#)
- optical modulation [1:291–293](#)
 - high-order modulation [1:293–295](#)
 - modulators [1:292–293](#)
- optical modulator [2:262](#)
- optical multiplex section (OMS) [1:287](#)
- optical multiplexer [1:357](#)
- optical multiplexing [4:35](#)
- optical network [1:342](#)

- optical network unit (ONU) [1:252](#), [2:73](#)
- optical nonlinearities [1:239](#), [4:25–26](#)
 - as factors in dispersion compensation [4:22–23](#)
 - FWM [4:23–24](#)
 - physical origin [1:346–347](#)
 - SPM [4:23](#)
 - XPM [4:23](#)
- optical packet switching [4:227–230](#), [4:229F](#)
 - core router growth [4:230](#)
 - HiPC switched intraconnect [4:230–231](#)
 - switch core growth [4:230](#)
- optical parameters
 - absorption cross-section [3:221](#)
 - photon migration in scattering media [3:221–222](#)
 - scattering cross-section [3:220–221](#)
- optical parametric amplification (OPA) [2:166](#), [2:290](#), [2:290F](#), [2:295F](#), [2:296F](#), [2:311](#), [2:314](#), [2:314](#)
 - architecture [2:294–295](#)
 - general OPA scheme [2:294–295](#)
 - optical parametric CPA [2:299–300](#)
 - parametric amplification [2:295–297](#)
 - phase matching [2:293](#)
 - pulse compression [2:297–298](#)
 - seed generation [2:295](#)
 - theory [2:290–292](#)
 - ultra-broadband OPAs [2:298–299](#)
- optical parametric chirped pulse amplification (OPCPA) [2:166](#), [2:290](#), [2:299–300](#), [2:300F](#), [2:311](#), [2:314](#), [2:315F](#)
- optical parametric generation (OPG) [2:295](#)
- optical path compensator (OPC) [4:500](#)
- optical path difference (OPD) [3:213](#)
- optical path lengths determination [5:112–113](#)
- optical perturbation [4:321](#)
- optical phase [5:39](#)
 - shifters [4:236–239](#), [4:237F](#), [4:239F](#)
 - space [3:164](#)
- optical phased locked loops (PLL) [4:145](#)
- optical phonons [2:84](#)
- optical PLL (OPLL) [1:296–297](#)
- optical polarization [4:264](#)
- optical pre-amplifiers (OPA) [1:248–249](#), [1:273](#)
- optical processing functions [1:24](#)
- optical propagation loss [4:217](#)
- optical pulse
 - chronocyclic tomography for characterization [4:179](#)
 - mathematical representation of [4:48–49](#)
- optical rectification (OR) [1:405–409](#), [1:406](#), [1:406F](#), [2:120](#), [2:121](#)
- optical regeneration by saturable absorbers [1:369–370](#)
- optical resonator [2:329–331](#), [2:330F](#), [2:371–372](#)
 - stability [4:161–162](#)
- optical response features in correlated electron materials [2:131](#)
- optical routing [4:243](#)
 - switches [1:147](#)
- optical scanning holography [1:377–378](#), [1:378F](#)
- optical sectioning, structured illumination to [5:199F](#), [5:200F](#), [5:201–202](#)
- optical sensing technology [3:86](#)
- optical setups for PST [4:178–179](#)
- optical signal
 - dispersion effect on [1:1–5](#)
 - processing [1:130](#)
 - regeneration [1:364](#)
- optical signal to noise ratio (OSNR) [1:274](#), [1:274F](#), [1:355](#), [1:360](#), [1:361](#), [1:366–367](#), [4:5](#)
- optical solitons [1:180](#)
 - formation [2:425–426](#)
- optical sources [4:38](#)
- optical spacer [5:302–303](#)
- optical spectroscopic methods and techniques [3:95](#)
- optical spectroscopy [2:130–131](#)
- optical spectrum [1:141](#)
- optical spectrum analyzer (OSA) [2:460](#)
- optical Stark effect [4:271](#)
- Optical Supervisory Channel (OSC) [1:287](#)
- optical susceptibility tensors [2:208](#)
- optical switch(es) [1:214–215](#), [4:224](#), [4:225F](#)
 - cell [4:234](#)
 - evolution from static links to optical switching [4:224](#)
 - existing and proposed deployments of optical switches [4:224–225](#)
 - families of optical switch technologies [4:226](#), [4:226F](#)
 - free-space optical switch technologies [4:231](#)
 - optical packet switching [4:227–230](#)
 - packet switch capacity growth [4:225](#)
 - patch panels [4:225](#)
 - planar lightwave circuit optical switch technologies [4:233](#)
 - protection and provisioning [4:225](#)
 - reconfiguration within data center [4:225](#)
 - ROADM, using wavelength selective switch [4:226–227](#)
 - switch characteristics [4:225](#)
 - terminology of optical switching [4:224](#)
 - tunable laser optical switch technologies [4:241](#)
- optical switching in PCF [1:181](#)
- optical systems [2:329](#), [4:174](#), [4:175](#)
 - analysis and design [3:164](#)
 - for DLP [3:36](#)
 - for single-panel projection applications [3:36](#), [3:36F](#)
- optical technique [3:95](#), [4:34](#), [4:103](#), [5:39](#)
- optical time division multiplexing (OTDM) [4:34](#), [4:35](#), [4:60](#), [4:61F](#)
 - BER performance [4:45–47](#), [4:47F](#)
 - bit interleaved OTDM [4:35–36](#), [4:36F](#)
 - clock and data synchronization in [4:44–45](#)
 - communication system [4:61–63](#)
 - components [4:38](#)
 - networking issues [4:65–66](#)
 - packet interleaved OTDM [4:36–38](#), [4:37F](#), [4:38F](#)
 - principle [4:35](#), [4:35F](#)
 - TDM [4:60–61](#)
- optical time domain measurement method [1:313](#)
- optical time domain reflectometer (OTDR) [1:309](#), [4:438](#)
- optical transfer function (OTF) [3:173](#), [4:169](#), [5:112](#)
 - with defocusing [4:169](#)
- optical transmission systems, noise in [1:272–275](#)
- Optical Transport Hierarchy (OTH) [1:287](#)
- optical transport network (OTN) [1:287](#)
 - layering and monitoring [1:288F](#)
 - layers [1:287](#), [1:288F](#)
 - mapping and multiplexing [1:287–288](#), [1:288F](#)
 - operations and monitoring aspects [1:288–289](#)
 - tandem connection monitoring example [1:289F](#)
- Optical Transport Section (OTS) [1:287](#)
- optical vortices [4:36–137](#)
- optical waveguides [1:10–11](#)
- optical wide-gap material [4:302](#), [4:303F](#)
- optical-optical-optical (OOO) [4:224](#)
- optical-pump, optical-probe (OPOP) [2:133](#)
- optical-pump THz-probe (OPTP) [2:136](#), [2:136–137](#), [2:137](#)
- optical-pump/THz-probe magnetospectroscopy of semiconductor quantum wells [2:76–77](#)
- optical-to-electrical-to-optical (OEO) [4:224](#), [4:224F](#)
- optically compensated bend mode (OCB mode) [3:27](#), [3:61](#)
- optically detected FIR resonance spectroscopy of semiconductor quantum wells [2:75–76](#)
- optically detected resonance (ODR) [2:75](#), [2:76F](#)
- optically pumped molecular gas lasers [1:379–380](#)
- optically pumped semiconductor lasers (OPLs) [2:280](#), [2:280F](#), [2:413–414](#), [2:448–450](#)
 - applications [2:287–288](#), [2:288F](#)
 - background [2:281](#)
 - cavity configurations [2:285](#)
 - design strategies [2:281–283](#)
 - gain structure [2:283](#)
 - high power, wavelength tunable, high brightness CW OPLs [2:285–286](#)
 - high power narrow linewidth CW OPLs sources [2:286](#)
 - intra-cavity SHG and DFG OPLs [2:287](#)
 - loss mechanisms [2:283–284](#)
 - low noise multi-gigahertz repetition rate femtosecond pulsed OPLs [2:286–287](#)
 - pumping schemes [2:284–285](#), [2:284F](#)
 - quantum dot CW OPL sources [2:286](#)
- optically-induced transitions [2:152](#)
- optics [3:137–138](#), [4:193](#)
- optimal control [1:487](#)
- optimal control experiments (OCEs) [5:31](#)
 - algorithms for implementing [5:34–38](#)

- closed-loop formulations for attaining laser control [5:37F](#)
dissociative rearrangement [5:36F](#)
optimal control theory (OCT) [5:31](#)
optimal spectral [1:487](#)
optimum energy distribution [2:334](#)
optimum mean value [2:334](#)
opto-acoustics *see* photo-acoustic
optoelectronic (OE) [5:6–7](#)
 applications [2:88–89](#)
 components [3:87–88](#)
 detector design [5:6–7](#)
 devices [1:141–142](#), [2:98–99](#), [2:105](#)
 LED [3:88](#)
 OLED [3:89](#)
 OPD [3:89](#)
 optoelectronics/electronics [3:94](#)
 PLLs [4:45](#)
 properties [4:217–220](#), [4:218F](#), [5:282](#)
 regeneration [1:364–365](#)
 sensing system [3:86–88](#), [3:87F](#)
 silicone photodiode [3:88](#)
 SiPM [3:88–89](#)
 VCSEL [3:88](#)
optoelectronic sensor [3:86–87](#)
 data processing algorithm for [3:89–91](#)
 quasi-spectral analysis [3:91](#)
 ratiometric analysis [3:90–91](#)
“order” of perturbation theory [2:153](#)
ordinary polarization directions [2:293](#), [2:293F](#)
ordinary ray [1:150](#)
organic crystalline materials [1:142](#)
organic crystals [1:409–411](#)
organic dielectrics [3:15](#)
organic dye lasers [5:309–310](#)
organic dyes [5:277](#)
organic electro-optic
 chromophore [1:144–146](#)
 materials [1:144](#)
organic electronic devices [5:220](#), [5:227](#)
organic field effect transistors (OFETs)
 [5:227–229](#)
organic hosts [2:346–347](#)
organic lasers
 conjugated polymers [5:309](#)
 gain measurement [5:311–313](#), [5:312F](#)
 organic semiconductor gain materials
 [5:310–311](#)
 organic semiconductor lasers [5:309–310](#)
 plastic diode lasers [5:314–316](#)
 polymer laser resonators [5:313–314](#)
organic layer–bulk heterojunction [5:322](#)
organic layer–single material [5:321–322](#),
 [5:321F](#)
organic light emitting diode (OLED) [1:445](#),
 [1:448F](#), [3:1](#), [3:5](#), [3:5F](#), [3:56](#), [3:64](#),
 [3:89](#), [5:220](#), [5:229F](#), [5:232](#), [5:232F](#),
 [5:240](#), [5:240F](#), [5:247](#), [5:249F](#), [5:250F](#),
 [5:252F](#)
 applications [3:67–69](#), [5:238–239](#)
 architectures [5:232](#)
 bottom-emission and *p-i-n* [3:65F](#)
 carrier injection and materials [5:232–233](#)
 carrier transport and materials [5:233–234](#)
 cavity effects in [1:447–448](#)
 degradation of OLEDs [5:238](#)
 devices [3:65–67](#)
 displays and lighting [3:68F](#)
 doped transport layer [5:234–235](#)
 electromagnetic modeling of light
 generation [3:68F](#)
 exciton formation and decay [5:235–237](#)
 host–guest system [5:237–238](#)
 materials [3:64–65](#), [3:66F](#)
 perspectives [3:69](#)
 small molecule and polymer OLEDs [5:232](#)
organic light-emitting diodes/light extraction
 light extraction of plasmonic mode
 [5:251–253](#)
 light extraction of substrate mode [5:248](#)
 light extraction of waveguiding mode
 [5:248–251](#)
organic light-emitting transistors (OLETs)
 [5:227](#), [5:230F](#)
organic macromolecular materials [1:143](#)
organic materials [4:331](#)
organic optoelectronics [3:89](#)
organic PD (OPD) [3:89](#)
organic photodetectors [5:317](#)
 organic photodiodes [5:317–319](#), [5:317F](#),
 [5:318F](#)
 organic phototransistors [5:319–320](#)
 with polymers [5:326–327](#)
 photosensitivity of [5:320–321](#)
 with polymers [5:322–324](#)
 with small molecules [5:321–322](#),
 [5:325–326](#)
organic photodiodes [5:317](#), [5:317–319](#),
 [5:317F](#), [5:318F](#)
 organic layer–bulk heterojunction [5:322](#)
 organic layer–single material [5:321–322](#),
 [5:321F](#)
 with polymers [5:322–324](#)
 with small molecules [5:321–322](#)
organic phototransistors [5:317](#), [5:318F](#),
 [5:319–320](#), [5:319F](#)
 organic layer–bulk heterojunction [5:326](#)
 organic layer–single material [5:325–326](#)
 with polymers [5:326–327](#)
 with small molecules [5:325–326](#)
organic photovoltaics (OPV) [5:256](#), [5:287](#)
 acceptor materials [5:265](#)
 devices [5:224](#), [5:259–260](#), [5:298](#)
 materials [2:194](#)
 strategy to improving BHJ device
 performance [5:260](#)
organic polymers [3:89](#)
 solar cell structure evolution [5:259](#)
organic semiconductors (OSCs) [3:15](#), [5:220](#),
 [5:226F](#), [5:256](#)
 concept of doping and p- and n-type OSCs
 [5:226–227](#)
 device applications [5:227–231](#)
 electronic structures [5:220–223](#)
 excitons [5:223–226](#)
 gain materials [5:310–311](#)
 lasers [5:309–310](#)
 materials [5:317](#)
 p-type [5:226–227](#)
 solitons, polarons, and bipolarons [5:223](#)
organic sensing channel layers [5:320](#)
organic sensing layer (OSIs) [5:317–319](#),
 [5:318F](#)
organic solar cells (OSCs) [5:256](#), [5:258F](#),
 [5:292](#), [5:293](#)
 AR effects [5:293–294](#)
 exciton diffusion bottleneck [5:292F](#)
 light harvesting [5:292](#)
 light trapping [5:294–296](#)
 metal nanostructures [5:296–302](#)
 other approaches for [5:304–306](#)
 spacer layer and cavity effects [5:302–304](#),
 [5:304F](#)
 ways of harnessing light [5:293F](#)
 characterization [5:258–259](#)
 principles [5:256–258](#)
organic solvents [5:277](#)
organic thin films [1:445](#), [3:15](#), [5:232](#)
organic thin-film transistors (OTFTs) [3:15](#),
 [3:53](#), [5:227](#)
organically modified ceramics (ORMOCERS)
 [2:346](#)
organically modified silanes (ORMOSILS)
 [2:346](#)
organic–inorganic hybrid perovskite [5:282](#)
 charge carriers transport materials
 [5:286–287](#), [5:286F](#)
 crystal structure [5:282–283](#), [5:282F](#)
 device structures of perovskite solar cell
 [5:284–286](#), [5:285F](#)
 ETM [5:287–288](#), [5:289F](#)
 HTM [5:287](#), [5:287F](#)
 optoelectronic properties [5:283–284](#),
 [5:284F](#), [5:285F](#)
 tandem solar cell [5:288–290](#)
 working mechanism of perovskite solar cell
 [5:286](#)
organometallic chemical catalysts [2:172](#)
 dynamics of homogeneous catalysts [2:172](#)
 surface attached catalysts [2:172–173](#)
orientation process [4:95–96](#)
ORMOSIL host matrices [2:347](#)
ortho-symplectic matrices [4:200](#)
orthodony, mode-locking as [2:303–304](#)
orthogonal frequency division multiplexing
 (OFDM) [1:255](#), [2:74](#), [2:83](#), [2:84F](#)
orthogonal frequency division multiplexing-
 passive optical networks (OFDM-
 PONs) [2:83–84](#)
orthogonal polarization [4:45](#)
orthogonal superposition [4:175–176](#)
orthoscopic real image production [4:152](#)
out-of-resonance values [4:287](#)
outer diameter (OD) [1:211](#)
outer scale [3:201](#)
output coupler (OC) [1:237–238](#), [2:329](#)
outside vapor deposition process (OVD
 process) [1:28](#), [1:29](#)
outside vapor phase oxidation (OVPO) *see*
 outside vapor deposition process
 (OVD) process)
overclad tube [1:196](#)
overcoupling/undercoupling [2:448](#)
overlap-projection back propagation [3:154](#)
overlapping [4:103](#)
oversampling smoothness (OSS) [3:148](#)

oxadiazole derivatives [2:346](#)
 oxide disorder [1:469–471](#)
 oxide heterostructures [2:129](#)
 oximetry images from patient eyes [4:522](#),
[4:522F](#)
 oxygen-dependent photobleaching [3:102](#)
 Oxymap Analyzer [4:522](#)
 Oxymap T1 instrument [4:521–522](#), [4:521F](#)
 oxymetry lidars [4:519–522](#), [4:519F](#)

P

- p-doped semiconductor [3:87–88](#)
 p-FFS [3:60–61](#)
 p-terphenyl [2:345](#)
 p-transmittance [1:153–154](#)
 p-type transition metal oxide–NiO_x
[5:265–266](#)
 packet interleaved OTDM [4:36–38](#), [4:37F](#),
[4:38F](#)
 packet interleaving *see* packet level
 packet level [4:35](#)
 packet switch capacity growth [1:225](#)
 pain, PBM for [3:118](#)
 pair braking
 detectors [1:427–428](#)
 photon detectors [1:427–428](#)
 panchromatic [4:88](#)
 pandrug-resistant (PDR) [3:102–103](#)
 panretinal photocoagulation (PRP)
[5:131–132](#), [5:132](#), [5:134](#)
 para-magnetic (PM) [2:127](#)
 parabolic gold coated mirror [2:329–331](#)
 parabolic quantum [2:19](#), [2:20–21](#)
 paradoxical hypertrichosis [3:119](#)
 parallax cues [3:208–209](#)
 parallel computing [4:178](#)
 parallel fibers, array of [1:214–215](#), [1:214F](#)
 parallel resistance (R_p) [5:259](#)
 parallel-serial conversion (P/S conversion)
[1:277](#)
 parallel-tangents condition [1:139](#)
 parametric (mixing) processes [1:181–183](#)
 parametric amplification of emission [1:460](#)
 parametric effects [1:345](#)
 parametric instabilities [1:442](#)
 parametric optical amplifiers [1:263](#)
 parametric superfluorescence (PSF) *see*
 optical parametric generation (OPG)
 paraxial approximation [4:175](#)
 paraxial focal plane [1:114–116](#)
 paraxial matrices [1:102–104](#)
 paraxial optical models [4:199](#)
 paraxial optical system
 propagation in [4:168](#)
 propagation through [4:167](#)
 paraxial ray vector [4:205–206](#)
 paraxial schematic eyes [5:48](#)
 paraxial system for design or analysis
[1:110–111](#)
 paraxial transverse profile [4:180](#)
 paraxial wave state [4:180](#)
 “paraxial” schematic eyes [5:44](#)
 parity-broken nematic order in Cd₂Re₂O₇
[2:216–217](#), [2:217F](#)
 parity-time symmetry (PT symmetry) [4:291](#),
[4:292F](#)
 anomalous scattering properties of PT-
 symmetric arrangements [4:300–301](#)
 mode management in PT-symmetric lasers
 [4:293–294](#)
 sensing at exceptional points [4:298–300](#)
 spontaneous symmetry breaking and EP
 [4:291–293](#)
 three photonic PT molecules [4:299F](#)
 Parkinson’s disease (PD) [3:124](#), [3:125](#)
 Parseval’s theorem [4:376](#)
 partial coherence algorithms [3:148–149](#)
 partial coherence KF-PR algorithm [3:151](#)
 partial coherence properties in image of
 incoherent source [4:169](#)
 partial differential equations [3:258](#)
 partial mirror (PM) [4:507](#)
 partial response pulse (PR pulse) [1:275](#)
 partial response signaling [1:278–280](#)
 partial Talbot effect [4:166–167](#)
 partially coherent illumination, PSF for
[3:263](#)
 partially coherent impulse response [3:263](#)
 partially coherent sources and fields [1:34](#)
 partially redundant pupil masking (PRPM)
[3:279](#)
 participation number [4:282](#)
 particle image velocimetry (PIV) [4:470](#),
[5:68–69](#)
 particle motion [1:139](#)
 particle scattering [4:317](#)
 passive component [1:223–224](#)
 passive dynamic holographic display devices
 (PDHDDs) [4:114](#)
 passive matrix (PM) [5:239](#)
 passive matrix driving method (PMLCD)
[3:30–31](#), [3:30F](#)
 passive matrix OLED (PMOLED) [5:239](#)
 passive mode-locking [2:337](#)
 passive optical components [1:336](#),
[1:342–343](#)
 branching devices [1:339–341](#)
 general considerations [1:336–338](#)
 mechanical splices [1:339](#)
 nonwavelength-selective branching devices
 [1:340–341](#)
 optical connectors [1:338](#)
 optical joint devices [1:336–338](#)
 WDM [1:341–342](#)
 passive optical networks (PON) [1:280–281](#),
[4:224](#)
 BPOs [2:73](#), [2:74](#), [2:75F](#)
 ER-PONs [2:82–83](#), [2:82F](#)
 GE-PON [2:73](#), [2:76](#)
 GPON [2:73](#), [2:74–76](#), [2:75F](#)
 NG-PON1 [2:73](#), [2:74](#), [2:76–78](#)
 NG-PON2 [2:78–79](#)
 OFDM-PONs [2:83–84](#)
 TDM-PONs [2:74](#)
 topology [2:73](#)
 WDM PONs [2:74](#), [2:81–82](#)
 passive Q switching [2:419–421](#), [2:420F](#)
 with bulk saturable absorbers [2:419–421](#)
 with semiconductor saturable-absorber
 mirrors [2:421](#)
 passive SA-based optical regeneration [1:370](#)
 passive static optical filters [1:266](#)
 passive waveguide and guided wave optics
[1:221](#)
 passive WDM filters and couplers [1:265–268](#)
 passive-matrix OLED (PM-OLED) [3:66](#)
 patch panels [4:225](#)
 path-independent loss (PILoss) [4:235](#),
[4:235F](#)
 pathogen classification [3:103–105](#)
 pattern
 effects [1:245](#)
 pattern-projecting matricial sensors
 [4:511–512](#)
 recognition and image filtering [4:332](#)
 Pattern Dependent Phase Distortion (PDPD)
[1:245](#), [1:247–248](#)
 pattern-recognition receptors (PRRs) [3:112](#)
 patterned vertical alignment LCD (PVA-LCD)
[3:62](#)
 Pauli blocking [1:460](#)
 Pauli exclusion principle [2:63](#), [2:84–85](#),
[2:99](#), [2:106](#), [2:464](#), [3:64](#)
 Pauli-blockade mechanism [1:473](#)
 pay as you grow approach [1:361](#)
 PDM-QPSK [1:295](#)
 transmission [1:295](#)
 2PE *see* two-photon excitation (TPE)
 peak to average power ratio (PAPR) [2:83](#)
 PEDOT doped with poly(4-styrene
 sulfonate) (PEDOT:PSS) [5:279–280](#)
 pentacene [5:321–322](#), [5:321F](#)
 pentacene photodiodes [5:321–322](#)
 pentaprism [1:125–126](#), [1:125F](#), [1:126F](#)
 perfectly coherent fields [1:41](#)
 periodic dispersion compensation [4:24](#)
 periodic multilayer structure [1:214](#)
 periodic PT-symmetric structures [4:301](#)
 periodic structure [1:207](#)
 periodically poled LiNbO₃ (PPLN) [2:121](#),
[4:39](#)
 periodically poled lithium niobate (PPLN)
[1:14](#), [2:287](#)
 permanent materials, holographic recording
 media and devices [4:89–90](#)
 DCG [4:90](#)
 holographic sensors [4:92–93](#)
 photopolymers [4:90–91](#)
 photoresists and embossed holograms [4:91](#)
 PIRG [4:91–92](#)
 silver halide [4:89–90](#)
 perovskites [2:71](#), [2:72F](#)
 crystals [4:235](#)
 materials [5:282](#)
 solar cell [5:283](#)
 device structures [5:284–286](#), [5:285F](#)
 working mechanism [5:286](#)
 persistent spectral hole burning [4:93–95](#)
 perturbation
 second order [2:88](#)
 theory [1:170–171](#), [1:172–173](#)
 perturbed free induction decay (PFID) [2:201](#)
 perylene bisimide (PBI) [5:265](#)
 perylene derivatives (PTC) [5:260](#)
 perylene diimide (PDI) [5:265](#)
 perylene dyes [2:345](#), [2:345F](#)

- Perylene Orange [2:347](#)
 Perylene Red [2:317](#)
 perylene-3,4,9,10-tetracarboxylic dianhydride (PTCDA) [5:325–326](#), [5:325F](#)
 perylenes [2:345](#)
 perylimide dyes [2:345](#), [2:345F](#)
 Petzval curvature [1:120–121](#), [1:121F](#)
 pharmacological agents [5:120](#)
 phase and frequency detector (PFD) [4:436–438](#)
 phase compensators [1:413–414](#)
 phase conjugation (PC) [3:205](#), [3:208](#), [4:313–314](#), [4:314F](#), [4:322](#), [4:325](#)
 formation of true and conjugate images from hologram [3:208F](#)
 holographic mechanism of [4:323](#)
 by holography [3:206–208](#)
 one way phase conjugation [3:209–210](#), [3:209F](#)
 phase conjugate reconstruction [4:331](#)
 phase conjugated wave [3:205](#)
 pseudoscopic image [3:208–209](#)
 phase conjugator [3:205–206](#), [3:205F](#)
 phase contrast [5:185](#), [5:187](#)
 filter [4:35](#), [4:36F](#)
 microscopy [5:192](#), [5:192](#)
 output [4:35](#), [4:36F](#)
 phase control [5:39](#)
 phase cycling [2:190](#)
 phase difference [4: 98](#)
 phase diversity [3:278](#)
 receiver [1:297–298](#), [1:298](#), [1:298F](#)
 phase factor [1:3–4](#)
 phase functions [4:183](#), [4:185–186](#), [4:185F](#), [4:189F](#)
 derivatives [4:186](#)
 general [4:186](#)
 phase gradient autofocus (PGA) [4:440](#)
 phase lock loop (PLL) [1:296–297](#), [4:6](#)
 phase mask (PM) [1:54](#), [1:236](#), [1:236F](#), [2:211](#)
 phase matching [1:11](#), [1:230](#), [1:230–231](#), [1:231–232](#), [1:231F](#), [2:166](#), [2:293](#)
 bandwidth of OPA [2:294](#)
 condition [1:13](#)
 gating [2:238](#), [2:239F](#)
 interaction types in OPA [2:293](#)
 methods [2:293–294](#)
 phase modulation (PM) [1:367–368](#), [2:304–305](#), [2:305F](#), [4:31–33](#)
 phase modulator (PM) [1:292](#), [2:311](#)
 phase noise [2:458–459](#), [2:459F](#)
 phase plates [5:119](#)
 phase problem [3:146](#), [3:162](#)
 phase retrieval [3:161–163](#), [3:170–171](#)
 algorithm [3:148](#), [3:150](#), [3:152](#), [4:55](#), [4:56F](#)
 application to [4:172](#)
 approach for 2D images [4:121–122](#)
 problem [3:162](#)
 Phase Retrieval by Omega Oscillation Filtering (PROOF) [2:322–323](#)
 phase retrieval transform function (PRTF) [3:148](#)
 Phase Sensitive Amplifier (PSA) [1:252](#)
 phase shift [3:242–243](#)
 Phase Shift Keying (PSK) [1:245](#)
 phase shifters [4:181](#)
 phase shifting recording geometry (PSRG) [4:369](#)
 phase singularities [4:185–186](#)
 derivatives of phase function [4:186](#)
 general phase function [4:186](#)
 phase functions [4:185–186](#)
 transmission function with [4:186–187](#)
 phase space
 analysis of classical geometric optical imaging [3:167](#)
 diagram [4:205](#)
 in geometrical optics [4:205–206](#), [4:207F](#)
 method [4:212](#)
 optics elements [3:164–166](#)
 rotations [4:176](#)
 rotator [4:176](#), [4:177](#), [4:178](#)
 phase space and imaging
 elements of phase space optics [3:164–166](#)
 holographic imaging [3:168–169](#), [3:169F](#)
 imaging aberrations [3:167](#)
 lightfield imaging [3:169–170](#), [3:170F](#)
 phase-space interpretation of image formation [3:166–167](#)
 phase-space tomography and phase retrieval [3:170–171](#)
 pupil design for extended depth of field [3:173](#)
 self-imaging condition [3:167–168](#), [3:168F](#)
 superresolution imaging [3:171–173](#), [3:172F](#)
 phase space tomography (PST) [4:174](#)
 chronocyclic tomography for optical pulse characterization [4:179](#)
 fundamentals [4:175–176](#)
 light beam characterization [4:174–175](#)
 for 1D beam characterization [4:176–177](#)
 optical setups for PST [4:178–179](#)
 optics [4:174](#)
 PST method [4:175](#), [4:176](#)
 quantum light characterization [4:179–181](#)
 for quantum light characterization [4:179–181](#)
 tools [4:174](#)
 for 2D Beam characterization [4:177–178](#)
 phase spectrum estimation methods [3:275–276](#)
 phase velocity dispersion [1:5](#)
 phase-added stereogram [4:123](#)
 phase-amplitude coupling [2:467–468](#)
 phase-conjugate mirror (PCM) [4:313–314](#)
 phase-cycling [2:153–154](#)
 phase-distorting media [3:205](#)
 phase-locked CW laser frequencies [5:41](#)
 phase-matching [2:153–154](#)
 angle [2:294](#)
 condition [1:349](#)
 phase-modulated wave, spectrum of [1:132F](#)
 phase-only transmission function [4:184](#)
 phase-reversal zone plate [1:95–96](#)
 phase-sensitive detection [5:116](#)
 phase-sensitive optical coherence
 elastography [5:146–147](#), [5:147F](#)
 applications of optical coherence elastography in ocular tissues [5:147–148](#)
 principles and implementations [5:146–147](#)
 phase-shift
 laser rangefinder [4:498](#), [4:498F](#)
 phase-shifted DFB single-frequency lasers [2:455–456](#), [2:456F](#)
 rangefinder using heterodyne technique [4:510](#), [4:510F](#)
 technique [1:315](#), [1:316F](#)
 phase-shift keying (PSK) [1:275](#), [1:277](#)
 phase-space interpretation [3:172](#)
 of image formation [3:166–167](#)
 of tomographic signal recovery [3:171](#)
 phase-space optics (PSO) [3:164](#), [4:164](#), [4:170F](#)
 phase-space representations of freeform optical systems [4:205](#)
 aberration contributions
 calculation at image [4:211–214](#)
 in freeform prism [4:214–215](#)
 aberration propagation in phase space [4:212F](#)
 difference in ray positions [4:208F](#)
 distortion analysis [4:213F](#)
 extension to 4D phase space [4:209–211](#)
 lens data of freeform prism of reference including additional dummy surfaces [4:210T](#)
 phase space in geometrical optics [4:205–206](#)
 ray-propagation and phase space trajectory in optical system [4:205F](#)
 real ray behavior in phase space [4:207F](#)
 transverse aberrations [4:214F](#)
 visualization of aberrations
 in freeform systems [4:208–209](#)
 in phase space [4:206–208](#)
 phase-space tomography [3:170–171](#), [4:169–170](#)
 phase-stepping technique [4:509](#)
 “phasing” 2DFS signal [2:188–189](#)
 phenyl-C61-butyric acid methyl ester (PCBM ester) [2:194](#), [5:261](#), [5:262F](#)
 [6,6]-phenyl-C₆₁-butyric acid methyl ester (PCBM) [5:288](#), [5:321–322](#), [5:321F](#)
 1-phenylnaphtalene (PN) [5:263](#)
 pheophytins (Pheo) [2:192](#)
 phosphate glasses [2:387–388](#), [2:392](#)
 phosphorescence [3:64](#)
 phosphorescent EL [5:241–244](#), [5:242F](#), [5:243F](#)
 phosphorous pentoxide (P₂O₅) [1:28](#)
 phosphors [3:3](#)
 phosphorus
 donors [1:469](#)
 impurities [1:467](#)
 system [1:469](#)
 photo activated chromophore for infectious keratitis (PACK-CXL) [5:136–138](#)
 photo diode (PD) [4:6](#)
 photo thermoplastic process [4:97](#), [4:97F](#)
 photo-acoustic
 imaging [3:226](#), [3:226F](#)
 microscopy [5:193](#)
 photo-Dember effect [1:404–405](#)
 photo-doping process [5:227](#)
 photo-elastic coefficient [1:131](#)
 photo-elastic constant [1:130](#)

- photo-elastic effect [1:130–131](#), [1:144](#), [1:160](#), [1:326](#)
- photo-emission microscopy [5:193](#)
- photo-induced carriers in solid materials
built-in field in semiconductor [1:404](#)
PC antenna [1:403–404](#)
- photo-orientation process [4:96](#)
- photo-physics of semiconducting polymers [5:311](#)
- photo-thermal refractive (PTR) [2:446](#)
- photo-thermo-refractive Glasses (PTRG) [4:91–92](#)
- photoacoustic spectroscopy [2:6](#)
- photo.mode [5:273](#)
conducting substrates [5:273](#)
semiconductor layers [5:273–275](#)
sensitizers [5:275–277](#)
- photobiomodulation (PBM) [3:114–115](#), [3:115F](#)
for cosmetic and esthetic applications [3:118–119](#)
Cox [3:115](#)
direct cell-free light-mediated effects on molecules [3:117](#)
for enhancement of cognitive performance [3:126–127](#)
enhancement of performance [3:126](#)
flavins [3:117](#)
for healing of wounds and tissue damage [3:117–118](#)
light/heat sensitive ion channels [3:116–117](#)
mechanisms of action [3:115](#)
for muscle and athletic performance [3:126](#)
for neurodegenerative diseases [3:124–125](#)
for pain and inflammation [3:118](#)
for psychiatric disorders [3:125–126](#)
retrograde mitochondrial signaling [3:115–116](#)
for stroke [3:121–123](#)
for TBI [3:121–123](#)
for vision and hearing problems [3:121](#)
- photocurrent generation [4:352](#)
- photocathode [3:242](#)
- photochemical interaction [5:131](#)
- photochemical therapy [5:136](#), [5:136–138](#)
- photochromic glasses [4:93](#)
- photochromic materials [4:93](#)
- photoconductive antenna (PC antenna) [1:403–404](#), [2:120](#), [2:121](#)
- photoconductivity [4:324](#)
- photocurrent gain of APD [4:423](#)
- photodetection [1:291](#), [4:243](#)
mechanism [3:175](#)
mode [1:374](#)
- photodetector (PD) [1:378](#), [1:418](#), [4:500](#), [4:523](#), [5:189](#), [5:194](#)
photodetector-array apertures [3:177](#), [3:177F](#)
- Photodetector Array Camera and Spectrometer (PACS) [1:425F](#), [1:426](#)
- photodiodes (PD) [1:141–142](#), [3:88](#), [4:509–510](#), [5:2](#)
- photodisruption *see* photomechanical interactions
- photodynamic alteration of tissue microcirculation [3:111](#)
- photodynamic therapy (PDT) [3:100–102](#), [3:101F](#)
antimicrobial photodynamic inactivation and PDT [3:102–103](#)
causes damage to biomolecules and intracellular organelles [3:107F](#)
mechanisms [3:101–102](#)
PDT for cancer [3:106–107](#)
- photoelectron [2:241](#)
with attosecond pulses [2:240–241](#)
hologram mean number of [4:373](#)
spectrum [2:322](#)
- photoemission electron microscopy (PEEM) [2:142](#)
- photoexcited bulk semiconductors, FIR magnetoabsorption in [2:73–75](#), [2:73F](#)
- photoexcited system [4:266](#)
- Photofrin [3:103F](#)
- photographic films [1:376](#)
plate [3:206–207](#), [3:250–251](#), [3:252–253](#)
techniques [5:64](#)
- photoinduced absorption (PIA) [2:142–143](#), [2:194](#)
- photoinduced phase transitions (PIPT) [2:139–141](#)
- photoionization [1:381](#)
- photoluminescence (PL) [1:461](#), [2:71](#), [2:73](#), [2:73F](#), [2:79F](#), [2:105–106](#), [2:108](#), [2:282](#), [5:252–253](#), [5:253F](#), [5:254F](#)
excitation [2:394](#)
quantum [5:309](#)
spectrum [2:108](#)
- photoluminescence quantum yield (PLQY) [5:237–238](#)
- photomechanical interaction [5:132–133](#), [5:138](#)
- photomedicine [3:100](#)
- photometric units [3:70](#)
- photomultiplier tube (PMT) [2:211](#), [3:86](#), [3:98](#), [4:513](#), [5:73](#), [5:74](#)
- photon density wave (PDW) [3:224](#), [3:256–258](#), [3:256](#), [3:258F](#)
- photon density wave imaging (PDWI) [3:256](#)
see also infrared imaging
applications [3:259–260](#), [3:260F](#)
forward problem [3:256–258](#)
inverse problem and imaging algorithm [3:258–259](#)
- photon echo [2:54–55](#), [4:272–275](#), [4:275F](#)
experiment [2:162](#)
scheme [2:154](#)
spectroscopy [2:154](#)
spectrum [2:56](#)
- photon transport equation *see* linear radiation transfer equation
- photon tunneling microscope (PTM) [5:190–191](#)
- photon(s) [1:224](#), [1:458](#), [2:63](#), [2:410](#), [3:225–226](#), [4:180](#)
absorption [1:487](#), [5:224](#), [5:256](#), [5:305–306](#)
avalanche [1:97](#), [2:402–403](#), [2:404T](#)
counting [4:446–447](#)
cutter [2:290](#)
detection [1:427](#)
- detectors [1:418](#)
distribution [2:106](#)
energies [2:19](#), [2:446](#)
lifetime [4:220](#)
migration [3:221](#), [3:221–222](#)
photon-atom binding in PC [2:12–15](#)
propagation [3:97](#)
statistics [1:452–153](#), [1:453](#)
switches [2:18](#)
teleportation [1:458–459](#)
- photonic arrays, devices with [4:507–509](#)
focus tracking micro-lidars [4:507–509](#)
pattern-projecting matricial sensors [4:511–512](#)
range imaging with matricial light-emitting diode array [4:509–510](#), [4:509F](#)
time-of-flight principle with matricial detectors [4:510–511](#)
- photonic band diagram [1:212–213](#), [1:212F](#)
- photonic band gaps (PBGs) [2:12](#)
- Photonic Bandgap-guided fibers (PBG-guided fibers) [1:263](#)
- photonic bandgap(s) [1:169](#), [1:207](#)
fibers [4:30](#)
origin [1:170–172](#)
- photonic bandstructures [1:185](#), [2:13](#)
- photonic crystal fiber (PCF) [1:177](#), [1:178](#), [1:195](#), [1:262](#), [1:263](#), [2:427–428](#)
see also microstructure fiber(s)
birefringence [1:179–180](#)
dispersion [1:178–179](#)
experiment examples [1:181](#)
nonlinear phenomena [1:180–181](#)
optical switching [1:181](#)
parametric (mixing) processes [1:181–183](#)
properties [1:178](#)
supercontinuum generation [1:181](#)
transverse mode structure [1:178](#)
- photonic crystal lasers
electromagnetics of photonic crystals [1:185–186](#)
photonic band diagram for photonic crystal slab [1:186F](#)
photonic crystal resonant cavities and lasers [1:186–189](#)
photonic crystal waveguides [1:189–193](#)
photonic crystal resonant cavities [1:187–188](#)
photonic crystals (PCs) [1:169](#), [1:170](#), [1:185](#), [1:190–191](#), [2:12](#)
cavities [1:187](#)
defect slab waveguides [1:193T](#)
electromagnetics [1:185–186](#)
slab [1:186F](#), [1:187](#)
structure [5:295](#)
waveguides [1:189–193](#)
- photonic devices and circuits [4:217](#)
- photonic element [4:221](#)
- Photonic Integrated Circuits (PICs) [1:242](#), [1:295–296](#), [4:216](#), [4:242](#)
- photonic integration [4:242](#), [4:242F](#)
- photonic lattice [1:185](#)
- photonic regime [2:119](#)
- photonic substrate texturing technique [5:296](#)
- photonic switching [4:224](#), [4:225](#)
- photonic techniques [2:118](#)
- photonic wire lasers [1:384](#)

- photonic-crystal devices [1:173](#)
slabs [1:173–175](#)
- photopolymers [4:90–91](#), [4:154](#)
- photoreceiver input [4:429](#)
- photoreceptor(s) [3:17](#), [5:43](#), [5:86](#)
bands [5:164](#)
- Photorefraction
adaptive signal processing [4:333](#)
applications [4:331](#)
basics of photorefractive effect [4:327](#)
coupled wave equations [4:329–330](#)
distortion compensation by phase conjugation [4:331–332](#)
holographic data storage [4:331](#)
inorganic materials [4:330–331](#)
materials [4:330](#)
optical limiting, novelty filter, and laser ultrasonic inspection [4:332–333](#)
organic materials [4:331](#)
pattern recognition and image filtering [4:332](#)
photorefractive solitons [4:333–334](#)
self-pumped phase conjugate mirrors [4:332](#)
standard rate equation model [4:327–329](#)
- photorefractive crystals (PRC) [4:154](#), [4:324](#)
- photorefractive effect [4:96](#), [4:327](#)
- photorefractive gratings [4:331–332](#)
- photorefractive intrastromal corneal collagen cross-linking (PiXL™) [5:136](#)
- photorefractive keratectomy (PRK) [5:100–101](#), [5:126](#), [5:135](#)
- photorefractive materials [4:96–97](#), [4:330](#)
- photorefractive nonlinear optics [4:327](#)
- photorefractive organic compounds [4:96–97](#)
- photorefractive polymer films (PRP films) [4:114](#)
- photorefractive solitons [4:333–334](#)
- photoresistive material [1:484](#)
- photoresists [4:91](#)
- photosensitive materials [1:54](#)
- photosensitivity [1:234–236](#)
of organic photodetectors [5:320–321](#)
- photosynthesis [2:191–192](#)
- photosystem II reaction center (PSII RC) [2:192](#), [2:193F](#)
- photothermal
cancer treatment [2:262–263](#)
interaction [5:131–132](#), [5:138](#)
microscopy [5:193](#)
- photovoltaic [5:256](#)
effect [1:423](#)
technologies [5:287](#)
- phthalocyanine (Pc) [5:260](#)
- physical contact solution [1:337](#)
- physical imaging system [3:136–137](#), [3:136F](#)
- physical interpretation states [1:87](#)
- Physical Layer Operation, Administration, and Maintenance (PLOAM) [2:74](#)
- physical layer operations administrations and management upstream (PLOAMu) [2:75–76](#)
- physical layer overhead (PLOu) [2:75–76](#)
- “physical processing” approach [3:137–138](#)
- physical transmission layer of worldwide telecommunications network [1:226](#)
- physical vapor deposition (PVD) [2:255](#), [2:260](#)
- physiological depth cue [3:44–45](#)
- π -phase shift (P_π) [4:237–238](#)
- pickup head (PUH) [4:507](#), [4:508F](#)
- pico-projector [3:32](#)
- picolinate (pic) [5:244](#)
- picosecond (ps) [2:124](#)
pulse pumped supercontinua [2:429–431](#)
- picture/hologram [4:89](#)
- piezo-optic
devices [1:159–160](#)
modulators [1:156–157](#)
- piezoelectric effect [2:105](#)
- piezoelectric force microscopy (PFM) [2:147](#)
- piezoelectric moving fiber optical switch [4:233](#)
- piezoelectric optical switch [4:233](#)
- piezoelectric transducer (PZT) [4:98](#), [4:502](#)
- pigmented epithelium [5:130](#)
- “pinch-off” phenomenon [5:227–229](#)
- pincushion distortion [1:120–121](#), [1:121F](#)
- pinhole effect [5:122](#), [5:123F](#)
- pixels per inch (PPI) [3:75](#)
- planar heterojunction *see* bilayer heterojunction
- planar laser induced fluorescence (PLIF) [4:470](#)
- planar lightwave circuit (PLC) [2:385](#), [4:39](#), [4:233](#), [4:234F](#)
blocker [4:233–234](#)
interferometer [4:233](#)
interferometric optical switch [4:236](#)
materials and physical principles [4:233](#)
materials for planar lightwave optical switches [4:235–236](#)
moving waveguide coupler [4:233](#)
optical switch technologies [4:233](#)
scaling up planar lightwave optical switch [4:234](#)
SOA switch [4:239–240](#)
waveguide-MEMS optical switch [4:240–241](#)
- planar lightwave optical switches materials [4:235–236](#)
- planar mirror [3:262](#)
- planar technique [1:340](#)
- planar waveguide [1:10–11](#), [1:11F](#), [2:384–385](#)
- planar waveguide lasers (PWL) [2:384](#), [2:385–387](#)
see also infrared imaging
DBR [2:388–389](#)
DFB [2:389–391](#)
lasers [2:385](#)
materials [2:387–388](#)
planar waveguides [2:384–385](#)
pumping [2:392–393](#)
PWL arrays [2:391–392](#)
stability [2:393](#)
systems [2:393](#)
waveguides [2:384](#)
- Planck’s black-body law [3:230](#), [3:231F](#)
- Planck’s constant (h) [5:89](#)
- plane of polarization [1:148](#), [1:162](#)
- plane wave [1:162](#), [1:376](#)
approximation [1:131–132](#)
basis [1:122](#)
- spatial structure [4:180](#)
- plasma
display [3:1](#), [3:4–5](#), [3:4F](#)
double sheath [2:381–382](#)
formation [5:92](#)
frequency [2:132](#)
gas dry etch processes [4:217](#)
nonlinearities and relativistic effects [4:337–338](#)
temperature [2:464–465](#)
- plasma-enhanced chemical vapor deposition (PECVD) [3:13](#)
- Plasmon mode [3:67](#)
- Plasmon oscillation [2:118](#)
- plasmonic(s) [5:296–302](#), [5:297F](#), [5:298F](#)
materials [2:252](#)
NP’s [5:298](#)
plasmonic mode, light extraction of [5:251–253](#)
resonators, molecules in [1:446–447](#)
structures [5:295](#)
- plastic(s) [5:309](#)
diode lasers [5:314–316](#)
electronics [5:220](#)
fillers [1:332](#)
films [5:317](#)
- plateau region [2:317](#)
- platinum (Pt) [5:279](#)
- plug and socket [1:338](#)
- plug stile attenuator [1:342](#)
- plug–adapter–plug configuration [1:338](#), [1:339F](#)
- p–n junction structures [5:317–319](#)
- pnictides [2:126](#)
- Pockel’s cell [4:351](#)
- Pockels effect [1:141](#)
- pocket-projector [3:32](#)
- Poincaré sphere analysis (PSA) [1:160](#), [1:317](#)
- point cloud
model [4:477](#), [4:477F](#)
resolution measurement [4:403](#)
- point object, hologram of [4:152](#)
- point response [3:253](#)
- point spread function (PSF) [3:137](#), [3:187](#), [3:263](#), [3:277](#), [4:533](#), [5:15F](#), [5:16](#), [5:59–60](#), [5:72](#), [5:90](#), [5:112](#), [5:176](#)
for coherent, incoherent, and partially coherent illumination [3:263](#)
diffraction calculation of 3D PSF [3:264–265](#)
spatial frequency content of 3D PSF [3:263–264](#)
- point-based methods [4:123–124](#)
- point-by-point methods [1:236](#)
- point-object hologram [1:377](#), [1:377F](#)
reconstruction of [1:377F](#)
- point-spread function (PSF) [3:149–150](#), [4:168](#), [4:370–371](#)
- point-target measurements [5:6](#)
- point-to-multipoint physical topologies (PiMP physical topologies) [2:73](#), [2:73F](#)
- point-to-point physical topologies (PiP physical topologies) [2:73](#), [2:73–74](#), [2:73F](#)
- points-of-presence (PoPs) [1:281](#)

- Poisson brackets [4:199, 4:200](#)
Poisson twice [3:277](#)
Poisson operators [4:200](#)
Poisson statistics [1:452](#)
Poisson-distributed random processes [4:373](#)
Poisson's equation [2:103–104](#)
Pockels effect [4:3](#)
polar coordinates, 2D Walsh functions in [5:18–19, 5:19F](#)
polar functional groups [3:8–9](#)
polarimetry [1:149](#)
polaritonic modes [1:449](#)
polaritonic systems [2:157](#)
polarizability [2:40](#)
polarization [1:148, 1:346, 1:347F, 4:180, 4:264–265, 4:310–311, 4:310F, 4:311F, 4:358, 5:169](#)
CCH [4:37, 4:37F, 4:38F](#)
control [2:418](#)
of quantum pathways in 2DCS measurements [2:154](#)
converter [3:33, 3:33F](#)
crosstalk [1:318–320](#)
dependent measurements [4:285, 4:285F](#)
diversity receiver [1:298](#)
ellipse [1:162–163, 1:163F](#)
holograms [4:95](#)
interferences [4:268](#)
maintaining [1:195](#)
microscopy [5:192](#)
modes [1:180](#)
preserving fibers [1:33](#)
selective transient grating [2:8–10](#)
sensitive materials [4:95–96](#)
TIR [5:169](#)
polarization beam combiner (PBC) [1:278](#)
polarization beam splitter (PBS) [1:181–183, 1:182F, 1:278, 1:298, 1:479–480, 1:480F, 3:35](#)
on and off for PBS and LCoS [3:35F](#)
3-PBS/x-prism LCoS projection system [3:35F](#)
prisms [1:151, 1:152F](#)
polarization dependent loss (PDL) [1:295](#)
equalization and polarization demultiplexing [1:302–304](#)
polarization division multiplexed systems (PDM systems) [1:293–294](#)
polarization gating (PG) [2:237, 2:238F, 2:319–320, 2:320F](#)
polarization mode dispersion (PMD) [1:1, 1:256–257, 1:295, 1:316–318, 1:317F, 1:363, 4:2–3, 4:3F](#)
equalization and polarization demultiplexing [1:302–304](#)
test methods for [1:316–318](#)
polarization switched systems (PS systems) [1:293–294, 3:101](#)
biodistribution [3:109–111](#)
pharmacokinetics [3:108–109](#)
structure [3:103–105](#)
polarization-based glucose content measurement [4:522–523](#)
polarization-dependent loss (PDL) [1:255, 4:226](#)
polarization-gating configuration [4:53–54](#)
polarization-insensitive modulators [1:142](#)
polarization-insensitive NOLMs [1:368](#)
polarization-maintaining fibers (PM fibers) [1:180](#)
polarization-mode dispersion (PMD) [1:178–179, 1:255](#)
polarizers [1:149–150, 2:211, 3:25–26, 4:523](#)
basic relations [1:149–150](#)
birefringent materials [1:150–151](#)
dichroic absorbers [1:151–153](#)
miscellaneous types [1:154](#)
reflection and transmission [1:153–154](#)
polarizing glass [3:45](#)
polarizing prisms [1:150–151, 1:151F](#)
polarons [5:223, 5:225F](#)
poly ether-imide (PEI) [1:210](#)
poly ether-sulfone (PES) [1:210](#)
poly-(3-hexylthiophene) films (P3HT films) [2:194, 5:323–324, 5:324–325, 5:324F](#)
poly-diamond carbon coated emitters *see* diamond-like carbon coated emitters (DLC coated emitters)
poly-L-lysine chloride(e6) (pL-ce6) [3:106](#)
poly-methylmethacrylate (PMMA) [3:15, 5:116–117](#)
fiber [4:279](#)
poly-Si thin film [3:13–14](#)
poly-vinylphenyle (PVP) [3:15](#)
poly(ethylene oxide) (PEO) [5:326–327](#)
poly(2-hydroxyethyl methacrylate) (pHEMA) [4:92](#)
poly(3-alkylthiophene) (P3AT) [3:15](#)
poly(3-hexylthiophene) (P3HT) [3:15](#)
poly(3-octylthiophene) (P3OT) [3:15](#)
poly(3,3-didodecylquaterthiophene) (PQTT-12) [5:326–327, 5:326F](#)
poly(3,4-ethylenedioxythiophene) (PEDOT) [5:279–280](#)
poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) (PEDOT:PSS) [5:233, 5:287, 5:287F, 5:298–300](#)
poly(3,4-propylenedioxythiophene) (PPD-OT) [5:279–280](#)
poly(9-vinylcarbazole) (PVK) [5:234](#)
poly(9,9-dioctylfluorene-co-benzothiadiazole) (F8BT) [5:327–328](#)
poly(acetylene) [5:309, 5:309F](#)
poly(acrylamide) (paam) [4:92](#)
poly(acrylic tetraphenyl-diaminobiphenol) (PATPD) [4:96–97](#)
poly(dimethylsiloxane) (PDMS) [4:92, 5:252](#)
poly(ethylene naphthalate) (PEN) [5:317](#)
poly(ethylene terephthalate) (PET) [5:317](#)
poly(methyl methacrylate) (PMMA) [3:60](#)
poly(N-vinylcarbazole) (PVK) [4:96–97](#)
poly(p-phenylenevinylene) (PPV) [5:240–241](#)
poly(vinyl alcohol) (PVA) [4:92](#)
poly[9,9-dioctylfluorene] (PFO) [5:309, 5:309F](#)
polyacetylene (PA) [5:223, 5:259](#)
polyaniline (PANI) [5:279–280](#)
polychromatic light [4:174](#)
polycrystalline silicon (poly-Si) [3:12–13](#)
polycrystalline silicon thin-film transistors (poly-Si TFTs) [3:13–14](#)
polydimethylsiloxane film (PDMS film) [5:293](#)
polyethylene naphthalate (PEN) [5:33](#)
polyethylene terephthalate (PET) [5:33](#)
polyglutamine fibrils structure [2:174–175](#)
polygon-based methods [4:123–124](#)
polyimide (PI) [3:56–57, 3:57, 5:248](#)
polymer light emitting diode (PLED) [3:56, 5:240–241](#)
polymer sustained alignment (PSA) [3:61](#)
polymer(s) [1:195, 1:196, 5:220, 5:221F, 5:310–311](#)
active materials [5:263–265](#)
coating [1:201–203](#)
fibers [1:210](#)
films [5:232](#)
laser resonators [5:313–314, 5:313F, 5:315F](#)
atomic force microscope images of “egg-box” DFB [5:314F](#)
●PVs [5:256](#)
organic photodiodes with organic layer–bulk heterojunction [5:324–325](#)
organic layer–single material [5:323–324](#)
organic phototransistors with organic layer–bulk heterojunction [5:327–328](#)
organic layer–single material [5:326–327](#)
polymer OLEDs, small molecule and [5:232](#)
polymer solar cells, active materials in [5:263–265](#)
polymer-based phototransistors [5:326–327](#)
polymeric photoresists [1:54](#)
polymeric processability [1:218](#)
polymerization [4:91](#)
polymyxin B nonapeptide (PMBN) [3:104](#)
polypyrrole (PPy) [5:279–280](#)
polystyrene (PS) [3:60](#)
fiber [4:279](#)
Polytec vibrometer, acoustic vibrations of violin acquired with [4:453, 4:455F](#)
polyurethane [1:331](#)
polyvinyl alcohol (PVA) [3:25–26](#)
polyvinylchloride (PVC) [1:331](#)
Ponderomotive Force [1:411](#)
population dynamics, investigation of [2:7–8](#)
population inversion [2:334, 2:350–351, 2:351, 2:372](#)
porphyrins [5:86](#)
Porro prisms [1:126, 1:127F](#)
portrait holography [4:155](#)
position-sensitive detector [5:190](#)
positive and negative affect schedule (PANAS-X) [3:126](#)
positive uniaxial crystal [2:293–294](#)
positive-intrinsic-negative (PIN) [4:6](#)
positron emission tomography (PET) [3:192, 3:196](#)
post-detection compensation [3:272](#)
blind deconvolution techniques [3:276–277](#)
DWFS [3:277–278](#)
speckle imaging techniques [3:274–276](#)
post-processing, adaptive optics and [3:278–279](#)
post-traumatic stress disorder (PTSD) [3:124](#)

- potassium dihydrogen phosphate (KDP) [2:190–191](#), [4:497](#)
- potassium titanyl phosphate (KTP) [1:16](#)
- Pouisselle's law [5:64](#)
- Pound–Drever–Hall method (PDH method) [2:227](#)
- power amplifier [1:354](#)
- power capability of laser material [2:413](#)
- power conversion efficiency (PCE) [2:355](#), [5:258](#), [5:283](#)
- power levelling sequence upstream (PLSu) [2:75–76](#)
- power management [1:283–284](#)
- power penalty (PP) [4:11](#)
- power ratio method [1:320](#)
- power recycled Michelson interferometer [1:436](#), [1:437F](#)
- power spectral density (PSD) [3:176](#), [3:176F](#)
- power times length process [2:426](#)
- power–current characteristic (P – I characteristic) [2:354–355](#)
- Powervision [5:121](#)
- Powervision FluidVision IOI [5:121](#)
- Poynting vector [1:164](#)
- PQBOC8 [5:326–327](#), [5:326F](#)
- Praseodymium-doped amplifiers [1:263](#)
- pre-detection compensation [3:272](#), [3:278](#)
- preamplifier [1:248](#), [1:354](#)
- preferential pumping [2:371](#)
- preform [1:195](#), [1:210](#), [4:14–15](#)
- construction process [1:211](#)
- fabrication [1:28](#), [1:195–196](#)
- MCVD process [1:28–29](#), [1:29F](#)
- other methods [1:30](#)
- OVD process [1:29](#)
- VAD process [1:29–30](#), [1:30F](#)
- preform-based fabrication approach [1:210](#)
- preform substrate [1:213](#)
- presbyopia [5:52](#), [5:61](#), [5:120–122](#), [5:121T](#), [5:141](#), [5:150–151](#)
- binocular approaches to presbyopia correction [5:125](#)
- extending eye's depth of focus [5:122](#)
- multifocal wavefront aberrations [5:122–125](#), [5:124F](#)
- pinhole effect [5:122](#), [5:123F](#)
- restoring accommodation [5:120–122](#), [5:122E](#)
- primary coating [1:328](#)
- primary sources [1:42–43](#)
- primitive cell [1:170](#)
- primitive reciprocal lattice vector [1:170](#)
- principal axes [1:387](#)
- Principal Component Generalized Projections Algorithm (PCGPA) [2:322–323](#)
- principal quantum number [2:40](#)
- principal self-terminating resonance-metastable metal vapor lasers [2:369T](#), [1:302–303](#), [1:316](#)
- principal wavelengths [2:368–369](#), [2:373–374](#)
- principle of equivalence [5:189](#)
- principle of reciprocity [5:189](#)
- print circuit board (PCB) [3:89](#)
- prior knowledge [3:157–159](#)
- priori information [4:176](#)
- prisms [1:124](#), [1:125F](#)
- deviation [1:126–128](#), [1:127F](#)
- double prism reflectors [1:126](#)
- as instruments [1:126–128](#)
- pre-expanded grazing incidence grating oscillators [2:337](#)
- prism-type devices [1:143](#)
- single prisms as reflectors [1:124–126](#)
- spectrometer [1:128F](#)
- pro-collagen [3:118–119](#)
- probability densities (PD) [1:365](#)
- probability density function (pdf) [3:277](#)
- probe microscopes
- see also* conventional microscope
- atomic force microscope [5:190](#)
- NSOM [5:190–191](#), [5:190F](#)
- scanning tunneling microscope [5:190](#)
- probe mode decomposition model [3:153](#)
- probe modulation [1:21](#)
- probe pulse [5:312–313](#)
- probing [2:3](#)
- ocular mechanics with light [5:141](#)
- air puff corneal dynamic imaging [5:142](#)
- Brillouin microscopy [5:148–149](#)
- phase-sensitive optical coherence elastography [5:146–147](#)
- programmable optical setup [4:178](#), [4:178F](#)
- projected Landweber method [3:191](#)
- projection
- CRT [3:1](#)
- re-scaling [4:178](#)
- sets [4:177–178](#), [4:178](#)
- subset [4:178](#)
- projection display [3:32](#), [3:32F](#)
- see also* display(s)
- color gamut and HDR [3:40–43](#), [3:42F](#)
- 4K × 2K [3:37](#)
- laser phosphor illumination projector [3:39–40](#), [3:40F](#)
- LCD [3:32–34](#)
- LCoS [3:34–35](#)
- light source [3:32](#)
- MEMS [3:35–37](#)
- microdisplay [3:32](#)
- new technologies, challenges, and applications [3:37](#)
- system [3:32](#)
- UST [3:37–39](#)
- projection-based blind deconvolution (PBBd technique) [3:276–277](#), [3:277](#)
- projection-slice theorem [2:161](#)
- prolate ellipsoids [5:45](#)
- prolate spheroidal wave functions [5:16](#)
- proliferative diabetic retinopathy (PDR) [5:132](#)
- prompt fluorescence (PF) [5:245](#)
- proof stressing [1:326](#)
- propagation
- of coherence [1:38–39](#)
- see also* coherent/coherence
- generalized propagation [1:39](#)
- Thompson and Wolf experiment [1:39–40](#), [1:39F](#)
- van Cittert–Zernike theorem [1:38–39](#)
- constant [1:6–7](#), [1:7F](#)
- in Doppler broadened medium [4:344](#)
- in free space [4:167](#)
- holographic phase retrieval from images [4:171–172](#)
- application to phase retrieval and X-ray holotomography [4:172](#)
- Fresnel diffraction images as in-line holograms [4:171–172](#)
- in paraxial optical system [4:168](#)
- through paraxial optical system in terms of AF
- transmission through thin object [4:167–168](#)
- proportionality constant [4:351–352](#)
- propylene carbonate (PC) [5:277](#)
- prostacyclin [3:111](#)
- proteins [1:445](#)
- function [3:102](#)
- structure and folding dynamics [2:173–175](#)
- amyloid proteins [2:174–175](#)
- membrane proteins [2:176](#)
- protons [2:40](#), [5:192](#)
- proximity effect [1:430](#)
- pseudo colour and manipulative artist [4:104–105](#)
- pseudo image [4:152](#)
- pseudo-3D image [4:106](#)
- pseudo-Doppler range pulling [4:439](#)
- pseudogap of YBa₂Cu₃O_x, magnetic quadrupole order in [2:220–223](#), [2:222F](#), [2:223F](#)
- pseudoscopic conjugate image [3:209–210](#), [3:209F](#)
- pseudoscopic image [3:208–209](#), [3:209](#)
- pseudoscopic real image [4:152](#)
- pseudospin [4:265](#)
- psychiatric disorders, PBM for [3:125–126](#)
- psychological depth cue [3:44](#), [3:44F](#)
- psychomotor vigilance task (PVT) [3:126](#)
- psychophysics [5:87](#)
- PtP-WDM channel overlay [2:79](#)
- ptychography [3:148](#), [3:152](#)
- pulsatility [5:65–66](#)
- pulse amplification model [1:248](#)
- pulse characterization techniques
- autocorrelation [4:49–50](#)
- FROG [4:52–54](#)
- inversion algorithms [4:54–57](#)
- measurements [4:57](#)
- mathematical representation of optical pulse [4:48–49](#)
- measuring pulses directly–SPIDER [4:57–58](#)
- measuring weak pulses [4:58–59](#)
- pulse measurement method [4:59](#)
- simplified FROG–GRENOUILLE [4:54](#)
- spectral interferometry–relative phase measurements [4:50–51](#)
- time-frequency representation [4:51–52](#)
- pulse compression [2:297–298](#), [2:297F](#)
- applications [1:233](#)
- pulse distortion *see* optical time domain measurement method
- pulse duration [2:372–373](#)
- pulse field [4:49](#)
- pulse generation technique [4:62](#), [4:62F](#), [4:63F](#)

- pulse generator [4:45](#)
- pulse laser deposition (PLD) [3:15](#)
- pulse LID [4:302](#)
- pulse measurement techniques [4:48](#), [4:59](#)
- pulse mode [4:497](#)
- pulse oximeter [3:89](#)
- pulse oximetry [4:519](#)
- pulse propagation [1:18–19](#)
 - for weak CW fields [4:345–346](#), [4:345F](#)
- pulse repetition frequency (PRF) [2:329](#), [2:334](#), [4:479](#)
- pulse response [4:88–89](#)
- pulse sequencing to control quantum [2:154](#)
 - double quantum 2D spectra [2:155](#), [2:155F](#)
 - nonrephasing 2D spectra [2:154–155](#)
 - rephasing 2D spectra [2:154](#), [2:155F](#)
- pulse shape [2:316](#)
- Pulse shaping [5:41](#)
- pulse synthesis [2:314–316](#)
- pulse tiling technique [1:409](#), [1:410F](#)
- 3-pulse transient four-wave mixing [2:156–157](#)
 - resolution in 2D spectroscopy [2:157](#)
- pulse width modulation (PWM) [3:81](#)
- pulse-amplitude modulation (PAM) [1:277](#)
- pulse-current electrodeposition technique [5:298–300](#)
- pulse-shapers [2:188](#)
- pulse-timing control methods [5:40](#)
- pulsed CO₂ laser [4:457](#)
- pulsed gas discharges [2:376](#)
- pulsed laser deposition (PLD) [2:258](#)
- pulsed lasers [4:347](#)
 - field [4:344](#)
 - holograms [4:102](#)
- pulsed narrow-linewidth tunable dye laser [2:337](#)
- pulsed operation [2:418–419](#)
 - active Q switching [2:418–419](#)
 - mode locking [2:421](#)
 - passive Q switching [2:419–421](#)
- pulsed pump lasers [5:314](#)
- pulsed repetition frequencies (prfs) [2:336](#)
- pulsed system [5:3](#)
- pulses
 - measuring pulses directly–SPIDER [4:58–59](#), [4:58F](#)
 - measuring weak pulses [4:58–59](#)
- pump
 - cavity [2:409](#)
 - chamber [2:409](#)
 - through design [1:358](#)
 - module [2:409](#)
 - pulse [1:181](#)
 - pump-dump control methods [5:40](#)
- pumping [2:392–393](#), [2:409–410](#), [2:410F](#), [2:416–417](#)
 - atoms [1:455](#)
 - of dye lasers [2:372](#)
 - OPSLs [2:284–285](#), [2:284F](#)
 - Ti:sapphire lasers [2:372](#)
- pump-probe
 - experiment [2:3](#), [4:270](#)
 - geometry [2:156](#), [2:167](#)
 - 2DES in [2:189](#)
 - pump spectroscopy [2:240–241](#)
 - Raman scattering [2:85](#)
 - techniques [2:242](#), [5:311](#)
 - pupil plane recording geometry (PPRG) [4:369](#)
 - pupil-Alt [4:169](#)
 - as generalization of OTF [4:169](#)
 - pupils [5:45–46](#), [5:46F](#)
 - design for extended depth of field [3:173](#)
 - fields [4:409](#), [4:410F](#)
 - plane [4:371–372](#)
 - purcell factor [1:446](#)
 - pure laser [3:39](#)
 - pure states [1:478](#)
 - pyrex glass tube walls [2:374](#)
 - pyroelectric detectors [1:418](#), [1:420](#)
 - pyromethene 567 dye (PM567 dye) [2:345](#), [2:345F](#), [2:348](#)
 - pyromethenes [2:344–345](#)
 - dyes [2:347](#)

Q

 - Q-switch(ing) [1:140](#), [2:412](#)
 - with bulk saturable absorbers [2:419–421](#)
 - with semiconductor saturable-absorber mirrors [2:421](#)
 - Q-switched lasers [2:424](#), [4:351](#)
 - microchip lasers [2:422](#)
 - single-frequency lasers [2:451](#)
 - QDash SOAs [1:245](#)
 - QinetiQ system [4:116](#), [4:116F](#)
 - QR-LPD[®] [3:82–84](#)
 - quadratic dependence of dispersion [4:9–10](#)
 - quadratic optical Stark effect [4:319–320](#)
 - quadrature amplitude modulation (QAM) [1:245](#)
 - 2-QAM [1:293](#)
 - 4-QAM [1:293](#)
 - quadrature amplitude modulation (QAM) [1:245](#), [1:277–278](#)
 - quadrature PR (QPR) [1:279](#)
 - quality factor (Q factor) [1:147](#), [1:187–188](#), [1:188](#), [1:215](#), [4:218](#), [4:218–219](#)
 - quality-of-service degradation [2:80](#)
 - quantitative information [2:165](#)
 - quantitative performance [4:225–226](#)
 - quantized states [1:459–461](#)
 - quantum beats [4:267–269](#)
 - quantum bits [1:478](#)
 - quantum cascade (QC) [2:19](#)
 - few cycle THz spectroscopy of [2:25–31](#)
 - heterostructure [2:25](#), [2:120](#)
 - lasers [1:381–382](#)
 - quantum computers [1:478](#), [1:484](#)
 - quantum computing methods [1:456–457](#), [1:462](#), [1:484](#), [4:174](#)
 - quantum confined Stark effect (QCSE) [2:96](#), [2:278–279](#), [4:243](#), [4:254](#), [4:255–256](#)
 - quantum confinement [2:19](#), [2:99](#)
 - influence [2:94–96](#)
 - quantum control [1:487](#), [5:31](#)
 - quantum cross-capacitance effect [1:474](#)
 - quantum cryptography [1:478](#), [1:481F](#)
 - quantum dash (QDash) [1:244](#)
 - quantum defect [2:46–47](#)
 - quantum dots (QD) [1:242](#), [1:245](#), [1:460–461](#), [1:469](#), [1:484](#), [1:491](#), [2:99](#), [2:280](#), [3:28–29](#), [3:29F](#), [3:61–62](#), [3:69](#), [3:211](#), [5:305–306](#)
 - CW OPSI sources [2:286](#)
 - SOAs [1:245](#), [1:245F](#), [1:247–248](#)
 - quantum dynamics phenomena [5:31](#), [5:36](#)
 - quantum efficiency (QE) [2:113–114](#), [4:419](#), [4:426F](#)
 - quantum electrodynamics (QED) [1:458](#), [4:180](#)
 - quantum information [1:456–457](#), [1:478](#)
 - application of optical entanglement [1:484](#)
 - basic notions [1:478](#)
 - creating entangled states experimentally [1:479–480](#)
 - entangled-photon source [1:480F](#)
 - multiqubit states and entanglement [1:478–479](#)
 - polarization entanglement-based quantum cryptography protocol [1:482T](#)
 - protocols [1:479](#)
 - quantum key distribution [1:480–482](#)
 - quantum superdense coding [1:482](#)
 - quantum teleportation [1:482–484](#)
 - qubits [1:478](#)
 - quantum interference phenomenon [1:487](#), [4:339](#)
 - quantum key distribution (QKD) [1:479](#), [1:481](#)
 - quantum kinetics [2:83](#), [2:84](#)
 - quantum light characterization [4:179–181](#)
 - quantum lithography [1:484](#)
 - quantum mechanical/mechanics [1:144–146](#), [1:169–170](#), [1:172](#), [1:458](#), [1:462](#), [1:479](#), [1:491](#), [4:199](#), [4:200](#), [5:30](#)
 - description of optical SHG susceptibility [2:208–209](#)
 - energy levels [2:99](#)
 - expressions [4:336](#)
 - interference [1:491](#)
 - laser field design [5:31–32](#)
 - objectives [5:34](#)
 - tunneling [2:101–102](#)
 - quantum noise [1:273](#), [1:442](#)
 - quantum optical correlations, observation of [1:461–463](#)
 - quantum optics [1:458](#), [4:180](#)
 - quantum optimal control theory for designing laser fields [5:31–34](#)
 - quantum pathway selection in 2D coherent spectra [2:153–154](#)
 - polarization control of quantum pathways in 2DCS measurements [2:154](#)
 - pulse sequencing to control quantum pathways in 2DCS measurements [2:154](#)
 - quantum pathways in 3DCS and other higher-order spectroscopies [2:155–156](#)
 - quantum phenomena [5:39](#)
 - quantum state teleportation [1:483–484](#), [1:483F](#)
 - quantum state tomography [1:478](#), [1:480](#)
 - quantum system [1:487](#), [5:30](#), [5:30F](#), [5:35F](#)

- quantum theory [2:112](#)
 of dyes [2:337](#)
 quantum well intermixing (QWI) [4:244–245](#)
 quantum wells (QW) [1:451](#), [1:459](#), [1:459F](#),
[1:491](#), [2:68–69](#), [2:84–85](#), [2:91](#),
[2:98–99](#), [2:103–104](#), [2:104](#), [2:201](#),
[2:202F](#), [2:278–279](#), [2:280](#), [2:281F](#),
[2:282F](#), [2:462–463](#)
 electron states in [2:100–102](#)
 formation [2:103–105](#)
 and GaAs-based structures
 absorption spectrum [2:107F](#)
 AlGaAs/GaAs quantum wells, optical
 properties of [2:105–107](#)
 electron energy diagram [2:98F](#), [2:101F](#),
[2:102F](#)
 electron states in bulk material
[2:99–100](#), [2:100–102](#)
 formation [2:103–105](#)
 GaAs/AlGaAs quantum well structures
[2:105](#)
 occupancy of states in [2:102–103](#)
 properties of electrons in large cubic
 sample [2:100F](#)
 intersub-band transitions in [1:488](#)
 lasers [2:350](#)
 nanostructures [2:58](#)
 occupancy of states in [2:102–103](#)
 SOAs [1:245](#)
 MZI logic gates [1:251](#)
 quantum wires and dots [2:71–72](#)
 quantum-beat spectroscopy [4:269](#)
 quarter wave [1:155](#)
 quarter-wave plate (QWP) [1:479–480](#),
[1:480F](#), [2:211](#), [2:212F](#), [3:35](#)
 quasi phase matching (QPM) [1:14–16](#),
[1:14F](#), [1:15F](#)
 quasi-2D nature [2:69](#)
 quasi-classical action [2:318](#)
 quasi-CW
 laser [2:382](#)
 mode-locked Ti:Sapphire laser [4:367](#)
 nanosecond duration laser pulses [5:39–40](#)
 quasi-DC polarization [1:406–407](#)
 quasi-equilibrium distributions [2:464–465](#)
 quasi-Fermi energy [2:81](#)
 quasi-Fermi levels [2:111](#), [2:355–356](#)
 quasi-free electrons in solids [2:118](#)
 quasi-homogeneous sources [1:44–45](#), [1:44F](#)
 quasi-monochromatic [3:263](#)
 coherent beam [4:174](#)
 fields [1:41](#)
 light [1:486](#)
 quasi-Newton's methods [3:150](#)
 quasi-periodic structure [5:252](#)
 quasi-phase-matching (QPM) [2:293](#), [2:399](#)
 quasi-solid-state electrolyte [5:277–278](#)
 quasi-spectral analysis [3:91](#)
 quasi-thermal equilibrium [2:94](#)
 quasi-three-level lasers [2:417](#)
 qubits [1:478](#), [1:478](#)
 couplings and hybridizations [1:473–474](#)
 capacitive and electric dipole coupling
[1:474–475](#)
 contact hyperfine interaction [1:473–474](#)
 magnetic dipole coupling [1:475](#)
 spin qubit coupling with exchange [1:474](#)
 creating entangled states experimentally
[1:479–480](#)
 multiqubit states and entanglement
[1:478–479](#)
 pair [1:482](#)
 quick response liquid powder display (QR-
 LPD) [3:82](#)
 quinolone [2:346](#)
- ## R
- rabbit small clot embolic model (RSCFM)
[3:121–122](#)
 Rabi coupling [4:339–340](#), [4:340](#)
 Rabi energy [2:48–49](#)
 Rabi flopping frequency [1:452](#), [2:54](#)
 Rabi frequency [4:339–340](#)
 radar [4:474](#)
 radial canonical transforms [4:203](#)
 radial pumping [2:409–410](#)
 radial shift variance [3:263](#)
 radial Walsh functions [5:19](#)
 radiance field [3:176](#), [3:176F](#)
 radiated field, propagating and evanescent
 spectra of [3:140–142](#)
 radiation
 acoustic [5:192](#)
 electromagnetic [5:191](#)
 electron microscope [5:191–192](#), [5:191F](#)
 generation of number states of radiation
 field [1:454–455](#)
 modes [1:229](#), [3:67](#)
 other matter [5:192](#)
 radiation-atom interaction [1:452](#)
 trapping [2:378](#)
 radiative decay [2:12–15](#), [4:266](#)
 rates [1:445](#), [1:447](#)
 radiative process [2:117](#)
 radiative relaxation [2:328](#)
 radiative transitions
 detailed balance [2:114](#)
 donor-acceptor pair recombination [2:113](#)
 quantum efficiency [2:113–114](#)
 spontaneous and stimulated emission
[2:112–113](#)
 radio frequency (RF) [1:32](#), [4:333](#), [4:431](#), [5:1](#)
 channel [2:73](#)
 coherent detection
 32x32 pixel FPA lidar breadboard using
 incoherent chirped amplitude
 modulation [5:6–7](#)
 incoherent FM/CW lidar architecture
[5:1–2](#)
 kinetic inductance detectors [1:427–428](#)
 RF-electrodes [2:334](#)
 signal power [1:139](#)
 spectrum [4:13](#)
 spectrum analyzer [2:304](#)
 radio-pharmaceuticals [3:195](#)
 radiometric units [3:70](#)
 Radon transform inversion [3:192–193](#),
[3:194](#), [4:177](#)
 Radon-Wigner display (RWD) [4:177](#), [4:179F](#)
- Radon-Wigner transform [4:177](#)
 rainbow effect [3:33](#), [3:37](#)
 rainbow-colour effect [4:103–104](#)
 Raleigh-leans theory of cavity radiation
[2:107](#)
 Raman absorption process [4:320](#)
 Raman amplification with co-and counter-
 propagating pumps [1:265F](#)
 Raman amplifier (RA) [1:362](#), [4:6](#)
 Raman lasers [2:265](#), [2:265–266](#)
see also microchip lasers; thin disk lasers
 advancements in
 crystalline Raman lasers [2:266–267](#)
 fiber Raman lasers [2:268–269](#)
 applications [2:269](#)
 energy diagram of stimulated Raman
 scattering [2:265F](#)
 on-chip Raman lasers [2:267–268](#)
 Raman lasing [2:269](#)
 Raman lidar techniques [4:456](#)
 Raman micro-lidars [4:516–518](#)
 Raman microscopy [4:354](#), [4:354–356](#),
[4:354F](#), [5:192–193](#)
 Raman optical activity (ROA) [4:358](#)
 Raman probe tip [4:517–518](#), [4:517F](#)
 Raman resonance, nonlinear optics with pair
 of strong coupling fields in
[4:344–345](#)
 Raman saturation phenomena [4:366](#)
 Raman scattering [1:180](#), [1:258](#), [2:265](#), [2:266](#),
[4:351](#), [4:467–468](#)
 Raman spectroscopy [2:164](#), [4:354](#)
 characteristics of laser sources [4:356–358](#)
 near infrared lasers [4:361–362](#)
 Raman microscopy [4:354–356](#)
 system [4:517–518](#), [4:518F](#)
 UV lasers [4:364–366](#)
 visible lasers [4:359–360](#)
 Raman-induced absorption [4:319–320](#)
 Raman-induced Kerr effect spectroscopy
 (RIKES) [2:247](#)
 Raman-Nath regime [4:87](#)
 Raman-Nath diffraction [1:131–132](#), [1:132F](#)
 Raman-Nath parameter [1:131–132](#)
 ramp-filter [3:194](#)
 Ramsey method [1:452](#)
 random and uniformly redundant arrays
[3:252](#)
 random sample consensus (RANSAC)
[5:79](#)
 random walk theory (RWT) [3:222](#)
 range imaging with matricial light-emitting
 diode array [4:509–510](#), [4:509F](#)
 range precision [4:430–431](#)
 range profiles studies
 airborne targets [4:479–482](#)
 atmospheric profiling [4:483](#)
 dimensions and structure of small boat
[4:481F](#)
 examples [4:477–479](#)
 land targets and mapping [4:482–483](#)
 naval targets [4:477–479](#)
 test vehicles for trial in Sweden [4:480F](#)
 range resolution [4:430–431](#), [4:430F](#)
 range responses [5:5](#)
 range-gate output [5:5](#), [5:5F](#)

- range-height indicator scan of absolute humidity field [4:464](#), [4:467F](#)
- range-resolved reflective tomography techniques [4:488](#)
- rapid thermal annealing (RTA) [4:245](#)
- rare gas halide lasers
- absorbers in [2:363F](#)
 - potential energy curves [2:362F](#)
- raster scanning approach [4:194–195](#), [4:194F](#)
- rate equations [4:304](#)
- ratiometric analysis [3:90–91](#)
- ray aberrations converting to wavefront aberrations [5:113–115](#)
- ray deflections, optical path lengths
- determining from [5:112–113](#)
- ray description of fiber propagation [1:223](#)
- ray theory [1:222](#)
- ray tracing aberrometry [4:531–533](#), [4:532F](#)
- commercial iTrace visual function analyzer [4:532–533](#), [4:532F](#)
 - locating visual axis [4:533](#)
 - reconstructed maps and functions [4:533](#), [4:533F](#)
- ray transformation matrix [4:176](#), [4:178](#)
- ray-optics limit [5:294–295](#)
- Rayleigh criterion [3:137](#), [3:139–140](#), [4:403](#), [5:186](#), [5:187](#)
- Rayleigh diffusion [3:202](#)
- Rayleigh expansions *see* grating equations
- Rayleigh imaging [2:409](#)
- Rayleigh lidar techniques [4:456](#)
- Rayleigh range [4:357](#), [4:357F](#)
- Rayleigh regime [3:220](#)
- Rayleigh resolution criterion [3:137–138](#), [4:430](#)
- Rayleigh scattering [1:149](#), [1:217–218](#), [1:255](#), [4:366](#)
- Rayleigh–Sommerfeld diffraction formula [3:142](#), [4:141–143](#)
- Raytheon Vision Systems (RVS) [1:426](#)
- Rb cell [2:229](#), [2:229F](#)
- re-amplifying, re-shaping, and re-timing (3rd R²) [1:364](#), [1:364F](#)
- reactive gases [2:382](#)
- reactive ion etching (RIE) [4:294](#)
- reactive oxygen species (ROS) [3:100–101](#), [3:115–116](#), [5:131](#)
- read out and integration circuit (ROIC) [4:373](#), [4:446](#)
- reading stones [5:120](#)
- readout
- mechanism [1:471–472](#)
 - charge qubits [1:471–472](#)
 - spin qubits [1:472–473](#)
 - points or strips [4:524](#)
- readout integrated circuit (ROIC) [4:419](#), [5:8](#)
- real 2D-dimensional space (R^{2D}) [4:199](#)
- real energy bands [2:87](#)
- real-time
- holography [4:327](#)
 - system [3:199](#)
- real-valued monochromatic optical wave [4:322](#)
- real-valued variant Discrete Multi-Tone [1:255](#)
- receiver circuitry [4:259–261](#)
- see also* transmitter circuitry
- bandwidth-limited frontends [4:261–263](#)
- wideband transimpedance amplifiers [4:260–261](#)
- receivers [1:270–271](#)
- circuits [4:222](#)
- receptor [3:250–251](#)
- recollision spectroscopy [2:239–240](#), [2:240F](#)
- recombination
- assumptions for nonequilibrium statistics [2:111](#)
 - coefficients [2:110](#), [2:110F](#)
 - effect of electron–boson interactions [2:110](#)
 - electron–electron interaction
 - for electrons in bands [2:110](#)
 - in traps [2:110–111](#)
 - lasers [2:372](#)
 - nonradiative processes [2:114–115](#)
 - radiative transitions [2:112–113](#)
- rates
- case of defects [2:112](#)
 - results and two-band case [2:111–112](#)
- reconfigurable optical add/drop multiplexers (ROADM) [1:260](#), [1:266](#), [1:268](#), [1:269–270](#), [1:269F](#), [4:225](#), [4:226–227](#), [4:228F](#)
- colorless [4:227](#)
- contentionless [4:227](#)
- degree [4:227](#)
- directionless [4:227](#)
- fixed grid [4:227](#)
- flex-grid [4:227](#)
- reconfigurable system [4:31](#)
- reconfiguration within data center [4:225](#)
- reconstructed 3D slice plot of blood flow in deep tissues [4:526](#), [4:527F](#)
- reconstruction algorithm [3:150](#)
- reconstruction of attosecond beating by interference of two-photon transitions (RABBITT) [2:241–242](#), [2:242F](#), [2:321–322](#), [2:322F](#)
- reconstruction principles in digital holography [4:143–145](#)
- Fresnel approximation [4:143–145](#)
- numerical reconstruction
- by convolution approach [4:144F](#), [4:145–146](#)
 - by lensless Fourier approach [4:146–147](#)
- reconstruction wave [1:377](#)
- recording
- emulsion [4:104](#)
 - materials [4:154](#)
 - step [3:250–251](#)
- rectangle function [4:385](#)
- rectangular apertures [1:87–88](#)
- diffraction by [1:69–70](#)
- rectangular coordinates, two dimensional
- Walsh functions in [5:18](#)
- rectification process [1:421–422](#)
- recursive algorithms [3:150–151](#)
- red, green, and blue (RGB) [3:18–19](#), [4:103–104](#)
- red AlInGaP laser diodes [2:278–279](#)
- red laser diodes [2:356](#)
- red green glasses *see* anaglyph
- reducible time delay [1:41–42](#)
- reduction–oxidation measurements (redox measurements) [5:93](#)
- reference [5:72](#)
- beam [4:151](#)
 - plane wave [1:374–375](#)
 - wave [1:376–377](#), [3:206–207](#)
- reference mirror (RM) [4:507](#)
- reference signal (RS) [4:500](#)
- reflectances [1:153](#)
- of polarizer [1:149](#)
- reflected beam in TIR [5:170–171](#)
- reflection [1:153–154](#), [1:172](#), [2:85](#)
- holograms [4:87](#)
 - holography [4:103](#), [4:104](#)
 - intensity [2:388](#)
 - microscopy [5:192](#)
 - polarizer [1:153](#)
 - spectrum [2:388](#), [2:389F](#), [2:391F](#)
- reflective display technologies [3:6F](#), [3:82–84](#), [3:83F](#)
- EPD [3:79–80](#)
 - EWB [3:80–82](#)
- reflective LCD [3:6–7](#), [3:6F](#)
- reflective objective (RO) [2:211](#)
- reflective semiconductor optical amplifier (RSOA) [1:252](#), [2:74](#), [2:82](#)
- based passive optical network [1:252](#)
- reflective SLMs [3:42–43](#)
- reflective tomography [4:485](#), [4:489](#)
- reflectivity [3:71](#)
- reflectors [1:111–113](#), [2:409](#)
- refracted near field technique (RNF technique) [1:322](#), [1:323](#)
- refracting power [1:114–116](#)
- refraction [5:61](#), [5:111](#)
- diffracted beam index [1:138–139](#)
 - index [4:310](#)
 - index of refraction profile [1:207](#), [1:207F](#)
 - matrix [1:102–104](#)
- refractive anomalies [5:51–52](#), [5:55](#)
- anisometropia [5:54](#)
 - astigmatism [5:52–54](#), [5:53F](#)
 - correction [5:54–55](#)
 - spherical [5:51–52](#)
- refractive anomalies, correction of [5:54–55](#)
- refractive correction [5:108–109](#), [5:108F](#)
- refractive errors [5:108](#), [5:111](#), [5:135](#)
- aberrometry and wavefront sensing [5:112–113](#)
 - excimer laser [5:135](#)
 - femtosecond laser [5:135–136](#)
 - higher-order aberrations specification of eyes [5:110–111](#)
 - methods for determining refractive error and refractive state [5:111](#)
 - eyes refractive state [5:111](#)
 - locating eyes far point with optical measurements [5:111](#)
 - refractive state from Zernike aberration coefficients [5:111–112](#)
- optimizing retinal image quality by correcting [5:115](#)
- refractive correction, and physical units [5:108–109](#), [5:108F](#)
- specification of sphero-cylindrical lenses and [5:109–110](#)

- refractive index [1:5–6](#), [1:27](#), [1:33](#), [2:5](#),
[4:238–239](#), [4:329](#), [4:342](#), [4:358](#)
effects [4:3](#)
of material [1:153](#)
refractive index profile (RIP) [1:311](#)
refractive lens system [4:205–206](#), [4:206F](#)
refractive multifocal wavefronts [5:123](#)
refractive state
eyes refractive state [5:111](#)
methods for determining [5:111](#)
from Zernike aberration coefficients
[5:111–112](#)
refractive surgery [5:125–126](#)
refractory metal(s) [2:262–263](#)
nitrides [2:260](#)
refresh rate [3:77–78](#)
refreshable materials
holographic recording media and devices
[4:93](#)
persistent spectral hole burning [4:93–95](#)
photochromic materials [4:93](#)
photorefractive materials [4:96–97](#)
photothermoplastic process [4:97](#)
polarization sensitive materials [4:95–96](#)
regenerative amplification [2:412](#)
regional cerebral blood flow (rCBF) [3:126](#)
regioregular poly(3-hexylthiophene) (RR-
P3HT) [5:261](#)
regrowth integration technology [4:250](#)
regular DFB [2:455–456](#), [2:456F](#)
regularization
parameter [3:189](#), [3:190F](#)
theory [3:187](#)
regulated cascode (RCC) [4:260–261](#)
relative horizontal accuracy [4:402](#)
accounting for [4:403–404](#)
relative intensity noise (RIN) [2:458](#), [2:458F](#),
[4:46–47](#)
relative vertical accuracy [4:401–402](#)
relaxation
effects [1:216](#)
oscillations [2:467](#), [4:62](#)
relaxed averaged alternating reflectors
(RAAR) [3:148](#)
reliability of lasers in fiber optic systems
[4:70–71](#)
remote node (RN) [1:252](#)
RN1 [2:82–83](#)
RN2 [2:82–83](#)
remotely operated vehicles (ROVs) [4:110](#)
repetition rate [2:306](#)
locking [2:307–308](#)
rephrasing [2:54–55](#)
complex signal fields [2:184–185](#)
signals [2:185–186](#)
time-ordering [2:57](#)
2D spectra [2:154](#), [2:155F](#)
residual stress [1:326](#)
resistance furnace [1:31–32](#)
resistive-feedback transimpedance amplifier
(RTIA) [4:429](#)
resolution [1:135–136](#), [3:137](#), [3:235–237](#),
[4:430](#), [5:158–159](#), [5:185–186](#),
[5:186F](#)
and image sharpness [3:75](#)
in imaging [3:136–137](#)
computational [3:136F](#)
inverse source and scattering problems
[3:137](#), [3:137F](#)
physical [3:136F](#)
scatterer [3:142–144](#)
source-field [3:138](#)
limit [3:144](#), [5:72](#)
superresolution [3:144–145](#)
resolving power of zone plate [1:96](#)
resonance
transition [2:368](#)
upper laser level [2:368](#)
vibration frequencies [4:455](#), [4:456F](#)
wavelengths [1:231–232](#)
resonance enhanced multiphoton ionization
(REMPI) [2:248](#)
resonance-metastable metal vapor lasers
[2:372–373](#)
resonant cavity [1:451](#), [2:385](#)
APDs [4:419](#)
resonant energy transfer [2:327–328](#)
resonant enhancement [4:350](#)
resonant excitation conditions [2:33](#)
resonant periodic gain (RPG) [2:281](#)
resonant phonon scheme (RP scheme)
[1:383](#)
resonant THz detection [1:423](#)
resonant-phonon designs (RP designs) [1:382](#)
resonated devices [1:143](#)
resonator [2:465–467](#)
spectrometer [1:395–396](#)
types of resonator modes [2:329–331](#)
response time [3:76–77](#)
responsivity of organic photodetectors
[5:320–321](#)
restoring accommodation [5:120–122](#),
[5:122F](#)
retardation plates *see* retarders
retarders [1:149](#), [1:154–156](#)
retina (R) [3:17](#), [3:18F](#), [4:500](#), [5:43](#), [5:130](#)
femtosecond laser effects on [5:91–92](#)
light at [5:50](#)
light loss at [5:61](#)
retinal blood flow measurement
capillary blood flow [5:65](#)
clinical measurements [5:69](#)
flow measurements [5:69](#)
total retinal blood flow [5:64–65](#)
retinal diseases [5:64](#)
retinal image quality [5:59–60](#), [5:124–125](#),
[5:161](#)
optimizing by correcting refractive errors
[5:115](#)
retinal imaging [5:72](#)
AO [5:73](#)
femtosecond lasers in
artifacts and implementation of TPE
fluorescence ophthalmoscopy [5:89–91](#)
effects on retina [5:91–92](#)
fluorophores [5:85F](#)
multiphoton retinal imaging theory
[5:87–89](#)
in vivo multiphoton retinal imaging
with adaptive optics [5:86–87](#)
retinal oximeter on platform of Canon CR
6–45NM [4:520](#), [4:520F](#)
retinal pigment epithelium (RPE) [5:75](#),
[5:75F](#), [5:164](#)
dark-field AOSLO [5:76F](#)
retro-reflective targets [4:439](#)
retro-scattering [4:456](#)
retrograde mitochondrial signaling
[3:115–116](#)
return loss (RL) [1:337](#)
return pulse [5:3](#)
return-to-zero (RZ) [4:39](#), [4:68](#)
format [4:21–22](#)
pulses [1:20–21](#), [1:275](#)
RZ33 [1:275](#)
RZ50 [1:275](#)
RZ67 [1:275](#)
transmission [1:294–295](#)
reverse dispersion fiber [4:24–25](#)
reverse intersystem crossing (RISC) [5:241](#)
rhemoglobin [5:86](#)
rhodamine 590 chloride *see* rhodamine 6G
(Rh6G)
rhodamine 6G (Rh6G) [2:344](#)
rhodamine dyes [2:344](#), [2:349](#)
rhodopsin densitometry [5:87](#)
ribbon cable *see* slotted core cables
riboflavin [5:126](#)
Richardson–Lucy method [3:191–192](#)
right sub-pixel (RSP) [3:62](#)
rigid gas permeable contact lenses (RGP
contact lenses) [5:116–117](#), [5:119](#)
rigorous diffraction theory [1:52](#), [1:55](#),
[1:58](#)
rigorous electromagnetic wave approach
[5:294–295](#)
rigorous grating analysis [1:52–53](#)
rigorous research (RIGOR) [3:122–123](#)
ring cavity
dye lasers [2:454](#)
fiber lasers [2:455](#)
ring micr. resonator devices [1:143](#), [1:147](#)
ring modulator transmitters
controllers [4:222](#)
fabrication [4:217](#)
laser sources [4:222](#)
link-level considerations [4:222](#)
nonlinear effects [4:220–221](#)
optoelectronic properties [4:217–220](#)
receiver circuits [4:222](#)
ring-based DWDM transceivers [4:221–222](#)
transmitter circuits [4:222](#)
ring resonant cavity [4:219](#)
ring resonator
modulator transmitters [4:257–258](#), [4:257F](#),
[4:258F](#), [4:259F](#)
structure [4:218](#)
ring-based DWDM transceivers [4:221–222](#),
[4:221F](#)
robustness [3:182–183](#)
rds [3:17](#), [5:43](#)
“rogue waves” [2:429](#)
roll-to-roll process (R2R process) [3:56](#), [3:57](#),
[3:80](#)
Ronchi grating [4:31](#), [4:31F](#)
room temperature (RT) [2:441](#)
detectors for room temperature THz
imaging [1:419–422](#), [1:421F](#)

room temperature (RT) (*continued*)

FET and CMOS-based detectors [1:422–423](#),
[1:424F](#)
microbolometers [1:423](#), [1:424F](#)
SBDs [1:421–422](#)

root mean squared difference (RMSD) [4:402](#)

root mean squared error (RMSE) [4:401–402](#)

root-mean square (RMS) [4:531](#)

Rostock cornea module (RCM) [5:99–100](#),
[5:103F](#)

Roswell Park Cancer Institute (RPCI) [3:101](#)

rotating-wave approximation (RWA) [2:12](#)

rotation [4:176](#)

transform [4:188](#)

vibration spectroscopy [1:389–390](#)

rotation anisotropy (RA) [2:207](#)

rotational anisotropy experimental designs
generations [2:209–212](#), [2:210F](#),
[2:211F](#)

rotational spectroscopy [1:387–389](#)

Rothwarf-Taylor equations (RT equations)
[2:133–134](#), [2:134](#)

rototype laser interferometric gravitational
wave detectors [1:433–434](#)

Royal Microscopical Society (RMS) [5:185](#)

Ru-based complex photodyes [5:276–277](#),
[5:276F](#)

rubidium (Rb) [5:282](#)

Rydberg atoms

energy shift of [1:452](#)

micromaser setup [1:453F](#)

Ruby laser operating [2:325–326](#)

rule of thumb [3:272](#)

Runge-Kutta method [2:329](#)

Rydberg atoms [1:451](#), [1:452](#), [2:40](#), [2:44–45](#)
energy shift [1:452](#)

Rydberg excitons in Cu_2O [2:41](#)

Rydberg series [2:33](#)

Rydberg states in semiconductors

coherent features [2:47–49](#)

experimental absorption spectra between
resonances [2:50F](#)

experimental absorption spectra of
Rydberg exciton series [2:48F](#)

incoherent exciton spectral
redistribution function [2:49F](#)

Rydberg excitons model as V-type three
level system [2:49F](#)

mixing of angular momentum states
[2:45–46](#)

quantum defect and non-parabolic bands
[2:46–47](#)

Rydberg blockade [2:41–44](#), [2:43F](#), [2:44F](#)

Rydberg excitons in Cu_2O [2:41](#), [2:42F](#),
[2:43F](#)

symmetry considerations [2:44–45](#)

S

saddle point approximation (SPA) [2:319](#)

sagittal fan [1:118–119](#), [1:120F](#)

Sagnac interferometers [4:41–42](#)

NOIM [4:41–42](#), [4:41F](#), [4:42F](#)

TOAD [4:42](#), [4:43F](#), [4:44F](#)

salts [2:346](#)

sampled-grating DBR (SG-DBR) [4:245](#)

sampling [1:299–300](#)

quotient

off-axis IPRC [4:375–376](#)

off-axis PPRG [4:378](#)

Sandercock's spectrometer [5:118–149](#)

sapphire [1:186](#)

laser [2:19–20](#)

regenerative amplifier lasers [2:166](#)

substrate [1:189](#)

"satin quartz" [4:279](#)

saturable absorbers (SA) [1:369](#), [1:369F](#),
[2:307–308](#), [2:308F](#), [2:468](#)

optical regeneration by [1:369–370](#)

saturable absorption [2:308](#)

crossover resonances [2:230](#)

Doppler free spectroscopy [2:229–230](#)

ECDL [2:227–228](#)

experimental setup [2:230](#), [2:231F](#), [2:232F](#)

principles of saturated absorption
spectroscopy [2:228–230](#)

spectroscopy for diode laser locking [2:227](#)

saturation irradiance (I_{sat}) [4:349](#)

saturation power (P_{sat}) [1:246](#)

SAVIS laser Doppler vibrometer [4:452](#),
[4:452F](#)

scalability [4:225](#)

scalar theory of diffraction [4:157](#)

scalar wave

equation [2:384](#)

model [3:138](#)

scale invariant features transform (SIFT)
[5:78](#)

scaling [2:411](#)

law [1:45–46](#), [2:371–372](#)

up planar lightwave optical switch from
small building blocks [4:234](#)

deterministic [4:234](#)

stages [4:234–235](#)

symmetric [4:234](#)

scanning

process [3:210](#)

systems [5:185](#), [5:189](#)

scanning electron microscopy (SEM) [1:211](#),
[1:212F](#), [1:218](#), [1:218F](#), [1:468](#),
[3:120–121](#), [5:248](#), [5:248F](#), [5:296](#)

scanning laser ophthalmoscope (SLO) [5:64](#),
[5:99](#), [5:163](#), [5:164F](#)

scanning microscopes [5:189](#)

image formation in [5:194–196](#)

scanning near-field optical microscopies
(SNOM) [2:148–149](#)

scanning tunneling microscopy (STM)
[1:468](#), [1:469](#), [2:147–148](#), [5:190](#)

scatterer imaging [3:137](#)

resolution in [3:142–144](#)

scattering [4:67](#), [4:88](#), [5:168](#)

amplitude [3:143](#)

coefficients of tissues [3:220](#)

cross-section [3:220–221](#)

effects [1:255](#)

objects [3:262](#)

phenomena [1:345](#)

photon migration in scattering media
[3:221–222](#)

problems [3:137](#), [3:137F](#)

processes [4:270](#)

TIR [5:171–172](#)

scattering potential (SP) [4:500](#)

scattering-based technique (sSNOM)
[2:148–149](#), [2:149](#)

Schawlow-Townes expression [4:358](#)

Scheiner's disk [5:112](#)

Schell-model sources [1:43–44](#)

schematic eyes [5:48–49](#)

Schottky barrier diodes (SBDs) [1:419](#),
[1:421–422](#), [1:422F](#)

Schottky emission [5:233](#), [5:233F](#)

Schottky type photodiodes [5:321–322](#)

Schrödinger equation [1:169–170](#), [1:458](#),
[2:12–13](#), [2:64–65](#), [2:82](#), [2:99](#), [2:100](#),
[2:103–104](#), [4:341](#), [5:30–31](#), [5:203](#),
[5:205](#)

sclera [5:130](#)

scleral lens prosthetic device (SLPD) [5:119](#)

scleral lenses [5:119](#)

sealed-off

CuBr systems [2:370](#)

and waveguide lasers [2:331](#), [2:331F](#), [2:332F](#)

second generation of AMVA (AMVA II) [3:61](#)

second harmonic imaging of cornea
[5:104–105](#)

second harmonics (SH) [1:412](#), [2:294](#),
[2:313–314](#)

2nd order nonlinear optical effects [1:491](#)

2nd order perturbation theory [1:488](#)

second ultrafast laser pulse [5:39–40](#)

second-harmonic generation (SHG) [1:10](#),
[1:11](#), [1:406](#), [2:121](#), [2:124](#), [2:145–146](#),
[2:146](#), [2:245](#), [2:280](#), [2:291](#), [2:316](#),
[2:398](#), [2:424](#), [2:446](#), [4:49](#), [4:284](#),
[4:285F](#), [5:89](#)

absolute measure [4:287–289](#)

for characterization of 2D materials

[4:284–285](#)

dependence on stacking and number of
layers [4:286–287](#)

polarization dependent measurements

[4:285](#), [4:285F](#)

in crystals [1:12–16](#), [1:13F](#), [2:207–209](#)

QPM [1:14–16](#), [1:14F](#), [1:15F](#)

electronic structure calculations [4:289–290](#)

imaging [5:74–75](#), [4:285–286](#), [4:286F](#), [5:98](#)
of corneal collagen fibrils [5:104–105](#),
[5:105F](#)

of human cornea [5:105F](#)

optical devices [4:290](#)

spectroscopy of hidden phases

generations of rotational anisotropy
experimental designs [2:209–212](#)

hidden phases in correlated electron
systems [2:212–214](#)

hidden phases revealed by SHG-based
techniques [2:214–216](#)

second-order

correlations [1:48–49](#)

DFB [1:384](#)

exponentials of second-order differential
operators [4:202](#)

nonlinear polarization [4:349](#)

spectroscopy [2:245–246](#)

- up-conversion from second-order optical nonlinearity [2:397–399](#)
 - second-order dispersion *see* dispersion slope
 - secondary fiber coating [1:328](#)
 - fiber identification [1:330](#)
 - loose jacket [1:328–330](#), [1:330F](#)
 - tight jacket [1:328](#), [1:329F](#)
 - secondary sources [1:43](#)
 - secondary wavelet [1:62](#)
 - seed generation [2:295](#)
 - Seidel aberration theory [4:205](#)
 - selective area growth (SAG) [4:243–244](#)
 - selective laser trabeculoplasty (SLT) [5:136](#)
 - self Q switching [2:417](#)
 - self-clutter signal [5:2](#)
 - self-focusing [5:92](#)
 - self-frequency doubling laser [2:404](#), [2:404F](#), [2:405F](#)
 - self-imaging
 - condition [3:167–168](#), [3:168F](#)
 - principle [3:168](#)
 - self-induced Kerr effect [1:348](#)
 - self-phase modulation (SPM) [1:10](#), [1:16](#), [1:16–18](#), [1:180](#), [1:245](#), [1:320](#), [1:321](#), [1:348](#), [2:311](#), [2:425](#), [4:3](#), [4:23](#)
 - spectral broadening by [1:19](#), [1:20F](#)
 - self-phase modulation [1:142](#), [1:348](#), [2:2–3](#), [4:278](#), [4:312–313](#), [4:320](#), [4:350–351](#), [4:351](#)
 - self-phase-conjugation (SPC) [4:321](#)
 - self-pumped phase conjugate mirrors [4:332](#)
 - self-similarity [5:27–28](#)
 - self-sum frequency mixing laser [2:404–405](#), [2:405F](#), [2:405F](#)
 - self-supporting aerial cable [1:333](#), [1:333F](#)
 - self-terminating
 - pulsed metal vapor laser [2:368–369](#)
 - resonance-metastable pulsed metal vapor lasers [2:368–369](#)
 - self-trapped excitons (STEs) [4:304–305](#)
 - Sellmeier expansion of refractive index [1:6](#)
 - semi-analytical methods [1:172–173](#)
 - semi-classical
 - model [2:318](#)
 - three-step model [2:316–317](#)
 - semi-polar
 - InGaIn layers [2:279](#)
 - planes [2:271](#)
 - semiconducting polymer lasers [5:310](#), [5:316](#), [5:323–324](#)
 - semiconductor disk laser (SDL) [2:413–414](#)
 - semiconductor laser amplifier (SLA) [4:38](#), [4:39](#), [4:40–41](#)
 - semiconductor optical amplifiers (SOA)
 - [1:242](#), [1:242F](#), [1:248–250](#), [1:263](#), [1:264](#), [1:365](#), [1:367–368](#), [1:368F](#), [4:64](#), [4:239](#), [4:250](#)
 - all-optical data regeneration [1:251–252](#)
 - bulk SOA [1:243–244](#), [1:244F](#)
 - functional applications [1:250–251](#)
 - IWM wavelength converter [1:250–251](#), [1:250F](#)
 - gain mechanism in [1:242](#)
 - gain-clamping [1:249–250](#)
 - microwave phase shifter [1:252–253](#), [1:252F](#)
 - models [1:245–247](#)
 - pulse amplification [1:248](#)
 - steady-state [1:245–247](#)
 - time-domain model and pattern effects [1:247–248](#)
 - MZI XOR
 - gate [1:251](#)
 - logic and truth table [1:251F](#)
 - optical logic [1:251](#)
 - phase shifter transfer function [1:252–253](#)
 - principles [1:242–243](#)
 - reflective SOA based passive optical network [1:252](#)
 - semiconductor nanostructure SOAs
 - [1:244–245](#)
 - structures [1:243–244](#)
 - switch [4:239–240](#)
 - semiconductor quantum structures, few cycle
 - THz spectroscopy of [2:19–21](#)
 - semiconductor saturable absorbing mirror (SESAM) [2:281](#), [2:412](#), [2:421](#)
 - Q switching with [2:421](#)
 - semiconductor-Bloch equations (SBEs)
 - [1:459](#), [2:34](#), [2:83](#), [4:265–266](#)
 - semiconductor(s) [1:224](#), [1:460](#), [2:33](#), [2:63](#), [2:83](#), [2:88–89](#), [2:98–99](#), [2:413–414](#)
 - amplifiers [1:351](#)
 - band gap [1:403](#)
 - bolometers [1:426–427](#)
 - cavity QED in
 - Heisenberg uncertainty principle [1:463–465](#)
 - quantized states [1:459–461](#)
 - superposition principle [1:461–463](#)
 - in coherence control [1:487](#), [1:487–490](#)
 - of electrical current [1:488–490](#), [1:489F](#), [1:491](#)
 - of carrier density, spin population, and spin current [1:490–491](#), [1:490F](#)
 - in semiconductors [1:487–490](#)
 - diode edge emitting lasers [2:356](#)
 - edge emitting diode laser [2:350](#)
 - excitation by co-circularly polarized ω and 2ω pulses [1:490F](#)
 - lasers [2:462](#), [2:462F](#)
 - applications [2:468](#)
 - dynamics [2:467–68](#)
 - gain medium physics [2:464–465](#)
 - resonator [2:465–467](#)
 - technology [2:336](#)
 - layers [5:273–275](#)
 - luminescence equations [1:461](#)
 - to metals [2:256–257](#), [2:256F](#)
 - III-V semiconductors [2:257](#)
 - TCOs [2:257–258](#)
 - microcavities [1:456](#)
 - nanostructures [2:19](#)
 - SOAs [1:244–245](#)
 - optical properties [2:99](#)
 - quantum wells [2:120](#)
 - optical-pump/THz-probe magnetospectroscopy [2:76–77](#)
 - optically detected FIR resonance spectroscopy [2:75–76](#)
- THz-driven AC stark effect of intersubband transitions in [2:200–203](#)
- strong-field terahertz excitations in crystalline semiconductor [2:33–34](#)
 - dynamical Bloch oscillations [2:36–37](#), [2:37F](#)
 - e-h recollisions [2:37–38](#), [2:38F](#)
 - excitations in semiconductor [2:34F](#)
 - HHG in two vs. three bands [2:35](#), [2:35F](#)
 - non-perturbative quantum interference [2:35–36](#), [2:35F](#)
 - resonant vs. non-resonant excitations [2:33F](#)
 - SREs [2:34](#)
 - time-resolved HHG [2:36](#), [2:36F](#), [2:37F](#)
- systems [1:185](#), [1:462](#)
- THz laser [2:119–120](#)
- ultrafast studies
 - coherent dynamics [2:82–83](#)
 - hot carrier regime [2:84–85](#)
 - hot phonons [2:85](#)
 - incoherent relaxation dynamics [2:83–84](#)
 - isothermal regime [2:85](#)
 - nonthermal regime [2:84](#)
 - novel coherent phenomena [2:86](#)
 - tunneling and transport dynamics [2:85–86](#)
- vacuum [1:467](#)
- semitransparent object [5:187](#)
- Sénarmont compensator [1:156–157](#)
- senile miosis [5:60](#)
- sensing at exceptional points [4:298–300](#)
- sensitivity [1:273](#), [3:235–237](#), [4:87–88](#), [5:159–160](#)
- sensitizer [2:394](#), [5:275–277](#), [5:276F](#), [5:276F](#)
- sensors [3:145](#), [3:239](#), [4:106](#)
- separate absorption, charge, and multiplication layer stack (SACM layer stack) [4:417](#)
- separate confinement heterostructure (SCH) [1:243](#), [2:351](#)
- sequential holograms [4:106](#)
- sequential two-photon absorption
 - [2:395–396](#), [2:396F](#), [2:400–402](#), [2:401F](#), [2:402F](#), [2:402F](#)
- series resistance (R_s) [5:259](#)
- Shack-Hartman wavefront sensors
 - [3:277–278](#), [5:72](#), [5:72–73](#), [5:113](#)
- Shack-Hartmann sensor [5:112–113](#), [5:113F](#)
- Shack-Hartmann wavefront sensing
 - [5:124–125](#)
- shadow mask [3:3](#), [3:3F](#)
- shadowgram [3:251](#)
- shallow ridge waveguide [4:247](#)
- shallow tissue imaging through selection of ballistic photons [3:222–223](#)
- shallow waveguide etch [4:217](#)
- shape factor [1:9](#), [1:74](#)
 - of lens [1:114–116](#)
- shearing interferometer wavefront sensors [3:277–278](#)
- sheaths [1:331–332](#), [1:331F](#)
- Shepp-Logan filter [3:194](#)
- shift variance [3:263](#)
- Shockley-Queisser theory [5:288–290](#)

- Shockley-Ramo theorem [4:422](#)
- Shor's factor-finding algorithm [1:484](#)
- short circuit current density [5:258](#)
- short pass filters (SPF) [2:211](#)
- short pulse generation [2:467–468](#)
- short trajectories [2:316–317](#)
- short wavelength infrared (SWIR) [3:229](#)
- short wire [1:244](#)
- short-circuit current density (J_{sc}) [5:270](#)
- short-exposure star images [3:274–275](#)
- short-haul transmitters [4:68](#)
- short-period FBCs [1:267](#)
- short-range WindScanner (SRWS) [4:462](#)
- short-term fluctuations [1:32](#)
- short-time-duration laser pulse [5:39](#)
- short-wavelength infrared detectors (SWIR detectors) [3:232](#), [4:415–416](#)
- short-wavelength-cone (S-cone) [3:17](#), [3:18F](#)
- shot-noise limit
- off-axis IPRG [4:377](#)
 - off-axis PPRG [4:379](#)
 - on-axis PSRG [4:380–381](#)
- shrinkage of emulsion [4:90](#)
- shutter glasses [3:46](#)
- side band technique [4:151](#)
- side mode suppression ratios (SMSR) [4:249](#)
- side pumping [2:409–410](#)
- side-looking airborne radar (SAR) [4:502](#)
- side-view technique/transverse interference [1:323](#)
- sidehands (SB) [2:321–322](#)
- signal [1:223–224](#)
- coding [3:175](#), [3:178–179](#)
 - distortion [1:249–250](#)
 - fidelity [1:223–224](#)
 - interferes [4:369](#)
 - plane wave [1:374](#)
 - reflections [1:356](#)
 - regenerator [1:365](#), [1:365F](#)
- signal carrier phase (SC phase) [1:296–297](#)
- signal processing techniques for range profiling and tomography [4:489–494](#)
- signal recycling mirror (SRM) [1:436–437](#)
- signal-ASE beating [1:353](#)
- signal-to-noise ratio (SNR) [1:242–243](#), [1:273](#), [1:298–299](#), [1:353](#), [3:87](#), [3:274–275](#), [4:5](#), [4:34](#), [4:369](#), [4:376–377](#), [4:415](#), [4:430–431](#), [4:475](#), [5:3](#), [5:64](#), [5:146](#), [5:320–321](#)
- off-axis IPRG [4:376–377](#)
 - off-axis PPRG [4:379](#)
 - on-axis PSRG [4:380](#)
- silica (SiO_2) [1:27](#), [1:195](#), [2:381](#)
- attenuation in silica single-mode fibers [1:255](#)
 - cladding [1:229](#)
 - fibers [1:257](#)
 - fiber attenuation [4:1](#), [4:1F](#)
 - glass fibers [2:268](#)
 - optical fibers [1:217–218](#)
 - PCF [1:217–218](#), [2:433–434](#)
 - silica-on-silicon [4:235](#)
 - silica-polymer-air structure forms [5:313–314](#)
- silicate [1:186](#), [2:347](#), [4:297](#)
- silicon (Si) [1:381](#), [2:267](#), [2:350](#), [5:292–293](#)
- impurity state lasers [1:381](#)
 - liquid-crystal on silicon wavelength-selective switch [4:231](#)
 - Raman lasers [2:267](#)
 - silicon carbide [2:271](#)
 - silicon oxynitride (SiON) [4:235](#)
 - silicon photonics (SiPh) [4:236](#)
 - controllers [4:222](#)
 - fabrication [4:217](#)
 - laser sources [4:222](#)
 - link-level considerations [4:222](#)
 - nonlinear effects [4:220–221](#)
 - optoelectronic properties [4:217–220](#)
 - receiver circuits [4:222](#)
 - ring-based DWDM transceivers [4:221–222](#)
 - transmitter circuits [4:222](#)
 - silicon qubits
 - future [1:475](#)
 - memory, control, and readout mechanisms [1:471–472](#)
 - motivation for [1:467–468](#)
 - principal material types [1:468–471](#), [1:468F](#), [1:470–471F](#)
 - qubit couplings and hybridizations [1:473–474](#)
 - silicon tetrachloride (SiCl_4) [1:28](#)
 - silicon-germanium (SiGe) [4:254](#)
 - silicon-on-insulator (SOI) [2:267](#), [3:14](#), [4:217](#)
 - silicon-perovskite tandem solar cell [5:290](#)
 - silicone photodiode [3:88](#)
 - silicone photomultiplier (SiPM) [3:88–89](#)
 - silver (Ag) [2:256](#), [4:91](#), [5:298](#)
 - silver halide [4:89–90](#), [4:89F](#)
 - crystals [4:93](#)
 - emulsion [4:90](#)
 - photographic emulsions [4:151](#)
 - simple dimensional analysis [3:160–161](#)
 - simple mathematical description of holography [4:152–153](#)
 - simple paraxial theory [5:195–196](#)
 - simple point-to-point WDM system [4:7–8](#), [4:9F](#)
 - simple two-band model [4:320](#)
 - simplified FROG-CRENQUILLE [4:54](#), [4:55F](#)
 - simulated and measured data [4:477](#), [4:478F](#)
 - simulating laser profiling data [4:476–477](#)
 - simulations [5:301](#)
 - sinc function [1:75](#), [4:385](#)
 - single A-scans [5:143](#)
 - single active electron (SAE) [2:318](#)
 - single beam technique [4:102–103](#)
 - single color beams, coherence control of electrical current using [1:491](#)
 - single configuration coordinate model (SCC model) [2:437–440](#), [2:438F](#)
 - single frequency gas laser [2:454](#)
 - single isotropic phase singularity [4:185](#)
 - single longitudinal mode lasing [4:294](#)
 - single mode connector attenuation [1:336–337](#), [1:337F](#)
 - single mode laser [1:223](#)
 - single mode operation
 - with eliminated spatial hole burning [2:452–453](#)
 - with mode suppression [2:453–454](#)
 - single molecules, TIRF on [5:172–173](#)
 - single phosphorus impurities [1:468](#)
 - single photon emission computerized tomography (SPECT) [3:124](#), [3:192](#), [3:196](#)
 - single polarization transmission (SP transmission) [1:293–294](#)
 - single polymer layer device [5:259](#)
 - single preform [1:28](#)
 - single prisms as reflectors [1:124–126](#)
 - single quantum cascade [2:27–28](#)
 - single spatial mode operation [4:298](#)
 - single telecommunications-grade optical fiber [1:355](#)
 - single-atom laser [1:455](#)
 - single-color method [1:411–412](#), [1:411F](#), [1:412F](#)
 - single-frequency [2:411](#)
 - DBR lasers [2:456](#), [2:457F](#)
 - dye lasers [2:454](#)
 - emission [4:16](#)
 - lasers [2:451](#), [2:451–452](#), [2:451F](#), [2:454–455](#)
 - applications [2:461](#)
 - characterization [2:458](#)
 - DBR lasers [2:456–457](#)
 - DFB lasers [2:455–456](#)
 - external cavity lasers [2:457–458](#)
 - NPRO [2:454–455](#)
 - phase noise [2:458–459](#)
 - RIN [2:458](#)
 - ring cavity fiber lasers [2:455](#)
 - single mode operation with eliminated spatial hole burning [2:452–453](#)
 - single mode operation with mode suppression [2:453–454](#)
 - spatial hole burning and multimode operation [2:451–452](#)
 - spectral linewidth [2:459–160](#)
 - operation [2:417](#)
 - ring-cavity fiber laser [2:455](#), [2:455F](#)

- singlet oxygen [3:102](#)
 singlet-singlet annihilation [5:310–311](#)
 singular phase function, gradient of [4:187–188](#)
 singular value decomposition (SVD) [3:157](#)
 sinusoidal structured pattern [4:453](#)
 sinusoidal surface-relief grating [1:51F](#), [1:52](#)
 Sjogren's syndrome [5:116–117](#)
 skew rays [3:35](#)
 skew-symmetric operation [4:199](#)
 skin lesions, treatment of [2:372](#)
 skin rejuvenation [3:118–119](#), [3:118F](#)
 slab coupled waveguide lasers (SCOWL) [4:293](#)
 slit scanning microscopy [5:99–100](#)
 slope tuning [4:31](#)
 slotted core cables [1:332–333](#), [1:332F](#), [1:333F](#)
 slow axial flow lasers [2:331–332](#), [2:332F](#)
 slow corrections [3:198](#)
 small area contrast [3:70](#)
 small footprint components soldered on conventional PCB [3:89](#)
 small incision lenticule extraction technique (SMILE technique) [5:126](#), [5:132](#), [5:133](#), [5:135](#), [5:138](#)
 small molecules
 organic photodiodes with [5:321–322](#)
 organic phototransistors with [5:325–326](#)
 and polymer OLEDs [5:232](#)
 small-signal modulation [4:69](#)
 smectic liquid crystals [3:9–10](#)
 smoke detectors [4:513–514](#), [4:514F](#)
 “smooth picture” technique [3:37](#)
 snake-like regime, imaging in [3:223–225](#)
 Snell's law [1:102–104](#), [1:114–116](#), [1:124–125](#)
 (SnO₂:F) *see* fluorine-doped tin oxide (FTO)
 Society of Automotive Engineers (SAE) [3:70](#)
 soft contact lenses [5:118–119](#), [5:119](#)
 soft glass fibers [2:433](#)
 soft lithography techniques [1:144–146](#)
 software [3:216–217](#), [3:217F](#)
 sol-gel
 process [2:346](#)
 slurry [1:196](#)
 techniques [2:347](#)
 solar cells [2:71](#), [5:270](#)
 solar-to-electricity conversion efficiency [5:270](#)
 solid phase crystallization (SPC) [3:14](#)
 solid-core fibers [1:217–218](#), [1:263](#)
 solid-state
 devices [2:265](#)
 DSSC [5:278–279](#)
 lasers [2:384](#), [2:407](#)
 materials [2:168](#)
 charge dynamics of dye-sensitized solar cells [2:168](#)
 nanoscale materials [2:168–170](#)
 microchip lasers [2:415](#)
 systems [2:194–196](#)
 solid-state dye laser [2:336](#), [2:337](#), [2:348–349](#)
 see also liquid dye lasers
 DFB solid-state dye lasers [2:344](#)
 matrices [2:346–347](#)
 inorganic and hybrid hosts [2:347–348](#)
 organic hosts [2:346–347](#)
 oscillators [2:340–342](#)
 cavity linewidth equation [2:344](#)
 multiple-prism dispersion grating theory [2:341–342](#), [2:342F](#)
 physics and architecture of [2:342–344](#), [2:343F](#)
 soliton(s) [2:427–428](#), [4:17](#), [5:223](#), [5:224F](#)
 operation [2:304–305](#)
 self-frequency shift [2:432–433](#)
 solution fabrication processes [5:220](#)
 solution-based deposition methods [3:15](#)
 solution-processing
 approaches [5:220](#)
 methods [5:313](#)
 solvent additive [5:262–263](#)
 solvent annealing (SA) [5:261](#), [5:261–262](#)
 solvent flipping (SF) [3:148](#)
 Sommerfeld enhancement factor [2:93](#)
 Sommerfeld Green's function [1:63](#), [1:64F](#), [1:66](#)
 Sommerfeld radiation condition [1:65](#)
 sonogram [3:194](#), [3:194F](#), [4:52](#), [4:52F](#)
 Sony and Konica/Minolta's holographic wave guide combiners [4:125](#)
 soot [1:28](#)
 sophisticated experimental techniques [2:36](#)
 sophisticated software tools (Simulase) [2:282–283](#)
 SOPRA CES5 [1:210](#)
 sound wave [4:98](#), [4:98F](#)
 source development lab (SDL) [1:405](#)
 source-field imaging [3:137](#)
 resolution in
 propagating and evanescent spectra of radiated field [3:140–142](#)
 scalar wave model [3:138](#)
 three-dimensional resolution of far fields [3:138–140](#)
 transverse resolution [3:142](#)
 Southeast Asia (SEA) [4:18](#)
 sp² hybrid orbitals [5:221](#), [5:222F](#)
 space bandwidth-product (SBP) [4:191](#)
 space coherence [1:35–36](#)
 space division multiplexing (SDM) [4:34](#), [4:45](#)
 space switches [4:34](#)
 space-bandwidth product (SBP) [3:168](#), [4:97](#)
 space-based far-infrared astronomy array
 technology (Si-As BIB array technology) [1:426](#)
 space-charge layer [5:233](#)
 space-charge-limited current (SCLC) [5:233](#)
 space-invariant (isoplanatic) imaging [1:168](#)
 spacer layer [5:302–304](#), [5:304F](#)
 spatial coherence [1:35–36](#), [1:35F](#), [1:37–38](#), [4:151](#)
 spatial color synthesis [3:19–20](#)
 spatial distribution [3:242](#)
 spatial filter (SF) [4:500](#)
 spatial Fourier domain [3:137](#)
 spatial frequency content of 3D PSF [3:263–264](#), [3:264F](#), [3:265F](#)
 spatial frequency response (SFR) [3:177–178](#), [3:178F](#), [3:181F](#), [3:183F](#)
 spatial gating [2:321](#)
 spatial heterodyne [4:369](#), [4:407](#)
 background material [4:369–371](#), [4:372–373](#)
 hologram irradiance [4:372–373](#)
 hologram mean number of photoelectrons [4:373](#)
 hologram photoelectron density [4:373–374](#)
 image plane [4:372](#)
 object plane [4:369–371](#)
 pupil plane [4:371–372](#)
 comparison of different recording geometries [4:381](#)
 off-axis IPRG [4:381](#), [4:382F](#)
 off-axis PPRG [4:381–385](#), [4:383F](#)
 on-axis PSRG [4:384F](#), [4:385](#)
 MATLAB® code to exploring off-axis IPRG, off-axis PPRG and on-axis PSRG [4:386–395](#)
 off-axis IPRG [4:374–375](#)
 off-axis PPRG [4:377–378](#)
 on-axis PSRG [4:379–380](#)
 recording geometries [4:370F](#)
 rectangle, comb, impulse, cylinder, and sinc functions [4:385](#)
 2-D convolution [4:386](#)
 spatial hole burning [2:340](#), [2:411](#), [2:451–452](#)
 spatial light modulator (SLM) [2:188](#), [3:42–43](#), [4:99](#), [4:113](#), [4:114](#), [4:115](#), [4:178](#), [4:179F](#), [4:191](#), [4:197–198](#), [5:91](#)
 LiCoS [4:99](#)
 MEMSs [4:99–100](#)
 technology [4:139](#)
 spatial modulation methods [5:41](#)
 spatial resolution [3:214](#), [3:216](#), [4:88](#)
 spatial variations [4:456](#)
 spatial-multiplexed display [3:48](#)
 spatially coherent light [3:249](#), [3:249F](#)
 spatially incoherent light [3:249–250](#), [3:249F](#)
 spatially incoherent objects, synthetic apertures with [3:250–252](#)
 spatially resolved spectroscopy [2:249–250](#)
 spatially varying complex amplitude transmittance [4:153](#)
 Spatio-Temporally Resolved Optical Laser-Induced Damage test (STEREO-LID test) [4:307F](#), [4:308](#)
 special cables [1:334](#)
 special encoding techniques [4:191](#)
 specific Doppler-free configuration [4:344](#)
 specimen-scanning method [5:194](#)
 speckle contrast tomography [4:526](#)
 speckle decorrelation [5:64](#), [5:64–65](#)
 “speckle effects” [4:109–110](#)
 speckle imaging techniques [3:274–276](#)
 speckle interferometry *see* Labeyrie's method
 speckle measurements [4:408](#), [4:408F](#)
 spectacles [5:116–119](#), [5:117F](#)
 contact lenses [5:116–119](#), [5:118F](#), [5:119F](#)
 spectacle-type of wearable display [3:51–52](#)
 spectral and angular Bragg selectivity [4:124](#)
 Spectral and Photometric Imaging Receiver (SPIRE) [1:427](#), [1:427F](#)

- spectral broadening [1:25](#), [1:181](#), [1:321](#), [2:424–425](#)
 - by SPM [1:19](#), [1:20F](#)
- spectral correlation [4:468–469](#)
- spectral decomposition [3:213](#)
- spectral density [4:179](#)
- spectral diffusion [2:160](#), [2:160F](#), [2:165](#), [2:176](#)
- spectral dispersion [4:92](#)
- spectral efficiency (SE) [1:299](#)
- spectral fluctuations [1:283](#)
- spectral hole burning (SHB) [1:250–251](#), [2:468](#)
- spectral interference law [1:41–42](#)
- spectral interferogram [4:51](#), [4:51F](#), [4:52](#)
- spectral interferometry (SI) [2:82](#), [2:316](#), [4:48](#), [4:50](#), [4:51](#)
 - spectral interferometry–relative phase measurements [4:50–51](#)
- Spectral Karyotyping system (SKY system) [3:213](#)
- spectral linewidth [2:459–460](#)
 - delayed self-heterodyne detection [2:460–461](#)
 - heterodyne detection [2:460](#)
- spectral oscillations [4:271–272](#), [4:274F](#)
- spectral phase interferometry for direct electric field reconstruction (SPIDER) [2:316](#), [2:317F](#), [4:48](#), [4:57–58](#), [4:58F](#)
- spectral properties
 - frequency tuning [2:417–418](#)
 - fundamental linewidth [2:417](#)
 - single-frequency operation [2:417](#)
- spectral resolution [3:214–215](#), [3:216](#)
- spectral selection [2:237](#)
- spectral sensitivity [4:88](#)
- spectral shear [2:316](#)
- spectral-domain optical low-coherence tomography (SD-OCT) [4:504–505](#), [4:505F](#), [5:156](#), [5:157F](#)
- spectral-domain Zeiss Cirrus instrument [4:506](#)
- spectrogram [4:52](#), [4:53F](#), [4:54](#)
- spectroscopic/spectroscopy [1:51](#), [1:139](#), [1:185](#), [2:244](#), [5:185](#)
 - gratings [1:55–57](#), [1:56F](#)
 - harmonic conversion [2:246](#)
 - higher-order and multidimensional [2:249](#)
 - information [2:244–245](#)
 - microscopies [5:193](#)
 - nonlinear optical spectroscopy in
 - chemistry [2:245–246](#)
 - nonlinear optics for [2:244–245](#)
 - second-order [2:245–246](#)
 - spatially resolved [2:249–250](#)
 - surface second-order [2:246–247](#)
 - third-order [2:247–248](#)
 - ultrafast time resolved [2:248–249](#)
- spherical aberration [1:114–116](#), [1:115F](#), [1:117F](#)
- spherical Bessel function [3:160](#)
- spherical emitter to spherical receiver, diffraction-propagation from [4:157](#)
- spherical refractive anomalies [5:51–52](#)
- spherical wavelet [1:62](#)
- sphero-cylindrical
 - correcting lens [5:115](#)
- specification of sphero-cylindrical lenses [5:109–110](#)
- SPIDER [4:52](#), [4:58F](#)
- spin [1:475](#), [1:487](#)
 - control
 - by ESR and EDSR [1:472](#)
 - by kinetic exchange [1:472–473](#)
 - current
 - coherence control [1:490–491](#)
 - using two color beams [1:490–491](#), [1:490F](#)
 - population [1:490–491](#), [1:490F](#)
 - coherence control [1:490–491](#)
 - qubits [1:472–473](#)
 - coupling with exchange [1:474](#)
 - spin-lattice relaxation dynamics in HoMnO₃ [2:138](#)
 - spin-orbit interaction [1:490](#)
- spin density wave (SDW) [2:134](#)
- Spitzer-Space Telescope [1:425](#)
- splice [1:336](#)
- split-off band [2:115](#)
- splitters [1:339–340](#)
- spontaneous decay rate [1:451](#), [1:455](#)
- spontaneous emission (SE) [1:451](#), [1:456](#), [2:18](#), [2:79](#), [2:105–106](#), [2:108](#), [2:112–113](#), [2:329](#)
- spontaneous parametric down-conversion (SPDC) [1:479](#)
- spontaneous radiation rate [1:451](#)
- spontaneous symmetry breaking [4:291–293](#)
- spontaneous transition rate in confined space, modification of [1:451](#)
- sputtering [1:210–211](#), [2:258](#)
- square wave modulating envelope [5:204–205](#)
- square-law detector [4:347](#)
- squeezing [1:464](#), [1:465](#)
- Sr₂IrO₄, odd-parity hidden order in [2:217–220](#), [2:218F](#), [2:219F](#)
- stability [4:88](#)
- stacking [1:196](#), [4:286–287](#), [4:287F](#)
- stainless steel (StSt) [5:273](#)
- standard deviation (SD) [3:124](#), [4:429](#)
- standard Fourier optics techniques [3:278](#)
- standard Fourier transform [4:157](#)
- standard laser amplification techniques [2:324](#)
- standard perturbation–theory techniques [1:173](#)
- standard planar hybrid architecture [1:425](#)
- standard rate equation model [4:327–329](#), [4:328F](#)
- standard reference material (SRM) [1:316](#)
- Standard Single-Mode Fibers (SSMF) [1:255](#), [1:261](#)
- standing light wave [5:214](#), [5:216F](#)
- standing waves [2:453](#)
 - interaction [5:213](#)
 - of light [5:207–213](#)
 - OCE [5:147](#)
- state-of-the-art
 - holographic recording technique [1:377–378](#)
 - semiconductor manufacturing process [2:446](#)
- static color breakup [3:21](#), [3:21F](#)
- static contrast ratio [3:72](#)
- static links to optical switching [4:224](#)
- static OCE [5:146–147](#)
- static random access memory (SRAM) [3:84](#)
- static technique [3:95](#)
- static wave [4:98](#)
- statistical CRB [3:137](#)
- steady-state model [1:245–247](#)
- steel [1:330](#)
- Stellar intensity interferometry [1:49](#)
- Stellar interferometry principles [1:47](#), [3:253–254](#), [3:254F](#)
- stencil-FSC methods [3:22–23](#), [3:22F](#), [3:23F](#)
- step index fiber [1:33](#)
- step index optical fiber [1:222](#)
- stereo effect [3:208](#)
- stereopsis [5:44](#)
- stereoscopic display [3:45](#)
 - anaglyph [3:45–46](#)
 - E-holography [3:46](#)
 - head mounted display [3:46](#)
 - integral imaging display [3:48](#)
 - LC lens for 2D/3D switchable display [3:48](#)
 - multi-planar [3:46–47](#)
 - polarizing glass [3:45](#)
 - shutter glasses [3:46](#)
 - SMV display [3:48](#)
 - spatial-multiplexed [3:48](#)
 - time-multiplexed [3:47–48](#)
 - volumetric display [3:46](#)
- Stevens-Johnson syndrome [5:116–117](#)
- stilbenes [2:345](#)
- Stiles-Crawford effect [5:50](#), [5:51F](#), [5:124–125](#)
- stimulated Brillouin scattering (SBS) [1:320](#), [1:321](#), [4:3](#), [4:322](#), [4:322F](#)
- stimulated emission [1:97](#), [2:112–113](#), [2:325–326](#), [2:328–329](#)
- stimulated Raman adiabatic passage (STRAP) [5:32–33](#)
- stimulated Raman scattering (SRS) [1:258](#), [1:320](#), [1:321](#), [1:357](#), [4:3](#), [4:34](#)
- stoichiometric lithium tantalite (SLT) [2:294](#)
- Stokes parameter evaluation (SPE) [1:316](#), [1:318](#), [1:318F](#)
- Stokes parameters [1:163–166](#)
- Stokes representation [1:167](#), [1:168T](#)
- Stokes vectors [1:160](#)
- Stokes wave [2:265](#)
- Stokes-Einstein picture [2:171–172](#)
- storage capacitor (Cs) [3:31](#)
- strain
 - amplitudes [1:130–131](#)
 - tuning [1:237](#)
- stranded structures [1:332](#)
- Stranski-Krastanov growth technique (SK growth technique) [1:245](#)
- Straubel functions [5:16](#)
- straylight [5:50](#)
- streak cameras [4:431–432](#)
- “streaming motion” [1:380](#)
- strength member [1:330–331](#), [1:331F](#)
- stress
 - birefringence [1:160](#)
 - corrosion susceptibility [1:326](#)

- stripline devices [1:143](#)
 stroke, PBM for [3:121–123](#)
 stroke therapy academic industry
 roundtable (STAIR) [3:122–123](#)
 strong field approximation (SFA) [2:318](#)
 strong light–matter coupling [1:448–450](#)
 strong-field terahertz excitations in
 semiconductors
 crystalline semiconductor [2:33–34](#)
 dynamical Bloch oscillations [2:36–37](#),
 [2:37F](#)
 e–h recollisions [2:37–38](#), [2:38F](#)
 excitations in semiconductor [2:34F](#)
 HHG in two vs. three bands [2:35](#), [2:35F](#)
 non-perturbative quantum interference
 [2:35–36](#), [2:35F](#)
 resonant vs. non-resonant excitations [2:33F](#)
 SBEs [2:34](#)
 time-resolved HHG [2:36](#), [2:36F](#), [2:37F](#)
 strontium ion laser [2:372–373](#), [2:373](#),
 [2:373F](#)
 strontium-barium niobate ($\text{Sr}_x\text{Ba}_{1-x}$
 Nb_2O_6) [4:96](#)
 structural in vivo multiphoton
 ophthalmoscopy [5:86–87](#)
 structured illumination microscopy (SIM)
 [5:168–169](#)
 sub-apertures [4:407–408](#), [4:408F](#)
 sub-band energy [2:101](#), [2:101–102](#)
 sub-carrier multiplexing (SCM) [1:255](#)
 sub-millimeter wave lasers *see* terahertz
 lasers
 sub-picosecond pulses [2:468](#)
 sub-Poisson statistics [1:452](#)
 sub-Poissonian atomic statistics [1:453](#)
 subcritical illumination [5:170](#)
 submarine cables [1:334](#), [1:334F](#)
 Submillimetre Common-User Bolometer
 Array (SCUBA) [1:430](#)
 SCUBA-2 array [1:430](#)
 subpixel divided method [3:62](#)
 subsea holography [4:109–110](#)
 subsonic advance cruise missile AGM 129A
 [4:450](#), [4:450F](#)
 substantia nigra pars compacta (SNc) [3:125](#)
 substitutional doping [2:257](#)
 substrate(s) [2:464](#)
 mode [3:67](#), [5:247](#)
 light extraction [5:248](#)
 role [4:289](#)
 successive pulse illumination in acoustical
 studies [4:528](#), [4:529F](#)
 sudden perturbation process [2:377–378](#)
 sum frequency generation (SFG) [1:406](#),
 [2:208](#), [2:247](#), [2:291](#), [2:314](#)
 sum-frequency mixing (SFM) [2:398](#),
 [2:446–447](#)
 super AMOLED products [3:67](#)
 super twisted nematic liquid crystal displays
 (STN LCDs) [3:2](#)
 super-heterodyne receiver [1:373](#)
 super-heterodyning [1:374](#)
 super-multi-view display (SMV display) [3:48](#)
 super-resolving method [3:190](#)
 superconducting detectors [1:427–428](#)
 superconducting HEB detectors [1:428–430](#)
 superconducting quantum interference
 devices (SQUIDs) [1:427–428](#), [1:428](#),
 [1:484](#)
 superconducting resonator [1:475](#)
 superconducting TES detectors [1:428–430](#)
 superconductivity (SC) [2:123](#)
 superconductor-insulator-normal metal
 (SIN) [1:427–428](#)
 superconductor-insulator-superconductor
 (SIS) [1:421–422](#), [1:427–428](#), [1:428](#),
 [1:428F](#)
 supercontinuum generation (SCG) [1:24–25](#),
 [1:25F](#), [1:181](#), [4:3](#) [6](#)
 continuous wave pumped supercontinua
 [2:431–432](#)
 femtosecond pulse pumped supercontinua
 [2:428–429](#)
 modeling [2:425–426](#)
 in optical fiber [2:426–427](#), [2:426F](#)
 picosecond pulse pumped supercontinua
 [2:429–431](#)
 solitons and photonic crystal fiber
 [2:427–428](#)
 spectral broadening and early
 supercontinuum sources [2:424–425](#)
 wavelength extension [2:432–434](#)
 supercritical angle emission [5:171](#)
 supercritical angle fluorescence (SAF) [5:176](#),
 [5:178–180](#), [5:178](#), [5:180F](#)
 superfluorescence (SF) [2:63](#), [2:79](#), [2:80F](#)
 from high-density 2D magnetoexcitons
 [2:79–81](#)
 superlens [3:145](#)
 superluminescent diode (SLD) [2:82](#), [4:507](#)
 superposition
 integral [3:263](#)
 principle [1:461–463](#), [1:486–487](#)
 superprism structures [1:147](#)
 superradiance (SR) [2:79](#), [4:266–267](#), [4:266](#)
 superresolution [3:137–138](#), [3:144–145](#), [5:16](#)
 depth measurements
 metal-induced energy transfer
 microscopy [5:180–183](#)
 super-critical angle fluorescence
 microscopy [5:178–180](#)
 variable-angle total internal reflection
 fluorescence microscopy [5:176–178](#)
 imaging [3:171–173](#), [3:172F](#)
 microscopy [5:175–176](#)
 superstructure grating (SSG) [4:502](#)
 surface attached catalysts [2:172–173](#)
 surface plasmon polariton (SPP) [2:260–261](#),
 [5:247](#)
 surface ridge *see* shallow ridge waveguide
 surface second-order spectroscopy
 [2:246–247](#)
 surface tensions [3:80–81](#)
 surface-plasmon waveguide (SP waveguide)
 [1:383](#)
 surface-sensitive 2D vibrational spectroscopy
 [2:177–178](#)
 susceptibility tensor elements [2:210–211](#)
 Su-Schrieffer-Heeger model (SSH model)
 [5:223](#)
 sutures [5:47](#)
 swept source (SS) [5:157](#), [5:158F](#)
 swept source optical coherence tomography
 (SS-OCT) [4:436](#), [4:504–505](#), [4:505](#),
 [4:505F](#), [5:142–143](#)
 switch/switching [4:34](#)
 characteristics
 granularity [4:225](#)
 quantitative performance [4:225–226](#)
 scalability [4:225](#)
 core growth [4:230](#)
 curves [1:181–183](#), [1:182F](#)
 symmetric topology [4:234](#)
 symmetry considerations [2:44–45](#)
 symplectic conditions [4:199](#)
 symplectic metric matrix [4:199](#)
 synchronization [4:45](#), [4:46F](#)
 synchronous digital hierarchy (SDH) [4:17](#),
 [4:17F](#), [4:60](#), [4:60F](#)
 synchronous disk mirroring [1:281](#)
 synchronous electrical detection [1:296–297](#)
 synchronous modulation technique
 [1:370–371](#)
 synchronous optical network (SONET) [4:17](#),
 [4:17F](#)
 synchronous transport module-level [1](#)
 (STM-1) [4:17](#), [4:60](#)
 synchronous transport signal-level [1](#) (STS-1)
 [4:17](#)
 Synopsis CodeV [4:210–211](#)
 synthetic aperture imaging lidar (SAIL)
 [4:431](#), [4:439](#)
 synthetic aperture lidar (SAL) [4:407](#), [4:439](#),
 [4:440F](#)
 synthetic aperture radar (SAR) [4:407](#),
 [4:439](#)
 synthetic apertures with spatially incoherent
 objects [3:250–252](#), [3:250F](#), [3:251F](#)
 FZP [3:251–252](#)
 random and uniformly redundant arrays
 [3:252](#)
 synthetic prototype [4:35](#)
 synthetic pupil [4:409](#)
 Sysmex XT-2000i and XT-1800i Automated
 Hematology Analyzers [4:514–515](#),
 [4:515F](#)
 system intensity transfer function (SITF)
 [3:236](#)
 system performance parameters [4:11–12](#)
 system point spread function (SPSF)
 [3:250–251](#)
 system-on-panel (SOP) [3:14](#), [3:27](#)
 systems approach to 3D imaging [3:262–263](#)
- ## T
- tabletop
 lasers [2:324](#)
 ultrahigh-intensity lasers [2:324](#), [2:324F](#)
 tail-biting [4:324](#)
 Talbot effect [3:168](#), [4:166](#)
 tandem connection monitoring (TCM)
 [1:289](#), [1:289F](#)
 tandem OPV devices [5:266–269](#)
 tandem scanning microscopy [5:99–100](#)
 tandem solar cell [5:288–290](#)
 tangential fan [1:118–119](#), [1:120F](#)

- taper region [1:200](#)
 tapered microstructure fiber (TMF)
 [1:199–200, 1:200, 1:201F](#)
 target acquisition model (TA model) [3:235](#)
 target of rapamycin kinase (TOR) [3:108](#)
 Taylor series [3:259, 4:21–22, 5:90](#)
 telecentric architectures [3:36](#)
 telecommunication(s) [1:142, 1:224–225, 1:459, 3:199](#)
 fibers [4:1](#)
 industry [1:355](#)
 networks [1:336](#)
 protocols [4:174](#)
 technology [1:226](#)
 telecommunications energy-efficiency ratings (TEER) [1:285–286, 1:286T](#)
 Teledyne Imaging Sensors [1:426](#)
 Teledyne Optech [4:483](#)
 telephone lines [4:16–17](#)
 teleportation [1:462, 1:479](#)
 telescope [3:192, 3:241](#)
 plane for derivation of Van Cittert–Zernike theorem [3:244F](#)
 resolution [1:83, 1:83F](#)
 TEM00 Gaussian beam [4:348](#)
 temperature [4:469, 4:470–471](#)
 temporal analysis by dispersing pair of light E-fields (TADPOLF) [4:58–59](#)
 temporal coherence [1:34, 1:37–38, 4:151](#)
 temporal color synthesis [3:20–21](#)
 temporal gating [2:237–238](#)
 temporal heterodyne [4:407, 4:413](#)
 temporal profile [2:234](#)
 temporal unwrapping by digital holography [4:148, 4:148F](#)
 temporal variations [4:456](#)
 tensor
 equation [1:131](#)
 quantity [1:139](#)
 terahertz (THz) [1:403, 1:418, 2:33–34, 2:73, 2:118, 2:124, 2:197, 2:198F, 2:201F](#)
 emission spectroscopy [2:138–139](#)
 gas lasers [2:119](#)
 generation [2:121](#)
 lamb-dip spectroscopy [1:396, 1:397F](#)
 to Mid-IR [2:135–139](#)
 molecular gas lasers [1:379–380](#)
 molecular spectroscopy [1:387, 1:387T, 1:389F, 1:390F, 1:391T](#)
 astrophysical molecular spectroscopy [1:396–399](#)
 databases and verification [1:400–401](#)
 influence of THz region on molecular spectroscopy [1:399–401](#)
 laboratory molecular spectroscopy [1:390–391](#)
 rotation vibration spectroscopy [1:389–390](#)
 rotational spectroscopy [1:387–389](#)
 output [2:121–122](#)
 PCA [2:120–121](#)
 photon energy [2:34](#)
 QCLs [1:382, 1:383F](#)
 frequency combs [1:384](#)
 quantum cascade lasers [2:19](#)
 radiation [1:412](#)
 nonlinear generation [1:384–385](#)
 source [2:121](#)
 spectral region [2:19](#)
 spectroscopy [2:19](#)
 THz-driven AC stark effect of intersubband transitions [2:200–203](#)
 THz-TDS [2:26–27](#)
 terahertz detectors [1:418](#)
 see also organic photodetectors
 BiR detectors [1:426](#)
 extrinsic semiconductor detectors [1:423–426](#)
 pair braking photon detectors [1:427–428](#)
 photon detectors [1:418](#)
 for room temperature THz imaging [1:419–422](#)
 semiconductor bolometers [1:426–427](#)
 superconducting HEB and TES detectors [1:428–430](#)
 thermal detectors [1:418–419, 1:419F, 1:420T](#)
 terahertz lasers [1:379](#)
 see also up-conversion lasers
 active regions [1:382–383, 1:382F](#)
 germanium intra-valence band lasers [1:380–381, 1:380F](#)
 nonlinear generation of THz radiation in mid-IR QCLs [1:384–385, 1:385F](#)
 optically pumped molecular gas lasers [1:379–380](#)
 QCLs [1:381–382](#)
 silicon impurity state lasers [1:381](#)
 THz QCL laser frequency combs [1:384](#)
 waveguides and cavities [1:383–384, 1:383F](#)
 terahertz optical asymmetric demultiplexer (TOAD) [4:41, 4:42, 4:43F, 4:44F, 4:64](#)
 terahertz physics of semiconductor heterostructures
 few cycle THz spectroscopy
 of QC heterostructures [2:25–31](#)
 of Semiconductor Quantum Structures [2:19–21](#)
 terahertz two-dimensional spectroscopy [2:150](#)
 terrain contour mapping (TERCOM) [4:450](#)
 Terumo OFDI system [4:507](#)
 testing infrared imagers [3:235–237](#)
 tetracene [5:325–326, 5:325F](#)
 tetrahydrofuran (THF) [5:261–262](#)
 texturing [5:292–293](#)
 theoretical minimum data rate [3:180](#)
 theoretical signal origin [2:152–153](#)
 Thereminvox [1:373](#)
 thermal annealing [5:261](#)
 thermal blooming effect [3:203, 3:203F](#)
 thermal deformation processing [1:213](#)
 thermal density phase grating [2:7](#)
 thermal detectors [1:418–419, 1:419F, 1:420T](#)
 thermal effect [3:197–198](#)
 mitigation [2:444–445](#)
 thermal energies [2:100](#)
 thermal evaporation [1:210–211](#)
 thermal imager [3:232](#)
 thermal lens(ing) [2:6, 2:444](#)
 thermal stability [3:87](#)
 thermal tuning [1:237, 1:260](#)
 thermally activated delayed fluorescence (TADF) [3:65, 5:241, 5:245F](#)
 thermally-tuned SGDBR laser [4:250](#)
 Thermionic emission [5:233](#)
 thermistors [1:427](#)
 thermo-optic switch [4:237–238](#)
 thermodynamic equilibrium [2:329](#)
 thermoelectric coolers (TECs) [1:356](#)
 thermophotovoltaics (TPV) [2:262–263](#)
 thermoplastic polymers [1:210](#)
 thermotropic LC phases [3:8](#)
 theta rotating interferometer [3:252–253](#)
 theta shearing interferometer [3:252, 3:253F](#)
 thick grating models [1:131–132](#)
 thieno[3,4-b]thiophene (TT) [5:263–264](#)
 thieno[3,4-c]pyrrole-4,6-dione (TPD) [5:263–264](#)
 thin disk laser (TDL) [2:407, 2:408F, 2:409F](#)
 see also microchip lasers
 amplified spontaneous emission [2:410–411](#)
 continuous wave [2:411](#)
 Cr doping [2:413](#)
 examples of disks with different coating [2:407F](#)
 gain [2:410](#)
 geometry [2:407F](#)
 Ho doping [2:413](#)
 materials [2:412–413](#)
 mode locking [2:412](#)
 mounting [2:410](#)
 Nd doping [2:413](#)
 performance in different operation modes [2:411](#)
 principle [2:408–410](#)
 pumping schemes [2:409–410, 2:410F](#)
 Q-switch and cavity dumping [2:412](#)
 regenerative and multipass amplification [2:412](#)
 scaling [2:411](#)
 semiconductor [2:413–414](#)
 temperature distribution in rod [2:408F](#)
 Ti doping [2:413](#)
 Yb:YAG [2:412–413](#)
 thin film EL displays (TFEL displays) [3:1](#)
 thin film transistor (TFT) [3:1, 3:5–6, 3:12, 3:27, 3:53, 4:99, 5:239](#)
 a-Si:H thin film and a-Si:H [3:13](#)
 AOS thin film and AOS [3:14–15](#)
 devices
 with bottom-gate ES [3:12F](#)
 with different semiconductor layers [3:12T](#)
 driver integrated circuit [3:29–31, 3:29F](#)
 LCD panel [3:5F](#)
 organic thin film and OTFT [3:15](#)
 poly-Si thin film and poly-Si [3:13–14](#)
 thin film transistor liquid crystal display (TFT-LCD) [3:25](#)
 thin grating [1:131–132](#)
 thin lenses, approximations for [1:109–110](#)
 thin phase holograms [4:154](#)
 thin-element approximation (TEA) [1:53, 2:334](#)
 thin-film transistor (TFT) [3:33](#)
 thin-film WDM filter [1:266F](#)

- thin-films filters (TFE) [1:266](#)
- third generation Advanced MVA III (AMVA III) [3:61](#)
- third harmonic (TH) [2:294](#)
- third harmonic generation imaging (THG imaging) [1:180](#), [2:218](#), [2:446–447](#), [4:313](#), [5:98](#), [5:98F](#)
- third order dispersion (TOD) [2:297](#)
- third-order nonlinear effects [1:16–18](#)
- FWM [1:21–24](#), [1:24F](#)
- propagation of pulse [1:18–19](#)
- SC generation [1:24–25](#)
- spectral broadening by SPM [1:19](#), [1:20F](#)
- SPM [1:16–18](#)
- XPM [1:19–21](#), [1:21F](#), [1:22F](#)
- polarization [4:313–314](#)
- susceptibility [4:311](#)
- third-order nonlinearly chirped FBG [4:31](#)
- 3rd order optical effects [1:491](#)
- third-order spectroscopy [2:247–248](#)
- third-order vibrational response functions [2:166](#)
- 32x32 pixel FPA lidar breadboard
- 32x32 pixel laboratory breadboard hardware [5:8–10](#)
- using incoherent chirped amplitude modulation [5:6–7](#)
- incoherent FM/CW FPA breadboard architecture [5:7–8](#)
- laboratory lidar breadboard block diagram [5:8F](#)
- MSM detector design and mounting scheme [5:10](#)
- results of laboratory breadboard tests [5:10–13](#), [5:12F](#)
- Thomas effect [1:98](#)
- Thompson and Wolf experiment [1:39–40](#), [1:39F](#)
- three beam spectral-domain optical low-coherence tomography [4:526](#), [4:527F](#)
- three phase-locked THz pulses [2:199–200](#)
- [360](#) degrees holographic display [4:119–120](#)
- three transmitter and five aperture array [4:411](#), [4:411F](#)
- three-beam lidar [4:363](#)
- three-channel light-field synthesizer [2:314](#), [2:315F](#)
- three-dimension (3D) [3:267](#), [5:155](#), [5:162F](#)
- case [2:64–66](#)
- CDI technique [3:149](#)
- completeness in [4:404–405](#), [4:405F](#)
- density in [4:404–405](#)
- IDTD model [5:301](#)
- field transformations
- diffraction calculation of 3D PSF [3:264–265](#), [3:265F](#), [3:266F](#)
- geometrical optics transformations [3:262](#)
- PSF for coherent, incoherent, and partially coherent illumination [3:263](#)
- spatial frequency content of 3D PSF [3:263–264](#), [3:264F](#), [3:265F](#)
- systems approach to 3D imaging [3:262–263](#)
- images [4:106](#), [4:123–124](#), [5:185](#)
- systems approach to [3:262–263](#)
- laser microvision [4:502](#), [4:502F](#), [4:503F](#)
- lidar scanner [4:439](#)
- measurement [4:496](#)
- photonic crystals [1:175](#)
- PSF
- diffraction calculation of [3:264–265](#), [3:265F](#), [3:266F](#)
- spatial frequency content of [3:263–264](#), [3:264F](#), [3:265F](#)
- random wave system [4:278](#)
- recovery problems [3:149](#)
- resolution of far fields [3:138–140](#)
- scattering process [3:149](#)
- vector [5:110](#)
- three-dimensional coherent spectroscopy (3DCS) [2:53](#), [2:156](#)
- quantum pathways and higher-order spectroscopies [2:155–156](#)
- three-level lasers [2:417](#)
- three-pulse
- transient FWM [2:53](#)
- 2D THz spectroscopy on InSb [2:203–205](#), [2:203F](#)
- three-step model of HHG [2:235–236](#)
- three-wave mixing (ThWM) [2:121–122](#), [4:325](#)
- 3D displays [3:44](#)
- see also wearable displays
- augmented and virtual reality display [3:48–49](#)
- classification [3:47F](#)
- health issue—visual fatigue [3:45](#)
- physiological depth cue [3:44–45](#)
- principle [3:44](#)
- psychological depth cue [3:44](#), [3:44F](#)
- stereoscopic and auto-stereoscopic display [3:45](#)
- 3D Elevation Program (3DEP) [4:401](#)
- 3D micro-electro-mechanical optical switch (3D-MEMS) [4:231–232](#), [4:232F](#)
- threshold
- energy [2:116](#)
- fluence [4:349](#)
- irradiance [4:349](#)
- threshold signal-to-noise ratio (SNR_{th}) [3:234](#)
- threshold voltage (V_{th}) [3:14](#)
- thresholdless laser [1:456](#)
- thromboxane [3:111](#)
- throughput SFR [3:182](#)
- throw ratio [3:38](#), [3:38F](#)
- thulium [2:413](#)
- doping [2:413](#)
- thulium-doped amplifiers [1:263](#)
- THz time-domain spectroscopy (THz-TDS) [2:136](#)
- THz-air-biased-coherent-detection (THz-ABCD) [1:414–415](#)
- Tisapphite
- crystal [2:290](#), [5:41](#)
- laser [4:347](#), [4:364](#), [4:365F](#)
- tight cables [1:334](#)
- tight jacket [1:328](#), [1:329F](#)
- tilted pulse front [1:409](#), [1:410F](#)
- tilting plate compensators [1:156–157](#)
- time [2:184](#)
- coherence [1:34](#), [1:34F](#)
- diffraction limit [4:357](#)
- dimension in velocity measurement [4:526](#)
- domain [3:87](#)
- electric laser field in [2:234](#)
- expressions [2:55](#)
- instrument [3:259](#)
- methods [5:39](#)
- model and pattern effects [1:247–248](#)
- gating [3:223–225](#)
- methods [5:40–41](#)
- resolved optical calorimetry [2:5–7](#)
- resolved photoemission spectroscopy [2:144–145](#), [2:145F](#)
- time-altitude diagrams [4:467](#), [4:468F](#)
- time-averaged intensity distribution [4:174](#)
- time-dependent
- diffusion approximation to transport equation [3:222](#)
- electric field [2:234](#)
- perturbation theory [2:53](#)
- variations of pulse [4:48](#)
- time-to-frequency mapping [4:434](#), [4:434F](#)
- time-bandwidth product [1:135–136](#)
- time and wavelength division multiplexed PON (TWDM-PON) [2:74](#), [2:78–79](#)
- time correlated single photon counting (TCSPC) [4:475](#)
- time division multiple access (TDMA) [1:255](#), [2:73](#), [2:74](#), [2:77](#)
- time division multiplexing (TDM) [1:226](#), [1:255](#), [1:428](#), [2:73](#), [2:77](#), [4:7](#), [4:34](#), [4:45](#), [4:60–61](#)
- see also wavelength division multiplexing (WDM)
- T1 TDM [1:226](#)
- T2 TDM [1:226](#)
- T3 TDM [1:226](#)
- time-dependent Fresnel zone plate (TDFZP) [1:377–378](#)
- Time-Division Multiplexed Passive Optical Networks (TDM-PONs) [2:74](#), [2:78](#)
- BPONs [2:74](#), [2:75F](#)
- comparison of standardized [2:78F](#)
- GE-PON [2:76](#)
- GPONs [2:74–76](#), [2:75F](#)
- time-domain optical low-coherence tomography (TD-OCT) [4:504](#), [4:504F](#), [5:155–156](#)
- time-domain spectroscopy (TDS) [1:418](#), [2:19](#)
- time-domain system (TDS) [2:25–26](#)
- time-frequency
- relationship in chirp signal [4:499–500](#), [4:499F](#)
- representation [4:51–52](#)
- time-integrated optical spectroscopy of correlated electron materials [2:130–131](#)
- Drude-Lorentz model [2:131–132](#)
- general features of optical response in correlated electron materials [2:131](#)
- time-multiplexed type [3:47–48](#), [3:62](#)
- time-of-flight (TOF) [2:322](#)
- camera sensor [4:510–511](#), [4:511F](#)
- measurement [4:502–503](#)
- principle with matricial detectors [4:510–511](#)

- time-of-flight (TOF) (*continued*)
time-resolved fluorescence spectroscopy [4:518–519](#), [4:519F](#)
time-resolved HHG [2:36](#), [2:36F](#), [2:37F](#)
time-resolved optical spectroscopy of correlated electron materials [2:132–135](#)
all-optical pump–probe spectroscopy of correlated electron materials [2:133–135](#)
angle-resolved photoemission spectroscopy [2:144–145](#)
mode-selective excitation [2:139–141](#), [2:141F](#)
PIPT [2:139–141](#)
UED [2:143–144](#)
ultrafast extreme UV and X-ray diffraction and spectroscopy [2:141–143](#)
ultrafast infrared spectroscopy [2:135–139](#)
ultrafast microscopy [2:146–149](#)
ultrafast second harmonic generation spectroscopy [2:145–146](#)
time-resolved photoemission spectroscopy [2:144–145](#), [2:145F](#)
time-resolved THz spectroscopy [2:21](#)
time-resolved SHG (trSHG) [2:146](#)
timing controller (T-Con) [3:29–30](#)
tin dioxide (SnO₂) [5:288](#)
tin-doped indium oxide (ITO) [5:273](#)
tip/tilt correction [3:202](#)
TIR-Fluorescence Correlation Spectroscopy (TIR-FCS) [5:172](#)
TIR-Fluorescence Recovery After Photobleaching (TIR-FRAP) [5:172](#)
tissue damage, PBM for [3:117–118](#)
tissue inhibitors of metalloproteinases (TIMPs) [3:118–119](#)
titania (TiO₂) [1:28](#), [5:288](#)
Tokyo university of agriculture and technology (TAUT) [4:118](#), [4:118F](#)
toll-like receptors (TLRs) [3:112](#)
tomographic/tomography analysis [4:177](#)
examples based on laser profiling [4:483–489](#)
modalities [4:174](#)
reconstruction algorithm [4:174](#), [4:176](#)
techniques [4:174](#)
toneburst modulation (TM) [1:392–393](#)
top cladding [1:186](#)
top emission OLED (TOLED) [3:5](#)
top gate structure (TG structure) [3:12–13](#)
Topcon TRC-50DX mydriatic retinal camera [4:521–522](#), [4:521F](#)
Topcon's KR-1W [4:531](#)
topological charge distribution [4:186–187](#)
topology, deterministic [4:234](#)
toricity [5:44–45](#)
total depth (TD) [4:507](#)
total dispersion [4:1–2](#), [4:2F](#)
total internal reflection (TIR) [1:124–125](#), [1:124F](#), [4:124](#), [5:247](#)
combined with FRAP or FCS or FLIM [5:172](#)
polarization [5:169](#)
reflected beam in [5:170–171](#)
scattering [5:171–172](#)
total internal reflection fluorescence (TIRF) [5:176](#), [5:176F](#)
on single molecules [5:172–173](#)
total internal reflection fluorescence microscopy (TIRFM) [5:168](#)
lateral resolution [5:168–169](#)
qualitative basic principles [5:168](#)
reflected beam in TIR [5:170–171](#)
supercritical angle emission [5:171](#)
two photon TIRF [5:170](#)
variable polar angle illumination, including subcritical [5:169–170](#)
total isolation [1:266](#)
total kinetic energy [2:116](#)
total noise factor (F_{total}) [1:249](#)
Toyozawa lineshape [2:48–49](#)
tracking control [5:32–33](#)
tracking micro-lidars [4:507–509](#)
traditional analysis interferometry (TINTY) [1:318](#)
traditional damage tests [4:306–307](#)
traditional double-clad fibers [1:201](#), [1:201–203](#)
traditional dye lasers [2:336](#)
traditional fiber [1:195](#), [1:201](#)
traditional light trapping [5:294](#)
traditional monovision [5:125](#)
traditional OCT (TD-OCT) *see* time-domain optical low-coherence tomography (TD-OCT)
traditional wide-field conventional microscope [5:194](#)
trailer effect [3:72](#)
transcranial Doppler ultrasound (TCD) [3:126–127](#)
transfer function [1:2](#)
of MZM [1:261](#)
transfer matrix method (TMM) [1:213](#), [5:303](#)
transflective LCD [3:6F](#), [3:7](#)
transform-limited (TL) [2:295](#)
pulse duration [2:295](#)
single-mode laser pulse [4:344](#)
transformation equations [4:182](#)
transformation optics (TO) [2:252–253](#), [2:263–264](#)
transient absorption changes [4:270–271](#), [4:272F](#), [4:273F](#)
transient density phase grating technique [2:4–5](#), [2:7](#)
transient grating (TG) [2:2F](#), [2:189](#)
principle [2:1F](#)
setup with pump–probe detection [2:3F](#)
signal [2:189](#)
techniques [2:10](#)
transient holographic grating techniques in chemical dynamics applications [2:3–5](#)
beam geometry for transient grating [2:4F](#)
classification of possible contributions to transient grating signal [2:2F](#)
determination of material properties [2:7](#)
experimental setup [2:3](#)
investigation of population dynamics [2:7–8](#)
polarization selective transient grating [2:8–10](#)
principle [2:1–3](#)
principle of transient grating technique [2:1F](#)
time resolved optical calorimetry [2:5–7](#)
transient density phase grating technique [2:4–5](#)
transient grating setup with pump–probe detection [2:3F](#)
transient photoluminescence (TRPL) [5:284](#)
transient receptor potential Ankyrin (TRPA) [3:116](#)
transient receptor potential channels (TRP channels) [3:116](#)
transient receptor potential Melastatin (TRPM) [3:116](#)
transient receptor potential Mucolipin (TRPML) [3:116](#)
transient receptor potential NOMP (TRPN) [3:116](#)
transient receptor potential Polycystin (TRPP) [3:116](#)
transient receptor potential Vanilloid (TRPV) [3:116](#)
transillumination [3:225–226](#)
transimpedance amplifier (TIA) [4:6](#), [4:423](#), [5:8](#)
transistors [2:119](#), [5:317](#)
transition amplitude [1:489–490](#)
transition metal (TM) [2:125](#), [2:436](#)
transition metal dichalcogenides (TMDs) [2:57](#), [2:69–71](#)
optical 2DCS [2:53–56](#)
optical properties of monolayer [2:57–58](#)
2DCS of monolayer [2:58–59](#)
transition metal laser advantages and issues [2:441–442](#)
concentration effects [2:441](#)
ESA [2:442](#)
mitigating thermal effects [2:444–445](#)
nonradiative relaxation [2:442–444](#)
thermal lensing [2:444](#)
transition metal oxides (TMOs) [2:124](#), [2:124–126](#)
ferroelectrics [2:128](#)
high-T_c superconductors [2:126](#)
magnetically ordered oxides [2:127–128](#)
MF multiferroics [2:128–129](#)
oxide heterostructures [2:129](#)
vanadium dioxide and related materials [2:126–127](#)
transition metal spectroscopy
crystal field theory [2:436–437](#), [2:436F](#)
SCC model [2:437–440](#), [2:438F](#)
transition process [1:224](#), [1:486–487](#), [2:106](#)
transition-edge sensor (TES) [1:428](#), [1:430](#)
superconducting TES detectors [1:428–430](#)
transition-edge superconducting bolometers (TES bolometers) [1:428](#)
Translating-Wedge-Based Identical Pulses eNcoding System (TWINS) [2:189](#)
transmembrane domain of M2 channel (M2TM) [2:176](#)
transmission [1:153–154](#), [1:172](#), [2:85](#)
effects in optical fibers [1:255–256](#)
linear [1:255–256](#)
nonlinear fiber [1:257–259](#)

- fiber [1:195](#), [1:351](#), [4:21–22](#), [4:30–31](#)
 function [1:76](#)
 continuous part of phase function [4:187](#)
 gradient of singular phase function [4:187–188](#)
 rotation transform [4:188](#)
 topological charge distribution [4:186–187](#)
 holograms [4:87](#)
 of OTDM signal over fiber [4:63](#)
 polarizers [1:153–154](#)
 through thin object [4:167–168](#)
 transmission computerized tomography (TCT) *see* X-ray tomography
 transmission electron microscopy (TEM) [1:468](#), [3:120–121](#), [5:192](#), [5:298–300](#)
 transmissive displays [3:79](#)
 transmissive liquid crystal displays panels (transmissive LCDs panels) [3:6F](#), [3:32](#)
 transmittance of polarizer [1:149](#)
 transmitted near field technique (TNF) [1:322](#)
 image technique [1:323–324](#)
 transmitter circuitry [4:254–255](#)
 see also receiver circuitry
 electroabsorption modulator transmitters [4:255–257](#)
 MZMs transmitters [4:258–259](#)
 ring resonator modulator transmitters [4:257–258](#)
 VCSEL transmitters [4:254–255](#)
 transmitter optical sub-assemblies (TOSAs) [4:252–253](#)
 transmitters
 circuits [4:222](#)
 requirements [4:68](#)
 WDM [1:260](#)
 transoceanic fiber optic cable systems [4:18](#)
 transparency [2:351](#)
 transparent conducting oxides (TCOs) [2:257–258](#), [2:257](#), [5:279](#)
 transparent cornea [5:43](#)
 transparent electrode [3:26](#)
 Transparent OLED (TOLED) [3:66](#), [3:67F](#), [5:252](#)
 transparent polymers [2:346](#)
 transport derivation of intensity equation [4:166](#)
 transport equation (TE) [3:257](#)
 polarization [1:55](#)
 transport of intensity equation (TIE) [3:151](#)
 transpupillary thermotherapy (TTT) [5:134](#)
 transverse aberration [5:55](#), [5:56F](#)
 transverse Anderson localization of light [4:278–279](#)
 in theory and experiment [4:279–281](#)
 transverse Cartesian coordinate [4:188](#)
 transverse chromatic aberration [5:59](#), [5:91](#)
 transverse coherence length [1:35–36](#)
 transverse electric (TE) [1:144](#), [1:173](#), [2:385](#)
 emission spectra [2:275](#)
 modes [1:207](#), [1:242](#)
 waves [1:185](#)
 transverse electromagnetic modes (TEM modes) [4:356](#), [4:356F](#)
 transverse image, intensity distribution on [5:19–22](#)
 transverse magnetic (TM) [1:144](#), [1:173](#), [2:385](#)
 emission spectra [2:275](#)
 modes [1:207](#), [1:242](#)
 waves [1:185](#)
 transverse mode(s) [2:329–331](#)
 definition [2:417](#)
 structure [1:178](#)
 transverse overlapping [4:322–323](#)
 transverse resolution [3:142](#), [5:158](#)
 transverse spherical aberration (TSA) [1:114–116](#), [1:115F](#)
 transverse translation-based techniques [3:151–153](#)
 alternating projection algorithms [3:152–153](#)
 illumination angle scanning-based Fourier ptychography [3:154–155](#)
 mode multiplexing algorithms [3:153](#)
 multi-slice algorithms [3:153–154](#)
 transverse waveguide component [2:351](#)
 transversely excited atmospheric pressure (TEA pressure) [2:334](#)
 trapping
 schemes [5:294](#)
 states [1:454](#), [1:455F](#)
 trapping/cooling laser for mercury atoms [2:448–450](#)
 traumatic brain injury (TBI) [3:123–124](#)
 traveling wave [4:98](#)
 laser [2:453](#)
 OCE [5:147](#)
 SOAs [1:242](#)
 triacetyl cellulose (TAC) [3:25–26](#)
 triangle orientation discrimination (TOD) [3:235](#)
 triangular grating [1:51F](#), [1:52](#)
 triangular lattice [1:190–191](#), [1:196](#)
 triazole (TAZ) [5:234](#)
 trifocal intraocular lenses [5:123](#)
 trigonometric functions [5:16](#)
 trimethylaluminum (TMA) [2:273](#)
 trimethylgallium (TMG) [2:273](#)
 Trinitron CRT [3:3](#)
 trion resonance spectrum [2:62](#)
 triple-beam
 configuration [4:533](#), [4:534F](#)
 spectral-domain OCT [4:526–528](#)
 triplet-triplet annihilation (TTA) [3:64](#), [5:242](#), [5:245F](#)
 tristimulus [3:17](#)
 trunk lines [1:226](#)
 TSI 3563 nephelometer [4:513](#), [4:513F](#)
 tuberculosis [3:102–103](#)
 tumor necrosis factor alpha (TNF- α) [3:118–119](#)
 tumor rejection antigen [3:113](#)
 tumor to normal tissue ratio [3:110](#)
 tumor-associated macrophages (TAMs) [3:110–111](#)
 tumors direct destruction by PDT [3:111](#)
 tunability [4:28](#), [4:28F](#), [4:29F](#)
 tunable attenuators [1:342](#)
 tunable device [1:203](#), [1:204F](#)
 tunable dispersion compensation
 see also fixed dispersion compensation
 approaches to [4:28–30](#), [4:30F](#), [4:31F](#)
 need for tunability [4:28](#), [4:28F](#), [4:29F](#)
 'tunable Fabry-Perot' fibers [1:213–214](#)
 tunable laser (TL) [1:260](#), [4:500](#)
 optical switch technologies [4:241](#)
 tunable microstructure fiber devices [1:203–205](#)
 tunable ONU transceivers (TRXs) [2:79](#)
 tunable thin-film filter (TTF) [2:455](#)
 tunable WDM filters [1:268](#)
 Tungstates [2:266](#)
 tuning range [2:296](#)
 tunnelling
 process [1:469](#)
 and transport dynamics [2:85–86](#)
 turbine compressors [2:333](#), [2:334](#)
 turn around point (TAP) [1:234](#)
 'Twin Janus' *see* Janus Geminus
 twin-guide technology (TG technology) [4:244](#)
 twist nematic liquid crystal displays (TN LCDs) [3:59](#), [3:72](#)
 'twisted-mode' laser [2:453](#)
 two color beams
 coherence control of carrier density, spin population, and spin current [1:490–491](#)
 coherence control of electrical current [1:488–490](#)
 two dimension (2D) [2:197](#), [4:177–178](#)
 approach [1:191–192](#)
 ATR spectroscopy [2:178](#)
 autocorrelation [4:375](#)
 case [2:66–68](#), [2:68F](#)
 characterization [4:177–178](#)
 coherent spectra [2:150](#)
 convolution [3:263](#), [4:375](#), [4:386](#)
 DFT [3:150](#)
 distribution of particle velocities [4:528](#), [4:529F](#)
 EV spectroscopy [2:179](#)
 Fourier model [3:149](#)
 Fourier transform spectroscopy [4:276](#)
 fractional Fourier transform [4:157](#)
 hole crystal [1:174](#)
 imaging supplemented by scanning in z direction [4:445–449](#)
 linear canonical integral transform [4:175](#)
 magnetoexcitons [2:78–79](#)
 materials [4:284](#), [4:284F](#)
 NMR spectroscopy [2:164](#)
 phase retrieval [4:52](#)
 problem [4:54](#)
 photonic crystals [1:173](#), [1:185](#), [1:190–191](#), [1:384](#)
 plasmonic materials [2:261](#)
 PST for [4:177–178](#)
 ranging supplemented by two-dimensional scanning [4:414–445](#)
 SFG spectroscopy [2:178](#)
 SHG for characterization [4:284–285](#)
 simultaneous differential equations [4:201](#)
 THz spectroscopy with three phase-locked THz pulses [2:199–200](#), [2:200F](#)

- two dimension (2D) (*continued*)
 time-frequency function [4:164](#)
 two-dimensionally periodic gratings [1:52](#)
 unbounded disordered systems [4:278](#)
 unit matrix [4:210](#)
 VE spectroscopy [2:179](#)
 Walsh functions [5:17](#)
 in polar coordinates [5:18–19](#), [5:19F](#)
 in rectangular coordinates [5:18](#)
- 2 f-imaging** [2:409](#)
- two photon TIRF [5:170](#)
- two sine modulation signals [4:499](#), [4:499F](#)
- two telescopes interferometer of Labeyrie [3:254](#)
- two-axis gimbal-less MEMS mirror [4:528](#)
- two-color method [1:412–416](#)
 peak THz electric field strength ranges [1:416F](#)
- phase scan [1:414F](#)
- THz
 emission from along, two-color laser-induced filament [1:415F](#)
 field as function of laser intensities [1:413F](#)
 wave generation efficiency [1:413F](#)
- two-color rapid acquisition coherence spectroscopy (T-RACS) [2:192](#)
- two-dimensional coherent spectroscopy (2DCS) [2:53](#), [2:150](#)
 experimental implementation [2:56–57](#)
 of monolayer TMDs [2:58–59](#)
 optical [2:53–56](#)
 optical properties of monolayer TMDs [2:57–58](#)
- polarization control of quantum pathways in [2:154](#)
- pulse sequencing to control quantum pathways in [2:154](#)
 double quantum 2D spectra [2:155](#), [2:155F](#)
 nonrephasing 2D spectra [2:154–155](#)
 rephasing 2D spectra [2:154](#), [2:155F](#)
- two-dimensional infrared spectroscopy (2D IR spectroscopy) [2:150](#), [2:164](#)
 applications [2:168](#)
 brief theory [2:164–166](#)
 broadband light sources [2:178](#)
 delay-lines and sample chamber [2:166–167](#)
 dynamics of water [2:176–177](#)
 experimental approaches [2:166](#)
 material sciences [2:168](#)
 microscopy [2:178–179](#)
 perspective [2:177–178](#)
 protein structure and folding dynamics [2:173–175](#)
 schematic 2D IR pulse sequence and spectra [2:165F](#)
 signal detection systems [2:167–168](#)
 surface- and interface-sensitive 2D vibrational spectroscopy [2:177–178](#)
- 2D EV spectroscopy [2:179](#)
- 2D VE spectroscopy [2:179](#)
- ultrafast IR light source [2:166](#)
- two-directional velocity measurement [4:525–526](#)
- “two-faced” or “lanus effect” [3:100](#)
- two-grating compressor [4:179](#)
- two-level
 model [2:89–90](#)
 quantum systems [1:478](#)
- two-path interference [5:41](#)
- two-photon absorption (2PA) [1:488](#), [2:267](#), [4:319–320](#), [4:320](#), [5:39](#), [5:89](#)
- two-photon excitation (TPE) [5:88](#), [5:98F](#)
 artifacts and implementation of TPE fluorescence ophthalmoscopy [5:89–91](#)
 microscopy [5:199](#)
- two-photon excited fluorescence (TPEF) [5:98](#)
 ophthalmoscopy [5:93](#)
- two-photon fluorescence microscopy
 method [2:249](#)
- two-photon imaging [1:484](#)
 of *Acanthamoeba keratitis* [5:104F](#)
 of corneal cross-linking [5:105–106](#)
 of human cornea [5:105F](#)
 of infectious keratitis [5:103–104](#)
 of limbal stem cells [5:103](#)
 of mice limbal stem cells [5:104F](#)
 of normal cornea [5:81F](#), [5:83](#)
- two-photon microscopy [5:75F](#), [5:80F](#)
 of enucleated rabbit eyes [5:106F](#)
 laser safety considerations [5:106–107](#)
 laser scanning microscopy [5:98–99](#), [5:102](#)
- two-photon transitions
 attosecond heating reconstruction by interference of [2:241–242](#)
 excitation of [2:248](#)
- two-photonexcited fluorescence microscopy (TPEF microscopy) [5:85](#)
- two-photonexcited fluorescence ophthalmoscopy (TPEFO) [5:86–87](#)
- two-point
 correlation function [4:174](#)
 resolution [5:186](#)
- two-pulse transient FWM [2:53](#)
- two-quantum spectrum [2:56](#)
- 2D electron gas systems (2DEG systems) [2:261](#)
- 2D electronic spectroscopy (2DES) [2:184](#), [2:185F](#), [2:187–188T](#)
see also attosecond spectroscopy;
 multidimensional THz spectroscopy
- applications [2:190–191](#)
 BOXCARS geometry [2:185–188](#)
 challenges [2:184–185](#)
 charge transfer in OPVs [2:194](#)
 coherence [2:192–193](#)
 early 2D studies [2:190–191](#)
 excitonic electronic structure and dynamics in I-aggregates [2:193–194](#)
 experimental implementations [2:185–188](#)
 fully collinear geometry [2:189–190](#)
 noise suppression [2:190](#)
 photosynthesis [2:191–192](#)
 single-shot measurements [2:190](#)
 solid state systems [2:194–196](#)
- 2D infrared (2DIR) [2:184](#)
 microscopy [2:178–179](#)
 spectra of Q24 [2:175](#), [2:175F](#)
- 2D micro-electro-mechanical optical switch (2D-MEMS) [4:232–233](#), [4:232F](#)
- 2D spectra/spectroscopy [2:150](#)
 connecting 2D and 1D spectra [2:161–162](#), [2:162F](#)
 of inhomogeneously-broadened systems, interpretation [2:158–160](#)
 measuring homogeneous and inhomogeneous linewidths [2:160](#)
 spectral diffusion [2:160](#), [2:160F](#)
 interpretation [2:157](#)
 anharmonic oscillator spectra interpretation [2:160–161](#), [2:161F](#)
 connecting 2D and 1D spectra [2:161–162](#), [2:162F](#)
 of coupled systems [2:157–158](#)
 homogeneous lineshapes [2:157](#)
 resolution in [2:157](#)
- 2D ultraviolet (2DUV) [2:184](#)
- type-I multiferroicity [2:129](#)
- type-I reaction [3:101](#), [3:102](#)
- type-II multiferroics [2:129](#)
- type-II reaction [3:101](#), [3:102](#)
- typical ICoS pixel pitch [4:99](#)

U

- ultra violet sources (UV sources) [1:234–235](#)
- ultra-broadband
 non-zero dispersion-shifted single-mode fibers [1:261](#), [1:262](#)
 OPAs [2:298–299](#)
- ultra-high performance (UHP) [3:40–41](#)
- ultra-realistic holographic imaging techniques [4:105](#)
- ultra-short-throw (UST) [3:37–39](#)
- US Patent 20130162957 A1 [3:39F](#)
- US Patent 7,009,765B2 [3:38F](#)
- US Patent 7441908 [3:39F](#)
- US Patent 9,372,388 B2 [3:40F](#)
- ultrafast and intense-field nonlinear optics
 high-harmonic generation [4:337](#)
 multiphoton absorption [4:336](#)
 nonlinear quantum electrodynamics [4:338](#)
 nonlinear refractive index [4:335–336](#)
 optical damage [4:336–337](#)
 plasma nonlinearities and relativistic effects [4:337–338](#)
 self focusing, supercontinuum generation, and filamentation [4:336](#)
- ultrafast control of magnetization dynamics [2:141](#)
- ultrafast electron diffraction (UED) [2:143–144](#)
- ultrafast extreme UV and X-ray diffraction and spectroscopy [2:141–143](#)
- ultrafast femtosecond pulse [1:405](#)
- ultrafast infrared spectroscopy [2:135–139](#)
- ultrafast laser pulses [4:48](#), [4:348](#)
- ultrafast microscopy [2:146–149](#)
- ultrafast mid-IR spectroscopy [2:135](#)
- ultrafast optical fiber lasers [2:121](#)
- ultrafast optical spectroscopy (UOS) [2:124](#), [2:132](#)
- ultrafast second harmonic generation spectroscopy [2:145–146](#)

ultrafast spectroscopy [2:19](#), [2:86](#)
 ultrafast time resolution to ARPES (trARPES) [2:144](#)
 ultrafast time resolved spectroscopy [2:248–249](#)
 ultrahigh Q microcavities [4:298–299](#)
 ultrahigh speed OTDM system [4:38](#)
 ultrahigh vacuum (UHV) [5:266](#)
 UltraQReflex™ system [5:134–135](#)
 ultrasensitive nonlinear phase-shifters [2:18](#)
 ultrashort electron pulses [2:233](#)
 ultrashort femtosecond laser pulses [1:414–415](#)
 ultrashort laser pulse technology [2:233](#)
 ultrashort optical pulse sources [4:48](#), [4:62–63](#)
 ultrashort pulse
 characterization [2:316](#)
 generation [2:283](#)
 lasers [4:348](#), [5:92](#)
 ultrashort-pulse dye lasers [2:340](#), [2:341F](#)
 ultraslow propagation of pulses [4:345](#)
 ultrasound technology [5:155](#)
 ultrathin metal film (UTMF) [5:303–304](#)
 ultraviolet (UV) [2:290](#), [2:373](#), [4:344](#), [5:85](#)
 absorption spectrum [1:235](#), [1:235F](#)
 AlGaIn laser diodes [2:271](#)
 lasers [2:356](#), [2:446](#), [4:364–366](#), [4:366F](#)
 dyes [2:348](#)
 enhancement factors as function of resonator loss [2:449F](#)
 frequency-doubled argon ion laser [4:366](#)
 frequency-doubled krypton ion laser [4:366](#)
 HeCd laser [4:366–367](#)
 nonlinear optical crystals and properties [2:447F](#)
 optical resonator with nonlinear crystal inside [2:448F](#)
 quasi-CW mode-locked Ti:Sapphire laser [4:367](#)
 reflected power at fundamental as function of input coupler [2:449F](#)
 second harmonic output power as function of input coupler [2:449F](#)
 light [5:131](#)
 photo-emission [2:246](#)
 photolithography [5:204](#)
 radiation [5:191](#), [5:191F](#)
 UV-excitable fluorophores [5:86](#)
 UVA [5:136](#)
 UV-visible spectroscopy [2:131](#)
 uncertainty principle [2:435](#)
 uncorrected myopes [5:51](#)
 uncoupling variables [4:162–163](#)
 uncyclic unsaturated hydrocarbons [2:345](#)
 under-critical angle fluorescence (UAF) [5:178](#)
 undersea optical cable systems [4:18](#)
 underwater cables [1:334](#)
 “undoping” process [5:227](#)
 unguided rays [1:222](#), [1:222F](#)
 unidirectional backscattering curves [1:310](#), [1:310F](#)
 uniform and chirped fiber gratings [1:233](#)
 uniform” BHJ structure [5:261](#)

uniformity [3:71–72](#)
 uniformly redundant arrays [3:252](#)
 unit interval (UI) [4:263](#)
 unit vectors [4:451](#), [4:451F](#)
 unitary transform [4:181](#)
 United States Geological Survey (USGS) [4:401](#)
 lidar base specification [4:401](#)
 Universal Lidar Error Model (ULiEM) [4:402](#)
 metadata [4:402–403](#)
 sensor-space error modeling [4:403F](#)
 unmanned aerial vehicle (UAV) [4:449](#), [4:474](#), [4:484F](#)
 unsampled surface fraction (USF) [4:405](#)
 unsaturated regime [1:352](#)
 unscattered photons [3:219](#)
 up conversion mechanisms [2:394](#)
 up-conversion lasers [2:394](#)
 see also noble gas ion lasers
 based on bi-functional crystals [2:403–404](#)
 based on luminescence mechanisms in materials [2:399–400](#)
 energy transfers [2:400](#), [2:401F](#)
 photon avalanche [2:402–403](#), [2:404F](#)
 sequential two-photon absorption [2:400–402](#), [2:401F](#), [2:402F](#), [2:402T](#)
 from luminescence mechanisms [2:394–395](#)
 energy transfers [2:394–395](#), [2:395F](#)
 materials [2:397](#)
 photon avalanche [2:396–397](#), [2:396F](#)
 sequential two-photon absorption [2:395–396](#), [2:396F](#)
 from second-order optical nonlinearity [2:397–399](#), [2:399F](#)
 upper exciton polaritons (UP) [1:448–449](#), [1:449F](#)
 upstream bandwidth map [2:75](#)
 uveal tract [5:43](#)

V

V-shaped cavity [2:285](#)
 Vac-Ion® vacuum pump [2:382](#)
 vacuum fluorescent display (VFD) [3:1](#), [3:3–4](#)
 vacuum Rabi frequency [1:448–449](#)
 vacuum ultraviolet (VUV) [2:359–360](#), [2:360–361](#), [4:344](#)
 valence band maximum (VBM) [5:256](#), [5:284](#), [5:286–287](#)
 valence band(s) [1:171](#), [2:33–34](#)
 electrodes [1:224](#)
 valence electrons [1:224](#)
 valley physics [1:467](#)
 valley superposition [1:467–468](#)
 van Cittert–Zernike theorem [1:38–39](#), [1:38F](#), [3:149](#), [3:245](#)
 van der Waals energy [1:452](#)
 van der Waals epitaxy (VDWE) [4:284](#)
 Vanadates [2:266](#)
 vanadium atoms [2:126–127](#)
 vanadium dioxide (VO₂) [2:126](#)
 and related materials [2:126–127](#)
 vanilla algorithm *see* frequency-resolved optical gating (FROG); inversion algorithms
 vapor-axial deposition process (VAD process) [1:28](#), [1:29–30](#), [1:30F](#)
 vapor-phase oxidation [4:14–15](#)
 variable frequency shifter [1:140](#)
 variable geometry pupil technique (VGP technique) [3:279](#)
 variable length records (VLRs) [4:402–403](#)
 variable polar angle illumination [5:169–170](#)
 variable retardation plates [1:156–159](#)
 variable-angle total internal reflection fluorescence microscopy (vaTIRF) [5:176–178](#), [5:177](#)
 varifocal lens(es) [4:177](#), [4:178–179](#)
 vascular endothelial growth factor (VEGF) [3:120](#), [5:132](#)
 VECSEL cavity geometries [2:285F](#)
 vector addition technique [1:93](#)
 vehicle borne laser radar [4:483](#)
 velocity [4:470–471](#)
 Velodyne Lidar [4:483](#)
 venerable algorithm [3:163](#)
 vertical alignment mode (VA mode) [3:9](#)
 vertical cavity surface emitting lasers (VCSEL) [1:186](#), [1:214](#), [1:227](#), [1:384](#), [1:459](#), [1:459F](#), [2:280](#), [2:281F](#), [2:283F](#), [2:413–414](#), [2:466](#), [2:466F](#), [2:467F](#), [3:88](#), [4:3](#), [4:67](#), [4:254](#), [4:293](#), [4:358](#)
 transmitters [4:254–255](#), [4:255F](#), [4:256F](#)
 very high range resolution lidars [4:431](#)
 applications [4:438–439](#)
 advanced coherent imaging with FMCW lidar [4:439–440](#)
 aspherical optical metrology [4:438–439](#)
 noncontact 3D metrology [4:439](#)
 brief-pulse, direct-detect lidar [4:431](#)
 coherent mode-locked lasers [4:432](#)
 FMCW lidar [4:434–435](#)
 range precision [4:430–431](#)
 range resolution [4:430–431](#)
 streak cameras [4:431–432](#)
 very high resolution FMCW lidar [4:435–436](#)
 active chirp laser stabilization for FMCW lidar [4:436](#)
 chirp nonlinearity [4:435–436](#), [4:435F](#)
 very large array (VLA) [3:254–255](#)
 very large scale integration (VLSI) [1:144–146](#), [2:115](#), [2:387–388](#)
 very large telescope interferometer (VITI) [3:245](#), [3:254–255](#), [3:254–255](#)
 very large telescopes (VLT) [3:253](#)
 very low density lipoprotein (VLDL) [3:110](#)
 vestigial singleside band (VSB) [4:7](#), [4:31–33](#)
 vibrating galvanometer-type mirrors [5:194](#)
 vibrational/vibration [2:328](#), [4:452–455](#)
 coherence [5:40](#)
 curve [1:92–93](#), [1:93F](#)
 energy groupings [2:326](#)
 frequency [2:217](#)
 intensity map [4:453](#), [4:454F](#)
 measurement of vocal folds [4:528–530](#)
 mode coupling [5:39](#)
 relaxation [2:328](#)
 transfer rates [2:328](#)

vibrational/vibration (*continued*)
 video camera [4:107](#)
 Video Electronics Standards Association
 Standards (VESA Standards) [3:70](#)
 video techniques [5:64](#)
 view angles of liquid crystal displays [3:59](#)
 controlled viewing angle of LCDs [3:62](#)
 improvement from LC modes [3:60–61](#)
 latest approaches of wide view angle LCDs
 [3:61–62](#)
 optical compensation film [3:59–60](#)
 viewing angle [3:74–75](#)
 viewing cone [3:74–75](#)
 viewing zone-scanning holographic display
 system [4:118, 4:119F](#)
 violet InGaN laser diodes [2:278](#)
 violet laser dyes [2:348](#)
 violet wavelengths [2:373](#)
 virtual image [3:208–209, 4:107](#)
 virtual reality (VR) [3:48, 3:49F, 4:124](#)
 devices [3:75](#)
 display [3:48–49](#)
 HMD [3:51, 3:51F](#)
 virtual SAI (vSAI) [5:180](#)
 virtually imaged phased array (VIPA) [4:29,](#)
 [4:432, 5:149](#)
 virulence [3:105](#)
 viscoelastic effect [5:145–146](#)
 visibility of interference fringes [1:37](#)
 visible and near-infrared light [2:150](#)
 visible lasers [4:359–360](#)
 argon ion laser [4:360–361](#)
 He-Ne laser [4:359–360](#)
 HeCd laser [4:361](#)
 krypton ion laser [4:361](#)
 visible light [5:86](#)
 “visible region” [3:141](#)
 vision correction methods
 chromatic aberration [5:120](#)
 future of [5:126–127](#)
 presbyopia [5:120–122](#)
 refractive surgery [5:125–126](#)
 spectacles [5:116–119](#)
 visual benefit of correcting higher order
 aberrations [5:119–120](#)
 vision problems, PRM for [3:121](#)
 visual analog scale (VAS) [3:125](#)
 visual benefit of correcting higher order
 aberrations [5:119–120](#)
 visual quality [3:183–185](#)
 visual stimulus [5:64, 5:111](#)
 visual Strehl ratio [5:115](#)
 visual tasks [5:119](#)
 visualization of aberrations
 in freeform systems [4:208–209](#)
 in phase space [4:206–208](#)
 vitreoretinal disorders [5:134–135](#)
 vitreous humor [5:43](#)
 vocal folds, vibration measurement of
 [4:528–530, 4:529F](#)
 voice coil actuator (VCA) [4:507](#)
 voice coil motor (VCM) [3:37](#)
 Voigt profile [2:55–56](#)
 voltage controlled oscillator (VCO) [4:6, 4:45](#)
 voltage standing wave ratio (VSWR) [4:69](#)
 volume Bragg grating (VBC) [2:416](#)

volume diffraction theory [3:267](#)
 volume holographic imaging (VHI) [3:267,](#)
 [3:267F](#)
 volume turbulence [4:413](#)
 volumetric display [3:46](#)
 Voronoi diagram [5:79–80, 5:80F](#)
 voxels [4:404](#)
 vulval intraepithelial neoplasia (VIN) [3:114](#)

W

wall-plug power efficiencies (WPE) [1:382](#)
 Walsh filters [5:27–28](#)
 Walsh functions
 Azimuth invariant case [5:19](#)
 intensity distribution
 along axis around image/focal plane
 [5:22–27](#)
 on neighboring transverse planes [5:27](#)
 on transverse image/focal plane [5:19–22](#)
 one dimensional Walsh functions [5:17–18](#)
 in optical masks [5:16, 5:16–17](#)
 radial Walsh functions [5:19](#)
 two dimensional Walsh functions
 in polar coordinates [5:18–19](#)
 in rectangular coordinates [5:18](#)
 Walsh filters as optical masks [5:19](#)
 Walsh-Hadamard single-pixel
 applications [4:196–197](#)
 imaging [4:195–196, 4:195F](#)
 Wannier excitons [2:93](#)
 Wannier–Mott exciton [5:223–224](#)
 warehouse scale computer (WSC) [1:227,](#)
 [1:227F](#)
 watch-type of display [3:53](#)
 water [5:277](#)
 attosecond pulse generation in water-
 window region [2:321](#)
 cooled ZnSe lens [2:329–331](#)
 dynamics [2:176–177](#)
 role in influenza A2 channel [2:176](#)
 vapor profiles [4:456](#)
 water-soluble dyes [2:349](#)
 wave [4:199](#)
 aberration [5:55, 5:56F, 5:59–60](#)
 equation [4:335](#)
 in nonlinear regime [4:311–312](#)
 field [4:139](#)
 front
 difference [4:410, 4:412F](#)
 errors [4:410](#)
 “reversal” beam [4:322](#)
 optics [4:200](#)
 propagation [1:169–170, 3:142](#)
 equation [5:203](#)
 process [3:139](#)
 theory [1:222](#)
 vector [4:180](#)
 wave plate (WP) [2:211](#)
 wave recording plane technique (WRP
 technique) [4:123](#)
 wave-mixing in Doppler broadened medium
 [4:344](#)
 wave-vector mismatch [2:291](#)
 wavefield [3:205](#)
 wavefront aberrations
 ray aberrations converting to [5:113–115](#)
 time-scale and associated aberrations
 [3:197–198, 3:198F](#)
 wavefront aberrometers [5:111–112](#)
 wavefront distortion [3:200](#)
 wavefront healing [3:207, 3:207F](#)
 wavefront reconstruction [4:531](#)
 wavefront sensing micro-lidars
 [4:530–531](#)
 Hartmann–Shack ocular wavefront sensing
 [4:530–531](#)
 ray tracing aberrometry [4:531–533](#)
 wavefront reconstruction [4:531](#)
 wavefront sensing technique [3:198,](#)
 [5:112–113](#)
 optical path lengths determining from ray
 deflections [5:112–113, 5:113F](#)
 optimizing retinal image quality by
 correcting refractive errors [5:115](#)
 ray aberrations converting to wavefront
 aberrations [5:113–115](#)
 wavefront sensor system (WS system)
 [3:198](#)
 wavefront sensors and control
 hybrid techniques
 adaptive optics and post-processing
 [3:278–279](#)
 PRPM [3:279](#)
 imaging and atmospheric turbulence
 [3:272–274](#)
 post-detection compensation
 [3:274–276](#)
 pre-detection compensation [3:278](#)
 wavefunctions [1:462, 2:66, 2:91, 2:99,](#)
 [2:100–101, 2:102F, 5:204](#)
 waveguides (WG) [1:383–384, 1:383F, 2:384,](#)
 [2:385F, 2:445](#)
 configuration [2:445](#)
 dispersion [1:6–7, 1:313, 1:314F, 4:1–2,](#)
 [4:2F](#)
 group velocity dispersion [1:6–9](#)
 laser materials [2:387–388](#)
 mode [1:190–191, 3:67](#)
 waveguide-MEMS optical switch
 [4:240–241, 4:240F](#)
 waveguiding mode [5:247](#)
 light extraction [5:248–251](#)
 wavelength division multiple access
 (WDMA) [1:255](#)
 wavelength division multiplexed PONs
 (WDM-PONs) [1:252, 1:252F, 2:74,](#)
 [2:81–82, 2:81F](#)
 wavelength division multiplexing (WDM)
 [1:23, 1:143, 1:147, 1:214–215, 1:226,](#)
 [1:237F, 1:245, 1:255, 1:291, 1:336,](#)
 [1:339–340, 1:341–342, 1:341, 1:355,](#)
 [1:356, 1:364, 1:371F, 2:81, 2:455,](#)
 [4:1, 4:17, 4:20–21, 4:20, 4:34, 4:45,](#)
 [4:60, 4:251](#)
 see also time division multiplexing (TDM)
 achievable bit-error rates [1:276–277](#)
 amplifier technologies [1:263F](#)
 chromatic dispersion compensation [1:265](#)
 components and sub-systems [1:259–260](#)

- in data center interconnects [1:281–282](#)
 demultiplexer isolation [1:272F](#)
 development of WDM systems transport capacity over time [1:255F](#)
 entrance [4:17–18](#)
 long-haul WDM systems [1:282–284](#)
 in metro and access networks [1:280–281](#)
 mixed 10 Gb/s/100 Gb/s design [1:281, 1:285T](#)
 modulation
 formats [1:275–276](#)
 of WDM signals [1:275–276](#)
 networks [4:18](#)
 non-fiber-related effects [1:271–272](#)
 optical amplifiers [1:263–265](#)
 optical multiplex section [1:263](#)
 optical transport network [1:287](#)
 passive WDM filters and couplers [1:265–268](#)
 receivers [1:270–271](#)
 ROADMs [1:269–270, 1:269F](#)
 single-mode fibers [1:261–263](#)
 superchannels [1:275](#)
 sustainability [1:284–287](#)
 system realization [1:280–281](#)
 transmission effects in optical fibers [1:255–256](#)
 transmitters [1:260](#)
 external modulators [1:260–261](#)
 LD [1:260](#)
 tunable WDM filters [1:268](#)
 wavelength switching devices [1:268–269](#)
 wavelength-selective switches (WSS) [1:269, 4:225, 4:226–227, 4:231F](#)
 wavelength(s) [1:99, 1:139](#)
 allocation [2:78](#)
 conversion [1:250–251](#)
 extension [2:432–434](#)
 filters [1:341](#)
 of light measurement [1:486](#)
 range [5:160–161](#)
 reuse schemes [2:82](#)
 scalability [1:218](#)
 scale [1:193](#)
 switching devices [1:268–269](#)
 tuning interferometry [2:265, 4:500, 4:501F](#)
 versatility and tenability [2:280](#)
 wavelength-blocking [1:270](#)
 wavelength-scalable hollow optical fibers [1:216–220, 1:220F](#)
 wavelength-scale microstructures [5:313](#)
 wavelength-selective
 branching device [1:341](#)
 directional couplers [1:266](#)
 wavepacket propagation methods [5:39](#)
 WaveScan model [4:531](#)
 WDM-PON Central Office (WDM-PON CO) [1:252](#)
 weak cavity [5:247](#)
 weak-coupling regime [1:460](#)
 wearable displays
 see also 3D displays
 clothes-integratable-type [3:52](#)
 flexible displays integrating
 with clothes [3:52](#)
 with human body [3:52–53](#)
 flexible E-paper [3:56](#)
 flexible liquid crystal display [3:56–57](#)
 flexible organic light emitting diode [3:54–56](#)
 flexible substrates [3:53](#)
 flexible TFT [3:53–54, 3:55F, 3:55T](#)
 flexible transparent conductive film [3:57](#)
 practical processes for flexible display [3:57](#)
 spectacle-type [3:51–52](#)
 watch-type [3:53](#)
 welding [2:329–331, 2:334](#)
 Western Europe (WE) [4:18](#)
 wet-coating processes [5:317](#)
 Weyl-symplectic group [4:202](#)
 whisker technology [1:422](#)
 white light emitting diodes (W-LEDs) [3:19–20, 3:27–28](#)
 “white light” spectrum [2:426](#)
 white-light continuum generation (WLC generation) [2:295](#)
 wide angle (WA) [5:262–263](#)
 schematic eyes [5:48](#)
 wide area network (WAN) [4:17–18](#)
 wide bandgap semiconductors types [2:271](#)
 wide view film (WV film) see optical compensation film
 wide viewing angle (WVA) [3:62](#)
 wideband amplifiers [1:362](#)
 wideband transimpedance amplifiers [4:260–261](#)
 Wien’s displacement law [3:230, 3:231F](#)
 Wiener filter [3:181, 3:184F, 3:185F](#)
 wiggler [2:119](#)
 Wigner distribution (WD) [4:174–175, 4:177F, 4:178F](#)
 parallel data processing for WD reconstruction [4:177–178, 4:177F](#)
 projections [4:176–177](#)
 Wigner distribution function (WDF) [3:164, 3:165, 3:166, 3:167, 4:164, 4:174](#)
 wind and flow velocity [4:456–463, 4:457F, 4:458F](#)
 Wind Infrared Doppler lidar (WIND) [4:457](#)
 Windcube 200S [4:459–461, 4:461F](#)
 Windcube of Doppler lidar [4:459–461](#)
 WindScanner system [4:461–462, 4:462, 4:462F, 4:463](#)
 windshear [4:457, 4:458F, 4:459F](#)
 WindTracer [4:458–459, 4:460F](#)
 WindTracer Terminal Doppler Solution (WTDS) [4:458–459](#)
 wire grid polarizer [1:153](#)
 WKB method [2:91](#)
 WOLED lighting [3:68–69](#)
 wound healing, PBM for [3:117–118](#)
 wurtzite material [1:491](#)
- X**
- X-ray absorption spectroscopy (XAS) [2:142–143](#)
 X-ray diffraction (XRD) [2:124, 2:142–143](#)
 X-ray free electron lasers (XFEL) [2:142](#)
 X-ray magnetic absorption dichroism (XMCD) [2:147](#)
 X-ray(s) [2:207, 5:175, 5:185, 5:191, 5:191F](#)
 analyzer-based imaging [4:171, 4:171F](#)
 crystallography studies [2:174, 3:8](#)
 CT principle [3:192–193](#)
 detector [5:192](#)
 free electron lasers [2:233](#)
 holotomography [4:172, 4:172F](#)
 microanalysis [5:193](#)
 photoelectron spectroscopy [2:246](#)
 tomography [3:187, 3:193F, 3:195](#)
 X-tion Pro Live [4:512](#)
 xanthene [2:344](#)
 xanthophyll [5:50–51](#)
 Xe* [2:360–361, 2:361–362](#)
 Xe₂ [2:360–361, 2:361F](#)
 XeI* [2:361–362](#)
 xenon ion lasers [2:379, 2:379T](#)
 XG-PON [2:77–78, 2:78E, 2:78T](#)
 XG-PON2 [2:77](#)
 XPM-induced intensity interference [1:21](#)
 XPM-induced Nonlinear Phase Noise (X-NLPN) [1:284](#)
- Y**
- Y123 dyes chemical structures [5:270, 5:271F](#)
 Y₃Al₅O₁₂ see yttrium aluminium garnet (YAG)
 Yb:YAG [2:412–413](#)
 YBa₂Cu₃O_{7-x} magnetic quadrupole order in pseudogap of [2:220–223, 2:222F, 2:223F](#)
 Yb–Pr-doped fiber up-conversion laser [2:402–403, 2:403F](#)
 YD2-o-C8 dyes chemical structures [5:270, 5:271F](#)
 Young’s double slit interferometer [1:74–75, 1:75F, 1:486F, 1:487](#)
 Young’s double-hole experiment [1:36–37, 1:36F](#)
 Young’s interference experiment, spectral changes in [1:46, 1:46F](#)
 young’s modulus [1:216](#)
 ytterbium ion [2:413](#)
 ytterbium laser ion (Yb laser ion) [2:408](#)
 yttrium aluminum garnet (YAG) [2:386, 2:412–413, 2:436, 4:363](#)
 Yttrium Iron Garnet (YIG) [1:356](#)
 yttrium lithium fluoride (YLF) [2:436, 4:364](#)
 Yttrium Orthovanadate wedge (VVO)₄ wedge [1:356, 2:436](#)
- Z**
- Zeeman term [2:65](#)
 Zernike aberration coefficients [5:110](#)
 estimating refractive state from Zernike aberration coefficients [5:111–112](#)

- Zernike aberration (*continued*)
 function profile [4:438–439](#)
- Zernike phase contrast [3:146](#)
- Zernike polynomials [4:409](#), [4:411](#), [4:531](#),
 [5:16](#), [5:57](#), [5:110](#), [5:111–112](#)
- Zernike-Van Cittert theorem derivation from
 propagation of AF [4:169](#)
- zero delay [5:157](#)
- zero dispersion wavelength [1:179](#),
 [1:195](#)
- zero-dimension (0D) [2:72](#)
- zero-phonon donor-bound exciton
 transitions [1:467](#)
- zinc oxide-based DSSCs [5:270](#)
- ZnCdSe quantum well, absorption spectrum
 of [2:94](#), [2:95F](#)
- ZnO layer [2:271](#), [5:303](#)
(ZnO:Al) *see* aluminum-doped zinc oxide
 (AZO)
- ZnSe [2:94](#), [2:271](#)
- ZnTe [1:407](#), [1:407F](#)
- zonal presentation [4:531](#)
- zone plate [1:95–96](#), [1:95F](#)
- zonules [5:47](#)